В.И. КРЫЛОВ, В.В. БОБКОВ, П.И. МОНАСТЫРНЫЙ

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

I

I

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

В. И. КРЫЛОВ, В. В. БОБКОВ,
П. И. МОНАСТЫРНЫЙ

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

ТОМ I

Допущено Министерством
высшего и среднего специального образования СССР
в качестве учебного пособия
для студентов высших учебных заведений

ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
МОСКВА 1976

518
К 85
УДК 519.95

Вычислительные методы, том I. В. И. Крылов, В. В. Бобков, П. И. Монастырный. Главная редакция физико-математической литературы изд-ва «Наука», М., 1976.

В книге дано изложение начал теории вычислительных методов математики и приведены наиболее часто применяемые в реальных вычислениях численные методы.

Как учебник книга предназначена для студентов высших технических учебных заведений. Она может быть также пособием для обучающихся на физических и механико-математических факультетах университетов. В качестве справочника книга рассчитана на работников вычислительных центров и лиц, которым приходится иметь дело с научными и техническими расчетами.

Первый том содержит теорию интерполирования, линейную алгебру, решение численных уравнений и численное интегрирование функций.

Илл. 11. Библ. 34 названия.

К $\frac{20204-075}{053(02)-76}$ 12-76

ОГЛАВЛЕНИЕ

Глава 3. **Вычисление собственных многочленов, значений и векторов матриц**

Глава 4. **Решение численных уравнений**

Глава 5. **Численное интегрирование**

ПРЕДИСЛОВИЕ

Сейчас электронные вычислительные машины (ЭВМ) разных классов — от простейших карманных до мощных быстродействующих — стали относительно дешевыми и доступными. Они имеются во многих научных центрах, конструкторских бюро и на предприятиях. Поэтому сильно расширился круг научных и инженерно-технических работников, имеющих возможность производить на ЭВМ расчеты, гораздо более сложные, чем они ранее производили вручную. Но более сложные расчеты требуют и более глубокого знакомства с численными методами.

В этой книге содержится изложение большого числа вычислительных методов, применяемых при решении задач на ЭВМ, и как справочник по этим методам она может служить руководством для работников вычислительных центров и тех научных и инженерно-технических работников, которым приходится иметь дело с численным решением научных и технических задач.

Книга может быть использована как учебник при изучении теории вычислительных методов математики студентами технических учебных заведений с расширенной программой курса математики. Для понимания почти всего содержания книги достаточно знания анализа и алгебры в объеме таких программ. Лишь в очень небольшой части изложения авторы выходили за их границы, используя понятие аналитической функции комплексной переменной и интегральные представления таких функций.

Как пособие при изучении вычислительных методов, книга будет полезной также студентам физических, механико-математических и математических факультетов университетов.

При составлении книги авторы использовали многолетний опыт преподавания вычислительной математики в Белорусском государственном университете и работы в вычислительных центрах университета и Института математики АН Белоруссии.

Первый том содержит задачи интерполирования функций, линейной алгебры, решения численных уравнений и численного интегрирования функций.

Второй том посвящен главным образом вопросам численного решения дифференциальных уравнений как обыкновенных, так и с частными производными. В нем будут также содержаться задачи решения интегральных уравнений и улучшения сходимости рядов и последовательностей.

Авторы стремились сделать изложение простым и достаточно наглядным, иногда даже отодвигая вопросы общности и логической строгости на второе место.

Книга разбита на главы, параграфы и пункты. В конце каждой главы указана литература, рекомендуемая для более глубокого изучения изложенной в главе темы. Нумерация рисунков — сквозная по всей книге. Нумерация формул — своя в каждом параграфе. Если ссылка на формулу не выводит за пределы данного параграфа, то указывается только номер формулы, например (18); если ссылка относится к другому параграфу данной главы, то указываются номера параграфа и формулы, например (3.18); если ссылка относится к другой главе, то указываются номера главы, параграфа и формулы, например (1.3.18).

Авторы

ВВЕДЕНИЕ

1. О вычислительных методах математики. Чтобы описать, какое место занимает теория вычислительных методов в научных знаниях, необходимо сказать, что современная вычислительная математика содержит в себе (как принято считать) три главнейших части: 1) теорию вычислительных методов; 2) приборы, позволяющие автоматизировать вычисления, осуществлять связь, хранить информацию и т. д.; среди них центральную роль играют вычислительные машины; 3) вспомогательные средства, облегчающие управление работой вычислительной машины; к ним относятся алгоритмические языки различных назначений, стандартные программы, содержащие изложения наиболее часто употребляемых вычислительных процессов, программы-диспетчеры и т. п. В настоящей книге будет в краткой форме дано изложение основ теории вычислительных методов и указаны те из них, которые часто применяются в вычислениях.

Теория вычислительных методов является очень разветвленной наукой и имеет применения всюду, где приходится встречаться с числами или рассматривать явления и процессы, подчиняющиеся количественным законам. Такие законы изучаются в очень многих науках: математике, физике, механике, астрономии, экономике, биологии, медицине, теории управления и регулирования и т. д. *). Этот неполный перечень наук приведен не только для того, чтобы показать, насколько большим является число отраслей деятельности людей, где

*) Математические методы проникли даже в такие отдаленные от естествознания науки, как общественные и гуманитарные. Сейчас нередко говорят, быть может, преждевременно, о математизации всех наук.

может применяться вычислительная математика, но скорее для следующей цели. В каждой из областей применения могут возникать свои специфические задачи, требующие выработки особых, приспособленных к ним вычислительных методов. Например, задачи, возникающие в вопросах организации производства, принадлежат математике, изучающей конечные множества, и решаются преимущественно с помощью некоторого направленного перебора части возможных вариантов. В проблемах же физики и механики очень часто бывает необходимо находить функции одного или нескольких аргументов, дающие описание изучаемого явления, и они разыскиваются, как правило, путем решения дифференциальных уравнений, являющихся записью законов, по которым протекает это явление.

Большое число видов задач, к решению которых может применяться вычислительная математика, потребовало создания многочисленных методов их решения.

Для книги были отобраны такие вычислительные методы, которые предназначены для решения задач, часто встречающихся в практике вычислений; полезность методов проверена многими годами их применений, и для их понимания достаточно знаний в объеме технических вузов с расширенной программой по математике.

Эти требования устанавливают выбор отделов теории вычислительных методов, включенных в книгу. Но осталось еще произвести выбор самих методов. Каждая задача может быть решена не одним, а несколькими вычислительными методами. Принципы выбора методов для наглядности изложения мы поясним на частном вопросе. Рассмотрим хорошо известную проблему интегрирования функций. Интеграл разыскивают обычно в форме линейной комбинации нескольких значений интегрируемой функции и строят формулы вида

$$\int_a^b f(x)\,dx = \sum_{k=1}^{n} A_k f(x_k) + R_n(f). \qquad (1)$$

Здесь сумма должна давать приближенное значение интеграла, а $R_n(f)$ есть погрешность этого значения.

При построении вычислительных формул (1) можно распорядиться выбором параметров A_k и x_k ($k=$

$= 1, \ldots, n$). Чтобы читатель мог представить себе, в каком большом количестве вариантов в практических условиях приходится решать эту проблему, мы перечислим сейчас главнейшие требования, которые предъявляют реальные задачи *).

При применении вычислительного правила (1) встречаются два основных случая, которые должны быть учтены во всем последующем изложении. Во-первых, когда функция F задана таблицей своих значений, тогда в качестве x_k можно брать только табличные значения аргумента x. В этом случае мы почти лишены возможности распоряжаться выбором x_k (можем лишь некоторые табличные значения пропустить). При построении вычислительной формулы (1) здесь можно выбирать без ограничений только коэффициенты A_k. Во-вторых, когда F задана аналитически (формулой). В этом случае можно произвольно выбирать как A_k, так и x_k.

Для приближенного вычисления интегрируемую функцию F обычно заменяют на другую, более простую, легко интегрируемую и принимающую в точках x_k ($k = 1, \ldots, n$) те же значения $F(x_k)$, что и F. В качестве таких более простых функций берут алгебраические или тригонометрические многочлены, рациональные функции и т. д. Чаще всего пользуются алгебраическими многочленами и заменяют F таким многочленом на всем отрезке $[a,b]$ или разбивают $[a,b]$ на части и на каждой из них берут свой алгебраический многочлен: например, заменяют F кусочно-линейной функцией или функцией, составленной из кусков многочленов второй степени, и т. д. Этого оказывается достаточно для очень широкого класса случаев, и этому способу будет уделено много внимания в книге. Но такая замена далеко не исчерпывает потребностей практики, и нередко приходится пользоваться другими простыми функциями, например, при интегрировании периодических функций естественнее применять не алгебраические, а тригонометрические многочлены, или при вычислении интегралов по полуоси $[a, \infty)$ заменять F другими функциями, стремящимися

*) Более теоретические аспекты проблемы и условия, рассматриваемые в них, нами опущены.

к нулю при $x \to \infty$ примерно с той же скоростью, что и $|F|$.

Далее, если функция F имеет особенности (например обращается в бесконечность в какой-либо точке) и интеграл является несобственным, то простейшие гладкие функции становятся непригодными для замены F. Тогда должны быть построены свои особые методы вычисления интеграла. В этом и предыдущем случае правило интегрирования (1) заменяют другим, более подходящим к свойствам интегрируемой функции.

Были перечислены лишь некоторые проблемы, которые возникают в задаче численного интегрирования, но их достаточно, чтобы увидеть, насколько многосторонней является эта задача и как многочисленны должны быть правила вычисления интегралов. Аналогичное положение имеет место и для других разделов теории методов вычислений.

Универсальным методам вычислений или методам, которые могут быть применены в очень широком классе случаев, всегда присущ неизбежный недостаток: они мало учитывают свойства каждой отдельной задачи и поэтому, как правило, имеют невысокую точность и медленную сходимость к точным значениям. Поэтому их применение обычно требует затраты большого труда или машинного времени. Специализированные же методы, позволяющие более полно учитывать свойства рассматриваемой задачи, обладают в этом отношении преимуществом перед универсальными и позволяют получать нужную точность с меньшей затратой работы.

Поэтому в книгу, кроме универсальных методов, был включен также набор более специализированных методов, которые позволяют часто облегчить вычислительный труд и знание которых, по мнению авторов, является весьма желательным.

Сделаем еще замечание, которое полезно иметь в виду всем, приступающим к изучению теории методов вычислений. Вычислительная математика очень часто позволяет получить решение там, где другие методы оказываются бессильными. Это — старая наука, и ее быстрое развитие за несколько последних десятилетий связано с изобретением быстродействующих вычислительных машин, освободивших человека от выполнения

большого числа арифметических и простых логических операций.

Вычислительная математика и машины сделали возможным решение таких задач, которые три десятилетия назад считались недоступными, и сильно расширили возможность исследований, расчета и предвидения. Нет сомнений, что эти возможности с течением времени будут быстро возрастать, и вместе с этим будет повышаться роль вычислительной математики во многих областях деятельности людей. Но было бы, по-видимому, ошибкой преувеличивать значение численных методов и отдавать им предпочтение перед другими математическими методами исследований, такими, как аналитический или логический методы. Численные методы применимы не во всех случаях, и надо научиться использовать в работе все математические методы, правильно применяя каждый метод в наиболее подходящей для него области.

2. О погрешностях при численном решении задач. Погрешности результата приближенного решения задачи вызываются следующими причинами.

а) *Неточность информации* о решаемой задаче. Ошибки в начальных данных определяют ту часть погрешности в решении, которая не зависит от математической стороны решения задачи и называется обычно *неустранимой погрешностью*. Сведения о границах таких погрешностей используются для возможного упрощения самой задачи, при выборе метода вычислений, точность которого должна быть согласована с точностью задачи, и, наконец, для определения точности вычислений.

б) *Погрешность аппроксимации*. При решении задачи численными методами необходимо считаться с тем, что неизбежно придется иметь дело только с конечным количеством чисел, и с ними можно выполнить только конечное количество операций. Граница для каждого из этих количеств определяется свойствами машин, используемых при решении, временем, важностью задачи и другими факторами, среди которых соображения целесообразности и стоимости имеют немалый вес.

Если количество чисел или операций превышает допустимые границы, то задачу приходится упрощать и заменять ее другой задачей, близкой к заданной, но уже

удовлетворяющей нужным требованиям. Поясним это простым примером. Пусть на отрезке $a \leqslant x \leqslant b$ нужно найти функцию $y(x)$, удовлетворяющую дифференциальному уравнению $y' = f(x, y)$ и начальному условию $y(a) = y_0$. Найти численно $y(x)$ во всех точках отрезка $a \leqslant x \leqslant b$ невозможно, так как таких точек бесконечное множество. Для решения задачи можно, вообще говоря, поступить так: взять на $[a, b]$ конечную сетку равноотстоящих точек $(x_n = a + nh,\ n = 0, 1, \ldots, N,\ a + Nh \leqslant b < a + (N+1)h)$ и находить значения $y(x_n) = y(a + nh) \approx y_n$ функции $y(x)$ только в точках этой сетки. Начальное значение $y(x_0) = y(a) = y_0$ нам задано. Для нахождения же остальных значений y_n $(n = 1, 2, \ldots, N)$ нужно построить систему уравнений, близкую к $y' = f(x, y)$. Для этого, следуя примеру Эйлера, рассмотрим дифференциальное уравнение не всюду на $[a, b]$, а только в точках сетки:

$$y'(x_n) = f(x_n, y(x_n)),$$

и заменим в нем производную $y'(x_n)$ ее приближенным значением: $y'(x_n) \approx \frac{1}{h}(y_{n+1} - y_n)$. Тогда получим конечную систему уравнений

$$\frac{1}{h}(y_{n+1} - y_n) = f(x_n, y_n) \quad (n = 0, 1, \ldots, N-1), \qquad (2)$$

которая позволит последовательно вычислять значения $y_1, y_2, \ldots, y_N$.

Возвратимся к общим рассуждениям. Заменяя заданную задачу близкой к ней приближенной, мы получим некоторую погрешность, которая и называется *погрешностью* избранного численного метода. В приведенном примере она имеет следующий смысл. Если в уравнение (2) вместо y_n и y_{n+1} подставить точные значения $y(x_n)$ и $y(x_{n+1})$ решения заданного дифференциального уравнения, то равенство удовлетворится лишь с некоторой погрешностью и будет иметь вид

$$\frac{1}{h}[y(x_{n+1}) - y(x_n)] = f(x, y(x_n)) + r_n. \qquad (3)$$

Величина r_n и есть погрешность аппроксимации. Уравнение (2) получается из (3), если отбросить величину r_n, полагая ее достаточно малой.

Замена исходной задачи аппроксимирующей предопределяет часть погрешности приближенного решения, которую называют обычно *погрешностью метода.*

в) *Погрешность округлений.* Наиболее часто для записи чисел в ЭВМ применяется двоичная система с плавающей запятой:

$$x \approx \pm 2^p \sum_{k=1}^{t} \alpha_k 2^{-k} = \pm 2^p (\alpha_1, \ldots, \alpha_t) \quad (|p| \leqslant p_0), \quad \alpha_1 = 1.$$

Типичными являются приводимые ниже значения параметра p_0, определяющего границу для порядка числа, и параметра t, определяющего количество знаков в числе:

$$p_0 = 64, \quad t = 35.$$

В десятичной системе счисления порядок и число знаков даются равенствами

$$2^{p_0} = 2^{64} \approx 7 \cdot 10^{19} \quad \text{и} \quad 2^{-t} = 2^{-35} \approx 3 \cdot 10^{-11}.$$

Напомним некоторые факты элементарной теории погрешностей. Пусть a есть точное и a^* — приближенное значение некоторой величины. Разность $a - a^* = \varepsilon$ называют *погрешностью приближенного значения* a^*. Так как точное значение a в большинстве случаев неизвестно, то неизвестно и точное значение ε. Но очень часто бывает известна *верхняя граница* Δ *абсолютной величины погрешности*:

$$|a - a^*| = |\varepsilon| \leqslant \Delta.$$

Ее мы ниже будем называть сокращенно *границей погрешности* ε. Точное значение a лежит, очевидно, в следующих пределах:

$$a^* - \Delta \leqslant a \leqslant a^* + \Delta.$$

Часто эти неравенства условно записывают в форме

$$a = a^* \pm \Delta.$$

При вводе числа a в машину его обычно округляют и приближенно записывают в виде

$$a \approx \pm 2^p \sum_{k=1}^{t} \alpha_k 2^{-k} = a^*. \tag{4}$$

Погрешность такой записи не больше единицы последнего разряда в a^*:

$$|a - a^*| = |\varepsilon| \leqslant 2^{p-t} = \Delta.$$

Относительной погрешностью величины a^* называют отношение $\frac{a^* - a}{a^*}$. Наряду с ней рассматривают верхнюю границу для абсолютного значения этой величины

$$\left|\frac{a^* - a}{a^*}\right| = \left|\frac{\varepsilon}{a^*}\right| \leqslant \frac{\Delta}{|a^*|} = \delta,$$

которая ниже называется *границей относительной погрешности*. Так, при приближенном вводе числа a в машину в виде (4) граница погрешности может быть найдена следующим способом:

$$\delta = \left|\frac{\varepsilon}{a^*}\right| \leqslant \frac{\Delta}{|a^*|} = \frac{2^{p-t}}{2^p \sum \alpha_k 2^{-k}} \leqslant 2^{-t},$$

Ошибки округлений сказываются на точности окончательного результата. Ту часть погрешности приближенного решения, которая зависит только от этих ошибок, называют обычно *вычислительной погрешностью*.

Остановим внимание на оценке погрешностей простейших действий с приближенными числами.

Погрешность суммы. Пусть $a = x_1 + x_2$ и известны приближенные значения x_1^*, x_2^* слагаемых и границы Δ_1, Δ_2 для их погрешностей. Обозначим погрешности слагаемых соответственно ε_1 и ε_2:

$$a^* = x_1^* + x_2^*, \quad a = (x_1^* + \varepsilon_1) + (x_2^* + \varepsilon_2) = \\ = a^* + \varepsilon_1 + \varepsilon_2 = a^* + \varepsilon.$$

Поэтому

$$|\varepsilon| \leqslant |\varepsilon_1| + |\varepsilon_2| \leqslant \Delta_1 + \Delta_2.$$

Это позволяет сформулировать правило: *граница погрешности суммы не больше суммы границ погрешностей слагаемых.*

Погрешность произведения. Рассмотрим теперь произведение

$$x = x_1 \cdot x_2 = (x_1^* + \varepsilon_1)(x_2^* + \varepsilon_2) = x_1^* x_2^* + x_1^* \varepsilon_2 + x_2^* \varepsilon_1 + \varepsilon_1 \varepsilon_2.$$

Так как приближенное значение произведения есть $x^* = x_1^* x_2^*$, то его погрешность будет следующей:

$$\varepsilon = x_1^* \varepsilon_2 + x_2^* \varepsilon_1 + \varepsilon_1 \varepsilon_2. \tag{5}$$

В большом числе реальных задач погрешность является малой величиной сравнительно с точным и приближенным значениями, и последний член правой части (5) будет более высокого порядка малости, чем два первых члена. Поэтому произведением $\varepsilon_1 \varepsilon_2$ в (5), как правило, можно пренебречь, и пользоваться для погрешности ε вместо (5) неточным равенством

$$\varepsilon \approx x_1^* \varepsilon_2 + x_2^* \varepsilon_1. \tag{6}$$

Из (5) и (6) получаются точная и приближенная оценки погрешности произведения и ее границы Δ:

$$|\varepsilon| \leqslant \Delta \leqslant |x_1^*| \Delta_2 + |x_2^*| \Delta_1 + \Delta_1 \Delta_2, \tag{7}$$

$$|\varepsilon| \leqslant \Delta \lesssim |x_1^*| \Delta_2 + |x_2^*| \Delta_1. \tag{8}$$

Здесь и ниже знак $\lesssim$ применен для обозначения приближенного неравенства.

Неравенства (6) и (7) позволяют сформулировать правило оценки Δ: *граница погрешности произведения по границам погрешности сомножителей определяется неравенством* (7); *при этом для малых* Δ_1 и Δ_2 *допустимо пользоваться приближенным неравенством* (8).

Более простым является правило оценки относительной погрешности произведения. Пусть границы относительных погрешностей множителей x_1^*, x_2^* есть δ_1 и δ_2. Если разделить обе части (5) на $x_1^* x_2^*$, получим

$$\frac{\varepsilon}{x_1^* x_2^*} = \frac{\varepsilon_1}{x_1^*} + \frac{\varepsilon_2}{x_2^*} + \frac{\varepsilon_1}{x_1^*} \cdot \frac{\varepsilon_2}{x_2^*}.$$

Отсюда вытекает оценка границы относительной погрешности

$$\left| \frac{\varepsilon}{x_1^* x_2^*} \right| \leqslant \delta \leqslant \delta_1 + \delta_2 + \delta_1 \delta_2. \tag{9}$$

Аналогично приближенное равенство (6) дает

$$\left| \frac{\varepsilon}{x_1^* x_2^*} \right| \leqslant \delta \lesssim \delta_1 + \delta_2. \tag{10}$$

Из (9) и (10) следует правило оценки δ: *граница δ относительной погрешности произведения оценивается через границы δ_1 и δ_2 относительных погрешностей множителей неравенством* (9); *если же относительные погрешности δ_1 и δ_2 малы, допустимо пользоваться приближенным неравенством* (10).

Погрешность отношения. Пусть $x = x_1/x_2$ и сохраняются прежние обозначения для приближенных значений x_1^*, x_2^*, погрешностей и их границ. Сделаем дополнительное предположение: $|\Delta_2| < |x_2^*|$. Если оно нарушается, то нецелесообразно рассматривать отношение, так как делитель x_2 может оказаться равным нулю.

Погрешность отношения здесь вычисляется почти столь же просто, как и выше:

$$\varepsilon = x - x^* = \frac{x_1}{x_2} - \frac{x_1^*}{x_2^*} = \frac{1}{x_2^*(x_2^* + \varepsilon_2)}(\varepsilon_1 x_2^* - \varepsilon_2 x_1^*),$$

что приводит к следующему правилу оценки границы погрешности:

$$|\varepsilon| \leqslant \Delta \leqslant \frac{1}{|x_2^*|(|x_2^*| - \Delta_2)}(\Delta_1 |x_2^*| + \Delta_2 |x_1^*|).$$

Если считать величину Δ_2 пренебрежимо малой сравнительно с $|x_2^*|$, эту оценку можно заменить более простой, но приближенной:

$$\Delta \lesssim \frac{1}{|x_2^*|^2}(\Delta_1 |x_2^*| + \Delta_2 |x_1^*|).$$

Что же касается относительной погрешности, то ее выражение и оценка являются более простыми:

$$\begin{gathered} \frac{\varepsilon}{x^*} = \varepsilon \frac{x_2^*}{x_1^*} = \frac{x_2^*}{x_2^* + \varepsilon_2}\left(\frac{\varepsilon_1}{x_1^*} - \frac{\varepsilon_2}{x_2^*}\right), \\ \left|\frac{\varepsilon}{x^*}\right| \leqslant \delta \leqslant \frac{|x_2^*|}{|x_2^*| - \Delta_2}(\delta_1 + \delta_2) = \frac{1}{1 - \delta_2}(\delta_1 + \delta_2). \end{gathered} \tag{11}$$

Когда Δ_2 будет мала сравнительно с $|x_2^*|$, последняя оценка может быть заменена приближенной:

$$\delta \lesssim \delta_1 + \delta_2. \tag{12}$$

Это приводит к правилу: *относительная погрешность дроби оценивается при помощи равенства* (11); *если относительная погрешность* δ_2 *делителя значительно меньше единицы, допустимо применять более простое правило* (12).

Упрощенные правила для произведения и дроби (10) и (12) оказываются одинаковыми.

Погрешность вычисления функции. Пусть в некоторой выпуклой области G n-мерного числового пространства рассматривается непрерывно дифференцируемая функция $y = f(x_1, x_2, \ldots, x_n)$. Предположим, что в точке $(x_1, \ldots, x_n)$ области G нужно вычислить значение y. Пусть нам известны лишь их приближенные значения $x_1^*, x_2^*, \ldots, x_n^*$ такие, что точка $(x_1^*, \ldots, x_n^*)$ принадлежит G. Необходимо найти погрешность приближенного значения функции $y^* = f(x_1^*, \ldots, x_n^*)$. Через погрешности $\varepsilon_i = x_i - x_i^*$ аргументов она выражается следующим образом:

$$\varepsilon = f(x_1^* + \varepsilon_1, \ldots, x_n^* + \varepsilon_n) - f(x_1^*, \ldots, x_n^*),$$

или, если воспользоваться формулой Лагранжа,

$$\varepsilon = \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(x_1^* + \theta\varepsilon_1, \ldots, x_n^* + \theta\varepsilon_n)\varepsilon_i, \quad 0 \leqslant \theta \leqslant 1.$$

Отсюда получается оценка для границы погрешности вычисления функции

$$|\varepsilon| \leqslant \Delta \leqslant \sum_{i=1}^{n} B_i \Delta_i, \tag{13}$$

где

$$|\varepsilon_i| \leqslant \Delta_i, \quad B_i = \max_{\theta} \left| \frac{\partial}{\partial x_i} f(x_1^* + \theta\varepsilon_1, \ldots, x_n + \theta\varepsilon_n) \right|.$$

Когда погрешности ε_i достаточно малы, а частные производные суть достаточно плавно изменяющиеся функции, коэффициенты B_i допустимо заменить на абсолютные значения производных в точке $(x_1^*, \ldots, x_n^*)$. После этого получится приближенное, более простое неравенство, оценивающее Δ:

$$\Delta \lesssim \sum_{i=1}^{n} \Delta_i \left| \frac{\partial}{\partial x_i} f(x_1^*, \ldots, x_n^*) \right|. \tag{14}$$

Указанные выше оценки погрешностей при арифметических операциях сложения, умножения и деления могут быть, вообще говоря, включены в программу работы машины, и при помощи них получено суждение о точности результатов, выданных машиной. Но обычно этого не делают на основании следующих соображений: во-первых, внесение оценок повлечет за собой усложнение программы, замедлит работу машины и отнимет дополнительно места памяти; во-вторых, указанные оценки могут дать удовлетворительное представление о погрешностях, если получение результата требует небольшого числа операций. Если же число операций является большим, эти оценки, рассчитанные на учет самых неблагоприятных случаев, дадут преувеличенную оценку погрешностей, часто сильно превосходящую действительные их значения. На машине же проводят, как правило, вычисления с большим числом операций, и тогда рациональность применения рассмотренных правил вызывает сомнение. Поэтому такие правила применяют в редких случаях малого числа операций, например для задач, решаемых на настольных машинах.

Проблема определения точности результата является трудной, и в полном ее виде в настоящей книге не рассматривается. В последующем изложении для многих вычислительных методов будут получены явные выражения для их погрешностей. Они позволяют составить представление о порядке малости этих погрешностей, либо в некоторых случаях вычислить их оценку или приближенное значение.

Мы ограничимся тем, что укажем на некоторые методы неполного решения проблемы, которые часто применяют в практике вычислений: 1) решение задачи другим методом или повторное применение того же метода, но с иной последовательностью операций; 2) малое изменение входных данных и решение задачи с измененными данными. Можно надеяться на то, что совпадающие знаки в результатах двух разных решений задачи будут верными.

ГЛАВА 1

ИНТЕРПОЛИРОВАНИЕ

§ 1. Содержание задачи; погрешность и сходимость

1. О задаче интерполирования. Под *аппроксимацией* в математике понимается операция нахождения неизвестных численных значений какой-либо величины по известным ее значениям и, может быть, численным значениям других величин, связанных с рассматриваемой. Задачи такого рода возникают во многих разделах науки и ее приложений, и проблема аппроксимации поэтому является многосторонней.

В этой главе будет изложен частный вопрос такого рода — задача об интерполировании значений функции. В виде, достаточно общем для многих случаев, она может быть сформулирована следующим образом. Пусть в k_0 точках $x_{01}, \ldots, x_{0k_0}$ известны значения некоторой функции: $f(x_{01}), \ldots, f(x_{0k_0})$; в k_1 точках $x_{11}, \ldots, x_{1k_1}$ известны значения первой производной от нее: $f'(x_{11}), \ldots, f'(x_{1k_1})$ и, наконец, в k_m точках $x_{m1}, \ldots, x_{mk_m}$ известны значения ее m-й производной: $f^{(m)}(x_{m1}), \ldots$ $\ldots, f^{(m)}(x_{mk_m})$.

Точки x_{ij} $(i = 1, \ldots, m;\ j = 1, \ldots, k_i)$ называются *узлами интерполирования*, а совокупность пар чисел $[x_{ij}, f^{(i)}(x_{ij})]$ — *исходными данными* интерполирования. Общее число исходных данных $k_0 + k_1 + \ldots + k_m$ обозначим через n.

Пусть значение x отлично от узлов x_{0j} $(j = 1, \ldots$ $\ldots, k_0)$, где известны значения f. Нужно, пользуясь исходными данными, найти значение $f(x)$ *).

*) Подобная задача может быть поставлена относительно нахождения значения производной любого порядка $f^{(i)}$ или других величин, связанных с f.

Если не делать относительно f дополнительных предположений, то такая задача является неопределенной, и в качестве $f(x)$ может быть взято любое число. Это становится особенно очевидным, если рассмотреть геометрическое значение задачи. Для пояснения существа вопроса достаточно взять простейший частный случай, когда интерполирование $f(x)$ выполняется по значениям только функции f. Этот частный случай и будет по преимуществу рассматриваться в дальнейшем.

Допустим, что в n узлах $x_1, x_2, \ldots, x_n$ даны значения функции $f(x_1), f(x_2), \ldots, f(x_n)$. Если в плоскости взять декартову систему координат, то геометрически это будет означать задание в плоскости n точек M_i с координатами $[x_i, f(x_i)]$ $(i = 1, \ldots, n)$. Будем считать, что функция f рассматривается на некотором отрезке $[a, b]$. Графиком ее будет множество точек $[x, f(x)]$, $x \in [a, b]$. Если f есть непрерывная функция, то такое множество будет некоторой линией над $[a, b]$. Она проходит через точки M_i; в остальном эта линия произвольна, и ее ордината $f(x)$ в точке с абсциссой x может быть любой.

Для интерполирования мы должны указать правило, позволяющее по x_i и $f(x_i)$ вычислить значение $f(x)$ точно или приближенно. Пусть избрано какое-либо правило, и его применение дало для $f(x)$ приближенное значение y. В погрешности интерполирования $\varepsilon = f(x) - y$ величина y зависит от исходных данных, в частности от числа n узлов x_i и от их расположения, от избранного правила интерполирования. Всеми указанными фактами (числом узлов x_i, их расположением и выбором правила вычисления y) в большинстве случаев можно распорядиться. Для краткости назовем их «параметрами» интерполирования.

При действительном построении формул интерполирования эту задачу видоизменяют. Так как нецелесообразно строить формулы для каждой функции в отдельности, так же как и для очень узких классов сходных функций, то строят формулы интерполирования, которые могут дать хорошие по точности результаты в некоторых достаточно широких классах функций f, и стараются избрать эти классы так, чтобы каждый класс содержал в себе все практически важные функции с некоторыми

общими для них свойствами. Более подробно об этом будет сказано ниже.

Пусть рассматривается некоторый класс F функций f, заданных на отрезке $[a, b]$. Для интерполирования f обычно выбирают семейство Φ функций φ, более простых, чем f, и достаточно легко вычисляемых; затем среди функций φ избирают ту, которая имеет такие же исходные данные интерполирования *), что и f, т. е. для которой выполняются равенства $\varphi(x_i) = f(x_i)$ $(i = 1, 2, \ldots, k)$. После этого приближенно полагают $f(x) \approx \approx \varphi(x)$ при всех x из $[a, b]$ или при заданном одном значении x, в зависимости от поставленной цели.

Семейство Φ обычно берут в форме линейной комбинации каких-либо простейших функций с таким расчетом, чтобы число их равнялось числу исходных данных интерполирования.

Рассмотрим последовательность функций ω_i $(i = 1, 2, \ldots)$, определенных на отрезке $[a, b]$ и линейно независимых там **). Возьмем n первых функций ω_i и образуем линейную комбинацию

$$s_n = a_1\omega_1 + a_2\omega_2 + \ldots + a_n\omega_n. \qquad (1)$$

Здесь a_i — произвольные постоянные коэффициенты. Их выберем так, чтобы выполнялись условия $s_n(x_i) = f(x_i)$ $(i = 1, 2, \ldots, n)$. Это дает для a_i систему n линейных уравнений

$$a_1\omega_1(x_i) + a_2\omega_2(x_i) + \ldots + a_n\omega_n(x_i) = f(x_i) \qquad (2)$$
$$(i = 1, 2, \ldots, n).$$

Решая ее относительно a_i и подставляя найденные значения в (1), получим линейную комбинацию s_n, интерполирующую функцию f по предписанным исходным данным.

Укажем на условия, которым должен быть подчинен выбор «координатных» функций ω_i, положенных в основание интерполирования.

---

*) Предполагается, что такой выбор возможен для каждой функции f и является единственным.

**) Функции бесконечной последовательности называются *линейно независимыми*, если взятые в любом конечном числе они оказываются линейно независимыми.

Чтобы система (2) имела единственное решение, нужно, чтобы ее определитель

$$D_n = D_n(x_1, \ldots, x_n) = \begin{vmatrix} \omega_1(x_1) & \omega_2(x_1) & \ldots & \omega_n(x_1) \\ \omega_1(x_2) & \omega_2(x_2) & \ldots & \omega_n(x_2) \\ \ldots & \ldots & \ldots & \ldots \\ \omega_1(x_n) & \omega_2(x_n) & \ldots & \omega_n(x_n) \end{vmatrix} \quad (3)$$

был отличен от нуля:

$$D_n(x_1, \ldots, x_n) \neq 0. \quad (4)$$

Левая часть неравенства зависит от x_i $(i = 1, \ldots, n)$, и функции ω_i принято выбирать так, чтобы неравенство (4) выполнялось при всяких $x_1, \ldots, x_n$, лежащих на $[a, b]$ и различных между собой. Последнее равносильно тому, что интерполирование f при помощи линейной комбинации s_n возможно и единственно при любом выборе узлов x_i $(i = 1, \ldots, n)$ на $[a, b]$, лишь бы они были несовпадающими.

Систему функций ω_i $(i = 1, \ldots, n)$, для которой выполняется условие (4) при любых различных $x_i \in [a, b]$, называют *системой Чебышева* на $[a, b]$.

Число n исходных данных интерполирования может быть произвольным. Поэтому можно сформулировать первое условие, которому должен быть подчинен выбор функций ω_i: *при всяких значениях* $n = 1, 2, \ldots$ *функции* ω_i $(i = 1, \ldots, n)$ *должны составлять систему Чебышева на отрезке* $[a, b]$.

Второе условие связано с понятием полноты. Семейство линейных комбинаций (1) называется *полным* в классе F функций f, если для всякой функции $f \in F$ и любого $\varepsilon > 0$ существует такое n и такие коэффициенты $a_1, \ldots, a_n$, что при всяких $x \in [a, b]$ выполняется неравенство

$$|f(x) - s_n(x)| < \varepsilon. \quad (5)$$

Если семейство s_n не обладает этим свойством, то невозможно рассчитывать на то, чтобы с помощью комбинаций s_k можно было выполнить сколь угодно точное интерполирование всех функций f. В этом случае следует признать систему функций ω_i $(i = 1, 2, \ldots)$ неудачно выбранной и неблагоприятной для интерполирования.

Условие, которому необходимо подчинить выбор функций ω_i, является следующим: *система функций* ω_i $(i = 1,$

2, ...) *должна быть такой, чтобы соответствующее ей семейство линейных комбинаций* (1) *было полным в классе F функций f, подлежащих интерполированию.*

Попутно отметим, что выполнение условия полноты еще не гарантирует возможность сколь угодно точного интерполирования f. Это становится ясным из того, что полнота семейства s_n в F дает возможность сколь угодно точного приближения f посредством s_n, когда на выбор s_n не налагается никакого ограничения. В проблеме же интерполирования выбор s_n, даже если считать n произвольным числом, является строго регламентированным: s_n определяется выбором узлов x_i $(i = 1, \ldots, n)$ и условиями совпадения значений $s_n(x_i)$ и $f(x_i)$. Вопрос о возможности сколь угодно точного приближения f при таком ограничении в построении s_n остается открытым и подлежит исследованию для каждого конкретного интерполяционного процесса.

Приведем примеры выбора систем функций ω_i.

I. Интерполирование с помощью алгебраических многочленов. Положим $\omega_1 = 1$, $\omega_2(x) = x, \ldots, \omega_n(x) = x^{n-1}, \ldots$ Линейная комбинация s_n будет многочленом степени $n-1$ от x:

$$s_n(x) = P_{n-1}(x) = a_1 + a_2 x + \ldots + a_n x^{n-1}.$$

Система уравнений (2), из которой должны быть найдены коэффициенты a_i, здесь имеет вид

$$P_{n-1}(x_i) = a_1 + a_2 x_i + \ldots + a_n x_i^{n-1} = f(x_i)$$
$$(i = 1, 2, \ldots, n).$$

Ее определитель является *определителем Вандермонда*

$$D_n(x_1, \ldots, x_n) = \begin{vmatrix} 1 & x_1 & \ldots & x_1^{n-1} \\ 1 & x_2 & \ldots & x_2^{n-1} \\ \cdot & \cdot & \cdot & \cdot \\ 1 & x_n & \ldots & x_n^{n-1} \end{vmatrix}.$$

Он отличен от нуля при всяких различных между собой значениях x_i, и интерполирование функции f по ее значениям в узлах x_i с помощью многочлена $P_{n-1}(x)$ всегда возможно и единственно.

В курсах математического анализа доказывается теорема *) Вейерштрасса: *если функция f непрерывна на конечном замкнутом отрезке $[a, b]$, то для всякого $\varepsilon > 0$ существует многочлен P_m некоторой степени m, для которого при любых $x \in [a, b]$ выполняется неравенство*

$$|f(x) - P_m(x)| < \varepsilon.$$

Иначе говоря, *для любого конечного замкнутого отрезка семейство алгебраических многочленов является полным в классе непрерывных на этом отрезке функций.*

Этот факт позволяет ожидать, что при надлежащем выборе узлов x_i и их числа n непрерывную на $[a, b]$ функцию можно на этом отрезке интерполировать сколь угодно точно при помощи алгебраического многочлена. В какой мере это ожидание оправдывается, будет выяснено в параграфе, посвященном проблеме сходимости интерполяционных процессов. Изложенные соображения относятся к интерполированию функций на любых конечных отрезках, могущих иметь сколь угодно большую длину. Вероятно, следует отметить более частный случай, встречающийся в практике, когда интерполирование выполняется на отрезках малой длины. Если функция имеет производные достаточно высоких порядков, то по своему поведению на малом участке она мало отличается от алгебраического многочлена, что сразу же видно из того, что она там представима по формуле Тейлора с малым остаточным членом. Поэтому интерполирование ее алгебраическим многочленом должно дать хорошую точность, если взять достаточно большое число узлов вблизи точки интерполирования x.

II. Интерполирование периодических функций тригонометрическими многочленами. Рассмотрим периодические функции периода 2π. Для их интерполирования естественно воспользоваться не алгебраическими, а тригонометрическими многочленами

$$T_n(x) = a_0 + \sum_{j=1}^{n} (a_j \cos jx + b_j \sin jx).$$

$T_n(x)$ содержит $2n + 1$ произвольных коэффициентов.

*) См., например, Л. Д. Кудрявцев, Математический анализ, т. II, стр. 262, изд-во «Высшая школа», М., 1970.

В соответствии с этим предположим, что для функции f известны ее значения в $2n+1$ узлах

$$0 \leqslant x_1 < x_2 < \ldots < x_{2n+1} \leqslant 2\pi. \qquad (6)$$

Коэффициенты многочлена должны быть определены из условий

$$T_n(x_p) = f(x_p) \quad (p = 1, 2, \ldots, 2n+1), \qquad (7)$$

дающих систему $2n+1$ уравнений относительно a_j и b_j.

Можно просто убедиться в том, что эта система имеет единственное решение. Заменим переменную x, положив $z = e^{ix}$. Если воспользоваться формулами Эйлера, которые в переменной z будут иметь вид $\cos jx = \frac{1}{2}(z^j + z^{-j})$, $\sin jx = \frac{1}{2i}(z^j - z^{-j})$, то для $T_n(x)$ получится следующее выражение через z:

$$T_n(x) = z^{-n}(c_0 + c_1 z + \ldots + c_{2n} z^{2n}) = z^{-n} P_{2n}(z).$$

Поэтому интерполирование $f(x)$ тригонометрическим многочленом $T_n(x)$ равносильно интерполированию $z^n f(x)$ алгебраическим многочленом $P_{2n}(z)$ по узлам $z_p = e^{ix_p}$ $(p = 1, 2, \ldots, 2n+1)$. Так как ввиду неравенства (6) все z_p различны между собой, последнее интерполирование, как доказано в первом примере, всегда возможно и единственно, и система (7) должна иметь решение и притом только одно.

В учебных книгах по математическому анализу доказывается*), что *семейство тригонометрических многочленов $T_n(x)$ является полным в классе непрерывных 2π-периодических функций.* Поэтому можно предполагать, что тригонометрическое интерполирование при надлежащем выборе узлов должно в широком классе случаев позволить достаточно точно интерполировать непрерывные периодические функции.

Сделаем еще одно дополнительное замечание. Выше говорилось о том случае, когда для интерполирования используется семейство функций вида (1), содержащее

*) См. сноску на стр. 26.

численные параметры a_i линейно. Такой выбор семейства оправдывается двумя обстоятельствами: во-первых, тем, что постоянные a_i определяются из простейшей линейной системы уравнений (2), и, во-вторых, тем, что правила интерполирования, полученные этим путем, позволяют производить вычисления в очень большом числе случаев с достаточно высокой точностью и при небольшой затрате вычислительного труда.

На практике значительно реже применяется интерполирование при помощи семейства функций, содержащих численные параметры нелинейно. Теория такого интерполирования почти не развивалась, и мы в книге на нем сейчас не будем останавливаться. С одним из его видов мы встретимся во второй части книги в главе об улучшении сходимости рядов и последовательностей.

2. Погрешность интерполирования и сходимость интерполяционного процесса. Пусть интерполирование функции f выполнено при помощи линейной комбинации (1). *Погрешностью интерполирования* называют разность $f(x) - s_n(x) = R_n(x)$. Эта величина, как указывалось выше, зависит от многих факторов — от свойств f, от всех параметров интерполирования, от положения точки интерполирования x; поэтому изучение $R_n(x)$ является сложной задачей. Остановим внимание на некоторых ее аспектах. Начнем с проблемы оценки погрешности. Здесь прежде всего следует определить выбор численной меры $m(R_n)$ погрешности. Если точка интерполирования x зафиксирована, то естественная численная мера погрешности определяется единственным образом: $\rho(\varepsilon_n) = |R_n(x)|$. Если же интерполирование выполняется всюду на отрезке $[a, b]$, то выбор численной меры $\rho(\varepsilon_n)$ может быть сделан многими способами. Когда рассматривается задача о равномерном интерполировании f на $[a, b]$, то за меру приближения принимают величину

$$\rho(\varepsilon_n) = \sup_x |R_n(x)|.$$

Если интерполирование выполняется на некотором отрезке $[\alpha, \beta]$, содержащемся в $[a, b]$, и представляет интерес средняя квадратичная погрешность интерполирова-

ния, то меру погрешности определяют равенством

$$\rho^2(R_n) = \int_a^\beta [f(x) - s_n(x)]^2\, dx.$$

Могут быть выбраны другие меры $m(R_n)$ в зависимости от задачи. Ниже для частных случаев будут приведены представления погрешности, которые дают возможность (по крайней мере теоретическую) вычислять значения погрешностей, находить их меру или получать оценку ее с той или иной степенью точности.

Величина погрешности R_n зависит, очевидно, от свойств функции f и от того, насколько с ними согласованы свойства линейной комбинации s_n. Поясним эту мысль на частном примере. Пусть выполняется интерполирование f при помощи алгебраического многочлена $P_{n-1}(x)$ степени $n-1$. Многочлен изменяется очень плавно, и если функция f будет иметь особенности (например, если ее производные невысоких порядков будут обращаться в бесконечность), то трудно ожидать, чтобы ее интерполирование многочленом P_{n-1} имело хорошую точность. Наоборот, если f имеет высокий порядок дифференцируемости или если f есть аналитическая функция и ее особые точки, нарушающие плавность изменения f, отстоят далеко от отрезка интерполирования $[a, b]$, то можно ожидать малую погрешность при интерполировании.

Когда оценивают меру погрешности, то это делают не для какой-либо индивидуально взятой функции, а получают оценку для класса функций, имеющих общими некоторые свойства. Представляют интерес как оценки погрешностей для широких классов функций (они используются, например, для выяснения условий сходимости интерполирования и установления порядка малости погрешности), так и оценки для узких специализированных классов функций, которые обладают многими общими свойствами (такие оценки могут быть полезны при решении специальных задач, например при определении числа верных знаков в полученном приближении или для определения числа узлов, необходимых для получения нужной точности интерполирования). Сформулируем еще одну постановку проблемы сходимости. Будем

считать, что функции ω_i, лежащие в основе интерполирования, выбраны и фиксированы. Допустим, что интерполирование f выполняется во всех точках отрезка $[\alpha, \beta]$, принадлежащего $[a, b]$*). Предположим, что число узлов n неограниченно возрастает. Мы должны указать также, как при этом будет изменяться расположение узлов. Допустим, что дана следующая бесконечная треугольная таблица узлов:

$$X = \begin{bmatrix} x_1^1 & & & \\ x_1^2, & x_2^2 & & \\ \cdots & \cdots & \cdots & \cdots \\ x_1^n, & x_2^n, & \ldots, & x_n^n \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}, \tag{8}$$

устроенная так, что на n-м шаге процесса интерполирование выполняется по n узлам, указанным в строке номера n. Вопрос о сходимости интерполирования исследуется не для отдельно взятой функции, а для классов функций с некоторыми общими свойствами. Пусть дан класс F функций f, и нужно выяснить, при каких условиях будет иметь место сходимость интерполирования в принятой мере погрешности, т. е. когда будет

$$\rho(\varepsilon_n) = \rho[f(x) - s_n(x)] \to 0 \tag{9}$$

при $f \in F$, $x \in [\alpha, \beta]$ и $n \to \infty$.

В сформулированной проблеме приходится иметь дело со следующими факторами: 1) область задания $[a, b]$ функций f, 2) область $[\alpha, \beta]$ интерполирования, 3) множество F функций f, 4) таблица узлов X интерполирования и 5) мера $m(\varepsilon_n)$ погрешности; нужно выяснить, как они должны быть связаны между собой, чтобы имела место сходимость (9). Некоторые результаты будут указаны в § 7, посвященном задаче сходимости.

§ 2. Конечные разности и разностные отношения

1. Конечные разности. Они являются рабочим аппаратом при изучении функций, заданных таблицей значений в равноотстоящих точках, и применяются в вычислениях с такими функциями.

*) Не исключается случай, когда $[\alpha, \beta]$ может выродиться в одну точку.

Предположим, что для равноотстоящих значений аргумента $x_k = x_0 + kh$ ($k = 0, 1, 2, \ldots$) известны соответствующие им значения функции: $y_k = f(x_0 + kh)$. *Конечными разностями первого порядка* называются величины

$$\Delta y_0 = y_1 - y_0, \quad \Delta y_1 = y_2 - y_1, \ \ldots, \ \Delta y_k = y_{k+1} - y_k, \ \ldots$$

Разности второго порядка определяются равенствами

$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0, \quad \Delta^2 y_1 = \Delta y_2 - \Delta y_1, \ \ldots$$

$$\ldots, \ \Delta^2 y_k = \Delta y_{k+1} - \Delta y_k, \ \ldots$$

и т. д. *Разности порядка* $n+1$ определяются через разности порядка n следующим образом:

$$\Delta^{n+1} y_k = \Delta^n y_{k+1} - \Delta^n y_k \quad (k = 0, 1, 2, \ldots).$$

Легко может быть найдено выражение разности любого порядка через значения функции:

$$\Delta y_0 = y_1 - y_0,$$

$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0 = (y_2 - y_1) - (y_1 - y_0) = y_2 - 2y_1 + y_0.$$

Продолжая вычисления и выполняя индукцию, убедимся, что верно равенство

$$\Delta^n y_0 = y_n - \frac{n}{1!} y_{n-1} + \frac{n(n-1)}{2!} y_{n-2} + \ldots + (-1)^n y_0. \quad (1)$$

Введем оператор E увеличения аргумента на шаг h, определив его и его степени равенствами $Ef(x) = f(x+h)$, $E^m f(x) = f(x+mh)$. Заметив, что $Ey_k = y_{k+1}$, $E^m y_k = y_{k+m}$, можно равенство (1) записать в краткой условной форме:

$$\Delta^n y_0 = (E-1)^n y_0. \quad (2)$$

Столь же просто можно получить выражения для значения функции y_n любого номера n через начальное значение y_0 и значения конечных разностей $\Delta^k y_0$ ($k = 0, 1, \ldots, n$), относящихся к начальной точке x_0. Из $\Delta y_0 = y_1 - y_0$ следует $y_1 = y_0 + \Delta y_0$. Далее $y_2 = y_1 + \Delta y_1 = (y_0 + \Delta y_0) + (\Delta y_0 + \Delta^2 y_0) = y_0 + 2\Delta y_0 + \Delta^2 y_0$. При

помощи несложно проводимой индукции доказывается, что

$$y_n = y_0 + \frac{n}{1!}\Delta y_0 + \frac{n(n-1)}{2!}\Delta^2 y_0 + \ldots + \Delta^n y_0 = \\ = (1+\Delta)^n y_0. \quad (3)$$

В конце следующего пункта будут приведены некоторые сведения о порядках малости конечных разностей.

2. Разностные отношения и связь их с конечными разностями. Разностные отношения, которые называют также «разделенными разностями» функции, применяются в вычислениях и для изучения функций в том случае, когда последние задаются на произвольной системе значений аргумента.

Предположим, что для любых, но различных между собой значений аргумента $x_0, x_1, x_2, \ldots$ даны значения функции $f(x_0), f(x_1), f(x_2), \ldots$ *Разностными отношениями первого порядка* называются величины

$$f(x_0, x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}, \quad f(x_1, x_2) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}, \ldots$$

Они имеют смысл средних скоростей изменения функции f на отрезках $(x_0, x_1), (x_1, x_2), \ldots$ По ним составляются *разностные отношения второго порядка*

$$f(x_0, x_1, x_2) = \frac{f(x_1, x_2) - f(x_0, x_1)}{x_2 - x_0},$$

$$f(x_1, x_2, x_3) = \frac{f(x_2, x_3) - f(x_1, x_2)}{x_3 - x_1}, \ldots$$

Разностные отношения любого *порядка* $n+1$ $(n = 1, 2, \ldots)$ определяются при помощи разностных отношений предыдущего порядка n по формуле

$$f(x_0, x_1, \ldots, x_n, x_{n+1}) = \frac{f(x_1, x_2, \ldots, x_{n+1}) - f(x_0, x_1, \ldots, x_n)}{x_{n+1} - x_0}.$$

Могут быть получены простые выражения разностных отношений всех порядков через значения функции. Действительно, по определению разностного отношения первого порядка

$$f(x_0, x_1) = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0}.$$

Поэтому для разностного отношения второго порядка получим

$$f(x_0, x_1, x_2) = \frac{1}{x_2 - x_0}\{f(x_1, x_2) - f(x_0, x_1)\} =$$

$$= \frac{1}{x_2 - x_0}\left\{\frac{f(x_1)}{x_1 - x_2} + \frac{f(x_2)}{x_2 - x_1} - \frac{f(x_0)}{x_0 - x_1} - \frac{f(x_1)}{x_1 - x_0}\right\} =$$

$$= \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}.$$

Выполнив несложную индукцию, можно показать, что при всяком n верно равенство

$$f(x_0, x_1, \ldots, x_n) =$$

$$= \sum_{i=0}^{n} \frac{f(x_i)}{(x_i - x_0)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)} =$$

$$= \sum_{i=0}^{n} \frac{f(x_i)}{\omega'(x_i)}, \qquad \omega(x) = \prod_{j=0}^{n}(x - x_j). \qquad (4)$$

Если в последней части этих равенств переставить какие-либо два узла, например x_k и x_l, то это равносильно тому, что поменяются местами слагаемые, отвечающие значениям $i = k$ и $i = l$; сумма же при этом не изменится. Так как всякая перестановка узлов может быть получена в результате перестановок пар узлов, то $f(x_0, x_1, \ldots, x_n)$ не будет изменяться при любой перестановке узлов x_i $(i=0, 1, \ldots, n)$. Это позволяет утверждать, что разностное отношение $f(x_0, x_1, \ldots, x_n)$ является симметрической функцией узлов $x_0, x_1, \ldots, x_n$.

Приведенные ниже две теоремы устанавливают связь между разностными отношениями и производными.

Теорема 1. *Пусть узлы $x_0, x_1, \ldots, x_n$ лежат на отрезке $[a, b]$ и функция f имеет на этом отрезке непрерывную производную порядка n. Верно следующее равенство, дающее выражение разностного отношения $f(x_0, x_1, \ldots, x_n)$ через производную порядка n от f:*

$$f(x_0, x_1, \ldots, x_n) =$$

$$= \int_0^1 dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{n-1}} dt_n f^{(n)}\left[x_0 + \sum_{i=1}^{n} t_i (x_i - x_{i-1})\right]. \qquad (5)$$

Так как аргумент производной $f^{(n)}(x)$, стоящей под знаком интегралов, можно представить в форме

$$x = (1-t_1)x_0 + (t_1-t_2)x_1 + \ldots + (t_{n-1}-t_n)x_{n-1} + t_n x_n = \\ = \frac{(1-t_1)x_0 + (t_1-t_2)x_1 + \ldots + (t_{n-1}-t_n)x_{n-1} + t_n x_n}{(1-t_1)+(t_1-t_2)+\ldots+(t_{n-1}-t_n)+t_n},$$

а область интегрирования определяется неравенствами $0 \leqslant t_n \leqslant t_{n-1} \leqslant \ldots \leqslant t_1 \leqslant 1$, то x есть среднее взвешенное значение, составленное из x_i $(i = 0, 1, \ldots, n)$ с неотрицательными весами. Поэтому x принадлежит отрезку $[a, b]$, и интеграл в (5) имеет смысл.

Ограничимся проверкой равенства (5) для $n = 1$ и $n = 2$. При $n = 1$ интеграл (5) вычисляется просто:

$$\int_0^1 dt\, f'[x_0 + t(x_1 - x_0)] = \left| \frac{1}{x_1 - x_0} f[x_0 + t(x_1 - x_0)] \right|_0^1 = \\ = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f(x_0, x_1),$$

и равенство (5) верно.

Для $n = 2$, если сначала выполнить интегрирование по t_2, а затем по t_1, получим

$$\int_0^1 dt_1 \int_0^{t_1} dt_2 f''[x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1)] = \\ = \int_0^1 dt_1 \left| \frac{1}{x_2 - x_1} f'[x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1)] \right|_0^{t_1} = \\ = \frac{1}{x_2 - x_1} \left\{ \int_0^1 dt_1 f'[x_0 + t_1(x_2 - x_0)] - \int_0^1 dt_1 f'[x_0 + t_1(x_1 - x_0)] \right\} = \\ = \frac{f(x_0, x_2) - f(x_1, x_0)}{x_2 - x_1} = f(x_1, x_0, x_2) = f(x_0, x_1, x_2),$$

и равенство (5) также верно.

Более простую, но менее точную связь между величинами $f(x_0, x_1, \ldots, x_n)$ и $f^{(n)}(x)$ дает

Теорема 2. *Если узлы $x_0, x_1, \ldots, x_n$ принадлежат отрезку $[a, b]$ и f имеет на $[a, b]$ непрерывную производ-*

ную порядка n, то на $[a, b]$ *существует такая точка* ξ, *для которой верно равенство*

$$f(x_0, x_1, \ldots, x_n) = \frac{1}{n!} f^{(n)}(\xi). \qquad (6)$$

(Недостатком (6) является то обстоятельство, что теорема не сообщает никаких сведений о положении точки ξ на $[a, b]$ и утверждает лишь существование такой точки.)

Доказательство. Достаточно применить к интегралу (5) теорему о среднем значении: он равен произведению значения интегрируемой функции в некоторой точке области интегрирования на величину объема области.

Выше было отмечено, что для любых t_i из области $0 \leqslant t_n \leqslant \ldots \leqslant t_1 \leqslant 1$ аргумент $f^{(n)}$ в интеграле принадлежит отрезку $[a, b]$. Обозначим его значение в указанной выше точке буквой ξ, так что значение интегрируемой функции в этой точке будет $f^{(n)}(\xi)$. Величина же объема области численно равна интегралу по области от единицы и просто вычисляется:

$$\int_0^1 dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{n-1}} dt_n = \frac{1}{n!}.$$

Поэтому применение теоремы о среднем значении к интегралу (5) приводит к (6).

Отметим одно простое следствие, вытекающее из (5) или (6). Пусть f есть многочлен степени n: $f(x) = a_0 x^n + a_1 x^{n-1} + \ldots$ Производная от него порядка n является постоянной величиной: $f^{(n)}(x) = n!\, a_0$, и равенства (5) и (6) для разностного отношения порядка n дадут значение

$$f(x_0, x_1, \ldots, x_n) = \frac{1}{n!} n!\, a_0 = a_0.$$

Заметим также, что все разностные отношения порядков $n+1$, $n+2$, ... будут равны нулю, и изложенное позволяет высказать приводимое ниже утверждение. Если f есть многочлен степени n от x, то его разностное отношение порядка n не зависит от положения узлов и равно коэффициенту при наивысшей степени n; разностные же отношения порядка выше n будут все равны нулю.

Приведем еще выражение любого значения $f(x_n)$ функции через начальное значение $f(x_0)$ и разностные отношения $f(x_0, x_1)$, $f(x_0, x_1, x_2)$, ... для начальной точки x_0. Можно показать, что при всяком $n = 1, 2, \ldots$ имеет место равенство

$$f(x_n) = f(x_0) + (x_n - x_0) f(x_0, x_1) + \\ + (x_n - x_0)(x_n - x_1) f(x_0, x_1, x_2) + \ldots \\ \ldots + (x_n - x_0) \ldots (x_n - x_{n-1}) f(x_0, x_1, \ldots, x_n). \quad (7)$$

Доказательство может быть получено с помощью несложной индукции и ввиду простоты вопроса мы ограничимся проверкой равенства для $n = 1$ и $n = 2$. По определению разностного отношения первого порядка $f(x_0, x_1) = \frac{1}{x_1 - x_0}[f(x_1) - f(x_0)]$, откуда $f(x_1) = f(x_0) + (x_1 - x_0) f(x_0, x_1)$, что доказывает правильность (7) для $n = 1$. На основании этого и на основании определения $f(x_0, x_1, x_2)$ можно написать:

$$f(x_2) = f(x_1) + (x_2 - x_1) f(x_1, x_2) = \\ = [f(x_0) + (x_1 - x_0) f(x_0, x_1)] + \\ + (x_2 - x_1)[f(x_0, x_1) + (x_2 - x_0) f(x_0, x_1, x_2)] = \\ = f(x_0) + (x_1 - x_0) f(x_0, x_1) + \\ + (x_2 - x_0)(x_2 - x_1) f(x_0, x_1, x_2).$$

Из этого результата следует, что равенство (7) верно для $n = 2$.

Когда значения аргумента x являются равноотстоящими, то разностные отношения должны быть связаны с конечными разностями. Пусть $x_k = x_0 + kh$ ($k = 0, 1, \ldots$) и известны значения $f(x_k) = f(x_0 + kh) = y_k$. Тогда

$$f(x_0, x_1) = f(x_0, x_0 + h) = \frac{f(x_0 + h) - f(x_0)}{x_0 + h - x_0} = \frac{\Delta y_0}{1!h}.$$

Для разностного отношения второго порядка

$$f(x_0, x_1, x_2) = f(x_0, x_0 + h, x_0 + 2h) = \\ = \frac{f(x_1, x_2) - f(x_0, x_1)}{x_2 - x_0} = \frac{1}{2h}\left(\frac{\Delta y_1}{1!h} - \frac{\Delta y_0}{1!h}\right) = \frac{\Delta^2 y_0}{2!h^2}$$

и т. д. При любом $n = 1, 2, \ldots$ будет

$$f(x_0, x_0 + h, \ldots, x_0 + nh) = \frac{\Delta^n y_0}{h^n n!}. \qquad (8)$$

Отсюда и из равенств (5) и (6) сразу же получается представление конечной разности через производную порядка n от f, которое мы сформулируем в виде теоремы.

Теорема 3. *Если функция f имеет непрерывную производную порядка n на отрезке $[x_0, x_0 + nh]$, то для $\Delta^n y_0$ верно представление*

$$\Delta^n y_0 = h^n n! \int_0^1 dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{n-1}} dt_n f^{(n)}\left(x_0 + h\sum_{i=1}^{n} t_i\right). \qquad (9)$$

Кроме того, на отрезке $[x_0, x_0 + nh]$ существует точка ξ такая, что для $\Delta^n y_0$ верно равенство

$$\Delta^n y_0 = h^n f^{(n)}(\xi). \qquad (10)$$

Обе приведенные формулы делают очевидным заключение о малости конечной разности при малом шаге h и позволяют сказать, что если шаг h есть бесконечно малая величина, то $\Delta^n y_0$ есть величина бесконечно малая порядка n сравнительно с h, когда $f^{(n)}(x_0) \neq 0$, и есть бесконечно малая величина порядка выше n, когда $f^{(n)}(x_0) = 0$.

§ 3. Алгебраическое интерполирование функций

В этом и в ближайших следующих параграфах будет рассматриваться задача интерполирования функции по нескольким ее значениям при помощи алгебраического многочлена. Сначала будут получены необходимые представления интерполирующего многочлена и погрешности интерполирования при произвольно расположенных узлах.

1. Некоторые представления интерполирующего многочлена. Пусть на отрезке $[a, b]$ рассматривается функция f, и пусть известны ее значения в $n + 1$ различных узлах $x_0, x_1, \ldots, x_n$, принадлежащих $[a, b]$. Эти значения

мы будем считать любыми конечными. Возьмем многочлен степени n:

$$P_n(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n, \qquad (1)$$

и его коэффициенты a_k $(k = 0, 1, \ldots, n)$ выберем так, чтобы выполнялись условия совпадения значений f и P_n в узлах x_i $(i = 0, 1, \ldots, n)$:

$$P_n(x_i) = f(x_i) \quad (i = 0, 1, \ldots, n). \qquad (2)$$

Такие равенства дают систему $n+1$ линейных уравнений для нахождения коэффициентов a_k.

В § 1 п. 1 мы обращали внимание на то, что определитель системы (2) отличен от нуля, и система имеет единственное решение при любых значениях правых частей $f(x_i)$ $(i = 0, 1, \ldots, n)$. Если a_k найти из системы (2) и подставить их значения в (1), получим явное выражение для $P_n(x)$. Последнее можно просто выписать без решения системы. Присоединим равенство (1) к системе (2) и запишем полученную новую систему в виде

$$\begin{aligned}
&-P_n(x) + a_0x^n + a_1x^{n-1} + \ldots + a_n = 0,\\
&-f(x_0) + a_0x_0^n + a_1x_0^{n-1} + \ldots + a_n = 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&-f(x_n) + a_0x_n^n + a_1x_n^{n-1} + \ldots + a_n = 0.
\end{aligned}$$

Ее можно рассматривать как однородную систему относительно $n+1$ неизвестных $-1, a_0, a_1, \ldots, a_n$, и так как они образуют ненулевое решение системы, то определитель системы должен быть равен нулю:

$$\begin{vmatrix}
P_n(x) & x^n & x^{n-1} & \ldots & 1\\
f(x_0) & x_0^n & x_0^{n-1} & \ldots & 1\\
\ldots & \ldots & \ldots & \ldots & \ldots\\
f(x_n) & x_n^n & x_n^{n-1} & \ldots & 1
\end{vmatrix} = 0.$$

Если первый столбец рассматривать как сумму двух столбцов: одного с элементами $[P_n(x), 0, \ldots, 0]$, а второго — с элементами $[0, f(x_0), \ldots, f(x_n)]$, и воспользоваться известной из курсов линейной алгебры теоремой

о сложении определителей, то получится приводимое ниже выражение для $P_n(x)$:

$$P_n(x) = -\frac{1}{W(x_0, x_1, \ldots, x_n)} \begin{vmatrix} 0 & x^n & x^{n-1} & \ldots & 1 \\ f(x_0) & x_0^n & x_0^{n-1} & \ldots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ f(x_n) & x_n^n & x_n^{n-1} & \ldots & 1 \end{vmatrix}, \quad (3)$$

$$W(x_0, x_1, \ldots, x_n) = \begin{vmatrix} x_0^n & x_0^{n-1} & \ldots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ x_n^n & x_n^{n-1} & \ldots & 1 \end{vmatrix}.$$

Оно редко применяется в приложениях, так как требует вычисления определителей. Укажем два других выражения $P_n(x)$, свободных от такого недостатка.

Из системы (2) видно, что коэффициенты a_k будут линейно зависеть от значений $f(x_i)$ $(i = 0, 1, \ldots, n)$. Поэтому и многочлен $P_n(x)$ линейно зависит от величин $f(x_i)$ и представим, следовательно, в форме

$$P_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i). \quad (4)$$

Такое выражение для P_n можно получить, например, если разложить определитель в (3) по элементам первого столбца. Но коэффициенты $l_i(x)$ можно найти, не выполняя операции разложения, а используя простые алгебраические соображения. Рассмотрим $l_k(x)$. Он будет совпадать с $P_n(x)$, как следует из (4), если функция $f(x)$, а стало быть, и многочлен $P_n(x)$ будут обладать следующими свойствами:

$$P_n(x_i) = f(x_i) = 0 \quad (i \neq k) \quad \text{и} \quad P_n(x_k) = f(x_k) = 1.$$

Таким образом, можно сказать, что $l_k(x)$ есть многочлен степени n, для которого все узлы x_i $(i = 1, \ldots, n;$ $i \neq k)$ являются корнями. Отметим, что все эти корни должны быть однократными, так как их число n такое же, как степень многочлена.

Если известны корни многочлена и их кратности, можно записать разложение многочлена на множители:

$$l_k(x) = C_k(x - x_0) \ldots (x - x_{k-1})(x - x_{k+1}) \ldots (x - x_n),$$

Постоянный множитель C_k может быть определен из условия $l_k(x_k) = 1$, что дает $C_k = [(x_k - x_0) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_n)]^{-1}$ и, следовательно,

$$l_k(x) = \frac{(x - x_0) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_n)}{(x_k - x_0) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_n)}.$$

Введем многочлен степени $n + 1$, положив $\omega(x) = (x - x_0)(x - x_1) \dots (x - x_n)$. Он связан с расположением всех узлов интерполирования, которые для него являются корнями первой кратности. При помощи $\omega(x)$ многочлен $l_k(x)$ записывается в виде

$$l_k(x) = \frac{\omega(x)}{(x - x_k)\omega'(x_k)},$$

и представление (4) интерполирующего многочлена $P_n(x)$ будет следующим:

$$P_n(x) = \sum_{k=0}^{n} l_k(x) f(x_k) = \sum_{k=0}^{n} \frac{\omega(x)}{(x - x_k)\omega'(x_k)} f(x_k). \quad (5)$$

Это равенство называют *формулой Лагранжа* для интерполирующего многочлена P_n, а множители $l_k(x)$ называют *лагранжевыми многочленами влияния* соответствующих узлов интерполирования или, более сокращенно, *множителями Лагранжа*.

Приведем еще *ньютоново представление* интерполирующего многочлена через разностные отношения:

$$P_n(x) = f(x_0) + (x - x_0) f(x_0, x_1) + \\ + (x - x_0)(x - x_1) f(x_0, x_1, x_2) + \dots \\ \dots + (x - x_0)(x - x_1) \dots (x - x_{n-1}) f(x_0, x_1, \dots, x_n). \quad (6)$$

В правильности этого равенства можно просто убедиться проверкой того, что многочлен (6) удовлетворяет всем требованиям, которые предъявляются к интерполирующему многочлену. Самую высокую степень x может содержать последний член правой части, и эта степень равна n. Поэтому правая часть есть многочлен либо степени n, либо меньшей степени, когда последний член отсутствует.

Остается еще проверить, что многочлен (6) удовлетворяет условиям (2). Это можно сделать без труда, вос-

пользовавшись выражением (2.7) для любого значения $f(x_k)$ $(k = 0, 1, \ldots, n)$ через разностные отношения. Если в (6) положить $x = x_0$, то равенство примет форму $P_n(x_0) = f(x_0)$, и многочлен (6) действительно удовлетворяет условиям (2) для $i = 0$. Если считать $x = x_1$, получим $P_n(x_1) = f(x_0) + (x_1 - x_0)f(x_0, x_1) = f(x_1)$; при этом последнее из равенств следует из (2.7) при $n = 1$. Полагая затем $x = x_2, x_3, \ldots, x_n$ и пользуясь соотношением (2.7) для $n = 2, 3, \ldots$, убедимся в том, что многочлен (6) удовлетворяет всем условиям (2).

Полезно сравнить между собой лагранжево и ньютоново представления $P_n(x)$. В формуле (5) в слагаемых суммы множители $l_k(x)$ зависят от выбора узлов x_i и точки x и не зависят от функции f. Множители же $f(x_i)$ позволяют учитывать влияние на $P_n(x)$ свойств функции f и ее значений. Это удобно в двух отношениях. Во-первых, когда по одной системе узлов x_i $(i = 0, \ldots, n)$ необходимо интерполировать несколько функций. Тогда можно вычислить множители $l_k(x)$ однажды и использовать их для всех функций f. Во-вторых, указанная «разделенность» влияния на $P_n(x)$ выбора узлов x_k и свойств функции f бывает полезной при изучении сходимости $P_n(x)$ к $f(x)$ при $n \to \infty$. Действительно, если эти факторы разделены, легче наблюдать за их влиянием и оценивать погрешность приближения $P_n(x)$ к $f(x)$.

Ньютоново представление (6) значительно менее удобно для исследований в такого рода вопросах, так как разностные отношения $f(x_0, x_1, \ldots, x_k)$ зависят от расположения узлов x_i и свойств f достаточно сложно. Это очень затрудняет использование формулы Ньютона в исследовании теоретических вопросов. Но формула Ньютона обладает другими чертами, делающими ее весьма полезной во многих вычислительных вопросах.

Когда собираются проводить интерполяционные вычисления, то, прежде всего, исходя из опыта и других соображений, выбирают число и расположение узлов, при которых можно ожидать получения принятой точности, и стараются обойтись возможно малым числом их. При этом редко бывает, что заранее можно гарантировать получение нужной точности, и поэтому необходимо бывает выполнить проверку результата и увеличение его точности, если она окажется недостаточной.

Делают это весьма часто путем добавления к взятым узлам $x_0, x_1, \ldots, x_n$ еще одного или нескольких новых узлов. Для определенности будем говорить о присоединении одного нового узла. В формуле Лагранжа это повлечет за собой не только добавление нового слагаемого в сумме (5), но потребует также исправления всех ранее найденных членов суммы. В формуле же Ньютона (6) при переходе от $n+1$ узлов к $n+2$ узлам все уже найденные члены сохраняются, и потребуется лишь добавление члена $(x-x_0)\ldots(x-x_n)f(x_0, \ldots, x_n, x_{n+1})$, имеющего смысл поправки к уже вычисленному значению.

Отметим также, что в вычислительной практике нередко, особенно при пользовании таблицами, приходится интерполировать на малых участках, когда узлы x_i $(i=0, 1, \ldots, n)$ и точка x принадлежат малому отрезку около точки x_0, длину которого обозначим h. В формуле Ньютона множители $(x-x_0)$, $(x-x_0)(x-x_1)$, ... будут величинами порядков соответственно $h, h^2, \ldots$, тогда как разностные отношения будут близки (что вытекает из равенства (2.6)) к величинам $f'(x_0)$, $\frac{1}{2!}f''(x_0)$, $\frac{1}{3!}f'''(x_0)$, ... Поэтому в (6) члены будут расположены в порядке их малости. Этот факт облегчает использование формулы Ньютона в вычислениях и в суждении о точности.

2. Погрешность интерполирования и ее представления для некоторых классов функций. Выше мы обращали внимание на то, что погрешность интерполирования

$$R_n(x) = f(x) - P_n(x) \tag{7}$$

зависит от многих фактов. Сейчас нас будет интересовать ее зависимость от свойств функции f. Напомним также, что приведенные в п. 1 представления (3), (5) и (6) многочлена P_n были получены при весьма общих предположениях: узлы x_i $(i=0, 1, \ldots, n)$ считались различными между собой, а функция f — имеющей конечные значения $f(x_i)$ в узлах, а в остальном произвольной.

Если в (7) внести вместо $P_n(x)$ любое из указанных трех выражений, то для погрешности $R_n(x)$ получится одно из возможных представлений, верное для всяких функций с конечными значениями $f(x_i)$ $(i = 0, 1, \ldots, n)$. Например, внесем в (7) вместо $P_n(x)$ лагранжево представление (5). Примем во внимание следующий простой факт. Пусть f есть многочлен степени не выше n. Интерполирующий многочлен P_n имеет одинаковое значение с ним в $n+1$ узлах x_i, и разность $f - P_n$ будет многочленом степени не больше n, обращающимся в нуль по меньшей мере в $n+1$ точке. Но тогда многочлен $f - P_n$ должен быть тождественным нулем, и f совпадает с P_n при всех значениях x. В частности, когда $f = 1$, будет выполняться равенство $1 = \sum_{i=0}^{n} l_i(x)$. С учетом этого погрешность интерполяции равна

$$R_n(x) = f(x) - \sum_{i=0}^{n} l_i(x) f(x_i) = \sum_{i=0}^{n} l_i(x)[f(x) - f(x_i)]. \quad (8)$$

Построим еще одно представление $R_n(x)$, также верное для всяких функций f с конечными значениями в узлах x_i и полезное нам для получения выражений погрешности $R_n(x)$ и ее оценок в классах дифференцируемых функций.

Рассмотрим последовательность значений $x_0, x_1, \ldots, x_n, x_{n+1} = x$ и применим для вычисления $f(x_{n+1}) = f(x)$ равенство (2.7), заменив в нем предварительно n на $n+1$:

$$f(x) = f(x_0) + (x - x_0) f(x_0, x_1) + \ldots$$
$$\ldots + (x - x_0) \ldots (x - x_{n-1}) f(x_0, x_1, \ldots, x_n) +$$
$$+ (x - x_0) \ldots (x - x_{n-1})(x - x_n) f(x_0, x_1, \ldots, x_n, x).$$

Сумма всех членов правой части равенства, кроме последнего, есть не что иное, как интерполяционный многочлен $P_n(x)$ в форме Ньютона (6). Поэтому последний член является погрешностью

$$R_n(x) = \omega(x) f(x_0, x_1, \ldots, x_n, x). \quad (9)$$

Множитель $\omega(x)$ зависит только от x и узлов x_i. Разностное же отношение не может быть вычислено, так как зависит от x и $f(x)$. Но в некоторых случаях может

быть получено представление о приближенном значении $f(x_0, \ldots, x_n, x)$. Так, например, будет, если функция f задана таблично и размеры таблицы позволяют вычислить одно или несколько значений этого разностного отношения путем замены в нем x табличным значением аргумента, близким *) к x. Такие вспомогательные значения разностного отношения позволяют составить некоторое представление об $R_n(x)$, хотя и неточное.

Равенство (9) дает возможность получить представления $R_n(x)$, рассчитанные на классы функций высокого порядка гладкости. Предположим, что узлы x_i $(i = 0, 1, \ldots, n)$ и точка x принадлежат отрезку $[a, b]$, и функция f имеет на $[a, b]$ непрерывную производную порядка $n+1$. В этих условиях для $f(x_0, x_1, \ldots, x_n, x)$ верны представления вида (2.5) и (2.6), если заменить в них n на $n+1$ и положить $x_{n+1} = x$. Это дает возможность высказать две приводимые ниже теоремы о представлении погрешности интерполирования.

Теорема 1. *Пусть выполнены условия*:

1) *точки* $x_0, x_1, \ldots, x_n$ *различны между собой и принадлежат отрезку* $[a, b]$;

2) *функция* f *имеет на* $[a, b]$ *непрерывную производную порядка* $n+1$.

Тогда погрешность $R_n(x)$ *интерполирования* f *по ее значениям в точках* x_i $(i = 0, 1, \ldots, n)$ *может быть представлена в форме*

$$R_n(x) = \omega(x) \int_0^1 dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{n-1}} dt_n f^{(n+1)}\left[x_0 + \sum_{\nu=1}^{n+1} t_\nu (x_\nu - x_{\nu-1})\right], \quad (10)$$

$$x_{n+1} = x, \quad \omega(x) = (x - x_0) \ldots (x - x_n).$$

Теорема 2. *Если выполнены условия предыдущей теоремы, то на* $[a, b]$ *существует такая точка* ξ, *что для погрешности интерполирования верно равенство*

$$R_n(x) = \frac{\omega(x)}{(n+1)!} f^{(n+1)}(\xi). \quad (11)$$

*) Предполагается, что $f(x_0, \ldots, x_n, x)$ мало изменяется при малых изменениях x, и табличное значение аргумента, заменяющее x, лежит близко к x.

Недостатком (11) является то обстоятельство, что о положении точки ξ на $[a, b]$ невозможно сказать ничего определенного.

Обратим внимание на одну из интерполяционных задач, связанных с построенными представлениями R_n. Пусть нужно выполнить интерполирование функции во всех точках отрезка $[a, b]$. Погрешность его зависит от выбора узлов x_i, точки x и свойств функций f.

Если интерполируется одна определенная функция f, то точность интерполирования характеризуется величиной $\max\limits_x |R_n(x)|$. Когда мы интерполируем не одну функцию f, а некоторое множество функций f, то точность может быть оценена величиной

$$\sup_f \max_x |R_n(x)| = m = m(x_0, x_1, \ldots, x_n). \qquad (12)$$

Эта величина зависит только от выбора узлов x_i $(i = 0, 1, \ldots, n)$.

Поставим задачу о выборе узлов x_i, которые можно было бы считать наилучшими при интерполировании всех функций f на $[a, b]$ из взятого множества. Такими узлами естественно считать те, для которых величина $m(x_0, x_1, \ldots, x_n)$ достигает наименьшего значения.

Найдем такие узлы для множества всех функций с непрерывной производной порядка $n+1$ на $[a, b]$. Изменим на время эту задачу и рассмотрим функции f, для которых при некотором произвольно взятом положительном M выполняется неравенство

$$|f^{(n+1)}(x)| \leqslant M. \qquad (13)$$

Для таких функций погрешность $R_n(x)$ может быть оценена следующим неравенством, которое сразу вытекает из (11):

$$\max_x |R_n(x)| \leqslant \frac{M}{(n+1)!} \max_x |\omega(x)|.$$

Эта оценка является неулучшаемой, так как в ней имеет место знак равенства, когда f есть следующий многочлен степени $n+1$:

$$f = \frac{M}{(n+1)!} x^{n+1} + c_1 x^n + \ldots$$

и, следовательно,

$$\sup_f \max_x |R_n(x)| = \frac{M}{(n+1)!} \max_x |\omega(x)|. \qquad (14)$$

Первый множитель правой части (14) не зависит от выбора узлов x_i, и поэтому наилучшими узлами при интерполировании функций f, удовлетворяющих условию (13), нужно признать те x_i, для которых

$$\max_x |\omega(x)| = \min. \qquad (15)$$

Это заключение верно при всяких M в неравенстве (13). Можно поэтому утверждать, что такие узлы будут наилучшими при интерполировании всяких функций f с непрерывной производной порядка $n+1$ на $[a, b]$.

Задача, указанная в условии (15), имеет наглядный смысл. Величина $\max_x |\omega(x)|$ есть отклонение от нуля на $[a, b]$ многочлена

$$\omega(x) = (x - x_0)(x - x_1) \ldots (x - x_n) = \\ = x^{n+1} + b_1 x^n + b_2 x^{n-1} + \ldots \qquad (16)$$

В задаче рассматриваются все многочлены $\omega(x)$, для которых корни x_i различны и все лежат на $[a, b]$. Среди таких многочленов нужно найти тот, который имеет наименьшее отклонение от нуля. Такой многочлен хорошо известен в конструктивной теории функций; им является многочлен Чебышева первого рода. Наиболее простой вид явное выражение для него принимает в случае отрезка *) $[-1, 1]$:

$$T_{n+1}(x) = \frac{1}{2^n} \cos[(n+1) \arccos x] = x^{n+1} + c_1 x^n + \ldots$$

Корни его вычисляются просто:

$$x_k = \cos \frac{2k+1}{2(n+1)} \pi \qquad (k = 0, 1, \ldots, n).$$

---

*) Переход к отрезку $[-1, 1]$ не является ограничением, так как всякий отрезок $a \leqslant x \leqslant b$ приводится к $[-1, 1]$ линейным преобразованием $x = \frac{1}{2}[a+b] + \frac{1}{2}[b-a]x'$, $-1 \leqslant x' \leqslant 1$.

Все они лежат внутри $[-1, 1]$ и различны между собой *).

Остановимся еще на представлении погрешности интерполирования аналитических функций, имеющих большое значение в прикладных вопросах. Пусть область D комплексной плоскости содержит внутри себя отрезок $[a, b]$ действительной оси, и функция $f(z)$ однозначна и регулярна в D, включая и ее контур l.

На $[a, b]$ возьмем $n+1$ различный узел $x_0, x_1, \ldots, x_n$ и интерполируем $f(z)$ по ее значениям $f(x_k)$ $(k = 0, 1, \ldots, n)$ алгебраическим многочленом степени n:

$$P_n(z) = \sum_{k=0}^{n} \frac{\omega(z)}{(z - x_k)\omega'(x_k)} f(x_k), \quad \omega(z) = (z - x_0)\ldots(z - x_n). \tag{17}$$

Нас будет интересовать погрешность интерполирования $R_n(z) = f(z) - P_n(z)$. Докажем теорему об ее представлении контурным интегралом.

Теорема 3. *Пусть для f, D и l выполняются высказанные выше предположения. При всяком z, лежащем внутри D, для погрешности $R_n(z)$ верно представление*

$$R_n(z) = \frac{\omega(z)}{2\pi i} \int_l \frac{f(t)}{\omega(t)(t - z)}\, dt. \tag{18}$$

Доказательство. Достаточно вычислить интеграл, стоящий справа, и убедиться в том, что он равен $f(z) - P_n(z)$. Если оставить в стороне множитель $\omega(z)$, то оставшийся контурный интеграл будет равен сумме вычетов функции $\frac{f(t)}{\omega(t)(t - z)}$ в особых точках, лежащих внутри D. Функция $f(z)$ не имеет в D особых точек, и такими точками будут только нули z, $x_0, \ldots, x_n$ знаменателя. При изучении $R_n(z)$ представляют интерес значения z, отличные от узлов x_k. Тогда все нули

*) Многочлен $T_{n+1}(x)$ есть решение следующей задачи: среди многочленов вида $x^{n+1} + b_1 x^n + \ldots$ найти тот, который имеет наименьшее отклонение от нуля на $[-1, 1]$. Дополнительное условие, чтобы корни многочлена лежали на отрезке $[-1, 1]$ и были различны, имеющееся в нашей задаче, здесь отсутствует. Но так как корни $T_{n+1}(x)$ лежат все внутри $[-1, 1]$ и различны, он решает и нашу задачу (15) на отрезке $[-1, 1]$.

знаменателя будут однократными, и вычеты могут быть найдены по известным правилам их вычисления:

$$\text{выч}\left[\frac{f(t)}{\omega(t)(t-z)}\right]_{t=z} = \frac{f(z)}{\omega(z)},$$

$$\text{выч}\left[\frac{f(t)}{\omega(t)(t-z)}\right]_{t=x_k} = \frac{f(x_k)}{\omega'(x_k)(x_k-z)} =$$

$$= -\frac{1}{(z-x_k)\omega'(x_k)} f(x_k).$$

Эти результаты дают для контурного интеграла следующее значение:

$$\frac{\omega(z)}{2\pi i}\int_l \frac{f(t)}{\omega(t)(t-z)}\,dt = f(z) - \sum_{k=0}^{n} \frac{\omega(z)}{(z-x_k)\omega'(x_k)} f(x_k) =$$

$$= f(z) - P_n(z) = R_n(z).$$

Утверждение теоремы этим доказано.

Получим при помощи (18) одну из простейших оценок погрешности $R_n(z)$ интерполирования аналитических функций. В представлении (18) от числа узлов x_k и их расположения на $[a, b]$ зависит величина

$$\frac{\omega(z)}{\omega(t)} = \prod_{k=0}^{n}\left(\frac{z-x_k}{t-x_k}\right).$$

Обозначим буквой δ наибольшее расстояние от z до точек отрезка $[a, b]$ и буквой r — расстояние от $[a, b]$ до l. Тогда при всякой точке t на l и для любого узла x_k будет верно неравенство $\left|\frac{z-x_k}{t-x_k}\right| \leqslant \frac{\delta}{r}$. Для $R_n(z)$ получится оценка

$$|R_n(z)| \leqslant \left(\frac{\delta}{r}\right)^{n+1} \frac{1}{2\pi}\int_l \frac{|f(t)|}{|t-z|}\,ds, \qquad ds = |dt|. \quad (19)$$

Отсюда видно, что если область D регулярности f будет достаточно широкой около $[a, b]$, а точка z расположена вблизи от $[a, b]$, то отношение δ/r будет меньше 1. В этом случае при неограниченном увеличении числа узлов $n+1$ погрешность $R_n(z)$ будет стремиться к нулю при всяком расположении узлов x_k на $[a, b]$, и интерполяционный процесс будет сходиться к $f(z)$ с такой же

по меньшей мере скоростью, как стремится к нулю $(\delta/r)^n$.

Например, пусть рассматривается сходимость интерполирования на отрезке $[a, b]$, так что точка z будет лежать на этом отрезке: $z = x \in [a, b]$. При всяком положении x и для всякого узла x_k верно неравенство $|x - x_k| \leqslant b - a = \delta$.

Рассмотрим линию λ, все точки которой удалены от $[a, b]$ на расстояние $b - a$. Эта линия состоит из двух полуокружностей радиуса $b-a$ с центрами в точках a и b и из двух прямолинейных отрезков $y = \pm(b - a)$ $(a \leqslant x \leqslant b)$, соединяющих полуокружности (рис. 1). Предположим, что $f(z)$ регулярна в области, ограниченной линией λ, и на самой линии λ. Тогда в качестве контура интегрирования l может быть взята линия, охватывающая λ, и расстояние от нее до $[a, b]$ будет равно $r = b - a + \gamma$ $(\gamma > 0)$, т. е. больше, чем $b - a$.

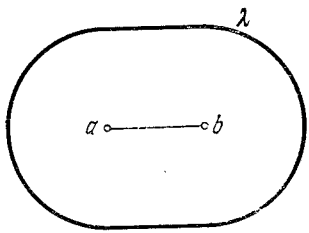

Рис. 1.

Отношение δ/r в этом случае будет следующим:

$$\frac{\delta}{r} = \frac{b-a}{b-a+\gamma} = q < 1,$$

и из оценки (19) вытекает

Теорема 4. *Если $f(z)$ есть аналитическая функция, регулярная в замкнутой области, ограниченной линией λ, то интерполяционный процесс для нее при неограниченном возрастании числа узлов x_k и любом расположении их на отрезке $[a, b]$ будет сходиться к $f(x)$ равномерно относительно x на $[a, b]$.*

При этом погрешность интерполирования $R_n(x)$ будет стремиться к нулю не медленнее, чем q^n $(n \to \infty)$.

§ 4. Интерполирование при равноотстоящих значениях аргумента

Такое интерполирование встречается в приложениях очень часто, выполнять его приходится во многих условиях и для разных целей, и в зависимости от них было

построено много интерполяционных формул. Мы ограничимся изложением небольшого набора этих формул, которого достаточно для вычислений в большинстве практически важных случаев.

1. Интерполирование в начале и в конце таблицы. Пусть функция f задана таблицей своих значений $f(x_k) = f(x_0 + kh) = y_k$ в равноотстоящих точках $x_k = x_0 + kh$ $(k = 0, 1, 2, \ldots)$, и точка интерполирования x находится близко от начальной точки x_0 или в любом месте слева от x_0.

При составлении формулы интерполирования воспользуемся формулой Ньютона (3.6). В рассматриваемой задаче для составления формулы узлы естественно брать в порядке их расположения в таблице: x_0, x_0+h, x_0+2h, ...:

$$f(x) = P_n(x) + R_n(x) = f(x_0) + (x - x_0) f(x_0, x_0 + h) + \\ + (x - x_0)(x - x_0 - h) f(x_0, x_0 + h, x_0 + 2h) + \ldots \\ \ldots + (x - x_0)(x - x_0 - h) \ldots \\ \ldots (x - x_0 - (k-1)h) f(x_0, x_0 + h, \ldots, x_0 + kh) + R_n(x).$$

Разностные отношения, стоящие в правой части равенства, выражаются через конечные разности функции $y = f$ при помощи соотношений вида (2.8)

$$f(x_0) = y_0, \quad f(x_0, x_0 + h) = \frac{\Delta y_0}{1!h},$$

$$f(x_0, x_0 + h, x_0 + 2h) = \frac{\Delta^2 y_0}{2!h^2}, \ldots$$

Введем новую переменную t, положив $x = x_0 + th$, $t = \frac{x - x_0}{h}$. Она имеет значение числа шагов h от x_0 до x:

$$x - x_0 = th, \quad (x - x_0)(x - x_0 - h) = t(t-1)h^2, \ldots$$

После внесения указанных величин в выражение для $f(x)$, получим формулу Ньютона для интерполирования вблизи начала таблицы:

$$y(x_0 + th) = y_0 + \frac{t}{1!}\Delta y_0 + \frac{t(t-1)}{2!}\Delta^2 y_0 + \ldots \\ \ldots + \frac{t(t-1)\ldots(t-k+1)}{k!}\Delta^k y_0 + R_k(x). \quad (1)$$

Остаточный член формулы $R_n(x)$ может быть представлен в любой из форм, указанных в § 2 п. 3, в зависимости от свойств функции. Мы ограничимся тем, что запишем его в лагранжевой форме (3.11). Пусть функция $y = f(x)$ имеет непрерывную производную порядка $k+1$ на отрезке $[a, b]$, содержащем точку x и узлы от x_0 до $x_k = x_0 + kh$. Тогда для $R_k(x)$ верно представление (3.11) с заменой n на k. В переменной t многочлен $\omega(x)$ имеет вид

$$\omega(x) = (x - x_0)(x - x_0 - h) \ldots (x - x_0 - kh) = \\ = h^{k+1} t(t-1) \ldots (t-k),$$

и для остаточного члена $R_k(x)$ в (1) получится

$$R_k(x) = h^{k+1} \frac{t(t-1) \ldots (t-k)}{(k+1)!} y^{(k+1)}(\xi), \qquad (2)$$

где ξ есть некоторая точка отрезка $[a, b]$, указанного выше.

Допустим теперь, что точка интерполирования лежит вблизи конечной точки x_n таблицы или где-то справа от нее. В этом случае узлы интерполирования следует брать в порядке x_n, $x_n - h$, $x_n - 2h$, ... Формула Ньютона тогда запишется в следующем виде:

$$f(x) = f(x_n) + (x - x_n) f(x_n, x_n - h) + \\ + (x - x_n)(x - x_n + h) f(x_n, x_n - h, x_n - 2h) + \ldots \\ \ldots + (x - x_n)(x - x_n + h) \ldots \\ \ldots (x - x_n + (k-1)h) f(x_n, x_n - h, \ldots, x_n - kh) + R_k(x).$$

Разностные отношения могут быть выражены через конечные разности, если воспользоваться возможностью переставлять в них аргументы и соотношением (2.8):

$$f(x_n) = y_n, \quad f(x_n, x_n - h) = f(x_{n-1}, x_n) = \frac{\Delta y_{n-1}}{1!h},$$

$$f(x_n, x_n - h, x_n - 2h) = \\ = f(x_n - 2h, x_n - h, x_n) = \frac{\Delta^2 y_{n-2}}{2!h^2}, \ldots$$

Введя переменную t и положив $x = x_n + tn$, получим для $f(x) = y(x)$ формулу Ньютона для интерполирования в конце таблицы или справа от x_n:

$$y(x_n + th) = y_n + \frac{t}{1!}\Delta y_{n-1} + \frac{t(t+1)}{2!}\Delta^2 y_{n-2} + \ldots$$
$$\ldots + \frac{t(t+1)\ldots(t+k-1)}{k!}\Delta^k y_{n-k} + R_k(x). \quad (3)$$

Если $y = f$ имеет непрерывную производную порядка $k+1$ на отрезке $[a, b]$, содержащем точки x, x_n, ..., x_{n-k}, то ее остаточный член представим в форме

$$R_k(x) = h^{k+1}\frac{t(t+1)\ldots(t+k)}{(k+1)!}y^{(k+1)}(\xi), \quad (4)$$

где ξ есть точка отрезка $[a, b]$.

2. Интерполирование внутри таблицы. Предположим, что точка x лежит вблизи внутреннего узла x_n таблицы с любой стороны от него. Тогда табличные узлы целесообразно привлекать для интерполирования в порядке удаленности от x_n, т. е. взять сначала узел x_n и присоединять к нему пары узлов $(x_n + h, x_n - h)$, $(x_n + 2h, x_n - 2h)$, ..., $(x_n + kh, x_n - kh)$. При таком порядке узлов интерполирования формула Ньютона будет иметь вид

$$f(x) = y(x) = f(x_n) + (x - x_n)f(x_n, x_n + h) +$$
$$+ (x - x_n)(x - x_n - h)f(x_n, x_n + h, x_n - h) +$$
$$+ (x - x_n)(x - x_n - h)(x - x_n + h)\times$$
$$\times f(x_n, x_n + h, x_n - h, x_n + 2h) + \ldots$$
$$\ldots + (x - x_n)(x - x_n - h)\ldots(x - x_n - kh)\times$$
$$\times f(x_n, x_n + h, x_n - h, \ldots, x_n - kh) + R_{2k}(x),$$
$$R_{2k}(x) = \frac{(x - x_n)(x - x_n - h)\ldots(x - x_n + kh)}{(2k+1)!}f^{(2k+1)}(\xi),$$

где ξ есть точка отрезка, содержащего $x_n - kh$, $x_n + kh$ и x.

Как и в п. 1, заменим разностные отношения их выражениями через конечные разности

$$f(x_n) = y_n, \quad f(x_n, x_n + h) = \frac{\Delta y_n}{1!h},$$

$$f(x_n, x_n + h, x_n - h) = f(x_n - h, x_n, x_n + h) = \frac{\Delta^2 y_{n-1}}{2!h^2}, \ldots$$

и введем переменную $t = \frac{x - x_n}{h}$; тогда получим

$$y(x_n + th) = y_n + \frac{t}{1!}\Delta y_n + \frac{t(t-1)}{2!}\Delta^2 y_{n-1} +$$
$$+ \frac{t(t-1)(t+1)}{3!}\Delta^3 y_{n-1} + \ldots + \frac{1}{(2k-1)!}(t-k+1)\ldots$$
$$\ldots t(t+1)\ldots(t+k-1)\Delta^{2k-1}y_{n-k+1} +$$
$$+ \frac{1}{(2k)!}(t+k-1)\ldots t\ldots(t-k)\Delta^{2k}y_{n-k+1} + R_{2k}(x).$$

Для придания правой части симметричного вида, перепишем равенство в форме

$$y(x_n + th) = y_n + t\left[\Delta y_n - \frac{1}{2}\Delta^2 y_{n-1}\right] + \frac{t^2}{2!}\Delta^2 y_{n-1} +$$
$$+ \frac{t(t^2-1^2)}{3!}\left[\Delta^3 y_{n-1} - \frac{1}{2}\Delta^4 y_{n-2}\right] + \ldots$$
$$\ldots + \frac{t(t^2-1^2)\ldots(t^2-(k-1)^2)}{(2k-1)!}\left[\Delta^{2k-1}y_{n-k+1} - \frac{1}{2}\Delta^{2k}y_{n-k}\right] +$$
$$+ \frac{t^2(t^2-1^2)\ldots(t^2-(k-1)^2)}{(2k)!}\Delta^{2k}y_{n-k} + R_{2k}(x).$$

Если из прямоугольных скобок исключить конечные разности четного порядка, пользуясь равенствами

$$\Delta^2 y_{n-1} = \Delta y_n - \Delta y_{n-1}, \quad \Delta^4 y_{n-2} = \Delta^3 y_{n-1} - \Delta^3 y_{n-2}, \ldots,$$

получим *интерполяционную формулу Ньютона — Стирлинга*:

$$y(x_n + th) = y_n + \frac{t}{1!}\frac{\Delta y_{n-1} + \Delta y_n}{2} + \frac{t^2}{2!}\Delta^2 y_{n-1} +$$
$$+ \frac{t(t^2-1^2)}{3!}\frac{\Delta^3 y_{n-2} + \Delta^3 y_{n-1}}{2} + \frac{t^2(t^2-1^2)}{4!}\Delta^4 y_{n-2} + \ldots$$
$$\ldots + \frac{t(t^2-1^2)\ldots(t^2-(k-1)^2)}{(2k-1)!}\frac{\Delta^{2k-1}y_{n-k} + \Delta^{2k-1}y_{n-k+1}}{2} +$$
$$+ \frac{t^2(t^2-1^2)\ldots(t^2-(k-1)^2)}{(2k)!}\Delta^{2k}y_{n-k} + R_{2k}(x), \qquad (5)$$
$$R_{2k}(x) = h^{2k+1}\frac{t^2(t^2-1^2)\ldots(t^2-k^2)}{(2k+1)!}y^{(2k+1)}(\xi).$$

Приведем еще *формулу Ньютона — Бесселя*. Она предназначена для интерполирования, когда x лежит вблизи середины между двумя соседними табличными узлами, которые мы обозначим x_n и x_{n+1}.

При построении вычислительной формулы табличные узлы целесообразно брать попарно в следующем порядке: (x_n, x_n+h), (x_n-h, x_n+2h), ..., $(x_n-kh+h,\ x_n+kh)$. *Формула Ньютона* здесь будет иметь вид

$$\begin{aligned} f(x) = {} & f(x_n) + (x-x_n) f(x_n, x_n+h) + \\ & + (x-x_n)(x-x_n-h) f(x_n, x_n+h, x_n-h) + \ldots \\ & \ldots + (x-x_n+kh-h) \ldots (x-x_n-kh+h) \times \\ & \times f(x_n, x_n+h, \ldots, x_n+kh-h, x_n-kh+h) + \\ & + (x-x_n+kh-h) \ldots (x-x_n-kh+h) \times \\ & \times f(x_n, x_n+h, \ldots, x_n-kh+h, x_n+kh) + R_{2k-1}(x). \end{aligned}$$

После использования равенств

$$f(x_n) = y_n, \quad f(x_n, x_n+h) = \frac{\Delta y_n}{1!h},$$

$$f(x_n, x_n+h, x_n-h) = f(x_n-h, x_n, x_n+h) = \frac{\Delta^2 y_{n-1}}{2!h^2}, \ldots$$

и замены переменной $x = x_n + th$ формула приводится к виду

$$\begin{aligned} y(x_n+th) = {} & y_n + \frac{t}{1!}\Delta y_n + \frac{t(t-1)}{2!}\Delta^2 y_{n-1} + \\ & + \frac{(t+1)t(t-1)}{3!}\Delta^3 y_{n-1} + \ldots \\ & \ldots + \frac{(t+k-2)\ldots(t-k+1)}{(2k-2)!}\Delta^{2k-2} y_{n-k+1} + \\ & + \frac{(t+k-1)\ldots(t-k+1)}{(2k-2)!}\Delta^{2k-1} y_{n-k+1} + R_{2k-1}(x). \end{aligned}$$

Чтобы привести правую часть равенства к симметричной относительно точки $x_n + \frac{1}{2}h$ форме, отделим от разностей четного порядка половины их значений: $\frac{1}{2}y_n$,

$\frac{1}{2}\Delta^2 y_{n-1}$, $\frac{1}{2}\Delta^4 y_{n-2}$, ..., и заменим эти величины следующими выражениями:

$$\frac{1}{2}y_n = \frac{1}{2}(y_{n+1} - \Delta y_n), \quad \frac{1}{2}\Delta^2 y_{n-1} = \frac{1}{2}(\Delta^2 y_n - \Delta^3 y_{n-1}),$$

$$\frac{1}{2}\Delta^4 y_{n-2} = \frac{1}{2}[\Delta^4 y_{n-1} - \Delta^5 y_{n-2}], \ldots$$

После объединения членов с одинаковыми разностями получится *интерполяционная формула Ньютона — Бесселя*

$$y(x_n + th) = \frac{y_n + y_{n+1}}{2} + \frac{t - \frac{1}{2}}{1!}\Delta y_n +$$

$$+ \frac{t(t-1)}{2!}\,\frac{\Delta^2 y_{n-1} + \Delta^2 y_n}{2} + \frac{\left(t - \frac{1}{2}\right)t(t-1)}{3!}\Delta^3 y_{n-1} +$$

$$+ \frac{(t+1)t(t-1)(t-2)}{4!}\,\frac{\Delta^4 y_{n-2} + \Delta^4 y_{n-1}}{2} + \ldots$$

$$\ldots + \frac{(t+k-2)\ldots(t-k+1)}{(2k-2)!}\,\frac{\Delta^{2k-2} y_{n-k+1} + \Delta^{2k-2} y_{n-k+2}}{2} +$$

$$+ R_{2k-1}(x), \qquad (6)$$

$$R_{2k-1}(x) = h^{2k}\frac{(t+k-1)\ldots(t-k)}{(2k)!}\,y^{(2k)}(\xi).$$

Здесь ξ есть точка отрезка, содержащего $x_n - kh + h$, $x_n + kh$, x.

§ 5. Интерполирование с кратными узлами

До сих пор рассматривалась задача интерполирования $f(x)$ по нескольким значениям только самой функции f. Кратко остановимся на немного более сложной задаче об интерполировании $f(x)$ по значениям f и производных от нее.

1. Содержание задачи; интерполирующий многочлен и погрешность. Пусть на отрезке $[a, b]$ даны m различных узлов x_i $(i = 1, \ldots, m)$. Предположим, что в точке x_1 известны значения $f(x_1)$, $f'(x_1)$, ..., $f^{(\alpha_1 - 1)}(x_1)$, в точке x_2 известны значения $f(x_2)$, $f'(x_2)$, ..., $f^{(\alpha_2 - 1)}(x_2)$ и т. д. Числа $\alpha_1, \alpha_2, \ldots, \alpha_m$ называются *кратностями*

узлов $x_1, x_2, \ldots, x_m$. Общее число известных данных о функции f обозначим $n+1$:

$$\alpha_1 + \alpha_2 + \ldots + \alpha_m = n+1.$$

Возьмем многочлен степени n с произвольными коэффициентами

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n \qquad (1)$$

и выберем его коэффициенты a_j так, чтобы выполнялись условия

$$P_n^{(i)}(x_k) = f^{(i)}(x_k) \quad (k = 1, \ldots, m;\ i = 0, 1, \ldots, \alpha_k - 1), \qquad (2)$$

которые дают для a_j $(j = 0, \ldots, n)$ систему $n+1$ линейных уравнений. Чтобы доказать существование и единственность решения, достаточно показать, что однородная система

$$P_n^{(i)}(x_k) = 0 \quad (k = 1, \ldots, m;\ i = 0, 1, \ldots, \alpha_k - 1) \qquad (3)$$

имеет только нулевое решение. Система говорит о том, что для многочлена $P_n(x)$ каждый узел x_k является корнем кратности не меньше α_k. Значит, для многочлена $P_n(x)$, степень которого равна n, сумма кратностей его корней не меньше, чем $\alpha_1 + \ldots + \alpha_m = n+1$. Такой многочлен тождественно равен нулю, и поэтому все его коэффициенты также равны нулю. Однородная система (2) имеет, следовательно, только нулевое решение, и существует лишь один многочлен (1), удовлетворяющий условиям (2). Явное выражение его будет дано в следующем пункте.

Рассмотрим погрешность интерполирования $R_n(x) = f(x) - P_n(x)$, и докажем сейчас одну из теорем о ее представлении, рассчитанную на функции f достаточно высокого порядка гладкости.

Теорема 1. *Пусть узлы* x_k $(k = 1, \ldots, m)$ *и точка* x *принадлежат отрезку* $[a, b]$ *и функция* f *имеет на* $[a, b]$ *непрерывную производную порядка* $n+1$. *Тогда на* $[a, b]$ *существует такая точка* ξ, *что для погрешности* $R_n(x)$ *интерполирования верно равенство*

$$R_n(x) = \frac{A_n(x)}{(n+1)!} f^{(n+1)}(\xi), \qquad (4)$$

$$A_n(x) = (x - x_1)^{\alpha_1} \ldots (x - x_m)^{\alpha_m}.$$

Доказательство. Будем считать x отличным от узлов x_k и рассмотрим вспомогательную функцию аргумента z:

$$F(z) = f(z) - P_n(z) - \frac{A_n(z)}{A_n(x)}[f(x) - P_n(x)].$$

Она имеет на $[a, b]$ непрерывную производную порядка $n+1$. Точки $x_1, \ldots, x_m$ и x для нее будут нулями, кратности которых не ниже соответственно $\alpha_1, \ldots, \alpha_m$ и 1. Сумма кратностей всех нулей F не меньше $\alpha_1 + \ldots + \alpha_m + 1 = n + 2$. Производная от нее $F'(z)$ будет иметь по теореме Ролля внутри каждого отрезка между смежными точками $x_1, \ldots, x_m, x$ не менее чем один нуль. Таких нулей будет m. Кроме того, точки $x_1, \ldots, x_m$ будут нулями кратностей не меньше $\alpha_1 - 1, \ldots, \alpha_m - 1$. Поэтому $F'(z)$ имеет на $[a, b]$ не меньше чем $(\alpha_1 - 1) + \ldots + (\alpha_m - 1) + m = n + 1$ нулей. Проведя сходные рассуждения для второй и следующих производных, придем к заключению, что производная порядка $n+1$ от F имеет на $[a, b]$ не меньше чем один нуль. Поэтому на $[a, b]$ существует такая точка ξ, что выполняется равенство

$$F^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{(n+1)!}{A_n(x)}[f(x) - P_n(x)] =$$

$$= f^{(n+1)}(\xi) - \frac{(n+1)!}{A_n(x)} R_n(x) = 0.$$

Отсюда следует (4).

2. Представление погрешности интерполирования в случае аналитической функции. Предположим, что отрезок $[a, b]$, на котором располагаются узлы x_k и точка x, является конечным, и функция f — аналитическая в замкнутой конечной области D, ограниченной контуром l и содержащей $[a, b]$ внутри себя.

Из системы (2), в частности, следует, что коэффициенты a_j будут линейными функциями свободных членов системы $f^{(i)}(x_k)$. Поэтому интерполирующий многочлен (1) также будет линейно выражаться через $f^{(i)}(x_k)$, и может быть записан в форме

$$P_n(f, x) = \sum_{k=1}^{m} \sum_{i=1}^{\alpha_k - 1} L_{k,i}(x) f^{(i)}(x_k).$$

Формула Коши $f(x) = \frac{1}{2\pi i}\int_l \frac{f(x)}{z-x}\,dz$ позволяет привести разыскание интерполирующего многочлена $P_n(f, x)$ к задаче построения многочлена $P_n(1/(z-x), x)$ для очень простой дробно-линейной функции $1/(z-x)$, где x является независимой переменной и z считается параметром, по которому затем выполняется интегрирование.

Рассмотрим погрешность интерполирования

$$R_n\left(\frac{1}{z-x}, x\right) = \frac{1}{z-x} - P_n\left(\frac{1}{z-x}, x\right) =$$

$$= \frac{1}{z-x} - \sum_{k=1}^{m}\sum_{i=1}^{\alpha_k - 1} L_{k,i}(x)\frac{i!}{(z-x_k)^{i+1}},$$

и сосредоточим сейчас внимание на зависимости ее от z. Это есть рациональная функция от x, и приведенное выше равенство есть разложение ее на простые дроби. Заметим, что точка $z = x$ для погрешности есть полюс первого порядка с вычетом, равным единице. Общий знаменатель всех членов в выражении для погрешности равен $(z-x)A_n(z)$ и погрешность представима, очевидно, в форме

$$R_n\left(\frac{1}{z-x}, x\right) = \frac{B(z, x)}{(z-x)A_n(z)}, \tag{5}$$

где $B(z, x)$ есть многочлен от z степени не выше $n-1$.

Убедимся сейчас, что $B(z, x)$ не зависит от z и равняется $A_n(x)$. Если $|z|$ имеет большое значение, то верно равенство

$$\frac{1}{z-x} = \sum_{\nu=1}^{\infty}\frac{x^\nu}{z^{\nu+1}}$$

и, так как погрешность линейно зависит от интерполируемой функции, то

$$R_n\left(\frac{1}{z-x}, x\right) = \sum_{\nu=0}^{\infty} z^{-\nu-1}R_n(x^\nu, x). \tag{6}$$

Степени x от нулевой до n интерполируются точно, и $R_n(x^\nu, x) = 0$ $(\nu = 0, 1, \ldots, n)$. Первые $n+1$ членов

в сумме справа исчезают, и разложение (6) должно начинаться с члена $z^{-n-2}R_n(x^{\nu+1}, x)$. Поэтому в (5) степень числителя $B(z, x)$ относительно z должна быть на $n+2$ единицы ниже степени знаменателя. Но знаменатель имеет степень $n+2$, значит, числитель должен иметь нулевую степень и не зависеть от z: $B(z, x) = B(x)$. Наконец, так как значение $z = x$ должно быть для погрешности простым полюсом с вычетом 1, то $B(x) = A_n(x)$ и

$$R_n\left(\frac{1}{z-x}, x\right) = \frac{A_n(x)}{(z-x)A_n(z)}.$$

Отсюда и из интеграла Коши для функции $f(x)$ получается представление погрешности R_n в форме контурного интеграла:

$$R_n(f, x) = \frac{1}{2\pi i}\int_l f(z) R_n\left(\frac{1}{z-x}, x\right) dz = \\ = \frac{A_n(x)}{2\pi i}\int_l \frac{f(z)}{(z-x)A_n(z)} dz. \quad (7)$$

Если вычислить полученный интеграл с помощью вычетов, мы найдем нужное представление интерполирующего многочлена $P_n(f, x) = f(x) - R_n(f, x)$.

Вычет функции $\frac{A_n(x)(z)}{(z-x)A_n(z)}$ в точке $z = x$ равен, очевидно, $f(x)$. Вычислим вычет в полюсе $z = x_k$. При z, близких к x_k, верны следующие разложения в степенные ряды:

$$f(z) = \sum_{s=0}^{\infty} \frac{1}{s!} f^{(s)}(x_k)(z-x_k)^s,$$

$$\frac{1}{z-x} = -\frac{1}{(x-x_k)-(z-x_k)} = -\sum_{s=0}^{\infty} \frac{(z-x_k)^s}{(x-x_k)^{s+1}},$$

$$\frac{(z-x_k)^{\alpha_k}}{A_n(z)} = \sum_{s=0}^{\infty} C_s^{(k)}(z-x_k)^s.$$

Вычет же функции

$$\frac{f(z)}{(z-x)A_n(z)} = \frac{1}{(z-x_k)^{\alpha_k}} \frac{(z-x_k)^{\alpha_k}}{A_n(z)} \frac{f(z)}{z-x}$$

может быть найден при помощи умножения приведенных трех степенных рядов и подсчета коэффициента при $(z-x_k)^{\alpha_k-1}$. Последний равен

$$-\sum_{i=0}^{\alpha_k-1} f^{(i)}(x_k)\frac{1}{i!}\sum_{s=0}^{\alpha_k-1-i} C_s^{(k)}(x-x_k)^{-\alpha_k-s+i}.$$

Найденные вычеты приведут к следующему *эрмитову представлению* многочлена P_n:

$$P_n(f, x) = \sum_{k=1}^{m}\sum_{i=0}^{\alpha_k-1} f^{(i)}(x_k)\frac{1}{i!}\frac{A_n(x)}{(x-x_k)^{\alpha_k}}\sum_{s=0}^{\alpha_k-1-i} C_s^{(k)}(x-x_k)^{i+s}. \qquad (8)$$

Например, при интерполировании с двукратными узлами условия, определяющие многочлен $P_{2m-1}(f, x)$, будут следующими:

$$P_{2m-1}(f, x_k) = f(x_k),\ P'_{2m-1}(f, x_k) = f'(x_k) \quad (k = 1, \ldots, m),$$

и формула (8) принимает вид

$$P_{2m-1}(f, x) = \sum_{k=1}^{m}\frac{\omega^2(x)}{(x-x_k)^2[\omega'(x_k)]^2}\times$$

$$\times\left\{\left[1-\frac{\omega''(x_k)}{\omega'(x_k)}(x-x_k)\right]f(x_k)+(x-x_k)f'(x_k)\right\}, \qquad (9)$$

$$\omega(x) = (x-x_1)\ldots(x-x_m).$$

Отметим, что представление (8) для $P_n(f, x)$ было получено в предположении аналитичности функции f, но в окончательный результат входят только значения $f(x_k)$ и $f'(x_k)$, и формула (8) остается верной для любой функции f с конечными значениями $f(x_k)$ и $f'(x_k)$.

§ 6. О вычислении значений производных с помощью интерполирования функций

1. Формула вычислений; представление погрешности формулы. Пусть на отрезке $[a, b]$ рассматривается функция f, имеющая непрерывную производную порядка $n+1$. Возьмем на $[a, b]$ $n+1$ различных узлов $x_0, x_1, \ldots$

..., x_n. Для упрощения записи предположим, что они перенумерованы слева направо так, что $x_0 < x_1 < \ldots$ $\ldots < x_n$. Интерполируем f по ее значениям в узлах x_k посредством многочлена $P_n(x)$ степени n и обозначим $R_n(x)$ погрешность интерполирования:

$$f(x) = P_n(x) + R_n(x).$$

Вычислим производную от f порядка m:

$$f^{(m)}(x) = P_n^{(m)}(x) + R_n^{(m)}(x).$$

Пренебрегая величиной $R_n^{(m)}$, получим формулу для приближенного вычисления производной:

$$f^{(m)}(x) \approx P_n^{(m)}(x). \tag{1}$$

Ее погрешность равна $R_n^{(m)}(x)$. Пользоваться ею целесообразно при небольших порядках m производной, во всяком случае, когда $m \leqslant n$, так как все производные от P_n порядка выше n тождественно равны нулю.

Вопрос о вычислении $P_n^{(m)}$ является простым, и мы отложим его до следующего пункта. Сейчас же остановимся на погрешности $R_n^{(m)}$. Необходимо найти ее представления, удобные для того, чтобы составить мнение об ее свойствах и, если можно, оценить ее численно. Построим представление $R_n^{(m)}$ лагранжева типа, являющееся простейшим. Как увидим ниже, оно будет верным не при всяком положении точки интерполирования x на $[a, b]$, и последняя должна подчиняться двум ограничениям, указываемым ниже.

Возьмем вспомогательную функцию аргумента t:

$$\begin{aligned} \varphi(t) &= R_n(t) - \frac{K}{(n+1)!}\omega(t), \\ \omega(t) &= (t - x_0) \ldots (t - x_n). \end{aligned} \tag{2}$$

Здесь K есть произвольная постоянная величина, значение которой будет выбрано ниже. При всяких значениях K функция φ обращается в нуль в точках x_i $(i=0, 1, \ldots, n)$. По теореме Ролля, φ' будет иметь по меньшей

мере один нуль между каждой парой соседних узлов (x_k, x_{k+1}). Таких нулей внутри отрезка $[x_0, x_n]$ будет не меньше n и т. д. Производная порядка m

$$\varphi^{(m)}(t) = R_n^{(m)}(t) - \frac{K}{(n+1)!}\omega^{(m)}(t)$$

будет иметь внутри $[x_0, x_n]$ не меньше чем $n+1-m$ нулей.

Выберем теперь величину K. Для этого нам потребуется наложить на положение точки x два ограничения.

А) Будем считать, что в точке x $\omega^{(m)}(x) \neq 0$. Попутно отметим, что если x не лежит внутри отрезка $[x_0, x_n]$, а находится вне его или на одном из его концов, то такое условие заведомо выполняется. Действительно, $\omega'(x)$ есть многочлен, имеющий по одному нулю внутри каждого из частичных отрезков $[x_i, x_{i+1}]$ $(i = 0, 1, \ldots, n-1)$. Таких нулей n штук, и никаких других нулей у $\omega'(x)$ нет. $\omega''(x)$ есть многочлен степени $n-1$, у которого есть по одному нулю между каждой парой соседних нулей ω'. Все эти нули ω'' лежат внутри $[x_0, x_n]$, и других нулей у ω'' нет, и т. д. Продолжая такие рассуждения, убедимся в том, что все нули $\omega^{(m)}$ лежат внутри $[x_0, x_n]$.

Выберем теперь K так, чтобы точка x была нулем $\varphi^{(m)}$, т. е. чтобы выполнялось равенство

$$\varphi^{(m)}(x) = R_n^{(m)}(x) - \frac{K}{(n+1)!}\omega^{(m)}(x) = 0. \qquad (3)$$

Рассмотрим наименьший отрезок, содержащий точки x_0, x_n, x, и обозначим его $[\alpha, \beta]$. Когда x лежит на $[x_0, x_n]$, то $[\alpha, \beta] = [x_0, x_n]$; когда же x лежит вне $[x_0, x_n]$, то $[\alpha, \beta]$ будет шире $[x_0, x_n]$.

Б) Будем считать, что на $[\alpha, \beta]$ производная $\varphi^{(m)}$ имеет не меньше чем $n+2-m$ нулей.

При выполнении этого требования все последующие рассуждения будут верными; если же оно нарушается, то рассуждения не могут быть проведены и формулируемая ниже теорема может оказаться неверной.

Отметим, что если x лежит вне $[x_0, x_n]$ или в точках x_0 или x_n, то $\varphi^{(m)}$ будет иметь не меньше $n+1-m$

нулей внутри и еще нуль $t = x$, отличный от них. При таком расположении x условие Б) заведомо выполнено. Если же x находится внутри $[x_0, x_n]$, то нуль $t = x$ может совпасть с одним из нулей функции $\varphi^{(m)}$, указанных выше при применении теоремы Ролля, и условие Б) может оказаться невыполненным.

Предположим, что условие Б) выполняется. Тогда можно утверждать, что $\varphi^{(m+1)}$ будет иметь внутри $[\alpha, \beta]$ не меньше $n+1-m$ различных нулей и т. д. и, наконец, производная порядка $n+1$ будет иметь не меньше одного нуля. На $[\alpha, \beta]$ существует, следовательно, такая точка ξ, для которой выполняется равенство

$$\varphi^{(n+1)}(\xi) = R_n^{(n+1)}(\xi) - K = f^{(n+1)}(\xi) - K = 0,$$

и, следовательно, $K = f^{(n+1)}(\xi)$. Из (3) тогда следует

$$R_n^{(m)}(x) = \frac{\omega^{(m)}(x)}{(n+1)!} f^{(n+1)}(\xi). \qquad (4)$$

Поэтому верна

Теорема 1. *Пусть на отрезке $[a, b]$, содержащем точки x_0, x_n, x, функция f имеет непрерывную производную порядка $n+1$, и для точки x выполняются условия* А) *и* Б), *указанные выше. Тогда на $[a, b]$ существует такая точка ξ, что для погрешности $R_n^{(m)}(x)$ вычислительной формулы* (1) *верно представление* (4).

2. Некоторые частные формулы вычисления производных. Каждая из формул для интерполяционного многочлена, указанных в предыдущих параграфах, может служить источником для получения формул вычисления производных. Таких формул можно получить большое число, но для выяснения идеи их построения достаточно ограничиться несколькими примерами.

Возьмем формулу Ньютона (3.6). Если ввести сокращенное обозначение $x - x_k = \alpha_k$, можно ее записать в виде

$$P_n(x) = f(x_0) + \alpha_0 f(x_0, x_1) + \alpha_0\alpha_1 f(x_0, x_1, x_2) + \ldots$$
$$\ldots + \alpha_0\alpha_1 \ldots \alpha_{n-1} f(x_0, x_1, \ldots, x_n).$$

Последовательное дифференцирование этого равенства дает следующие приближенные выражения производных *) от f:

$$f'(x) \approx P_n'(x) = f(x_0, x_1) + (\alpha_0 + \alpha_1) f(x_0, x_1, x_2) + \\ + (\alpha_0\alpha_1 + \alpha_0\alpha_2 + \alpha_1\alpha_2) f(x_0, x_1, x_2, x_3) + \ldots,$$

$$\frac{1}{2!} f''(x) \approx \frac{1}{2!} P_n''(x) = f(x_0, x_1, x_2) + \\ + (\alpha_0 + \alpha_1 + \alpha_2) f(x_0, x_1, x_2, x_3) + \ldots,$$

$$\frac{1}{3!} f'''(x) \approx \frac{1}{3!} P'''(x) = f(x_0, x_1, x_2, x_3) + \\ + (\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3) f(x_0, x_1, x_2, x_3, x_4) + \ldots,$$

$$\frac{1}{4!} f^{\mathrm{IV}}(x) \approx \frac{1}{4!} P_n^{\mathrm{IV}}(x) = f(x_0, x_1, x_2, x_3, x_4) + \\ + (\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) f(x_0, x_1, x_2, x_3, x_4, x_5) + \ldots$$

Сходным образом могут быть получены формулы для вычисления производных в случае равноотстоящих точек. Если взять, например, формулу Ньютона для интерполирования в начале таблицы (4.1) и вычислить производные по переменной t, получатся приводимые ниже выражения для производных **):

$$hy'(x_0 + th) \approx \Delta y_0 + \frac{2t-1}{2!} \Delta^2 y_0 + \frac{3t^2 - 6t + 2}{3!} \Delta^3 y_0 + \\ + \frac{4t^3 - 18t^2 + 22t - 6}{4!} \Delta^4 y_0 + \ldots,$$

$$h^2 y''(x_0 + th) \approx \Delta^2 y_0 + (t-1) \Delta^3 y_0 + \\ + \frac{6t^2 - 18t + 11}{12} \Delta^4 y_0 + \ldots,$$

$$h^3 y'''(x_0 + th) \approx \Delta^3 y_0 + \frac{2t-3}{2} \Delta^3 y_0 + \ldots$$

---

*) В правых частях приведенных равенств могут стоять суммы любого конечного числа слагаемых в зависимости от степени интерполирующего многочлена $P_n(x)$. Приведены же только первые слагаемые этих сумм, по которым можно судить о виде всех следующих членов правых частей.

**) Большое число формул для вычисления производных можно найти в книгах [4, 5, 6].

§ 7. О сходимости интерполяционных процессов

1. Введение. Пусть на отрезке $[a, b]$ рассматривается функция f с конечными значениями. Предположим, что задана бесконечная треугольная таблица узлов, определяющая интерполяционный процесс:

$$X = \begin{bmatrix} x_1^1 & & & \\ x_1^2, & x_2^2 & & \\ \cdots & & & \\ x_1^n, & x_2^n, & \ldots, & x_n^n \\ \cdots & & & \end{bmatrix}, \qquad x_i^n \in [a, b]. \tag{1}$$

На шаге номера n за узлы интерполирования принимаются элементы x_k^n $(k = 1, \ldots, n)$, стоящие в строке таблицы того же номера n. Интерполяционный многочлен имеет следующее выражение:

$$P_n(x) = \sum_{j=1}^{n} \frac{\omega_n(x)}{(x - x_j^n)\,\omega_n'(x_j^n)} f(x_j^n), \tag{2}$$
$$\omega_n(x) = (x - x_1^n) \ldots (x - x_n^n).$$

Ниже рассматривается либо поточечная, либо равномерная сходимость $P_n(x)$ к $f(x)$ и в соответствии с этим за меру близости $P_n(x)$ к $f(x)$ принимается либо $|f(x) - P_n(x)|$, $x \in [a, b]$, либо $\sup\limits_{[a, b]} |f(x) - P_n(x)|$.

Погрешность интерполирования

$$\varepsilon_n(x) = f(x) - P_n(x) \tag{3}$$

зависит, что является вполне очевидным, от следующих факторов: от свойств функции f, таблицы узлов X и, наконец, от числа n узлов. В проблеме сходимости основным является вопрос о том, как между собой должны быть связаны свойства f и таблица X, чтобы в принятой мере приближения имела место сходимость $P_n(x)$ к $f(x)$. Практическая полезность этой задачи является очевидной: в ней выясняются условия, при которых возможно сколь угодно точное вычисление $f(x)$, если число узлов взято достаточно большим. Отметим также, что здесь решается лишь принципиальный вопрос о возможности сколь угодно точного нахождения

f, но не дается пока никакого правила для нахождения числа узлов n, при котором погрешность становится меньше заданной границы. Чтобы ответить на последний вопрос о выборе n, потребовалось бы более глубокое численное изучение закона стремления к нулю меры погрешности $\varepsilon_n(x)$.

В практике вычислений, когда возникает потребность интерполирования, очень часто бывают заранее известны некоторые свойства функции f: является ли она аналитической на $[a, b]$ и какова область ее регулярности около отрезка $[a, b]$ (насколько широка эта область и как расположены на ее границе особые точки f); если f не является аналитической, то каким будет порядок ее непрерывной дифференцируемости и т. д. Кроме того, часто заранее бывает указана граница допустимой погрешности интерполирования.

По всем имеющимся заранее сведениям необходимо избрать способ интерполирования. Каждый такой способ определяется числом узлов и их расположением на $[a, b]$, и этими параметрами нужно распорядиться так, чтобы мера погрешности была не больше заданной величины.

При выборе способа интерполирования естественно взять интерполяционный процесс, сходящийся к f. Чтобы правильно сделать его выбор, полезно иметь представление о теории сходимости интерполирования, но изложение ее требует от читателя запаса знаний большего, чем тот, на который рассчитан наш учебник. Мы вынуждены ограничиться изложением лишь некоторых результатов о сходимости с небольшими их пояснениями.

2. Сходимость на множествах непрерывных и дифференцируемых функций. Остановимся на проблеме равномерной сходимости интерполирования. Рассмотрим множество непрерывных на отрезке $[a, b]$ функций f. Выберем какую-либо определенную функцию f. Если мы хотим ее интерполировать равномерно на $[a, b]$, то прежде всего следует выяснить, существует ли такая таблица узлов (1), которая обеспечивала бы равномерную сходимость интерполяционного процесса к f. Здесь следует дать утвердительный ответ, так как может быть доказана

Теорема 1. *Для каждой функции f, непрерывной на конечном отрезке $[a, b]$, существует такая таблица узлов* (1), *что соответствующий ей интерполяционный процесс равномерно на $[a, b]$ сходится к f.*

Обратим внимание на то, что эта таблица зависит в какой-то мере от функции f. Она может быть выбрана, например, следующим образом. Рассмотрим произвольный многочлен $P_n(x) = a_0x^n + a_1x^{n-1} + \ldots$ степени n. Его отклонение от функции f на $[a, b]$ характеризуется величиной $\max\limits_{a \leqslant x \leqslant b} |f(x) - P_n(x)|$. Возьмем многочлен, для которого эта величина имеет наименьшее значение $\min\limits_{a_0 \ldots a_n} \max\limits_{a \leqslant x \leqslant b} |f(x) - P_n(x)| = m_n$. Такой многочлен называется *многочленом, наименее уклоняющимся от f* на $[a, b]$. Обозначим его $P_n^*(x)$.

Ясно, что число m_n убывает с увеличением n, и так как по теореме Вейерштрасса к f можно с помощью многочлена приблизиться сколь угодно точно и равномерно на $[a, b]$, то $m_n \to 0$ при $n \to \infty$, и последовательность многочленов $P_n^*(x)$ будет сходиться к f равномерно на $[a, b]$. В теории приближения доказывается, что $P_n^*(x)$ принимает одинаковые значения с $f(x)$ на $[a, b]$ не меньше чем в $n + 1$ точках. Пусть эти точки суть $x_1^{(n+1)}$, $x_2^{(n+1)}$, ..., $x_{n+1}^{(n+1)}$. Поместим их в строку номера $n + 1$ таблицы (1) и примем за узлы интерполирования. Интерполяционный многочлен, так построенный, совпадет с $P_n^*(x)$. Если это проделать для $n = 1, 2, \ldots$, а $P_0(x)$ взять совпадающим с f в любой точке, то получится интерполяционный процесс, сходящийся к f равномерно.

В связи с теоремой 1 возникает вопрос: существует ли таблица узлов (1) такая, чтобы соответствующий ей интерполяционный процесс сходился равномерно на $[a, b]$ для всякой непрерывной там функции f. Приводимая ниже теорема дает отрицательный ответ.

Теорема 2. *Не существует таблицы узлов* (1), *для которой интерполяционный процесс был бы равномерно сходящимся на $[a, b]$ для всякой непрерывной там функции f.*

Множество непрерывных функций является слишком широким, чтобы для него существовала единственная таблица узлов, обеспечивающая равномерную сходимость

интерполяционного процесса для всех функций множества. Существование такой таблицы возможно только для более узких множеств функций.

Приведем две теоремы такого типа; для формулировки их необходимо будет ввести некоторые понятия. Функция f называется абсолютно непрерывной на $[a, b]$, если она представима в виде неопределенного интеграла

$$f(x) = C + \int_a^x F(t)\,dt, \qquad x \in [a, b]. \tag{4}$$

Интеграл здесь понимается в смысле Лебега, и $F(x)$ есть любая функция, абсолютно интегрируемая на $[a, b]$. Если читатель не знаком с этими понятиями, то можно понимать интеграл в смысле Римана и считать F абсолютно интегрируемой по Риману. Но тогда в последующем изложении необходимо заменить слова «абсолютно непрерывная функция» словами «функция, представимая в виде (4)».

Теорема 3. *Существуют таблицы узлов* (1) *такие, что соответствующие им интерполяционные процессы сходятся равномерно на* $[a, b]$ *к* f *для всякой функции* f, *абсолютно непрерывной там.*

Таким свойством обладает, например, таблица, у которой в строке номера n стоят корни многочлена Чебышева первого рода степени n. Для отрезка $[-1, 1]$ многочлен Чебышева есть $T_n(x) = \cos(n \arccos x)$ и корни его имеют значения $x_k^n = \cos\dfrac{2(n-k)+1}{2n}\pi$, $k = 1, 2, \ldots, n$.

Многочлены Чебышева для произвольного отрезка $[a, b]$ получаются из $T_n(x)$ при помощи линейного преобразования $x' = \dfrac{1}{2}(b+a) + \dfrac{1}{2}(b-a)x$, переводящего $[-1, 1]$ в $[a, b]$. Таблица корней многочленов $T_n(x)$ ($n = 1, 2, \ldots$) есть только пример, и несомненно, что всякая другая таблица, «достаточно близкая» к ней, будет приводить к равномерно сходящемуся интерполяционному процессу для множества абсолютно непрерывных функций.

Говорят, что функция f удовлетворяет на $[a, b]$ *условию Липшица с показателем* α, если при некотором

$M > 0$ для всяких x и y из $[a, b]$ выполняется неравенство

$$|f(x) - f(y)| \leqslant M|x - y|^{\alpha}, \quad 0 < \alpha \leqslant 1. \tag{5}$$

Множество функций f, для которых выполняется такое условие, обозначают обычно $\mathrm{Lip}_M \alpha$.

Теорема 4. *Если за узлы интерполирования принимаются корни многочлена Чебышева первого рода для отрезка* $[a, b]$, *то интерполяционный процесс сходится равномерно на* $[a, b]$ *для всякой функции* f, *удовлетворяющей условию Липшица* (5), *каким бы ни было* $\alpha > 0$.

Как простое следствие из теорем 4 или 3 может быть получена теорема о сходимости интерполирования для дифференцируемых функций, достаточная для некоторых приложений. Пусть функция f имеет всюду на $[a, b]$ первую производную, и эта производная ограничена числом M. Теорема Лагранжа о конечном приращении позволяет сказать, что из сделанного предположения вытекает принадлежность функции f классу Липшица при $\alpha = 1$:

$$|f(x) - f(y)| = |(x - y)f'(\xi)| \leqslant M|x - y|.$$

Это дает возможность высказать следующую теорему.

Теорема 5. *Если функция* f *имеет ограниченную производную на* $[a, b]$, *то интерполяционный процесс, в котором за узлы принимаются корни многочленов Чебышева первого рода, сходится равномерно к* f *на* $[a, b]$.

Вопрос о сходимости интерполяционных процессов, ввиду его большой принципиальной и прикладной важности, привлекал к себе в текущем столетии внимание многих ученых, и было опубликовано большое количество результатов для различных возникающих здесь задач; в частности, были найдены как необходимые, так и достаточные условия, которым должна удовлетворять таблица узлов (1) для сходимости процесса на множестве функций любого фиксированного порядка гладкости. Равным образом для некоторых типов таблиц (1) были указаны классы функций, в которых такие таблицы приводят к сходящимся интерполяционным процессам. Здесь невозможно дать систематический обзор даже основных результатов ввиду их многочисленности. Мы

ограничимся только небольшим количеством приведенных выше теорем, но дополним их пояснениями.

Прежде всего отметим, что в теоремах 3—5 говорится о сходимости интерполирования в очень широких классах функций и, так как такие теоремы должны предусматривать наличие функций с «плохими свойствами», то они обязаны налагать на таблицы узлов (1) весьма ограничительные условия. В теоремах в качестве (1) указывались таблицы корней многочленов Чебышева для отрезка $[a, b]$, но, как отмечалось выше, ясно, что утверждение теорем должно быть верным для других таблиц (1), «достаточно близких» к ним. Но остается неясным, насколько обязательной является такая близость: быть может существуют другие таблицы, близость к которым также достаточна для сходимости?

Заходя немного вперед, укажем, что это может случиться только для много более узких классов функций, чем в теоремах 3—5: такие функции должны быть не только аналитическими на $[a, b]$, но и регулярными в достаточно широкой области около $[a, b]$. В следующем пункте будет приведена теорема, из которой следует, что если требовать сходимости интерполирования на $[a, b]$ для всякой функции, аналитической на $[a, b]$, то таблица узлов должна быть такой, чтобы при больших n узлы $x_1^n, x_2^n, \ldots, x_n^n$, находящиеся в строке номера n, были распределены на $[a, b]$ почти так же, как и корни многочлена Чебышева *).

Отступать от такой близости в отношении распределения узлов допустимо лишь в том случае, когда мы заинтересованы в сходимости интерполирования не для всех аналитических на $[a, b]$ функций, а рассматриваем классы функций, регулярных в некоторой фиксированной области, содержащей $[a, b]$ внутри себя **).

Узлы, близкие к корням многочленов Чебышева, не всегда удобно, а иногда и невозможно применять в практике интерполирования. Для получения сходящегося интерполяционного процесса тогда часто отрезок $[a, b]$ разделяют на части точками $a_0 = a < a_1 < a_2 < \ldots < a_p = b$ и интерполируют f не при помощи одного мно-

*) Точный смысл этой близости указан в теореме 7 п. 3.
**) См. теоремы 8 и 9 п. 3.

гочлена на всем отрезке $[a, b]$, а для каждого частичного отрезка $[a_i, a_{i+1}]$ избирают свои узлы и строят свой интерполирующий многочлен $P^i_{k_i}(x)$ некоторой степени k_i.

Построенные так многочлены $P^i_{k_i}$ дают интерполирование f на всем отрезке; но необходимо отметить, что в точках a_i, где стыкуются соседние отрезки, интерполирование будет, вообще говоря, негладким и может быть даже разрывным, если точка a_i не принята за узел интерполирования при составлении многочленов $P^{i-1}_{k_{i-1}}$ и $P^i_{k_i}$.

Чтобы устранить этот недостаток и построить составную функцию, гладко интерполирующую f, обычно увеличивают степени мночленов $P^i_{k_i}$ и берут их больше, чем нужно для интерполирования по взятым на $[a_i, a_{i+1}]$ узлам. При этом некоторые коэффициенты или другие параметры остаются произвольными. Их выбирают так, чтобы следующий многочлен $P^i_{k_i}$ был достаточно гладким продолжением с $[a_{i-1}, a_i]$ на $[a_i, a_{i+1}]$ предыдущего многочлена $P^{i-1}_{k_{i-1}}$.

Для пояснения рассмотрим частный пример. Пусть на отрезке [0, 1] функция f задана своими значениями $f_k = f(k/n)$ в равноотстоящих точках $x_k = k/n$ $(k = 0, 1, \ldots, n)$. Рассмотрим многочлен второй степени на отрезке $[i/n, (i+1)/n]$ $P_i(x) = ax^2 + bx + c$ и выберем его так, чтобы он принимал в точках $x = i/n$ и $x = (i+1)/n$ такие же значения, как и f. Многочлен P_i можно, очевидно, представить в форме суммы линейного многочлена, принимающего на концах отрезка такие же значения, как и f, и многочлена второй степени, обращающегося в нуль на концах отрезка. Легко проверить, что такое представление имеет вид

$$P_i(x) = (nx - i) f_{i+1} - (nx - i - 1) f_i + \\ + A_i (nx - i)(nx - i - 1). \quad (6)$$

В каждом многочлене P_i остается произвольным параметр A_i. Выберем эти параметры так, чтобы производные от многочленов P_{i-1} и P_i в точке $x = i/n$, где соединяются отрезки $[(i-1)/n, i/n]$ и $[i/n, (i+1)/n]$, были равными. Это даст уравнение

$$P'_i(i/n) = nf_{i+1} - nf_i - nA_i = F'_{i-1}(i/n) = nf_i - nf_{i-1} + nA_{i-1},$$

или

$$A_{i-1} + A_i = f_{i+1} - 2f_i + f_{i-1} = \Delta^2 f_{i-1}, \quad i = 1, 2, \ldots, n-1. \tag{7}$$

Число параметров A_i равно n, тогда как число уравнений на единицу меньше. Один из параметров A_i остается произвольным и может быть выбран из дополнительного условия или задан произвольно. Например, если известно значение производной от f в точке $x = 0$, то параметр A_0 может быть найден из условия

$$P_0'(0) = nf_1 - nf_0 - nA_0 = f'(0), \quad -A_0 = \frac{1}{n} f_0' + f_0 - f_1.$$

Интерполирование, которое было описано сейчас, может быть названо *сглаженным кусочным интерполированием*, но его часто называют *сплайн-интерполированием* *), используя английский термин.

3. Сходимость интерполирования для аналитических функций. Предположим, что $f(z)$ есть аналитическая функция комплексной переменной $z = x + iy$, регулярная в замкнутой области D, содержащей отрезок $[a, b]$ действительной оси x внутри себя. Как и выше, будем считать, что узлы x_i^n таблицы (1) принадлежат промежутку $[a, b]$.

Рассмотрим следующий вопрос: как между собой должны быть связаны область D и таблица (1), чтобы интерполяционный процесс сходился на отрезке $[a, b]$ равномерно относительно x для всякой функции f, регулярной в D?

Сначала укажем область D, регулярность в которой гарантирует сходимость интерполирования при любой матрице (1) узлов интерполирования.

Построим два круга радиуса $|b-a|$ с центрами в точках a и b и назовем ω замкнутую область, являющуюся суммой этих кругов (рис. 2).

Теорема 6. *Для всякой функции f, регулярной в замкнутой области ω, интерполяционный процесс будет сходиться равномерно на отрезке $[a, b]$ при любой таблице узлов* (1).

*) *Сплайн* есть приспособление, позволяющее плавно соединять дуги разных кривых и аналогичное по роли лекалу.

Область ω является наименьшей замкнутой областью, регулярность в которой функции f гарантирует сходимость интерполирования (1).

Чтобы сформулировать дальнейшие теоремы, нам необходимо ввести понятие о предельной функции распределения. Пусть единичная масса любым образом распределена на отрезке $[a, b]$. Возьмем на $[a, b]$ произвольную точку x, отличную от b, и обозначим $\mu(x)$ сумму масс, лежащих на отрезке строго левее точки x. При $x = b$ положим $\mu(b) = 1$. Так определенная функция $\mu(x)$, очевидно, обладает следующими свойствами:

1) $\mu(a) = 0$;

2) $\mu(x)$ есть монотонная неубывающая функция от x на $[a, b]$; при этом она непрерывна слева в любой точке, лежащей внутри $[a, b]$;

3) $\mu(b) = 1$.

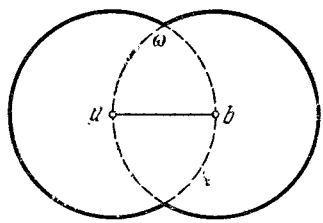

Рис. 2.

Всякую функцию $\mu(x)$, обладающую этими тремя свойствами, будем называть *функцией распределения* на отрезке $[a, b]$. Она может быть разрывной и не везде иметь производную. В частных же случаях $\mu(x)$ может быть дифференцируемой:

$$\frac{d\mu(x)}{dx} = \rho(x). \tag{8}$$

Функция $\rho(x)$ есть *плотность распределения масс*, и через нее $\mu(x)$ имеет следующее выражение:

$$\mu(x) = \int_a^x \rho(t)\, dt. \tag{9}$$

Пусть дана последовательность функций распределения $\mu_n(x)$ $(n = 1, 2, \ldots)$. Принято называть последовательность $\mu_n(x)$ *сходящейся в основном* к функции распределения $\mu(x)$, если во всякой точке непрерывности $\mu(x)$, лежащей внутри $[a, b]$, будет $\mu_n(x) \to \mu(x)$; функцию $\mu(x)$ называют *предельной функцией распределения* для последовательности μ_n.

Возвратимся к матрице узлов интерполирования (1) и рассмотрим ее строку номера n. Каждому из узлов x_k^n припишем массу $1/n$ и обозначим через $\mu_n(x)$ соответствующую функцию распределения. Будем считать, что последовательность $\mu_n(x)$ имеет предельную функцию распределения $\mu(x)$. Ее называют предельной функцией распределения узлов таблицы (1). Предположим, что такая функция существует. Как будет следовать из приводимых ниже теорем, она определяет область регулярности f, достаточную для равномерной сходимости интерполяционного процесса с таблицей узлов (1).

Теорема 7. *Пусть узлы таблицы* (1) *принадлежат отрезку* $[-1, 1]$ *и таблица имеет следующую предельную функцию распределения узлов* *):

$$\mu(x) = \frac{1}{\pi} \int_{-1}^{x} \frac{dt}{\sqrt{1-x^2}}, \quad \rho(x) = \pi^{-1}(1-x^2)^{-\frac{1}{2}}. \tag{10}$$

Тогда интерполяционный процесс будет сходиться равномерно на $[-1, 1]$ *для всякой функции* f, *аналитической на замкнутом отрезке* $[-1, 1]$.

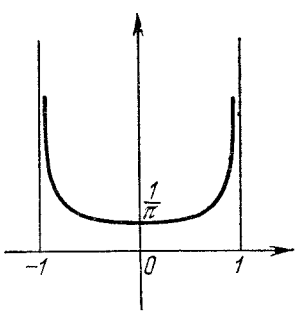

Рис. 3.

Верна также теорема, которую можно считать обратной для теоремы 7.

Теорема 8. *Пусть узлы таблицы* (1) *лежат на отрезке* $[-1, 1]$, *и пусть интерполяционный процесс по узлам этой таблицы* (1) *сходится во всех точках* $x \in [-1, 1]$ *для любой функции, аналитической при* $-1 \leqslant x \leqslant 1$. *Тогда таблица* (1) *имеет* (10) *своей предельной функцией распределения.*

Рассмотрим теперь важный для приложений случай равноотстоящих узлов. Пусть отрезок интерполирования есть $[0, 1]$, и за узлы интерполирования приняты точки $x_k = k/n$ $(k = 0, 1, \ldots, n)$. Предельная плотность распределения узлов при $n \to \infty$ здесь есть, очевидно, вели-

*) Функцию $\mu(x)$, указанную в (10), называют *функцией Чебышева*. График соответствующей плотности приведен на рис. 3.

чина постоянная, $\rho(x) = 1$, и функция распределения есть $\mu(x) = x$.

Представляет интерес выяснить, для каких функций можно быть уверенным в сходимости интерполирования при узлах $x_k = k/n$. Ответ дает

Теорема 9. *Если аналитическая функция $f(z)$ регулярна в замкнутой области D, содержащей отрезок $[0, 1]$ и ограниченной линией с уравнением*

$$x \ln \sqrt{x^2 + y^2} + (1 - x) \ln \sqrt{(1 - x)^2 + y^2} - \\ - y \operatorname{arctg} \frac{y}{x - x^2 - y^2} = 1, \quad (11)$$

то интерполяционный процесс с узлами $x_k = k/n$ $(k = 0, 1, \ldots, n)$ сходится к f равномерно на $[0, 1]$ при $n \to \infty$.

Область, ограниченная линией (11), изображена на рис. 4.

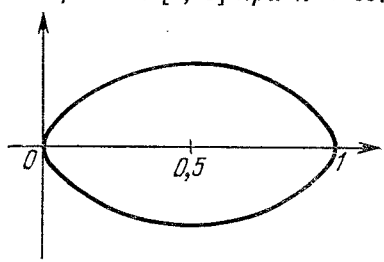

Рис. 4.

Наконец, укажем теорему, устанавливающую связь между областью D регулярности f и функцией $\mu(x)$ в более общем случае. Будем считать, что таблица (1) имеет предельную функцию распределения узлов $\mu(x)$, и построим при ее помощи логарифмический потенциал

$$U(z) = U(x, y) = \int_a^b \ln \frac{1}{|t - z|} d\mu(t), \quad z = x + iy. \quad (12)$$

Интеграл здесь понимается в смысле Стилтьеса. Читатель, не знакомый с такими интегралами, может понимать интеграл в обычном смысле Римана, что можно сделать в случае, когда $\mu(x)$ имеет непрерывную производную, и заменить $d\mu(x)$ на $\rho(x)dx$, где $\rho(x)$ есть плотность распределения узлов.

Комплексная переменная z считается лежащей вне $[a, b]$. Функция $U(z)$ — гармоническая вне отрезка $[a, b]$. Рассмотрим ее линии уровня l_c:

$$U(z) = U(x, y) = c. \quad (13)$$

Когда z удаляется на бесконечность, то $U(z)$ будет стремиться к $-\infty$. При отрицательных c, больших по абсолютному значению, l_c будет близка к окружности большого радиуса с центром в середине $[a, b]$. При возрастании c линия l_c будет сжиматься к отрезку $[a, b]$, и при некотором значении $c = \lambda$ она коснется впервые отрезка $[a, b]$. Возьмем линию l_λ. Ее уравнение есть

$$U(x, y) = \lambda. \tag{14}$$

Она содержит отрезок $[a, b]$ внутри себя, но в каких-то точках — одной или нескольких — касается его. Примем область, ограниченную линией уровня l_λ, за область D, присоединив к ней саму линию l_λ.

Теорема 10. *Если аналитическая функция f регулярна в замкнутой области D, то интерполяционный процесс, определяемый таблицей узлов* (1), *будет сходиться к f на отрезке $[a, b]$ равномерно относительно z* *).

ЛИТЕРАТУРА

1. Гончаров В. Л., Теория интерполирования и приближения функций, изд. 2, Физматгиз, М., 1954.
2. Марков А. А., Исчисление конечных разностей, Одесса, 1910.
3. Натансон И. П., Конструктивная теория функций, Гостехиздат, М. — Л., 1949.
4. Стеффенсен И. Ф., Теория интерполяций, ОНТИ, М. — Л., 1935.
5. Уиттекер Э., Робинсон Г., Математическая обработка результатов наблюдений, ОНТИ, М. — Л., 1935.
6. Березин И. С. и Жидков Н. П., Методы вычислений, т. I, изд. 3, «Наука», 1966.
7. Крылов В. И., Бобков В. В., Монастырный П. И., Вычислительные методы высшей математики, т. I, «Вышэйшая школа», Минск, 1972.
8. Мак-Кракен Д., Дорн У., Численные методы и программирование на Фортране, «Мир», М., 1969.

*) Можно показать, что он будет сходиться равномерно по z в замкнутой области D.

ГЛАВА 2

СИСТЕМЫ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

§ 1. Введение

Линейные системы имеют в вычислениях очень большое значение, так как к ним может быть приведено приближенное решение широкого круга задач. Теория этих систем сравнительно проста и доведена во многих частях до совершенства. Что же касается практики решения систем, то наши возможности еще сильно отстают от потребностей. Здесь многое зависит от порядка системы, т. е. от числа уравнений и неизвестных в ней. С увеличением порядка число операций, нужных для решения системы, быстро растет.

Число операций, требующихся для решения, зависит не только от порядка системы, но также от выбора метода вычислений. Поясним это примером. Предположим, что дана система n уравнений с n неизвестными и с определителем, отличным от нуля. По теореме Крамера система имеет единственное решение. В этой теореме указывается явное выражение для значений неизвестных в виде отношения двух определителей порядка n, при этом число различных определителей в отношениях равно $n+1$.

Пусть для нахождения решения мы хотим воспользоваться теоремой Крамера, при этом детерминанты будем вычислять по их обычному определению, как сумму со знаками $n!$ произведений элементов по одному из каждой строки и каждого столбца. Легко можно подсчитать, что для нахождения решения нужно будет

приблизительно $n^2n!$ умножений и делений *). Уже при $n = 20$ это число приблизительно равно 10^{21} и является настолько большим, что становится ясной невозможность решать указанным путем на современных машинах систему даже двадцати уравнений.

Чтобы было возможным решение систем большого числа уравнений, необходимо изменить метод вычислений и сделать его менее трудоемким. Такая задача привлекала внимание очень большого числа лиц и было указано много методов решения линейных систем, преследующих не только основную цель уменьшения числа операций, но и другие цели, о которых будет идти речь ниже. Эти методы строились как для систем общего вида с любыми коэффициентами, так и для систем специальных форм, например, получающихся при численном решении дифференциальных уравнений.

Какого сокращения вычислительной работы здесь можно достигнуть, мы поясним следующим примером: в методе исключения при применении схемы единственного деления для решения системы n уравнений требуется, как выяснится ниже, $\frac{n}{6}(2n^2 + 9n + 1)$ умножений и делений. При больших n эта величина во много раз меньше, чем $n^2n!$. Так, для $n = 20$ она имеет значение 3270, и поэтому система двадцати уравнений по этому методу может быть быстро решена.

Возможность уменьшить число арифметических и логических операций, необходимых для решения систем, имеет не только и даже не столько экономическое значение. Она позволяет увеличить порядок систем, поддающихся численному решению на ЭВМ, и расширяет круг прикладных задач, которые можно решать путем вычислений. Последнее же является весьма важным обстоятельством.

Методы решения систем уравнений обычно разделяют на две большие группы. К первой группе относят методы, которые принято называть *точными*: они позволяют для любых систем найти точные значения неизвестных

*) Мы подсчитываем эти операции, как более длительные по сравнению со сложением и вычитанием.

после конечного числа арифметических операций, каждая из которых выполняется точно.

Ко второй группе относят все методы, не являющиеся точными. Их называют обычно *итерационными*, и точные решения в них получают в результате бесконечного процесса приближений. Особое место среди них занимают *вероятностные методы*, в основу которых положены соображения, взятые из теории вероятностей. На таких методах мы не будем останавливаться и заметим лишь, что они могут быть особенно полезны в том случае, когда число неизвестных в системе и число уравнений будут весьма большими.

§ 2. Некоторые сведения о векторах и матрицах

Здесь будут рассмотрены дополнительные к обычным курсам алгебры сведения о нормах векторов и матриц и о предельном переходе для них, необходимые для понимания последующего изложения.

1. О сходимости последовательностей векторов и матриц. Пусть в n-мерном векторном пространстве дана последовательность векторов $x^k = (x_1^k, x_2^k, \ldots, x_n^k)$, $k = 1, 2, \ldots$

Вектор $x = (x_1, x_2, \ldots, x_n)$ называют *пределом* этой последовательности, если существует каждый из n указанных ниже пределов и

$$\lim_{k\to\infty} x_i^k = x_i \quad (i = 1, 2, \ldots, n).$$

Аналогично, если дана последовательность квадратных матриц $A_k = (a_{ij}^k)$ $(k = 1, 2, \ldots)$, то матрица $A = (a_{ij})$ называется *пределом* этой последовательности, если существуют n^2 указываемых ниже пределов и верны равенства

$$\lim_{k\to\infty} a_{ij}^k = a_{ij} \quad (i, j = 1, 2, \ldots, n).$$

Сходимость последовательности позволяет определить сходимость векторных и матричных рядов. Например, если дан бесконечный матричный ряд $A_1 + A_2 + \ldots$, то говорят, что он сходится к сумме $S = (s_{ij})$, если

конечная сумма $S_k = A_1 + A_2 + \ldots + A_k$ будет сходиться к S при $k \to \infty$; иначе говоря, если верны n^2 численных равенств

$$\sum_{k=1}^{\infty} a_{ij}^k = s_{ij} \quad (i, j = 1, 2, \ldots, n).$$

Определенная так сходимость часто называется *сходимостью по составляющим.*

В следующем пункте будет определена сходимость другого вида — по норме, но в случае векторов и матриц конечной размерности эти две сходимости оказываются, как будет выяснено ниже, равносильными. Вторая же сходимость будет введена потому, что во многих случаях ею пользоваться более удобно, чем первой.

2. Нормы векторов и матриц. Норма вектора может быть определена многими способами в зависимости от условий задачи и целей исследования, но при всяком определении она должна удовлетворять указываемым ниже трем условиям, которые являются аксиомами общей или абстрактной нормы.

Нормой вектора x называют действительное число $\|x\|$, удовлетворяющее условиям

1) $\|x\| > 0$, если $x \neq 0$, и $\|0\| = 0$;
2) $\|cx\| = |c| \|x\|$ при любом численном множителе c;
3) $\|x + y\| \leqslant \|x\| + \|y\|$.

Отметим полезное неравенство, просто вытекающее из аксиом:

$$\|x - y\| \geqslant |\|x\| - \|y\||. \tag{1}$$

Действительно,

$$\|x\| = \|x - y + y\| \leqslant \|x - y\| + \|y\|, \quad \|x - y\| \geqslant \|x\| - \|y\|.$$

Аналогично

$$\|x - y\| = \|y - x\| \geqslant \|y\| - \|x\| = -(\|x\| - \|y\|).$$

Сравнение двух полученных результатов доказывает справедливость (1).

Говорят, что последовательность векторов x^k *сходится к вектору* x *по норме*, если $\|x - x^k\| \to 0$ $(k \to \infty)$. Это определение сходимости не зависит от того, будет ли векторное пространство иметь конечную или бесконеч-

ную размерность. Если размерность имеет конечное значение n, то сходимость по норме и сходимость по компонентам равносильны, как видно из приводимой ниже теоремы.

Теорема 1. *Для того чтобы последовательность векторов* $x^k = (x_1^k, \ldots, x_n^k)$ *сходилась к вектору* $x = (x_1, \ldots, x_n)$ *по составляющим, необходимо и достаточно, чтобы она сходилась к* x *по норме*: $\|x - x^k\| \to 0$ $(k \to \infty)$.

Доказательство. Проверим необходимость. Пусть $x_i^k \to x_i$ $(k \to \infty,\ i = 1, \ldots, n)$. Введем векторы $e_1 = (1, 0, \ldots, 0), e_2 = (0, 1, 0, \ldots, 0), \ldots, e_n = (0, \ldots, 0, 1)$. При помощи них можно записать равенство

$$x - x^k = \sum_{i=1}^{n} (x_i - x_i^k) e_i.$$

Если обозначить $\max_i \|e_i\| = N$, то на основании аксиом нормы получим

$$\|x - x^k\| \leqslant N \sum_{i=1}^{n} |x_i - x_i^k|.$$

Отсюда видно, что $\|x - x^k\| \to 0$ $(k \to \infty)$, и последовательность x^k сходится к x по норме.

Достаточность проверяется почти так же просто.

Предположим, что $\|x - x^k\| \to 0$. Так как

$$\|x^k\| = \|x + (x^k - x)\| \leqslant \|x\| + \|x^k - x\|,$$

то $\|x^k\|$ при $k = 1, 2, \ldots$ будет ограниченной величиной: $\|x^k\| \leqslant M$. Покажем, что будет также ограниченной величина $c_k = |x_1^k| + \ldots + |x_n^k|$ $(k = 1, 2, \ldots)$. Допустим противоположное и предположим, что существует такая последовательность номеров $k_1, k_2, \ldots$, что $c_{k_m} \to \infty$ $(m \to \infty)$. Изменяя, если нужно, нумерацию, можно считать, что $c_k \to \infty$ $(k \to \infty)$.

По x^k построим новую систему векторов $y^k = \frac{1}{c_k} x^k$. Для нее, очевидно,

$$|y_1^k| + \ldots + |y_n^k| = \frac{1}{c_k} |x_1^k| + \ldots + \frac{1}{c_k} |x_n^k| = 1.$$

Поэтому составляющие векторов y^k ограничены в совокупности, и можно выбрать из y^k частичную последовательность y^{k_m}, сходящуюся по составляющим к некоторому вектору y. Изменяя нумерацию, можно считать, что этой сходимостью обладает последовательность y^k:

$$y_i^k \to y_i \quad (i = 1, 2, \ldots, n).$$

Так как $|y_1| + \ldots + |y_n| = 1$, то предельный вектор $y = (y_1, \ldots, y_n)$ не является нулевым. Но

$$\|y\| = \|y^k + (y - y^k)\| \leqslant \|y^k\| + \|y - y^k\| \leqslant$$
$$\leqslant \frac{1}{c_k}\|x^k\| + \|y - k^k\|.$$

Левая часть не зависит от k; в первом члене правой части $\|x^k\| \leqslant M$ и $c_k \to \infty$ при $k \to \infty$, второй же член стремится к нулю, так как из сходимости y^k по составляющим к y следует сходимость по норме. Поэтому правая часть стремится к нулю; следовательно, $\|y\| = 0$ и $y = 0$, что противоречит предыдущему.

Последнее доказывает ограниченность в совокупности величин c_k и x_i^k $(k = 1, 2, \ldots)$. Отсюда вытекает возможность выбора последовательности индексов k, для которой существуют конечные пределы $\xi_i = \lim\limits_k x_i^k$ $(i = 1, \ldots, n)$ и, стало быть, существует для x^k предельный вектор $\xi = (\xi_1, \ldots, \xi_n)$ в смысле сходимости по составляющим.

Осталось убедиться в том, что всякий предельный вектор ξ совпадает с x. Оценим норму разности $x - \xi$:

$$\|x - \xi\| = \|(x - x^k) + (x^k - \xi)\| \leqslant \|x - x^k\| + \|\xi - x^k\|.$$

Левая часть не зависит от k, что же касается правой части, то когда k пробегает последовательность значений, при которой $x^k - \xi$ по составляющим, то $\|\xi - x^k\| \to 0$, как выяснено в начале доказательства теоремы, и $\|x - x^k\| \to 0$ по предположению о сходимости x^k к x по норме. Поэтому $\|x - \xi\| = 0$ и $\xi = x$. Из доказанной теоремы и из неравенства (1) вытекает непрерывность зависимости нормы вектора $\|x\|$ от составляющих $x_1, \ldots, x_n$. В самом деле, если последовательность x^k векторов сходится

по составляющим x_i к вектору x, то она будет сходиться и по норме $\|x - x^k\|$. Но тогда ввиду (1) будет

$$\|x^k\| - \|x\| \to 0 \quad (k \to \infty) \quad \text{и} \quad \|x^k\| \to \|x\|.$$

Укажем теперь наиболее часто применяемые нормы векторов.

1. Первая, *кубическая норма*

$$\|x\|_{\text{I}} = \max_i |x_i|.$$

Эта норма называется *кубической* в связи с тем, что множество точек действительного пространства, удовлетворяющих условию $\|x\|_{\text{I}} \leqslant 1$, образует единичный куб

$$-1 \leqslant x_i \leqslant 1 \quad (i = 1, \ldots, n).$$

2. Вторая, *октаэдрическая норма*

$$\|x\|_{\text{II}} = \sum_{i=1}^{n} |x_i|.$$

Название связано с тем, что множество векторов, для которых $\|x\|_{\text{II}} \leqslant 1$, образует n-мерный аналог октаэдра.

3. Третья, *сферическая*, или *евклидова норма*

$$\|x\|_{\text{III}} = \left[\sum_{i=1}^{n} |x_i|^2\right]^{1/2} = [(x, x)]^{1/2} = |x|.$$

Множество векторов x, для которых $\|x\|_{\text{III}} \leqslant 1$, образует в пространстве шар единичного радиуса.

Можно просто проверить, что во всех трех приведенных примерах аксиомы нормы выполняются. Для первой и второй норм выполнение их является очевидным. Для третьей нормы выполнение первой и второй аксиом также является очевидным; что же касается третьей аксиомы — неравенства треугольника — то это есть известное в математическом анализе неравенство Коши

$$\|x + y\| = \left[\sum_{i=1}^{n} |x_i + y_i|^2\right]^{1/2} \leqslant$$

$$\leqslant \left[\sum_{i=1}^{n} |x_i|^2\right]^{1/2} + \left[\sum_{i=1}^{n} |y_i|^2\right]^{1/2} = \|x\| + \|y\|.$$

Пусть A есть матрица размерности $n \times n$. *Нормой матрицы* A называют число $\|A\|$, удовлетворяющее условиям

1) $\|A\| > 0$, если $A \neq 0$ и $\|0\| = 0$;
2) при всяком численном множителе c

$$\|cA\| = |c| \|A\|;$$

3) $\|A + B\| \leqslant \|A\| + \|B\|$;
4) $\|AB\| \leqslant \|A\| \cdot \|B\|$.

Как и в случае векторов, можно показать, что для нормы матрицы верно неравенство

$$\|A - B\| \geqslant |\|A\| - \|B\|| \tag{2}$$

и что сходимость последовательности A_m матриц к A по составляющим равносильна сходимости по норме: $\|A - A_m\| \to 0$ $(m \to \infty)$.

Отсюда и из (2) вытекает непрерывность зависимости нормы матрицы $\|A\|$ от элементов a_{ij} матрицы. В большом числе вопросов приходится одновременно рассматривать матрицы и векторы, и поэтому нормы их рационально вводить так, чтобы они были в какой-то мере согласованными. Обычно говорят, что норма матрицы A *согласована* с нормой вектора, если для всякого вектора x размерности n выполняется неравенство

$$\|Ax\| \leqslant \|A\| \cdot \|x\|. \tag{3}$$

Среди согласованных норм матрицы часто, особенно при получении оценок, чтобы сделать их точными, избирают наименьшую. В неравенстве случай нулевого вектора x интереса не имеет, так как тогда обе части (3) обращаются в нуль. Можно считать $\|x\| \neq 0$, а (3) взять в форме $\|Ay\| \leqslant \|A\|$, $y = x/\|x\|$, $\|y\| = 1$. Последнее означает, что согласованная норма матрицы должна быть верхней границей норм векторов Ay, при условии $\|y\| = 1$. Наименьшей же согласованной нормой будет точная верхняя граница множества значений $\|Ay\|$ при $\|y\| = 1$:

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

Она называется нормой матрицы, *подчиненной* норме вектора.

Отметим, что в определении (4) Ax есть непрерывная функция от x, и она рассматривается на ограниченном замкнутом множестве векторов, для которых $\|x\| = 1$. Поэтому максимальное значение $\|Ax\|$ достигается, и существует такой вектор x_0, что $\|x_0\| = 1$ и $\|Ax_0\| = \max \|Ax\| = \|A\|$.

Указанные выше четыре аксиомы абстрактной нормы матрицы оставляют широкие возможности выбора нормы, и она может быть задана многими способами.

Приведем примеры подчиненных норм матрицы.

III. Третья, или сферическая, норма вектора

$$\|x\|_{\text{I}} = \max_i |x_i|.$$

Оказывается, что подчиненная этой норме вектора норма матрицы есть

$$\|A\|_{\text{I}} = \max_i \sum_{j=1}^{n} |a_{ij}|. \tag{5}$$

II. Вторая, или октаэдрическая, норма вектора

$$\|x\|_{\text{II}} = \sum_{i=1}^{n} |x_i|.$$

Подчиненная ей норма матрицы есть

$$\|A\|_{\text{II}} = \max_j \sum_{i=1}^{n} |a_{ij}|. \tag{6}$$

III. Третья, или сферическая, норма вектора

$$\|x\|_{\text{III}} = \left[\sum_{i=1}^{n} |x_i|^2\right]^{1/2}.$$

Подчиненная ей норма матрицы является следующей:

$$\|A\| = \sqrt{\Lambda_1}. \tag{7}$$

Здесь Λ_1 есть наибольшее собственное значение матрицы A^*A, где A^* — эрмитово сопряженная с A.

Во всех трех примерах утверждение о подчиненной норме матрицы были высказаны без доказательств. В примерах I и II доказательства получаются просто, и мы на них не станем останавливаться. Немного более

сложно доказывается утверждение в примере III, и на нем мы остановим внимание.

Покажем, что все собственные значения матрицы A^*A являются действительными и неотрицательными числами. Для этого воспользуемся известным правилом перестановки матричного сомножителя в скалярном произведении: $(Bx, x) = (x, B^*x)$. Пусть Λ есть собственное значение A^*A, и $x \neq 0$ — соответствующий ему собственный вектор. Умножим почленно равенство $A^*Ax = \Lambda x$ на x справа. Воспользовавшись тем, что $(A^*Ax, x) = (Ax, (A^*)^*x) = (Ax, Ax) = \|Ax\|^2_{III}$ и $(\Lambda x, x) = \Lambda(x, x) = \Lambda \|x\|^2_{III}$, получим

$$\|Ax\|^2_{III} = \Lambda \|x\|^2_{III}.$$

Отсюда сразу следует $\Lambda \geqslant 0$.

Напомним, что матрица B называется эрмитовой, если $B^* = B$. Интересующая нас матрица A^*A является эрмитовой, так как $(A^*A)^* = (A)^*(A^*)^* = A^*A$. Относительно же эрмитовых матриц в алгебре доказывается, что они обладают полной системой n собственных векторов, которые можно взять ортогональными.

Пусть собственные значения A^*A перенумерованы в порядке убывания

$$\Lambda_1 \geqslant \Lambda_2 \geqslant \ldots \geqslant \Lambda_n \geqslant 0$$

и отвечающие им собственные векторы суть

$$x^1, x^2, \ldots, x^n.$$

Можно считать их ортонормированными.

Возьмем произвольный вектор x с единичной нормой и разложим его по собственным векторам:

$$x = \sum_{i=1}^{n} a_i x^i.$$

Для коэффициентов a_i разложения должно выполняться равенство

$$\|x\|^2_{III} = (x, x) = \sum_{i=1}^{n} |a_i|^2 = 1.$$

Оценим норму матрицы сверху и снизу:

$$\| Ax \|^2_{III} = (Ax, Ax) = (x, A^*Ax) = (x, \Lambda x) =$$
$$= \left(\sum_i a_i x^i, \Lambda \sum_i a_i x^i\right) = \sum_i \Lambda_i |a_i|^2 \leqslant \Lambda_1 \sum_i |a_i|^2 = \Lambda_1.$$

Значит,

$$\| A \|_{III} = \max_x \| Ax \|_{III} \leqslant \sqrt{\Lambda_1}.$$

Если в качестве x взять первый собственный вектор x^1, то

$$\| Ax^1 \|^2_{III} = (x^1, A^*Ax^1) = (x^1, \Lambda_1 x^1) = \Lambda_1 \| x^1 \|^2_{III} = \Lambda_1.$$

Поэтому

$$\| A \|_{III} = \max_x \| Ax \|_{III} \geqslant \| Ax^1 \|_{III} = \sqrt{\Lambda_1}.$$

Сравнение полученных оценок позволяет сказать, что верно равенство (7).

3. Сходимость матричной геометрической прогрессии. Относительно численной геометрической прогрессии

$$1 + a + a^2 + \ldots + a^m + \ldots$$

известно, что она сходится при условии $|a| < 1$, и сумма ее равна $(1-a)^{-1}$. Мы постараемся выяснить условия, при которых будет сходиться *матричная геометрическая прогрессия*

$$E + A + A^2 + \ldots + A^m + \ldots, \tag{8}$$

и найти ее сумму.

Лемма 1. *Для стремления A^m к нулю при $m \to \infty$ необходимо и достаточно, чтобы все собственные значения матрицы A по модулю были меньше единицы.*

Доказательство. Воспользуемся известной в алгебре теоремой о приведении матрицы к канонической форме Жордана: для каждой матрицы A существует неособенная матрица C такая, что A представима в форме

$$A = CIC^{-1}, \tag{9}$$

где I есть квазидиагональная матрица $[I_{t_1}(\lambda_1), I_{t_2}(\lambda_2), \ldots, I_{t_r}(\lambda_r)]$, $\lambda_1, \lambda_2, \ldots, \lambda_r$ — собственные значения

матрицы A, r — число линейно независимых собственных векторов матрицы A и $I_t(\lambda)$ есть ящик Жордана порядка t:

$$I_t(\lambda) = \begin{bmatrix} \lambda & 0 & 0 & \dots & 0 & 0 \\ 1 & \lambda & 0 & \dots & 0 & 0 \\ 0 & 1 & \lambda & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 & \lambda \end{bmatrix}.$$

При этом, если n есть порядок матрицы A, то $\sum_{i=1}^{r} t_i = n$.

Из (9) вытекает

$$A^m = CIC^{-1} \cdot CIC^{-1} \cdot \dots \cdot CIC^{-1} = CI^mC^{-1}.$$

Поэтому A^m и I^m одновременно стремятся к нулю при $m \to \infty$. С другой стороны, $I^m = \left[I^m_{t_1}(\lambda_1), \dots, I^m_{t_r}(\lambda_r)\right]$, и условия сходимости $I^m \to 0$ равносильны условиям сходимости $I^m_t(\lambda) \to 0$ $(m \to \infty)$. При помощи несложных вычислений можно показать, что

$$I^m_t(\lambda) = \begin{bmatrix} \lambda^m & 0 & 0 & \dots & 0 & 0 \\ \frac{1}{1!}(\lambda^m)' & \lambda^m & 0 & \dots & 0 & 0 \\ \frac{1}{2!}(\lambda^m)'' & \frac{1}{1!}(\lambda^m)' & \lambda^m & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{1}{(t-1)!}(\lambda^m)^{(t-1)} & \frac{1}{(t-2)!}(\lambda^m)^{(t-2)} & \dots & & \frac{1}{1!}(\lambda^m)' & \lambda^m \end{bmatrix}.$$

Диагональными элементами являются λ^m, а чтобы они стремились к нулю, должно выполняться условие $|\lambda| < 1$. Вместе с тем это условие достаточно для сходимости $I^m_t(\lambda) \to 0$, так как $(\lambda^m)^{(i)} \to 0$ для $i = 0, 1, \dots, t-1$.

Условие $|\lambda_i| < 1$ должно выполняться при $i = 1, \dots, r$. Это доказывает лемму.

Полученный признак хотя и дает полное решение вопроса о сходимости $A^m \to 0$, но требует достаточно хороших сведений о собственных значениях λ_i. Укажем более удобный достаточный во многих случаях признак.

Лемма 2. *Для стремления A^m к нулю при $m \to \infty$ достаточно, чтобы какая-либо норма матрицы A была меньше единицы.*

Доказательство. Для всякой нормы выполняются неравенства

$$\|0 - A^m\| = \|A^m\| \leqslant \|A^{m-1}\| \cdot \|A\| \leqslant \ldots \leqslant \|A\|^m,$$

и если $\|A\| < 1$, то $\|0 - A^m\| \to 0$ $(m \to \infty)$, т. е. $A^m \to 0$.

Докажем еще одну лемму.

Лемма 3. *Модули собственных значений матрицы не превосходят любой из ее норм.*

Доказательство. Возьмем любое $\varepsilon > 0$ и построим вспомогательную матрицу

$$B = \frac{1}{\|A\| + \varepsilon} A.$$

Очевидно, что

$$\|B\| = \frac{\|A\|}{\|A\| + \varepsilon} < 1,$$

и, стало быть, по лемме 2, $B^m \to 0$ $(m \to \infty)$. Поэтому, ввиду леммы 1, все собственные значения B по модулю меньше единицы. Но они отличаются от собственных значений A множителем $(\|A\| + \varepsilon)^{-1}$. Поэтому для каждого собственного значения λ матрицы A выполняется неравенство $(\|A\| + \varepsilon)^{-1}|\lambda| < 1$, или $|\lambda| < \|A\| + \varepsilon$. Так как ε есть произвольное положительное число, то должно быть $|\lambda| \leqslant \|A\|$.

Перейдем к геометрической прогрессии (8).

Теорема 2. *Для того чтобы ряд* (8) *сходился, необходимо и достаточно, чтобы все собственные значения матрицы A были по модулю меньше единицы. При выполнении этого условия матрица $E - A$ имеет обратную и верно равенство*

$$E = A + A^2 + \ldots + A^m + \ldots = (E - A)^{-1}. \quad (10)$$

Доказательство. Во всяком сходящемся числовом ряду его член номера m стремится к нулю при $m \to \infty$. Сходимость же матричного ряда равносильна сходимости n^2 числовых рядов, и поэтому в сходящемся матричном ряду его член номера m также стремится к нулю при $m \to \infty$. Следовательно, если прогрессия,

стоящая слева в (10), сходится, то $A^m \to 0$ $(m \to \infty)$; последнее же равносильно тому, что для собственных значений λ_i $(i = 1, \ldots, n)$ матрицы A выполняется неравенство $|\lambda_i| < 1$. Этим доказана необходимость условия теоремы.

Проверим достаточность условия. Пусть неравенства $|\lambda_i| < 1$ $(i = 1, \ldots, n)$ выполнены. Собственные значения $E - A$ равны $1 - \lambda_i$ $(i = 1, \ldots, n)$ и будут отличны от нуля. Значит, определитель матрицы $E - A$, равный произведению всех собственных значений ее, взятому со знаком $+$ или $-$, также отличен от нуля, и обратная матрица $(E - A)^{-1}$ существует.

Если обе части тождества

$$(E + A + \ldots + A^m)(E - A) = E - A^{m+1}$$

умножить справа на $(E - A)^{-1}$, получим следующее выражение конечной суммы прогрессии:

$$E + A + \ldots + A^m = (E - A)^{-1} - A^{m+1}(E - A)^{-1}.$$

Когда $m \to \infty$, отсюда, ввиду $A^{m+1} \to 0$, следует равенство

$$E + A + \ldots + A^m + \ldots = (E - A)^{-1},$$

показывающее, что геометрическая прогрессия сходится и имеет указанную в теореме сумму.

Приведем еще теорему, дающую достаточное условие сходимости прогрессии, более простое для проверки и являющееся простым следствием леммы 3 и теоремы 2.

Теорема 3. *Если какая-либо из норм матрицы A меньше единицы, то прогрессия* (8) *сходится.*

Скорость сходимости прогрессии позволяет оценить

Теорема 4. *Если* $\|A\| < 1$, *то верно неравенство*

$$\|(E - A)^{-1} - (E + A + \ldots + A^m)\| \leqslant \frac{1}{1 - \|A\|} \|A\|^{m+1}.$$

Доказательство. При условии $\|A\| < 1$ из теоремы 3 и (10) следует равенство

$$(E - A)^{-1} - (E + A + \ldots + A^m) = A^{m+1} + A^{m+2} + \ldots$$

и, стало быть,

$$\|(E-A)^{-1}-(E+A+\ldots+A^m)\|\leqslant$$
$$\leqslant\|A\|^{m+1}+\|A\|^{m+2}+\ldots=\frac{1}{1-\|A\|}\|A\|^{m+1}.$$

§ 3. Методы исключения неизвестных

1. Метод Гаусса, схема единственного деления. Начнем с изложения *метода Гаусса*, идеи которого известны из школьных курсов математики. Он может осуществляться при помощи многих вычислительных схем и приводится, как будет видно из дальнейшего, к преобразованию заданной системы к эквивалентной системе с верхней треугольной матрицей.

Рассмотрим достаточно удобную для вычислений схему единственного деления. Предположим, что дана система

$$\begin{aligned}
a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n&=f_1,\\
a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n&=f_2,\\
\cdots\cdots\cdots\cdots\cdots&\\
a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n&=f_n.
\end{aligned}\tag{1}$$

Ее преобразование связано с выбором порядка, в котором будут исключаться неизвестные. Оставим сейчас выбор произвольным и о принципах его будем говорить ниже.

Выберем какое-либо уравнение и какое-либо неизвестное в нем. Единственное условие, которое должно быть соблюдено при выборе, состоит в том, что коэффициент при избранном неизвестном должен быть отличен от нуля. Переставляя, если необходимо, уравнения и изменяя места неизвестных, можно считать, что избрано первое уравнение, неизвестное x_1, и при этом $a_{11}\neq 0$. Разделим уравнение на a_{11} и приведем его к виду

$$x_1+b_{12}x_2+\ldots+b_{1n}x_n=g_1.\tag{2}$$

Исключим теперь x_1 из остальных уравнений системы. Будем умножать (2) последовательно на a_{21}, a_{31}, ..., $\ldots,a_{n1}$ и вычитать из второго, третьего ..., последнего

уравнения системы. Преобразованные так уравнения будут иметь форму

$$\begin{gathered} a_{22.1}x_2 + a_{23.1}x_3 + \ldots + a_{2n.1}x_n = f_{2.1}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n2.1}x_2 + a_{n3.1}x_3 + \ldots + a_{nn.1}x_n = f_{n.1}, \\ a_{ij.1} = a_{ij} - b_{1j}\cdot a_{i1} \quad (i, j \geqslant 2), \qquad f_{i1} = f_i - a_{i1}g_1. \end{gathered} \tag{3}$$

Они образуют систему $n-1$ уравнений с неизвестными $x_2, \ldots, x_n$. Порядок ее на единицу меньше, чем у исходной системы. К ней можно применить такое же преобразование: выбрать в ней уравнение и неизвестное с коэффициентом, отличным от нуля, привести этот коэффициент к единице, исключить неизвестное из прочих уравнений и т. д. до тех пор, пока такие преобразования возможны, т. е. либо когда мы переберем все уравнения системы, либо когда в оставшихся уравнениях не будет коэффициентов, отличных от нуля.

Предположим, что было проделано m шагов преобразований, где $m < n$, и при этом новая система приняла вид

$$\begin{gathered} x_1 + b_{12}x_2 + b_{13}x_3 + \ldots + b_{1n}x_n = g_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_m + b_{mm+1}x_{m+1} + \ldots + b_{mn}x_n = g_m, \\ 0 = f_{m+1,\,m}, \\ \ldots\ldots \\ 0 = f_{n,\,m}. \end{gathered} \tag{4}$$

Если среди свободных членов $f_{m+1,\,m}, \ldots, f_{n,\,m}$ будут отличные от нуля, система не имеет решения, и уравнения заданной системы (1) являются несовместными. Если же свободные члены $f_{m+1,\,m}, \ldots, f_{n,\,m}$ все равны нулю, последние $n-m$ уравнений системы (4) выполняются тождественно, и система приводится к первым m уравнениям. Из них могут быть найдены неизвестные $x_1, \ldots, x_m$. Другие неизвестные $x_{m+1}, \ldots, x_n$ остаются произвольными и им можно придавать любые значения. В этом случае заданная система является неопределенной.

Пусть $m = n$. Тогда после преобразований получится система

$$\begin{aligned} x_1 + b_{12}x_2 + b_{12}x_3 + \ldots + b_{1n}x_n &= g_1, \\ x_2 + b_{23}x_3 + \ldots + b_{2n}x_n &= g_2, \\ \cdots\cdots\cdots\cdots\cdots & \\ x_{n-1} + b_{n-1n}x_n &= g_{n-1}, \\ x_n &= g_n. \end{aligned} \tag{5}$$

Последнее уравнение дает значение x_n, из предпоследнего уравнения находится x_{n-1} и т. д.

Приведение системы (1) к виду (5) называют *прямым ходом* метода Гаусса, нахождение же x_n, x_{n-1}, ..., x_1 из системы (5) — *обратным ходом* этого метода.

Число операций умножения и деления, которые выполняются в схеме единственного деления *), равно $\frac{n}{6}(2n^2 + 9n + 1)$, что легко может быть проверено.

В вычислениях, чтобы избежать ошибок, обычно применяют контрольные операции, которые позволяют сделать весьма мало вероятным наличие арифметических погрешностей. В методе Гаусса таблицу коэффициентов a_{ij} и свободных членов f_i принято дополнять столбцом сумм коэффициентов отдельных уравнений и соответствующих свободных членов: $s_i = \sum_{j=1}^{n} a_{ij} + f_i$. Столбец таких сумм помещают рядом со столбцом свободных членов f_i и проделывают с s_i такие же действия, как и с f_i. Контролем будет то обстоятельство, что в каждом вновь полученном уравнении сумма его коэффициентов и свободного члена должна совпадать с вновь полученным значением вспомогательной величины s_i.

Напомним, что на каждом шаге прямого хода выбираются уравнение и неизвестное, подлежащее исключению из прочих уравнений. Это равносильно выбору коэффициента в матрице системы, и его называют *ведущим* для выполняемого шага преобразований. Он должен быть отличным от нуля. Последовательность

*) Считая и операции с очевидным результатом, такие, как деление числа на себя и умножение единицы на число.

ведущих коэффициентов зависит от системы и указывается в ходе вычислений.

Обратим внимание на то, что если ведущий коэффициент близок к нулю, то деление уравнения на него может привести к большим новым коэффициентам. В обратном же ходе метода Гаусса на эти большие коэффициенты будут множиться неточные значения неизвестных, что может привести к росту погрешностей и быстрой потере точности. Чтобы избежать этого, за ведущий коэффициент принимают либо максимальный по модулю коэффициент для всей системы, либо максимальный по модулю коэффициент в избранном уравнении.

2. Метод оптимального исключения. Этот метод можно рассматривать как видоизменение метода Гаусса. Он требует при осуществлении меньше элементов памяти; это связано с тем, что на каждом шаге преобразуется не вся система, как в методе Гаусса, а только часть ее уравнений. Если использовать только оперативную память ЭВМ, то методом оптимального исключения можно решать системы с числом неизвестных x приблизительно в два раза большим, чем по методу Гаусса.

Возвратимся к системе (1). Для выполнения первого шага преобразований изберем ведущий коэффициент. Он должен быть отличным от нуля, и его выбирают обычно наибольшим по модулю во взятом уравнении или в системе. Можно считать, что ведущим является коэффициент a_{11} в первом уравнении. Разделив на него уравнение, приведем последнее к виду (2). Этим заканчивается, как и в методе Гаусса, первый шаг. Выберем новое уравнение для преобразований совместно с первым на втором шаге. Пусть это будет второе уравнение системы. Исключим из него x_1 с помощью первого уравнения; для этого умножим (2) на a_{21} и вычтем из второго уравнения, после чего оно примет вид

$$a_{22.1}x_2 + a_{23.1}x_3 + \ldots + a_{2n.1}x_n = f_{2.1}. \qquad (6)$$

Выберем в нем ведущий коэффициент. Может оказаться, что этого нельзя сделать, и все коэффициенты в (6) будут равны нулю. Тогда уравнение (6) будет иметь форму $0 = f_{2.1}$. Когда $f_{2.1} \neq 0$, два первых уравнения системы оказываются несовместными; когда же $f_{2.1} = 0$, второе

уравнение будет отличаться от первого только множителем и может быть при решении опущено.

Предположим, что в (6) существует ведущий коэффициент. Можно считать, что это есть $a_{22.1}$. Разделив на него уравнение, мы приведем (6) к виду

$$x_2 + b_{23}x_3 + \ldots + b_{2n}x_n = g_2.$$

Пользуясь им, можно исключить из (2) неизвестное x_2. Тогда получим первые два уравнения системы в форме

$$\begin{aligned} x_1 + b_{13.1}x_3 + \ldots + b_{1n.1}x_n &= g_{1.1}, \\ x_2 + b_{23.1}x_3 + \ldots + b_{2n.1}x_n &= g_{2.1}. \end{aligned} \tag{7}$$

Этим заканчивается второй шаг преобразований.

Затем из оставшихся уравнений выбирается одно для преобразований с (7). Пусть это есть третье уравнение системы (1). Исключим из него x_1 и x_2 с помощью (7) и придадим ему вид

$$a_{33.2}x_3 + \ldots + a_{3n.2}x_n = f_{3.2}. \tag{8}$$

Может оказаться, что все коэффициенты полученного уравнения равны нулю, и оно будет либо тождеством $0 = 0$, когда $f_{3.2} = 0$, либо — невыполнимым равенством $0 = f_{3.2}$ при $f_{3.2} \neq 0$. В первом случае третье уравнение системы будет следствием первых двух и может быть опущено; во втором же случае оно несовместно с первыми двумя уравнениями, и заданная система не имеет решения.

Предположим, что не все коэффициенты (8) равны нулю, и среди них можно выбрать ведущий, например, наибольший по модулю. Пусть ведущим является $a_{33.2}$. После деления на него уравнение (8) перейдет в следующее:

$$x_3 + b_{34.2}x_4 + \ldots + b_{3n.2}x_n = g_{3.2}.$$

Его используют для исключения из (7) неизвестного x_3 и приведения системы трех первых уравнений к виду

$$\begin{aligned} x_1 + b_{14.2}x_4 + \ldots + b_{1n.2}x_n &= g_{1.2}, \\ x_2 + b_{24.2}x_4 + \ldots + b_{2n.2}x_n &= g_{2.2}, \\ x_3 + b_{34.2}x_4 + \ldots + b_{3n.2}x_n &= g_{3.2}. \end{aligned} \tag{9}$$

После этого приступают к четвертому шагу: выбирают четвертое уравнение и исключают из него при помощи (9) неизвестные x_1, x_2, x_3, принимают во вновь полученном уравнении какой-либо коэффициент за ведущий и т. д.

Если возможны все n шагов исключений, то в результате получится n уравнений вида

$$x_i = g_{i,\,n-1} \quad (i = 1,\ 2,\ \ldots,\ n),$$

дающие численные значения неизвестных. Если же возможны меньше чем n шагов, то это будет свидетельствовать, как указывалось выше, либо о том, что в системе меньше чем n независимых уравнений, либо о том, что система не имеет решений.

Контроль вычислений может проводиться так же, как в методе Гаусса.

Можно показать, что число умножений и делений, нужных для решения системы порядка n по методу оптимального исключения, равно $\frac{1}{3}n(n^2+3n+2)$. Это почти столько же, сколько в методе Гаусса.

Метод оптимального исключения близок к методу Гаусса и отличается существенно лишь тем, что обратный ход метода Гаусса здесь видоизменен и соединен с прямым ходом.

3. Вычисление определителей. Идея способа Гаусса последовательного исключения неизвестных в системе уравнений может быть перенесена на задачу *вычисления определителей*, и здесь она переходит в способ последовательного понижения порядка n определителя. Рассмотрим схему единственного деления. Пусть дан определитель

$$D = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}.$$

Выберем как-либо ведущий элемент первого шага преобразований. Он должен быть отличным от нуля; чтобы избежать сильного разброса в порядках чисел, за него принимают либо наибольший по модулю элемент D, либо наибольший элемент в избранной строке или

избранном столбце. Выполняя, если нужно, перестановку строк и столбцов, можно считать, что за ведущий элемент принят a_{11}. Вынося a_{11} из первой строки (первого столбца) за знак D, приведем определитель к виду

$$D = a_{11}\begin{vmatrix} 1 & b_{12} & \dots & b_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}.$$

Умножая первую строку последовательно на a_{21}, a_{31}, ..., ..., a_{n1} и вычитая из второй, третьей и т. д. строк, получим

$$D = a_{11}\begin{vmatrix} 1 & b_{12} & \dots & b_{12} \\ 0 & a_{22.1} & \dots & a_{2n.1} \\ \dots & \dots & \dots & \dots \\ 0 & a_{n2.1} & \dots & a_{nn.1} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22.1} & a_{23.1} & \dots & a_{2n.1} \\ a_{32.1} & a_{33.1} & \dots & a_{3n.1} \\ \dots & \dots & \dots & \dots \\ a_{n2.1} & a_{n3.1} & \dots & a_{nn.1} \end{vmatrix}. \quad (10)$$

Этим мы понизим порядок определителя на единицу и можем перейти ко второму шагу преобразований, применяя к полученному определителю порядка $n-1$ такие же преобразования. Выполняя все n шагов, найдем определитель D как произведение ведущих элементов:

$$D = a_{11} \cdot a_{22.1} \cdot a_{33.2} \cdots a_{nn.n-1}. \quad (11)$$

Можно было бы применить к вычислению определителя идеи метода оптимального исключения неизвестных, но, как легко видеть, в этом случае будет проделана излишняя, по сравнению с изложенным методом, вычислительная работа. Это связано с тем, что в описанных выше преобразованиях таблица элементов, если не понижать порядка определителей, будет приведена к правой треугольной; определитель по этой причине будет вычислен просто, так как он равен произведению диагональных элементов. При применении же метода оптимального исключения таблица будет приведена не к треугольной, а к диагональной, что при вычислении определителя является излишним упрощением.

4. Обращение матриц и уточнение приближенной обратной матрицы. Пусть A есть неособенная матрица порядка n. Известно, что элементы обратной для нее матрицы A^{-1} имеют следующее выражение через

матрицу A:

$$\{A^{-1}\}_{ik} = \frac{1}{D(A)} A_{ki}, \tag{12}$$

где A_{ki} есть алгебраическое дополнение элемента a_{ki} матрицы A. Но это представление неудобно для нахождения A^{-1}, так как оно требует вычисления одного определителя D порядка n и n^2 определителей A_{ik} порядка $n-1$.

Обозначим A^{-1} буквой X. По определению обратной матрицы $AX = E$. Столбцы матрицы X будем рассматривать как векторы порядка n и обозначать их $x^k = (x_1^k, \ldots, x_n^k)$ $(k = 1, \ldots, n)$. Аналогично столбцы единичной матрицы E также будем рассматривать как векторы и назовем их $e^k = (0, \ldots, 0, \overset{(k)}{1}, 0, \ldots, 0)$. Вектор x^k является решением системы

$$Ax^k = e^k. \tag{13}$$

Для нахождения обратной матрицы $X = A^{-1}$ достаточно решить n таких систем, отвечающих значениям $k = 1, 2, \ldots, n$. Эти системы имеют одну и ту же матрицу коэффициентов и различаются между собой только векторами свободных членов e^k. Решать их можно, например, по схеме единственного деления или методом оптимального исключения. При этом преобразования систем надо вести параллельно, пользуясь объединенной таблицей коэффициентов и столбцов свободных членов

A	e^1	e^2	$\ldots$	e^n
$a_{11}a_{12} \ldots a_{1n}$	1	0		0
$a_{21}a_{22} \ldots a_{2n}$	0	1		0
$\ldots\ldots\ldots$	$\ldots$	$\ldots$		$\ldots$
$a_{n1}a_{n2} \ldots a_{nn}$	0	0		1

Если обратная матрица X найдена недостаточно точно и необходимо ее улучшить, то можно воспользоваться итерационным процессом

$$X_k = X_{k-1}(2E - AX_{k-1}). \tag{14}$$

За X_0 принимается некоторое приближенное значение обратной матрицы.

Для процесса (14) можно просто указать достаточные условия сходимости. Рассмотрим матрицу $G_k = E - AX_k$. Для нее верно получаемое ниже правило уменьшения индекса k и выражение через начальное значение G_0:

$$G_k = E - AX_k = E - AX_{k-1}(2E - AX_{k-1}) = (E - AX_{k-1})^2 = \\ = G_{k-1}^2 = G_{k-2}^4 = \ldots = G_0^{2^k}.$$

Погрешность приближения имеет следующее выражение через G_0:

$$A^{-1} - X_k = A^{-1}(E - AX_k) = A^{-1}G_k = A^{-1}G_0^{2^k}.$$

Отсюда получается оценка

$$\|A^{-1} - X_k\| \leqslant \|A^{-1}\| \cdot \|G_0\|^{2^k}. \qquad (15)$$

Поэтому можно сказать, что если приближенное значение X_0 будет настолько близко к A^{-1}, что $\|G_0\| = \|E - AX_0\| < 1$, то $X_k \to A^{-1}$; при этом сходимость будет весьма быстрой, так как показатель степени 2^k в (15) с ростом k увеличивается очень быстро.

§ 4. Методы, основанные на разложении матрицы коэффициентов

Такие методы основаны на представлении матрицы A коэффициентов системы в форме произведения треугольных матриц. Это позволяет свести решение заданной системы к последовательному решению двух систем с треугольными матрицами, что является задачей более простой, так как в них неизвестные находятся последовательно. Иногда для придания вычислениям единообразия вводят еще диагональную матрицу.

1. Случай эрмитовой матрицы*). Пусть матрица A системы (3.1) является эрмитовой, т. е. совпадает с комплексно-сопряженной транспонированной: $A^* = A$.

*) Излагаемый метод часто называют «методом квадратного корня».

Найдем такую правую треугольную матрицу S и диагональную матрицу D, элементы которой равны 1 или -1, чтобы A имела представление

$$A = S^* DS, \tag{1}$$

$$S = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ 0 & s_{22} & \dots & s_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & s_{nn} \end{bmatrix}, \quad D = \begin{bmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_{nn} \end{bmatrix}.$$

Если матрицы S и D найдены, то заданная система (3.1) может быть решена следующим путем:

$$Ax = S^* DSx = (S^* D)\, Sx = By = f,$$

где $S^*D = B$ есть нижняя треугольная матрица и $y = Sx$ — вспомогательный вектор. Решение системы $Ax = f$ равносильно решению двух треугольных систем

$$By = f \quad \text{и} \quad Sx = y.$$

Выясним теперь условия, при которых возможно представление (1), и укажем правила вычисления элементов s_{ij} матрицы S и знаков элементов D.

В равенстве (1) сравним элементы матриц, стоящих слева и справа. Так как матрицы эрмитовы, то достаточно произвести сравнение элементов, стоящих на главной диагонали и справа от нее. Это даст следующую систему $\frac{1}{2} n(n+1)$ численных уравнений:

$$\begin{aligned} &\bar{s}_{1i} \cdot d_{11} s_{1j} + \bar{s}_{2i} d_{22} s_{2j} + \dots + \bar{s}_{ii} d_{ii} s_{ij} = a_{ij} \quad (i < j), \\ &|s_{1i}|^2 d_{11} + |s_{2i}|^2 d_{22} + \dots + |s_{ii}|^2 d_{ii} = a_{ii} \quad (i = j), \\ &\qquad j = 1, 2, \dots, n. \end{aligned} \tag{2}$$

При $i = j = 1$ второе уравнение будет $|s_{11}|^2 d_{11} = a_{11}$. Чтобы удовлетворить ему, нужно положить $d_{11} = \operatorname{sign} a_{11}$ и $s_{11} = \sqrt{|a_{11}|}$. Знак корня остается произвольным, и мы выберем $+$. Из первого уравнения (2) при $i = 1$ получим $\bar{s}_{1j} = \frac{a_{1j}}{d_{11} s_{11}}$, при этом нужно считать $s_{11} \neq 0$. Для $i = 2$ второе уравнение (2) дает $|s_{12}|^2 d_{11} + |s_{22}|^2 d_{22} = a_{22}$. Отсюда находим

$$d_{22} = \operatorname{sign}(a_{22} - |s_{12}|^2 d_{11}), \quad s_{22} = \sqrt{|a_{22} - |s_{12}|^2 d_{11}|}.$$

Здесь вновь берется положительное значение корня. Из первого уравнения (2) находим, если $s_{22} \neq 0$,

$$\bar{s}_{2j} = \frac{1}{d_{22}s_{22}}(a_{2j} - \bar{s}_{12}d_{11}s_{1j}) \qquad (j = 3, 4, \ldots, n).$$

Продолжая вычисления дальше, получим следующие общие формулы:

$$d_{ii} = \operatorname{sign}\left(a_{ii} - \sum_{p=1}^{i-1} |s_{pi}|^2 d_{pp}\right),$$

$$s_{ii} = \sqrt{\left|a_{ii} - \sum_{p=1}^{i-1} |s_{pi}|^2 d_{pp}\right|} \qquad (i > 1),$$

$$\bar{s}_{ij} = \frac{1}{d_{ii}s_{ii}}\left(a_{ij} - \sum_{p=1}^{i-1} \bar{s}_{pi}d_{pp}s_{pj}\right) \quad (i < j,\ j = i+1, \ldots, n).$$

Условием возможности нахождения s_{ij} является неравенство $s_{ii} \neq 0$.

2. Случай матрицы с отличными от нуля главными минорами. Как будет показано ниже, такая матрица всегда может быть представлена в форме

$$A = BC, \tag{3}$$

где B есть нижняя треугольная матрица

$$B = \begin{bmatrix} \beta_{11} & 0 & \ldots & 0 \\ \beta_{21} & \beta_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ \beta_{n1} & \beta_{n2} & \ldots & \beta_{nn} \end{bmatrix}$$

и C — верхняя треугольная матрица

$$C = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \ldots & \gamma_{1n} \\ 0 & \gamma_{22} & \ldots & \gamma_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \gamma_{nn} \end{bmatrix}.$$

Отсюда сразу же следует возможность приведения решения системы с такой матрицей к решению двух треугольных систем. Действительно,

$$Ax = BCx = B(Cx) = By = f, \quad \text{где} \quad y = Cx.$$

Поэтому решение системы $Ax = f$ приводится к решению двух следующих треугольных систем:

$$By = f, \quad Cx = y.$$

Осталось убедиться в справедливости представления (3).

Теорема 1. *Если главные миноры матрицы A отличны от нуля*:

$$a_{11} \neq 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0, \ \ldots, \ \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix} \neq 0,$$

то матрицу A можно представить в форме (3).

Доказательство. Заметим, что разложение (3) не может быть единственным. Если D есть любая диагональная матрица с элементами, отличными от нуля, то наряду с равенством $A = BC$ верно также равенство $A = (BD^{-1})(DC)$. Поэтому диагональные элементы одной из матриц B или C можно задавать произвольными, отличными от нуля. После этого, как будет видно в последующем, все остальные элементы матриц B и C определятся единственным образом.

Если в равенстве (3) произвести сравнение элементов матриц, стоящих справа и слева, получится система уравнений для определения β_{ik} и γ_{ik}:

$$\sum_{k=1}^{\min(i,\, j)} \beta_{ik}\gamma_{kj} = a_{ij} \qquad (i, j = 1, \ldots, n). \tag{4}$$

При $i = j = 1$ отсюда получается $\beta_{11}\gamma_{11} = a_{11}$. Можно задать, например, γ_{11} и тогда найти β_{11}. Так как $a_{11} \neq 0$, то $\beta_{11}\gamma_{11} \neq 0$.

Для $i = 2$ и $j = 1$ уравнение (4) будет $\beta_{21}\gamma_{11} = a_{21}$, и для $i = 1$, $j = 2$ оно имеет вид $\beta_{11}\gamma_{12} = a_{12}$. Отсюда находятся β_{21} и γ_{12}. При $i = j = 2$ будем иметь $\beta_{22}\gamma_{22} + \beta_{21}\gamma_{12} = a_{22}$, откуда, задавая для γ_{22} любое значение, не равное нулю, сможем определить β_{22}.

Убедимся теперь в том, что $\beta_{22}\gamma_{22} \neq 0$. Для этого рассмотрим главные миноры второго порядка матриц A, B и C:

$$A_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \quad B_2 = \begin{vmatrix} \beta_{11} & 0 \\ \beta_{21} & \beta_{22} \end{vmatrix}, \quad C_2 = \begin{vmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{vmatrix}.$$

Составим произведение B_2C_2 по правилу умножения строки на столбец:

$$B_2C_2 = \begin{vmatrix} \beta_{11}\gamma_{11} & \beta_{11}\gamma_{12} \\ \beta_{21}\gamma_{11} & \beta_{21}\gamma_{12} + \beta_{22}\gamma_{22} \end{vmatrix}.$$

Ввиду приведенных выше уравнений B_2C_2 совпадает с A_2. С другой стороны, $B_2 = \beta_{11}\beta_{22}$ и $C_2 = \gamma_{11}\gamma_{22}$, поэтому $A_2 = \beta_{11}\gamma_{11}\beta_{22}\gamma_{22}$, и так как $A_2 \neq 0$, то должно быть $\beta_{22}\gamma_{22} \neq 0$ и т. д. Продолжая вычисления и выполняя индукцию, убедимся в том, что, во-первых, задавая произвольно отличными от нуля диагональные элементы B либо C, мы единственным образом найдем все элементы матриц B и C; во-вторых, при этом все их диагональные элементы будут отличны от нуля.

§ 5. Метод ортогонализации

Будем считать, что определитель системы (3.1) отличен от нуля и, следовательно, система имеет единственное решение. Запишем ее в виде

$$\begin{aligned} a_{11}x_1 + \ldots + a_{1n}x_n + a_{1n+1} &= 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + \ldots + a_{nn}x_n + a_{nn+1} &= 0, \\ a_{in+1} &= -f_i. \end{aligned} \tag{1}$$

Исключим из рассмотрения неинтересный случай, когда система (1) является однородной и все a_{in+1} равны нулю.

Введем $n+1$-мерные векторы $a^i = (a_{i1}, \ldots, a_{in}, a_{in+1})$ и $y = (x_1, \ldots, x_n, 1)$. Это позволит придать уравнениям (1) форму условий ортогональности вектора y к векторам a^i $(i = 1, \ldots, n)$:

$$(a^i, y) = 0 \quad (i = 1, \ldots, n), \tag{2}$$

что даст возможность применить к нахождению y хорошо известный процесс ортогонализации. Единственное ограничение, которому нужно заранее подчинить вектор y, — это равенство единице его последней составляющей.

Ортогональность y к a^i равносильна ортогональности y к линейному подпространству P_n, натянутому на векторы a^i, и, следовательно, к любому базису этого

подпространства. Построим ортогональный базис при помощи известного *процесса ортогонализации* системы векторов: положим *) $u^1 = a^1$, $v^k = \frac{1}{\|u^k\|} u^k$, ..., $u^k = a^k - \sum_{i=1}^{k-1} (a^k, v^i) v^i$, ... Векторы v^i $(i = 1, 2, \ldots, n)$ образуют в P_n ортонормированный базис.

Присоединим теперь к v^i $(i = 1, \ldots, n)$ еще любой вектор a^{n+1}, от них линейно не зависящий. Таким будет, например, вектор $a^{n+1} = (0, \ldots, 0, 1)$. Выделим из него часть, ортогональную ко всем $v^i (1, \ldots, n)$, положив $u^{n+1} = a^{n+1} - \sum_{i=1}^{n} (a^{n+1}, v^i) v^i$. Для u^{n+1} выполняются равенства $(u^{n+1}, v^i) = 0$ или равносильные им $(u^{n+1}, a^i) = 0$ $(i = 1, \ldots, n)$. Запишем их более подробно:

$$\begin{aligned} &(u^{n+1}, a^1) = a_{11}u_1^{n+1} + \ldots + a_{1n}u_n^{n+1} + u_{n+1}^{n+1}a_{1n+1} = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &(u^{n+1}, a^n) = a_{n1}u_1^{n+1} + \ldots + a_{nn}u_n^{n+1} + u_{n+1}^{n+1}a_{nn+1} = 0. \end{aligned} \tag{3}$$

Последняя составляющая u_{n+1}^{n+1} вектора u^{n+1} отлична от нуля, ибо если она равна нулю, то составляющие $u_1^{n+1}, \ldots, u_n^{n+1}$ были бы решением однородной системы с матрицей $A = \{a_{ij}\}$. Но так как эта система может иметь только нулевое решение, то все составляющие вектора u^{n+1} были бы равны нулю и вектор a^{n+1} являлся бы линейной комбинацией векторов v^i, что противоречит выбору a^{n+1}.

Если уравнения системы (3) разделить на u_{n+1}^{n+1}, то будет видно, что вектор $y = (x_1, \ldots, x_n, 1)$, для которого $x_i = \frac{u_i^{n+1}}{u_{n+1}^{n+1}}$ $(i = 1, \ldots, n)$, будет решением системы (2), а вектор $x = (x_1, \ldots, x_n)$ — решением заданной системы (3.1).

---

*) Здесь и ниже под *нормой вектора* u понимается величина

$$\|u\| = \sqrt{(u, u)}.$$

Метод ортогонализации требует выполнения большего числа умножений и делений, чем метод Гаусса, и в этом отношении уступает ему.

§ 6. Метод простой итерации

1. Об итерационных методах. Они дают возможность найти решение системы, как предел бесконечного вычислительного процесса, позволяющего по уже найденным приближениям к решению построить следующее, более точное приближение. Привлекательной чертой таких методов является их самоисправляемость и простота реализации на ЭВМ. Если в точных методах ошибка в вычислениях, когда она не компенсируется случайно другими ошибками, неизбежно ведет к ошибкам в результате, то в случае сходящегося итерационного процесса ошибка в каком-то приближении исправляется в последующих вычислениях, и такое исправление требует, как правило, только нескольких лишних шагов единообразных вычислений. Итерационный метод, для того чтобы начать по нему вычисления, требует знания одного или нескольких начальных приближений к решению.

Условия и скорость сходимости каждого итерационного процесса существенно зависят от свойств уравнений, т. е. от свойств матрицы системы, и от выбора начальных приближений.

Итерационных процессов для систем линейных уравнений построено большое число. Мы же остановимся только на двух основных и начнем рассмотрение с метода простой итерации.

2. Описание метода простой итерации и условия его сходимости. Пусть дана система линейных алгебраических уравнений $Ax = f$ с неособенной матрицей. В *методе простой итерации* ее предварительно приводят к виду

$$x = Bx + b, \tag{1}$$

где

$$B = -C^{-1}D, \quad b = C^{-1}f, \quad C + D = A, \quad \det C \neq 0.$$

Предположим, что известно приближение $x^0 = (x_1^0, \ldots, x_n^0)$ к точному решению $x^* = (x_1^*, \ldots, x_n^*)$ системы. Все следующие приближения определим правилом

$$x^{k+1} = Bx^k + b, \quad k = 0, 1, 2, \ldots \tag{2}$$

Если последовательность приближений x^k сходится к некоторому предельному вектору x', то он будет решением системы. Действительно, если в равенстве (2) перейти к пределу при $k \to \infty$, считая, что $x^k \to x'$, то в пределе получим $x' = Bx' + b$. Выясним условия сходимости последовательности x^k.

Теорема 1. *Для того чтобы последовательность приближений x^k сходилась, достаточно, чтобы все собственные значения матрицы B были по модулю меньше единицы*:

$$|\lambda_i| < 1 \quad (i = 1, 2, \ldots, n). \tag{3}$$

Доказательство. Найдем выражение любого приближения x^k через x^0:

$$x^k = Bx^{k-1} + b = B[Bx^{k-2} + b] + b = B^2x^{k-2} +$$
$$+ (E + B)b = \ldots = B^kx^0 + (E + B + \ldots + B^{k-1})b. \tag{4}$$

Отсюда и из (3) с учетом (2.10) сразу следует, что при $k \to \infty$ $B^k \to 0$ и

$$E + B + \ldots + B^{k-1} \to E + B + B^2 + \ldots = (E - B)^{-1},$$

откуда $x^k \to (E - B)^{-1} b = x^*$.

Что касается необходимости условия (3), то ответ на такой вопрос дает

Теорема 2. *Если требовать, чтобы последовательность x^k сходилась к x^* при любом начальном приближении x^0, то условие* (3) *является и необходимым.*

Доказательство. Пусть для всякого начального приближения x^0 будет $x^k \to x^*$. Имеем

$$x^* - x^k = (Bx^* + b) - (Bx^{k-1} + b) = B(x^* - x^{k-1}) = \ldots$$
$$\ldots = B^k(x^* - x^0).$$

При $k \to \infty$ разность $x^* - x^k$ стремится к нулю, поэтому последний член цепи равенства должен стремиться к нулю, каким бы ни был вектор $x^* - x^0$. Отсюда сле-

дует, что $B^k \to 0$, последнее же будет лишь в том случае, когда верно неравенство (3) (см. § 2, п. 3, лемма 1).

Применение теорем 1 и 2 требует знания границ собственных значений матрицы B; нахождение их часто является нелегкой задачей. Укажем более простые, но только достаточные признаки сходимости.

Теорема 3. *Для того чтобы последовательность приближений x^k в методе простой итерации сходилась, достаточно, чтобы какая-либо норма матрицы B была меньше единицы.*

Доказательство. Если $\|B\| < 1$, то по лемме 3 из § 2, п. 3 все собственные значения матрицы B по модулю меньше единицы, и по теореме 1 последовательность x^k сходится.

Непосредственным следствием теоремы 3 и равенств (2.5) и (2.6), определяющих кубическую и октаэдрическую норму матрицы, является

Теорема 4. *Последовательность x^k в методе простой итерации сходится, если для матрицы B выполняется одно из неравенств*:

$$1)\ \sum_{j=1}^{n} |b_{ij}| < 1 \quad (i = 1, 2, \ldots, n),$$

$$2)\ \sum_{i=1}^{n} |b_{ij}| < 1 \quad (j = 1, 2, \ldots, n).$$

Для многих приложений важно знать, какой является скорость сходимости $x^k \to x^*$, и уметь оценить погрешность $x^* - x^k$ замены точного решения x^* системы приближением x^k.

Теорема 5. *Если какая-либо норма матрицы B, согласованная с рассматриваемой нормой вектора x, меньше единицы, то верна следующая оценка погрешности приближения в методе простой итерации*:

$$\|x^* - x^k\| \leqslant \|B\|^k \|x^0\| + \frac{1}{1 - \|B\|} \|B\|^k \|b\|. \qquad (5)$$

Доказательство. Для x^k выше дано выражение (4), и так как $\|B\| < 1$, то $x^* = (E + B + B^2 + \ldots) b$. Поэтому

$$x^* - x^k = (B^k + B^{k+1} + \ldots) b - B^k x^0 \qquad (6)$$

и, стало быть,

$$\|x^* - x^k\| \leqslant (\|B\|^k + \|B\|^{k+1} + \ldots)\|b\| + \|B\|^k\|x^0\| =$$
$$= \|B\|^k\|x^0\| + \frac{1}{1-\|B\|}\|B\|^k\|b\|.$$

Часто за x^0 принимают вектор b. В этом случае оценка (5) немного упростится; подставляя $x^0 = b$ в (6), получим

$$\|x^* - x^k\| \leqslant \frac{1}{1-\|B\|}\|B\|^{k+1}\|b\|. \qquad (7)$$

3. Об уточнении приближений и ускорении сходимости в методе простой итерации. Предположим, что выполнены условия (3), и поэтому последовательные приближения x^k сходятся к решению x^* системы.

Сравнительно простой закон сходимости x^k к x^* и простое представление погрешности $x^* - x^k$, указываемое в равенстве (6), позволяют в некоторых случаях выделить из погрешности главную часть и, добавив ее к x^k, уменьшить тем самым погрешность приближения. Кроме того, простота закона сходимости позволяет указать правило ускорения сходимости x^k к x^*.

Предположим, что за начальное приближение x^0 принят свободный вектор системы (6). Такое допущение не уменьшит общности изложения, но позволит не выписывать отдельно в (6) слагаемое $B^k x^0$. При $x^0 = b$ погрешность $x^* - x^k$ имеет представление

$$x^* - x^k = (B^{k+1} + B^{k+2} + \ldots)b. \qquad (8)$$

Выясним наглядную картину поведения погрешности при больших значениях k. Предположим, что матрица B обладает полной системой собственных векторов в n-мерном пространстве. Обозначим их $\xi^1, \xi^2, \ldots, \xi^n$, а отвечающие им собственные значения — $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Будем считать, что λ_i перенумерованы в порядке убывания модулей: $|\lambda_1| \geqslant |\lambda_2| \geqslant \ldots$ Разложим b по собственным векторам ξ^i:

$$b = \alpha_1\xi^1 + \ldots + \alpha_n\xi^n.$$

Ввиду соотношения

$$B^m\xi^i = B^{m-1}(B\xi^i) = B^{m-1}(\lambda_i\xi^i) = \ldots = \lambda_i^m\xi^i$$

для $x^* - x^k$ получится следующее выражение:

$$x^* - x^k = \sum_{m=k+1}^{\infty} \sum_{i=1}^{n} \alpha_i B^m \xi^i = \sum_{i=1}^{n} \alpha_i \sum_{m=k+1}^{\infty} \lambda_i^m \xi^i =$$

$$= \sum_{i=1}^{n} \alpha_i \xi^i \frac{1}{1-\lambda_i} \lambda_i^{k+1}. \quad (9)$$

Так как $|\lambda_i| < 1$ $(i = 1, \ldots, n)$, то $\lambda_i^k \to 0$ при $k \to \infty$, и x^k будет стремиться к x^*. Но сходимость может быть весьма медленной, если среди λ_i будут близкие по модулю к единице. В связи с этим возникают две задачи, которые в первоначальной формулировке кажутся различными: во-первых, как при фиксированном k можно уточнить приближенное значение x^k и, во-вторых, как можно при изменяющемся k ускорить сходимость x^k к x^*?

Рассмотрим первую из задач. Уточнить приближенное значение x^k — это означает, что из погрешности нужно выделить главную часть и добавить ее к x^k. Когда говорят о выделении главной части погрешности, то подразумевают, что новая погрешность, полученная после выделения, должна стремиться к нулю быстрее первоначальной. Поэтому правило выделения, которое будет указано ниже, можно применять лишь при достаточно больших k, когда глазная часть погрешности лежит еще в границах принятой точности, тогда как новая погрешность становится меньше этой границы и ее можно отбросить. Очевидно, такой способ уточнения x^k тесно связан с задачей ускорения сходимости x^k к x^*.

В равенстве (9) нам удобнее от векторов перейти к составляющим:

$$x_m^* - x_m^k = \sum_{i=1}^{n} \alpha_i \frac{1}{1-\lambda_i} \xi_m^i (\lambda_i)^{k+1} \quad (m = 1, 2, \ldots, n). \quad (10)$$

Предположим, что среди собственных значений λ_i есть одно наибольшее по модулю, так что

$$|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots$$

В сумме, стоящей справа в (10), отдельные слагаемые стремятся к нулю, как степени соответствующих λ_i,

и слагаемые расположены в порядке скорости убывания. При больших k главным членом в сумме будет слагаемое с λ_1, если только коэффициент α_1 не окажется случайно равным нулю в пределах принятой точности, и будет верно равенство

$$x_m^* - x_m^k \approx \alpha_1 \frac{1}{1-\lambda_1} \xi_m^1 (\lambda_1)^{k+1} = C_{1m} (\lambda_1)^{k+1}. \qquad (11)$$

Заметим, что оно выполняется тем более точно, чем больше k.

Будем считать k настолько большим, чтобы погрешность равенства была меньше принятой границы для нее. Признак, по которому можно судить о том, что такое значение k достигнуто, укажем ниже.

Если известно собственное значение λ_1, то для получения правила уточнения x_m^k достаточно из равенства (11) исключить не зависящий от k коэффициент C_{1m}. Для этого запишем аналогичное равенство для индекса $k+1$:

$$x_m^* - x_m^{k+1} \approx C_{1m} (\lambda_1)^{k+2} \qquad (12)$$

и вычтем его из (11):

$$\begin{aligned} x_m^{k+1} - x_m^k &\approx C_{1m} (1-\lambda_1) (\lambda_1)^{k+1}, \\ C_{1m} &\approx \frac{1}{1-\lambda_1} (\lambda_1)^{-k-1} (x_m^{k+1} - x_m^k). \end{aligned} \qquad (13)$$

Если это значение C_{1m} внести в (11), то мы найдем приближенное значение для x_m^*:

$$x_m^* \approx x_m^k + \frac{1}{1-\lambda_1} (x_m^{k+1} - x_m^k). \qquad (14)$$

Так как равенство (11) не было точным, то (14) позволит вычислить не точное, а только улучшенное значение для x_m^*.

Возвратимся к вопросу о выборе k. Из (13) следует, что

$$\frac{1}{x_m^{k+1} - x_m^k} C_{1m} \approx \frac{1}{1-\lambda_1} (\lambda_1)^{-k-1} \quad (m = 1, 2, \ldots, n). \qquad (15)$$

Правая часть полученного соотношения не зависит от m; если k выбрано достаточно большим, не должна за-

висеть в принятой точности от m и левая часть этого равенства. Поэтому (15) может служить контролем правильности выбора k, хотя и недостаточно полным, так как из независимости левой части от m в принятой точности еще не следует, принципиально говоря, что k взято достаточно большим.

Рассмотрим теперь задачу уточнения x_m^k в том случае, когда λ_1 является неизвестным. В этом случае правая часть (11) будет содержать два параметра (C_{1m} и λ_1), и для исключения их потребуется рассмотреть равенство вида (11) для трех значений k. Возьмем его для значений $k-1$, k, $k+1$ и присоединим поэтому к равенствам (11), (12) еще равенство, отвечающее значению $k-1$:

$$x_m^* - x_m^{k-1} \approx C_{1m}(\lambda_1)^k. \tag{16}$$

Путем вычитания (11) из (16) и (12) из (11) получим два равенства, свободные от x^*:

$$\begin{aligned} x_m^k - x_m^{k-1} &\approx C_{1m}(1-\lambda_1)\lambda_1^k, \\ x_m^{k+1} - x_m^k &\approx C_{1m}(1-\lambda_1)\lambda_1^{k+1}. \end{aligned} \tag{17}$$

Разделив почленно второе из них на первое, найдем приближенное значение λ_1:

$$\lambda_1 \approx (x_m^{k+1} - x_m^k)(x_m^k - x_m^{k-1})^{-1} = \Delta x_m^k (\Delta x_m^{k-1})^{-1}$$
$$(m = 1, 2, \ldots, n).$$

Затем последовательно можно получить

$$1 - \lambda_1 \approx (2x_m^k - x_m^{k+1} - x_m^{k-1})(\Delta x_m^{k-1})^{-1} = -\Delta^2 x_m^{k-1}(\Delta x_m^{k-1})^{-1},$$

$$C_{1m}\lambda_1^{k+1} \approx (x_m^{k+1} - x_m^k)(1-\lambda_1)^{-1} = -\Delta x_m^k \cdot \Delta x_m^{k-1}(\Delta^2 x_m^{k-1})^{-1}.$$

Подстановка этих величин в (11) позволяет получить приближенное значение x_m^*:

$$x_m^* \approx x_m^k - \Delta x_m^k \cdot \Delta x_m^{k-1}(\Delta^2 x_m^{k-1})^{-1} =$$
$$= [x_m^{k+1}x_m^{k-1} - (x_m^k)^2] \cdot [\Delta^2 x_m^{k-1}]^{-1}. \tag{18}$$

Оно является уточненным значением x_m^k.

В заключение сделаем несколько замечаний о проблеме ускорения сходимости последовательных приближений

к решению системы уравнений. По существу дела это есть задача ускорения сходимости последовательности векторов $x^0, x^1, \ldots, x^k, \ldots$ В общей форме она состоит в построении новой последовательности векторов

$$y^k = f_k(x^0, x^1, \ldots, x^m, \ldots),$$

обладающей следующими свойствами: 1) если последовательность x^k сходится к вектору x^*, то последовательность y^k $(k = 0, 1, \ldots)$ также должна сходиться к x^*; 2) сходимость $y^k \to x^*$ должна быть более быстрой, чем сходимость x^k к x^*. Такая задача есть проблема совместного преобразования n числовых последовательностей — составляющих x_m^k $(m = 1, 2, \ldots, n;$ $k = 0, 1, \ldots)$ векторов x^k — в n новых последовательностей величин

$$y_m^k \quad (m = 1, 2, \ldots, n; \; k = 0, 1, \ldots),$$

являющихся составляющими векторов y^k. Упростим эту проблему: будем рассматривать независимо отдельные последовательности составляющих x_m^k $(k = 0, 1, \ldots)$ и для них поставим задачу об ускорении сходимости. Если считать, как выше, λ_1 наибольшим по модулю собственным значением матрицы B, то из представления (10) для x_m^k следует равенство

$$x_m^k = x_m^* - C_{1m}\lambda_1 (\lambda_1)^k [1 + \delta_k], \tag{19}$$

$$\delta_k = \sum_{i=2}^{n} \frac{1}{C_{1m}} \frac{1}{1-\lambda_i} \xi_m^i \left(\frac{\lambda_i}{\lambda_1}\right)^{k+1}, \quad \delta_k \to 0 \quad (k \to \infty).$$

Правило (18) уточнения приближенного значения x_m^k, если его переписать в форме

$$y_m^k \approx \left[x_m^{k+1} x_m^{k-1} - (x_m^k)^2\right] \cdot \left[x_m^{k+1} - 2x_m^k + x_m^{k-1}\right]^{-1}, \tag{20}$$

является правилом преобразования последовательности x_m^k в последовательность y_m^k, ускоряющим сходимость последовательности x_m^k к x_m^*. Это правило основано на учете свойств только одной главной части погрешности приближения и не учитывает всех прочих частей погрешности, включенных в δ_k. Оно не зависит от пара-

метров C_{1m}, λ_1 и применимо для ускорения сходимости при любых их значениях.

Аналогично правило (14) уточнения x_m^k приводит к преобразованию последовательности x_m^k $(k = 0, 1, 2, \ldots)$ в последовательность

$$y_m^k = x_m^* - \frac{1}{1-\lambda_1}(x_m^{k+1} - x_m^k). \tag{21}$$

Оно также предназначено для ускорения сходимости последовательностей вида (19), но для своего применения требует знания *) λ_1.

§ 7. Метод Зейделя

1. Описание и сходимость метода. *Метод Зейделя* применяют в двух видоизменениях. Рассмотрим сначала случай канонической формы системы для метода итерации

$$x = Bx + b. \tag{1}$$

В методе простой итерации следующее приближение $x^{k+1} = (x_1^{k+1}, \ldots, x_n^{k+1})$ находится по предыдущему $x^k = (x_1^k, \ldots, x_n^k)$ путем подстановки x^k в правую часть всех уравнений системы (1). Для нас сейчас удобнее записать результат подстановки не в векторной форме, а в развернутом виде по составляющим:

$$\begin{aligned}
x_1^{k+1} &= b_{11}x_1^k + b_{12}x_2^k + \ldots + b_{1n}x_n^k + b_1,\\
x_2^{k+1} &= b_{21}x_1^k + b_{22}x_2^k + \ldots + b_{2n}x_n^k + b_2,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
x_n^{k+1} &= b_{n1}x_1^k + b_{n2}x_2^k + \ldots + b_{nn}x_n^k + b_n.
\end{aligned} \tag{2}$$

В этой операции порядок выбора уравнений значения не имеет. Здесь, очевидно, опускаются две возможности улучшения итераций: разумный выбор порядка уравнений для подстановок и немедленный ввод в вычисления каждого из полученных исправленных значений неизвестных.

*) С преобразованием вида (20) читатель встретится в гл. 4, общая же теория преобразований вида (23) и (21) и их обобщений будет кратко изложена во втором томе.

О принципах выбора порядка уравнений будет говориться ниже, а сейчас предположим, что для перехода от приближения x^k к следующему — x^{k+1} нами выбран какой-то порядок привлечения уравнений для подстановок. Изменяя, если необходимо, нумерацию уравнений и неизвестных, можно считать, что уравнения для подстановок берутся в порядке роста их номеров. Для каждого шага от приближения k до $k+1$ порядок привлечения уравнений может быть своим, и должны быть выполнены свои изменения нумерации и перестановки, что влечет за собой свое изменение матрицы B системы и свободного вектора b. Чтобы отметить это, обозначим матрицу B для рассматриваемого шага B^k и свободный вектор b^k. В этих обозначениях итерация в методе Зейделя выполняется в следующем порядке:

$$\begin{aligned}
x_1^{k+1} &= b_{11}^k x_1^k + b_{12}^k x_2^k + b_{13}^k x_3^k + \ldots + b_1^k,\\
x_2^{k+1} &= b_{21}^k x_1^{k+1} + b_{22}^k x_2^k + b_{23}^k x_3^k + \ldots + b_2^k,\\
x_3^{k+1} &= b_{31}^k x_1^{k+1} + b_{32}^k x_2^{k+1} + b_{33}^k x_3^k + \ldots + b_3^k,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned} \tag{3}$$

После нахождения вектора x^{k+1} устанавливают порядок подстановок в уравнения значений x_i^{k+1} $(i = 1, \ldots, n)$ и переходят к вычислению вектора x^{k+2} и т. д.

Приведем теперь пример принципа, на основании которого можно устанавливать порядок привлечения уравнений для подстановок x_i^k $(i = 1, \ldots, n)$. Можно пытаться в первую очередь улучшить ту составляющую решения, которая найдена наименее точно, чтобы при нахождении всех других составляющих употреблять улучшенное ее значение. О точности x^k можно было бы судить по вектору погрешности $\varepsilon^k = x^* - x^k$, но так как вектор точного решения x^* неизвестен, то ε^k в вычислениях заменяют другим вектором, по которому можно, хотя бы неполно, судить о погрешности ε^k. Чаще всего для этой цели пользуются вектором поправки на последнем шаге $\delta^k = (\delta_1^k, \ldots, \delta_n^k)$, где $\delta_i^k = x_i^k - x_i^{k-1}$. Величины поправок составляющих нумеруют в порядке убывания их абсолютных значений, и в том же порядке вычисляют составляющие следующего приближения

x^{k+1}: сначала ту составляющую, которая отвечает наибольшей по модулю поправке, и т. д.

Остановимся более подробно на стационарном методе Зейделя; когда при итерациях порядок уравнений сохраняется, матрица B будет одинаковой на всех шагах и составляющие следующего приближения находятся при всяком k по правилу (3).

Разложим матрицу B на сумму двух матриц H и F, где

$$H = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ b_{21} & 0 & \dots & 0 & 0 \\ b_{31} & b_{32} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ b_{n1} & b_{n2} & \dots & b_{nn-1} & 0 \end{bmatrix}, \quad F = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n-1} & b_{1n} \\ 0 & b_{22} & \dots & b_{2n-1} & b_{2n} \\ 0 & 0 & \dots & b_{3n-1} & b_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & b_{nn} \end{bmatrix}.$$

Тогда равенства (2) можно записать в форме матричного равенства

$$x^{k+1} = Hx^{k+1} + Fx^k + b,$$

откуда следует, что $(E - H)x^{k+1} = Fx^k + b$, а так как определитель матрицы

$$E - H = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ -b_{21} & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -b_{n1} & -b_{n2} & \dots & -b_{nn-1} & 1 \end{bmatrix}$$

равен единице и она имеет обратную, то равенство (3) равносильно

$$x^{k+1} = (E - H)^{-1} Fx^k + (E - H)^{-1} b. \qquad (4)$$

Поэтому стационарный метод Зейделя равносилен методу простой итерации, примененному к системе

$$x = (E - H)^{-1} Fx + (E - H)^{-1} b.$$

Это сразу дает возможность на основании теоремы 1 из § 6, п. 3 сказать, что для сходимости стационарного процесса Зейделя (2) при любом векторе x^0 начального приближения необходимо и достаточно, чтобы все собственные значения матрицы $(E - H)^{-1}F$, т. е. корни уравнения $|(E - H)^{-1}F - \lambda E| = 0$, были по модулю меньше единицы.

Этот признак может быть высказан в форме, не требующей обращения $E-H$. Если воспользоваться тем, что $|(E-H)^{-1}|=|E-H|^{-1}=1$, то можно написать систему равенств

$$|(E-H)^{-1}F-\lambda E|=|(E-H)^{-1}[F-\lambda(E-H)]|=$$
$$=|(E-H)^{-1}|\cdot|F+\lambda H-\lambda E|=|F+\lambda H-\lambda E|.$$

Поэтому верна

Теорема 1. *Для того чтобы стационарный метод Зейделя сходился при любом начальном векторе приближения x^0, необходимо и достаточно, чтобы все корни уравнения*

$$|F+\lambda H-\lambda E|=\begin{vmatrix} b_{11}-\lambda & b_{12} & \dots & b_{1n} \\ \lambda b_{21} & b_{22}-\lambda & \dots & b_{2n} \\ \dots & \dots & \dots & \dots \\ \lambda b_{n1} & \lambda b_{n2} & \dots & b_{nn}-\lambda \end{vmatrix}=0 \quad (5)$$

были по модулю меньше единицы.

Укажем еще более простой достаточный признак сходимости. Предварительно докажем лемму.

Лемма 1. *Если в матрице $A=\{a_{ij}\}$ диагональные элементы a_{ii} $(i=1,\dots,n)$ доминируют по строкам или по столбцам, т. е. если*

$$\sum_{j=1,\ j\neq i}^{n}|a_{ij}|<|a_{ii}| \quad (i=1,\ 2,\ \dots,\ n)$$

или

$$\sum_{i=1,\ i\neq j}^{n}|a_{ij}|<|a_{jj}| \quad (j=1,\ 2,\ \dots,\ n),$$

то определитель матрицы A отличен от нуля.

Доказательство. Для определенности предположим, что имеет место доминирование по строкам. Достаточно показать, что однородная система

$$Ax=0 \quad (6)$$

имеет только нулевое решение.

Допустим противоположное и будем считать, что система имеет ненулевое решение $x^*=(x_1^*,\ x_2^*,\ \dots,\ x_n^*)$. Среди составляющих решения выберем наибольшую по модулю:

$$|x_i^*|\geqslant|x_j^*| \quad (j=1,\ \dots,\ n).$$

Положим $x = x^*$ и оценим снизу левую часть уравнения номера i системы (6):

$$|a_{i1}x_1^* + \ldots + a_{ii}x_i^* + \ldots + a_{in}x_n^*| \geqslant$$

$$\geqslant |a_{ii}| \cdot |x_i^*| - \sum_{j=1,\, j\neq i}^{n} |a_{ij}| \cdot |x_j^*| \geqslant$$

$$\geqslant |x_i^*| \cdot \left(|a_{ii}| - \sum_{j=1,\, j\neq i}^{n} |a_{ij}| \right) > 0,$$

так как $|x_i^*| > 0$ и $|a_{ii}| > \sum_{j=1,\, j\neq i}^{n} |a_{ij}|$ по условию леммы. Этот результат противоречит тому, что x^* есть решение системы, и доказывает неверность допущения.

Теорема 2. *Для сходимости стационарного метода Зейделя* (4) *достаточно, чтобы выполнялось какое-либо одно из условий*:

$$\|B\|_{\mathrm{I}} = \max_i \sum_{j=1}^{n} |b_{ij}| < 1 \tag{7}$$

или

$$\|B\|_{\mathrm{II}} = \max_j \sum_{i=1}^{n} |b_{ij}| < 1. \tag{8}$$

Доказательство. Достаточность первого и второго условий проверяется аналогично, и можно ограничиться рассмотрением только первого условия.

Нужно показать, что при выполнении условия (7) все корни уравнения (5) будут по модулю меньше единицы. Будем считать, что $|\lambda| \geqslant 1$, и рассмотрим сумму модулей недиагональных элементов строки номера i матрицы $F + \lambda H - \lambda E$:

$$|\lambda||b_{i1}| + \ldots + |\lambda||b_{ii-1}| + |b_{ii+1}| + \ldots + |b_{in}| \leqslant$$

$$\leqslant |\lambda|(|b_{i1}| + \ldots + |b_{ii-1}| + |b_{ii+1}| + \ldots + |b_{in}|) =$$

$$= |\lambda| \left(\sum_{j=1}^{n} |b_{ij}| - |b_{ii}| \right) < |\lambda|(1 - |b_{ii}|) =$$

$$= |\lambda| - |\lambda||b_{ii}| \leqslant |\lambda| - |b_{ii}| \leqslant |\lambda - b_{ii}|.$$

Таким образом, диагональные элементы матрицы $F + \lambda H - \lambda E$ доминируют по строкам, и на основании леммы 1 определитель этой матрицы отличен от нуля,

а значение λ, для которого $|\lambda| \geqslant 1$, не может быть корнем уравнения (6). Корни этого уравнения все по модулю меньше единицы, и по теореме 1 стационарный метод Зейделя сходится.

2. Другая форма метода Зейделя. В ней требуется предварительное преобразование системы $Ax = b$ к виду, в котором все диагональные коэффициенты отличны от нуля. Такое приведение стремятся выполнить, если это возможно, так, чтобы диагональные коэффициенты были наибольшими или даже доминирующими в соответствующих уравнениях.

Мы ограничимся описанием только стационарного процесса. Пусть взято какое-либо исходное приближение $x^0 = (x_1^0, \ldots, x_n^0)$ к решению системы. Приближение номера $k+1$ находят по приближению номера k с помощью системы соотношений

$$\begin{aligned} a_{11}x_1^{k+1} + a_{12}x_2^k + a_{13}x_3^k + \ldots + a_{1n}x_n^k &= b_1, \\ a_{21}x_1^{k+1} + a_{22}x_2^{k+1} + a_{23}x_3^k + \ldots + a_{2n}x_n^k &= b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1}x_1^{k+1} + a_{n2}x_2^{k+1} + a_{n3}x_3^{k+1} + \ldots + a_{nn}x_n^{k+1} &= b_n. \end{aligned} \tag{9}$$

Если разложить матрицу A на сумму двух матриц

$$B = \begin{bmatrix} a_{11} & 0 & 0 & \ldots & 0 \\ a_{21} & a_{22} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{nn} \end{bmatrix} \quad \text{и} \quad C = \begin{bmatrix} 0 & a_{12} & a_{13} & \ldots & a_{1n} \\ 0 & 0 & a_{23} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix},$$

то равенства (9) можно записать в матричном виде

$$Bx^{k+1} + Cx^k = b,$$

или

$$x^{k+1} = -B^{-1}Cx^k + B^{-1}b.$$

Поэтому ясно, что метод Зейделя в форме (9) равносилен методу простых итераций, примененному к системе в каноническом виде:

$$x = -B^{-1}Cx + B^{-1}b.$$

Для сходимости метода при любом векторе b необходимо и достаточно, как следует из теорем 2 и 3,

чтобы все собственные значения матрицы $-B^{-1}C$, т. е. все корни уравнения $|-B^{-1}C-\lambda E|=0$, были по модулю меньше единицы. Это условие можно упростить и высказать в форме, не требующей обращения матрицы B. В самом деле,

$$|-B^{-1}C-\lambda E|=|(-B^{-1})[C+\lambda B]|=|-B^{-1}|\cdot|C+\lambda B|,$$

и можно формулировать следующую теорему.

Теорема 3. *Для того чтобы процесс Зейделя, определяемый равенствами* (9), *сходился при любых свободных членах* b_i $(i=1,\ldots,n)$, *необходимо и достаточно, чтобы корни уравнения*

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

все были меньше единицы по модулю.

§ 8. Связь с задачей об экстремуме многочлена второй степени

1. Введение. Эта связь очень проста, а методы решения систем, использующие ее, сравнительно редко применяются в вычислениях. Несмотря на это, краткое изложение такой связи сохранено в книге как элементарное введение к более общей и глубокой связи между задачей об экстремуме функционалов и дифференциальными уравнениями, где эта связь используется более часто и более результативно.

Рассмотрим систему уравнений

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

с действительной, симметричной и положительной матрицей A. Векторы x и b будем считать также действительными. Возьмем многочлен второй степени от $x=(x_1,\ldots,x_n)$, сопряженный с системой (1):

$$F(x)=(Ax,\ x)-2(b,\ x)=\sum_{i,\ j=1}^{n}a_{ij}x_ix_j-2\sum_{i=1}^{n}b_ix_i. \tag{2}$$

Ввиду положительности матрицы A он имеет единственный минимум.

Теорема 1. *Если в некоторой точке n-мерного пространства $x=(x_1,\ldots,x_n)$ многочлен $F(x)$ имеет минимум, то координаты этой точки $(x_1,\ldots,x_n)$ удовлетворяют системе* (1).

Доказательство. Пусть $x=(x_1,\ldots,x_n)$ есть произвольная точка пространства. Придадим вектору x приращение $\lambda\Delta x$, где λ — произвольный численный множитель, Δx — произвольный вектор, и рассмотрим соответствующее измененное значение многочлена F:

$$F(x+\lambda\Delta x)=(Ax+\lambda A\Delta x,\ x+\lambda\Delta x)-2(b,\ x+\lambda\Delta x)=$$
$$=(Ax,\ x)+\lambda(Ax,\ \Delta x)+\lambda(A\Delta x,\ x)+$$
$$+\lambda^2(A\Delta x,\ \Delta x)-2(b,\ x)-2\lambda(b,\ \Delta x).$$

Так как $(A\Delta x,\ x)=(\Delta x,\ Ax)=(Ax,\ \Delta x)$, то $F(x+\lambda\Delta x)=$ $=F(x)+2\lambda(Ax-b,\ \Delta x)+\lambda^2(A\Delta x,\ \Delta x)$, и

$$\Delta F=F(x+\lambda\Delta x)-F(x)=2\lambda(Ax-b,\Delta x)+\lambda^2(A\Delta x,\Delta x). \quad (3)$$

Если в точке x многочлен F имеет минимум, то скалярное произведение $(Ax-b,\Delta x)$ должно быть равно нулю при любом Δx. Действительно, если при некотором Δx будет $(Ax-b,\Delta x)\neq 0$, то при малых $|\lambda|$ в правой части (3) первый член будет главным, и если изменить знак у λ, изменит свой знак и вся правая часть, тогда как ΔF должно сохранять знак плюс при всяких λ, малых по абсолютному значению. Ввиду же того, что равенство $(Ax-b,\Delta x)=0$ должно выполняться при любом Δx, отсюда следует $Ax-b=0$, и x действительно есть решение уравнения (1).

Теорема 2. *Если x есть решение уравнения* (1), *то многочлен F имеет в точке x собственный абсолютный минимум во всем пространстве $(x_1,\ldots,x_n)$.*

Доказательство. Если x есть решение системы (1), то изменение F, указанное в (3), будет при любых λ и Δx иметь значение

$$\Delta F=\lambda^2(A\Delta x,\ \Delta x).$$

Матрица A, по предположению, является положительной, правая часть равенства всегда неотрицательна и может обращаться в нуль только при $\Delta x=0$. Это доказывает теорему.

Приведенные две теоремы показывают, что решение системы (1) и нахождение точки минимума многочлена F суть задачи равносильные.

2. Метод покоординатного спуска. Он является общим методом приближения к точке минимума функции нескольких аргументов. Для многочлена второй степени этот метод имеет особенно простую форму, как это будет выяснено ниже.

Пусть имеется приближение $x^0 = (x_1^0, \ldots, x_n^0)$ к точке минимума F. Рассмотрим функцию одного аргумента x_1: $F(x_1, x_2^0, \ldots, x_n^0)$, и найдем точку ее минимума. Для этого нужно решить линейное уравнение

$$\frac{\partial}{\partial x_1} F(x_1, x_2^0, \ldots, x_n^0) = \\ = 2[a_{11}x_1 + a_{12}x_2^0 + \ldots + a_{1n}x_n^0 - b_1] = 0.$$

Оно, по существу, совпадает с первым уравнением системы (7.9), которая рассматривалась во втором видоизменении метода Зейделя для $k = 0$.

Обозначим найденное x_1 через x_1^1, рассмотрим функцию от x_2: $F(x_1^1, x_2, x_3^0, \ldots, x_n^0)$, и найдем точку ее минимума. Уравнение для ее нахождения

$$\frac{\partial}{\partial x_2} F(x_1^1, x_2, x_3^0, \ldots, x_n^0) = \\ = 2[a_{21}x_1^1 + a_{22}x_2 + a_{23}x_3^0 + \ldots + a_{2n}x_n^0 - b_2] = 0$$

совпадает, по существу, со вторым уравнением системы (7.9). Продолжим этот процесс до нахождения x_n^1. После этого повторяем цикл вычислений, отправляясь от значений $(x_1^1, x_2^1, \ldots, x_n^1)$.

Геометрическое значение изложенного метода поясним на случае $n = 2$ и системы двух уравнений. Здесь

$$F(x_1, x_2) = a_{11}(x_1)^2 + 2a_{12}x_1x_2 + a_{22}(x_2)^2 - 2(b_1x_1 + b_2x_2).$$

Линиями уровня $F = C$ для него будут подобные эллипсы с общим центром и совпадающими направлениями осей. Они изображены на рис. 5. При начале

вычислений мы исходим из точки x^0 и движемся параллельно оси x_1 до точки, в которой значение многочлена F становится наименьшим: $F = \min = m_1$. Это случится, очевидно, в том месте прямой $x_2 = x_2^0$, где она коснется линии уровня $F = m_1$. После этого перемещаемся параллельно оси x_2 до места касания прямой $x_1 = x_1^1$ с более низкой линией уровня и т. д.

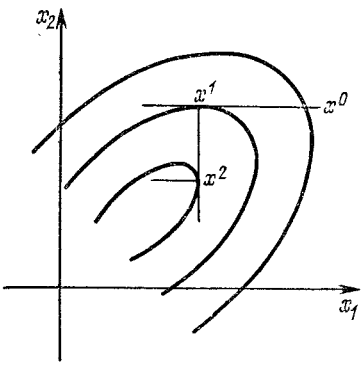

Рис. 5.

3. Понятие о методе градиентного спуска. Этот метод имеет общее значение и может быть применен для нахождения точки минимума любого функционала, для которого имеет смысл понятие градиента. Мы рассмотрим его простую форму, когда решается задача о минимуме функции нескольких аргументов. Пусть в некоторой области n-мерного векторного пространства $x = (x_1, \ldots, x_n)$ дана произвольная функция $F(x)$ и нужно найти точку ее минимума. Предположим, что известно приближенное положение $x^0 = (x_1^0, \ldots, x_n^0)$ этой точки. Если мы хотим возможно быстро, отправляясь из x^0, дойти до места минимума, то разумно выйти из x^0 в направлении, противоположном градиенту F в точке x^0. Путь, по которому мы будем перемещаться, имеет уравнение $x = x^0 - t \operatorname{grad} F(x^0)$ $(t \geqslant 0)$. При этом мы должны наблюдать за изменением F при нашем перемещении, т. е. следить за изменением функции одного аргумента

$$\varphi(t) = F(x^0 - t \operatorname{grad} F(x^0))$$

и остановиться там, где $\varphi(t)$ достигнет своего наименьшего значения. Соответствующее значение t должно быть решением уравнения $\varphi'(t) = 0$.

Для функции $F(x) = (Ax, x) - 2(x, b)$, сопряженной с системой (1), уравнение $\varphi'(t) = 0$ находится просто. Здесь $\operatorname{grad} F(x) = 2(Ax - b)$ и $\varphi(t) = (Ax, x) - 2(x, b)$,

где $x = x^0 - 2t(Ax^0 - b)$. Кроме того, ввиду симметричности A

$$\frac{d}{dt}(Ax, x) = \left(A\frac{dx}{dt}, x\right) + \left(Ax, \frac{dx}{dt}\right) = 2\left(\frac{dx}{dt}, Ax\right) =$$
$$= 2(-2[Ax^0 - b], A\{x^0 - 2t[Ax^0 - b]\}) =$$
$$= 8t(Ax^0 - b, A[Ax^0 - b]) - 4(Ax^0 - b, Ax^0),$$
$$\frac{d}{dt}(x, b) = \left(\frac{dx}{dt}, b\right) = (-2[Ax^0 - b], b) = -2(Ax^0 - b, b),$$
$$\varphi'(t) = \frac{d}{dt}(Ax, x) - 2\frac{d}{dt}(x, b) =$$
$$= 8t(Ax^0 - b, A[Ax^0 - b]) - 4(Ax^0 - b, Ax^0 - b) = 0$$

Отсюда получаем значение t:

$$t = \frac{(Ax^0 - b, Ax^0 - b)}{2(Ax^0 - b, A[Ax^0 - b])}.$$

Его мы должны подставить в выражение для x, и результат принять за первое улучшенное значение x_1 для

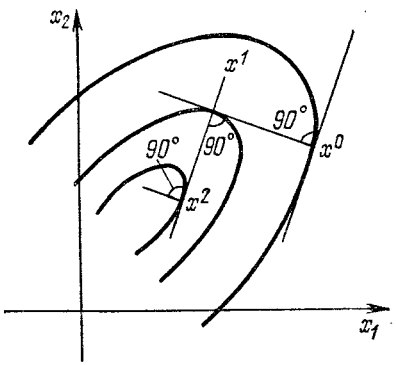

Рис. 6.

точки минимума. После этого, отправляясь от x^1, выполняем вычисления так же, как для x^0, и т. д.

Метод градиентного спуска часто называют *методом наискорейшего спуска*. Расположение приближений и перемещений для случая двух переменных изображено на рис. 6.

§ 9. Оценка погрешности приближенного решения и мера обусловленности

1. Оценка погрешности решения системы. Пусть рассматривается система

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

и X есть ее точное решение. В действительности решается система с измененными матрицей и свободным вектором

$$A_1 x = b_1, \quad A_1 = A + \delta, \quad b_1 = b + \eta. \tag{2}$$

Обозначим ее решение X_1 и рассмотрим погрешность решения $X_1 - X = r$. Нашей ближайшей задачей будет получение оценки r в зависимости от оценок δ и η. Построим уравнение для r; с этой целью подставим в (2) вместо A_1, b_1 и X_1 их выражения через A, b, X:

$$(A + \delta)(X + r) = b + \eta.$$

Вычитая отсюда почленно равенство $AX = b$, получим

$$Ar + \delta X + \delta r = Ar + \delta X_1 = \eta,$$

откуда находим

$$Ar = \eta - \delta X_1 \quad \text{и} \quad r = A^{-1}(\eta - \delta X_1).$$

Из этого выражения для r сразу же получается нужная оценка

$$\|r\| \leqslant \|A^{-1}\|(\|\eta\| + \|\delta\|\|X_1\|).$$

Если матрица A системы участвует в вычислениях точно и, следовательно, можно считать $\delta = 0$, то выражение и оценка для r упрощаются:

$$Ar = \eta, \quad r = A^{-1}\eta, \quad \|r\| \leqslant \|A^{-1}\| \cdot \|\eta\|.$$

2. Мера обусловленности системы и матрицы. Для количественной характеристики зависимости погрешности r решения системы от погрешности η свободного вектора вводятся понятия обусловленности системы и обусловленности матрицы системы. Ниже в изложении под *нормой матрицы* понимается норма, подчиненная норме вектора.

Рассмотрим отношения $\frac{\|r\|}{\|X\|}$ и $\frac{\|\eta\|}{\|b\|}$, являющиеся аналогами относительных погрешностей, и определим меру обусловленности системы равенством

$$\mu = \sup_{\eta}\left[\frac{\|r\|}{\|X\|} : \frac{\|\eta\|}{\|b\|}\right] = \frac{\|b\|}{\|X\|}\sup_{\eta}\frac{\|r\|}{\|\eta\|}, \quad \|\eta\| \leqslant 1.$$

По определению μ верна оценка для погрешности решения

$$\frac{\|r\|}{\|X\|} \leqslant \mu \frac{\|\eta\|}{\|b\|}. \tag{3}$$

Значение μ может быть просто вычислено. Так как $r = A^{-1}\eta$, то $\|r\| \leqslant \|A^{-1}\| \cdot \|\eta\|$; при этом существует такой вектор η, что разница между правой и левой частями неравенства будет сколь угодно малой. Поэтому

$$\sup_{\eta}\frac{\|r\|}{\|\eta\|} = \|A^{-1}\|$$

и

$$\mu = \frac{\|b\|}{\|X\|}\|A^{-1}\|. \tag{4}$$

В некоторых случаях предпочитают свойства системы характеризовать только при помощи свойств матрицы A. Для этого вводят величину $\nu(A) = \nu = \sup_{b}\mu$. Тогда неравенство (3) может быть заменено неравенством

$$\frac{\|r\|}{\|X\|} \leqslant \nu(A)\frac{\|\eta\|}{\|b\|},$$

дающим оценку относительной погрешности решения через относительную погрешность свободного вектора с коэффициентом $\nu(A)$, зависящим только от A и называемым *мерой обусловленности* матрицы A.

Ввиду того, что

$$\sup_{b}\frac{\|b\|}{\|X\|} = \sup_{X}\frac{\|AX\|}{\|X\|} = \|A\|,$$

для $\nu(A)$, если принять во внимание (4), получаем значение

$$\nu(A) = \|A\| \cdot \|A^{-1}\|.$$

Рассмотрим, например, третью, или сферическую, норму вектора и предположим матрицу A симметричной. Тогда $\|A\| = \max|\lambda_A|$. Так как собственные значения A и A^{-1} взаимно обратны, то

$$\|A^{-1}\| = \max \frac{1}{|\lambda_A|} = \frac{1}{\min|\lambda_A|}$$

и

$$\nu(A) = \frac{\max|\lambda_A|}{\min|\lambda_A|}.$$

Принято называть системы уравнений и матрицы с большими значениями мер обусловленности μ и ν *плохо обусловленными.* Если же эти меры имеют небольшие значения, то системы и матрицы называют *хорошо обусловленными.*

ЛИТЕРАТУРА

1. Воеводин В. В., Численные методы алгебры (теория и алгорифмы), «Наука», М., 1966.
2. Фаддеев Д. К., Фаддеева В. Н., Вычислительные методы линейной алгебры, изд. 2-е, Физматгиз, М. — Л., 1963.
3. Хаусхолдер А. С., Основы численного анализа, ИЛ, М., 1956.
4. Мак-Кракен Д., Дорн У., Численные методы и программирование на Фортране, «Мир», М., 1969.
5. Форсайт Д., Молер К., Численное решение систем линейных алгебраических уравнений, «Мир», М., 1974.
6. Сборник научных программ на Фортране, вып. 1, 2, «Статистика», М., 1974.

ГЛАВА 3

ВЫЧИСЛЕНИЕ СОБСТВЕННЫХ МНОГОЧЛЕНОВ, ЗНАЧЕНИЙ И ВЕКТОРОВ МАТРИЦ

§ 1. Введение

1. О содержании задачи. Во многих задачах одновременно с матрицей A приходится рассматривать связанное с ней уравнение

$$|A-\lambda E|=\begin{vmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{vmatrix}=0, \qquad (1)$$

которое называют *характеристическим* (или *вековым*) *уравнением* матрицы A. Определитель $|A-\lambda E|$ есть алгебраический многочлен степени n от λ со старшим коэффициентом $(-1)^n$ и его обычно записывают в виде

$$|A-\lambda E|=(-1)^n(\lambda^n-p_1\lambda^{n-1}-\dots-p_n)= \\ =(-1)^n P_n(\lambda). \qquad (2)$$

Многочлен $P_n(\lambda)$ называют *собственным многочленом* матрицы A. Корни его, которые мы будем обозначать λ_1, λ_2, ..., λ_n получили название *собственных* (или *характеристических*) *значений* (чисел) матрицы A. Они характеризуются тем, что однородная система

$$Ax=\lambda x \qquad (3)$$

имеет ненулевое решение в том и только в том случае, когда λ есть собственное значение A. Отвечающие ему

ненулевые решения системы (3) получили название *собственных векторов* матрицы, соответствующих значению λ.

Запись характеристического уравнения с помощью определителя (1) не всегда удобна для вычислений, и собственный многочлен предпочитают приводить к виду

$$P_n(\lambda) = \lambda^n - p_1\lambda^{n-1} - \ldots - p_n, \qquad (4)$$

что можно сделать путем разложения определителя $|A - \lambda E|$ по элементам. Затруднение в таких вычислениях вызвано тем, что диагональные элементы его содержат параметр λ в буквенном виде, и это делает вычисления громоздкими.

Было предложено несколько способов, позволяющих упрощать вычисления коэффициентов собственного многочлена по элементам матрицы A. Некоторые из них будут изложены ниже.

Разыскание собственных значений матрицы A — это задача решения алгебраического уравнения степени n в форме (1) или (4). После того, как собственные значения будут вычислены, соответствующие им собственные векторы могут быть найдены, вообще говоря, как решения однородной системы (3).

При решении многих научных и технических задач необходимо бывает находить все собственные значения матрицы. В этом случае задача называется *полной проблемой собственных значений*. Но существует ряд задач, где полные сведения не являются необходимыми и можно ограничиться меньшим объемом знаний: например, достаточно указать границы, в которых лежат все собственные значения, как это иногда бывает при изучении устойчивости или неустойчивости процессов, или найти собственное значение, близкое к известному числу, когда рассматриваются явления резонанса, и т. д. Все такого рода задачи называются *частичными проблемами собственных значений* и для каждой из них, чтобы избежать излишней затраты труда, создаются свои методы решения. С частью их мы ознакомимся ниже.

Изложение начнем в следующем параграфе с полной проблемы собственных значений и нахождения многочленов (2) или (4).

2. Аннулирующий и минимальный многочлены матрицы. Многочлен

$$P(\lambda) = a_0\lambda^m + a_1\lambda^{m-1} + \ldots + a_m$$

называется *аннулирующим* для матрицы A, если

$$P(A) = a_0A^m + a_1A^{m-1} + \ldots + a_mE = 0.$$

Чтобы исключить из рассмотрения многочлен, тождественно равный нулю, будем считать, что старший коэффициент a_0 в $P(\lambda)$ равен единице.

Если $P_n(\lambda)$ есть собственный многочлен (4) для матрицы A, то по известной в алгебре теореме Гамильтона — Кели $P_n(A) = 0$, и собственный многочлен является одним из аннулирующих многочленов A.

Многочлен $P_n(\lambda)$ имеет степень n, но в некоторых случаях может оказаться, что существует многочлен $\psi(\lambda)$ степени меньшей n, аннулирующий для матрицы A. Такой многочлен наименьшей степени называют *минимальным многочленом* матрицы A. Укажем на некоторые свойства его.

1) Всякий аннулирующий многочлен нацело делится на минимальный многочлен. В самом деле, пусть $P(\lambda)$ есть такой многочлен. Если разделить его на $\psi(\lambda)$, то он может быть представлен в форме $P(\lambda) = \psi(\lambda)Q(\lambda) + r(\lambda)$, где $r(\lambda)$ имеет степень, меньшую чем $\psi(\lambda)$. Так как $P(A) = 0$ и $\psi(A) = 0$, то отсюда следует, что $r(A) = 0$, что возможно только в случае $r(\lambda) \equiv 0$, и $P(\lambda)$, следовательно, нацело делится на $\psi(\lambda)$.

2) Отсюда, в частности, вытекает, что собственный многочлен нацело делится на минимальный многочлен. Поэтому корни минимального многочлена являются собственными числами матрицы A.

3) Минимальный многочлен матрицы — единственный. Если бы существовало два минимальных многочлена $\psi_1(\lambda)$ и $\psi_2(\lambda)$, то разность между ними $r(\lambda) = \psi_1(\lambda) - \psi_2(\lambda)$ была бы аннулирующим многочленом для A, степень которого меньше, чем у ψ_1 и ψ_2, что противоречит их минимальности.

Нам потребуется еще одно понятие. Пусть C есть произвольный ненулевой вектор. Рассмотрим всевозмож-

ные многочлены $g(\lambda)$ со старшим коэффициентом, приведенным к единице, такие, что $g(A)C = 0$. К этому множеству принадлежат, в частности, все аннулирующие многочлены для матрицы A, но могут принадлежать также многочлены $g(\lambda)$, для которых $g(A) \neq 0$.

Среди взятых многочленов выберем тот, который имеет наименьшую степень. Его называют *минимальным аннулирующим вектор C многочленом* матрицы A. Обозначим его $\varphi(\lambda)$. Он обладает свойствами, аналогичными свойствам минимального многочлена $\psi(\lambda)$:

1) если $g(A)C = 0$, то $g(\lambda)$ нацело делится на $\varphi(\lambda)$;

2) все корни многочлена $\varphi(\lambda)$ являются собственными значениями матрицы A;

3) минимальный аннулирующий вектор C многочлен матрицы A — единственный.

§ 2. Метод, основанный на подобном преобразовании матрицы

1. Нахождение собственного многочлена. Известно, что подобное преобразование матрицы A не изменяет ее собственного многочлена. В самом деле, если $B = S^{-1}AS$, то

$$|B - \lambda E| = |S^{-1}AS - \lambda S^{-1}ES| =$$
$$= |S^{-1}| \cdot |A - \lambda E| \cdot |S| = |A - \lambda E|.$$

Естественно пытаться подобно преобразовать матрицу A к такой форме, в которой в явном виде присутствуют коэффициенты $p_1, p_2, \ldots, p_n$ собственного многочлена. Для этого удобна *каноническая форма Фробениуса*:

$$\Phi = \begin{bmatrix} p_1 & p_2 & p_3 & \cdots & p_{n-1} & p_n \\ 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 & 0 \end{bmatrix}.$$

То, что элементы первой строки действительно являются коэффициентами собственного многочлена матрицы Φ,

можно просто проверить при помощи разложений по элементам столбцов:

$$|\Phi-\lambda E|=\begin{vmatrix} p_1-\lambda & p_2 & p_3 & \dots & p_{n-1} & p_n \\ 1 & -\lambda & 0 & \dots & 0 & 0 \\ 0 & 1 & -\lambda & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & -\lambda \end{vmatrix}=$$

$$=(p_1-\lambda)(-\lambda)^{n-1}-\begin{vmatrix} p_2 & p_3 & \dots & p_{n-1} & p_n \\ 1 & -\lambda & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & -\lambda \end{vmatrix}=$$

$$=(p_1-\lambda)(-\lambda)^{n-1}-p_2(-\lambda)^{n-2}+$$

$$+\begin{vmatrix} p_3 & p_4 & \dots & p_{n-1} & p_n \\ 1 & -\lambda & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & -\lambda \end{vmatrix}=\dots$$

$$\dots=(p_1-\lambda)(-\lambda)^{n-1}-p_2(-\lambda)^{n-2}+$$

$$+p_3(-\lambda)^{n-3}-\dots+(-1)^{n+1}p_n=$$

$$=(-1)^n(\lambda^n-p_1\lambda^{n-1}-p_2\lambda^{n-2}-\dots-p_n)=(-1)^nP_n(\lambda)$$

В преобразовании $\Phi=S^{-1}AS$ матрицу S целесообразно находить путем последовательного приведения строк к каноническому виду *). Начнем с приведения последней строки. Предположим, что элемент a_{nn-1} последней строки отличен от нуля. Разделим на него $(n-1)$-й столбец матрицы A. Вновь полученный $(n-1)$-й столбец умножим на a_{ni} и вычтем из столбца номера i. Проделав это для $i=1, 2, \dots, n-2, n$, приведем последнюю строку к виду Фробениуса. Можно легко проверить, что такое преобразование равносильно умножению A справа на матрицу

$$M_{n-1}=\begin{vmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -\frac{a_{n1}}{a_{nn-1}} & -\frac{a_{n2}}{a_{nn-1}} & \dots & \frac{1}{a_{nn-1}} & -\frac{a_{nn}}{a_{nn-1}} \\ 0 & 0 & \dots & 0 & 1 \end{vmatrix}.$$

*) Данилевский А. М., Матем. сб., 1937, 2, [44] 169—171.

Полученная матрица AM_{n-1} не будет подобна A, и, чтобы сделать ее подобной A, нужно AM_{n-1} слева умножить на M_{n-1}^{-1}. Такая матрица существует, так как $|M_{n-1}| = \frac{1}{a_{nn-1}} \neq 0$, и можно проверить, что

$$M_{n-1}^{-1} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn-1} & a_{nn} \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}.$$

Заметим, что преобразование $M_{n-1}^{-1}(AM_{n-1})$ не изменит последней строки в AM_{n-1}, и она останется в виде Фробениуса. Этим заканчивается первый шаг преобразования, и в результате его получится матрица вида

$$A_1 = M_{n-1}^{-1}AM_{n-1} = \begin{bmatrix} a_{11}^1 & a_{12}^1 & \dots & a_{1n-1}^1 & a_{1n}^1 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n-11}^1 & a_{n-12}^1 & \dots & a_{n-1n-1}^1 & a_{n-1n}^1 \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}.$$

Второй шаг преобразований аналогичен первому и состоит в приведении предпоследней строки матрицы A_1 к виду Фробениуса при условии неизменности последней ее строки. Предположим, что элемент a_{n-2n-1}^1 в A_1 отличен от нуля. Нужное преобразование можно записать в форме

$$A_2 = M_{n-2}^{-1}A_1M_{n-2} = M_{n-2}^{-1}M_{n-1}^{-1}AM_{n-1}M_{n-2} =$$

$$= \begin{bmatrix} a_{11}^2 & \dots & a_{1n-2}^2 & a_{1n-1}^2 & a_{1n}^2 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n-21}^2 & \dots & a_{n-2n-2}^2 & a_{n-2n-1}^2 & a_{n-2n}^2 \\ 0 & \dots & 1 & 0 & 0 \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix}.$$

Здесь

$$M_{n-2} =$$

$$= \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -\dfrac{a^1_{n-11}}{a^1_{n-1n-2}} & \dots & -\dfrac{a^1_{n-1n-3}}{a^1_{n-1n-2}} & \dfrac{1}{a^1_{n-1n-2}} & -\dfrac{a^1_{n-1n-1}}{a^1_{n-1n-2}} & -\dfrac{a^1_{n-1n}}{a^1_{n-1n-2}} \\ 0 & \dots & 0 & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$M^{-1}_{n-2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ a^1_{n-11} & a^1_{n-12} & \dots & a^1_{n-1n-1} & a^1_{n-1n} \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}.$$

Правило построения матриц M_{n-2} и M^{-1}_{n-2} по A_1 аналогично правилу построения M_{n-1} и M^{-1}_{n-1} по матрице A. Эта аналогия правила сохраняется на всех следующих шагах.

Таким образом, в регулярном случае, когда $a_{nn-1} \neq 0$, $a^1_{n-1n-2} \neq 0, \dots, a^{n-1}_{21} \neq 0$, после выполнения $n-1$ шагов преобразований матрица A будет приведена к каноническому виду Фробениуса

$$A_{n-1} = M_1^{-1}M_2^{-1} \dots M^{-1}_{n-1}AM_{n-1}M_{n-2} \dots M_1 = S^{-1}AS =$$

$$= \begin{bmatrix} a^{n-1}_{11} & a^{n-1}_{12} & \dots & a^{n-1}_{1n-1} & a^{n-1}_{1n} \\ 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} = \begin{bmatrix} p_1 & p_2 & \dots & p_{n-1} & p_n \\ 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} = \Phi.$$

По первой строке полученной матрицы Φ составляется собственный многочлен A: $P_n(\lambda) = \lambda^n - p_1\lambda^{n-1} - \dots - p_n$.

Обратимся теперь к нерегулярному случаю. Предположим, что процесс приведения A к виду Фробениуса доведен до строки номера k и выполнено, следовательно, $n-k$ шагов преобразований; при этом оказалось, что

$a_{kk-1}^{n-k} = 0$. Дальнейшие преобразования зависят от того, существует ли в строке номера k слева от места $(k, k-1)$ отличный от нуля элемент, или такого элемента нет.

Допустим сначала, что $a_{ki} \neq 0$ $(i < k-1)$. В этом случае продолжение преобразований может быть приведено к регулярному случаю. Для этого достаточно поменять местами столбцы с номерами i и $k-1$ и строки с теми же номерами. Как легко видеть, такое преобразование может быть записано в виде

$$TA_{n-k}T, \tag{1}$$

$$T = \begin{bmatrix}
1 & 0 & & & & & & & & & & & & & & \\
0 & 1 & & & & & & & & & & & & & & \\
 & & \ddots & & & & & & & & & & & & & \\
 & & & \ddots & & & & & & & & & & & & \\
 & & & & 1 & & & & & & & & & & & \\
 & & & & & 0 & \cdots & \cdots & \cdots & 1 & \cdots & \cdots & \cdots & \cdots & \cdots & (i) \\
 & & & & & & 1 & & & & & & & & & \\
 & & & & & & & \ddots & & & & & & & & \\
 & & & & & & & & 1 & & & & & & & \\
 & & & & & 1 & \cdots & \cdots & \cdots & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & (k-1) \\
 & & & & & & & & & & 1 & & & & & \\
 & & & & & & & & & & & \ddots & & & & \\
 & & & & & & & & & & & & & 1 & 0 & \\
 & & & & & & & & & & & & & 0 & 1 &
\end{bmatrix}$$

Можно просто проверить, что (1) есть преобразование подобия. В самом деле, после двукратной перестановки строк и столбцов получится исходная матрица A_{n-k}, поэтому $(T)^2 = E$, $(T)^{-1} = T$, и (1) есть действительно преобразование подобия.

После преобразования (1) следующий шаг может быть выполнен так же, как в регулярном случае.

Теперь рассмотрим случай, когда все элементы строки номера k, предшествующие a_{kk-1}^{n-k}, равны нулю: $a_{ki}^{n-k} =$

$= 0$, $i = 1, \ldots, k-2$. Матрица A_{n-k} имеет форму

$$A_{n-k} = \left[\begin{array}{ccc|ccc} a_{11}^{n-k} & \ldots & a_{1k-1}^{n-k} & a_{1k}^{n-k} & \ldots & a_{1n-1}^{n-k} \quad a_{1n}^{n-k} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{k-11}^{n-k} & \ldots & a_{k-1k-1}^{n-k} & a_{k-1k}^{n-k} & \ldots & a_{k-1n-1}^{n-k} \quad a_{k-1n}^{n-k} \\ \hline 0 & \ldots & 0 & a_{kk}^{n-k} & \ldots & a_{kn-1}^{n-k} \quad a_{kn}^{n-k} \\ 0 & \ldots & 0 & 1 & \ldots & 0 \quad\quad 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & 0 & \ldots & 1 \quad\quad 0 \end{array}\right] = \left[\begin{array}{cc} B_{n-k} & C_{n-k} \\ 0 & \Phi_{n-k} \end{array}\right],$$

где матрицы B_{n-k}, C_{n-k}, Φ_{n-k}, 0 в A_{n-k} отмечены разделительными прямыми. В этом случае, ввиду теоремы Лапласа *) о разложении определителя, имеет место равенство, в правой части которого индексами снизу обозначены порядки единичных матриц:

$$|A_{n-k} - \lambda E| = |B_{n-k} - \lambda E_{k-1}| \cdot |\Phi_{n-k} - \lambda E_{n-k+1}|.$$

Так как Φ_{n-k} имеет канонический вид Фробениуса, ее собственный многочлен выписывается по элементам первой строки. Остается привести к каноническому виду Фробениуса только матрицу B_{n-k}, имеющую порядок $k-1 < n$, и задача преобразований упрощается.

Можно подсчитать, что при изложенных выше преобразованиях в регулярном случае необходимо бывает выполнить приблизительно n^3 операций умножения и деления. Как будет вытекать из последующего изложения, рассматриваемый метод по числу операций является одним из наиболее экономных.

Сделаем еще замечание о практической стороне вычислений. Чтобы избежать операций с числами, сильно различающимися по своим порядкам, стараются перед каждым шагом преобразований при помощи перестановок столбцов и строк или равносильных им преобразований вида (1) на места элементов a_{nn-1}, a_{n-1n-2}^{1}, $\ldots$, a_{kk-1}^{n-k} ставить наибольшие из элементов, находящихся левее и выше их.

*) См., например, А. Г. Курош, Курс высшей алгебры, М., 1962, гл. I, § 6.

2. Вычисление собственных векторов. Когда известны собственные значения λ_i $(i = 1, \ldots, n)$ матрицы A, то ее собственные векторы могут быть найдены путем решения однородных систем $Ax = \lambda_i x$ $(i = 1, \ldots, n)$. Но если построена матрица S, с помощью которой A приводится к виду Фробениуса $\Phi = S^{-1}AS$, то нахождение собственных векторов значительно упрощается. Такое упрощение основано на следующих фактах.

В начале предыдущего пункта отмечалось, что если матрицы A и B подобны: $B = S^{-1}AS$, то собственные многочлены их, а стало быть и собственные числа, совпадают. Что же касается собственных векторов A и B, то связь между ними устанавливается следующим утверждением: если $y = (y_1, \ldots, y_n)$ есть собственный вектор матрицы B, соответствующий собственному значению λ, то $x = (x_1, \ldots, x_n) = Sy$ есть собственный вектор матрицы A, отвечающий тому же значению λ.

Действительно, пусть $By = \lambda y$; тогда $S^{-1}ASy = \lambda y$; после умножения слева на S получается равенство $A(Sy) = \lambda Sy$, говорящее о справедливости высказанного утверждения. Оно позволяет находить собственные векторы A, если известны y, S и λ_i.

Возвратимся к соотношению $\Phi = S^{-1}AS$. Собственные значения λ_i будем считать известными. Собственные векторы Φ находятся просто. В самом деле, система уравнений

$$\Phi y = \lambda_i y,$$

если записать ее в составляющих вектора y, имеет вид

$$\begin{aligned} p_1y_1 + p_2y_2 + \ldots + p_ny_n &= \lambda_i y_1, \\ y_1 &= \lambda_i y_2, \\ \ldots\ldots\ldots & \\ y_{n-1} &= \lambda_i y_n. \end{aligned}$$

Так как собственный вектор y находится заведомо лишь с точностью до численного множителя, можно положить $y_n = 1$. Тогда все составляющие вектора y найдутся последовательно, начиная с последнего уравнения системы до первого, и для y получим

$$y = (\lambda_i^{n-1}, \lambda_i^{n-2}, \ldots, \lambda_i, 1).$$

Что же касается первого уравнения, то оно приведется к равенству

$$p_1\lambda_i^{n-1} + p_2\lambda_i^{n-2} + \ldots + p_n = \lambda_i^n$$

и будет удовлетворяться, так как λ_i есть корень собственного многочлена $P_n(\lambda)$ матрицы A.

Напомним, что матрица S находится в регулярном случае изложенного метода и в нерегулярном случае, приводящемся к регулярному. Например, в регулярном случае для нее получается представление вида

$$S = M_{n-1}M_{n-2} \ldots M_1.$$

Умножая на S собственный вектор y, получим для x выражение

$$x = Sy = M_{n-1}M_{n-2} \ldots M_1 y.$$

Каждая матрица M_i $(i = 1, \ldots, n-1)$ отличается от единичной матрицы только одной строкой. При умножении вектора на M_i будет изменяться только одна составляющая вектора. Поэтому в полученном выражении вектора x удобно y умножать на M_1, M_2, ... последовательно.

§ 3. Применение минимального многочлена матрицы, аннулирующего заданный вектор *)

1. Построение многочлена. Возьмем произвольный ненулевой вектор **) C^0 размерности n и по нему составим рекуррентную последовательность векторов $C^1 = AC^0$, $C^2 = AC^1 = A^2C^0$, ... Вычисления будем продолжать до того места, когда впервые получим вектор, являющийся линейной комбинацией предшествующих векторов. Пусть это будет вектор C^m, и пусть

$$C^m = q_mC^0 + q_{m-1}C^1 + \ldots + q_1C^{m-1}. \qquad (1)$$

Среди векторов размерности n может быть лишь n линейно независимых, поэтому $m \leqslant n$.

*) Идея метода была высказана А. Н. Крыловым в статье «О численном решении уравнения, ...», ИАН ОМЕН, 1931, 4, 491—539.

**) Верхний индекс у вектора C есть порядковый номер, у матрицы A — показатель степени.

Чтобы определить число m и вычислить коэффициенты q_i $(i = 1, \ldots, m)$, обычно записывают равенство (1) для наибольшего возможного m, полагая $m = n$:

$$C^n = q_n C^0 + q_{n-1} C^1 + \ldots + q_1 C^{n-1}. \tag{2}$$

Рассмотрим это равенство в составляющих векторов, положив $C^i = (c_1^i, c_2^i, \ldots, c_n^i)$:

$$\begin{aligned} c_1^0 q_n + c_1^1 q_{n-1} + \ldots + c_1^{n-1} q_1 &= c_1^n, \\ c_2^0 q_n + c_2^1 q_{n-1} + \ldots + c_2^{n-1} q_1 &= c_2^n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ c_n^0 q_n + c_n^1 q_{n-1} + \ldots + c_n^{n-1} q_1 &= c_n^n. \end{aligned} \tag{3}$$

Эти равенства образуют неоднородную систему уравнений для нахождения q_i $(i = 1, \ldots, n)$. Определитель системы зависит от матрицы A, которая в вычислениях считается фиксированной, и от начального вектора C^0, который можно выбирать произвольно. В некоторых случаях оказывается, что определитель системы близок к нулю, и численное решение системы является затруднительным. Тогда можно пытаться улучшить свойства системы, изменяя C^0. По построению системы ясно, что она всегда является разрешимой. Для определения ранга ее обычно приводят к треугольному виду при помощи метода Гаусса.

Предположим, что выполнимы все n шагов прямого хода метода Гаусса, и система приведена к следующей форме:

$$\begin{aligned} q_n + b_{12} q_{n-1} + b_{13} q_{n-2} + \ldots + b_{1n} q_1 &= g_1, \\ q_{n-1} + b_{23} q_{n-2} + \ldots + b_{2n} q_1 &= g_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ q_1 &= g_n. \end{aligned} \tag{4}$$

Это говорит о том, что определитель системы

$$\Delta = \begin{vmatrix} c_1^0 & \ldots & c_1^{n-1} \\ \ldots & \ldots & \ldots \\ c_n^0 & \ldots & c_n^{n-1} \end{vmatrix}$$

отличен от нуля, и векторы $C^0, \ldots, C^{n-1}$ являются линейно независимыми.

Коэффициенты $q_1, \ldots, q_n$ найдутся единственным образом из системы. Покажем, что они совпадут с соответствующими коэффициентами собственного многочлена $P_n(\lambda) = \lambda^n - p_1\lambda^{n-1} - \ldots - p_n$, т. е. что $p_i = q_i$ $(i = 1, 2, \ldots, n)$.

Подставив в (2) вместо C^i их значения $C^i = A^iC^0$, получим равенство

$$(A^n - q_1A^{n-1} - \ldots - q_nE)C^0 = 0. \tag{5}$$

С другой стороны, ввиду соотношения Кели

$$P_n(A)C^0 = (A^n - p_1A^{n-1} - \ldots - p_nE)C^0 = 0.$$

Вычитая это равенство почленно из (5), найдем

$$[(p_1 - q_1)A^{n-1} + \ldots + (p_n - q_n)E]C^0 =$$
$$= (p_1 - q_1)C^{n-1} + (p_2 - q_2)C^{n-2} + \ldots + (p_n - q_n)C^0 = 0.$$

Ввиду линейной независимости векторов C^i $(i = 0, 1, \ldots, n-1)$ последнее возможно только при $p_i = q_i$ $(i = 1, \ldots, n)$.

Перейдем теперь к случаю, когда в прямом ходе метода Гаусса для системы (3) возможно $m < n$ шагов преобразований. После их выполнения, если принять во внимание, что система имеет решение, она будет приведена к виду

$$\begin{aligned}
&q_n + b_{12}q_{n-1} + b_{13}q_{n-2} + \ldots + b_{1n}q_1 = g_1,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&q_{n-m+1} + b_{mm+1}q_{n-m} + \ldots + b_{mn}q_1 = g_m,\\
&0 = 0,\\
&\cdots\\
&0 = 0.
\end{aligned}$$

Это говорит о том, что ранг системы (3) равен m и, стало быть, среди векторов C^i $(i = 0, 1, \ldots, n-1)$ есть лишь m линейно независимых. Ими могут быть только векторы $C^0, \ldots, C^{m-1}$, так как из того, что вектор C^i линейно выражается через $C^0, \ldots, C^{i-1}$, следует, очевидно, что векторы C^{i+1} $C^{i+2}, \ldots$ также будут линейно

выражаться через $C^0, \ldots, C^{i-1}$. Поэтому должно выполняться равенство (1).

Если вновь воспользоваться тем, что $C^i = A^i C^0$, то (1) можно переписать в форме

$$(A^m - q_1 A^{m-1} - q_2 A^{m-2} - \ldots - q_m E) C^0 = \varphi(\lambda) C^0 = 0.$$

Поэтому $\varphi(\lambda) = \lambda^m - q_1\lambda^{m-1} - \ldots - q_m$ есть многочлен матрицы A, аннулирующий вектор C^0. Коэффициенты его $q_1, \ldots, q_m$ должны быть найдены из системы уравнений, получающейся, если соотношение (1) записать в составляющих векторов C^i:

$$\begin{gathered}
c_1^0 q_m + c_1^1 q_{m-1} + \ldots + c_1^{m-1} q_1 = c_1^m, \\
c_2^0 q_m + c_2^1 q_{m-1} + \ldots + c_2^{m-1} q_1 = c_2^m, \\
\cdots\cdots\cdots\cdots\cdots\cdots \\
c_n^0 q_m + c_n^1 q_{m-1} + \ldots + c_n^{m-1} q_1 = c_n^m.
\end{gathered}$$

Такая система имеет ранг m и преобразованиями прямого хода метода Гаусса приводится к треугольному виду

$$\begin{aligned}
q_m + b_{12} q_{m-1} + b_{13} q_{m-2} + \ldots + b_{1m} q_1 &= g_1, \\
q_{m-1} + b_{23} q_{m-2} + \ldots + b_{2m} q_1 &= g_2, \\
\cdots\cdots\cdots\cdots\cdots\cdots \\
q_1 &= g_m.
\end{aligned}$$

После нахождения q_i ($i = 1, \ldots, m$) построим многочлен $\varphi(\lambda)$. Покажем, что он является минимальным аннулирующим вектор C^0. Действительно, если бы существовал многочлен $\psi(\lambda)$, удовлетворяющий условию $\psi(A) C^0 = 0$ и имеющий степень меньше m, то это противоречило бы линейной независимости векторов $C^0, \ldots, C^{m-1}$.

Корни многочлена $\varphi(\lambda)$ являются собственными значениями матрицы A, и, решая уравнение $\varphi(\lambda) = 0$, мы найдем либо все собственные значения A, либо их часть. В последнем случае для нахождения недостающих собственных значений необходимо изменить начальный вектор C^0.

2. Нахождение собственных векторов. Напомним, что собственный вектор, соответствующий собственному значению λ_i, может быть найден как решение однородной

системы $Ax = \lambda_i x$. В рассматриваемом методе задача разыскания собственных векторов в частных случаях может быть упрощена путем использования промежуточных результатов вычислений.

Пусть λ_i есть корень минимального многочлена $\varphi(\lambda) = \lambda^m - q_1\lambda^{m-1} - \ldots - q_m$, аннулирующего вектор C^0. Собственный вектор x^i, принадлежащий λ_i, будем находить в форме линейной комбинации построенных выше независимых векторов $C^0, \ldots, C^{m-1}$:

$$x^i = \beta_{i1}C^{m-1} + \beta_{i2}C^{m-2} + \ldots + \beta_{im}C^0. \qquad (6)$$

Отметим, что такой выбор представления собственного вектора x^i является ограничением, и мы не сможем при помощи (6) получить собственные векторы A, отвечающие λ_i, которые не принадлежат линейному подпространству $C^0, \ldots, C^{m-1}$.

Коэффициенты β_{ij} $(j = 1, \ldots, m)$ нужно выбрать так, чтобы было $Ax^i = \lambda_i x^i$. Если подставить сюда вместо x^i его представление (6) и воспользоваться тем, что $AC^i = C^{i+1}$, то получим

$$\beta_{i1}C^m + \beta_{i2}C^{m-1} + \ldots + \beta_{im}C^1 = \\ = \lambda_i(\beta_{i1}C^{m-1} + \beta_{i2}C^{m-2} + \ldots + \beta_{im}C^0).$$

Исключим отсюда C^m при помощи его выражения (1) и соберем все члены налево:

$$(q_1\beta_{i1} + \beta_{i2} - \lambda_i\beta_{i1})C^{m-1} + (q_2\beta_{i1} + \beta_{i3} - \lambda_i\beta_{i2})C^{m-2} + \ldots \\ \ldots + (q_{m-1}\beta_{i1} + \beta_{im} - \lambda_i\beta_{im-1})C^1 + (q_m\beta_{i1} - \lambda_i\beta_{im})C^0 = 0.$$

Ввиду линейной независимости векторов $C^0, \ldots, C^{m-1}$ такое равенство возможно только в том случае, когда все его коэффициенты обращаются в нуль, что приводит к следующей системе уравнений для β_{ij}:

$$\begin{aligned} q_1\beta_{i1} + \beta_{i2} - \lambda_i\beta_{i1} &= 0, \\ q_2\beta_{i1} + \beta_{i3} - \lambda_i\beta_{i2} &= 0, \\ \cdots\cdots\cdots\cdots\cdots & \\ q_{m-1}\beta_{i1} + \beta_{im} - \lambda_i\beta_{im-1} &= 0, \\ q_m\beta_{i1} - \lambda_i\beta_{im} &= 0. \end{aligned}$$

Из нее последовательно находим

$$\beta_{i2} = (\lambda_i - q_1)\beta_{i1},$$
$$\beta_{i3} = (\lambda_i^2 - q_1\lambda_i - q_2)\beta_{i1},$$
$$\cdots\cdots\cdots\cdots$$
$$\beta_{im} = (\lambda_i^{m-1} - q_1\lambda_i^{m-2} - q_2\lambda_i^{m-3} - \ldots - q_{m-1})\beta_{i1},$$
$$(\lambda_i^m - q_1\lambda_i^{m-1} - q_2\lambda_i^{m-2} - \ldots - q_m)\beta_{i1} = \varphi(\lambda_i)\beta_{i1} = 0.$$

Последнее из равенств является тождеством, так как $\varphi(\lambda_i) = 0$. Коэффициент β_{i1} остается произвольным.

Здесь λ_i есть корень минимального многочлена $\varphi(\lambda)$. Проделав указанные вычисления для всех корней $\varphi(\lambda)$, мы получим несколько собственных векторов, которые могут не образовать полную систему независимых собственных векторов A. Недостающие векторы, как и при вычислении собственных значений, можно искать, изменяя начальный вектор C^0.

§ 4. Два видоизменения правила применения минимального многочлена

Правило вычисления собственного или минимального многочленов посредством простой итерации исходного вектора C^0 при помощи умножения на заданную матрицу A требует решения системы (3.3), и если последняя плохо обусловлена, то это является трудной или даже практически невозможной задачей. Чтобы избежать этого затруднения, предлагались видоизменения правила; о двух из них, сходных по идее, будет сказано ниже.

1. Метод ортогонализации. По произвольно выбранному вектору $C^0 \neq 0$ построим вектор

$$C^1 = AC^0 - g_{10}C^0,$$

при этом g_{10} выберем так, чтобы C^1 был ортогонален C^0. Условие ортогональности $(C^1, C^0) = 0$ дает для g_{10} значение

$$g_{10} = \frac{(AC^0, C^0)}{(C^0, C^0)}.$$

Если $C^1 = 0$, то многочлен

$$\varphi_1(\lambda) = \lambda - g_{10}$$

является минимальным многочленом матрицы A, аннулирующим C^0. Если же $C^1 \neq 0$, то вычисляем AC^1 и по нему образуем вектор

$$C^2 = AC^1 - g_{21}C^1 - g_{20}C^0;$$

при этом коэффициенты g_{21} и g_{20} выбираем так, чтобы C^2 был ортогонален к C^0 и C^1. Условия ортогональности $(C^2, C^1) = 0$ и $(C^1, C^0) = 0$ дадут значения

$$g_{21} = \frac{(AC^1, C^1)}{(C^1, C^1)}, \quad g_{20} = \frac{(AC^0, C^0)}{(C^0, C^0)}.$$

В случае когда $C^2 = 0$, равенство

$$(A - g_{21})C^1 - g_{20}C^0 = (A - g_{21})(A - g_{10})C^0 - g_{20}C^0 =$$
$$= A^2C^0 - (g_{10} + g_{21})AC^0 - (g_{20} - g_{21}g_{10})C^0 = 0$$

дает нулевую линейную комбинацию векторов C^0, AC^0, A^2C^0. Поэтому многочлен

$$\varphi_2(\lambda) = (\lambda - g_{21})(\lambda - g_{10}) - g_{20} = (\lambda - g_{21})\varphi_1(\lambda) - g_{20}$$

будет минимальным многочленом матрицы A, аннулирующим C_0. Когда же C^2 есть ненулевой вектор, процесс ортогонализации продолжается.

Предположим, что построены ненулевые, попарно ортогональные векторы $C^1, C^2, \ldots, C^{m-1}$. По ним строится вектор

$$C^m = AC^{m-1} - g_{mm-1}C^{m-1} - \ldots - g_{m0}C^0, \qquad (1)$$

и коэффициенты g_{mi} выбираются так, чтобы он был ортогонален к $C^0, \ldots, C^{m-1}$, т. е. чтобы выполнялись условия $(C^m, C^i) = 0$ $(i = 0, 1, \ldots, m-1)$. Для g_{mi} получаются значения

$$g_{mi} = \frac{(AC^i, C^i)}{(C^i, C^i)} \quad (i = 0, 1, \ldots, m-1). \qquad (2)$$

Параллельно с векторами C^i по коэффициентам g_{ij} составляются многочлены

$$\begin{aligned}
&\varphi_0(\lambda) = 1,\\
&\varphi_1(\lambda) = (\lambda - g_{10})\varphi_0(\lambda),\\
&\varphi_2(\lambda) = (\lambda - g_{21})\varphi_1(\lambda) - g_{20}\varphi_0(\lambda),\\
&\varphi_3(\lambda) = (\lambda - g_{32})\varphi_2(\lambda) - g_{31}\varphi_1(\lambda) - g_{30}\varphi_0(\lambda),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\varphi_m(\lambda) = (\lambda - g_{mm-1})\varphi_{m-1}(\lambda) -\\
&\qquad - g_{mm-2}\varphi_{m-2}(\lambda) - \ldots - g_{m0}\varphi_0(\lambda).
\end{aligned} \tag{3}$$

Так как в n-мерном пространстве существует лишь n попарно ортогональных векторов, то при некотором $m \leqslant n$ получится нулевой вектор C^m. Равенство

$$\begin{aligned}
0 = AC^{m-1} - g_{mm-1}C^{m-1} - g_{mm-2}C^{m-2} - \ldots - g_{m0}C^0 =\\
= (A - g_{mm-1})\varphi_{m-1}(A)C^0 - g_{mm-2}\varphi_{m-2}(A)C^0 - \ldots\\
\ldots - g_{m0}\varphi_0(A)C^0 = \varphi_m(A)C^0
\end{aligned} \tag{4}$$

говорит о том, что векторы C^0, AC^0, ..., A^mC^0 являются линейно зависимыми, и многочлен $\varphi_m(A)$ есть минимальный многочлен матрицы A, аннулирующий C^0.

При помощи векторов C^i $(i = 0, 1, \ldots, m-1)$ могут быть найдены собственные векторы матрицы A. Пусть λ_i есть корень многочлена $\varphi_m(\lambda)$. Отвечающий ему собственный вектор x^i будем искать в форме

$$x^i = \beta_{i1}C^{m-1} + \beta_{i2}C^{m-2} + \ldots + \beta_{im}C^0. \tag{5}$$

Коэффициенты β_{ij} выберем так, чтобы выполнялось условие

$$Ax^i = \lambda_i x^i.$$

Подставляя сюда вместо x^i его выражение (5) и учитывая следующее равенство, вытекающее из (1),

$$AC^l = C^{l+1} + g_{l+1l}C^l + \ldots + g_{l+10}C^0,$$

после несложных группировок членов получим

$$(\lambda_i\beta_{i1}-\beta_{i1}g_{mm-1}-\beta_{i2})C^{m-1}+$$
$$+(\lambda_i\beta_{i2}-\beta_{i1}g_{mm-2}-\beta_{i2}g_{m-1m-2}-\beta_{i3})C^{m-2}+\ldots$$
$$\ldots+(\lambda_i\beta_{im-1}-\beta_{i1}g_{m1}-\beta_{i2}g_{m-11}-\ldots-\beta_{im-1}g_{21}-\beta_{im})C^1+$$
$$+(\lambda_i\beta_{im}-\beta_{i1}g_{m0}-\beta_{i2}g_{m-10}-\ldots-\beta_{im}g_{10})C^0=0.$$

Ввиду линейной независимости векторов C^i ($i=0, 1, \ldots, m-1$) последнее равенство возможно только в том случае, когда коэффициенты при этих векторах равны нулю:

$$\begin{aligned}\lambda_i\beta_{i1}-\beta_{i1}g_{mm-1}-\beta_{i2}&=0,\\ \lambda_i\beta_{i2}-\beta_{i1}g_{mm-2}-\beta_{i2}g_{m-1m-2}-\beta_{i3}&=0,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots&\\ \lambda_i\beta_{im-1}-\beta_{i1}g_{m1}-\beta_{i2}g_{m-11}-\ldots-\beta_{im-1}g_{21}-\beta_{im}&=0,\\ \lambda_i\beta_{im}-\beta_{i1}g_{m0}-\beta_{i2}g_{m-10}-\ldots-\beta_{im}g_{10}&=0.\end{aligned} \quad (6)$$

Из полученных уравнений последовательно находятся $\beta_{i2}, \beta_{i3}, \ldots, \beta_{im}$, при этом коэффициент β_{i1} остается произвольным. Это связано с тем, что собственный вектор заведомо определен лишь с точностью до численного множителя.

При нахождении β_{ij} ($j=2, \ldots, m$) последнее уравнение остается неиспользованным, и можно показать, что оно является следствием первых $m-1$ уравнений и равенства $\varphi_m(\lambda_i)=0$.

Если указанным путем будут найдены не все собственные числа и векторы матрицы A, то изменяют начальный вектор C^0 и повторяют вычисления.

2. Метод приведения к нулю составляющих итерированных векторов. Вновь возьмем некоторый вектор $C^0\neq 0$ и построим вектор

$$C^1=AC^0-g_{10}C^0,$$

где g_{10} выберем так, чтобы первая составляющая вектора C^1 была равна нулю. Затем образуем вектор

$$C^2=AC^1-g_{21}C^1-g_{20}C^0,$$

и коэффициенты g_{21}, g_{20} подберем из условия, чтобы у C^2 обратились в нуль две первые составляющие, и т. д. Если

известны векторы C^0, C^1, ..., C^{m-1}, то C^m строится в виде

$$C^m = AC^{m-1} - g_{mm-1}C^{m-1} - g_{mm-2}C^{m-2} - \ldots - g_{m0}C^0, \quad (7)$$

коэффициенты же g_{mj} $(j = 0, \ldots, m-1)$ выбираются так, чтобы первые m составляющих C^m обратились в нуль. В этом процессе не всегда можно выполнить n шагов. Мы остановимся на регулярном случае, когда все n шагов являются выполнимыми, и будем считать

$$c_1^0 \neq 0, \; c_2^1 \neq 0, \; \ldots, \; c_n^{n-1} \neq 0. \quad (8)$$

Векторы C^0, C^1, ..., C^{n-1}, очевидно, линейно независимы, и их составляющие образуют треугольную матрицу с диагональными элементами, отличными от нуля.

Заметим, что правило их вычисления (7) имеет ту же форму, что и правило (1) в методе ортогонализации, но с другим смыслом коэффициентов g_{ij}: в методе ортогонализации они находятся из требования ортогональности C^m к C^i $(i < m)$, тогда как в рассматриваемом — из условия обращения в нуль первых m составляющих вектора C^m.

Можно утверждать, что все формальные следствия, не зависящие от смысла g_{ik}, которые были извлечены в п. 1 из правила (1), будут верны и в методе обращения в нуль составляющих итерированных векторов C^i; например, можно утверждать, что и здесь векторы C^m представимы в форме

$$C^m = \varphi_m(A)\, C^0 \quad (m = 1, 2, \ldots, n-1),$$

где многочлены $\varphi_m(\lambda)$ вычисляются по формулам (5) при значениях g_{ij}, которые определены в настоящем пункте. Многочлен $\varphi_n(\lambda)$, найденный по тому же правилу (5), будет минимальным многочленом матрицы A, аннулирующим C^0. Старший коэффициент его равен единице, он должен нацело делиться на собственный многочлен $P_n(\lambda)$, и поэтому должен совпадать с $P_n(\lambda)$.

Пусть λ_i есть любой корень $\varphi_n(\lambda)$. Соответствующий ему собственный вектор x^i есть решение однородной системы $Ax^i = \lambda_i x^i$; но его можно найти, пользуясь векторами C^i $(i = 0, 1, \ldots, n-1)$, вычисляемыми при по-

строении φ_n. Будем искать x^i в виде линейной комбинации из C^i:

$$x^i = g_{i1}C^{n-1} + g_{i2}C^{n-2} + \ldots + g_{in}C^0. \qquad (9)$$

Численные множители g_{ij} $(j = 1, \ldots, n)$ должны быть выбраны так, чтобы удовлетворялась однородная система для x^i, указанная выше. Подставляя в нее вместо x^i разложение (9) и воспользовавшись соотношением, вытекающим из (7),

$$AC^{m-1} = C^m + g_{mm-1}C^{m-1} + \ldots + g_{m0}C^0 \quad (m < n)$$

и тем, что $C^n = 0$ и $\varphi(\lambda_i) = 0$, придем к системе уравнений для коэффициентов β_{ij} $(j = 1, \ldots, n)$, отличающейся от (6) значениями величин g_{mm-1} и тем, что m заменено на n:

$$\begin{aligned}
\lambda_i\beta_{i1} - \beta_{i1}g_{nn-1} - \beta_{i2} &= 0,\\
\lambda_i\beta_{i2} - \beta_{i1}g_{nn-2} - \beta_{i2}g_{n-1n-2} - \beta_{i3} &= 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
\lambda_i\beta_{in-1} - \beta_{i1}g_{n1} - \beta_{i2}g_{n-11} - \ldots - \beta_{in-1}g_{21} - \beta_{in} &= 0,\\
\lambda_i\beta_{in} - \beta_{i1}g_{n0} - \beta_{i2}g_{n-10} - \ldots - \beta_{ni}g_{10} &= 0.
\end{aligned}$$

Если β_{i1} считать произвольным, то из уравнений последовательно могут быть найдены β_{ij} $(j = 2, \ldots, n)$. Последнее уравнение есть следствие предыдущих и равенства $\varphi_n(\lambda_i) = 0$.

§ 5. Интерполяционный метод нахождения собственного многочлена

Интерполяционный метод является общим и позволяет вычисление определителя, элементы которого суть многочлены от некоторого буквенного параметра, привести к вычислению нескольких определителей с численными элементами, что может значительно упростить задачу. В частности, этот метод применим к вычислению собственного многочлена матрицы

$$\begin{aligned}
P_n(x) = \Delta(x) = D(xE - A) &= (-1)^n D(A - xE) =\\
= x^n - p_1x^{n-1} - p_2x^{n-2} - \ldots - p_n &= x^n - \Pi(x), \qquad (1)\\
\Pi(x) &= p_1x^{n-1} + p_2x^{n-2} + \ldots + p_n.
\end{aligned}$$

Но интерполяционный метод ввиду его общности не учитывает некоторых свойств определителя $D(xE-A)$ и поэтому в отношении числа операций несколько уступает изложенным выше методам. Несмотря на это, в книге приведена идея интерполяционного метода, так как определители $D(A-\lambda E)$ являются не единственными, содержащими произвольные параметры в своих элементах, которые встречаются в приложениях*) и к вычислению которых интерполяционный метод может быть применен.

1. Приведение к линейной системе уравнений. Для нахождения n коэффициентов $p_1, \ldots, p_n$ собственного многочлена достаточно вычислить n значений определителя $\Delta(x)$. Выберем n различных значений $(x_1, \ldots, x_n)$ параметра x и рассмотрим соответствующие им значения $\Delta(x_i) = (-1)^n D(A - x_i E)$ исходного определителя.

Для нахождения p_i $(i = 1, \ldots, n)$ получим систему n линейных уравнений

$$\begin{aligned} x_1^n - p_1 x_1^{n-1} - p_2 x_1^{n-2} - \ldots - p_n &= \Delta(x_1), \\ x_2^n - p_1 x_2^{n-1} - p_2 x_2^{n-2} - \ldots - p_n &= \Delta(x_2), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ x_n^n - p_1 x_n^{n-1} - p_2 x_n^{n-2} - \ldots - p_n &= \Delta(x_n). \end{aligned} \tag{2}$$

Ее определитель является определителем Вандермонда; он отличен от нуля, так как все x_i $(i = 1, \ldots, n)$ различны между собой; следовательно, система имеет единственное решение.

2. Связь с задачей интерполирования. Можно указать простое явное выражение $P_n(x)$ через значения определителя Δ. Рассмотрим многочлен $\Pi(x) = x^n - P_n(x) = x^n - \Delta(x) = p_1 x^{n-1} + \ldots + p_n$. Условия (2) для коэффициентов p_i можно переписать в форме

$$\Pi(x_i) = x_i^n - \Delta(x_i) \quad (i = 1, 2, \ldots, n)$$

и истолковать их как условия совпадения значений многочлена $\Pi(x)$ с функцией $x^n - \Delta(x)$ в избранных n уз-

*) К их числу принадлежит, например, определитель $D(A - xB)$ (где B — произвольная квадратная матрица), близкий к $D(A - xE)$ по строению и свойствам.

лах. Поэтому многочлен П должен быть для этой функции интерполирующим многочленом и через значения функции иметь следующее представление:

$$\Pi(x) = \sum_{i=1}^{n} \frac{\omega(x)}{(x - x_i)\,\omega'(x_i)} [x_i^n - \Delta(x_i)],$$

$$\omega(x) = (x - x_1)(x - x_2) \ldots (x - x_n),$$

что дает приводимое ниже выражение собственного многочлена $P_n(x)$ через значения определителя $\Delta(x)$ в узлах x_i:

$$P_n(x) = x^n - \sum_{i=1}^{n} \frac{\omega(x)}{(x - x_i)\,\omega'(x_i)} [x_i^n - \Delta(x_i)]. \qquad (3)$$

§ 6. Итерационный степенной метод нахождения собственных значений и собственных векторов

Этот метод применяется в частичной проблеме собственных значений и дает, вообще говоря, удобное средство для нахождения одного или небольшого числа наибольших по модулю собственных значений.

1. Нахождение наибольшего по модулю собственного значения и соответствующего собственного вектора. Будем считать, что матрица A обладает полной системой n линейно независимых собственных векторов, или, что равносильно, имеет только линейные элементарные делители. Так будет, например, в тех случаях, когда собственные значения A все различны между собой, или когда матрица A является симметричной.

Обозначим собственные значения и векторы A соответственно $\lambda_1, \lambda_2, \ldots, \lambda_n$ и $x^1, x^2, \ldots, x^n$. Предположим также, что λ_i перенумерованы в порядке невозрастания модулей:

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|.$$

Возьмем произвольный вектор $y^0 = (y_1^0, \ldots, y_n^0)$ и построим рекуррентную последовательность векторов:

$$y^0,\ y^1 = Ay^0,\ y^2 = Ay^1 = A^2y^0, \ldots$$

$$\ldots,\ y^k = Ay^{k-1} = \ldots = A^ky^0, \ldots$$

Чтобы выяснить, каким будет y^k при больших k, разложим y^0 по собственным векторам x^i:

$$y^0 = \alpha_1 x^1 + \alpha_2 x^2 + \ldots + \alpha_n x^n. \tag{1}$$

Приняв во внимание, что $A^k x^i = \lambda_i^k x^i$, получим

$$y^k = A^k y^0 = \alpha_1 \lambda_1^k x^1 + \alpha_2 \lambda_2^k x^2 + \ldots + \alpha_n \lambda_n^k x^n. \tag{2}$$

Предположим, что $|\lambda_1| > |\lambda_2|$ и $\alpha_1 \neq 0$. Тогда при больших значениях k в правой части (2) первое слагаемое будет, очевидно, главным. Для нахождения λ_1 векторное равенство (2) удобнее записать в составляющих. Введем следующие обозначения:

$$y^k = (y_1^k, y_2^k, \ldots, y_n^k), \qquad x^i = (x_1^i, x_2^i, \ldots, x_n^i).$$

Равенство (2) равносильно n численным равенствам

$$y_s^k = \beta_{1s}\lambda_1^k + \beta_{2s}\lambda_2^k + \ldots + \beta_{ns}\lambda_n^k, \quad \beta_{is} = \alpha_i x_s^i, \tag{3}$$
$$s = 1, 2, \ldots, n.$$

Отношение составляющих y_s^{k+1} и y_s^k будет равно

$$\frac{y_s^{k+1}}{y_s^k} = \frac{\beta_{1s}\lambda_1^{k+1} + \beta_{2s}\lambda_2^{k+1} + \ldots + \beta_{ns}\lambda_n^{k+1}}{\beta_{1s}\lambda_1^k + \beta_{2s}\lambda_2^k + \ldots + \beta_{ns}\lambda_n^k} =$$
$$= \lambda_1 \frac{1 + \gamma_{2s}\mu_2^{k+1} + \ldots + \gamma_{ns}\mu_n^{k+1}}{1 + \gamma_{2s}\mu_2^k + \ldots + \gamma_{ns}\mu_n^k}, \tag{4}$$

где

$$\gamma_{is} = \frac{\beta_{is}}{\beta_{1s}}, \qquad \mu_i = \frac{\lambda_i}{\lambda_1}. \tag{5}$$

Так как $|\mu_i| < 1$ $(i > 1)$, то при неограниченном росте k верно соотношение

$$\frac{y_s^{k+1}}{y_s^k} = \lambda_1 + O(|\mu_2|^k), \tag{6}$$

и для достаточно больших k с принятой точностью будет

$$\lambda_1 \approx \frac{y_s^{k+1}}{y_s^k}. \tag{7}$$

Правая часть зависит от номера s взятой составляющей, и если эта часть будет иметь одинаковое значение при всяких s в пределах принятой точности, то это является некоторой, правда, не достоверной, гарантией того, что взято достаточно большое значение k.

При достаточно больших k в представлении (2) вектора y^k все слагаемые справа, начиная со второго, будут иметь значения менее принятой погрешности вычислений, и сохранится лишь первое слагаемое. Отсюда получается правило для приближенного нахождения собственного вектора x^1:

$$x^1 \approx \frac{1}{\alpha_1 \lambda_1^k} y^k.$$

Так как x^1 определен только с точностью до численного множителя, то постоянный множитель $(\alpha_1 \lambda_1^k)^{-1}$ можно заменить любым числом и считать

$$x^1 = C_k y^k.$$

В случае симметричной матрицы A можно указать другой вычислительный процесс, более быстро сходящийся к наибольшему по модулю собственному значению λ_1. Симметричная матрица A имеет полную систему собственных векторов $x^1, \ldots, x^n$, и их всегда можно считать ортонормированными: $(x^i, x^j) = \delta_{ij} (i, j = 1, \ldots, n)$.

Представление (3) позволяет вычислить скалярные произведения

$$(y^k, y^k) = \alpha_1^2 \lambda_1^{2k} + \alpha_2^2 \lambda_2^{2k} + \ldots + \alpha_n^2 \lambda_n^{2k},$$
$$(y^{k+1}, y^k) = \alpha_1^2 \lambda_1^{2k+1} + \alpha_2^2 \lambda_2^{2k+1} + \ldots + \alpha_n^2 \lambda_n^{2k+1}.$$

Поэтому при увеличении k будет

$$\frac{(y^{k+1}, y^k)}{(y^k, y^k)} = \lambda_1 + O(|\mu_2|^{2k}), \tag{8}$$

а для достаточно больших k имеет место, следовательно, равенство

$$\lambda_1 \approx \frac{(y^{k+1}, y^k)}{(y^k, y^k)}.$$

Сравнение (6) с (8) показывает, что первое из рассмотренных правил вычисления λ_1 имеет скорость сходимости не медленнее геометрической прогрессии со знаменателем $|\mu_2|$, тогда как во втором способе аналогичный знаменатель есть $|\mu_2|^2$.

2. Некоторые более сложные случаи.

1) Кратное доминирующее собственное значение. Пусть $\lambda_1 = \lambda_2 = \ldots = \lambda_r$ и $|\lambda_1| > |\lambda_{r+1}| \geqslant \ldots$ В этом случае разложение (2) принимает вид

$$y^k = \lambda_1^k (\alpha_1 x^1 + \alpha_2 x^2 + \ldots + \alpha_r x^r) + \alpha_{r+1}\lambda_{r+1}^k x^{r+1} + \ldots$$

Составляющие же y^k будут иметь выражения

$$y_s^k = \beta_{1s}\lambda_1^k + \beta_{r+1\,s}\lambda_{r+1}^k + \ldots + \beta_{ns}\lambda_n^k, \tag{9}$$

$$\beta_{1s} = \alpha_1 x_s^1 + \ldots + \alpha_r x_s^r, \quad \beta_{is} = \alpha_s x_s^i \quad (i > r).$$

При возрастании k верно равенство

$$\frac{y_s^{k+1}}{y_s^k} = \lambda_1 + O(|\mu_{r+1}|^k), \tag{10}$$

из которого получается правило вычисления λ_1, рассчитанное на достаточно большие значения k:

$$\lambda_1 \approx \frac{y_s^{k+1}}{y_s^k}. \tag{11}$$

Для больших k в разложении y^k сохранится в принятой точности вычислений лишь слагаемое с λ_1, остальные будут пренебрежимо малы, и разложение будет иметь вид

$$y^k \approx \lambda_1^k (\alpha_1 x^1 + \ldots + \alpha_r x^r).$$

Это равенство позволяет найти только один собственный вектор, отвечающий значению λ_1. Для нахождения остальных $r-1$ собственных векторов для λ_1 нужно изменять начальный вектор C^0 и проделывать вновь указанные вычисления.

2) Случай двух наибольших собственных значений, отличающихся знаками. Предполо-

жим, что $\lambda_2 = -\lambda_1$ и $|\lambda_1| > |\lambda_3| \geqslant \ldots$ Тогда (2) будет иметь форму

$$y^k = \lambda_1^k [\alpha_1 \lambda_1^k x^1 + (-1)^k \lambda_1^k x^2] + \alpha_3 \lambda_3^k x^3 + \ldots$$

или, в зависимости от четности и нечетности числа итераций k,

$$\left.\begin{aligned} y^{2k} &= \lambda_1^{2k}(\alpha_1 x^1 + \alpha_2 x^2) + \alpha_3 \lambda_3^{2k} x^3 + \ldots \\ y^{2k+1} &= \lambda_1^{2k+1}(\alpha_1 x^1 - \alpha_2 x^2) + \alpha_3 \lambda_3^{2k+1} x^3 + \ldots \end{aligned}\right\} \quad (12)$$

Следовательно, если отличны от нуля коэффициенты при степенях λ_1, верны равенства

$$\frac{y_s^{2k+2}}{y_s^{2k}} = \lambda_1^2 + O(|\mu_3|^{2k}) \quad \text{и} \quad \frac{y_s^{2k+1}}{y_s^{2k-1}} = \lambda_1^2 + O(|\mu_3|^{2k}),$$

откуда вытекают следующие правила вычисления λ_1 при больших k:

$$\lambda_1^2 \approx \frac{y_s^{2k+2}}{y_s^{2k}}, \quad \lambda_1^2 \approx \frac{y_s^{2k+1}}{y_s^{2k-1}}.$$

Когда k есть достаточно большое число, то в выражениях для y^{2k+1} и y^{2k+2} в границах точности сохранятся только члены с λ_1 и $\lambda_2 = -\lambda_1$:

$$y^{2k+1} \approx \lambda_1^{2k+1}(\alpha_1 x^1 - \alpha_2 x^2),$$
$$y^{2k+2} \approx \lambda_1^{2k+2}(\alpha_1 x^1 + \alpha_2 x^2).$$

Отсюда, составляя комбинации $y^{2k+2} + \lambda_1 y^{2k+1}$ и $y^{2k+2} - \lambda_1 y^{2k+1}$, можно найти собственные векторы x^1 и x^2:

$$x^1 = C_k^1 (y^{2k+2} + \lambda_1 y^{2k+1}),$$
$$x^2 = C_k^2 (y^{2k+2} - \lambda_1 y^{2k+1}).$$

3) Случай двух комплексно-сопряженных наибольших по модулю собственных значений. Матрицу A и начальный вектор y^0 предполагаем действительными. Пусть $\lambda_1 = re^{i\theta}$, $\lambda_2 = re^{-i\theta} = \bar{\lambda}_1$, $|\lambda_1| = r > |\lambda_3| \geqslant \ldots$ Собственные векторы x^1 и x^2 можно считать комплексно-сопряженными: $x^2 = \bar{x}^1$. Так как y^0 и A действительны, то будут действительными и векторы $y^k = A^k y^0$ $(k = 0, 1, \ldots)$. Как и раньше, будет

верным представление (3) для составляющих y_s^k вектора y^k; при этом коэффициенты β_{is}, соответствующие комплексно-сопряженным собственным значениям, будут комплексно сопряжены. В частности, $\beta_{2s} = \bar{\beta}_{1s}$. Пусть *)

$$\beta_{1s} = Re^{i\varphi} \quad \text{и} \quad \beta_{2s} = Re^{-i\varphi}.$$

Предположим, что $R \neq 0$. Равенство (3) в рассматриваемом случае будет иметь вид

$$y_s^k = 2Rr^k \cos(k\theta + \varphi) + \beta_{3s}\lambda_3^k + \ldots + \beta_{ns}\lambda_n^k.$$

Отсюда вытекает, что при возрастании k верным является соотношение

$$y_s^k = 2Rr^k \cos(k\theta + \varphi) + O(|\lambda_3|^k). \tag{13}$$

Аналогично

$$y_s^{k+1} = 2Rr^{k+1} \cos[(k+1)\theta + \varphi] + O(|\lambda_3|^k), \tag{14}$$

$$y_s^{k+2} = 2Rr^{k+2} \cos[(k+2)\theta + \varphi] + O(|\lambda_3|^k). \tag{15}$$

Рассмотрим определитель

$$I_s^k = \begin{vmatrix} y_s^k & y_s^{k+1} \\ y_s^{k+1} & y_s^{k+2} \end{vmatrix}$$

и вычислим его значение, пользуясь приведенными выше выражениями для y_s^k, y_s^{k+1}, y_s^{k+2}:

$$\begin{aligned} I_s^k = 4R^2r^{2k+2}\{&\cos[(k+2)\theta + \varphi]\cos[k\theta + \varphi] - \\ &- \cos^2[(k+1)\theta + \varphi] + r^k O(|\lambda_3|^k)\} = \\ &= -4R^2r^{2k+2}\sin^2\theta + r^k O(|\lambda_3|^k). \end{aligned}$$

При замене k на $k-1$ получим

$$I_s^{k-1} = -4R^2r^{2k}\sin^2\theta + r^k O(|\lambda_3|^k).$$

Следовательно, $|\lambda_1| = r$ можно при достаточно больших k вычислить на основании формулы

$$r^2 \approx \frac{I_s^k}{I_s^{k-1}} \approx \frac{y_s^k y_s^{k+2} - (y_s^{k+1})^2}{y_s^{k-1} y_s^{k+1} - (y_s^k)^2}. \tag{16}$$

*) Для простоты записи индекс s у величин R и φ опущен.

Для нахождения $\cos\theta$ достаточно обратить внимание на то, что

$$y_s^{k+2} + r^2 y_s^k =$$
$$= 2Rr^{k+2}\{\cos[(k+2)\theta+\varphi] + \cos[k\theta+\varphi]\} + O(|\lambda_3|^k) =$$
$$= 4Rr^{k+2}\cos[(k+1)\theta+\varphi]\cos\theta + O(|\lambda_3|^k) =$$
$$= 2ry_s^{k+1}\cos\theta + O(|\lambda_3|^k),$$

и поэтому при больших k с принятой точностью будет

$$\cos\theta \approx \frac{y_s^{k+2} + r^2 y_s^k}{2ry_s^{k+1}}. \qquad (17)$$

В правых частях приближенных равенств (16) и (17) стоят величины, зависящие от s, т. е. от выбора составляющей векторов y^k, y^{k+1}, y^{k+2}. Если правые части в пределах принятой точности будут одинаковыми для всех s, то это может служить некоторой гарантией того, что взято достаточно большое значение k.

После нахождения λ_1 и λ_2 можно без труда найти собственные векторы x^1 и x^2. Действительно, для двух смежных итерированных векторов имеем равенства

$$y^k = \alpha_1\lambda_1^k x^1 + \alpha_2\lambda_2^k x^2 + O(|\lambda_3|^k),$$
$$y^{k+1} = \alpha_1\lambda_1^{k+1}x^1 + \alpha_2\lambda_2^{k+1}x^2 + O(|\lambda_3|^k).$$

При помощи их находим

$$y^{k+1} - \lambda_2 y^k = \alpha_1\lambda_1^k(\lambda_1 - \lambda_2)x^1 + O(|\lambda_3|^k),$$
$$y^{k+1} - \lambda_1 y^k = \alpha_2\lambda_2^k(\lambda_2 - \lambda_1)x^2 + O(|\lambda_3|^k).$$

Если отбросить справа остатки $O(|\lambda_3|^k)$, будет видно, что за векторы x^1 и x^2 можно принять соответствующие части этих равенств.

3. Нахождение собственного значения, второго по величине модуля. В принципе степенным методом может быть найдено любое собственное значение матрицы. Так, если мы хотим найти λ_2 при условии $|\lambda_2| < |\lambda_1|$, то в последовательности y^k ($k = 0, 1, \ldots$) мы должны исключить из y^k часть, содержащую λ_1, тогда главной

частью будет слагаемое, содержащее λ_2, и его можно находить сходным путем, как и выше. Однако при больших k слагаемое с λ_1 является главной частью, и исключение ее связано с большой потерей точности вычислений. Поэтому вычисление λ_2 бывает связано с необходимостью значительного увеличения точности вычислений, в частности с необходимостью находить λ_1 с большим числом верных значащих цифр, и с большой потерей точных знаков при определении λ_2. Еще большая потеря верных знаков будет происходить при нахождении собственных значений, меньших по модулю λ_2, и такие значения степенным методом находятся в редких случаях.

Рассмотрим вопрос о нахождении λ_2 в наиболее простом случае, когда выполняются неравенства

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots \tag{18}$$

Как и выше, пусть рассматриваются составляющие номера s в y^k и y^{k+1}:

$$\left.\begin{aligned} y_s^k &= \beta_{1s}\lambda_1^k + \beta_{2s}\lambda_2^k + \beta_{3s}\lambda_3^k + \ldots, \\ y_s^{k+1} &= \beta_{1s}\lambda_1^{k+1} + \beta_{2s}\lambda_2^{k+1} + \beta_{3s}\lambda_3^{k+1} + \ldots \end{aligned}\right\} \tag{19}$$

Исключим член с λ_1, для чего составим следующую комбинацию:

$$y_s^{k+1} - \lambda_1 y_s^k = \beta_{2s}\lambda_2^k(\lambda_2 - \lambda_1) + \beta_{3s}\lambda_3^k(\lambda_3 - \lambda_1) + \ldots \tag{20}$$

Заменой k на $k-1$ получим еще одно нужное нам равенство

$$y_s^k - \lambda_1 y_s^{k-1} = \beta_{2s}\lambda_2^{k-1}(\lambda_2 - \lambda_1) + \beta_{3s}\lambda_3^{k-1}(\lambda_3 - \lambda_1) + \ldots \tag{21}$$

При $\beta_{2s} \neq 0$ из двух последних равенств получим

$$\frac{y_s^{k+1} - \lambda_1 y_s^k}{y_s^k - \lambda_1 y_s^{k-1}} = \lambda_2 \frac{1 + \gamma_{3s}^* \mu_3^k + \gamma_{4s}^* \mu_4^k + \ldots}{1 + \gamma_{3s}^* \mu_3^{k-1} + \gamma_{4s}^* \mu_4^{k-1} + \ldots} = \lambda_2 [1 + O(\mu_3^k)],$$

$$\gamma_{is}^* = \frac{\beta_{is}(\lambda_i - \lambda_1)}{\beta_{2s}(\lambda_2 - \lambda_1)}, \quad \mu_i = \frac{\lambda_i}{\lambda_2} \quad (i = 3, 4, \ldots).$$

Ввиду $|\mu_3| < 1$ отсюда следует правило приближенного вычисления λ_2, верное с тем большей точностью, чем большее значение имеет k:

$$\lambda_2 \approx \frac{y_s^{k+1} - \lambda_1 y_s^k}{y_s^k - \lambda_1 y_s^{k-1}}. \tag{22}$$

Необходимо сделать некоторые пояснения техники применения формулы (22). Напомним, что при нахождении λ_1 число k выбирается столь большим, чтобы при принятом числе верных знаков в правых частях равенств вида (19) все члены, начиная со вторых, находились вне принятой точности и их можно было бы отбросить. Тогда справа сохраняются только первые члены и λ_1 найдется как отношение y_s^{k+1} к y_s^k. Пусть λ_1 найдено.

При нахождении λ_2 мы должны уменьшить значение k настолько, чтобы в правых частях равенств (19) в принятой точности еще сохранились вторые члены с λ_2, и вне этой точности оказались члены с λ_3, λ_4. ... Аналогичное должно произойти и в равенстве, полученном при замене k на $k-1$. Но тогда в правых частях равенств (19) первые члены будут превышать вторые, а вторые — превышать третие в меньшее число раз, чем при нахождении λ_1. Соответственно определяемая из (22) величина λ_2 будет содержать меньше достоверных знаков, чем λ_1.

В формуле (22) правая часть зависит от номера s выбранной составляющей, и число совпадающих цифр в значениях λ_2, полученных для разных s, дает некоторую неполную возможность судить о действительном числе полученных верных знаков.

§ 7. Итерационный метод вращений для полной проблемы собственных значений

Этот метод применим к эрмитовым матрицам с комплексными элементами, но для простоты мы рассмотрим его для действительных симметричных матриц.

1. Введение. Всякая симметричная действительная матрица A может быть приведена к диагональному виду подобным преобразованием

$$A = U\Lambda U^{-1}, \tag{1}$$

где U — ортогональная матрица и Λ — диагональная, элементами которой являются собственные значения $\lambda_1, \ldots, \lambda_n$ матрицы A. Так как для ортогональной матрицы обратная совпадает с транспонированной ($U^{-1} = U'$), то равенство (1) равносильно следующему:

$$U'AU = \Lambda. \tag{2}$$

Оно дает возможность построить много алгоритмов для приближенного вычисления матрицы Λ, отличающихся между собой способами построения матрицы U. В основании их лежит следующий простой факт. Пусть каким-либо ортогональным преобразованием с матрицей $\tilde{U}$ мы привели A к некоторой матрице $\tilde{\Lambda}$, мало отличающейся от диагональной, и получили равенство

$$\tilde{U}'A\tilde{U} = \tilde{\Lambda}, \tag{3}$$

$$\tilde{\Lambda} = \begin{bmatrix} \tilde{\lambda}_1 & \lambda_{12} & \dots & \lambda_{1n} \\ \lambda_{21} & \tilde{\lambda}_2 & \dots & \lambda_{2n} \\ \dots & \dots & \dots & \dots \\ \lambda_{n1} & \lambda_{n2} & \dots & \tilde{\lambda}_n \end{bmatrix}. \tag{4}$$

Собственные значения A и $\tilde{\Lambda}$ совпадают между собой. Если бы оказалось, что все недиагональные элементы λ_{ij} $(i \neq j)$ в $\tilde{\Lambda}$ равны нулю, то равенства (2) и (3) совпали бы, и собственные значения A были бы равны диагональным элементам $\tilde{\lambda}_i$ в $\tilde{\Lambda}$. Если же недиагональные элементы λ_{ij} $(i \neq j)$ не все равны нулю, но все будут иметь малые значения, то следует ожидать, что собственные значения A будут близкими к λ_i $(i = 1, \dots, n)$ и $\tilde{\lambda}_i$ могут быть приняты за приближенные величины этих значений.

Для использования равенства (2) нужно построить последовательность ортогональных преобразований, позволяющих неограниченно уменьшать модули недиагональных элементов матрицы A. Меру близости A к диагональному виду целесообразно определить следующим образом. Введем суммы квадратов модулей недиагональных элементов по строкам

$$\sigma_i(A) = \sum_{j=1,\ j\neq i}^{n} |a_{ij}|^2, \quad i = 1, \dots, n, \tag{5}$$

и за нужную нам величину примем число

$$t(A) = \sigma_1 + \dots + \sigma_n = \sum_{i\neq j} |a_{ij}|^2. \tag{6}$$

Пусть с помощью преобразования подобия с ортогональными матрицами построена последовательность матриц $A^0 = A$, A^1, $\dots$, A^k, $\dots$ Процесс построения

называется монотонным, если

$$t(A^k) < t(A^{k-1}).$$

Таких процессов может быть построено большое число; мы остановимся лишь на одном из них — методе вращений. Он достаточно прост по вычислительной схеме и обладает быстрой сходимостью.

2. Метод вращений. По заданной матрице A будем строить последовательность матриц A^k такую, что каждая следующая матрица A^{k+1} получается из предыдущей A^k при помощи преобразования подобия со следующей ортогональной *матрицей вращения*:

$$U_{ij}(\varphi) = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos\varphi & \cdots & -\sin\varphi & \cdots & (i) & \\ & & & \vdots & & \vdots & & & \\ & & & \sin\varphi & \cdots & \cos\varphi & \cdots & (j) & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{bmatrix} \qquad (7)$$

Предположим, что преобразования доведены до шага номера k, и построена матрица $A^k = [a_{ij}^k]$. Найдем в ней наибольший по модулю недиагональный элемент. Пусть это есть a_{ij}^k. Ввиду симметричности A^k можно считать $i < j$. Если таких элементов не один, а несколько, можно взять любой из них. По индексам i, j строим матрицу вращения $U_{ij}^k = U_{ij}^k(\varphi^k)$, в которой угол φ^k определим ниже. Образуем после этого матрицу

$$A^{k+1} = U_{ij}^{k\prime} A^k U_{ij}^k. \qquad (8)$$

Для упрощения записи введем обозначения

$$B^k = A^k U_{ij}^k, \quad B^k = [b_{ij}^k].$$

Ввиду определения (7) матрицы U_{ij} все столбцы B^k,

кроме i-го и j-го, будут такими же, как и в A^k, элементы же столбцов номеров i и j будут вычисляться по формулам

$$\begin{aligned} b_{\nu i}^k &= a_{\nu i}^k \cos\varphi^k + a_{\nu j}^k \sin\varphi^k, \\ b_{\nu j}^k &= -a_{\nu i}^k \sin\varphi^k + a_{\nu j}^k \cos\varphi^k \end{aligned} \qquad (\nu = 1, 2, \ldots, n). \tag{9}$$

Аналогично строки матрицы $A^{k+1} = U_{ij}^{k\prime} B^k$, кроме i-й и j-й, будут такими же, как в B^k, а элементы строк i-й и j-й вычисляются по формулам

$$\begin{aligned} a_{i\nu}^{k+1} &= b_{i\nu}^k \cos\varphi^k + b_{j\nu}^k \sin\varphi^k, \\ a_{j\nu}^{k+1} &= -b_{i\nu}^k \sin\varphi^k + b_{j\nu}^k \cos\varphi^k \end{aligned} \qquad (\nu = 1, 2, \ldots, n). \tag{10}$$

Равенства (9) и (10) позволяют легко вычислить a_{ij}^{k+1}:

$$\begin{aligned} a_{ij}^{k+1} = b_{ij}^k \cos\varphi^k + b_{jj}^k \sin\varphi^k = \\ = \left(-a_{ii}^k \sin\varphi^k + a_{ij}^k \cos\varphi^k\right)\cos\varphi^k + \\ + \left(-a_{ji}^k \sin\varphi^k + a_{jj}^k \cos\varphi^k\right)\sin\varphi^k \end{aligned}$$

или, так как $a_{ij}^k = a_{ji}^k$, то

$$a_{ij}^{k+1} = a_{ij}^k \cos 2\varphi^k + \frac{1}{2}\left(a_{jj}^k - a_{ii}^k\right)\sin 2\varphi^k. \tag{11}$$

Выберем теперь угол φ^k так, чтобы элемент a_{ij}^{k+1} обратился в нуль. Это требование дает

$$\operatorname{tg} 2\varphi^k = \frac{2a_{ij}^k}{a_{ii}^k - a_{jj}^k} = p_k, \quad \varphi^k = \frac{1}{2}\operatorname{arctg} p_k. \tag{12}$$

Что касается значения меры $t(A^k)$ близости A^k к диагональной форме, то, пользуясь симметричностью A^k и соотношениями (9)—(11), можно показать, что верно равенство

$$t(A^{k+1}) = t(A^k) - 2\left(a_{ij}^k\right)^2, \tag{13}$$

и, так как a_{ij}^k есть наибольший недиагональный элемент и он предполагается отличным от нуля *), верно нера-

*) Если бы $a_{ij}^k = 0$, то матрица A^k была бы диагональной, и ее днагональные элементы были собственными значениями A. Переход к A^{k+1} являлся бы излишним.

венство $t(A^{k+1}) < t(A^k)$, и мера $t(A^k)$ уменьшается при переходе к A^{k+1}. Что же касается скорости стремления $t(A^k)$ к нулю, то просто может быть получена приводимая ниже оценка. По выбору элемента a_{ij}^k справедливо неравенство

$$t(A^k) \leqslant n(n-1)\left(a_{ij}^k\right)^2$$

и, следовательно,

$$\left(a_{ij}^k\right)^2 \geqslant \frac{1}{n(n-1)} t(A^k). \tag{14}$$

С помощью этого неравенства из (13) получается

$$t(A^{k+1}) = t(A^k) - 2\left(a_{ij}^k\right)^2 \leqslant t(A^k) - \frac{2}{n(n-1)} t(A^k) =$$

$$= qt(A^k), \quad q = 1 - \frac{2}{n(n-1)};$$

при этом $0 \leqslant q < 1$ ввиду $n \geqslant 2$. Отсюда вытекает цепь неравенств

$$t(A^k) \leqslant qt(A^{k-1}) \leqslant q^2 t(A^{k-2}) \leqslant \ldots \leqslant q^k t(A^0),$$

и так как $A^0 = A$, то для $t(A^k)$ будем иметь оценку

$$t(A^k) \leqslant q^k t(A). \tag{15}$$

В частности, отсюда вытекает, что $t(A^k) \to 0$ $(k \to \infty)$ со скоростью, не меньшей скорости сходимости геометрической прогрессии со знаменателем $q < 1$.

Рассмотренный метод подобных преобразований с ортогональной матрицей требует выбора среди недиагональных элементов наибольшего по модулю, для чего необходимо выполнить приблизительно n^2 операций. Можно указать много других способов выбора a_{ij}^k, например можно взять наибольшую из сумм (5), и если это есть $\sigma_i(A^k)$, то в качестве a_{ij}^k можно взять наибольший по модулю элемент строки номера i. При таком правиле выбора a_{ij}^k необходимо будет выполнить приблизительно $2n$ операций. Оценка (15) при этом сохраняется.

§ 8. Увеличение точности приближенных собственных значений и векторов и ускорение сходимости вычислительных процессов

1. Уточнение отдельного собственного значения и соответствующего собственного вектора. Пусть известно приближенное собственное значение $\tilde{\lambda}$ матрицы A и соответствующий ему собственный вектор $\tilde{x}$. Точные их значения обозначим λ, x и положим

$$\lambda = \tilde{\lambda} + \Delta\tilde{\lambda}, \quad x = \tilde{x} + \Delta\tilde{x}. \tag{1}$$

Уравнение для приближенных значений $\tilde{\lambda}$, $\tilde{x}$ получится из точного уравнения $Ax = \lambda x$, если в него вместо λ, x поставить их выражения через $\tilde{\lambda}$ и $\tilde{x}$:

$$A(\tilde{x} + \Delta\tilde{x}) = (\tilde{\lambda} + \Delta\tilde{\lambda})(\tilde{x} + \Delta\tilde{x}). \tag{2}$$

Введем *вектор невязки* приближенных значений: $A\tilde{x} - \tilde{\lambda}\tilde{x} = r$, $r = (r_1, \ldots, r_n)$. В реальных вычислениях погрешности $\Delta\tilde{\lambda}$, $\Delta\tilde{x}$, r бывают обычно малыми величинами. Произведение $\Delta\tilde{\lambda}$, $\Delta\tilde{x}$ будет сравнительно с ними малой величиной более высокого порядка, и им можно пренебречь. После этого уравнение (2) примет вид *)

$$-\Delta\tilde{\lambda}\tilde{x} - \tilde{\lambda}\Delta\tilde{x} + A\Delta\tilde{x} = -r. \tag{3}$$

Если его записать в составляющих векторов $\tilde{x}$, r, получится следующая линейная система для погрешностей $\Delta\tilde{\lambda}$, $\Delta\tilde{x}_1$, $\ldots$, $\Delta\tilde{x}_n$:

$$\begin{aligned}
&-\Delta\tilde{\lambda}\tilde{x}_1 + (a_{11} - \tilde{\lambda})\Delta\tilde{x}_1 + a_{12}\Delta\tilde{x}_2 + \ldots + a_{1n}\Delta\tilde{x}_n = -r_1,\\
&-\Delta\tilde{\lambda}\tilde{x}_2 + a_{21}\Delta\tilde{x}_1 + (a_{22} - \tilde{\lambda})\Delta\tilde{x}_2 + \ldots + a_{2n}\Delta\tilde{x}_n = -r_2,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&-\Delta\tilde{\lambda}\tilde{x}_n + a_{n1}\Delta\tilde{x}_1 + a_{n2}\Delta\tilde{x}_2 + \ldots + (a_{nn} - \tilde{\lambda})\Delta\tilde{x}_n = -r_n.
\end{aligned} \tag{4}$$

В системе содержится n уравнений и $n+1$ неизвестных величин $\Delta\tilde{\lambda}$, $\Delta\tilde{x}_i$ $(i = 1, \ldots, n)$. Так как собственный вектор x определен только с точностью до постоянного множителя, можно одну из составляющих $\Delta\tilde{x}$ избирать произвольно. Положим, например, $\Delta x_1 = 0$ и после этого

*) Как будет видно в гл. 4, переход от (2) к (3) равносилен одному шагу метода Ньютона для нелинейного уравнения (7.2).

решим систему (4). При этом мы найдем не точные значения составляющих погрешностей $\Delta\tilde{\lambda}$ и $\Delta\tilde{x}$, а только их главные части. Прибавив $\Delta\tilde{\lambda}$ к $\tilde{\lambda}$ и $\Delta\tilde{x}$ к $\tilde{x}$, получим исправленные их значения $\lambda^* = \tilde{\lambda} + \Delta\tilde{\lambda}$, $x^* = \tilde{x} + \Delta\tilde{x}$. Если новые значения λ^* и x^* неудовлетворительны по точности, можно повторить процесс уточнения, принимая λ^* и x^* за исходные неточные значения.

2. Увеличение точности в полной проблеме собственных значений и векторов. Для простоты будем считать, что матрица A имеет попарно различные собственные значения, и предположим, что известны их приближенные величины $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n$ и приближенные собственные векторы $\tilde{x}^1, \ldots, \tilde{x}^n$. Кроме того, предположим, что известны приближенные собственные векторы $\tilde{y}^1, \ldots, \tilde{y}^n$ сопряженной матрицы A^*. Точные значения всех этих величин обозначим λ_i, x^i, y^i и положим

$$\lambda_i = \tilde{\lambda}_i + \Delta\tilde{\lambda}_i, \quad x^i = \tilde{x}^i + \Delta\tilde{x}^i, \quad y^i = \tilde{y}^i + \Delta\tilde{y}^i. \tag{5}$$

Здесь $\Delta\tilde{\lambda}_i, \Delta\tilde{x}^i, \Delta\tilde{y}^i$ являются погрешностями приближенных значений. Их мы будем считать малыми величинами, и нашей задачей будет нахождение главных частей их значений. Нам потребуются для этой цели невязки

$$\begin{aligned} A\tilde{x}^i - \tilde{\lambda}_i\tilde{x}^i &= r^i, \\ A^*\tilde{y}^i - \tilde{\lambda}_i\tilde{y}^i &= s^i. \end{aligned} \tag{6}$$

Разложим x^i и y^i по соответствующим приближенным собственным векторам:

$$\begin{aligned} x^i = \tilde{x}^i + \Delta\tilde{x}^i &= \sum_{j=1}^{n}{}^{\sim} h_{ij}\tilde{x}^j, \\ y^i = \tilde{y}^i + \Delta\tilde{y}^i &= \sum_{j=1}^{n} g_{ij}\tilde{y}^j. \end{aligned} \tag{7}$$

Так как собственные векторы x^i и y^i определяются с точностью до численных множителей, всегда можно

считать $h_{ii} = 1$ и $g_{ii} = 1$. Равенства (7) при этом примут вид

$$\begin{aligned} \Delta\tilde{x}^i &= \sum_{j \neq i} h_{ij}\tilde{x}^j, \\ \Delta\tilde{y}^i &= \sum_{j \neq i} g_{ij}\tilde{y}^j. \end{aligned} \tag{8}$$

Уравнение для погрешностей $\Delta\tilde{\lambda}_i$, $\Delta\tilde{x}^i$ получится, если в уравнение $Ax^i - \lambda_i x^i = 0$ для точных величин x^i, λ_i подставить их выражения (5):

$$A(\tilde{x}^i + \Delta\tilde{x}^i) = (\tilde{\lambda}_i + \Delta\tilde{\lambda}_i)(\tilde{x}^i + \Delta\tilde{x}^i). \tag{9}$$

Пользуясь предположением о малости погрешностей, сохраним в уравнении лишь линейные члены, отбросив справа слагаемое второго порядка малости $\Delta\tilde{\lambda}_i$, $\Delta\tilde{x}^i$. После этого останется линейное уравнение для погрешностей. Оно является неточным, но из него, как следует ожидать, могут быть найдены главные части погрешностей *)

$$A\Delta\tilde{x}^i - \tilde{\lambda}_i \Delta\tilde{x}^i = -r^i + \Delta\tilde{\lambda}_i \tilde{x}^i. \tag{10}$$

Для решения уравнения обе части этого равенства умножим скалярно на $\tilde{y}^j$:

$$(A\Delta\tilde{x}^i, \tilde{y}^j) - \tilde{\lambda}_i(\Delta\tilde{x}^i, \tilde{y}^j) = -(r^i, \tilde{y}^j) + \Delta\tilde{\lambda}_i(\tilde{x}^i, \tilde{y}^j). \tag{11}$$

Проверим, что левая часть равенства при $i = j$ есть малая величина порядка выше линейного и может быть, следовательно, отброшена. Действительно,

$$\begin{aligned} (A\Delta\tilde{x}^i, \tilde{y}^j) &= (\Delta\tilde{x}^i, A^*\tilde{y}^j) = (\Delta\tilde{x}^i, A^*(y^j - \Delta\tilde{y}^j)) = \\ &= (\Delta\tilde{x}^i, A^*y^j) - (\Delta\tilde{x}^i, A^*\Delta\tilde{y}^j) = (\Delta\tilde{x}^i, \tilde{\lambda}_j y^j) - (\Delta\tilde{x}^i, A^*\Delta\tilde{y}^j) = \\ &= \tilde{\lambda}_j(\Delta\tilde{x}^i, y^j) - (\Delta\tilde{x}^i, A^*\Delta\tilde{y}^j). \end{aligned}$$

Значит,

$$(A\Delta\tilde{x}^i, \tilde{y}^j) - \tilde{\lambda}_j(\Delta\tilde{x}^i, y^j) = -(\Delta\tilde{x}^i, A^*\Delta\tilde{y}^j). \tag{12}$$

Правая часть последнего равенства есть малая величина выше первого порядка. Что же касается левых частей

*) Такая линеаризация уравнения равносильна применению метода Ньютона к уравнению (9) (см. гл. 4).

(11) и (12), то при $i = j$ они отличаются на малую величину второго порядка: λ_i $(\Delta\tilde{x}^i, \Delta\tilde{y}^j)$. Поэтому для $i = j$ равенство (11) может быть заменено следующим:

$$-(r^i, \tilde{y}^i) + \Delta\tilde{\lambda}_i(\tilde{x}^i, \tilde{y}^i) \approx 0. \tag{13}$$

Отсюда получаем правило для вычисления главной части погрешности $\Delta\tilde{\lambda}_i$:

$$\Delta\tilde{\lambda}_i \approx \frac{(r^i, \tilde{y}^i)}{(\tilde{x}^i, \tilde{y}^i)}. \tag{14}$$

Для $i = 1, \ldots, n$ оно позволит вычислить главные части $\Delta\tilde{\lambda}_1, \ldots, \Delta\tilde{\lambda}_n$ и найти уточненные значения $\lambda_1, \ldots, \lambda_n$. Выясним теперь правила для нахождения главных частей погрешностей $\Delta\tilde{x}^i$ и $\Delta\tilde{y}^i$. Для этого достаточно указать правила вычисления коэффициентов h_{ij} и g_{ij} в (8). Способ получения их аналогичен тому, который был применен для установления (14), и мы ограничимся поэтому только объяснением наглядной стороны дела.

Возвратимся к разложениям (8). Они близки к так называемым биортогональным разложениям, последние же во многом аналогичны ортогональным разложениям Фурье. Напомним, что системы собственных векторов $x^1, \ldots, x^n$ матрицы A и $y^1, \ldots, y^n$ сопряженной с ней матрицы A^* обладают следующим *свойством биортогональности*:

$$(x^i, y^j) = 0, \quad i \neq j. \tag{15}$$

Оно позволяет находить коэффициенты Фурье разложений произвольного вектора по векторам x^i или y^i. В самом деле, пусть x есть произвольный вектор, и рассматривается его разложение по x^i:

$$x = a_1x^1 + \ldots + a_nx^n. \tag{16}$$

Для нахождения коэффициента a^i любого номера i умножим это равенство скалярно на y^i. Ввиду свойства биортогональности (15) все члены правой части разложения, кроме одного члена номера i, обратятся в нуль; получится равенство $(x, y^i) = a_i(x^i, y^i)$, из которого сразу находится a_i.

В рассматриваемых нами разложениях (8) особенность их в том, что вместо точных собственных векторов x^i, y^j в них участвуют приближенные векторы $\tilde{x}^i$, $\tilde{y}^j$.

Последние мы считаем близкими к x^i, y^j, и поэтому соотношение биортогональности (15) должны заменить приближенным

$$(\tilde{x}^i, \tilde{y}^j) \approx 0, \quad i \neq j. \tag{17}$$

Если первое из равенств (8) умножить на $\tilde{y}^j$ и пренебречь членами с малыми скалярными произведениями, согласно (17) получим

$$(\Delta\tilde{x}^i, \tilde{y}^j) \approx h_{ij}(\tilde{x}^i, \tilde{y}^j). \tag{18}$$

Для нахождения левой части воспользуемся равенством (11), внеся в него следующие упрощения. В произведении $\Delta\tilde{\lambda}_i(\tilde{x}^i, \tilde{y}^j)$ оба множителя являются малыми величинами; ими можно пренебречь в наших вычислениях. В равенстве (12) правая часть $(\Delta\tilde{x}^i, A^*\Delta\tilde{y}^j)$ есть малая величина выше первого порядка; ею также можно пренебречь и считать

$$(A\Delta\tilde{x}^i, \tilde{y}^j) \approx \tilde{\lambda}_j(\Delta\tilde{x}^i, \tilde{y}^j) \approx \tilde{\lambda}_j h_{ij}(\tilde{x}^i, \tilde{y}^j). \tag{19}$$

Ввиду (18) и (19) равенство (11) примет вид

$$(\tilde{\lambda}_i - \tilde{\lambda}_j) h_{ij}(\tilde{x}^i, \tilde{y}^j) \approx (r^i, \tilde{y}^j);$$

отсюда находится коэффициент h_{ij}:

$$h_{ij} \approx \frac{(r^i, \tilde{y}^j)}{(\tilde{\lambda}_i - \tilde{\lambda}_j)(\tilde{x}^i, \tilde{y}^j)}. \tag{20}$$

Аналогично можно получить формулу для вычисления g_{ij}:

$$g_{ij} \approx \frac{(s^i, \tilde{x}^j)}{(\tilde{\tilde{\lambda}}_i - \tilde{\tilde{\lambda}}_j)(\tilde{x}^j, \tilde{y}^i)}. \tag{21}$$

Отметим, наконец, что когда матрица A является самосопряженной и $A^* = A$, формулы вычислений (14) и (20) упрощаются:

$$\Delta\tilde{\lambda}_i = \frac{(r^i, \tilde{x}^i)}{(\tilde{x}^i, \tilde{x}^i)}, \quad h_{ij} = \frac{(r^i, \tilde{x}^j)}{(\tilde{\lambda}_i - \tilde{\lambda}_j)(\tilde{x}^i, \tilde{x}^i)}. \tag{22}$$

3. Ускорение сходимости с помощью преобразования последовательности. Начнем с задачи улучшения сходимости вычислительного процесса для наибольшего по мо-

дулю собственного значения λ_1. Напомним, что для нахождения его было использовано равенство (6.4), которое сейчас запишем в форме

$$u_k = \frac{y_s^{k+1}}{y_s^k} = \lambda_1 \frac{1 + \gamma_{2s}\mu_2^{k+1} + \gamma_{3s}\mu_3^{k+1} + \dots}{1 + \gamma_{2s}\mu_2^k + \gamma_{3s}\mu_3^k + \dots} =$$

$$= \lambda_1 - C\mu_2^k(1 + \varepsilon_k) \quad (\varepsilon_k \to 0,\ k \to \infty)$$

или

$$\lambda_1 - u_k = C\mu_2^k(1 + \varepsilon_k) \approx C\mu_2^k. \tag{23}$$

Переменная u_k сходится к пределу λ_1 приблизительно по закону стремления к нулю показательной функции μ_2^k, где $|\mu_2| < 1$. С аналогичным законом сходимости мы встречались в гл. 2, § 2, п. 6, когда рассматривали вопрос о решении системы линейных уравнений методом простой итерации. Сходимость приближенного решения x^k системы к точному решению x^* тогда определялась равенством (2.6.9) или, если от векторов x^* и x^k перейти к составляющим, равенствами (2.6.10) и (2.6.11). Там же была рассмотрена задача об ускорении сходимости $x_m^k \to x_m^*$, при этом вопрос изучался в двух формах: во-первых, когда основание λ_1 показательной функции, стоящей справа в (2.6.11), считалось известным, и тогда можно было пользоваться правилом ускорения (2.6.14); во-вторых, когда основание λ_1 предполагалось неизвестным; тогда могло быть применено правило (2.6.18).

Для переменной u_k основание μ_2 степенной функции неизвестно, и для улучшения ее сходимости мы должны воспользоваться правилом (2.6.18). Применительно к u_k оно будет иметь вид

$$v_k = \frac{u_{k+1}u_{k-1} - u_k^2}{u_{k+1} - 2u_k + u_{k-1}} \approx \lambda_1, \quad u_k = \frac{y_s^{k+1}}{y_s^k}. \tag{24}$$

Это есть преобразование последовательности u_k в новую последовательность v_k, сходящуюся к λ_1 более быстро.

Аналогичное можно сказать о вычислительном процессе λ_1 для симметричной матрицы A, основанном на

равенстве (6.8). Запишем это равенство более подробно:

$$w_k = \frac{(y^{k+1}, y^k)}{(y^k, y^k)} = \frac{\alpha_1^2\lambda_1^{2k+1} + \alpha_2^2\lambda_2^{2k+1} + \dots}{\alpha_1^2\lambda_1^{2k} + \alpha_2^2\lambda_2^{2k+1} + \dots} =$$

$$= \lambda_1 \frac{1 + \gamma_2\mu_2^{2k+1} + \gamma_3\mu_3^{2k+1} + \dots}{1 + \gamma_2\mu_2^{2k} + \gamma_3\mu_3^{2k} + \dots} = \lambda_1 + C\mu_2^{2k}(1 + \varepsilon_k)$$

$$\left(C = \gamma_2\mu_2 - \gamma_2, \quad \mu_2 = \frac{\lambda_2}{\lambda_1}\right).$$

Здесь переменная w_k сходится к λ_1 с такой же скоростью, как показательная функция μ_2^{2k} сходится к нулю. Основание μ_2 и коэффициент C здесь также неизвестны, и для ускорения сходимости может быть применено правило вида (24)

$$W_k = \frac{w_{k+1}w_{k-1} - w_k^2}{w_{k+1} - 2w_k + w_{k-1}}, \quad w_k = \frac{(y^{k+1}, y^k)}{(y^k, y^k)}. \qquad (25)$$

Сходимость $W_k \to \lambda_1$ будет более быстрой, сравнительно со сходимостью $w_k \to \lambda_1$.

Теперь рассмотрим вопрос об ускорении сходимости итерационного степенного процесса нахождения собственного вектора x^1, отвечающего наибольшему по модулю собственному значению λ_1; при этом будем считать, что λ_1 известно.

Необходимо прежде всего освободиться от одной неопределенности, которая до сих пор допускалась нами в задаче нахождения x^1. Для этой цели было использовано представление (6.2) для y^k. Выбиралось настолько большое значение k, чтобы в пределах нужной точности все члены правой части, начиная со второго, были пренебрежимо малы сравнительно с первым членом. Тогда y^k от x^1 отличалось только численным множителем. Этот множитель был оставлен неизвестным, и мы полагали $x^1 = C_k y^k$, где C_k — произвольное число.

Нас сейчас будет интересовать вопрос о сходимости, и поэтому необходимо произвести выбор определенного собственного вектора x^1. Сделать это можно многими способами: фиксируя, например, норму собственного вектора и знак одной из составляющих, или фиксируя значение одной из составляющих и т. д. Чтобы получить

сходящийся процесс приближений к собственному вектору, мы, зная λ_1, рассмотрим вектор

$$u^k = \lambda_1^{-k} y^k = \alpha_1 x^1 + \alpha_2 \mu_2^k x^2 + \alpha_3 \mu_3^k x^3 + \ldots, \quad \mu_i = \frac{\lambda_i}{\lambda_1}. \quad (26)$$

Так как $|\mu_i| < 1$ $(i = 2, \ldots, n)$, то при $k \to \infty$ u^k будет стремиться к собственному вектору $\alpha_1 x^1$.

Для каждой составляющей вектора u^k будет своя числовая последовательность приближений

$$u_s^k = \beta_{1s} + \beta_{2s}\mu_2^k + \beta_{3s}\mu_3^k + \ldots, \quad \beta_{is} = \alpha_i x_s^i. \quad (27)$$

Если $\beta_{2s} \neq 0$, то при $k \to \infty$ последовательность u_s^k будет стремиться к β_{1s} столь же быстро, как показательная функция μ_2^k стремится к нулю. Тип сходимости здесь такой же, как в предыдущих задачах, и к ускорению сходимости может быть применено правило вида (24)

$$V_k = \frac{u_s^{k+1} u_s^{k-1} - (u_s^k)^2}{u_s^{k+1} - 2u_s^k + u_s^{k-1}}. \quad (28)$$

ЛИТЕРАТУРА

1. Воеводин В. В., Численные методы алгебры (теория и алгорифмы), «Наука», М., 1966.
2. Фаддеев Д. К., Фаддеева В. Н., Вычислительные методы линейной алгебры, изд. 2-е, Физматгиз, М.—Л., 1963.
3. Уилкинсон Дж. Х., Алгебраическая проблема собственных значений, «Наука», М., 1970.
4. Мак-Кракен Д., Дорн У., Численные мстоды и программирование на Фортране, «Мир», М., 1969.
5. Форсайт Д., Молер К., Численное решение систем линейных алгебраических уравнений, «Мир», М., 1969.
6. Сборник научных программ на Фортране, вып. 1, 2. «Статистика», М., 1974.

ГЛАВА 4

РЕШЕНИЕ ЧИСЛЕННЫХ УРАВНЕНИЙ

§ 1. Введение

В общем виде задача решения уравнений может быть сформулирована в следующих словах. Пусть рассматриваются множество X, элементы которого обозначаются x, и множество Y с элементами y. Природа элементов обоих множеств может быть любой: это могут быть числа, функции, линии и т. д. Говорят, что на множестве X задан *оператор* A, если каждому элементу x из X ставится в соответствие некоторый элемент $y = A(x)$ множества Y. Элемент x часто называют *оригиналом* и y — *изображением*.

Предположим теперь, что взят какой-либо элемент y_0 из Y и нужно найти такие элементы x из X, для которых элемент y_0 является изображением. Такую задачу можно записать в форме операторного уравнения

$$A(x) = y_0. \tag{1}$$

Вопросы об условиях существования решения и об условиях его единственности принадлежат общей теории уравнений и не затрагиваются в книге. В ней будут рассматриваться только правила, позволяющие точно или приближенно найти, в зависимости от поставленной цели, все или некоторые решения уравнения*) (1). При этом ограничимся изучением только вычислительных методов, позволяющих найти приближенное решение при

*) Во многих случаях трудно отделить проблемы существования решения и эффективного его нахождения, так как весьма часто методы эффективного решения позволяют доказать существование решения и выяснить условия, при которых можно быть в этом уверенным.

помощи конечного числа арифметических действий над числами. Многие другие методы, основанные на моделировании уравнений средствами геометрии, механики, электромагнитных явлений и т. д., мы оставим в стороне, так как точность их ограничена точностью черчения и физических измерений и, кроме того, каждый из методов моделирования применим к ограниченному кругу уравнений. Вычислительные же методы являются универсальными и, вообще говоря, не ограничены по точности.

Для нас особое значение будут иметь уравнения с численными неизвестными и системы таких уравнений. Они являются частным случаем операторных уравнений, когда множества X и Y являются числовыми пространствами конечной размерности. В этом случае уравнения можно привести к виду

$$f(x)=0 \tag{2}$$

или в случае системы к виду

$$\begin{aligned} &f_1(x_1, \ldots, x_n)=0,\\ &\ldots\ldots\ldots\\ &f_n(x_1, \ldots, x_n)=0. \end{aligned} \tag{3}$$

Настоящая глава посвящена рассмотрению методов решения численных уравнений. Необходимо заметить, что в теории вычислительных методов математики не меньшее место занимает проблема приведения нечисленных операторных уравнений к численным уравнениям, что является обязательным, если для решения используются вычислительные машины. Поясним это простым примером. Пусть необходимо решить следующую граничную задачу для дифференциального уравнения второго порядка:

$$\begin{aligned} &L(x)=x''+p(t)x'+q(t)x=f(t),\\ &a\leqslant t\leqslant b, \quad x(a)=0, \quad x(b)=0. \end{aligned} \tag{4}$$

Вычислить значения функции $x(t)$ во всех точках отрезка $[a, b]$ невозможно, и необходимо прежде всего выбрать конечное число точек, в которых нужно вычислить x. Мы выберем систему равноотстоящих точек

$t_k = a + kh\left(h = \frac{b-a}{n},\ k = 0, 1, \ldots, n\right)$, так как у нас сейчас нет оснований выбирать более сложную систему.

Теперь нам необходимо построить систему численных уравнений для $n+1$ значений $x(t_k) = x_k$ $(k = 0, 1, \ldots, n)$, заменяющую точно или приближенно дифференциальное уравнение и граничные условия. Относительно последних отметим сразу же, что они дают $x_0 = 0$ и $x_n = 0$, и поэтому замене подлежит только дифференциальное уравнение. Рассмотрим его во внутренних точках t_k $(k = 1, 2, \ldots, n-1)$ отрезка $[a, b]$ и заменим значения первой и второй производных $x'(t_k)$ и $x''(t_k)$ их симметричными приближенными выражениями через значения функции *)

$$x'(t_k) \approx \frac{x_{k+1} - x_{k-1}}{2h}, \quad x''(t_k) \approx \frac{x_{k+1} - 2x_k + x_{k-1}}{h^2}.$$

Все это дает возможность приближенно заменить граничную задачу (4) для функции $x(t)$ системой численных уравнений

$$x_{k+1} - 2x_k + x_{k-1} + \frac{h}{2} p_k (x_{k+1} - x_{k-1}) + h^2 q_k x_k = h^2 f_k,$$

$$k = 1, 2, \ldots, n-1,$$

$$x_0 = 0, \quad x_n = 0,$$

$$p(t_k) = p_k, \quad q(t_k) = q_k, \quad f(t_k) = f_k.$$

Приведенный пример является простым, в более же сложных задачах такое сведение к численным уравнениям может потребовать глубоких знаний самой задачи и большой изобретательности в выборе метода сведения.

§ 2. Метод итерации; одно численное уравнение

Метод итераций или метод повторных подстановок является общим методом решения уравнений и применим к широкому классу операторных уравнений. Общность и во многих случаях хорошая сходимость позволяют часто применять его в практике вычислений. Тео-

*) При этом мы внесем погрешность, порядок малости которой такой же, как h^2.

рия метода в настоящее время исследована с большой общностью и подробностью.

Мы рассмотрим метод итерации подробно в случае одного численного уравнения, и значительно более кратко дадим его описание для систем уравнений.

Применение метода итераций требует предварительного приведения уравнения к каноническому виду

$$x = \varphi(x). \tag{1}$$

Область изменения аргумента x на числовой оси назовем X. В прикладных задачах X есть обычно конечный

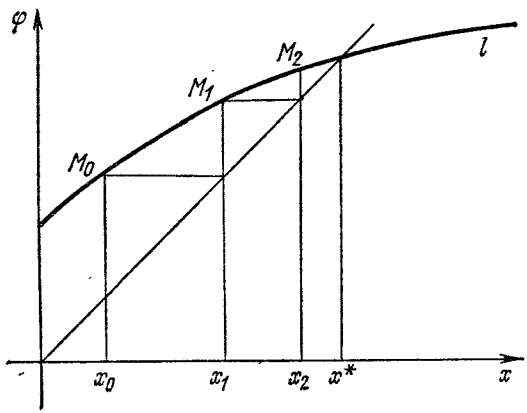

Рис. 7.

или бесконечный отрезок числовой оси. Область значений функции $y = \varphi(x)$ обозначим Y. Функцию φ можно рассматривать как оператор, преобразующий X в Y. Уравнение (1) говорит о том, что нужно найти такие точки области X, которые при преобразовании оператором φ переходят в себя, т. е. точки, остающиеся неподвижными при преобразовании X в Y.

Для нас полезным будет также изображение уравнения в координатных осях на плоскости. Построим график обеих частей уравнения (1). Для левой части это будет прямая линия $y = x$, являющаяся биссектрисой первого координатного угла. Для правой части график есть некоторая линия с уравнением $y = \varphi(x)$, обозначенная на рис. 7 буквой l. Решением уравнения является

абсцисса x^* точки M^* пересечения l и биссектрисы. Точек M^* может быть не одна, а несколько.

Допустим, что для x^* нами каким-либо способом указано начальное приближение x_0. В простейшем методе итерации все дальнейшие приближения строятся по формуле

$$x_{n+1} = \varphi(x_n), \quad n = 0, 1, \ldots \tag{2}$$

Этот процесс называется простой *одношаговой итерацией.*

Геометрическое значение процесса вычислений x_n указано на рис. 7. По x_0 на l находится точка $M_0[x_0, \varphi(x_0)]$, через нее проводится прямая, параллельная оси x, и находится точка пересечения ее с биссектрисой. Абсцисса этой точки принимается за следующее приближение x_1 к x^*, на l находится точка $M_1[x_1, \varphi(x_1)]$, дальше для M_1 повторяют такое же построение, как и для M_0 и т. д.

Следующее приближение x_{n+1} может быть построено, когда x_n принадлежит X. Допустим, что вся последовательность x_n $(n = 0, 1, 2, \ldots)$ может быть построена. Так будет в том случае, когда множество Y содержится в X, иначе говоря, когда оператор φ отражает X в себя.

Выясним не строго, но наглядно поведение приближений x_n, когда они находятся вблизи решения x^*. Нам удобнее иметь дело не с приближениями x_n, а с их погрешностями $\varepsilon_n = x_n - x^*$, так как это дает право воспользоваться малостью ε_n.

Зависимость между ε_n и ε_{n+1} получится, если в (2) вместо x_n и x_{n+1} подставить их выражения $x_n = x^* + \varepsilon_n$, $x_{n+1} = x^* + \varepsilon_{n+1}$:

$$x^* + \varepsilon_{n+1} = \varphi(x^* + \varepsilon_n) = \varphi(x^*) + \varepsilon_n \varphi'(x^*) + o(\varepsilon_n).$$

Если воспользоваться равенством $x^* = \varphi(x^*)$ и пренебречь малой величиной более высокого порядка малости $o(\varepsilon_n)$, то зависимость между ε_n и ε_{n+1} запишется в виде приближенного равенства

$$\varepsilon_{n+1} \approx \varphi'(x^*)\, \varepsilon_n. \tag{3}$$

1. Когда $|\varphi'(x^*)| > 1$, погрешность ε_{n+1} по абсолютному значению больше ε_n, и приближение x_{n+1} будет отстоять от x^* дальше, чем x_n. Решение x^* будет «точ-

кой отталкивания» для приближений x_n, близких к нему, и в этом случае не будет сходимости последовательности x_n к x^*.

2. Если $|\varphi'(x^*)|<1$, то $|\varepsilon_{n+1}|$ будет меньше $|\varepsilon_n|$, и можно ожидать, что последовательность x_n, если x_0 взято достаточно близким к x^*, будет сходиться к x^* приблизительно со скоростью геометрической прогрессии со знаменателем $q=\varphi'(x^*)$.

При $\varphi'(x^*)>0$ ε_{n+1} и ε_n будут иметь одинаковые знаки, и сходимость x_n к x^* будет монотонной. Когда же $\varphi'(x^*)<0$, погрешности ε_{n+1} и ε_n имеют разные знаки, и приближение x_n будет сходиться к x^*, колеблясь около x^*. Последнее часто облегчает суждение о точности вычислений.

3. Случай $\varphi'(x^*)=0$ требует специального рассмотрения, так как тогда ε_{n+1} будет малой величиной высшего порядка сравнительно с ε_n. Можно поэтому здесь ожидать, что если x_0 взято достаточно близко к x^*, то x_n будет весьма быстро сходиться к x^*: при возрастании n погрешность ε_n будет стремиться к нулю со скоростью, превосходящей сходимость геометрической прогрессии со сколь угодно малым знаменателем. Это часто используют для ускорения сходимости последовательности x_n к x^* путем преобразования заданного уравнения (1) к новому $x=\psi(x)$, имеющему то же решение x^*, но такому, что $\psi'(x^*)=0$.

Можно просто указать порядок малости ε_{n+1} сравнительно с ε_n. Пусть φ имеет непрерывную производную порядка m вблизи x^*, и выполняются равенства

$$\varphi'(x^*)=\ldots=\varphi^{(m-1)}(x^*)=0 \quad \text{и} \quad \varphi^{(m)}(x^*)\neq 0.$$

В этом случае разложение $\varphi(x_n)=\varphi(x^*+\varepsilon_n)$ около x^* будет иметь форму

$$\varphi(x_n)=\varphi(x^*+\varepsilon_n)=\varphi(x^*)+\frac{1}{m!}\varphi^{(m)}(x^*)\varepsilon_n^m+o(\varepsilon_n^m).$$

Подстановка его в (2) и отбрасывание $o(\varepsilon_n^m)$ даст следующее соотношение между ε_{n+1} и ε_n:

$$\varepsilon_{n+1}\approx\frac{1}{m!}\varphi^{(m)}(x^*)\varepsilon_n^m. \tag{4}$$

Отсюда видно, что ε_{n+1} будет малой величиной порядка m относительно ε_n.

Докажем теперь простую теорему о сходимости итерационной последовательности, где указываются достаточные для этого условия.

Теорема 1. *Пусть выполняются условия:*

1) *функция* $\varphi(x)$ *определена на отрезке*

$$|x - x_0| \leqslant \delta, \tag{5}$$

непрерывна там и удовлетворяет условию Липшица с постоянным коэффициентом, меньшим единицы:

$$|\varphi(x) - \varphi(x')| \leqslant q|x - x'| \quad (0 \leqslant q < 1); \tag{6}$$

2) *для исходного приближения* x_0 *верно неравенство*

$$|x_0 - \varphi(x_0)| \leqslant m;$$

3) *числа* δ, q, m *удовлетворяют условию*

$$\frac{m}{1-q} \leqslant \delta. \tag{7}$$

Тогда

1) *уравнение* (1) *в области* (5) *имеет решение;*

2) *последовательность* x_n *приближений, построенная по правилу* (2), *принадлежит отрезку* (5), *является сходящейся* ($\lim x_n = x^*$), *и предел последовательности* x^* *удовлетворяет уравнению* (1);

3) *скорость сходимости* x_n *к* x^* *оценивается неравенством*

$$|x^* - x_n| \leqslant \frac{m}{1-q} q^n, \quad n = 1, 2, \ldots \tag{8}$$

Перед доказательством поясним условия теоремы. Функция φ преобразует отрезок $x_0 - \delta \leqslant x \leqslant x_0 + \delta$ числовой оси в некоторый отрезок той же оси. Возьмем две точки x и x' на $[x_0 - \delta, x_0 + \delta]$. Расстояние между ними есть $|x - x'|$, а $|\varphi(x) - \varphi(x')|$ есть расстояние между их изображениями. Отношение $\frac{|\varphi(x) - \varphi(x')|}{|x - x'|}$ есть коэффициент увеличения этих расстояний при преобразовании. По условию (6) он не превосходит числа q. Но так как число q меньше единицы, то при отображении оператором φ происходит не растяжение, а сжатие всех отрезков с коэффициентом, не большим q.

Входящая в условие величина m связана с близостью начального приближения x_0 к решению x^*. Если окажется, что $x_0 = x^*$ и, стало быть, $x_0 - \varphi(x_0) = 0$, то можно считать $m = 0$. Когда $x_0 \neq x^*$, но x_0 близко к x^*, то разность $x_0 - \varphi(x_0)$ будет иметь малое значение, и m может быть взята малой величиной.

Неравенство (7) налагает на δ, q и m ограничение, достаточное для того, чтобы были верны утверждения теоремы.

Доказательство. Покажем, прибегнув к индукции, что при всяких значениях $n = 1, 2, \ldots$ приближения x_n лежат на отрезке (5) и для них верно неравенство

$$|x_{n+1} - x_n| \leqslant mq^n. \tag{9}$$

При $n = 0$ неравенство проверяется просто. Приближение $x_1 = \varphi(x_0)$, очевидно, может быть найдено, так как $x = x_0$ принадлежит отрезку (5). Кроме того, $|x_1 - x_0| = |\varphi(x_0) - x_0| \leqslant m$ по второму условию теоремы, и неравенство (9) для $n = 0$ выполнено. Наконец, так как $m \leqslant \frac{m}{1-q} \leqslant \delta$, то x_1 принадлежит отрезку (5).

Предположим, что $x_0, x_1, \ldots, x_n$ принадлежат области (5) и выполняются условия

$$|x_{k+1} - x_k| \leqslant mq^k \quad (k = 0, 1, \ldots, n-1).$$

Так как x_n, по предположению, принадлежит области (5), приближение $x_{n+1} = \varphi(x_n)$ может быть построено.

По сделанному допущению $|x_n - x_{n-1}| \leqslant mq^{n-1}$, поэтому

$$|x_{n+1} - x_n| = |\varphi(x_n) - \varphi(x_{n-1})| \leqslant q|x_n - x_{n-1}| \leqslant qmq^{n-1} = mq^n,$$

и для приближений x_{n+1}, x_n неравенство (9) выполнено. Осталось еще проверить принадлежность x_{n+1} области (5).

$$|x_{n+1} - x_0| = |(x_{n+1} - x_n) + (x_n - x_{n-1}) + \ldots + (x_1 - x_0)| \leqslant$$

$$\leqslant mq^n + mq^{n-1} + \ldots + m = \frac{m - mq^{n+1}}{1-q} < \frac{m}{1-q} \leqslant \delta.$$

Этим завершается индукция.

Покажем теперь, что для последовательности x_n выполняется условие Больцано — Коши

$$|x_{n+p}-x_n| = |(x_{n+p}-x_{n+p-1}) + (x_{n+p-1}-x_{n+p-2}) + \ldots + (x_{n+1}-x_n)| \leqslant$$

$$\leqslant mq^{n+p-1} + mq^{n+p-2} + \ldots + mq^n = m\frac{q^n - q^{n+p}}{1-q} < \frac{m}{1-q}q^n.$$

Последняя часть цепочки неравенств не зависит от p и, ввиду $0 < q < 1$, при всяких достаточно больших n будет меньше любого заданного заранее числа. Признак сходимости для x_n действительно выполняется, и поэтому существует

$$\lim_{n\to\infty} x_n = x^*.$$

Принадлежность x^* замкнутому отрезку $|x-x_0| \leqslant \delta$ следует из того, что ему принадлежат все x_n.

Покажем, что x^* есть решение заданного уравнения. Для этого в правиле вычислений $x_{n+1} = \varphi(x_n)$ устремим n к бесконечности. Тогда будет $x_{n+1} \to x^*$ и $x_n \to x^*$, ввиду же непрерывности $\varphi(x)$ во всех точках отрезка (5) при этом будет $\varphi(x_n) \to \varphi(x^*)$. В пределе получится равенство $x^* = \varphi(x^*)$, говорящее о том, что x^* есть решение рассматриваемого уравнения.

Дополнительно укажем еще теорему единственности решения.

Теорема 2. *На всяком множестве точек, где для функции $\varphi(x)$ выполняется условие*

$$|\varphi(x) - \varphi(y)| < |x-y|, \quad x \neq y,$$

уравнение $x = \varphi(x)$ может иметь не более одного решения.

Доказательство. Допустим противоположное и будем считать, что на указанном множестве существуют два разных решения x и y $(x \neq y)$. Оценим разность $x-y$:

$$|x-y| = |\varphi(x) - \varphi(y)| < |x-y|,$$

и мы приходим к невозможному неравенству $|x-y| < |x-y|$. Поэтому предположение о существовании двух разных решений неверно.

§ 3. Об ускорении сходимости итерационного метода

1. О задаче ускорения сходимости. В предыдущем параграфе обращалось внимание на то, что в простом одношаговом итерационном процессе погрешность $\varepsilon_n = x_n - x^*$, если приближенные значения x_n находятся вблизи решения x^*, изменяется приблизительно по закону геометрической прогрессии

$$\varepsilon_{n+1} \approx \varphi'(x^*)\,\varepsilon_n,$$

знаменатель которой есть $\varphi'(x^*)$.

Последовательность x_n сходится к решению x^*, если $|\varphi'(x^*)| < 1$ и если начальное приближение взято достаточно близко к x^*. Но скорость сходимости зависит от $|\varphi'(x^*)|$. Если $|\varphi'(x^*)|$ близок к единице, то сходимость может быть очень медленной, и для получения нужной точности потребуется проделать много шагов вычислений.

Как всякий процесс последовательных приближений, простой одношаговый итерационный процесс можно улучшать, преследуя при этом две цели: ускорение сходимости процесса и ослабление условий сходимости. Для процесса (2.2) последнее означало бы замену требования $|\varphi'| < 1$ другим, менее ограничительным условием.

Для осуществления этих целей можно воспользоваться двумя средствами: изменять заданное уравнение (1) и изменять итерационный процесс. Начнем с описания изменений процесса и укажем два возможных способа. Правило итерации (2.2) является одношаговым, и, как всякая одношаговая итерация, не использует многих возможностей, содержащихся в вычислительном процессе.

Поясним это обстоятельство более подробно. Пусть в некотором итерационном процессе для уравнения (2.1) вычисления доведены до приближения x_n и составлена таблица x_i $(i = 0, 1, \ldots, n)$ и соответствующих значений $\varphi(x_i)$.

В формуле (2.2) мы используем только одно предшествующее значение x_n, полагаем $x_{n+1} = \varphi(x_n)$ и не пользуемся ни одним предшествующим приближением x_{n-1}, x_{n-2}, ... или значением $\varphi(x_{n-1})$, $\varphi(x_{n-2})$, ... Для

геометрической картины, изображенной на рис. 7, это означает, что мы находим не точку пересечения линии l с биссектрисой $y = x$, а линию l заменяем прямой $y = \varphi(x_n)$, проходящей через точку $M_n[x_n, \varphi(x_n)]$, и определяем ее пересечение с биссектрисой.

Если пользоваться языком теории интерполирования, то можно сказать, что при нахождении x_{n+1} мы выполняем интерполирование функции $\varphi(x)$ постоянной величиной — ее значением $\varphi(x_n)$.

Можно пытаться улучшить точность нахождения x_{n+1}, повысив степень интерполирования и привлекая для этого не одну точку M_n, а несколько таких точек $M_n[x_n, \varphi(x_n)]$, $M_{n-1}[x_{n-1}, \varphi(x_{n-1})]$, ..., $M_{n-m}[x_{n-m}, \varphi(x_{n-m})]$. С интерполяционными методами решения уравнений в общей форме мы ознакомимся в одном из следующих параграфов.

2. Метод линейного интерполирования или метод секущих. Сейчас остановим внимание на случае линейного интерполирования φ по двум парам чисел $[x_n, \varphi(x_n)]$ и $[x_{n-1}, \varphi(x_{n-1})]$. С геометрической точки зрения это означает, что линию l мы заменим секущей, проходящей через точки M_n и M_{n-1}.

Точка пересечения секущей и биссектрисы определится системой уравнений

$$\frac{x - x_{n-1}}{x_n - x_{n-1}} = \frac{y - \varphi(x_{n-1})}{\varphi(x_n) - \varphi(x_{n-1})}, \quad y = x.$$

Решение ее относительно x приведет к правилу вычисления следующего приближения:

$$x_{n+1} = \frac{x_{n-1}\varphi(x_n) - x_n\varphi(x_{n-1})}{\varphi(x_n) - x_n - \varphi(x_{n-1}) + x_{n-1}}. \qquad (1)$$

Оно является двухшаговым, и применение его требует знания двух начальных приближений x_0 и x_1 к корню.

Теорему о сходимости правила секущих не станем приводить и ограничимся описанием наглядной картины поведения погрешности $\varepsilon_n = x_n - x^*$, когда x_n будет близким к решению x^*, а ε_n — малой величиной. Из (1) легко получается соотношение между тремя последовательными значениями погрешности ε_{n-1}, ε_n, ε_{n+1}, если

в (1) вместо x_{n-1}, x_n, x_{n+1} внести их выражения вида $x_n = x^* + \varepsilon_n$ через погрешности

$$x^* + \varepsilon_{n+1} = \frac{(x^* + \varepsilon_{n-1})\varphi(x^* + \varepsilon_n) - (x^* + \varepsilon_n)\varphi(x^* + \varepsilon_{n-1})}{\varphi(x^* + \varepsilon_n) - x^* - \varepsilon_n - \varphi(x^* + \varepsilon_{n-1}) + x^* + \varepsilon_{n-1}}.$$

Заменим величины $\varphi(x^* + \varepsilon_k)$ $(k = n, n-1)$ разложениями по формуле Тейлора около точки x^* и примем во внимание, что $\varphi(x^*) = x^*$:

$$\varphi(x^* + \varepsilon_k) = x^* + \alpha\varepsilon_k + \beta\varepsilon_k^2 + \gamma\varepsilon_k^3 + o(\varepsilon_k^3),$$

$$\alpha = \varphi'(x_k), \quad \beta = \frac{1}{2}\varphi''(x_k), \quad \gamma = \frac{1}{6}\varphi'''(x_k).$$

Тогда будет

$$x^* + \varepsilon_{n+1} = \frac{x^*[(\alpha-1)(\varepsilon_n - \varepsilon_{n-1}) + \beta(\varepsilon_n^2 - \varepsilon_{n-1}^2) + \gamma(\varepsilon_n^3 - \varepsilon_{n-1}^3)]}{d} +$$

$$+ \frac{\beta\varepsilon_{n-1}\varepsilon_n(\varepsilon_n - \varepsilon_{n-1}) + o(\varepsilon_n^3) - o(\varepsilon_{n-1}^3)}{d},$$

где $d = (\alpha-1)(\varepsilon_n - \varepsilon_{n-1}) + \beta(\varepsilon_n^2 - \varepsilon_{n-1}^2) + \gamma(\varepsilon_n^3 - \varepsilon_{n-1}^3) + o(\varepsilon_n^3) + o(\varepsilon_{n-1}^3)$.

Отбрасывая в правой части малые величины выше третьего порядка $o(\varepsilon_n^3)$ и $o(\varepsilon_{n-1}^3)$, выделяя целую часть x^*, сокращая в оставшейся дроби общий множитель $\varepsilon_n - \varepsilon_{n-1}$ и, наконец, сохраняя в ней только главную часть, получим приближенное равенство

$$\varepsilon_{n+1} \approx \frac{\beta\varepsilon_{n-1}\varepsilon_n}{\alpha - 1} = \frac{1}{2}\frac{\varphi''(x^*)}{\varphi'(x^*) - 1}\varepsilon_{n-1}\varepsilon_n, \qquad (2)$$

дающее достаточно простое описание закона изменения погрешности ε_n на одном шаге вычислений. Равенство (2) является приближенным, и все заключения об ε_n, вытекающие из него, можно считать лишь ориентировочными.

Для ε_n равенство (2) является уравнением в конечных разностях второго порядка. Нас будет интересовать абсолютное значение погрешности $|\varepsilon_n| = E_n$, и вместо (2) мы будем рассматривать уравнение

$$E_{n+1} \approx AE_nE_{n-1}, \quad A = \left|\frac{1}{2}\frac{\varphi''(x^*)}{\varphi'(x^*) - 1}\right|.$$

Для решения приведем его к более хорошо известному линейному уравнению при помощи перехода к логарифмам. Если обозначить $\ln E_n = \alpha_n$ и $\ln A = a$, для α_n получим линейное уравнение с постоянными коэффициентами

$$\alpha_{n+1} - \alpha_n - \alpha_{n-1} \approx a. \tag{3}$$

Это уравнение также является приближенным. Мы найдем решение этого уравнения, заменяя приближенное равенство (3) на точное и считая, что решение от этого мало изменится. Решение можно найти, воспользовавшись известными правилами решения конечноразностных уравнений *). Сразу же видно, что $\alpha_n^0 = -a$ есть решение неоднородного уравнения. Характеристическое уравнение для (3) есть $\lambda^2 - \lambda - 1 = 0$. Его корни имеют значения $\lambda_1 = \frac{\sqrt{5}+1}{2} \approx 1{,}618$, $\lambda_2 = -\frac{\sqrt{5}-1}{2} \approx$ $\approx -0{,}618$, и функции λ_1^n, λ_2^n являются линейно независимыми решениями уравнения. Можно поэтому ожидать, что для α_n верно приближенное представление

$$\alpha_n \approx C_1\lambda_1^n + C_2\lambda_2^n - a, \tag{4}$$

и, следовательно,

$$|\varepsilon_n| = E_n \approx A^{-1} e^{C_1\lambda_1^n} e^{C_2\lambda_2^n}. \tag{5}$$

*) Пусть дано линейное разностное уравнение любого порядка k

$$L_k(y_n) = a_0 y_{n+k} + a_1 y_{n+k-1} + \ldots + a_k y_n = f_n, \quad n = 0, 1, \ldots$$

Известно, что общее решение неоднородного уравнения есть сумма частиого решения неоднородного уравнения y_n^0 и общего решения z_n соответствующего однородного уравнения $L_k(z_n) = 0$. Если коэффициенты a_i $(i = 0, 1, \ldots, k)$ являются величиниами постоянными, то для нахождения z_n нужно составить характеристическое уравнение $\varphi_k(\lambda) = a_0\lambda^k + a_1\lambda^{k-1} + \ldots + a_k = 0$ и найти его корни λ_1, $\lambda_2, \ldots, \lambda_k$. Если они различны: $\lambda_i \neq \lambda_j$ $(j \neq i)$, то $\lambda_1^n, \lambda_2^n, \ldots, \lambda_k^n$ образуют фундамеитальную систему решений, и общее решение однородиого уравнения $L_k(z_n) = 0$ есть нх линейная комбинация с постоянными коэффициентами, общим же решением неодиородного уравнения будет

$$y_n = y_n^0 + \sum_{i=1}^{k} C_i\lambda_i^n,$$

где C_i — произвольные постоянные.

Постоянные C_1 и C_2 должны быть найдены при помощи начальных значений ε_0 и ε_1 погрешности. Напомним, что равенство (2) получено при предположении малости значений погрешности ε_n. Изменяя, если нужно, нумерацию значений ε_n, мы можем считать, что такое предположение выполняется для всех значений ε_n $(n = 0, 1, \ldots)$. Если в (4) положить $n = 0$ и $n = 1$, то для C_1 и C_2 получится линейная система уравнений

$$C_1 + C_2 = \alpha_0 + a, \quad C_1\lambda_1 + C_2\lambda_2 = \alpha_1 + a,$$

откуда находятся C_1 и C_2:

$$C_1 = \frac{a + \alpha_1 - (a + \alpha_0)\lambda_2}{\lambda_1 - \lambda_2} = \frac{1}{\sqrt{5}}[a + \alpha_1 - (a + \alpha_0)\lambda_2],$$

$$C_2 = \frac{1}{\sqrt{5}}[(a + \alpha_0)\lambda_1 - a - \alpha_1].$$

Полученное выражение (5) для $|\varepsilon_n|$ позволяет судить о вероятном законе изменения $|\varepsilon_n|$ при росте n. Множитель A^{-1} в правой части (5) не зависит от n. Так как $|\lambda_2| \approx 0{,}62$, то $\lambda_2^n \to 0$ при $n \to \infty$, и множитель $\exp(C_2\lambda_2^n)$ будет стремиться к единице. Наконец, ввиду $\lambda_1 > 1{,}6$ величина λ_1^n будет быстро возрастать при $n \to \infty$, поведение же множителя $\exp(C_1\lambda_1^n)$ зависит от знака C_1. Когда $C_1 < 0$, то $\exp(C_1\lambda_1^n)$ будет быстро стремиться к нулю.

Отрицательность же C_1 равносильна выполнению неравенства $a + \alpha_1 - (a + \alpha_0)\lambda_2 < 0$ или, если умножить обе части его на λ_1 и принять во внимание, что $\lambda_1\lambda_2 = -1$, неравенству $(a + \alpha_1)\lambda_1 + (a + \alpha_0) < 0$, что равносильно

$$|\varepsilon_0| \cdot |\varepsilon_1|^{\lambda_1} A^{1+\lambda_1} < 1. \tag{6}$$

Это есть условие, которому должны удовлетворять погрешности приближенных значений x_0 и x_1 для того, чтобы можно было ожидать сходимости итерационного процесса (1).

Так как решение x^* является неизвестным, то число $A = \frac{1}{2}\varphi''(x^*)(\varphi'(x^*) - 1)^{-1}$ точно не может быть найдено, но приближенное значение A можно получить, если заменить x^* величиной x_n любого номера n.

3. Применение преобразования Эйткена. Приведем пример ускорения сходимости итерационного процесса путем введения вспомогательных значений функции φ. Излагаемый ниже метод связан по идее с задачей ускорения сходимости последовательности в некотором частном случае. Эта задача будет более подробно рассматриваться во второй части, а сейчас мы введем только понятие о нужном нам преобразовании. В общем виде задача формулируется следующими словами. Пусть дана последовательность $s_1, s_2, \ldots, s_n, \ldots$, сходящаяся к пределу s. По ней нужно построить новую последовательность

$$\sigma_n = f_n(s_1, s_2, \ldots), \quad n = 1, 2, \ldots,$$

сходящуюся к тому же значению s, что и s_n, но более быстро. Достичь такого ускорения сходимости можно, очевидно, если согласовать характер преобразования со свойствами сходимости s_n к s.

В процессе простой одношаговой итерации (2.2) приходится иметь дело со сравнительно простым законом изменением погрешности $\varepsilon_n = x_n - x^*$. Предположим, что $|\varphi'(x^*)| < 1$, и приближение x_0 взято близким к корню x^*. Тогда последовательность x_n сходится к x^*, и погрешность ε_n на одном шаге при переходе от x_{n-1} к x_n изменится по правилу вида (2.3):

$$\varepsilon_n \approx \varphi'(x^*)\,\varepsilon_{n-1} \approx \ldots \approx [\varphi'(x^*)]^n \varepsilon_0.$$

Поэтому приближение x_n есть показательная функция от n следующего вида:

$$x_n \approx x^* + \varepsilon_0 q^n, \quad q = \varphi'(x^*) \tag{7}$$

(равенство приближенное). Чтобы получить преобразование, которое могло бы ускорить сходимость x_n, рассмотрим показательную функцию $s_n = s + Aq^n$ $(n = 0, 1, \ldots)$. При $|q| < 1$ s_n, очевидно, сходится к s. В этом простом случае легко построить преобразование, которое давало бы выражение предельного значения s через три значения s_n. В самом деле, если $s_n - s = Aq^n$ разделить на $s_{n-1} - s = Aq^{n-1}$, то получится $\frac{s_n - s}{s_{n-1} - s} = q$. Аналогично $\frac{s_{n+1} - s}{s_n - s} = q$. Сравнение этих

двух значений для q приведет к равенству $(s_{n+1} - s)(s_{n-1} - s) = (s_n - s)^2$, откуда следует

$$s = \frac{s_{n+1}s_{n-1} - s_n^2}{s_{n+1} - 2s_n + s_{n-1}}.$$

В соответствии с этим результатом рассмотрим принадлежащее Эйткену преобразование произвольной последовательности s_n в другую последовательность σ_n:

$$\sigma_n = \frac{s_{n+1}s_{n-1} - s_n^2}{s_{n+1} - 2s_n + s_{n-1}}. \tag{8}$$

Если его применить к последовательности $s_n = s + Aq^n$, имеющей показательный тип сходимости, то при всяком значении n будет $\sigma_n = s = \lim s_n$. Можно ожидать, что если s_n будет иметь не точно показательный тип сходимости, а будет изменяться по закону, близкому к нему, то преобразование (8), вообще говоря, не будет давать при всяком n предельное значение s_n, но приведет к новой последовательности σ_n, сходящейся к $s = \lim s_n$ более быстро, чем s_n.

Как показывает приближенное равенство (7), последовательные приближения x_n в методе простой одношаговой итерации с возрастанием n изменяются по закону, близкому к показательной функции. Поэтому для улучшения сходимости x_n естественно воспользоваться преобразованием Эйткена. При этом целесообразно каждое улучшенное значение сразу же вводить в вычисление, чтобы в последующих вычислениях найденное улучшение было учтено. Поясним это на одном шаге вычислений. Пусть вычисления доведены до значения x_n; по нему вычисляем два вспомогательных значения $x_n' = \varphi(x_n)$, $x_n'' = \varphi(x_n') = \varphi[\varphi(x_n)]$. К трем значениям x_n, x_n', x_n'' применяем правило улучшения (8) и результат принимаем за следующее приближение x_{n+1}:

$$x_{n+1} = \frac{x_n x_n'' - [x_n']^2}{x_n'' - 2x_n' + x_n} = \frac{x_n \varphi[\varphi(x_n)] - \varphi^2(x_n)}{\varphi[\varphi(x_n)] - 2\varphi(x_n) + x_n}. \tag{9}$$

Полученное равенство называют итерационной формулой Стеффенсена, она является одношаговой и требует вычисления двух значений φ на каждом шаге. Ее

можно истолковать как простой итерационный процесс для вспомогательного уравнения

$$x = \Phi(x), \qquad \Phi(x) = \frac{x\varphi[\varphi(x)] - \varphi^2(x)}{\varphi[\varphi(x)] - 2\varphi(x) + x}. \tag{10}$$

Остановимся еще на вопросе о поведении погрешности $\varepsilon_n = x_n - x^*$ формулы (9) вблизи решения x^*. Для этой цели все величины, входящие в правую часть (9), разложим по степеням ε_n, подставив всюду вместо x_n его выражение $x_n = x^* + \varepsilon_n$. Для упрощения записи в вычислениях будем опускать у x^* и ε_n индексы * и n и писать $x_n = x + \varepsilon$. Воспользуемся также тем, что $\varphi(x) = x$. Тогда:

$$\varphi(x_n) = \varphi(x + \varepsilon) = x + \alpha\varepsilon + \beta\varepsilon^2 + \gamma\varepsilon^3 + \ldots,$$

$$\alpha = \varphi'(x), \quad \beta = \frac{1}{2}\varphi''(x), \quad \gamma = \frac{1}{6}\varphi'''(x).$$

$$\varphi[\varphi(x_n)] = x + \alpha(\alpha\varepsilon + \beta\varepsilon^2 + \gamma\varepsilon^3 + \ldots) +$$
$$+ \beta(\alpha\varepsilon + \beta\varepsilon^2 + \ldots)^2 + \gamma(\alpha\varepsilon + \ldots)^3 + \ldots =$$
$$= x + \alpha^2\varepsilon + \alpha\beta(1+\alpha)\varepsilon^2 + \alpha(\gamma + 2\beta^2 + \alpha^2\gamma)\varepsilon^3 + \ldots$$

$$x_n\varphi[\varphi(x_n)] = (x+\varepsilon)[x + \alpha^2\varepsilon + \alpha\beta(1+\alpha)\varepsilon^2 +$$
$$+ \alpha(\gamma + 2\beta^2 + \alpha^2\gamma)\varepsilon^3 + \ldots] = x^2 + x(1+\alpha^2)\varepsilon +$$
$$+ [\alpha^2 + x\alpha\beta(1+\alpha)]\varepsilon^2 + [\alpha\beta(1+\alpha) + x\alpha(\gamma + 2\beta^2 + \alpha^2\gamma)]\varepsilon^3 + \ldots$$

$$\varphi^2(x_n) = (x + \alpha\varepsilon + \beta\varepsilon^2 + \gamma\varepsilon^3 + \ldots)^2 =$$
$$= x^2 + 2x\alpha\varepsilon + (\alpha^2 + 2\beta x)\varepsilon^2 + 2(\gamma x + \alpha\beta)\varepsilon^3 + \ldots$$

$$x_n\varphi[\varphi(x_n)] - \varphi^2(x_n) = x\{(\alpha-1)^2\varepsilon + \beta(\alpha+2)(\alpha-1)\varepsilon^2 +$$
$$+ [(\alpha-2)\gamma + 2\alpha\beta^2 + \alpha^3\gamma]\varepsilon^3\} + \alpha\beta(\alpha-1)\varepsilon^3 + \ldots$$

$$\varphi[\varphi(x_n)] - 2\varphi(x_n) + x_n = (\alpha-1)^2\varepsilon + \beta(\alpha+2)(\alpha-1)\varepsilon^2 +$$
$$+ [(\alpha-2)\gamma + 2\alpha\beta^2 + \alpha^3\gamma]\varepsilon^3 + \ldots$$

$$x_{n+1} = \frac{x_n\varphi[\varphi(x_n)] - \varphi^2(x_n)}{\varphi[\varphi(x_n)] - 2\varphi(x_n) + x_n} =$$
$$= \frac{x\{(\alpha-1)^2\varepsilon + \beta(\alpha+2)(\alpha-1)\varepsilon^2 + [(\alpha-2)\gamma + 2\alpha\beta^2 + \alpha^3\gamma]\varepsilon^3\}}{d\varepsilon} +$$
$$+ \frac{\alpha\beta(\alpha-1)\varepsilon^3 + \ldots}{d\varepsilon} = x + \frac{\alpha\beta(\alpha-1)\varepsilon^2 + \ldots}{d},$$

где $d = (\alpha-1)^2 + \beta(\alpha+2)(\alpha-1)\varepsilon + [(\alpha-2)\gamma + 2\alpha\beta^2 + \alpha^3\gamma]\varepsilon^2 + \ldots$

Слева заменим x_{n+1} на $x^* + \varepsilon_{n+1}$, справа же возвратимся к прежним обозначениям, заменив x на x^* и ε на ε_n, и, кроме того, выделим там лишь главный член, сохранив в числителе дроби только слагаемое с ε_n^2, а в знаменателе — член $(\alpha - 1)^2$, свободный от ε_n. После этого получим приводимое ниже приближенное выражение ε_{n+1} через ε_n:

$$\varepsilon_{n+1} \approx \frac{1}{2} \frac{\alpha\beta}{\alpha - 1} \varepsilon_n^2 = \frac{1}{2} \frac{\varphi'(x^*)\,\varphi''(x^*)}{\varphi'(x^*) - 1} \varepsilon_n^2 = B\varepsilon_n^2. \qquad (11)$$

Равенство является приближенным, и так как нумерацию погрешностей можно начать с любого шага вычислений, мы вправе, не ограничивая общности, считать, что оно выполняется с нужной точностью при $n = 0, 1, 2, \ldots$ Его можно рассматривать кок нелинейное разностное уравнение первого порядка. Если умножить (11) почленно на B и ввести новую переменную $\eta_n = B\varepsilon_n$, то это равенство можно записать в форме $\eta_{n+1} = \eta_n^2$. Применив его несколько раз, начиная со значения η_n, получим следующую систему равенств:

$$\eta_n \approx \eta_{n-1}^2 \approx (\eta_{n-2}^2)^2 \approx \ldots \approx \eta_0^{2^n}.$$

Отсюда получаем

$$\varepsilon_n \approx B^{-1} (B\varepsilon_0)^{2^n}. \qquad (12)$$

Если погрешность $\varepsilon_0 = x_0 - x^*$ начального значения x_0 настолько мала, что для нее выполняется неравенство

$$|B\varepsilon_0| = \left| \frac{1}{2} \frac{\varphi'(x^*)\,\varphi''(x^*)}{\varphi'(x^*) - 1} \varepsilon_0 \right| < 1, \qquad (13)$$

то можно ожидать, что при неограниченном возрастании n погрешность ε_n будет стремиться к нулю и итерационный процесс Стеффенсена, определяемый формулой (9), будет сходиться к решению x^*; при этом сходимость будет весьма быстрой, как это следует из (12).

4. Ускорение сходимости при помощи преобразования уравнения. В основании преобразования уравнения лежит следующий факт, который был отмечен в § 2:

если в уравнении $x = \varphi(x)$ функция $\varphi(x)$ такова, что при $x = x^*$ будут выполняться равенства $\varphi'(x^*) = \ldots$ $\ldots = \varphi^{(m-1)}(x^*) = 0$ и $\varphi^{(m)}(x^*) \neq 0$ и, следовательно, разложение $\varphi(x)$ по степеням разности $x - x^*$ имеет вид

$$\varphi(x) = \varphi(x_0) + \frac{1}{m!}\varphi^{(m)}(x^*)(x - x^*)^m + o((x - x^*)^m),$$

то для погрешностей $\varepsilon_n = x_n - x^*$ приближений в простой одношаговой итерации, когда x_n будут близки к x^*, верно соотношение

$$\varepsilon_{n+1} = \frac{1}{m!}\varphi^{(m)}(x^*)\varepsilon_n^m + o(\varepsilon_n^m),$$

и можно ожидать тем более быстрой сходимости x_n к x^*, чем больше m. Когда указанные условия не выполняются, например, когда $\varphi'(x^*) \neq 0$, то можно попытаться ускорить сходимость итерационного процесса, если заменить заданное уравнение $x = \varphi(x)$ другим уравнением $x = \Phi(x)$, в котором функция $\Phi(x)$ удовлетворяет следующим двум требованиям: 1) уравнение $x = \Phi(x)$ имеет те же решения x^*, что и заданное уравнение, и 2) для каждого из решений x^* выполняются условия $\Phi^{(k)}(x^*) = 0$ $(k = 1, 2, \ldots, m-1)$.

Такую функцию можно построить, очевидно, многими способами. Мы приведем один из известных способов.

Наряду с $\varphi(x)$ рассмотрим функцию $f(x) = \varphi(x) - x$. С помощью ее уравнение $x = \varphi(x)$ запишется как $f(x) = 0$. Будем искать $\Phi(x)$ в виде многочлена степени $m-1$ от f:

$$\Phi(x) = x + a_1(x)f(x) + a_2(x)f^2(x) + \ldots$$
$$\ldots + a_{m-1}(x)f^{m-1}(x). \qquad (14)$$

Свободный член многочлена взят равным x, так как уравнение $\Phi(x) = x$ должно быть равносильным $f(x) = 0$.

Выберем теперь коэффициенты $a_i(x)$ $(i = 1, \ldots$ $\ldots, m-1)$ так, чтобы производные $\Phi^{(p)}(x)$ $(p = 1, 2, \ldots, m-1)$ обращались в нуль всякий раз, когда x

есть решение уравнения $f(x)=0$. Эти требования дадут для $a_i(x)$ систему $m-1$ линейных уравнений

$$\begin{aligned}
&\Phi'(x)|_{f=0} = 1 + a_1(x) f'(x) = 0,\\
&\Phi''(x)|_{f=0} = 2a_1'(x) f'(x) + a_1(x) f''(x) + \\
&\qquad + 2a_2(x) f'^2(x) = 0,\\
&\Phi'''(x)|_{f=0} = a_1(x) f'''(x) + 3a_1'(x) f''(x) + \\
&\qquad + 3a_1''(x) f'(x) + 6a_2'(x) f'^2 + \\
&\qquad + 6a_2(x) f'(x) f''(x) + 6a_3(x) f'^3(x) = 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\Phi^{(m-1)}(x)|_{f=0} = (m-1)\, a_1^{(m-2)}(x) f'(x) + \ldots\\
&\qquad \ldots + a_1(x) f^{(m-1)}(x) + \ldots\\
&\qquad \ldots + (m-1)!\, a_{m-1}(x) [f'(x)]^{m-1} = 0.
\end{aligned} \tag{15}$$

Из нее последовательно могут быть найдены коэффициенты $a_1(x)$, $a_2(x)$, ..., $a_{m-1}(x)$.

При $m=2$ в системе сохранится только первое уравнение, и будет

$$a_1(x) = -\frac{1}{f'(x)}, \quad \Phi(x) = x - \frac{f(x)}{f'(x)} = \frac{x\varphi'(x) - \varphi(x)}{\varphi'(x_n) - 1}.$$

Соответствующий итерационный процесс имеет вид

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = \frac{x_n\varphi'(x_n) - \varphi(x_n)}{\varphi'(x) - 1}.$$

Как мы увидим ниже, он совпадает с процессом Ньютона, который будет получен на основе иных соображений.

Для $m=3$ система (15) даст

$$a_1(x) = -\frac{1}{f'(x)}, \quad a_2(x) = -\frac{f''(x)}{2f'^3(x)};$$

$$\begin{aligned}
\Phi(x) &= x - \frac{f(x)}{f'(x)} - \frac{f''(x) f^2(x)}{2f'^3(x)} = \\
&= \frac{x\varphi'(x) - \varphi(x)}{\varphi'(x) - 1} - \frac{\varphi''(x)[\varphi(x) - x]^2}{2[\varphi'(x) - 1]^3}.
\end{aligned}$$

Итерационный процесс примет вид

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{f''(x_n) f^2(x_n)}{2f'^3(x_n)}.$$

Заметим, наконец, что метод Стеффенсена (9) можно рассматривать как переход от уравнения $x = \varphi(x)$ к уравнению (10) с иным, правда, правилом составления функции $\Phi(x)$, чем (14).

§ 4. Метод итерации для системы уравнений

1. Описание метода простой одношаговой итерации. Пусть для нахождения значений численных неизвестных $x_1, x_2, \ldots, x_n$ задана система n уравнений. Применение к ее решению метода итерации требует приведения системы к виду

$$\begin{aligned} x_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ x_2 &= \varphi_2(x_1, x_2, \ldots, x_n), \\ &\cdots\cdots\cdots\cdots \\ x_n &= \varphi_n(x_1, x_2, \ldots, x_n). \end{aligned} \tag{1}$$

Для облегчения записи обычно рассматривают n-мерное числовое пространство R_n, элементами x которого являются упорядоченные совокупности n чисел $x = (x_1, x_2, \ldots, x_n)$. Один элемент x этого пространства будет служить для изображения значений аргументов $x_1, x_2, \ldots, x_n$ функций φ_i, второй же элемент, который мы обозначим φ, — для изображения соответствующих x значений функций $\varphi_1, \varphi_2, \ldots, \varphi_n$. Система (1) коротко запишется в виде

$$x = \varphi(x). \tag{2}$$

Зависимость $y = \varphi(x)$ можно рассматривать как отображение пространства R_n или части его в R_n. Уравнение (2) равносильно нахождению таких элементов R_n, которые переходят в себя, т. е. являются неподвижными при отображении.

Точное решение (2) обозначим $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ и предположим, что выбрано исходное приближение $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$. Последующие приближения нахо-

дятся по правилу

$$x^{k+1} = \varphi(x^k), \quad k = 0, 1, \ldots \tag{3}$$

Как и в предыдущем параграфе, мы ограничимся описанием наглядной приближенной картины поведения вектора погрешности $\varepsilon^k = x^k - x^* = (x_1^k - x_1^*, \ldots, x_n^k - x_n^*)$, когда приближение x^k будет близким к точному решению x^* и компоненты погрешности ε^k будут малыми величинами. Соотношение между ε^{k+1} и ε^k получится, если в (3) вместо x^k и x^{k+1} подставить их выражения $x^k = x^* + \varepsilon^k$ и $x^{k+1} = x^* + \varepsilon^{k+1}$:

$$x^* + \varepsilon^{k+1} = \varphi(x^* + \varepsilon^k). \tag{4}$$

Будем считать функции φ_i непрерывно дифференцируемыми в общей области их задания, и решение x^*, так же как и приближения x^k $(k = 0, 1, \ldots)$, лежащими внутри этой области. Соотношение (4) — векторное, мы выделим в нем равенство компонент номера i

$$x_i^* + \varepsilon_i^{k+1} = \varphi_i(x_1^* + \varepsilon_1^k, \ldots, x_n^* + \varepsilon_n^k).$$

Разложим правую часть по степеням ε_j^k $(j = 1, \ldots, n)$, выделив в разложении линейную часть. Если принять во внимание, что $\varphi_i(x_1^*, \ldots, x_n^*) = x_i^*$, получим

$$\varepsilon_i^{k+1} = \sum_{j=1}^{n} \frac{\partial}{\partial x_j} \varphi_i(x_1^*, \ldots, x_n^*)\, \varepsilon_j^k + o\left(\max_j |\varepsilon_j^k|\right),$$
$$i = 1, \ldots, n.$$

Отсюда следует, что вектор погрешности $\varepsilon^k = (\varepsilon_1^k, \ldots, \varepsilon_n^k)$ на одном шаге итерации испытывает линейное преобразование

$$\varepsilon^{k+1} \approx A\varepsilon^k. \tag{5}$$

Здесь A есть значение матрицы Якоби системы функций φ_i на точном решении x^*:

$$A = \begin{bmatrix} \left(\frac{\partial}{\partial x_1}\varphi_1\right)^* & \cdots & \left(\frac{\partial}{\partial x_n}\varphi_1\right)^* \\ \cdots\cdots\cdots\cdots & & \\ \left(\frac{\partial}{\partial x_1}\varphi_n\right)^* & \cdots & \left(\frac{\partial}{\partial x_n}\varphi_n\right)^* \end{bmatrix},$$

$$\left(\frac{\partial}{\partial x_j}\varphi_i\right)^* = \frac{\partial}{\partial x_j}\varphi_i(x_1^*, \ldots, x_n^*).$$

Равенство (5) позволяет наглядно истолковать закон изменения ε_k^i при итерации. Приведем матрицу A к каноническому виду Эрмита. Для упрощения записи предположим, что все элементарные делители A являются простыми, но заметим, что заключения, к которым мы придем ниже, остаются верными для любых элементарных делителей A. Пусть

$$A = S^{-1}[\lambda_1, \lambda_2, \ldots, \lambda_n] S.$$

Здесь $\lambda_1, \ldots, \lambda_n$ суть собственные значения A, и знаком $[\lambda_1, \ldots, \lambda_n]$ обозначена диагональная матрица с элементами $\lambda_1, \ldots, \lambda_n$. Соотношение (5) запишется в виде

$$\varepsilon^{k+1} \approx S^{-1}[\lambda_1, \ldots, \lambda_n] S\varepsilon^k$$

или, если ввести новый вектор η^k, положив $\eta^k = S\varepsilon^k$, то

$$\eta^{k+1} \approx [\lambda_1, \ldots, \lambda_n] \eta^k.$$

Это равносильно n численным равенствам

$$\eta_i^{k+1} \approx \lambda_i \eta_i^k, \quad i = 1, 2, \ldots, n.$$

Каждая величина η_i^k будет изменяться с увеличением k приблизительно по геометрической прогрессии, имеющей знаменателем λ_i. Когда все λ_i по модулю меньше единицы:

$$|\lambda_i| < 1, \quad i = 1, 2, \ldots, n,$$

то можно ожидать *), что $\eta_i^k \to 0$ при $k \to \infty$; так как $\varepsilon^k = S^{-1}\eta^k$, то ε^k также будет стремиться к нулю, и последовательность итерационных приближений x^k будет сходиться к точному решению x^*.

2. Аналог правила Зейделя. Возвратимся к записи системы в виде (1). Предположим, что вычисления доведены до приближения номера k: $x^k = (x_1^k, \ldots, x_n^k)$. В правиле Зейделя для нахождения следующего приближения $x^{k+1} = (x_1^{k+1}, \ldots, x_n^{k+1})$ должен быть прежде всего установлен порядок вычисления его компонен-

*) Теоремы о сходимости можно найти, например, в книге: В. И. Крылов и др., Вычислительные методы высшей математики, т. I, гл. 1, § 1.8, «Вышэйшая школа», Минск, 1972.

тов x_i^{k+1} $(i = 1, 2, \ldots, n)$. Такой порядок может быть своим для каждой системы и для каждого шага. Так как всякий порядок расположения x_i^{k+1} может быть приведен путем изменения нумерации к натуральному порядку x_1^{k+1}, x_2^{k+1}, ..., x_n^{k+1}, то правило Зейделя достаточно записать для этого последнего порядка:

$$\begin{aligned} x_1^{k+1} &= \varphi_1\left(x_1^k, x_2^k, \ldots, x_n^k\right), \\ x_2^{k+1} &= \varphi_2\left(x_1^{k+1}, x_2^k, \ldots, x_n^k\right), \\ &\ldots\ldots\ldots\ldots\ldots \\ x_n^{k+1} &= \varphi_n\left(x_1^{k+1}, \ldots, x_{n-1}^{k+1}, x_n^k\right). \end{aligned} \tag{6}$$

После вычисления x_i^{k+1} $(i = 1, \ldots, n)$ переходят к нахождению следующего приближения x^{k+2}: выбирают последовательность вычисления его компонентов x_i^{k+2} и выполняют счет при помощи равенств, аналогичных (6), и т. д.

§ 5. Метод Ньютона

Подобно методу итерации, метод Ньютона является общим и применимым к решению очень широкого класса нелинейных операторных уравнений. Значение его заключается в том, что он позволяет привести решение нелинейных уравнений к решению последовательности линейных задач. Достигается это при помощи выделения из нелинейного уравнения его главной линейной части.

1. Метод Ньютона для одного численного уравнения. Рассмотрим уравнение $f(x) = 0$, где x есть численная переменная и f — достаточно гладкая функция от x. Назовем x^* точное решение уравнения и предположим, что для x^* каким-либо путем указано исходное приближение x_0. Построим линейное уравнение для уточнения x_0.

Удобнее рассматривать не x_0, а погрешность $\varepsilon = x^* - x_0$ ввиду того, что x_0 стараются выбрать близким к x^*; тогда ε — оказывается малой величиной, что позволяет легко выделить из заданного уравнения главную часть.

Уравнение для ε сразу же получится, если в равенство $f(x^*)=0$ вместо x^* внести его значение $x^*=x_0+\varepsilon$:

$$f(x_0+\varepsilon)=0. \tag{1}$$

Разложим левую часть по степеням ε и выделим в разложении линейную часть, относя остальные члены в остаток:

$$f(x_0)+\varepsilon f'(x_0)+o(\varepsilon)=0.$$

Отбросив остаток $o(\varepsilon)$, получим линейное уравнение для ε, близкое к (1), если ε есть малая величина:

$$f(x_0)+\varepsilon f'(x_0)=0. \tag{2}$$

Решая его относительно ε, мы найдем лишь приближенное значение погрешности, которое обозначим ε_0. Можно ожидать, что ε_0 является главной частью погрешности ε по меньшей мере в том случае, когда ε есть малая величина. Добавляя ε_0 к x_0, получим улучшенное значение для решения $x_1=x_0+\varepsilon_0$, относительно которого можно ожидать, что оно будет ближе к x^*, чем x_0:

$$x_1=x_0-\frac{f(x_0)}{f'(x_0)}. \tag{3}$$

Пользуясь теми же соображениями, x_1 можно также улучшить и т. д. В результате получится последовательность приближений к x^*, в которой каждое следующее приближение будет находиться по правилу, принадлежащему Ньютону:

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}, \qquad n=0,\ 1,\ \ldots \tag{4}$$

Чтобы такие последовательные приближения могли быть построены, необходимо выполнение условий: 1) все x_n принадлежат области определения функции f и 2) $f'(x_n)\neq 0$ $(n=0,1,\ldots)$.

Геометрический смысл правила (4) весьма прост. В плоскости xy построим график l функции f. Точное решение x^* уравнения будет абсциссой точки пересечения l с осью Ox (рис. 8). Рассмотрим на l точку $M_n[x_n, f(x_n)]$ и проведем касательную прямую T_n к l в точке M_n. Уравнение касательной T_n есть $y-f(x_n)=$

$= f'(x_n)(x - x_n)$. Найдем точку пересечения касательной T_n с осью Ox и обозначим x_{n+1} абсциссу этой точки. Она должна быть найдена из уравнения $-f(x_n) = f'(x_n)(x_{n+1} - x_n)$, откуда следует $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, что совпадает с (4). Таким образом, правило Ньютона геометрически означает, что следующее приближение x_{n+1} находится, если линию l заменить прямой T_n, касающейся l в точке M_n.

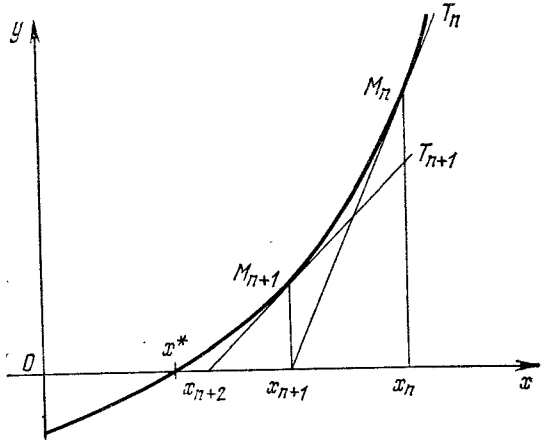

Рис. 8.

Выясним теперь наглядную приближенную картину поведения погрешности $\varepsilon_n = x_n - x^*$, когда приближение x_n будет близким к точному решению x^*, а ε_n — малой величиной. С целью получения соотношения между погрешностями ε_n и ε_{n+1} достаточно в (4) вместо x_n и x_{n+1} внести их выражения $x_n = x^* + \varepsilon_n$ и $x_{n+1} = x^* + \varepsilon_{n+1}$:

$$\varepsilon_{n+1} = \frac{\varepsilon_n f'(x^* + \varepsilon_n) - f(x^* + \varepsilon_n)}{f'(x^* + \varepsilon_n)}.$$

Для выделения главной части из правой части равенства можно воспользоваться следующими выражениями для f, f' и принять во внимание, что $f(x^*) = 0$:

$$f(x^* + \varepsilon_n) = \varepsilon_n f'(x^*) + \frac{1}{2}\varepsilon_n^2 f''(x^*) + o(\varepsilon_n^2),$$
$$f'(x^* + \varepsilon_n) = f'(x^*) + \varepsilon_n f''(x^*) + o(\varepsilon_n).$$

Поэтому

$$\varepsilon_{n+1} = \frac{1}{2}\frac{f''(x^*)}{f'(x^*)}\varepsilon_n^2 + o(\varepsilon_n^2). \tag{5}$$

Отбросив справа величину $o(\varepsilon^{2n})$, получим следующее простое приближенное равенство, дающее достаточно наглядное описание поведения погрешности ε_n:

$$\varepsilon_{n+1} \approx \frac{1}{2}\frac{f''(x^*)}{f'(x^*)}\varepsilon_n^2 = \alpha\varepsilon_n^2. \tag{6}$$

Отсюда сразу же получается, что $\varepsilon_{n+1}\varepsilon_n^{-2} \approx \alpha \approx \varepsilon_n\varepsilon_{n-1}^2$ и $\frac{\varepsilon_{n+1}}{\varepsilon_n} \approx \left(\frac{\varepsilon_n}{\varepsilon_{n-1}}\right)^2$, поэтому отношение смежных значений погрешностей на одном шаге вычислений приблизительно возводится в квадрат.

Так как нумерацию погрешностей можно начать с любого места, то мы вправе считать, что равенство $\varepsilon_i \approx \alpha\varepsilon_{i-1}^2$ выполняется для $i = 1, 2, \ldots$. При помощи этих равенств можно найти выражение ε_n через ε_0:

$$\varepsilon_n \approx (\alpha\varepsilon_0)^{2^n-1}\varepsilon_0 = \left[\frac{1}{2}\frac{f''(x^*)}{f'(x^*)}\varepsilon_0\right]^{2^n-1}\varepsilon_0. \tag{7}$$

Это равенство показывает, что когда $|\alpha\varepsilon_0| > 1$, то $|\varepsilon_n|$ будет увеличиваться при возрастании n, и трудно ожидать сходимости x_n к x^*. Когда же $|\alpha\varepsilon_0| < 1$, то весьма вероятно, что ε_n будет быстро стремиться к нулю и x_n — сходиться к x^*. Мы говорим не о достоверной, а только о вероятной сходимости на том основании, что соотношение получено путем отбрасывания в (5) малой величины $o(\varepsilon_n^2)$, и мы не оценили ее влияние на ε_n при возрастании n.

Приведем одну из простых теорем, которая дает условия, достаточные для существования решения уравнения и сходимости метода Ньютона.

Теорема 1. *Пусть для уравнения* $f(x) = 0$ *выполняются условия*:

1) *функция определена и дважды непрерывно дифференцируема на отрезке*

$$|x - x_0| \leqslant \delta, \tag{8}$$

при этом $|f''(x)| \leqslant K$ *для любых* x *на этом отрезке*;

2) $$f'(x_0) \neq 0 \quad \text{и} \quad \frac{1}{|f'(x_0)|} \leqslant B; \tag{9}$$

3) *в точке* x_0 *выполняется неравенство*

$$\left|\frac{f(x_0)}{f'(x_0)}\right| \leqslant \eta; \tag{10}$$

4) *для величин* δ, B, K, η *соблюдены условия*

$$h = BK\eta \leqslant \frac{1}{2}, \tag{11}$$

$$\frac{1-\sqrt{1-2h}}{h}\eta \leqslant \delta. \tag{12}$$

Тогда:

1) *последовательность* (4) *может быть построена и является сходящейся к некоторому значению* x^*, *принадлежащему отрезку* (8);

2) *предельное значение* x^* *есть решение заданного уравнения*;

3) *выполняется неравенство, характеризующее скорость сходимости,*

$$|x^* - x_n| \leqslant t^* - t_n, \tag{13}$$

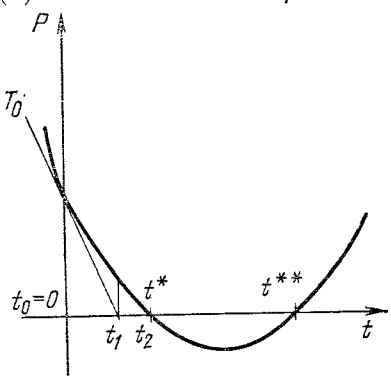

Рис. 9.

где t_n $(n = 0, 1, \ldots)$ *есть последовательность приближений к меньшему корню* t^* *вспомогательного уравнения*

$$P(t) = \frac{1}{2}Kt^2 - \frac{1}{B}t + \frac{\eta}{B} = 0, \tag{14}$$

построенная по правилу $t_{n+1} = t_n - \dfrac{P(t_n)}{P'(t_n)}$ *при* $t_0 = 0$.

Доказательство. Корни многочлена $P(t)$ имеют значения $t^* = \dfrac{1-\sqrt{1-2h}}{h}\eta$ и $t^{**} = \dfrac{1+\sqrt{1-2h}}{h}\eta$, оба при $0 < h \leqslant 1/2$ действительны и положительны. Графиком $P(t)$ является парабола, проходящая через точки t^*, t^{**} оси t и с осью симметрии, параллельной оси ординат (рис. 9). При первом взгляде на график становится ясным, что если построить ньютоновы

приближения t_n, начиная с $t_0 = 0$, то получится монотонно возрастающая последовательность, сходящаяся к t^*.

Обратимся теперь к последовательности x_n и покажем, что для $n = 0, 1, \ldots$ все x_n принадлежат области (8) и для них выполняется неравенство

$$|x_{n+1} - x_n| \leqslant t_{n+1} - t_n. \tag{15}$$

Воспользуемся методом индукции. Проверим (15) для $n = 0$. Ввиду того, что x_0 принадлежит (8) и $f'(x_0) \neq 0$, приближение $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$ может быть найдено. При $0 < h \leqslant \frac{1}{2}$ дробь

$$\frac{1-\sqrt{1-2h}}{h} = \frac{2}{1+\sqrt{1-2h}}$$

изменяется в области $(1, 2]$. Следовательно, выполняются оценки

$$|x_1 - x_0| = \left|\frac{f(x_0)}{f'(x_0)}\right| \leqslant \eta < \frac{1-\sqrt{1-2h}}{h}\eta \leqslant \delta.$$

Поэтому x_1 принадлежит (8). По предположению, $t_0 = 0$ и, стало быть, $t_1 = t_0 - \frac{P(t_0)}{P'(t_0)} = \eta$, $|t_1 - t_0| = \eta$, и ввиду $|x_1 - x_0| \leqslant \eta$ неравенство (15) для $n = 0$ выполняется.

Предположим теперь, что $x_0, x_1, \ldots, x_n$ принадлежат области (8) и для них

$$|x_{k+1} - x_k| \leqslant t_{k+1} - t_k \quad (k = 0, 1, \ldots, n-1).$$

Для проверки возможности построения x_{n+1} достаточно показать, что $f'(x_n) \neq 0$. Рассмотрим неравенства

$$|f'(x_n)| = \left|f'(x_0) + \int_{x_0}^{x_n} f''(t)\,dt\right| \geqslant \frac{1}{B} - K|x_n - x_0| \geqslant$$

$$\geqslant \frac{1}{B} - K|(x_n - x_{n-1}) + \ldots + (x_1 - x_0)| \geqslant$$

$$\geqslant \frac{1}{B} - K[(t_n - t_{n-1}) + \ldots + (t_1 - t_0)] =$$

$$= \frac{1}{B} - K(t_n - t_0) = \frac{1}{B} - Kt_n = -P'(t_n).$$

Так как $t_n < t^*$, то (см. рис. 9) $P'(t_n) < 0$ и, стало быть, $|f'(x_n)| > 0$.

Оценим величину $f(x_n)$. Ввиду $x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$ будет

$$f(x_n) = f(x_{n-1}) + (x_n - x_{n-1}) f'(x_{n-1}) + \int_{x_{n-1}}^{x_n} (x_n - t) f''(t)\, dt =$$

$$= \int_{x_{n-1}}^{x_n} (x_n - t) f''(t)\, dt.$$

$$|f(x_n)| \leqslant K \left| \int_{x_{n-1}}^{x_n} (x_n - t)\, dt \right| =$$

$$= \frac{1}{2} K (x_n - x_{n-1})^2 \leqslant \frac{1}{2} K (t_n - t_{n-1})^2.$$

Сходные вычисления, проведенные для многочлена $P(t)$, показывают, что $P(t_n) = \frac{1}{2} K (t_n - t_{n-1})^2$. Значит, $|f(x_n)| \leqslant P(t_n)$ и

$$|x_{n+1} - x_n| = \left| \frac{f(x_n)}{f'(x_n)} \right| \leqslant -\frac{P(t_n)}{P'(t_n)} = t_{n+1} - t_n.$$

Этим неравенство (15) доказано. Осталось лишь проверить, что x_{n+1} принадлежит области (5):

$$|x_{n+1} - x_0| = |(x_{n+1} - x_n) + (x_n - x_{n-1}) + \ldots + (x_1 - x_0)| \leqslant$$

$$\leqslant (t_{n+1} - t_n) + (t_n - t_{n-1}) + \ldots + (t_1 - t_0) =$$

$$= t_{n+1} - t_0 = t_{n+1} < t^* = \frac{1 - \sqrt{1 - 2h}}{h} \eta \leqslant \delta$$

и x_{n+1} действительно лежит внутри (8).

Проверим, что для x_n выполняется признак сходимости Больцано — Коши

$$|x_{n+p} - x_n| = |(x_{n+p} - x_{n+p-1}) + \ldots + (x_{n+1} - x_n)| \leqslant$$

$$\leqslant (t_{n+p} - t_{n+p-1}) + \ldots + (t_{n+1} - t_n) = t_{n+p} - t_n.$$

Последовательность t_n сходится, и для нее признак Больцано — Коши выполняется; из полученного же неравенства следует, что он будет выполняться и для

последовательности x_n, и эта последовательность также будет сходящейся. Предел x_n обозначим x^*: $\lim x_n = x^*$.

Утверждение теоремы о скорости сходимости t_n к t^* легко получается, если в неравенстве $|x_{n+p} - x_n| \leqslant t_{n+p} - t_n$ перейти к пределу при $p \to \infty$. Ввиду $x_{n+p} \to x^*$ и $t_{n+p} \to t^*$ в пределе получится (13).

Отметим попутно, что в условиях теоремы оценка (13) является неулучшаемой, так как знак неравенства в ней переходит в знак равенства, когда $f(x) = P(x)$. Но она не является наглядной, так как требует вычисления величины $t^* - t_n$. Ниже (13) будет заменено другим неравенством, более наглядным, но менее точным.

Осталось показать, что $x^* = \lim x_n$ есть решение уравнения $f(x) = 0$. Пусть в равенстве $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ номер n неограниченно возрастает. Так как при этом $x_{n+1} \to x^*$, $x_n \to x^*$, то, значит, $\frac{f(x_n)}{f'(x_n)} \to 0$. Ввиду того, что $f'(x_n)$ есть ограниченная величина, отсюда вытекает, что $f(x_n) \to 0$ и, так как $\lim x_n = x^*$ принадлежит отрезку (8), где f есть непрерывная функция, то в пределе получится $f(x^*) = 0$.

Теорема 2. *Если выполняются условия теоремы* 1, *то для погрешности приближений верна оценка*

$$|x^* - x_n| \leqslant 2^{-n+1} (2h)^{2^n - 1} \eta. \tag{16}$$

Доказательство. Для упрощения записи в многочлене (14) заменим аргумент t, положив $t = \eta\tau$:

$$P(t) = \frac{\eta}{B}\left(\frac{1}{2} h\tau^2 - \tau + 1\right) = \frac{\eta}{B} \varphi(\tau). \tag{17}$$

Для уравнения $\varphi(\tau) = 0$ рассмотрим правило Ньютона и возьмем последовательность приближений к меньшему корню $\tau^* = \frac{1 - \sqrt{1 - 2h}}{h}$:

$$\tau_0 = 0, \quad \tau_{n+1} = \tau_n - \frac{\varphi(\tau_n)}{\varphi'(\tau_n)}, \quad n = 0, 1, \ldots$$

С последовательностью t_n она связана, очевидно, соотношением $t_n = \eta\tau_n$. Для доказательства теоремы достаточно установить неравенство

$$\tau^* - \tau_n \leqslant 2^{-n+1} (2h)^{2^n - 1}. \tag{18}$$

Воспользовавшись $\varphi(\tau^*)=0$, получим

$$\tau^*-\tau_n=\tau^*-\tau_{n-1}+\frac{\varphi(\tau_{n-1})}{\varphi'(\tau_{n-1})}=$$

$$=-\frac{1}{\varphi'(\tau_{n-1})}[\varphi(\tau^*)-\varphi(\tau_{n-1})-(\tau^*-\tau_{n-1})\varphi'(\tau_{n-1})].$$

Но

$$\varphi(\tau^*)=\varphi[\tau_{n-1}+(\tau^*-\tau_{n-1})]=$$

$$=\varphi(\tau_{n-1})+(\tau^*-\tau_{n-1})\varphi(\tau_{n-1})+\frac{1}{2}\varphi''(\xi)(\tau^*-\tau_{n-1})^2,$$

$$\varphi'(\tau_{n-1})=h\tau_{n-1}-1,\quad \varphi''(\tau)=h,$$

и поэтому

$$\tau^*-\tau_n=\frac{1}{1-h\tau_{n-1}}\cdot\frac{1}{2}h(\tau^*-\tau_{n-1})^2. \tag{19}$$

Для оценки первого множителя в правой части воспользуемся формулой Тейлора для $\varphi(\tau_n)$

$$\varphi(\tau_n)=\varphi[\tau_{n-1}+(\tau_n-\tau_{n-1})]=$$

$$=\varphi(\tau_{n-1})+(\tau_n-\tau_{n-1})\varphi'(\tau_{n-1})+\frac{1}{2}\varphi''(\xi)(\tau_n-\tau_{n-1})^2$$

и, так как $\varphi(\tau_n)+(\tau_n-\tau_{n-1})\varphi'(\tau_{n-1})=0$, а $\varphi''(\xi)=h$, получим

$$\tau_{n+1}-\tau_n=-\frac{\varphi(\tau_n)}{\varphi'(\tau_n)}=\frac{1}{2}\frac{h}{1-h\tau_n}(\tau_n-\tau_{n-1})^2.$$

При $n=1$, ввиду $\tau_0=0$, $\tau_1=1$ и $h\leqslant 1/2$, отсюда получим

$$\tau_2-\tau_1=\frac{1}{2}\frac{h}{1-h}\leqslant\frac{1}{2},$$

$$\tau_2=\tau_1+(\tau_2-\tau_1)\leqslant 1+\frac{1}{2}=\frac{3}{2}.$$

Для $n=2$

$$\tau_3-\tau_2=\frac{1}{2}\frac{h}{1-h\tau_2}(\tau_2-\tau_1)^2\leqslant\frac{1}{2}\frac{h}{1-\frac{3}{2}h}\frac{1}{2^2}\leqslant\frac{1}{4},$$

$$\tau_3=\tau_2+(\tau_3-\tau_2)\leqslant\frac{3}{2}+\frac{1}{4}=\frac{7}{4}.$$

Если продолжить эти оценки дальше, найдем

$$\tau_{n-1}\leqslant 2-2^{2-n},\quad 1-h\tau_{n-1}\geqslant 1-1/2(2-2^{2-n})=2^{1-n},$$

что позволяет неравенство (19) заменить следующим:

$$\tau^* - \tau_n \leqslant 2^{n-2} h (\tau^* - \tau_{n-1})^2. \qquad (20)$$

При $n = 1$, ввиду $\tau_0 = 0$, $h \leqslant \frac{1}{2}$ и $\tau^* = \frac{1 - \sqrt{1-2h}}{h} = \frac{2}{1 + \sqrt{1-2h}} \leqslant 2$, отсюда следует $\tau^* - \tau_1 \leqslant 2^{-1} h 2^2 = 2h$, и нужное нам неравенство (18) для $n = 1$ действительно верно. Выполним несложную индукцию и допустим, что для значения $n = k - 1$ неравенство (18) является верным. Если применить (20) для $n = k$, то получится

$$\tau^* - \tau_k \leqslant 2^{k-2} h \left[2^{-k+2} (2h)^{2^{k-1}-1}\right]^2 = 2^{-k+1} (2h)^{2^k - 1}.$$

Это доказывает выполнение (18) для значения $n = k$.

2. Некоторые видоизменения метода Ньютона. Этот метод является одним из старейших вычислительных методов решения уравнений, и ввиду его большого значения было сделано много попыток его изменения с целью уточнения или упрощения вычислений. Ниже мы остановимся на трех простейших вопросах такого типа.

1) Метод секущих. В формуле Ньютона (4) на каждом шаге вычисляются три величины: x_n, $f(x_n)$ и $f'(x_n)$; при этом главная часть труда затрачивается на нахождение $f(x_n)$ и $f'(x_n)$. Можно уменьшить вычислительную работу, отказавшись частично или полностью от вычисления одной из этих величин. Так как по значению $f(x_n)$ можно точнее судить о значении погрешности $x_n - x^* = \varepsilon_n$, чем по $f'(x_n)$, то естественно отказаться от вычисления $f'(x_n)$ и исключить эту величину из вычислительного процесса. Сделать это можно, заменив, например, в формуле (4) $f'(x_n)$ любым приближенным выражением $\tilde{f}'$ через значения функции f. Мы остановимся только на самой простой замене, когда значение $f'(x_n)$ вычисляется по двум парам чисел $[x_{n-1}, f(x_{n-1})]$ и $[x_n, f(x_n)]$:

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}.$$

После замены формула (4) перейдет в следующую

формулу:

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1})\, f(x_n)}{f(x_n) - f(x_{n-1})} = \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}. \quad (21)$$

Она называется формулой секущих, и причина такого названия легко может быть объяснена. На графике l функции f возьмем две точки $M_{n-1}[x_{n-1}, f(x_{n-1})]$ и $M_n[x_n, f(x_n)]$ и проведем через них секущую прямую $S_{n,\,n-1}$ (рис. 10). Уравнение ее есть

$$\frac{x - x_n}{x_{n-1} - x_n} = \frac{y - f(x_n)}{f(x_{n-1}) - f(x_n)}.$$

Формула (21) означает, что за x_{n+1} берется абсцисса точки пересечения секущей $S_{n,\,n-1}$ с осью x; иначе

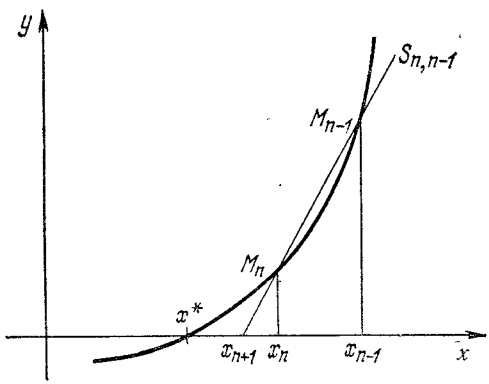

Рис. 10.

говоря, при нахождении следующего приближения x_{n+1} линия l заменяется секущей $S_{n,\,n-1}$.

Метод секущих является двухшаговым, и нахождение следующего значения x_{n+1} в нем требует знания двух предыдущих значений x_n, x_{n-1}. В частности, начало расчета требует знания двух начальных значений x_0, x_1.

Выясним еще закон изменения погрешности $\varepsilon_n = x_n - x^*$ вблизи решения x^*. Это можно просто сделать*), если в равенство (21) внести вместо x_{n-1}, x_n,

*) Полученное ниже соотношение (22) является, по существу, другой записью равенства (3.2).

$f(x_n)$ и $f(x_{n-1})$ их выражения через погрешности $\varepsilon_{n-1}, \varepsilon_n$:

$$x_{n-1} = x^* + \varepsilon_{n-1}, \quad x_n = x^* + \varepsilon_n,$$

$$f(x_{n-1}) = f(x^* + \varepsilon_{n-1}) = \varepsilon_{n-1} f'(x^*) + \frac{1}{2}\varepsilon_{n-1}^2 f''(x^*) + \ldots,$$

$$f(x_n) = f(x^* + \varepsilon_n) = \varepsilon_n f'(x^*) + \frac{1}{2}\varepsilon_n^2 f''(x^*) + \ldots,$$

а затем выделить из результата главный член, имеющий наименьшую размерность относительно ε_{n-1} и ε_n:

$$\varepsilon_{n+1} = \varepsilon_n - \frac{(\varepsilon_n - \varepsilon_{n-1})\left[\varepsilon_n f'(x^*) + 1/2\varepsilon_n^2 f''(x^*) + \ldots\right]}{(\varepsilon_n - \varepsilon_{n-1}) f'(x^*) + 1/2\left(\varepsilon_n^2 - \varepsilon_{n-1}^2\right) f''(x^*) + \ldots},$$

$$\varepsilon_{n+1} \approx \frac{1}{2}\frac{f''(x^*)}{f'(x^*)}\varepsilon_n \varepsilon_{n-1}. \tag{22}$$

Если сравнить полученное равенство с аналогичным для основного метода Ньютона (6), то можно видеть, что погрешность ε_n в методе секущих убывает по закону, близкому к убыванию в основном методе Ньютона, но только с несколько меньшей скоростью стремления к нулю.

2) Видоизменение правила Ньютона с постоянным значением производной. Для уменьшения вычислений можно не находить производную в каждой точке x_n, а воспользоваться только одним ее начальным значением и вычислять приближения по правилу

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)} \qquad (n = 0, 1, \ldots). \tag{23}$$

Геометрически это означает, что за x_{n+1} принимают координату точки персечения с осью x прямой линии, проведенной через $M_n[x_n, f(x_n)]$ на l, параллельно касательной T_0 к l в точке $M_0[x_0, f(x_0)]$ (рис. 11).

Заменим в (23) x_n, x_{n+1} их выражениями через погрешности $x_n = x^* + \varepsilon_n$, $x_{n+1} = x^* + \varepsilon_{n+1}$ и воспользуемся разложением

$$f(x_n) = f(x^* + \varepsilon_n) = \varepsilon_n f'(x^*) + \frac{1}{2}\varepsilon_n^2 f''(x^*) + \ldots$$

Если считать, что x_n и x_{n+1} близки к x^*, и погрешности ε_n, ε_{n+1} являются малыми величинами, то после того,

как в результате замены будут сохранены в равенстве члены первой степени с ε_n, ε_{n+1} и отброшены члены с более высоким порядком малости, получится следующее сотношение, дающее приближенное описание закона изменения погрешности вблизи точного решения:

$$\varepsilon_{n+1} \approx \varepsilon_n \left[1 - \frac{f'(x^*)}{f'(x_0)}\right]. \qquad (24)$$

Закон изменения ε_n будет близок к геометрической прогрессии со знаменателем $q = 1 - \frac{f'(x^*)}{f'(x_0)}$. Относительно q заметим, что x_0 берется обычно близким к x^*, поэтому $f'(x_0)$ будет мало отличаться от $f'(x^*)$, отношение

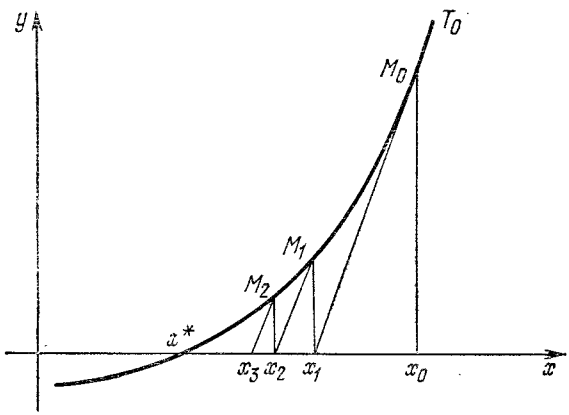

Рис. 11.

$f'(x^*)/f'(x_0)$ — мало отличаться от единицы и q будет иметь небольшое значение. Можно ожидать, следовательно, что если x_0 взято достаточно близко к решению x^*, то x_n будет сходиться к x^*, но скорость сходимости будет более медленной, чем для формулы Ньютона (4).

3) Уточнение правила Ньютона для случая кратного корня уравнения. При описании правила Ньютона и его видоизменений всегда предполагалось, что первая производная f' отлична от нуля как на самом решении x^*, так и в некоторой его окрестности. Теперь нашей целью будет выяснить, как

следует изменить формулу (4) для случая кратного корня уравнения.

Предположим, что кратность решения x^* равна $m>1$ и тейлорово разложение f вблизи x^* имеет вид

$$f(x)=a_m(x-x^*)^m+\ldots+a_{m+p}(x-x^*)^{m+p}+R_{m+p}, \quad (25)$$

$$a_k=\frac{1}{k!}f^{(k)}(x^*), \quad a_m\neq 0.$$

Как и выше, предположим, что x_n лежит вблизи x^* и погрешность $\varepsilon_n=x_n-x^*$ есть малая величина. Формула (4) дает следующую связь между ε_n и ε_{n+1}:

$$\varepsilon_{n+1}=\varepsilon_n-\frac{f(x^*+\varepsilon_n)}{f'(x^*+\varepsilon_n)}.$$

Представление (25) дает приводимые ниже разложения для f, f' и $[f']^{-1}$ по степеням ε_n:

$$f(x^*+\varepsilon_n)=a_m\varepsilon^m+a_{m+1}\varepsilon^{m+1}+\ldots,$$

$$f'(x^*+\varepsilon_n)=ma_m\varepsilon_n^{m-1}+(m+1)a_{m+1}\varepsilon_n^m+\ldots,$$

$$[f'(x^*+\varepsilon_n)]^{-1}=\frac{1}{ma_m}\varepsilon_n^{-m+1}\left[1-\frac{(m+1)a_{m+1}}{ma_m}\varepsilon_n+\ldots\right].$$

Значит,

$$\frac{f(x^*+\varepsilon_n)}{f'(x^*+\varepsilon_n)}=\frac{1}{m}\varepsilon_n\left[1-\frac{a_{m+1}}{ma_m}\varepsilon_n+\ldots\right],$$

$$\varepsilon_{n+1}=\left(1-\frac{1}{m}\right)\varepsilon_n+\frac{a_{m+1}}{m^2a_m}\varepsilon_n^2+\ldots, \quad (26)$$

$$\varepsilon_{n+1}\approx\left(1-\frac{1}{m}\right)\varepsilon_n. \quad (27)$$

Как видно отсюда, погрешность ε_n вблизи решения x^* будет с ростом n изменяться приблизительно по закону геометрической прогрессии со знаменателем $q=1-\frac{1}{m}$, меньшим единицы. Убывание ε_n будет значительно медленнее, чем в случае, когда производная $f'(x^*)$ отлична от нуля и где закон убывания описывается равенством (6).

Изменим правило вычислений и возьмем его в форме

$$x_{n+1}=x_n-k\frac{f(x_n)}{f'(x_n)}, \quad (28)$$

где k есть численный параметр, выбором которого сейчас займемся. Погрешности, соответствующие этому правилу, будут изменяться по закону

$$\varepsilon_{n+1} = \left(1 - \frac{k}{m}\right)\varepsilon_n + k\frac{a_{m+1}}{m^2 a_m}\varepsilon_n^2 + \ldots$$

Достаточно положить $k = m$, чтобы справа обратить в нуль линейный член и увеличить скорость убывания погрешности ε_n. Формула вычислений тогда будет такой:

$$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)}. \tag{29}$$

Соответствующий ему закон изменения ε_n запишется в виде приближенного равенства

$$\varepsilon_{n+1} \approx \frac{a_{m+1}}{m a_m}\varepsilon_i^2 = \frac{f^{(m+1)}(x^*)}{m(m+1)f^{(m)}(x^*)}\varepsilon_n^2.$$

Он сходен с законом (6) изменения погрешности для основной формулы Ньютона.

3. Метод Ньютона для системы уравнений. Предположим, что значения ν численных неизвестных $x_1, x_2, \ldots, x_\nu$ должны быть найдены из системы ν уравнений

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_\nu) &= 0, \\ \ldots\ldots\ldots\ldots & \\ f_\nu(x_1, x_2, \ldots, x_\nu) &= 0. \end{aligned} \tag{30}$$

Рассмотрим ν-мерное векторное пространство R_ν, и совокупности значений неизвестных будем рассматривать как элементы R_ν: $x = (x_1, x_2, \ldots, x_\nu)$. Каждую функцию $f_i(x_1, \ldots, x_\nu)$ коротко можно обозначить $f_i(x)$. Если, кроме того, ввести вектор-функцию $f(x) = [f_1(x), \ldots, f_\nu(x)]$, то система (30) запишется в форме

$$f(x) = 0. \tag{31}$$

Допустим, что известно некоторое исходное приближение $x^0 = (x_1^0, \ldots, x_\nu^0)$ к решению системы $x^* = (x_1^*, \ldots, x_\nu^*)$. Для выделения главной части из системы (31) удобнее рассматривать не точное решение x^*, а вектор-погрешность $x^* - x^0 = (x_1^* - x_1^0, \ldots, x_\nu^* - x_\nu^0) = \varepsilon = (\varepsilon_1, \ldots, \varepsilon_\nu)$.

Уравнение для ε получится, если в равенстве $f(x^*) = 0$ заменить x^* на $x^* = x^0 + \varepsilon$:

$$f(x^0 + \varepsilon) = 0. \tag{32}$$

Предполагая все составляющие вектора-погрешности малыми величинами, выделим в системе (30) главную линейную часть. Для этого рассмотрим уравнение любого номера $f_i(x^0 + \varepsilon) = 0$, разложим функцию f_i с помощью формулы Тейлора по степеням погрешностей $\varepsilon_1, \ldots, \varepsilon_\nu$ и сохраним в разложении линейную часть, отбросив все члены более высокого измерения. После этого получится линейная система уравнений относительно погрешностей, приближенно заменяющая систему (30):

$$\sum_{j=1}^{\nu} \varepsilon_j \frac{\partial}{\partial x_j} f_i(x^0) \approx -f_i(x^0), \qquad i = 1, 2, \ldots, \nu. \tag{33}$$

Решая ее относительно ε_i, мы не найдем точных значений погрешностей, а получим только приближенные значения для них, и можно ожидать, что они будут главными частями ε_i. Назовем их ε_i^0. Мы улучшим исходные значения неизвестных x_i^0, если прибавим к ним ε_i^0:

$$x_1^1 = x_1^0 + \varepsilon_1^0, \quad x_2^1 = x_2^0 + \varepsilon_2^0, \quad \ldots, \quad x_\nu^1 = x_\nu^0 + \varepsilon_\nu^0.$$

Новые приближенные значения x_i^1 аналогичными вычислениями в свою очередь могут быть улучшены и т. д.

В результате для каждого значения x_i^* получится последовательность приближений x_i^n такая, что каждое следующее приближение x_i^{n+1} ($i = 1, \ldots, \nu$) будет находиться из линейной системы по предыдущему приближению x_i^n:

$$\sum_{j=1}^{\nu} (x_j^{n+1} - x_j^n) \frac{\partial}{\partial x_j} f_i(x^n) = -f_i(x^n); \tag{34}$$

$$i = 1, \ldots, \nu, \quad n = 0, 1, \ldots$$

Рассмотрим матрицу Якоби системы функций f_i $(i = 1, \ldots, \nu)$

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_\nu} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_\nu} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_\nu}{\partial x_1} & \frac{\partial f_\nu}{\partial x_2} & \cdots & \frac{\partial f_\nu}{\partial x_\nu} \end{bmatrix} = f'(x). \tag{35}$$

Значение ее при $x = x^n$ есть матрица системы (34). Система будет иметь единственное решение, когда ее определитель отличен от нуля: $D[f'(x^n)] \neq 0$.

Как и во многих предшествующих случаях выясним приближенную наглядную картину поведения погрешностей $\varepsilon_i^{*n} = x_i^n - x_i^*$, когда x^n мало отличается от точного решения x^*, и ε_i^{*n} являются малыми величинами. Соотношение между ε_i^{*n} и ε_i^{*n+1} получается, если в (34) заменить x^n и x^{n+1} их выражениями $x^n = x^* + \varepsilon^{*n}$ и $x^{n+1} = x^* + \varepsilon^{*n+1}$. Считая функции f_i дважды непрерывно дифференцируемыми, примем при замене во внимание следующие равенства:

$$x_i^{n+1} - x_i^n = \varepsilon_i^{*n+1} - \varepsilon_i^{*n}, \quad f_i(x^*) = 0,$$

$$f_i(x^n) = f_i(x^* + \varepsilon^{*n}) = \sum_{j=1}^{\nu} \varepsilon_j^{*n} \frac{\partial}{\partial x_j} f_i(x^*) + \frac{1}{2} \sum_{j,k=1}^{\nu} \varepsilon_j^{*n} \varepsilon_k^{*n} \frac{\partial^2}{\partial x_j \partial x_k} f_i(x^*) + \ldots,$$

$$\frac{\partial}{\partial x_j} f_i(x^n) = \frac{\partial}{\partial x_j} f_i(x^* + \varepsilon^{*n}) = \frac{\partial}{\partial x_j} f_i(x^*) + \sum_{k=1}^{\nu} \varepsilon_k^{*n} \frac{\partial^2}{\partial x_j \partial x_k} f_i(x^*) + \ldots$$

Пользуясь предположением о малости погрешностей, сохраним в результате замены только главные члены и

после этого получим

$$\sum_{j=1}^{\nu} \varepsilon_j^{*n+1} \frac{\partial}{\partial x_j} f_i(x^*) \approx \frac{1}{2} \sum_{j,\,k=1}^{\nu} \varepsilon_j^{*n} \varepsilon_k^{*n} \frac{\partial^2}{\partial x_j \partial x_k} f_i(x^*),$$

$$i = 1, \ldots, \nu.$$

Если воспользоваться вектор-функцией $f(x)$ и матрицей Якоби $f'(x)$, полученную систему можно записать в виде

$$f'(x^*)\varepsilon^{*n+1} \approx \frac{1}{2} \sum_{j,\,k=1}^{\nu} \varepsilon_j^{*n} \varepsilon_k^{*n} \frac{\partial^2}{\partial x_j \partial x_k} f(x^*). \qquad (36)$$

Если матрица Якоби $f'(x^*)$ является неособенной, отсюда можно найти ε^{*n+1}:

$$\varepsilon^{*n+1} \approx \frac{1}{2} \sum_{j,\,k=1}^{\nu} \varepsilon_j^{*n} \varepsilon_k^{*n} [f'(x^*)]^{-1} \frac{\partial^2}{\partial x_j \partial x_k} f(x^*). \qquad (37)$$

Это равенство показывает, что погрешности ε_i^{*n+1} следующего приближенного значения x^{n+1} будут приблизительно квадратично зависеть от погрешностей ε_j^{*n} предыдущего приближения x^n. Это обстоятельство позволяет ожидать, что если определитель $D[f'(x)]$ матрицы Якоби отличен от нуля в некоторой окрестности решения x^* системы и если исходное приближение x^0 взято достаточно близким к x^*, то последовательность x^n будет сходиться к решению x^*; при этом сходимость $x^n \to x^*$ будет быстрой, так как с возрастанием n погрешность ε^{*n} будет убывать по квадратичному закону (37).

Во всех приведенных выше рассуждениях не было изучено влияние отбрасываемых членов более высокого порядка малости, поэтому высказанное заключение, как и аналогичные предыдущие, может иметь только ориентировочное значение.

§ 6. Интерполяционные методы решения уравнений

1. О построении интерполяционных методов. Такие методы можно рассматривать, в определенной степени, как уточнения и обобщения методов итерации и Нью-

тона. О связи интерполирования с итерационным методом кратко говорилось в § 3. Несколько более подробно мы остановимся сейчас на связи интерполирования с методом Ньютона. Для пояснения идеи построения интерполяционных методов достаточно ограничиться этим случаем.

В основном методе Ньютона вычисляются значения трех величин x_n, $f(x_n)$, $f'(x_n)$. Метод Ньютона — одношаговый, и для вычисления следующей строки таблицы, содержащей x_{n+1}, $f(x_{n+1})$, $f'(x_{n+1})$, используется лишь одна предыдущая строка с x_n, $f(x_n)$, $f'(x_n)$; при этом для нахождения x_{n+1} выполняется линейное интерполирование $f(x)$ по значениям $f(x_n)$, $f'(x_n)$, и за x_{n+1} принимается нуль линейной интерполирующей функции.

Можно рассчитывать увеличить точность следующего приближения x_{n+1}, повышая степень интерполирующего многочлена; сделать же это можно, например, двумя следующими путями.

Во-первых, можно в уже найденных приближениях x_n, x_{n-1}, ... вычислять не только f и f', но находить также значения производных более высокого порядка f'', f''', ... и привлекать их к интерполированию f. Во-вторых, можно для интерполирования f привлекать эти значения не только в точке x_n, но и в нескольких предшествующих точках x_{n-1}, ..., x_{n-k}.

Пусть избран какой-либо тип интерполирования f и построен для нахождения x_{n+1} соответствующий интерполирующий многочлен $P_m(x)$ некоторой степени $m>1$. После этого заданное уравнение $f(x)=0$ заменяется приближенным уравнением

$$P_m(x)=0. \tag{1}$$

Оно требует некоторых пояснений. Так как $m>1$, уравнение будет иметь, как правило, более одного решения. В практике вычислений наиболее часто бывает, что приближения x_n, начиная с какого-то места, будут лежать на некотором малом отрезке $[\alpha, \beta]$, содержащем точное решение x^*. На $[\alpha, \beta]$ многочлен $P_m(x)$ будет близким к функции $f(x)$, и графики f и P_m будут мало отличаться между собой. В наиболее распространенном случае f имеет на $[\alpha, \beta]$ один корень, являющийся простым. Очень часто бывает, что $P_m(x)$ имеет на $[\alpha, \beta]$

также один простой корень. Все другие корни P_m лежат вне $[\alpha, \beta]$, не близко от этого отрезка, и не имеют для вычислений интереса. Вычисляют лишь корень P_m, ближайший к x_n.

Нахождение такого корня затруднено тем обстоятельством, что уравнение (1) является нелинейным. Оно решается обычно путем построения последовательных приближений к разыскиваемому корню P_m. Так как он близок к x_n, то значение x_n может быть принято за исходное приближение к нему. Все следующие приближения находятся из линейного уравнения, которое строится путем выделения из $P_m(x)$ главной линейной части, в предположении, что x близок к x_n. Форма этой главной части зависит от способа интерполирования f; для пояснения рассмотрим следующий случай: предположим, что в интерполировании f участвовали величины $f(x_n)$ и $f'(x_n)$. Тогда главной линейной частью $P_m(x)$ при x, близком к x_n, будет, очевидно, $f(x_n)+(x-x_n)f'(x_n)$, и уравнение (1) приведется к виду

$$P_m(x) = f(x_n) + (x - x_n) f'(x_n) + F_m(x) = 0, \qquad (2)$$
$$F_m(x) = P_m(x) - f(x_n) - (x - x_n) f'(x_n).$$

Если решить уравнение (2) относительно x, содержащегося в выделенной линейной части, то оно приводится к следующему каноническому для итерации виду

$$x = x_n - \frac{1}{f'(x_n)}[f(x_n) + F_m(x)] = \varphi(x).$$

Решают его обычно при помощи какого-либо итерационного метода. Найденное значение для x принимают за x_{n+1}, по нему находят нужные значения функции f и производных от нее и переходят к вычислению x_{n+2}.

Остановим еще внимание на методе, основанном на интерполировании обратной функции. Как будет видно из изложения, он не требует для нахождения приближений решения уравнений, и в этом состоит его преимущество по сравнению с методом, основанным на интерполировании f.

Поясним идею метода на простом примере. Пусть решается уравнение $f(x) = 0$, и пусть при вычислениях составляется таблица приближений x_n к решению и со-

ответствующих значений функции $y = f(x)$ и производных от нее:

$$\begin{array}{l} x_0,\ x_1,\ \ldots,\ x_n, \\ y_0,\ y_1,\ \ldots,\ y_n, \\ y_0',\ y_1',\ \ldots,\ y_n', \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \end{array} \qquad [y_k = f(x_k)]. \tag{3}$$

Обычно бывает, что на некотором отрезке вблизи решения x^* функция f изменяется монотонно, и зависимость $y = f(x)$ может быть обращена. Обратную функцию обозначим $x = F(y)$. Когда решается уравнение $f(x) = 0$, то для обратной функции это означает, что нужно найти значение $x = F(y)$, отвечающее значению $y = 0$. При помощи чисел, имеющихся в таблице (3), это может быть сделано путем известных средств интерполирования. В рассматриваемом случае это есть задача интерполирования $x^* = F(0)$ по значениям функции F и производных от нее*) в точках y_n, y_{n-1}, ...

Пусть выбрано какое-либо интерполяционное правило, и при его помощи составлен интерполяционный многочлен $\Pi_m(y)$ степени m. Полагая в нем $y = 0$, найдем приближенное значение $F(0) \approx \Pi_m(0)$, которое примем за x_{n+1}:

$$x_{n+1} = \Pi_m(0).$$

После этого вычисляем значения f и нужных производных, дополняем таблицу (3) столбцом номера $n+1$ и переходим к нахождению x_{n+2}.

Остановимся теперь немного более подробно на простейших задачах в каждом из указанных методов.

2. Метод, основанный на интерполировании функции. Рассмотрим наиболее простую форму этого метода, когда для интерполирования f используются только значения самой функции f. Предположим, что вычисления доведены до шага номера n и составлена таблица приближений x_k к решению и соответствующих значений f.

*) Производные от обратной функции F вычисляются по значениям производных от f без труда по известным из курса анализа правилам.

Интерполируем f по ее значениям в $k+1$ узлах x_n, x_{n-1}, ..., x_{n-k} посредством алгебраического многочлена степени k:

$$f(x)=P(x)+R(x), \quad P(x)=\sum_{j=0}^{k}\frac{\omega(x)}{(x-x_{n-j})\,\omega'(x_{n-j})}f(x_{n-j}), \quad (4)$$

$$R(x)=\frac{\omega(x)}{(k+1)!}f^{(k+1)}(\xi),$$

$$\omega(x)=(x-x_n)(x-x_{n-1})\ldots(x-x_{n-k}).$$

Точка ξ принадлежит наименьшему отрезку, содержащему x, x_n, ..., x_{n-k}. Отбрасывая остаток $R(x)$, заменим $f(x)=0$ приближенным уравнением

$$P(x)=0. \quad (5)$$

Чтобы сделать дальнейшее изложение наглядным, воспользуемся не лагранжевым представлением $P(x)$, приведенным в (4), а ньютоновым представлением

$$P(x)=f(x_n)+(x-x_n)f(x_n, x_{n-1})+$$
$$+(x-x_n)(x-x_{n-1})f(x_n, x_{n-1}, x_{n-2})+\ldots$$
$$\ldots+(x-x_n)\ldots(x-x_{n-k+1})f(x_n, x_{n-1}, \ldots, x_{n-k}). \quad (6)$$

Оно имеет то преимущество, что когда x и приближения x_n, x_{n-1}, ..., x_{n-k} близки к решению x^*, разности $x-x_n$, $x-x_{n-1}$, ..., $x-x_{n-k}$ будут малыми величинами и слагаемые в правой части равенства будут расположены по возрастанию порядка малости. Выделение главной линейной части здесь делается без затруднений — в качестве нее можно, очевидно, взять

$$f(x_n)+(x-x_n)f(x_n, x_{n-1})=$$
$$=\frac{x-x_n}{x_{n-1}-x_n}f(x_{n-1})+\frac{x-x_{n-1}}{x_n-x_{n-1}}f(x_n).$$

График ее есть прямая линия, проходящая через точки $M_n[x_n, f(x_n)]$ и $M_{n-1}[x_{n-1}, f(x_{n-1})]$ и являющаяся секущей для графика функции f.

Уравнение (5) запишется в форме

$$f(x_n)+(x-x_n)f(x_n, x_{n-1})+F(x)=0, \quad (7)$$
$$F(x)=(x-x_n)(x-x_{n-1})f(x_n, x_{n-1}, x_{n-2})+\ldots$$

или, если его решить относительно x, входящего в линейную часть,

$$x = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}[f(x_n) + F(x)]. \qquad (8)$$

Решают его обычно простой одношаговой итерацией, принимая за исходное приближение $x^0_{n+1} = x_n$. Подставляют $x = x^0_{n+1}$ в правую часть и результат подстановки принимают за первое приближение x^1_{n+1} и т. д.

Часто соединяют итерацию с последовательным уточнением уравнения: отбрасывают справа в (8) $F(x)$ и результат, не зависящий от x, принимают за x^1_{n+1}; затем сохраняют в $F(x)$ только первый член $(x - x_n)(x - x_{n-1}) \times$ $\times f(x_n, x_{n-1}, x_{n-2})$, полагают справа $x = x^1_{n+1}$ и результат принимают за x^2_{n+1}; после этого сохраняют справа в $F(x)$ два первых члена, подставляют $x = x^2_{n+1}$, принимают результат за x^3_{n+1} и т. д.

После того как найдено значение x_{n+1}, вычисляют $f(x_{n+1})$, составляют интерполяционный многочлен типа (4) для узлов $x_{n+1}, x_n, \ldots, x_{n-k+1}$ и приступают к вычислению x_{n+2} и т. д.

Остановимся еще на выяснении закона изменения погрешности $\varepsilon_n = x_n - x^*$, когда приближения x_n являются близкими к x^*. Так как $P(x_{n+1}) = 0$, будет

$$f(x_{n+1}) = P(x_{n+1}) + R(x_{n+1}),$$

$$R(x_{n+1}) = \frac{(x_{n+1} - x_n) \ldots (x_{n+1} - x_{n-k})}{(k+1)!} f^{(k+1)}(\xi),$$

где ξ лежит на отрезке, содержащем точки $x_{n-k}, \ldots, x_{n+1}$. С другой стороны,

$$R_{n+1}(x_{n+1}) = f(x_{n+1}) = f(x_{n+1}) - f(x^*) = \varepsilon_{n+1} f'(x^* + \theta\varepsilon_{n+1}),$$
$$0 < \theta < 1,$$

и, значит,

$$\varepsilon_{n+1} = \frac{1}{f'(x^* + \theta\varepsilon_{n+1})} R_{n+1}(x_{n+1}).$$

Если воспользоваться приведенным выше выражением для $R(x_{n+1})$ и принять во внимание равенство $x_{n+1} - x_m = \varepsilon_{n+1} - \varepsilon_m$, для ε_{n+1} получим

$$\varepsilon_{n+1} = \frac{f^{(k+1)}(\xi)}{(k+1)! f'(x^* + \theta\varepsilon_{n+1})} (\varepsilon_{n+1} - \varepsilon_n) \ldots (\varepsilon_{n+1} - \varepsilon_{n-k}).$$

Это равенство показывает, в частности, что следующее значение ε_{n+1} погрешности будет малой величиной более высокого порядка, чем каждая из погрешностей $\varepsilon_n, \ldots, \varepsilon_{n-k}$. Поэтому каждую из скобок $(\varepsilon_{n+1} - \varepsilon_n), \ldots, (\varepsilon_{n+1} - \varepsilon_{n-k})$ можно приближенно заменить на $-\varepsilon_n, \ldots, -\varepsilon_{n-k}$. Наконец, величины $f^{(k+1)}(\xi)$ и $f'(x^* + \theta\varepsilon_{n+1})$ можно, внося малые погрешности, заменить соответственно на $f^{(k+1)}(x^*)$ и $f'(x^*)$.

Все такие замены дадут следующее приближенное равенство, достаточно верно описывающее закон изменения погрешности ε_n, когда x_n будут близкими к x^*:

$$\varepsilon_{n+1} \approx \frac{(-1)^{k+1}}{(k+1)!} \cdot \frac{f^{(k+1)}(x^*)}{f'(x^*)} \varepsilon_n \varepsilon_{n-1} \ldots \varepsilon_{n-k}. \qquad (9)$$

Применение изложенного метода к вычислениям требует составления начальной таблицы, содержащей значения $x_0, x_1, \ldots, x_k$, и необходимых значений функции f и производных от нее.

3. Метод, основанный на интерполировании обратной функции. Будем считать, что при решении уравнения $f(x) = 0$ составляется таблица приближений к решению x_n и соответствующих значений $y_n = f(x_n)$ функции f. Предположим также, что зависимость $y = f(x)$ обратима на некотором отрезке, содержащем приближения x_n и решение x^*. Рассмотрим обратную функцию $x = F(y)$ и интерполируем ее по $k+1$ значениям $x_n, \ldots, x_{n-k}$ в точках $y_n, \ldots, y_{n-k}$:

$$F(y) \approx \Pi(y) = \sum_{j=0}^{k} \frac{\Omega(y)}{(y - y_{n-j})\,\Omega'(y_{n-j})} x_{n-j}, \qquad (10)$$

$$\Omega(y) = \prod_{j=0}^{k} (y - y_{n-j}).$$

За следующее приближение x_{n+1} принимается значение многочлена $\Pi(y)$ при $y = 0$:

$$x_{n+1} = \Pi(0) = -\sum_{j=0}^{k} \frac{\Omega(0)}{y_{n-j}\,\Omega'(y_{n-j})} x_{n-j}. \qquad (11)$$

Обратим внимание на то, что в рассматриваемом методе дается явное выражение для x_{n+1}, достаточно удобное для вычислений.

Остановимся еще на законе изменения погрешностей $\varepsilon_n = x_n - x^*$, когда приближения x_n становятся близкими к решению x^* и ε_n — малыми величинами. Рассмотрим для этого погрешность интерполирования $F(y)$:

$$r(y) = F(y) - \Pi(y) = \frac{\Omega(y)}{(k+1)!} F^{(k+1)}(\eta).$$

Здесь η есть некоторая средняя точка на наименьшем отрезке, содержащем $y_n, \ldots, y_{n-k}$ и y.

Так как $x^* = F(0)$ и $x_{n+1} = \Pi(0)$, то

$$\varepsilon_{n+1} = x_{n+1} - x^* = \Pi(0) - F(0) = -r(0) =$$

$$= -\frac{\Omega(0)}{(k+1)!} F^{(k+1)}(\eta_0) =$$

$$= \frac{(-1)^k}{(k+1)!} F^{(k+1)}(\eta_0) \prod_{j=0}^{k} y_{n-j} = \frac{(-1)^k}{(k+1)!} F^{(k+1)}(\eta_0) \prod_{j=0}^{k} f(x_{n-j}).$$

Знаком η_0 названа некоторая точка отрезка, содержащего $y_n, \ldots, y_{n-k}, 0$. Ввиду $f(x^*) = 0$ и $f(x_m) = f(x^* + \varepsilon_m) - f(x^*) = \varepsilon_m f'(x^* + \theta\varepsilon_m)$ $(0 < \theta < 1)$, будет

$$\varepsilon_{n+1} = \frac{(-1)^k}{(k+1)!} F^{(k+1)}(\eta_0) \prod_{j=0}^{k} \varepsilon_{n-j} f'(x^* + \theta\varepsilon_{n-j}).$$

В правой части равенства выделим главную часть, выполняя замены $F^{(k+1)}(\eta_0) \approx F^{(k+1)}(0)$ и $f'(x^* + \theta\varepsilon_{n-j}) \approx f'(x^*)$, после чего получим приближенное равенство, дающее достаточно простую и наглядную картину изменения погрешности вблизи x^*:

$$\varepsilon_{n+1} \approx \frac{(-1)^k}{(k+1)!} [f'(x^*)]^{k+1} F^{(k+1)}(0) \varepsilon_n \varepsilon_{n-1} \ldots \varepsilon_{n-k}. \quad (12)$$

Оно позволяет думать, что если обратная функция имеет непрерывную производную порядка $k+1$ в окрестности точного решения x^* и если приближения $x_0, x_1, \ldots, x_k$ взяты достаточно близко к x^*, то вычислительный процесс, определяемый правилом (11), будет быстро сходиться к x^*.

§ 7. Упрощение алгебраических уравнений путем выделения множителей

1. Введение. Алгебраические уравнения являются уравнениями простейшего вида; они часто встречаются в приложениях, и вопрос о численном решении их уже давно привлекает внимание. Общие методы решения, о которых говорилось выше, разумеется, применимы и к алгебраическим уравнениям, но кроме них были развиты другие методы, предназначенные специально для таких уравнений и учитывающие их особенности. Таких методов существует немало, мы же остановимся только на одном из них — на методе разложения на множители, так как он часто применяется в вычислениях.

Пусть дано алгебраическое уравнение степени $n > 1$:

$$P(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n = 0 \quad (a_0a_n \neq 0). \quad (1)$$

Если удастся разложить многочлен $P(x)$ на множители, то мы понизим степень уравнения и в той или иной мере упростим задачу его решения. Для определенности изложения рассмотрим случай, когда коэффициенты a_k многочлена $P(x)$ суть действительные числа. Как известно из курсов алгебры, корни такого многочлена могут быть действительными, при этом кратности их могут быть любыми; комплексные же корни должны быть попарно сопряжены, и кратности сопряженных корней обязательно одинаковы. Многочлен $P(x)$ может быть представлен в виде произведения нескольких действительных множителей, линейных или квадратичных, совпадающих или различных; при этом линейные множители отвечают действительным корням многочлена, квадратичные же отвечают парам сопряженных комплексных корней. Поэтому, принципиально говоря, достаточно ограничиться рассмотрением алгоритмов выделения только линейных и квадратичных множителей.

2. Нахождение линейного множителя многочлена. Пусть нужно найти действительный корень α многочлена $P(x)$, и для этой цели мы хотим выделить из $P(x)$ линейный множитель $x - \alpha$. Предположим, что мы знаем приближенное значение x_0 этого корня и по нему можем составить лишь приближенное значение $x - x_0$ линей-

ного множителя. Для его уточнения делят по хорошо известным правилам $P(x)$ на $x-x_0$. Если деление выполнить до конца, то в остатке получится постоянная величина. Остановим деление на предпоследнем шаге, когда остаток будет линейным d_0x+d_1, и запишем его в форме $d_0(x-x_1)$. При этом предполагается $d_0\neq 0$. Разность $x-x_1$ называют приведенным предпоследним остатком.

Такая операция деления приведет к следующему представлению $P(x)$:

$$P(x)=(x-x_0)(b_0x^{n-1}+\ldots+b_{n-2}x)+d_0(x-x_1). \quad (2)$$

Число x_1 принимают за первое улучшенное значение корня. Легко построить явное выражение x_1 через многочлен P. Из равенства (2) получается

$$P(x_0)=d_0(x_0-x_1), \quad P(0)=-d_0x_1.$$

Исключая отсюда d_0, найдем

$$x_1=\frac{P(0)\,x_0}{P(0)-P(x_0)}=-\frac{a_n}{a_0x_0^{n-1}+\ldots+a_{n-1}}. \quad (3)$$

Для построения x_2 делим $P(x)$ на $x-x_1$, находим «предпоследний» остаток и, представляя его в приведенной форме $d_1(x-x_2)$, найдем x_2 и т. д.

Следующее приближение x_{k+1} строится по предыдущему x_k путем деления и представления $P(x)$ в виде

$$P(x)=(x-x_k)(c_0x^{n-1}+c_1x^{n-2}+\ldots+c_{n-2}x)+ \\ +d_k(x-x_{k+1}). \quad (4)$$

Условием возможности неограниченного продолжения такого алгоритма, указанного Лином, является соблюдение неравенства $d_k\neq 0$ $(k=0,1,\ldots)$.

Явное выражение x_{k+1} через x_k получается из (2):

$$x_{k+1}=\frac{P(0)\,x_k}{P(0)-P(x_k)}=-\frac{a_n}{a_0x_k^{n-1}+\ldots+a_{n-1}}. \quad (5)$$

По существу изложенный алгоритм является простой одношаговой итерацией для уравнения

$$x=\varphi(x)=\frac{P(0)\,x}{P(0)-P(x)}=x+\frac{xP(x)}{P(0)-P(x)}. \quad (6)$$

Заметим, что если последовательность x_k сходится, и $\lim x_k = x^*$, то x^* есть корень многочлена $P(x)$. Действительно, если предположить, что $x_k \to x^*$ $(k \to \infty)$, то из рекуррентного соотношения (5), взятого в форме

$$x_{k+1} = x_k + \frac{x_k}{P(0) - P(x_k)} P(x_k),$$

следует, что при $k \to \infty$ должно быть $P(x_k) \to 0$ и, ввиду непрерывности $P(x)$, будет $P(x^*) = 0$.

Что же касается сходимости последовательных приближений x_k, то ее гарантировать можно не во всех случаях. Напомним, что для уравнения $x = \varphi(x)$ при нахождении решения x^* в методе простой итерации, если исходное приближение x_k взято достаточно близко к x^* и если $|\varphi(x^*)| < 1$, то итерационная последовательность x_n сходится к x^*, а если $|\varphi(x^*)| > 1$, то решение x^* будет «точкой отталкивания» для последовательности x_n и сходимости, вообще говоря, не будет.

Как видно из (5),

$$\varphi'(\alpha) = 1 + \alpha \frac{P'(\alpha)}{P(0)}.$$

Поэтому можно утверждать, что последовательность x_k, определяемая процессом (5), будет сходиться к корню α, если выполнены условия: во-первых, $\left|1 + \alpha \frac{P'(\alpha)}{P(0)}\right| < 1$ и, во-вторых, исходное приближение x_0 взято достаточно близким к α.

3. Нахождение квадратичного множителя. Будем искать множитель в виде $x^2 + px + q$. Предположим, что нам известны или как-то заданы приближенные значения p_0 и q_0 для p и q. Разделим $P(x)$ на $x^2 + p_0 x + q_0$ и остановим процесс деления на предпоследнем остатке, который будет, вообще говоря, многочленом второй степени $ax^2 + bx + c$. Предполагая $a \neq 0$, приведем его к виду $a(x^2 + p_1 x + q_1)$ и примем p_1, q_1 за первые улучшенные значения для p, q. После этого разделим $P(x)$ на $x^2 + p_1 x + q_1$ и проделаем весь цикл вычислений для нахождения вторых улучшений значений p_2, q_2 и т. д.

Выясним характер зависимости p_k и q_k от коэффициентов уравнения. Для этого будем считать, что деление $P(x)$ на трехчлен x^2+px+q доведено до конца, т. е. до линейного остатка. После деления получится следующее равенство:

$$P(x)=a_0x^n+\ \dots\ +a_n=$$
$$=(x^2+px+q)(b_0x^{n-2}+b_1x^{n-3}+\ \dots\ +b_{n-3}x+b_{n-2})+$$
$$+b_{n-1}(x+p)+b_n. \qquad (7)$$

Коэффициенты b_{n-1}, b_n линейного остатка записаны в особой форме, чтобы получить единообразные равенства, связывающие a_i и b_i $(i=0,\ 1,\ \dots,\ n)$. Если сравнить коэффициенты при степенях x, получится система уравнений

$$\begin{aligned}
a_0&=b_0,\\
a_1&=pb_0+b_1,\\
a_2&=qb_0+pb_1+b_2,\\
&\dots\dots\dots\dots\\
a_{n-1}&=qb_{n-3}+pb_{n-2}+b_{n-1},\\
a_n&=qb_{n-2}+pb_{n-1}+b_n,
\end{aligned} \qquad (8)$$

из которой последовательно могут быть найдены b_i. Очевидно, при этом b_i будут многочленами относительно p и q, и легко видеть, что b_i будет иметь степень i относительно p и степень $i-1$ относительно q. Ниже коэффициенты b_i будем обозначать $b_i(p,\ q)$.

При делении $P(x)$ на x^2+px+q предпоследний остаток есть

$$b_{n-2}(x^2+px+q)+b_{n-1}(x+p)+b_n=$$
$$=b_{n-2}x^2+(pb_{n-2}+b_{n-1})x+qb_{n-2}+pb_{n-1}+b_n.$$

Коэффициенты же p_{k+1} и q_{k+1} в приведенном предпоследнем остатке номера $k+1$ должны находиться по значениям p_k и q_k по правилу

$$\begin{aligned}
p_{k+1}&=p_k+b_{n-1}(p_k,\ q_k)/b_{n-2}(p_k,\ q_k),\\
q_{k+1}&=q_k+[p_kb_{n-1}(p_k,\ q_k)+b_n(p_kq_k)]/b_{n-2}(p_k,\ q_k).
\end{aligned} \qquad (9)$$

Укажем еще на связь алгоритмов (9) для нахождения последовательных приближений p_k и q_k с точными

уравнениями для нахождения p и q. Для того чтобы трехчлен x^2+px+q был делителем многочлена $P(x)$, необходимо и достаточно, чтобы последний остаток $b_{n-1}(x+p)+b_n$ в представлении (7) был тождественно равен нулю, что равносильно равенствам

$$pb_{n-1}(p, q)+b_n(p, q)=0, \qquad b_n(p, q)=0$$

или

$$b_{n-1}(p, q)=0, \qquad b_n(p, q)=0.$$

Последние можно записать в виде

$$\begin{aligned} p &= \varphi(p, q)=p+\frac{b_{n-1}(p, q)}{b_{n-2}(p, q)}, \\ q &= \psi(p, q)=q+\frac{b_n(p, q)+pb_{n-1}(p, q)}{b_{n-2}(p, q)}. \end{aligned} \tag{10}$$

Равенства же (9) суть не что иное, как формулы простой одношаговой итерации для системы (10).

В § 4 мы обращали внимание на то, что такой метод для систем уравнений сходится не всегда, и указывали условия, при соблюдении которых можно ожидать сходимости последовательностей p_k и q_k, полученных по правилам (9).

ЛИТЕРАТУРА

1. Бахвалов Н. С., Численные методы, «Наука», М., 1973.
2. Березин И. С., Жидков Н. П., Методы вычислений. I, «Наука», М., 1966.
3. Загускин В. Л., Справочник по численным методам решения уравнений, Физматгиз, М., 1960.
4. Ланс Дж. Н., Численные методы для быстродействующих вычислнтельных машин, ИЛ, М., 1962.
5. Мысовских И. П., Лекции по методам вычислений, Физматгиз, М., 1962.
6. Островскнй А., Решение уравненнй и систем уравнений, ИЛ, М., 1963.
7. Хаусхолдер А. С., Основы численного анализа, ИЛ, М., 1956.
8. Мак-Кракен Д., Дорн У., Численные методы и программирование на Фортране, «Мир», М., 1969.
9. Сборник научных программ на Фортране, вып. 1, 2, «Статистика», М., 1974.

ГЛАВА 5

ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

В этой главе будет рассмотрена задача о вычислении однократного интеграла с помощью конечного числа значений интегрируемой функции. Изложение начнем с рассмотрения определенного интеграла.

§ 1. Введение

1. О форме, придаваемой интегралу при вычислении. Пусть $[a, b]$ есть любой конечный или бесконечный отрезок числовой оси, и рассматривается интеграл $\int_a^b F(x)\,dx$. Предположим, что мы ставим своей задачей найти приближенное его значение по n значениям $F(x_i)$ функции F в точках x_i $(i = 1, \ldots, n)$.

Многие правила приближенных квадратур основаны на замене интегрируемой функции F на всем отрезке $[a, b]$ или на его частях на более простую функцию φ, близкую к F, легко интегрируемую точно и принимающую в узлах x_i те же значения $F(x_i)$, что и F. В качестве такой функции берут либо алгебраический многочлен от x, либо рациональную функцию, либо тригонометрический многочлен и т. д. в зависимости от задачи. Все эти вспомогательные функции φ являются аналитическими и обладают большой гладкостью изменения.

Когда отрезок интегрирования конечный и интегрируемая функция F имеет высокую гладкость, то можно рассчитывать хорошо приблизить ее многочленом невысокой степени или несложной рациональной функцией. Если же сама функция F или ее производные невысоких порядков имеют особенности или даже обращаются

в бесконечность, то это затруднит приближение F или сделает его вообще невозможным. В этом случае мы должны будем заранее освободиться от таких особенностей путем их выделения. Делается это при помощи разложения F на два сомножителя $F(x) = p(x)f(x)$, где $p(x)$ имеет такие же особенности, как и $F(x)$, а $f(x)$ есть достаточно гладкая функция, и интеграл рассматривается в форме

$$\int_a^b p(x)f(x)\,dx. \tag{1}$$

Такое представление важно также, например, в задаче вычисления несобственных интегралов по бесконечным отрезкам. Нередко приходится вычислять интегралы вида $\int_a^\infty F(x)\,dx$. Для них большое значение имеет закон, по которому убывает F при $x \to \infty$. Здесь F целесообразно разложить на множители $F(x) = p(x)f(x)$, из которых первый $p(x)$ характеризует закон убывания F, второй же $f(x)$ является гладкой функцией, допускающей хорошее приближение алгебраическими многочленами или рациональными функциями.

Функция $p(x)$ в (1) называется весовой функцией или весом. При построении вычислительного правила для (1) она считается фиксированной, и поэтому правила, о которых будет говориться ниже, в большинстве своем будут специализированными, и каждое из них будет предназначено для численного интегрирования функций, имеющих особенности одного и того же типа, определяемого весом $p(x)$. Функция $f(x)$ предполагается любой достаточно гладкой на $[a, b]$.

2. Квадратурная сумма и связанные с ней задачи. Будут строиться формулы вычислений, имеющие вид

$$\int_a^b p(x)f(x)\,dx \approx \sum_{k=1}^n A_k f(x_k), \qquad x_k \in [a, b]. \tag{2}$$

Величины A_k называются квадратурными коэффициентами, x_k — квадратурными узлами и правая часть (2) — квадратурной суммой. Формула имеет $2n+1$ парамет-

ров: n, A_k, x_k $(k = 1, 2, \ldots, n)$, и их следует выбрать так, чтобы формула давала возможно лучший результат при интегрировании избранного класса функций f.

Назначение параметра n является очевидным: чем больше n, тем больше слагаемых в квадратурной сумме и тем большей точности можно достигнуть путем выбора A_k и x_k. Поэтому при построении формулы число n считают закрепленным и рассматривают лишь задачу о выборе A_k и x_k. Отметим попутно, что эти параметры не всегда являются произвольными и в некоторых случаях на их значения необходимо бывает наложить ограничения, например, при интегрировании таблично заданных функций за узлы x_k могут быть взяты только табличные значения аргумента. В ближайшем изложении мы будем считать A_k и x_k произвольными. Правом выбора их обычно пользуются для следующих целей.

1) Увеличение степени точности. Рассмотрим систему линейно независимых функций $\omega_m(x)$ $(m = 0, 1, 2, \ldots)$ таких, чтобы произведения $p(x)\omega_n(x)$ были абсолютно интегрируемы на $[a, b]$.

Предположим, что рассматривается некоторое семейство F функций f. Будем приближать функции f при помощи линейных комбинаций

$$s_n(x) = \sum_{k=0}^{n} a_k \omega_k(x).$$

За меру близости между f и s_n примем величину

$$\rho(f, s_n) = \int_a^b |p(f - s_n)|\, dx.$$

Система $\omega_m(x)$ называется *полной* в множестве F, если для всякого $\varepsilon > 0$ существует такая линейная комбинация s_n, что выполняется неравенство $\rho(f, s_n) < \varepsilon$.

Если система ω_m обладает свойством полноты, то из неравенства

$$\left| \int_a^b pf\, dx - \int_a^b ps_n\, dx \right| \leqslant \int_a^b |p(f - s_n)|\, dx = \rho(f, s_n) < \varepsilon$$

вытекает, что интеграл (1) может быть вычислен сколь угодно точно при помощи замены f на линейную комби-

нацию s_n, составленную при надлежащем выборе n и a_k $(k = 0, 1, \ldots, n)$. Это позволяет ожидать, что если мы сможем возможно точнее интегрировать функции ω_m, то мы одновременно сможем хорошо интегрировать всякую функцию $f \in F$.

Говорят, что формула (2) имеет степень точности m, если она является точной для функций ω_i $(i = 0, 1, \ldots, m)$:

$$\int_a^b p(x)\,\omega_i(x)\,dx = \sum_{k=1}^{n} A_k \omega_i(x_k) \qquad (i = 0, 1, \ldots, m),$$

и не является точной для ω_{m+1}.

Можно стремиться к тому, чтобы при помощи выбора параметров A_k и x_k сделать степень точности формулы (2) наивысшей возможной. Такие формулы впервые были рассмотрены Гауссом, и их часто называют формулами гауссова типа или формулами наивысшей степени точности. Так как число параметров A_k и x_k равно $2n$, то есть надежда достигнуть того, чтобы формула (2) имела степень точности $2n - 1$, и можно предполагать, что такая степень точности является, как правило, наивысшей возможной. Ниже будут указаны случаи, когда эти ожидания оправдываются.

2) Минимизация погрешности. Рассмотрим остаточный член или погрешность квадратурной формулы (2)

$$R_n(f) = \int_a^b pf\,dx - \sum_{k=1}^{n} A_k f(x_k) = \int_a^b pf\,dx - Q_n(f). \qquad (3)$$

За величину, характеризующую точность формулы на множестве F функций f может быть принята верхняя грань абсолютного значения R_n:

$$\sup_f |R_n(f)| = M(A_1, \ldots, A_n; x_1, \ldots, x_n).$$

Она зависит от выбора узлов x_k и коэффициентов A_k, и их можно стараться выбрать так, чтобы величина M имела бы наименьшее значение. Такая задача интересна главным образом в теоретическом отношении и не при-

вела еще к полезным для практических вычислений результатам.

3) Упрощение вычислений. Можно при помощи выбора параметров A_k и x_k стремиться сделать возможно более простыми вычисления по формуле (2). Например, можно выбрать, что делают нередко, узлы x_k равноотстоящими, положив $x_k = a + kh$, $h = (b-a)/n$ и применять формулу

$$\int_a^b p(x) f(x)\, dx \approx \sum_{k=0}^{n} A_k f(a + kh),$$

выбирая в ней коэффициенты A_k по каким-либо условиям.

Для упрощения счета можно потребовать равенства коэффициентов и рассматривать следующую формулу:

$$\int_a^b p(x) f(x)\, dx \approx C[f(x_1) + \ldots + f(x_n)].$$

Она содержит $n+1$ параметров C и x_k $(k = 1, \ldots, n)$, которые можно попытаться выбрать так, чтобы такая формула имела степень точности не ниже n.

3. Погрешность квадратуры и сходимость квадратурного процесса. Погрешность квадратурной формулы (2) указана в равенстве (3). Величина погрешности зависит, очевидно, от свойств функции f и от выбора формулы, т. е. от ее узлов x_k и коэффициентов.

В исследовании погрешности основными являются две следующие задачи.

Во-первых, оценка погрешности для функций с теми или иными свойствами, среди которых наибольший интерес имеют классы функций, часто встречающиеся в приложениях: функции с конечным числом разрывов, непрерывные, имеющие заданный порядок дифференцируемости, аналитические и т. д. Здесь имеют значение как точные оценки для узких классов, так и более грубые оценки для широких классов, полезные в вопросах сходимости.

Во-вторых, выяснение условий сходимости квадратурных процессов, т. е. условий, при которых $R_n(f) \to 0$ $(n \to \infty)$.

Квадратурный процесс, или последовательность квадратурных формул, определяется двумя бесконечными треугольными таблицами: таблицей узлов

$$X = \begin{Bmatrix} x_1^1 & & & \\ x_1^2 & x_2^2 & & \\ x_1^3 & x_2^3 & x_3^3 & \\ \dots & \dots & \dots & \dots \end{Bmatrix} \tag{4}$$

и таблицей коэффициентов

$$A = \begin{Bmatrix} A_1^1 & & & \\ A_1^2 & A_2^2 & & \\ A_1^3 & A_2^3 & A_3^3 & \\ \dots & \dots & \dots & \dots \end{Bmatrix}. \tag{5}$$

В проблеме сходимости приходится иметь дело с тремя объектами: классом F функций f и двумя таблицами X и A; необходимо бывает выяснить, как они должны быть связаны между собой для того, чтобы остаток $R_n(f)$ стремился к нулю при $n \to \infty$. Несколько теорем такого вида будет приведено ниже.

§ 2. Интерполяционные квадратурные правила

Здесь будут изложены правила численных квадратур, основанные на алгебраическом интерполировании функции на всем отрезке интегрирования.

1. Общая интерполяционная квадратура. Предположим, что узлы x_k $(k = 1, \dots, n)$ как-либо выбраны, и мы имеем право при построении квадратурной формулы распоряжаться выбором лишь коэффициентов A_k. Остановимся сначала на их определении, основанном на алгебраическом интерполировании функции f на всем отрезке $[a, b]$. Выполним такое интерполирование f по ее значениям в узлах x_k $(k = 1, 2, \dots, n)$ при помощи многочлена $P(x)$ степени $n-1$:

$$f(x) = P(x) + r(x), \quad P(x) = \sum_{k=1}^{n} \frac{\omega(x)}{(x - x_k)\,\omega'(x_k)} f(x_k), \tag{1}$$

$$\omega(x) = (x - x_1) \dots (x - x_n).$$

Если это представление f внести в интеграл (1), получим равенство

$$\int_a^b p(x) f(x)\,dx = \int_a^b p(x) P(x)\,dx + \int_a^b p(x) r(x)\,dx.$$

Отбросив справа интеграл с остаточным членом $r(x)$, найдем приближенное правило интегрирования, называемое интерполяционным:

$$\int_a^b p(x) f(x)\,dx \approx \sum_{k=1}^{n} A_k f(x_k),$$
$$A_k = \int_a^b p(x) \frac{\omega(x)}{(x-x_k)\omega'(x_k)}\,dx. \tag{2}$$

Его погрешность следующим образом выражается через остаточный член интерполирования:

$$R_n(f) = \int_a^b pf\,dx - \sum_{k=1}^{n} A_k f(x_k) = \int_a^b p(x) r(x)\,dx. \tag{3}$$

Интерполяционная формула (2) характеризуется следующей теоремой о степени точности.

Теорема 1. *Для того чтобы квадратурная формула* (1.2) *была точной для алгебраических многочленов степени* $n-1$, *необходимо и достаточно, чтобы она была интерполяционной.*

Доказательство. Проверим необходимость. Возьмем функцию

$$f(x) = \frac{\omega(x)}{(x-x_i)\omega'(x_i)} = \omega_i(x).$$

Это есть многочлен степени $n-1$, и если правило верно для всяких многочленов степени $n-1$, то оно должно быть точным и для $\omega_i(x)$. Поэтому верны равенства

$$\int_a^b p(x)\omega_i(x)\,dx = \int_a^b p(x) \frac{\omega(x)}{(x-x_i)\omega'(x_i)}\,dx =$$
$$= \sum_{k=1}^{n} A_k \omega_i(x_k) = A_i,$$

и формула (1.2) действительно является интерполяционной, так как ее коэффициенты имеют значения, указанные в (2).

Докажем достаточность. Пусть f есть произвольный многочлен степени $n-1$. Убедимся в том, что для него равенство (2) будет выполняться точно. Интерполируем f по значениям в узлах x_k $(k=1,\ldots,n)$. Интерполирование будет точным:

$$f(x)=\sum_{k=1}^{n}\frac{\omega(x)}{(x-x_k)\,\omega'(x_k)}f(x_k).$$

Кроме того, если A_k имеют значения, указанные в (2), то верны равенства

$$\int_a^b pf\,dx=\sum_{k=1}^{n}f(x_k)\int_a^b p(x)\frac{\omega(x)}{(x-x_k)\,\omega'(x_k)}\,dx=\sum_{k=1}^{n}A_kf(x_k), \quad (4)$$

и равенство (2) для $f(x)$ выполняется точно.

Как показывает теорема 1, интерполяционные квадратурные формулы характеризуются тем, что при всяком расположении узлов x_k алгебраическая степень точности их не меньше $n-1$. Если мы хотим увеличить степень точности, то это можно сделать только за счет выбора узлов x_k. Ниже будет показано, что при помощи такого выбора степень точности, по крайней мере для знакопостоянных весовых функций $p(x)$, может быть повышена на n единиц и доведена до $2n-1$.

Обратимся к рассмотрению погрешности квадратуры (2). Ее представление через $r(x)$ указано в равенстве (3). Что же касается выражения для остаточного члена интерполирования $r(x)$, то его представления для некоторых классов функций указывались в гл. I. Например, для любых функций f с конечными значениями остаточный член $r(x)$ может быть записан в виде

$$r(x)=\omega(x)f(x_1,\ x_2,\ \ldots,\ x_n,\ x).$$

Это позволяет утверждать, что для любой функции f с конечными значениями на $[a,b]$ и такой, что инте-

гралы, входящие в (4), имеют смысл, погрешность квадратуры представима в форме

$$R_n(f) = \int_a^b p(x)\,\omega(x)\,f(x_1, x_2, \ldots, x_n, x)\,dx. \qquad (5)$$

Ввиду большой общности это представление редко используется в приложениях и имеет главным образом теоретическое значение. Укажем еще два более специализированных представления, но более полезные в практике вычислений. Если функция f имеет непрерывную производную порядка n на $[a, b]$, то на отрезке, содержащем точки $x_1, \ldots, x_n, x$, существует точка ξ такая, что для $r(x)$ верно равенство

$$r(x) = \frac{\omega(x)}{n!} f^{(n)}(\xi).$$

Оно позволяет получить для $R_n(f)$ следующее выражение:

$$R_n(f) = \frac{1}{n!} \int_a^b p(x)\,\omega(x)\,f^{(n)}(\xi)\,dx. \qquad (6)$$

Для функций, имеющих на $[a, b]$ непрерывную производную порядка n, удовлетворяющую условию $|f^{(n)}(x)| \leqslant M_n$ $(a \leqslant x \leqslant b)$, отсюда получается оценка погрешности квадратуры

$$|R_n(f)| \leqslant \frac{1}{n!} M_n \int_a^b |p(x)\,\omega(x)|\,dx.$$

Здесь равенство достижимо только при условии, когда произведение $p(x)\omega(x)$ сохраняет знак на $[a, b]$, в других же случаях оценка может быть далека от оптимальной.

Для получения точной оценки $R_n(f)$ для рассматриваемых функций воспользуемся другими соображениями. Всякая функция с непрерывной производной порядка m на $[a, b]$ представима по формуле Тейлора, остаточный член которой взят в интегральной

форме *) с гасящей функцией E:

$$f(x) = \sum_{i=0}^{m-1} \frac{1}{i!}(x-a)^i f^{(i)}(a) + \\ + \frac{1}{(m-1)!}\int_a^b f^{(m)}(t)(x-t)^{m-1} E(x-t)\,dt.$$

Если это представление внести в остаточный член

$$R_n(f) = \int_a^b p(x) f(x)\,dx - \sum_{k=1}^{n} A_k f(x_k)$$

и изменить порядок интегрирования по x и t, для $R_n(f)$ получится равенство

$$R_n(f) = \sum_{i=0}^{m-1} \frac{1}{i!} f^{(i)}(a) R_n[(x-a)^i] + \frac{1}{(m-1)!}\int_a^b f^{(m)}(t) K_m(t)\,dt,$$

$$K_m(t) = \int_t^b p(x)(x-t)^{m-1}\,dx - \sum_{x_k > t} A_k (x_k - t)^{m-1}$$

$$\{t \neq a,\ x_k\ (k = 1, 2, \ldots, n)\}.$$

Оно дает представление остатка $R_n(f)$ для любого квадратурного правила в классе функций, имеющих на $[a, b]$ непрерывные производные до порядка m. В случае интерполяционной квадратуры многочлены до степеней $n-1$ интегрируются точно, и поэтому $R_n[(x-a)^i] = 0$ $(i = 0, 1, \ldots, n-1)$. Если, кроме того, интегрируемая функция f имеет непрерывную производную порядка n, удовлетворяющую условию

$$|f^{(n)}(x)| \leqslant M_n, \qquad x \in [a, b], \tag{7}$$

то в представлении $R_n(f)$ сохранится справа лишь интегральный член

$$R_n(f) = \frac{1}{(n-1)!}\int_a^b f^{(n)}(x) K_n(x)\,dx \tag{8}$$

*) См. гл. 1, § 3, п. 2. При записи формулы мы для упрощения считали a конечной величиной. Если же $a = -\infty$, то вместо a может быть взята любая точка C на $[a, b]$ и незначительно изменен интеграл.

и оценка $R_n(f)$ в классе функций, определяемом условием (7), будет

$$|R_n(f)| \leqslant M_n \frac{1}{(n-1)!} \int_a^b |K_n(x)|\, dx.$$

Оценка является точной.

2. Квадратурные формулы с равноотстоящими узлами. В вычислениях часто узлы x_k берутся равноотстоящими. Интерполяционные квадратуры с такими узлами принято называть формулами Ньютона — Котеса в память того, что впервые они в достаточно общей форме были рассмотрены Ньютоном, коэффициенты же для них в случае постоянной весовой функции найдены Котесом при $n = 1, 2, \ldots, 10$.

Предположим, что отрезок интегрирования конечен; разделим его на n равных частей длины $h = \frac{1}{n}(b-a)$, и точки деления $a + kh$ $(k = 0, 1, \ldots, n)$ примем за узлы интерполяционной формулы, которую запишем в форме

$$\int_a^b p(x) f(x)\, dx \approx (b-a) \sum_{k=0}^{n} B_k^n f(a+kh), \tag{9}$$

$$B_k^n = (b-a)^{-1} A_k = (b-a)^{-1} \int_a^b p(x) \frac{\omega(x)}{(x-a-kh)\,\omega'(a+kh)}\, dx,$$

$$\omega(x) = (x-a)(x-a-h) \ldots (x-a-nh).$$

Если ввести вместо x переменную t, положив $x = a + th$ $(0 \leqslant t \leqslant n)$, можно упростить выражение для B_k^n:

$$\omega(x) = \omega(a+th) = h^{n+1} t(t-1) \ldots (t-n),$$

$$x - a - kh = h(t-k), \quad \omega'(a+kh) = (-1)^{n-k} h^n k!\,(n-k)!,$$

$$B_k^n = \frac{(-1)^{n-k}}{nk!\,(n-k)!} \int_0^n p(a+th) \frac{t(t-1) \ldots (t-n)}{t-k}\, dt. \tag{10}$$

Для постоянной весовой функции $p(x) \equiv 1$ формула Ньютона — Котеса имеет вид

$$\int_a^b f(x)\,dx \approx (b-a)\sum_{k=0}^{n} B_k^n f(a+kh), \tag{11}$$

$$B_k^n = \frac{(-1)^{n-k}}{nk!\,(n-k)!}\int_0^n \frac{t(t-1)\ldots(t-n)}{t-k}\,dt.$$

В таблице приведены значения коэффициентов B_k^n формулы (11) для ряда значений n вплоть до $n=7$. Поскольку при каждом n имеет место соотношение симметрии $B_k^n = B_{n-k}^n$, то в таблицу включены только коэффициенты с индексом $k \leqslant n/2$.

n \ k	0	1	2	3
1	1/2			
2	1/6	4/6		
3	1/8	3/8		
4	7/90	32/90	12/90	
5	19/288	75/288	50/288	
6	41/840	216/840	27/840	272/840
7	751/17 280	3 577/17 280	1 323/17 280	2989/17 280

Коэффициенты B_k^n вычислены (см. [3]) до $n=20$. Они являются рациональными числами с большими числителями и знаменателями, и это заставило нас ограничиться лишь приведенной краткой таблицей. Относительно B_k^n отметим без доказательства следующие факты.

1) Среди B_k^n $(k=0, 1, \ldots, n)$ для $n \geqslant 10$ существуют отрицательные.

Чтобы пояснить значение этого факта, отметим, что при вычислении суммы, стоящей справа в (9), пользуются, как правило, приближенными значениями $f(a+kh)$. Пусть все они известны с погрешностью ε. Погрешность,

которая может получиться при составлении суммы $\sum B_k^n f(a+kh)$, должна быть оценена величиной $\varepsilon \sum |B_k^n|$.

Заметим, что сумма $\sum B_k^n$ при всяком n равна единице; это сразу же следует из (11) при $f \equiv 1$. Наличие же среди B_k^n $(k = 0, 1, \ldots, n)$ отрицательных чисел вызывает увеличение $\sum |B_k^n|$ и возможной погрешности $\varepsilon \sum |B_k^n|$.

Насколько быстрым является рост $\sum |B_k^n|$ при увеличении n, можно судить по следующим цифрам:

$$\text{при } n = 10 \qquad \sum |B_k^{10}| \approx 3{,}1;$$
$$\text{при } n = 15 \qquad \sum |B_k^{15}| \approx 8{,}3;$$
$$\text{при } n = 20 \qquad \sum |B_k^{20}| \approx 560.$$

Поэтому при пользовании формулой Ньютона — Котеса (9) для $n = 15$ при составлении квадратурной суммы можно потерять в точности один десятичный разряд, тогда как для $n = 20$ могут быть потеряны в точности три десятичных разряда.

2) При больших значениях n коэффициенты B_k^n имеют следующие представления:

$$B_k^n = \frac{(-1)^{k-1} n!}{k!\,(n-k)!\,n \ln^2 n}\left[\frac{1}{k} + \frac{(-1)^k}{n-k}\right] \cdot \left[1 + O\left(\frac{1}{\ln n}\right)\right] =$$
$$= \frac{(-1)^{k-1} n^{k-1}}{k! \ln^2 n}\left[\frac{1}{k} + \frac{(-1)^k}{n-k}\right] \cdot \left[1 + O\left(\frac{1}{\ln n}\right)\right] \qquad (12)$$
$$(1 \leqslant k \leqslant n-1),$$
$$B_0^n = B_n^n = \frac{1}{n \ln n}\left[1 + O\left(\frac{1}{\ln n}\right)\right].$$

Как видно отсюда, для таких n будут часто встречаться случаи, когда смежные коэффициенты B_k^n и B_{k+1}^n будут иметь большие абсолютные значения и противоположные знаки. Поэтому при $n \gg 1$ сумма $\sum |B_k^n|$ будет большой и быстро возрастающей величиной. Это позволяет ожидать, что при больших n формулы Ньютона — Котеса становятся малопригодными для вычислений.

Обратим еще внимание на некоторые свойства формулы (10) без их подробного рассмотрения.

1) Коэффициенты B_k^n, отвечающие узлам, равноотстоящим от концов отрезка интегрирования a и b, равны между собой:

$$B_j^n = B_{n-j}^n \quad (j = 0, 1, \ldots). \tag{13}$$

Это свойство может быть проверено вычислениями, но его легко можно предвидеть заранее, так как при всякой весовой функции, значения которой распределены симметрично относительно середины отрезка $[a, b]$, в частности при постоянной весовой функции, у коэффициентов B_j^n и B_{n-j}^n нет оснований иметь различные значения.

2) Выше мы обращали внимание на то, что интерполяционная формула квадратур с n узлами характеризуется тем, что она является точной для многочленов степени $n-1$, какова бы ни была весовая функция $p(x)$ и как бы ни были расположены узлы. В формуле (11) имеется $n+1$ узел и она является точной для всяких многочленов степени n. Естественно выяснить, может ли она быть точной для всех многочленов некоторой степени, большей n. Оказывается, что ответ здесь зависит от четности или нечетности числа узлов $n+1$.

Если число узлов $n+1$ является четным, то квадратурная формула (11) не может быть верной для многочленов степени $n+1$, и n является степенью точности правила (11).

Когда число узлов $n+1$ нечетное, то один из узлов располагается на середине $c = \frac{1}{2}(a+b)$ отрезка интегрирования $[a, b]$, остальные же узлы лежат симметрично относительно c. Рассмотрим многочлен $P(x) = (x-c)^{n+1}$, имеющий степень $n+1$. Он является нечетным относительно точки c:

$$P(c-t) = -P(c+t),$$

и для него $\int_a^b P(x)\,dx$ поэтому равен нулю. Ввиду же нечетности $P(x)$ и свойства (13) для коэффициентов B_k^n, для $P(x)$ будет равна нулю правая часть (11) и, следовательно, для $P(x)$ равенство (11) выполняется

точно. Но так как это равенство является точным для всякого многочлена степени n, то оно будет точным для любых многочленов степени $n+1$.

Можно проверить, что (11) не может быть точным для многочленов степени $n+2$. Поэтому $n+1$ является степенью точности формулы (11).

§ 3. Простейшие формулы Ньютона — Котеса и применение их к повышению точности интегрирования путем разделения отрезка на части

Для повышения точности интегрирования отрезок $[a, b]$ часто делят на несколько частей, затем применяют избранную квадратурную формулу к каждой отдельной части и результаты складывают. Этот метод является общим, и им можно пользоваться при применении всякой квадратурной формулы. Он основан на простых соображениях, которые можно пояснить на примере любой формулы. Для многих формул приближенных квадратур погрешность $R_n(f)$ зависит от величины отрезка интегрирования, как будет видно в дальнейшем изложении, по следующему простому закону:

$$R_n(f) = (b-a)^k C(a, b), \tag{1}$$

где $C(a, b)$ есть медленно изменяющаяся функция от a, b, и k есть целое положительное число. Такая зависимость показывает, что если мы уменьшим отрезок интегрирования в p раз, то $R_n(f)$ при этом уменьшится приблизительно в p^k раз, и если k есть достаточно большое число, это уменьшение может быть значительным.

Для вычисления интеграла по всему отрезку $[a, b]$ разделим его на p равных частей и вычислим при помощи выбранной формулы интегралы по всем частичным отрезкам. В каждом случае погрешность будет приблизительно в p^k раз меньше, чем (1). При сложении всех таких интегралов получится результат, погрешность которого будет приблизительно в p^{k-1} раз меньше, чем погрешность (1), когда формула квадратур применяется для вычисления интеграла по всему отрезку $[a, b]$. Если $k > 1$, произойдет уменьшение погрешности тем большее, чем большее k. Описанный

способ увеличения точности применим сейчас к простейшим формулам Ньютона — Котеса.

1. Формула трапеций. Пусть линейное интерполирование выполняется по двум значениям $f(a)$ и $f(b)$, принимаемым функцией f на концах отрезка $[a, b]$. Равенство (2.11) в этом случае имеет форму

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}[f(a)+f(b)]. \tag{2}$$

Это есть известная формула трапеций. Погрешность ее, указываемая в (2.6), ввиду $\omega(x)=(x-a)(x-b)$ и $p(x)\equiv 1$ есть

$$R(f)=\frac{1}{2}\int_a^b (x-a)(x-b)f''(\xi)\,dx.$$

Если вторая производная f'' непрерывна на $[a, b]$, то, так как множитель $(x-a)(x-b)$ сохраняет знак на $[a, b]$, существует на $[a, b]$ такая точка η, для которой

$$R(f)=f''(\eta)\int_a^b (x-a)(x-b)\,dx$$

или, после вычисления интеграла,

$$R(f)=-\frac{(b-a)^3}{12}f''(\eta) \qquad (a\leqslant\eta\leqslant b). \tag{3}$$

Для увеличения точности формулы трапеции (2), разделим $[a, b]$ на n равных частей длины $h=\frac{1}{n}(b-a)$ и рассмотрим частичный отрезок $[a+kh, a+(k+1)h]$. Формула (2) для него дает

$$\int_{a+kh}^{a+(k+1)h} f(x)\,dx=\frac{h}{2}[f_k+f_{k+1}]+R_k \quad [f_k=f(a+kh)],$$

$$R_k=-\frac{h^3}{12}f''(\eta_k), \quad a+kh\leqslant\eta_k\leqslant a+(k+1)h.$$

Сумма интегралов по всем частичным отрезкам даст общую квадратурную формулу трапеций

$$\int_a^b f(x)\,dx = \frac{b-a}{n}\left[\frac{1}{2}f_0 + f_1 + \ldots + f_{n-1} + \frac{1}{2}f_n\right] + R, \quad (4)$$

$$R = R_0 + \ldots + R_{n-1} = -\frac{h^3}{12}[f''(\eta_0) + \ldots + f''(\eta_{n-1})] =$$

$$= -\frac{(b-a)^3}{12n^2}\cdot\frac{1}{n}[f''(\eta_0) + \ldots + f''(\eta_{n-1})].$$

Величина $\frac{1}{n}[f''(\eta_0) + \ldots + f''(\eta_{n-1})]$ есть среднее арифметическое, составленное из n значений второй производной f'', и оно лежит где-то между этими значениями. Вторую производную f'' мы предполагаем непрерывной на $[a, b]$ функцией, и она принимает все промежуточные значения. Существует поэтому такая точка ξ, что

$$R = -\frac{(b-a)^3}{12n^2}f''(\xi) \qquad (a \leqslant \xi \leqslant b).$$

2. Формула парабол. Пусть $n = 2$ и интерполирование f выполняется по трем ее значениям в точках a, $c = {}^1/_2(a+b)$ и b. Интерполирующий многочлен имеет, вообще говоря, вторую степень, и его графиком является парабола. Формула (2) здесь имеет вид

$$\int_a^b f(x)\,dx \approx \frac{b-a}{6}[f(a) + 4f(c) + f(b)] \qquad (5)$$

и называется формулой парабол или формулой Симпсона.

Она точна для всякого многочлена второй степени и так как она, очевидно, является точной для $f(x) = (x-c)^3$ ввиду того, что в этом случае левая и правая части в (5) обращаются в нуль, то она точна для всяких многочленов третьей степени.

Для нахождения погрешности формулы (5) рассмотрим многочлен $P_3(x)$ третьей степени, удовлетворяющий условиям

$$P_3(a) = f(a), \quad P_3(c) = f(c), \quad P_3'(c) = f'(c), \quad P_3'(b) = f(b).$$

Он интерполирует $f(x)$ по значениям в двух однократных узлах a и b и по значениям $f(c)$, $f'(c)$ в двойном узле c:

$$f(x) = P_3(x) + r(x).$$

Для $P_3(x)$ равенство (5) является точным, и поэтому

$$\int_a^b f(x)\,dx = \int_a^b P_3(x)\,dx + \int_a^b r(x)\,dx =$$

$$= \frac{b-a}{6}[P_3(a) + 4P_3(c) + P_3(b)] + \int_a^b r(x)\,dx =$$

$$= \frac{b-a}{6}[f(a) + 4f(c) + f(b)] + \int_a^b r(x)\,dx.$$

Погрешность формулы парабол имеет, следовательно, значение

$$R(f) = \int_a^b r(x)\,dx.$$

В гл. 1, § 5 для некоторых случаев получены представления остаточного члена $r(x)$ интерполирования с кратными узлами. В частности, из (1.5.4), если считать, что f имеет на $[a, b]$ непрерывную производную четвертого порядка, следует для $r(x)$ равенство

$$r(x) = \frac{1}{4!}(x-a)(x-c)^2(x-b)f^{(4)}(\xi) \qquad (a \leqslant x,\ \xi \leqslant b).$$

Поэтому

$$R(f) = \frac{1}{24}\int_a^b (x-a)(x-c)^2(x-b)f^{(4)}(\xi)\,dx. \qquad (6)$$

Так как множитель $(x-a)(x-c)^2(x-b)$ не изменяет знак на отрезке $[a, b]$ и $f^{(4)}$ есть непрерывная функция при $a \leqslant x \leqslant b$, на $[a, b]$ существует точка η такая,

что

$$R(f)=\frac{1}{24}f^{(4)}(\eta)\int_a^b(x-a)(x-c)^2(x-b)\,dx=$$
$$=-\frac{1}{90}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\eta). \quad (7)$$

Получим теперь общую формулу парабол. Разделим $[a, b]$ на четное число n равных частей длины $h=\frac{1}{n}(b-a)$ и возьмем сдвоенный частичный отрезок $[a+(k-1)h,\ a+(k+1)h]$. Формула парабол, примененная к нему, будет следующей:

$$\int_{a+(k-1)h}^{a+(k+1)h} f(x)\,dx=\frac{h}{3}[f_{k-1}+4f_k+f_{k+1}]+R_k$$
$$[f_k=f(a+kh)].$$

Если такие равенства составить для отрезков $[a, a+2h]$, $[a+2h, a+4h]$, ... и их сложить, получим общую формулу парабол

$$\int_a^b f(x)\,dx=\frac{b-a}{3n}[f_0+f_n+2(f_2+f_4+\ \dots\ +f_{n-2})+$$
$$+4(f_1+f_3+\ \dots\ +f_{n-1})]+R(f), \quad (8)$$
$$R(f)=-\frac{1}{90}h^5[f^{(4)}(\eta_1)+f^{(4)}(\eta_3)+\dots+f^{(4)}(\eta_{n-1})].$$

Ввиду непрерывности $f^{(4)}$ на $[a, b]$ существует точка ξ такая, что

$$\frac{2}{n}[f^{(4)}(\eta_1)+\ \dots\ +f^{(4)}(\eta_{n-1})]=f^{(4)}(\xi).$$

Для $R(f)$, следовательно, верно равенство

$$R(f)=-\frac{(b-a)^5}{180n^4}f^{(4)}(\xi) \qquad (a\leqslant\xi\leqslant b). \quad (9)$$

3. Формула «трех восьмых». При $n=3$ для построения формулы Ньютона — Котеса (2.11) интерполирование f выполняется по значениям ее в точках a, $a+\frac{1}{3}H$,

$a+\frac{2}{3}H$, b $(H=b-a)$. Формула будет следующей:

$$\int_a^b f(x)\,dx \approx H\left[\frac{1}{8}f(a)+\frac{3}{8}f\left(a+\frac{H}{3}\right)+\right.$$
$$\left.+\frac{3}{8}f\left(a+\frac{2H}{3}\right)+\frac{1}{8}f(b)\right]. \quad (10)$$

Ее часто называют «формулой трех восьмых». Степень точности формулы равна трем.

Можно показать, проделав рассуждения немного более сложные, чем в случае формулы Симпсона, что погрешность формулы (10) может быть, если $f^{(4)}$ есть непрерывная и мало меняющаяся на $[a, b]$ функция, представлена в виде

$$R^*(f) \approx -\frac{(b-a)^5}{6480}f^{(4)}(\eta), \quad a \leqslant \eta \leqslant b. \quad (11)$$

Остановимся еще на применении формулы трех восьмых к построению формул приближенного интегрирования при большом числе равноотстоящих узлов.

Пусть n есть число, кратное трем. Разделим $[a, b]$ на n равных частей длины $h=\frac{1}{n}(b-a)$. Возьмем строенный отрезок $[a+kh, a+(k+3)h]$ и к интегрированию по нему применим правило (10):

$$\int_{a+kh}^{a+(k+3)h} f(x)\,dx = \frac{3h}{8}\{f(a+kh)+3f[a+(k+1)h]+$$
$$+3f[a+(k+2)h]+f[a+(k+3)h]\}+R_{k+1}(f),$$
$$R_{k+1}(f) \approx -\frac{(3h)^5}{6480}f^{(4)}(\eta_{k+1}).$$

Если такие равенства записать для всех строенных отрезков $[a, a+3h]$, $[a+3h, a+6h]$, ... и сложить их почленно, построим общую формулу трех восьмых:

$$\int_a^b f(x)\,dx = \frac{3h}{8}[(f_0+f_n)+2(f_3+f_6+\ldots+f_{n-3})+$$
$$+3(f_1+f_2+f_4+f_5+\ldots+f_{n-2}+f_{n-1})]+R(f), \quad (12)$$
$$[f_k=f(a+kh)].$$

Погрешность $R(f)$ имеет следующее значение:

$$R(f) = R_1(f) + R_2(f) + \ldots + R_{n/3}(f) \approx$$
$$\approx -\frac{(3h)^5}{6480}[f^{(4)}(\eta_1) + \ldots + f^{(4)}(\eta_{n/3})] =$$
$$= -\frac{(b-a)^5}{80n^4} \cdot \frac{3}{n}[f^{(4)}(\eta_1) + \ldots + f^{(4)}(\eta_{n/3})].$$

Множитель $\frac{3}{n}[f^{(4)}(\eta_1) + \ldots + f^{(4)}(\eta_{n/3})]$ есть среднее арифметическое, составленное из $n/3$ значений четвертой производной, и так как $f^{(4)}$ считается непрерывной функцией, то на $[a, b]$ существует точка ξ такая, что этот множитель равен $f(\xi)$. Для погрешности $R(f)$ окончательно получится выражение

$$R(f) \approx -\frac{(b-a)^5}{80n^4} f(\xi), \quad a \leqslant \xi \leqslant b. \tag{13}$$

Представляет интерес сравнение формул парабол и «трех восьмых» по точности. Когда число отрезков кратно 6, для вычисления интеграла могут быть взяты обе эти формулы. Они имеют одинаковую степень точности, и вопрос о выборе одной из них здесь решается возможной погрешностью результата. Сравнение погрешностей (9) и (13) этих формул говорит о том, что при применении «правила трех восьмых» можно ожидать погрешности, большей приблизительно в два раза по сравнению с формулой парабол. Это заставляет отдать предпочтение формуле (8) и применять правило трех восьмых лишь как контрольное. Но необходимо отметить, что последнее правило имеет самостоятельное значение, так оно может быть применено в случае нечетного числа узлов, кратного трем, тогда как формула парабол здесь неприменима.

§ 4. Квадратурные формулы наивысшей алгебраической степени точности

1. Некоторые общие понятия и теоремы. Как и выше, будем рассматривать квадратурную формулу вида

$$\int_a^b p(x) f(x)\, dx \approx \sum_{k=1}^{n} A_k f(x_k). \tag{1}$$

Весовую функцию предположим такой, чтобы интегралы $\int_a^b p(x)\,x^m\,dx$ были абсолютно сходящимися при $m = 0, 1, 2, \ldots$, и, кроме того, будем считать вес $p(x)$ не равным тождественно нулю:

$$\int_a^b |p(x)|\,dx > 0.$$

Формула (1) имеет $2n$ параметров A_k и x_k, и можно ожидать, что при помощи выбора их можно сделать равенство точным для всяких алгебраических многочленов степени $2n-1$ или, что равносильно, чтобы оно было точным для степеней x от нулевой до $2n-1$:

$$\int_a^b p(x)\,x^m\,dx = \sum_{k=1}^{n} A_k x_k^m \quad (m = 0, 1, \ldots, 2n-1).$$

Выясним сначала условия, при которых равенство (1) будет точным для всех многочленов степени $2n-1$.

Теорема 1. *Для того чтобы формула* (1) *была точной для всех многочленов степени* $2n-1$, *необходимо и достаточно выполнение условий*:

1) *формула* (1) *должна быть интерполяционной, т. е. ее коэффициенты* A_k *должны иметь значения, указанные в* (2.2);

2) *узлы* x_k *формулы* (1) *должны быть такими, чтобы многочлен* $\omega(x) = (x-x_1)(x-x_2)\ldots(x-x_n)$ *был ортогонален с весом* $p(x)$ *ко всякому многочлену* $Q(x)$ *степени меньше* n:

$$\int_a^b p(x)\,\omega(x)\,Q(x)\,dx = 0. \tag{2}$$

Доказательство. Убедимся в необходимости условий. Если равенство (1) является точным для всякого многочлена степени $2n-1$, то оно точно и для многочленов степени $n-1$, а тогда по теореме 1 § 2 формула должна быть интерполяционной, и первое условие должно выполняться.

Предположим теперь, что $Q(x)$ есть любой многочлен степени $n-1$. Произведение $\omega(x)Q(x)=f(x)$ является многочленом степени $2n-1$, и для него (1) должно выполняться точно. Но $f(x_k)=\omega(x_k)Q(x_k)=0$ ввиду $\omega(x_k)=0$, и правая часть (1) обращается в нуль, что приводит к (2).

Проверим достаточность условий. Возьмем произвольный многочлен $f(x)$ степени $2n-1$. Разделим его на $\omega(x)$ по обычным правилам и представим f в форме $f(x)=\omega(x)Q(x)+r(x)$, где Q и r есть многочлены степеней не выше $n-1$. Так как $\omega(x_k)=0$, то, очевидно, $f(x_k)=r(x_k)$ $(k=1, 2, \ldots, n)$;

$$\int_a^b p(x)f(x)\,dx=\int_a^b p(x)\omega(x)Q(x)\,dx+\int_a^b p(x)r(x)\,dx.$$

Первый из интегралов, стоящих в равенстве справа, равен нулю по второму условию. Степень $r(x)$ не выше $n-1$, и для него формула (1) должна выполняться точно по первому условию:

$$\int_a^b p(x)r(x)\,dx=\sum_{k=1}^{n}A_k r(x_k),$$

а так как $r(x_k)=f(x_k)$, то точным является равенство

$$\int_a^b p(x)f(x)\,dx=\sum_{k=1}^{n}A_k f(x_k)$$

и формула (1) для f действительно выполняется точно.

Выясним вопрос о существовании многочлена $\omega(x)$, удовлетворяющего условию ортогональности (2). Сейчас удобнее рассматривать не корни x_k многочлена $\omega(x)$, а его разложение по степеням x и коэффициенты разложения

$$\omega(x)=x^n+a_1x^{n-1}+a_2x^{n-2}+\ldots+a_n. \qquad (3)$$

Теорема 2. *Пусть весовая функция $p(x)$ не изменяет знак на отрезке $[a, b]$ и является неотрицательной. Тогда при всяком n существует и при этом единственный многочлен $\omega(x)$ вида (3), ортогональный по весу $p(x)$ ко всякому многочлену степени меньшей n.*

Доказательство. Ортогональность к любому многочлену степени меньшей n равносильна ортогональности к $1, x, x^2, \ldots, x^{n-1}$. Условия же ортогональности к этим степеням дадут следующую систему линейных уравнений для нахождения коэффициентов $a_1, \ldots, a_n$:

$$\int_a^b p(x)\,[x^n + a_1 x^{n-1} + \ldots + a_n]\,x^i\,dx = 0 \qquad (4)$$

$$(i = 0, 1, \ldots, n-1).$$

Чтобы убедиться в существовании и единственности у нее решения, достаточно проверить, что однородная система

$$\int_a^b p(x)\,[a_1 x^{n-1} + \ldots + a_n]\,x^i\,dx = 0 \qquad (i = 0, 1, \ldots, n-1)$$

имеет только нулевое решение. Если выписать эти уравнения для $i = 0, 1, \ldots, n-1$, умножить их последовательно на $a_n, a_{n-1}, \ldots, a_1$ и результаты сложить, получится равенство

$$\int_a^b p(x)\,[a_1 x^{n-1} + \ldots + a_n]^2\,dx = 0.$$

Если бы многочлен $a_1 x^{n-1} + \ldots + a_n$ был отличен от нуля, он обращался бы в нуль не больше чем в $n-1$ точках. Но тогда полученное равенство не могло бы выполняться, так как $p(x) \geqslant 0$ и $\int_a^b p(x)\,dx > 0$. Поэтому многочлен тождественно равен нулю, равны нулю все его коэффициенты, и у однородной системы существует только нулевое решение.

Узлы x_k квадратурной формулы предполагаются лежащими на отрезке интегрирования $[a, b]$, и нам осталось проверить, будут ли корни многочлена $\omega(x)$, найденного по условиям ортогональности (4), принадлежать этому отрезку.

Теорема 3. *Пусть весовая функция знакопостоянна на* $[a, b]$ *и многочлен* $\omega(x)$ *ортогонален на* $[a, b]$ *с весом* $p(x)$ *ко всякому многочлену* $Q(x)$ *степени не выше*

$n-1$. *Тогда корни многочлена* $\omega(x)$ *все лежат внутри* $[a, b]$ *и различны между собой.*

Доказательство. Пусть многочлен $\omega(x)$ имеет всего m различных корней, лежащих внутри $[a, b]$ и имеющих нечетную кратность. Назовем эти корни $\xi_1, \ldots, \xi_m$. Достаточно показать, что $m = n$. Так как $\omega(x)$ может иметь не более n различных корней, из равенства $m = n$ будет следовать, что все корни ξ_i $(i = 1, \ldots, m)$ однократные, и никаких других корней у $\omega(x)$ нет.

Предположим, что $m < n$, и рассмотрим многочлен

$$\rho(x) = (x - \xi_1) \ldots (x - \xi_m).$$

Его степень m меньше n, и по свойству ортогональности должно быть

$$\int_a^b p(x)\,\omega(x)\,\rho(x)\,dx = 0. \tag{5}$$

С другой стороны, ω и ρ имеют внутри $[a, b]$ одинаковые точки перемены знака, произведение $\omega\rho$ не изменяет знака внутри $[a, b]$ и, кроме того, оно обращается в нуль лишь в конечном числе точек. Поэтому интеграл (5) не может быть равен нулю, так как $p(x)$ также сохраняет знак на $[a, b]$ и $\int_a^b p(x)\,dx \neq 0$.

Из теорем 2 и 3 вытекает, что при условии сохранения на $[a, b]$ знака весовой функцией $p(x)$, квадратурная формула вида (1), точная для всех многочленов степени $2n-1$, может быть построена при любых $n = 1, 2, \ldots$, и такая формула является единственной.

Покажем также, что при том же условии сохранения знака весовой функцией $p(x)$ число $2n-1$ является наивысшей алгебраической степенью точности формулы (1).

Теорема 4. *Если вес* $p(x)$ *сохраняет знак на* $[a, b]$, *то ни при каких* A_k *и* x_k *равенство* (1) *не может быть точным для всех многочленов степени* $2n$.

Доказательство. По узлам x_k построим $\omega(x) = (x - x_1) \ldots (x - x_n)$ и рассмотрим многочлен $f(x) = \omega^2(x)$. Он положителен всюду на $[a, b]$, кроме точек x_k. Инте-

грал $\int_a^b p(x) f(x)\,dx$ не может быть равен нулю, так как произведение pf сохраняет знак на $[a, b]$ и $\int_a^b p(x)\,dx \neq 0$. Правая же часть (1) равна нулю ввиду $f(x_k) = 0$ $(k = 1, 2, \ldots, n)$. Поэтому равенство в (1) для $f = \omega^2$ не может быть точным.

2. О положительности квадратурных коэффициентов. Знаки коэффициентов устанавливаются на основании следующей теоремы.

Теорема 5. *Если* $p(x) \geqslant 0$ $[x \in [a, b]]$ *и квадратурная формула* (1) *имеет наивысшую алгебраическую степень точности* $2n - 1$, *то все ее коэффициенты* A_k *положительны.*

Формулированная теорема является простым следствием приводимой ниже простой леммы.

Лемма 1. *Если* $p(x) \geqslant 0$ $[x \in [a, b]]$ *и равенство* (1) *является точным для следующих многочленов степени* $2n - 2$:

$$f_i(x) = \left[\frac{\omega(x)}{x - x_i}\right]^2 \qquad (i = 1, 2, \ldots, n),$$

то все квадратурные коэффициенты A_k *равенства положительны.*

Доказательство. Так как в узлах x_k многочлены $f_i(x)$ принимают значения

$$f_i(x_k) = \begin{cases} [\omega'(x_i)]^2, & k = i, \\ 0, & k \neq i, \end{cases}$$

то равенство (1) для $f_i(x)$ приводит к результату

$$0 < \int_a^b p(x) f_i(x)\,dx = \sum_{k=1}^{n} A_k f_i(x_k) = A_i [\omega'(x_i)]^2.$$

Отсюда следует положительность A_i.

3. Погрешность квадратуры наивысшей степени точности. Будет получена лишь простейшая теорема о представлении погрешности квадратуры наивысшей степени

точности, предполагающая достаточно высокий порядок дифференцируемости f.

Теорема 6. *Пусть весовая функция сохраняет знак на* $[a, b]$ *и* f *имеет на* $[a, b]$ *непрерывную производную порядка* $2n$. *Тогда на* $[a, b]$ *существует такая точка* ξ, *что для погрешности*

$$R(f) = \int_a^b p(x) f(x)\, dx - \sum_{k=1}^{n} A_k f(x_k)$$

квадратуры наивысшей алгебраической степени точности верно представление

$$R(f) = \frac{1}{(2n)!} f^{(2n)}(\xi) \int_a^b p(x)\, \omega^2(x)\, dx. \tag{6}$$

Доказательство. Пусть $H(x)$ есть многочлен степени $2n-1$, интерполирующий $f(x)$ по условиям

$$H(x_k) = f(x_k), \quad H'(x_k) = f'(x_k) \qquad (k = 1, \ldots, n).$$

Это есть интерполирование с двукратными узлами. В § 5 гл. 1 было рассмотрено интерполирование с узлами любой кратности. В нашей задаче, если предположить, что f имеет на $[a, b]$ непрерывную производную порядка $2n$, остаточный член такой интерполяции может быть записан в виде

$$r(x) = \frac{1}{(2n)!} \omega^2(x) f^{(2n)}(\eta),$$

где η есть некоторая точка отрезка, содержащего x и x_k $(k = 1, \ldots, n)$. Преобразуем искомый интеграл

$$\int_a^b p(x) f(x)\, dx =$$

$$= \int_a^b p(x) H(x)\, dx + \frac{1}{(2n)!} \int_a^b p(x)\, \omega^2(x) f^{(2n)}(\eta)\, dx. \tag{7}$$

Так как квадратурная формула верна для всяких многочленов степени $2n-1$ и $H(x_k) = f(x_k)$, то

$$\int_a^b p(x) H(x)\, dx = \sum_{k=1}^{n} A_k H(x_k) = \sum_{k=1}^{n} A_k f(x_k).$$

Следовательно, погрешность приближенного интегрирования имеет значение

$$R(f) = \frac{1}{(2n)!} \int_a^b p(x)\,\omega^2(x)\,f^{(2n)}(\eta)\,dx.$$

Отсюда при помощи обычных рассуждений, которые несколько раз проделывались в предыдущем изложении, можно прийти к заключению о том, что существует точка $\xi \in [a, b]$, для которой верно равенство (6).

4. Замечание о связи с ортогональной системой многочленов. Говорят, что последовательность многочленов $P_0(x)$, $P_1(x)$, ..., $P_n(x)$, ... $[P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0,\ a_n \neq 0]$ образует ортогональную систему по весу $p(x)$ на $[a, b]$, если

$$\int_a^b p(x)\,P_m(x)\,P_n(x)\,dx = 0, \qquad m \neq n. \tag{8}$$

Многочлен $P_n(x)$ называют нормированным, когда

$$\int_a^b p(x)\,P_n^2(x)\,dx = 1, \qquad a_n > 0.$$

Если все многочлены ортогональной системы нормированы, то систему сокращенно называют ортонормированной.

В условии ортогональности (8), ввиду равноправности индексов m и n, достаточно требовать его выполнения для $m < n$. С другой стороны, так как всякий многочлен $Q(x)$ степени не выше $n-1$ может быть разложен по многочленам $P_m(x)$ $(m = 0, 1, \ldots, n-1)$, условие ортогональности (8) равносильно требованию, чтобы каждый многочлен $P_n(x)$ системы был ортогонален к любым многочленам степени $m < n$.

Возвратимся к квадратурным формулам наивысшей степени точности. Такая формула может быть построена для каждого $n = 1, 2, \ldots$ Когда n фиксировано, ей будут отвечать свой многочлен ω_n и свои квадратурные узлы x_k^n $(k = 1, \ldots, n)$

$$\omega_n(x) = (x - x_1^n) \ldots (x - x_n^n).$$

По теореме 1 он должен быть ортогонален ко всяким многочленам $Q(x)$ степени $m < n$, следовательно, он может отличаться от $P_n(x)$ только численным множителем, который, очевидно, должен быть равен коэффициенту при старшей степени $P_n(x)$:

$$P_n(x) = a_n\omega_n(x) \quad (n = 1, 2, \ldots).$$

Поэтому узлы x_k^n квадратурной формулы должны быть корнями $P_n(x)$.

§ 5. Квадратурные формулы, отвечающие простейшим весовым функциям

Здесь будут рассмотрены квадратурные формулы наивысшей алгебраической степени точности, отвечающие классическим весам: постоянному весу $p(x) \equiv 1$, весу Якоби $p(x) = (b-x)^\alpha(x-a)^\beta$ $[a < x < b]$, весам Лагерра $p(x) = x^\alpha e^{-x}$ $(0 < x < \infty)$ и Эрмита $p(x) = e^{-x^2}$ $(-\infty < x < \infty)$. Эти весовые функции позволяют заранее учитывать наиболее часто встречающиеся в приложениях особенности у интегрируемых функций F путем представления их в виде произведения $F(x) = p(x)f(x)$.

Постоянная весовая функция выбирается в том случае, когда $F(x)$ не имеет особенностей и является достаточно гладкой как внутри отрезка $[a, b]$, так и на его концах. Функция Якоби позволяет учитывать степенные особенности $F(x) = (b-x)^\alpha(x-a)^\beta f(x)$ на концах a и b отрезка интегрирования или на одном из его концов при достаточной гладкости F внутри отрезка. Вес Лагерра учитывает степенную особенность в точке $x = 0$ и связан со скоростью убывания $F(x) = x^\alpha f(x)$ при $x \to \infty$. Наконец, вес Эрмита связан со скоростью убывания $F(x) = e^{-x^2}f(x)$ при стремлении x к $+\infty$ и $-\infty$ в случае интегрирования по всей числовой оси.

1. Постоянная весовая функция. Отрезок интегрирования $[a, b]$ предполагается конечным. Линейным преобразованием аргумента его можно перевести в $[-1, 1]$. Интеграл берется в форме

$$\int_{-1}^{1} f(x)\,dx, \tag{1}$$

при этом f предполагается достаточно гладкой функцией всюду на $[-1, 1]$, включая его концы.

Известно, что ортогональную на $[-1, 1]$ систему многочленов с постоянным весом образуют многочлены Лежандра

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n}(x^2 - 1)^n = \frac{(2n)!}{2^n (n!)^2} x^n - \ldots \quad (2)$$

Они нормированы не на единицу, так как

$$\int_{-1}^{1} P_n^2(x)\, dx = \frac{2}{2n+1}. \quad (3)$$

В формуле квадратур с n узлами

$$\int_{-1}^{1} f(x)\, dx \approx \sum_{k=1}^{n} A_k f(x_k), \quad (4)$$

имеющей наивысшую степень точности $2n - 1$ и впервые полученной Гауссом, узлы x_k должны быть корнями многочлена Лежандра степени n:

$$P_n(x_k) = 0 \quad (k = 1, \ldots, n).$$

Коэффициенты A_k могут быть вычислены при помощи равенства (2.2), но для них существует более простое выражение через многочлен Лежандра, которое мы приведем без вывода:

$$A_k = \frac{2(1 - x_k^2)}{n^2 P_{n-1}^2(x_k)}.$$

Соответствующий формуле (3) многочлен $\omega(x)$ отличается от $P_n(x)$ постоянным множителем, равным обратной величине старшего коэффициента в $P_n(x)$:

$$\omega(x) = \frac{2^n (n!)^2}{(2n)!} P_n(x).$$

Если воспользоваться этим равенством и (3), то можно на основании (4.6) сказать, что в случае существования у f непрерывной на $[-1, 1]$ производной по-

рядка $2n$ погрешность $R(f)$ формулы (4) представима в виде

$$R(f)=\frac{2^{2n+1}}{(2n+1)(2n)!}\left[\frac{(n!)^2}{(2n)!}\right]^2 f^{(2n)}(\xi) \qquad (-1\leqslant\xi\leqslant 1). \tag{5}$$

Подробные таблицы A_k и x_k можно найти в книгах [3] и [4].

2. Интегралы вида $\int_a^b (b-x)^\alpha (x-a)^\beta f(x)\,dx$. Линейным

преобразованием $x=\frac{1}{2}(a+b)+\frac{1}{2}(b-a)t$, $-1\leqslant t\leqslant 1$, отрезок интегрирования $[a, b]$ приводится к каноническому отрезку $[-1, 1]$, и достаточно рассмотреть интеграл вида

$$\int_{-1}^{1}(1-x)^\alpha (1+x)^\beta f(x)\,dx \qquad (\alpha,\ \beta>-1).$$

Ортогональными на $[-1, 1]$ с весом $p(x)=(1-x)^\alpha(1+x)^\beta$ являются многочлены Якоби

$$P_n^{(\alpha,\beta)}(x)=$$
$$=\frac{(-1)^n}{2^n n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}\left[(1-x)^{\alpha+n}(1+x)^{\beta+n}\right]=$$
$$=\frac{\Gamma(\alpha+\beta+2n+1)}{2^n n!\,\Gamma(\alpha+\beta+n+1)}x^n-\ \ldots \tag{6}$$

В квадратурной формуле наивысшей степени точности

$$\int_{-1}^{1}(1-x)^\alpha(1+x)^\beta f(x)\,dx\approx\sum_{k=1}^{n}A_k f(x_k) \tag{7}$$

узлы располагаются в корнях многочлена Якоби степени n:

$$P_n^{(\alpha,\beta)}(x_k)=0, \qquad k=1,\ 2,\ \ldots,\ n.$$

Коэффициенты A_k могут быть найдены по общей формуле (2.2), но для них известно более удобное для вычислений выражение через многочлен Якоби $P_n^{(\alpha, \beta)}(x)$:

$$A_k = 2^{\alpha+\beta+1} \frac{\Gamma(\alpha+n+1)\,\Gamma(\beta+n+1)}{n!\,\Gamma(\alpha+\beta+n+1)\left(1-x_k^2\right)\left[P_n^{(\alpha, \beta)'}(x_k)\right]^2}. \qquad (8)$$

Соответствующий формуле (7) многочлен $\omega(x)$ связан с $P_n^{(\alpha, \beta)}(x)$ равенством

$$\omega(x) = \frac{2^n n!\,\Gamma(\alpha+\beta+n+1)}{\Gamma(\alpha+\beta+2n+1)} P_n^{(\alpha, \beta)}(x).$$

Известно, что для многочлена Якоби имеет место соотношение

$$\int_{-1}^{1} (1-x)^{\alpha}(1+x)^{\beta}\left[P_n^{(\alpha, \beta)}(x)\right]^2 dx = \frac{2^{\alpha+\beta+1}\,\Gamma(\alpha+n+1)\,\Gamma(\beta+n+1)}{(\alpha+\beta+2n+1)\,\Gamma^2(\alpha+\beta+2n+1)}.$$

Два последних равенства позволяют вычислить для формулы (7) интеграл, который входит в общее представление (4.6) погрешности приближенной квадратуры наивысшей степени точности. После простых вычислений получим

$$R(f) = \frac{f^{(2n)}(\xi)}{(2n)!} \frac{2^{\alpha+\beta+2n+1}\,\Gamma(\alpha+n+1)\,\Gamma(\beta+n+1)\,\Gamma(\alpha+\beta+n+1)}{(\alpha+\beta+2n+1)\,\Gamma^2(\alpha+\beta+2n+1)}, \quad -1 \leqslant \xi \leqslant 1. \qquad (9)$$

Равенство (7) содержит два произвольных параметра α и β и является, по сути дела, записью семейства квадратурных формул. Придавая α и β численные значения, будем из (7) получать частные формулы, учитывающие те или иные особенности поведения интегрируемой функции в точках $x = \pm 1$. Например, формула Гаусса (4), рассчитанная на интегрирование функций, не имеющих особенностей на $[a, b]$, получается из (7) при $\alpha = \beta = 0$. Остановимся на частном случае, когда $\alpha = \beta = -\frac{1}{2}$ и $p(x) = (1-x^2)^{-1/2}$.

Многочлен Якоби $P_n^{(-0,5;\ -0,5)}(x)$ только постоянным множителем отличается от многочлена Чебышева первого рода:

$$P_n^{(-0,5;\ -0,5)}(x) = CT_n(x) = C\cos(n \arccos x).$$

Узлы x_k должны быть нулями многочлена $T_n(x)$ и имеют значения

$$x_k = \cos\frac{2k-1}{2n}\pi \qquad (k = 1, \ldots, n).$$

Коэффициенты A_k можно найти с помощью равенства (8), и они оказываются следующими:

$$A_k = \frac{\Gamma^2\left(n+\frac{1}{2}\right)}{n!\Gamma(2n)\,C^2n^2}.$$

Правая часть равенства не зависит от k, и все коэффициенты оказываются одинаковыми. Общую величину их назовем A. Ее наиболее просто можно найти, если вспомнить, что квадратурная формула должна быть точной для функции $f \equiv 1$:

$$\sum_{k=1}^{n} A_k = nA = \int_{-1}^{1} \frac{dx}{\sqrt{1-x^2}} = \pi, \qquad A = \frac{\pi}{n}.$$

Поэтому квадратурная формула наивысшей степени точности, отвечающая весовой функции $p(x) = (1 - x^2)^{-1/2}$, имеет вид

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx = \frac{\pi}{n}\sum_{k=1}^{n} f\left(\cos\frac{2k-1}{2n}\pi\right) + R(f), \qquad (10)$$

$$R(f) = \frac{\pi}{2^{2n-1}(2n)!} f^{(2n)}(\xi) \qquad (-1 \leqslant \xi \leqslant 1). \qquad (11)$$

В связи с этой формулой П. Л. Чебышев рассмотрел вопрос о построении квадратурных формул с равными коэффициентами для произвольной весовой функции $p(x)$. Некоторые результаты, полученные в этом направлении, будут даны в одном из следующих параграфов.

3. Интегралы вида $\int_0^\infty x^\alpha e^{-x} f(x)\,dx$. Ортогональную систему на полуоси $[0, \infty)$ с весом $p(x) = x^\alpha e^{-x}$ $(\alpha > -1)$ образуют многочлены Чебышева — Лагерра

$$L_n^{(\alpha)}(x) = (-1)^n x^{-\alpha} e^x \frac{d^n}{dx^n}\left(x^{\alpha+n} e^{-x}\right) = \\ = x^n - \frac{n(n+\alpha)}{1!} x^{n-1} + \ldots$$

В формуле приближенного интегрирования наивысшей степени точности

$$\int_0^\infty x^\alpha e^{-x} f(x)\,dx = \sum_{k=1}^{n} A_k f(x_k) + R(f) \qquad (12)$$

за узлы x_k должны быть взяты корни многочлена $L_n^{(\alpha)}$:

$$L_n^{(\alpha)}(x_k) = 0 \quad (k = 1, \ldots, n).$$

Для вычисления коэффициентов A_k здесь также известна более простая, чем (2.2), формула

$$A_k = \frac{\Gamma(n+1)\,\Gamma(\alpha+n+1)}{x_k \left[L_n^{(\alpha)'}(x_k)\right]^2} = \frac{x_k \Gamma(n)\,\Gamma(\alpha+n)}{n(\alpha+n)\left[L_{n-1}^{(\alpha)}(x_k)\right]^2}.$$

Укажем еще представление погрешности $R(f)$. Для этого воспользуемся ее общим выражением (6). Для формулы (12) многочлены $\omega(x)$ и $L_n^{(\alpha)}(x)$ совпадают, так как старшие коэффициенты у них равны единице. Корни x_k у них одинаковые. Кроме того, $p(x) = x^\alpha e^{-x}$. Поэтому

$$\int_0^\infty p(x)\,\omega^2(x)\,dx = \int_0^\infty x^\alpha e^{-x} \left[L_n^{(\alpha)}(x)\right]^2 dx.$$

Численное значение последнего интеграла находится легко, и в теории многочленов Лагерра известно, что оно равно $n!\Gamma(\alpha+n+1)$. Поэтому, если f имеет непрерывную производную порядка $2n$ на $[0, \infty)$, то для $R(f)$ верно равенство

$$R(f) = \frac{n!\Gamma(\alpha+n+1)}{(2n)!} f^{(2n)}(\xi) \quad (0 \leqslant \xi < \infty). \qquad (13)$$

4. Интегралы вида $\int_{-\infty}^{\infty} e^{-x^2} f(x)\,dx$. На всей числовой оси с весом $p(x) = e^{-x^2}$ ортогональными являются многочлены Чебышева — Эрмита

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2} = 2^n x^n - \ldots$$

В формуле приближенного интегрирования

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)\,dx = \sum_{k=1}^{n} A_k f(x_k) + R(f), \tag{14}$$

имеющей степень точности $2n-1$, узлы x_k должны быть корнями многочлена $H_n(x)$:

$$H_n(x_k) = 0.$$

Многочлен $\omega(x)$ для формулы (14) отличается от $H_n(x)$ численным множителем 2^{-n}:

$$\omega(x) = 2^{-n} H_n(x).$$

Как и во всех предыдущих случаях, для вычисления коэффициентов A_k известно представление более удобное для вычислений, чем (2.2):

$$A_k = \frac{2^{n+1} n! \sqrt{\pi}}{H_n'^2(x_k)} = \frac{2^{n-1}(n-1)! \sqrt{\pi}}{n H_{n-1}^2(x_k)}.$$

Для нахождения выражения погрешности $R(f)$ в формуле (14) воспользуемся вновь равенством (4.6) и, кроме того, известным в теории многочленов $H_n(x)$ соотношением $\int_{-\infty}^{\infty} e^{-x^2} H_n^2(x)\,dx = 2^n n! \sqrt{\pi}$. Тогда

$$\int_a^b p(x)\,\omega^2(x)\,dx = 2^{-2n} \int_{-\infty}^{\infty} e^{-x^2} H_n^2(x)\,dx = \frac{n!\sqrt{\pi}}{2^n},$$
$$R(f) = \frac{n!\sqrt{\pi}}{2^n (2n)!} f^{(2n)}(\xi), \quad -\infty < \xi < \infty. \tag{15}$$

Для формул (12) и (14) таблицы значений узлов x_k и коэффициентов A_k можно найти в книгах [3] и [4].

§ 6. Формулы численного интегрирования, содержащие заранее предписанные узлы

С такими формулами приходится встречаться, например, в следующих задачах. Рассмотрим для дифференциального уравнения граничную или даже многоточечную задачу с заданными значениями функции в нескольких точках рассматриваемого отрезка числовой оси. Такие задачи во многих случаях могут быть сведены к решению интегральных уравнений. К последним же для приведения их к системе численных уравнений может быть применен метод квадратур, состоящий в том, что интегралы, входящие в уравнение, вычисляются посредством какой-либо из квадратурных формул. Но тогда при построении квадратурной формулы естественно воспользоваться теми точками, в которых известны значения функции, приняв эти точки за узлы формулы и взяв еще несколько узлов, выбором которых можно распоряжаться для увеличения точности.

1. Содержание задачи и общие теоремы. Будет рассматриваться квадратурная формула

$$\int_a^b p(x) f(x)\, dx \approx \sum_{k=1}^{n} A_k f(x_k) + \sum_{i=1}^{m} B_i f(a_i), \qquad (1)$$

которая содержит m фиксированных узлов a_i $(i = 1, \ldots, m)$. В нее входят также $2n + m$ произвольных параметров A_k, x_k $(k = 1, \ldots, n)$ и B_i $(i = 1, \ldots, m)$. Их можно надеяться выбрать так, чтобы формула (1) была точной для любых многочленов степени меньшей $2n + m$.

Известно (см. § 2, теорему 1), что при всяких x_k и a_i равенство (1) можно сделать точным для всех многочленов степени $n + m - 1$ при помощи выбора A_k и B_i; для этого необходимо сделать формулу (1) интерполяционной. Для (1) это означает, что ее коэффициенты

должны иметь следующие значения:

$$A_k = \int_a^b p(x) \frac{\omega(x)\,\Omega(x)}{(x-x_k)\,\omega'(x_k)\,\Omega(x_k)}\,dx, \qquad k = 1, \ldots, n,$$
$$B_i = \int_a^b p(x) \frac{\omega(x)\,\Omega(x)}{(x-a_i)\,\omega(a_i)\,\Omega'(a_i)}\,dx, \qquad i = 1, \ldots, m, \tag{2}$$
$$\omega(x) = (x-x_1)\ldots(x-x_n), \quad \Omega(x) = (x-a_1)\ldots(x-a_m).$$

Теорема 1. *Чтобы формула* (1) *была точной для многочленов степени* $2n+m-1$, *необходимо и достаточно выполнение условий*:

1) *формула является интерполяционной, т. е. ее коэффициенты* A_k *и* B_i *имеют значения* (2);

2) *узлы* x_k $(k = 1, \ldots, n)$ *таковы, что соответствующий им многочлен* $\omega(x)$ *ортогонален с весом* $p(x)\Omega(x)$ *на* $[a, b]$ *ко всякому многочлену* $Q(x)$ *степени меньшей* n:

$$\int_a^b p(x)\,\Omega(x)\,\omega(x)\,Q(x)\,dx = 0. \tag{3}$$

Для проверки справедливости утверждений теоремы достаточно повторить с несущественными изменениями — с заменой веса $p(x)$ на вес $p(x)\Omega(x)$ и квадратурной суммы (4.1) на сумму (1) — доказательство теоремы 1 § 5.

Рассмотренная теорема 1 требует существования многочлена $\omega(x)$, обладающего свойством ортогональности (3). Кроме того, по существу задачи корни $\omega(x)$ должны все принадлежать отрезку $[a, b]$.

Как видно из доказательства теоремы 2 § 4, наличие обоих фактов можно гарантировать в случае, когда весовая функция (в нашей задаче это есть $p(x)\Omega(x)$) сохраняет знак на отрезке $[a, b]$. Последнее же может произойти в том и только в том случае, когда точки перемены знака $p(x)$, лежащие внутри $[a, b]$, совпадают с такими же точками $\Omega(x)$, т. е. с фиксированными узлами a_i $(i = 1, \ldots, m)$.

Наибольший интерес в приложениях имеет случай, когда $p(x) \geqslant 0$ и, следовательно, внутри $[a, b]$ весовая

функция перемен знака не имеет. Поэтому особое значение имеют случаи, когда фиксированными узлами являются концы отрезка, оба или один из них.

Предположим теперь, что многочлен $\omega(x)$ существует, и формула (1), имеющая степень точности $2n+m-1$, может быть построена. Рассмотрим погрешность этой формулы и получим для нее представление, рассчитанное на функции высокого порядка гладкости.

Построим интерполирование функции f при помощи многочлена $H(x)$ степени $2n+m-1$ по следующим условиям:

$$H(x_k)=f(x_k),\quad H'(x_k)=f'(x_k),\qquad k=1,\ \ldots,\ n,$$
$$H(a_i)=f(a_i),\qquad i=1,\ \ldots,\ m.$$

Это есть интерполирование с n двукратными узлами x_k и m простыми узлами a_i.

Если функция f имеет на отрезке $[a,b]$, где располагаются x, x_k, a_i, непрерывную производную порядка $2n+m$, остаточный член интерполирования $r(x)$ имеет следующее представление (см. гл. 1, § 5, п. 1):

$$r(x)=\omega^2(x)\,\Omega(x)\,\frac{f^{(2n+m)}(\xi)}{(2n+m)!}\qquad (a\leqslant\xi\leqslant b).$$

Так как $f(x)=H(x)+r(x)$, то для погрешности $R(f)$, ввиду $R(H)=0$, будет верно равенство $R(f)=R(H)+R(r)=R(r)$ и, следовательно,

$$R(f)=R(r)=\int_a^b p(x)\,r(x)\,dx-\sum_{k=1}^{n}A_k r(x_k)-\sum_{i=1}^{m}B_i r(a_i),$$

или, по причине $r(x_k)=0$ и $r(a_i)=0$,

$$R(f)=\int_a^b p(x)\,r(x)\,dx=$$
$$=\frac{1}{(2n+m)!}\int_a^b p(x)\,\Omega(x)\,\omega^2(x)\,f^{(2n+m)}(\xi)\,dx. \qquad (4)$$

Формула (4) для $R(f)$ позволяет ответить на вопрос о степени алгебраической точности формулы (1) в наиболее важном случае.

Теорема 2. *Если узлы* x_k *и* a_i *формулы* (1) *таковы, что*

$$\int_a^b p\Omega\omega^2\,dx \neq 0, \tag{5}$$

то формула (1) *не может быть точной для многочленов степени* $2n+m$.

Доказательство. Оно очевидным образом вытекает из (5), так как если f есть многочлен степени $2n+m$, то для f производная $f^{(2n+m)}$ есть величина постоянная, и

$$R(f) = \frac{f^{(2n+m)}}{(2n+m)!}\int_a^b p\Omega\omega^2\,dx \neq 0.$$

Поэтому в (1) левая и правая части не могут совпадать.

2. Частные случаи. Будем считать весовую функцию постоянной: $p(x) \equiv 1$, и рассмотрим формулы с одним и двумя узлами, лежащими на концах отрезка интегрирования, на обоих или одном. Отрезок интегрирования предположим приведенным к $[-1, 1]$ и начнем с рассмотрения формулы с одним фиксированным узлом в точке -1:

$$\int_{-1}^{1} f(x)\,dx = Af(-1) + \sum_{k=1}^{n} A_k f(x_k) + R(f). \tag{6}$$

Здесь $\Omega(x) = x+1$. Весовая функция $p(x)\Omega(x)$, участвующая в условии ортогональности (3), равна $1\cdot(x+1) = x+1$; она положительна внутри отрезка $[-1, 1]$, и поэтому многочлен $\omega(x)$ существует и единствен при всяких $n = 1, 2, \ldots$ Он должен быть ортогональным на $[-1, 1]$ с весом $x+1$ ко всем многочленам степени меньше n. Такой многочлен может отличаться от многочлена Якоби $P_n^{(0,1)}(x)$ только численным множителем, и так как коэффициент в $P_n^{(0,1)}$ при старшей степени равен $\frac{\Gamma(2n+2)}{2^n n!\Gamma(n+2)}$, то

$$\omega(x) = \frac{2^n n!\Gamma(n+2)}{\Gamma(2n+2)} P_n^{(0,1)}(x).$$

Ввиду того, что в рассматриваемом случае $m = 1$, наивысшая степень точности формулы (6) равна $2n$. Для ее достижения узлы x_k должны совпадать с корнями многочлена $P_n^{(0,\,1)}(x)$, коэффициенты же вычислены по формулам (2), которые в нашем случае принимают вид

$$A_k = \frac{1}{1+x_k}\int_{-1}^{1}(1+x)\frac{\omega(x)}{(x-x_k)\,\omega'(x_k)}\,dx \qquad (k=1,\ \ldots,\ n),$$

$$A = [P_n^{(0,\,1)}(-1)]^{-1}\int_{-1}^{1}P_n^{(0,\,1)}(x)\,dx.$$

Оба интеграла просто вычисляются при помощи известных свойств многочленов Якоби, и в результате получаются следующие значения коэффициентов:

$$A_k = \frac{4}{(1+x_k)\left(1-x_k^2\right)\left[P_n^{(0,\,1)'}(x_k)\right]^2}, \qquad (7)$$

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

Погрешность $R(f)$ формулы (6) можно найти с помощью ее общего выражения (4):

$$R(f) = \frac{1}{(2n+1)!}\int_{-1}^{1}(1+x)\,\omega^2(x)\,f^{(2n+1)}(\xi)\,dx.$$

Множитель $(1+x)\omega^2(x)$, стоящий под знаком интеграла, сохраняет знак на отрезке интегрирования. Когда $f^{(2n+1)}(x)$ есть непрерывная функция на $[-1, 1]$, то на этом отрезке существует такая точка η, что будет верно следующее представление погрешности:

$$R(f) = \frac{f^{(2n+1)}(\eta)}{(2n+1)!}\int_{-1}^{1}(1+x)\,\omega^2(x)\,dx =$$

$$= \frac{f^{(2n+1)}(\eta)}{(2n+1)!}\left[\frac{2^n n!\,(n+1)!}{(2n+1)!}\right]^2\int_{-1}^{1}(1+x)\left[P_n^{(0,\,1)}(x)\right]^2 dx =$$

$$= \frac{2}{2n+1}\left[\frac{2^n n!\,(n+1)!}{(2n+1)!}\right]^2\frac{f^{(2n+1)}(\eta)}{(2n+1)!} \qquad (-1\leqslant\eta\leqslant 1). \qquad (8)$$

Перейдем теперь к формуле с двумя фиксированными узлами в точках -1 и $+1$

$$\int_{-1}^{1} f(x)\,dx = Af(-1) + \sum_{k=1}^{n} A_k f(x_k) + Bf(1) + R(f). \quad (9)$$

В этом случае $\Omega(x) = x^2 - 1$. Весовая функция $p(x)\Omega(x)$ в условиях ортогональности равна $1\,(x^2-1) = x^2 - 1$ и сохраняет знак на $[-1, +1]$; $\omega(x)$ может отличаться от многочлена Якоби $P_n^{(1,1)}(x)$ постоянным множителем:

$$\omega(x) = \frac{2^n n!\,\Gamma(n+3)}{\Gamma(2n+3)} P_n^{(1,1)}(x).$$

Чтобы достичь наивысшей степени точности, равной $2n+1$ для формулы (9), необходимо в качестве узлов x_k взять корни многочлена Якоби $P_n^{(1,1)}(x)$, и коэффициенты A, B, A_k вычислить по их представлениям (2). Вычисления дадут для них следующие значения:

$$A_k = 8\,\frac{n+1}{n+2}\,\frac{1}{(1-x_k^2)\left[P_n^{(1,1)'}(x_k)\right]^2}, \quad A = B = \frac{2}{(n+1)(n+2)}.$$

Погрешность формулы (9) при этом равна

$$R(f) = \frac{8(n+1)}{(n+2)(2n+3)}\left[\frac{2^n n!\,(n+2)!}{(2n+2)!}\right]^2 \frac{f^{(2n+2)}(\eta)}{(2n+2)!}$$
$$(-1 \leqslant \eta \leqslant 1).$$

§ 7. Квадратурные формулы с равными коэффициентами

1. Построение формулы. Формулы квадратур с одинаковыми коэффициентами

$$\int_a^b p(x) f(x)\,dx \approx C_n \sum_{k=1}^{n} f(x_k) \quad (1)$$

особенно удобны при работе с чертежами, когда ординаты легко снимаются с чертежа и столь же просто суммируются.

Формула (1) содержит $n+1$ параметров C_n, x_k $(k=1,\ldots,n)$, и их естественно выбрать так, чтобы равенство (1) выполнялось точно для всех многочленов степени*) n или, что равносильно, выполнялось точно для степеней x от нулевой до n:

$$\int_a^b p(x)\,x^i\,dx = C_n \sum_{k=1}^{n} x_k^i, \qquad i=0,\ 1,\ \ldots,\ n. \tag{2}$$

К принятому ранее предположению об абсолютной интегрируемости на $[a,b]$ произведений $p(x)x^m$ $(m=0,1,\ldots)$ мы сделаем еще одно предположение о весе $p(x)$, естественное в рассматриваемой задаче. Будем считать, что выполняется условие

$$I_0 = \int_a^b p(x)\,dx \neq 0. \tag{3}$$

Если окажется, что $I_0=0$, и если, кроме того, предположить, что равенство (1) выполняется точно в случае, когда $f(x)\equiv 1$, т. е. что верно равенство

$$\int_a^b p(x)\,dx = nC_n, \tag{4}$$

то должно быть $C_n=0$, и формула (1) тогда теряет всякое значение.

Выясним возможность решения системы (2) относительно C_n и x_i. При $i=0$ получится равенство (4), и из него найдем

$$C_n = \frac{1}{n}\int_a^b p(x)\,dx. \tag{5}$$

*) Формулы квадратур вида (1), обладающие этим свойством, называют *формулами Чебышева*.

Полагая последовательно $i = 1, 2, \ldots, n$, получим систему уравнений для нахождения x_i

$$\left.\begin{aligned} s_1 &= x_1 + x_2 + \ldots + x_n = C_n^{-1} \int_a^b p x\, dx = C_n^{-1}\mu_1, \\ s_2 &= x_1^2 + x_2^2 + \ldots + x_n^2 = C_n^{-1} \int_a^b p x^2\, dx = C_n^{-1}\mu_2, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ s_n &= x_1^n + x_2^n + \ldots + x_n^n = C_n^{-1} \int_a^b p x^n\, dx = C_n^{-1}\mu_n. \end{aligned}\right\} \quad (6)$$

Рассмотрим многочлен $\omega(x)$, для которого x_i являются корнями:

$$\omega(x) = (x - x_1) \ldots (x - x_n) = \\ = x^n + a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_n. \quad (7)$$

Уравнения (6) дают численные значения сумм степеней корней от s_1 до s_n. В алгебре многочленов известны соотношения между коэффициентами многочлена $a_1, \ldots, a_n$ и суммами степеней корней

$$\left.\begin{aligned} s_1 + a_1 &= 0, \\ s_2 + a_1 s_1 + 2a_2 &= 0, \\ s_3 + a_1 s_2 + a_2 s_1 + 3a_3 &= 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ s_n + a_1 s_{n-1} + a_2 s_{n-2} + \ldots + n a_n &= 0. \end{aligned}\right\} \quad (8)$$

Значения s_i $(i = 1, \ldots, n)$ известны, с их помощью мы можем, и при этом единственным образом, найти коэффициенты a_i многочлена. После этого, решая уравнение $\omega(x) = 0$, найдем узлы x_i квадратурной формулы (1). Но необходимо заметить, что корни многочлена $\omega(x)$ могут оказаться комплексными или выходить за границы $[a, b]$.

Все изложенное позволяет формулировать приводимую ниже теорему.

Теорема 1. *Если* $\int_a^b p(x)\,dx \neq 0$, *квадратурная формула вида* (1) *с действительными или комплексными узлами* x_k, *точная для любых алгебраических многочленов степени* n, *может быть построена, и при этом единственным образом, при всяких* $n = 1, 2, \ldots$

Отметим, что если среди узлов x_i есть комплексные, то квадратурная формула (1) может быть полезной лишь для интегрирования функций f, аналитических в области, содержащей внутри себя отрезок $[a, b]$ и все узлы x_i. Поэтому для формул Чебышева особое значение имеет случай, когда все узлы в этой формуле являются действительными и лежат на $[a, b]$.

2. Случай постоянной весовой функции. Положим $p(x) \equiv 1$ и отрезок интегрирования приведенным к $[-1, 1]$. Тогда

$$\int_{-1}^{1} f(x)\,dx \approx C_n \sum_{k=1}^{n} f(x_k). \tag{9}$$

Коэффициент C_n определится при помощи равенства (8):

$$C_n = \frac{1}{n}\int_{-1}^{1} dx = \frac{2}{n}.$$

Так как $\int_{-1}^{1} x^i\,dx = [1-(-1)^{i+1}]/(i+1)$, уравнения (6) дадут для сумм степеней узлов x_i значения

$$s_1 = \frac{n}{2}\int_{-1}^{1} x\,dx = 0, \quad s_2 = \frac{n}{2}\int_{-1}^{1} x^2\,dx = \frac{n}{3}, \ldots,$$

$$s_n = \frac{n}{2}\,\frac{1}{n+1}\,[1-(-1)^{n+1}].$$

Поэтому система уравнений для коэффициентов a_k многочлена $\omega(x)$ имеет вид

$$\begin{aligned} a_1 &= 0, \\ \frac{n}{3} + 2a_2 &= 0, \\ a_3 &= 0, \\ \frac{n}{5} + \frac{n}{3}a_2 + 4a_4 &= 0, \\ a_5 &= 0, \\ \frac{n}{7} + \frac{n}{5}a_2 + \frac{n}{3}a_4 + 6a_6 &= 0 \\ &\cdots\cdots\cdots\cdots \end{aligned}$$

При $n = 1$ $\omega(x) = x$, $x_1 = 0$, $C_1 = 2$, поэтому

$$\int_{-1}^{1} f(x)\,dx \approx 2f(0).$$

Это есть формула прямоугольника с высотой, равной ординате в середине отрезка интегрирования.

Для $n = 2$ $\omega(x) = x^2 + \frac{1}{3}$, $x_1 = -\frac{1}{\sqrt{3}}$, $x_2 = \frac{1}{\sqrt{3}}$, $C_3 = 1$, откуда

$$\int_{-1}^{1} f(x)\,dx \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

Эта формула совпадает с формулой Гаусса для двух узлов.

Приведем еще таблицу узлов формулы Чебышева для значений n от $n = 1$ до $n = 9$ с пятью значащими цифрами:

$n = 1$,
$x_1 = 0$;

$n = 2$,
$x_2 = -x_1 = -0{,}57735$;

$n = 3$,
$x_3 = -x_1 = 0{,}70711$,
$x_2 = 0$;

$n = 4$,
$x_4 = -x_1 = 0{,}79465$,
$x_3 = -x_2 = 0{,}18759$;

$n = 5$,
$x_5 = -x_1 = 0{,}83250$,
$x_4 = -x_2 = 0{,}37454$,
$x_3 = 0$;

$$n = 6,$$
$$x_6 = -x_1 = 0{,}86625,$$
$$x_5 = -x_2 = 0{,}42252,$$
$$x_4 = -x_3 = 0{,}26664;$$

$$n = 7,$$
$$x_7 = -x_1 = 0{,}88386,$$
$$x_6 = -x_2 = 0{,}52966,$$
$$x_5 = -x_3 = 0{,}32391,$$
$$x_4 = 0;$$

$$n = 9,$$
$$x_9 = -x_1 = 0{,}91159,$$
$$x_8 = -x_2 = 0{,}60102,$$
$$x_7 = -x_3 = 0{,}52876,$$
$$x_6 = -x_4 = 0{,}16791,$$
$$x_5 = 0.$$

При $n = 8$ среди узлов x_k существуют два комплексных, и при всяких $n \geqslant 10$ среди x_k также существуют комплексные.

§ 8. Задача увеличения точности квадратурных формул; формула Эйлера

1. О содержании задачи. Пусть рассматривается какой-либо интеграл, и необходимо вычислить его значение с предписанной точностью. Обычно бывает, что многие из формул приближенных квадратур, о которых говорилось выше, позволяют вычислить интеграл со сколь угодно высокой точностью, если взять в них достаточно большое число n узлов. Но определить, какое именно n следует взять для получения нужной точности, часто бывает очень трудно. Значение n указывают, как правило, используя имеющийся вычислительный опыт, сравнивая заданный интеграл с другими, более простыми и уже вычисленными, прикидывая заранее приблизительную величину погрешности и т. д. При таком способе выбора n обычно не бывает полной убежденности в том, что полученный результат имеет нужную точность. При этом возникает необходимость проверки результата и увеличения его точности, если она оказывается недостаточной.

Чтобы увеличить точность выбранной формулы, нужно выяснить какое дополнительное слагаемое следует прибавить к квадратурной сумме, чтобы полученная после этого новая квадратурная формула была более точной, чем взятая первоначально.

Добавляемый новый член формулы должен быть главной частью погрешности $R(f)$ выбранной формулы и, кроме того, удовлетворять некоторым дополнительным условиям, среди которых обязательно должно присутствовать требование его возможной простоты и эффективной вычислимости.

Пусть новый член формулы найден. Может оказаться, что добавление его к квадратурной сумме не исправит результат до принятой точности, и одного шага улучшения правила окажется недостаточно. Тогда мы вынуждены будем найти погрешность новой квадратурной формулы, выделить из нее в свою очередь главную часть и т. д.

Число шагов уточнения может быть своим в каждой задаче, и в общем случае ставится проблема о разложении погрешности $R(f)$ на сумму любого числа «главных частей» возрастающих порядков малости.

Такое разложение можно сделать для каждой квадратурной формулы, но вид последовательных главных частей будет своим для всякой формулы.

Ниже идея построения таких разложений будет пояснена на примере самой простой квадратурной формулы — формулы трапеций *). Это приведет нас к формуле Эйлера — Маклорена, давно известной и часто применяемой.

2. Формула Эйлера — Маклорена. Возьмем элементарную формулу трапеций

$$\int_a^b f(x)\,dx = \frac{b-a}{2}[f(a)+f(b)] + R(f) \qquad (1)$$

и для получения ее погрешности $R(f)$ в нужном нам виде, удобном для выделения из $R(f)$ простейшей главной части, воспользуемся формулой Тейлора для f,

*) Для ознакомления с другими формулами, позволяющими увеличивать точность квадратурных формул, можно воспользоваться книгой [3].

ограничившись в ней только линейными членами и остаточным членом:

$$f(x) = f(a) + (x-a) f'(a) + \int_a^x f''(t)(x-t)\,dt =$$

$$= L(x) + \int_a^b f''(t) E(x-t)(x-t)\,dt = L(x) + r(x).$$

Линейная часть по формуле трапеций (1) будет проинтегрирована точно, и погрешность $R(f)$ интегрирования f поэтому совпадает с погрешностью интегрирования остаточного члена $r(x)$:

$$R(r) = \int_a^b r(x)\,dx - \frac{b-a}{2}[r(a) + r(b)],$$

$$\int_a^b r(x)\,dx = \int_a^b dx \int_a^b f''(t) E(x-t)(x-t)\,dt =$$

$$= \int_a^b dt f''(t) \int_a^b E(x-t)(x-t)\,dx =$$

$$= \int_a^b dt f''(t) \int_t^b (x-t)\,dx = \int_a^b f''(t) \frac{(b-t)^2}{2}\,dt,$$

$$r(a) = 0, \quad r(b) = \int_a^b f''(t)(b-t)\,dt.$$

Поэтому

$$R(f) = \int_a^b f''(t)\left[\frac{(b-t)^2}{2} - \frac{b-a}{2}(b-t)\right]dt =$$

$$= -\frac{1}{2}\int_a^b f''(t)(b-t)(t-a)\,dt,$$

или, если привести $[a, b]$ к каноническому отрезку $[0, 1]$, положив $t = a + (b - a)u \quad (0 \leqslant u \leqslant 1)$, то

$$R(f) = -\frac{(b-a)^3}{2}\int_0^1 f''[a + (b-a)u]\,u(1-u)\,du. \quad (2)$$

Рассмотрим отдельно интеграл и постараемся выделить из него последовательность главных частей возрастающих порядков малости. Для упрощения записи введем обозначения

$$f''[a + u(b-a)] = \varphi(u) \quad \text{и} \quad u(1-u) = K_0(u).$$

Ядро интеграла $K_0(u)$ есть плавно и мало изменяющаяся функция на $[0, 1]$, близкая к нулю только в малых окрестностях точек $u = 0$ и $u = 1$. Рассмотрим среднее значение $K_0(u)$ на отрезке интегрирования

$$C_0 = \int_0^1 K_0(u)\,du = \int_0^1 u(1-u)\,du = \frac{1}{6}.$$

Все значения $K_0(u) \quad (0 \leqslant u \leqslant 1)$ будут как-то колебаться около C_0. Поэтому для выделения главной части интеграла ядро K_0 естественно разложить на две следующих части:

$$K_0(u) = C_0 + [K_0(u) - C_0],$$

$$\int_0^1 \varphi(u) K_0(u)\,du = C_0 \int_0^1 \varphi(u)\,du - \int_0^1 \varphi(u)[C_0 - K_0(u)]\,du.$$

Первый из интегралов, стоящих справа, вычисляется просто:

$$\int_0^1 \varphi(u)\,du = \int_0^1 f''[a + (b-a)u]\,du = \frac{1}{b-a}[f'(b) - f'(a)].$$

Во втором интеграле выполним интегрирование по частям, введя функцию

$$K_1(u) = \int_0^u [C_0 - K_0(t)]\,dt = \frac{1}{3}\left[u^3 - \frac{3}{2}u^2 + \frac{1}{2}u\right]$$

и приняв во внимание, что $K_1(0) = K_1(1) = 0$, $\varphi'(u) = (b-a) f'''[a+(b-a)u]$. Тогда

$$\int_0^1 \varphi(u)[C_0 - K_0(u)]\,du = \Big|_0^1 \varphi(u) K_1(u) - \int_0^1 \varphi'(u) K_1(u)\,du =$$

$$= -(b-a)\int_0^1 f'''[a+(b-a)u] K_1(u)\,du.$$

К вновь полученному интегралу, содержащему f''', могут быть применены те же преобразования, с заменой f на $(b-a)f'''[a+(b-a)u] = \varphi'(u)$ и K_0 на K_1:

$$C_1 = \int_0^1 K_1(u)\,du = \int_0^1 \frac{1}{3}\left[t^3 - \frac{3}{2}t^2 + \frac{1}{2}t\right]dt = 0,$$

$$K_2(u) = \int_0^u \left[-\frac{1}{3}\left(t^3 - \frac{3}{2}t^2 + \frac{1}{2}t\right)\right]dt = -\frac{1}{12}(u^4 - 2u^3 + u^2),$$

$$(b-a)\int_0^1 f'''[a+(b-a)u][C_1 - K_1(u)]\,du =$$

$$= -(b-a)^2 \int_0^1 f^{IV}[a+(b-a)u] K_2(u)\,du.$$

Подстановка полученных результатов в (2) приведет к следующему правилу выделения из $R(f)$ главной части:

$$R(f) = -\frac{(b-a)^2}{12}[f'(b) - f'(a)] -$$

$$-\frac{(b-a)^5}{2}\int_0^1 f^{IV}[a+(b-a)u] K_2(u)\,du.$$

К интегралу, стоящему справа, в свою очередь можно применить сходные рассуждения и выделить из него главную часть, получающуюся при замене ядра $K_2(u)$ его средним значением, и т. д. После $\nu+1$ шагов таких выделений будет получено нужное для наших

целей разложение погрешности $R(f)$ простейшей формулы трапеции:

$$R(f) = -\frac{(b-a)^2}{12}[f'(b) - f'(a)] +$$
$$+\frac{(b-a)^4}{720}[f'''(b) - f'''(a)] - \ldots + \rho_{2\nu+2}(f) =$$
$$= -\sum_{k=1}^{\nu} \frac{(b-a)^{2k}}{(2k)!} B_{2k}[f^{(2k-1)}(b) - f^{(2k-1)}(a)] + \rho_{2\nu+2}(f), \quad (3)$$
$$\rho_{2\nu+2}(f) =$$
$$= \frac{(b-a)^{2\nu+3}}{(2\nu+2)!} \int_0^1 f^{(2\nu+2)}[a + (b-a)u][B_{2\nu+2}(u) - B_{2\nu+2}]\,du.$$

В этой записи под B_n понимаются числа Бернулли, которые могут быть определены равенством *)

$$\frac{t}{e^t - 1} = \sum_{n=0}^{\infty} \frac{B_n}{n!} t^n, \quad |t| < 2\pi,$$

и $B_n(x)$ — многочлены Бернулли, определяемые разложением по степеням t следующей производящей функции:

$$e^{xt}\frac{t}{e^t - 1} = \sum_{n=0}^{\infty} \frac{B_n(x)}{n!} t^n, \quad |t| < 2\pi.$$

О полученном разложении (3) необходимо сказать, что слагаемые правой части будут расположены по возрастанию порядков их малости, если длина $b - a$ отрезка интегрирования будет малой величиной. Этого оказывается достаточно для нашей цели.

Перейдем теперь к общей формуле трапеций. Разделим отрезок интегрирования $[a, b]$ на n равных частей длины $h = \frac{1}{n}(b-a)$ точками $a + kh$ $(k = 1, \ldots, n-1)$. К интегралу по частичному отрезку $[a + ph, a + (p+1)h]$

*) B_n и $B_n(x)$ можно найти, например, в книге [2].

применим формулу трапеций (1) с остаточным членом $R(f)$, взятым в форме (3):

$$\int_{a+ph}^{a+(p+1)h} f(x)\,dx =$$

$$= \frac{h}{2}[f_p + f_{p+1}] - \sum_{k=1}^{\nu} \frac{h^{2k}}{(2k)!} B_{2k}[f_{p+1}^{(2k-1)} - f_p^{(2k-1)}] + \rho_{2\nu+2}^{p}(f),$$

$$f_p = f(a+ph), \qquad f_p^{(2k-1)} = f^{(2k-1)}(a+ph),$$

$$\rho_{2\nu+2}^{p}(f) = \frac{h^{2\nu+1}}{(2\nu)!} \int_0^1 f^{(2\nu+2)}[a+h(p+u)][B_{2\nu}(u) - B_{2\nu}]\,du.$$

Если сложить такие равенства по всем частичным отрезкам ($p = 0, 1, \ldots, n-1$), то при этом слагаемые в сумме $\sum$, отвечающие точкам деления, лежащим внутри $[a, b]$, сократятся, и останутся только члены суммы, соответствующие концам a и b отрезка.

После сложения получится формула Эйлера — Маклорена

$$\int_a^b f(x)\,dx = T_n - \sum_{k=1}^{\nu} \frac{h^{2k}}{(2k)!} B_{2k}[f^{(2k-1)}(b) - f^{(2k-1)}(a)] + \rho_{2\nu+2}(f) =$$

$$= T_n - \frac{h^2}{12}[f'(b) - f'(a)] + \frac{h^4}{720}[f'''(b) - f'''(a)] -$$

$$- \frac{h^6}{30\,240}[f^{(5)}(b) - f^{(5)}(a)] + \frac{h^8}{1\,209\,600}[f^{(7)}(b) - f^{(7)}(a)] -$$

$$- \frac{h^{10}}{47\,900\,160}[f^{(9)}(b) - f^{(9)}(a)] + \ldots + \rho_{2\nu+2}(f), \qquad (4)$$

где

$$T_n = h\left[\frac{1}{2}f(a) + f(a+h) + f(a+2h) + \ldots + \frac{1}{2}f(b)\right],$$

$$\rho_{2\nu+2}(f) = \frac{h^{2\nu+3}}{(2\nu+2)!} \int_0^1 [B_{2\nu}(u) - B_{2\nu}] \sum_{p=0}^{n-1} f^{(2\nu+2)}(a+ph+uh)\,du.$$

Когда n неограниченно возрастает и, следовательно, h стремится к нулю, для каждого фиксированного ν

члены в правой части (4) будут расположены по возрастающим порядкам их малости. При этом остаточный член формулы $\rho_{2v+2}(f)$, имеющий смысл погрешности в формуле (4), является малой величиной порядка не ниже h^{2v+2}.

3. Разностные видоизменения формулы Эйлера — Маклорена. Формула (4) для применения требует вычисления значений производных от функции f на концах отрезка, что не всегда удобно. Известны видоизменения этой формулы, позволяющие уточнять формулу трапеций только при помощи значений функции f, и не требующие вычисления производных от f. Все такие видоизменения получаются при помощи замены в (4) производных f', f''', $f^{(5)}$, ... приближенными выражениями их через значения f в точках $a+kh$ и $b+kh$ ($k=\pm 1, \pm 2, \ldots$). Такая замена может быть сделана многими способами в зависимости от выбора узлов и способа интерполирования производных. Наиболее часто используется видоизменение, принадлежащее Грегори, когда используются узлы, не выходящие за границу отрезка $[a, b]$. Для вычисления значений производных в точке a интерполируем f по ее значениям в точках $a+kh$ ($k=0, 1, \ldots$):

$$f(x)=f(a+th)=f_0+\frac{t}{1!}\Delta f_0+\frac{t(t-1)}{2!}\Delta^2 f_0+\ldots+r(x),$$

$$f_k=f(a+kh).$$

Если от обеих частей равенства вычислить производные по t и положить затем $t=0$, $x=a$, получим *)

$$hf'(a)=\Delta f_0-\frac{1}{2}\Delta^2 f_0+\frac{1}{3}\Delta^3 f_0-\frac{1}{4}\Delta^4 f_0+\frac{1}{5}\Delta^5 f_0-\ldots+r'(a),$$

$$h^3 f'''(a)=\Delta^3 f_0-\frac{3}{2}\Delta^4 f_0+\frac{7}{4}\Delta^5 f_0-\ldots+r'''(a),$$

$$h^5 f^{(5)}(a)=\Delta^5 f_0-\ldots+r^{(5)}(a),$$

. . . . . .

*) См. гл. 1, § 6, п. 2.

Аналогично для нахождения значений производных в точке $x = b$, интерполируем f по значениям в точках $b + kh = a + nh + kh$ $(k = 0, -1, \ldots)$:

$$f(x) = f(b + th) =$$

$$= f_n + \frac{t}{1!}\Delta f_{n-1} + \frac{t(t+1)}{2!}\Delta^2 f_{n-2} + \ldots + \rho(x).$$

После вычисления производных по t при $t = 0$, $x = b$ отсюда найдем

$$hf'(b) = \Delta f_{n-1} + \frac{1}{2}\Delta^2 f_{n-2} + \frac{1}{3}\Delta^3 f_{n-3} + \frac{1}{4}\Delta^4 f_{n-4} +$$

$$+ \frac{1}{5}\Delta^5 f_{n-5} + \ldots + \rho'(b),$$

$$h^3 f'''(b) = \Delta^3 f_{n-3} + \frac{3}{2}\Delta^4 f_{n-4} + \frac{7}{4}\Delta^5 f_{n-5} + \ldots + \rho^{(3)}(b),$$

$$h^5 f^{(5)}(b) = \Delta^5 f_{n-5} + \ldots + \rho^{(5)}(b),$$

. . . . . .

Подстановка полученных отсюда значений производных в (4) приведет к формуле Грегори

$$\int_a^{a+nh} f(x)\,dx = T_n - \frac{h}{12}(\Delta f_{n-1} - \Delta f_0) - \frac{h}{24}(\Delta^2 f_{n-2} + \Delta^2 f_0) -$$

$$- \frac{19h}{720}(\Delta^3 f_{n-3} - \Delta^3 f_0) - \frac{3h}{160}(\Delta^4 f_{n-4} + \Delta^4 f_0) -$$

$$- \frac{863h}{60\,480}(\Delta^5 f_{n-5} - \Delta^5 f_0) - \frac{275h}{24\,192}(\Delta^6 f_{n-6} + \Delta^6 f_0) - \ldots$$

$$\ldots - C_k h[\Delta^k f_{n-k} + (-1)^k \Delta^k f_0] + R_1(f), \qquad (5)$$

$$C_k = \frac{(-1)^k}{(k+1)!}\int_0^1 x(x-1)\ldots(x-k)\,dx.$$

§ 9. Некоторые теоремы о сходимости квадратурных процессов

Напомним, что сходимость или расходимость квадратурного процесса зависит от следующих факторов: 1) от последовательности квадратурных формул или,

что равносильно, от бесконечных треугольных матриц X, A узлов x_k^n и коэффициентов A_k^n этих формул (1.4) и (1.5); 2) от класса F интегрируемых функций f.

В проблеме сходимости необходимо выяснить, как должны быть между собой связаны X, A и F, чтобы погрешность

$$R_n(f) = \int_a^b p(x) f(x)\, dx - \sum_{k=1}^{n} A_k^n f(x_k^n) \qquad (1)$$

стремилась к нулю при $n \to \infty$ для всякой функции f из F. Ниже будут приведены некоторые теоремы о сходимости. Для доказательства части их потребовалось бы большое число страниц, и такие теоремы даны без доказательств, только в формулировках с некоторыми пояснениями их.

1. О сходимости общего квадратурного процесса. Будем считать отрезок интегрирования $[a, b]$ конечным, и возьмем множество $C[a, b]$ функций f, непрерывных на $[a, b]$. Рассмотрим квадратурный процесс

$$\int_a^b p(x) f(x)\, dx = \sum_{k=1}^{n} A_k^n f(x_k^n) + R_n(f) = Q(f) + R_n(f). \qquad (2)$$

Теорема 1. *Для того чтобы квадратурный процесс* (2) *сходился для всякой функции* f, *непрерывной на* $[a, b]$, *необходимо и достаточно выполнение условий*:

1) *квадратурный процесс сходится для любого алгебраического многочлена*;

2) *существует число* M *такое, что при всяких* $n = 1, 2, \ldots$ *выполняется неравенство*

$$B_n = \sum_{k=1}^{n} |A_k^n| \leqslant M < \infty. \qquad (3)$$

Эта теорема дает условия сходимости в очень широком, хотя и не исчерпывающем все потребности практики, классе функций.

По поводу теоремы 1 могут быть высказаны два следующих пожелания: 1) указать частные случаи теоремы, условия в которых являются более простыми

и удобными для проверки; 2) выяснить условия сходимости в более узких классах функций, часто встречающихся в приложениях.

Приведем сначала теорему о сходимости квадратурного процесса с неотрицательными коэффициентами A_k^n.

Теорема 2. *Если квадратурная формула* (2) *имеет при всех* $n = 1, 2, \ldots$ *неотрицательные коэффициенты:* $A_k^n \geqslant 0$ $(k = 1, \ldots, n;\ n = 1, 2, \ldots)$, *то для сходимости процесса* (2) *при любой непрерывной функции* f *необходима и достаточна сходимость процесса для всякого алгебраического многочлена.*

Доказательство. Высказанная теорема является весьма простым следствием теоремы 1. Действительно, необходимость условия теоремы очевидна, так как, если процесс сходится для любой непрерывной функции, то он должен сходиться для всякого многочлена ввиду его непрерывности. Достаточность проверяется почти так же просто. Если процесс сходится для всякого многочлена, то он сходится и в случае $f \equiv 1$, и, следовательно,

$$Q_n(1) = \sum_{k=1}^{n} A_k^n \to \int_a^b p(x)\,dx = \mu_0 \qquad (n \to \infty).$$

С другой стороны, ввиду $A_k^n \geqslant 0$ будет

$$B_n = \sum_{k=1}^{n} |A_k^n| = \sum_{k=1}^{n} A_k^n \to \mu_0 \qquad (n \to \infty).$$

Поэтому числа B_n ограничены в совокупности: $B_n \leqslant M < \infty$. В условиях теоремы 2 выполняются оба условия теоремы 1, и квадратурный процесс (2) будет сходиться для всех непрерывных на $[a, b]$ функций f.

Среди классов функций, более узких чем $C[a, b]$, в первую очередь имеют интерес функции, обладающие непрерывными производными того или иного порядка.

Возьмем класс $C_r[a, b]$ функций f, имеющих на $[a, b]$ непрерывные производные порядка r.

Чтобы сформулировать теорему о сходимости процесса для такого класса, нам потребуется ввести некоторые вспомогательные функции, характеризующие рас-

пределение узлов x_k^n, значения коэффициентов A_k^n и порядок r гладкости функций f.

Предположим, что узлы перенумерованы в порядке роста:

$$a \leqslant x_1^n < x_2^n < \ldots < x_n^n \leqslant b.$$

Построим кусочно-постоянную функцию, связанную с распределением узлов и величинами коэффициентов A_k^n:

$$F_{n0}(t) = \sum_{k=1}^{n} A_k^n E\left(t - x_k^n\right), \qquad a \leqslant t \leqslant b.$$

Значения, которые принимает F_{n0} на $[a, b]$, даются в следующей таблице*):

$$F_{n0}(t) = \begin{cases} 0 & \text{при} \quad a \leqslant t < x_1^n, \\ A_1^n & \text{при} \quad x_1^n < t < x_2^n, \\ A_1^n + A_2^n & \text{при} \quad x_2^n < t < x_3^n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ A_1^n + \ldots + A_n^n & \text{при} \quad x_n^n < t \leqslant b. \end{cases} \tag{4}$$

Одновременно с функцией F_{n0} будем рассматривать неопределенный интеграл для нее $F_{n,r}(t)$ любого порядка r с нулевыми начальными значениями в точке a: $F_{n,r}^{(j)}(a) = 0$ $(j = 0, 1, \ldots, r-1)$

$$F_{n,r}(t) = \sum_{k=1}^{n} A_k^n E(t - x_k) \frac{1}{r!} (t - x_k)^r.$$

Функция $F_{n,r}(t)$ является, очевидно, непрерывной с производными до порядка $r-1$ включительно на $[a, b]$, производная же от нее порядка r, совпадающая с F_{n0}, есть кусочно-постоянная функция с разрывами в узлах x_k^n $(k = 1, \ldots, n)$ и имеет в них скачки соответственно A_k^n $(k = 1, \ldots, n)$. Она играет роль функции влияния (функции Грина) в рассматриваемой задаче сходимости.

*) Предполагается, что узлы x_k^n лежат внутри $[a, b]$, но можно без труда указать, как изменится эта таблица, если один или два узла будут лежать в точках a и b.

Теорема 3. *Для сходимости квадратурного процесса* (2) *для всякой функции* f, *имеющей на* $[a, b]$ *непрерывную производную порядка* r, *необходимо и достаточно выполнение условий*:

1) *процесс* (2) *сходится всякий раз, когда* f *есть алгебраический многочлен*;

2) *существует число* $M < \infty$ *такое, что при значениях* $n = 1, 2, \ldots$ *выполняется неравенство*

$$\int_a^b |F_{n, r-1}(t)|\, dt \leqslant M. \tag{5}$$

Заметим, что второму условию может быть придана другая, по-видимому, более наглядная форма. Если функция $\varphi(x)$ имеет на $[a, b]$ непрерывную первую производную $\varphi'(x)$, то интеграл по $[a, b]$ от $|\varphi'(t)|$ есть не что иное, как полная вариация функции $\varphi(x)$ на $[a, b]$:

$$\operatorname*{var}_{[a, b]} \varphi(x) = \int_a^b |\varphi'(x)|\, dx.$$

Если принять во внимание соотношение $F'_{n, r}(t) = F_{n, r-1}(t)$, то можно сказать, что условие (5) означает ограниченность в совокупности полных вариаций на $[a, b]$ функций $F_{n, r}(t)$ $(n = 1, 2, \ldots)$:

$$\operatorname*{var}_{[a, b]} F_{n, r}(t) \leqslant M < \infty \qquad (n = 1, 2, \ldots). \tag{6}$$

Сделаем замечание об условиях теоремы 3. Пусть мы переходим от r к $r+1$ и, в соответствии с этим, от класса функций $C_r[a, b]$ к классу $C_{r+1}[a, b]$, который, очевидно, является частью класса $C_r[a, b]$. Произойдет уменьшение множества функций, для которого должен сходиться квадратурный процесс, и условия сходимости должны замениться на менее ограничительные.

Первое из условий теоремы 3 не зависит от r. Второе условие возьмем в форме (6). Для его проверки мы должны рассмотреть функции $F_{n, t}(t)$, взять их полные вариации на $[a, b]$ и определить, будут ли они ограничены в совокупности. При переходе к $C_{r+1}[a, b]$ функция $F_{n, r}(t)$ должна быть заменена на $F_{n, r+1}(t)$,

которая является первообразной для $F_{n,r}$ с нулевым начальным значением в точке $t = a$. Поэтому

$$F_{n,r+1}(t) = \int_a^t F_{n,r}(u)\,du \quad \text{и} \quad \operatorname*{var}_{[a,b]} F_{n,r+1} = \int_a^b |F_{n,r}(u)|\,du.$$

Необходимым и достаточным условием сходимости квадратурного процесса для всех функций из $C_{r+1}[a,b]$ является существование такого числа $M_1 < \infty$, чтобы при $n = 1, 2, \ldots$ выполнялось неравенство

$$\operatorname*{var}_{[a,b]} F_{n,r+1} = \int_a^b |F_{n,r}(u)|\,du \leqslant M_1 < \infty. \tag{7}$$

Таким образом, переход от множества C_r к C_{r+1} для второго условия теоремы равносилен замене требования (6) более слабым *) требованием (7).

Остановим еще внимание на случае $r = 1$ и множестве $C_1[a,b]$ непрерывно дифференцируемых функций.

Теорема 4. *Чтобы квадратурный процесс* (2) *сходился для всякой функции* f, *непрерывно дифференцируемой на* $[a,b]$, *необходимо и достаточно, чтобы выполнялись условия*:

1) *квадратурный процесс сходится для всякого многочлена*;

2) *существует число* $M < \infty$ *такое, что при* $n = 1, 2, \ldots$ *выполняется неравенство*

$$\begin{aligned} &|A_1^n|(x_2^n - x_1^n) + |A_1^n + A_2^n|(x_3^n - x_2^n) + \ldots \\ &\quad \ldots + |A_1^n + \ldots + A_{n-1}^n|(x_n^n - x_{n-1}^n) + \\ &\quad + |A_1^n + \ldots + A_n^n|(b - x_n^n) \leqslant M < \infty. \end{aligned} \tag{8}$$

---

*) Легко проверить, что из (6) следует (7). Действительно,

$$|F_{n,r}(t)| = \left|\int_a^t F_{n,r-1}(u)\,du\right| \leqslant \int_a^b |F_{n,r-1}(u)|\,du = \operatorname*{var}_{[a,b]} F_{n,r} \leqslant M,$$

поэтому $\operatorname*{var}_{[a,b]} F_{n,r+1} \leqslant \int_a^b |F_{n,r}(t)|\,dt \leqslant \int_a^b M\,dt = M(b-a) = M_1 < \infty.$

Сравним рассматриваемую теорему с теоремой 3 для случая $r = 1$. Первые условия в них одинаковы, и остается проверить, что неравенство (5) при $r = 1$ совпадает с (8). Но это является очевидным, так как таблица (4) значений $F_{n0}(t)$ позволяет вычислить интеграл, входящий в (5) ($r = 1$), и дает для него следующее значение:

$$\int_a^b |F_{n0}(t)|\,dt = |A_1^n|(x_2^n - x_1^n) + |A_1^n + A_2^n|(x_3^n - x_2^n) + \ldots$$
$$\ldots + |A_1^n + \ldots + A_n^n|(b - x_n^n),$$

что доказывает совпадение неравенств (5) и (8).

2. О сходимости интерполяционных квадратур. В таких квадратурных формулах коэффициенты A_k^n вычисляются по узлам x_k^n при помощи равенств

$$A_k^n = \int_a^b p(x)\frac{\omega_n(x)}{(x - x_k^n)\,\omega_n'(x_k^n)}\,dx \qquad (k = 1, \ldots, n),$$
$$\omega_n(x) = (x - x_1^n) \ldots (x - x_n^n),$$

и матрица A коэффициентов (5) определяется матрицей X узлов (4). Поэтому в проблеме сходимости интерполяционных квадратурных процессов необходимо бывает рассматривать два объекта — матрицу X узлов и класс F функции f, и нужно выяснить, как они должны быть связаны между собой, чтобы процесс сходился.

Рассмотрим сначала простейшие теоремы, дающие достаточное условие сходимости и часто применяемые в приложениях. Они основаны на результатах, полученных в теории интерполирования.

Теорема 5. *Пусть отрезок $[a, b]$ конечный и весовая функция $p(x)$ абсолютно интегрируема на $[a, b]$. Если функция f и таблица узлов X таковы, что интерполяционный процесс сходится к f равномерно относительно x на $[a, b]$, то интерполяционный квадратурный процесс*

$$\int_a^b p(x)f(x)\,dx = \sum_{k=1}^n A_k^n f(x_k^n) + R_n(f) \qquad (9)$$

сходится при $n \to \infty$ к точному значению интеграла.

Доказательство. Обозначим $r_n(x)$ погрешность интерполирования f по значениям в узлах x_k^n ($k=1, \ldots, n$). Погрешность приближенной квадратуры связана с $r_n(x)$ равенством

$$R_n(f) = \int_a^b p(x) r_n(x)\, dx.$$

Так как по условию теоремы $r_n(x)$ равномерно на $[a, b]$ стремится к нулю при $n \to \infty$, можно перейти к пределу под знаком интеграла, что приведет к заключению $\lim_{n\to\infty} R_n(f) = 0$. Это доказывает теорему.

Отметим два частных случая теоремы.

1) Сходимость квадратурной формулы Ньютона—Котеса

$$\int_0^1 p(x) f(x)\, dx \approx \sum_{k=0}^{n} B_k^n f\left(\frac{k}{n}\right), \tag{10}$$

$$B_k^n = \frac{(-1)^{n-k}}{nk!\,(n-k)!} \int_0^1 \frac{t(t-1)\ldots(t-n)}{t-k}\, dt.$$

Соответствующая ей таблица узлов есть

$$X = \begin{Bmatrix} 0 & 1 & & & \\ 0 & \frac{1}{2} & 1 & & \\ \cdot & \cdot & \cdot & \cdot & \\ 0 & \frac{1}{n} & \frac{2}{n} & \ldots & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{Bmatrix}.$$

В главе 1 (см. § 7, п. 3) говорилось, что если функция f является аналитической и регулярной в замкнутой области, ограниченной линией

$$\operatorname{Re}[z \ln z + (1-z)\ln(1-z)] = 0, \tag{11}$$

то интерполяционный процесс сходится к f равномерно на $[0, 1]$.

Это позволяет высказать следующую теорему о сходимости квадратурной формулы (10).

Теорема 6. *Пусть f есть аналитическая функция, регулярная в замкнутой области, ограниченной линией с уравнением* (11); *тогда погрешность квадратурной формулы Ньютона — Котеса* (10) *стремится к нулю при неограниченном увеличении n.*

2) Сходимость квадратурного процесса с предельной функцией Чебышева распределения узлов. В гл. 1, § 7, п. 2 говорилось о том, что интерполяционный процесс, для которого предельной функцией распределения узлов является функция Чебышева для отрезка $[a, b]$, сходится к интерполируемой функции f равномерно на $[a, b]$, если f есть аналитическая функция всюду на этом отрезке, включая его концы. Отсюда следует

Теорема 7. *Если матрица* (1.4) *узлов интерполяционного процесса имеет функцию Чебышева своей предельной функцией распределения и если функция f является аналитической на $[a, b]$, тогда интерполяционный квадратурный процесс* (9) *сходится к точному значению интеграла.*

В предыдущем пункте были сформулированы некоторые теоремы о сходимости общего квадратурного процесса. Проследим сейчас, как упрощаются эти теоремы для интерполяционных квадратур. Здесь дело заключается в том, что первым условием в указанных теоремах было требование сходимости квадратурного процесса для всяких многочленов. Для интерполяционных квадратурных процессов это требование может быть опущено, так как оно всегда выполняется. В самом деле, напомним, что интерполяционная квадратурная формула с n узлами характеризуется тем, что она является точной для всякого многочлена степени меньше n. Поэтому, если интегрируемая функция f есть многочлен некоторой степени m, то при всяких $n > m$ в интерполяционном квадратурном процессе всегда будет получаться точное значение вычисляемого интеграла, и процесс будет, очевидно, сходящимся.

Так, например, аналог теоремы 1 для интерполяционных квадратурных процессов будет следующим.

Теорема 8. *Для сходимости интерполяционного квадратурного процесса*

$$\int_a^b p(x) f(x)\, dx \approx \sum_{k=1}^{n} A_k^n f(x_k), \qquad A_k^n = \int_a^b p(x) \frac{\omega_n(x)}{(x - x_k^n)\,\omega'(x_k^n)}\, dx \tag{12}$$

при всякой непрерывной на $[a, b]$ *функции* f *необходимо и достаточно существование числа* M, *для которого при* $n = 1, 2, \ldots$ *выполняется неравенство*

$$B_n = \sum_{k=1}^{n} |A_k^n| \leqslant M < \infty.$$

Аналогом теоремы 2 является

Теорема 9. *Если в интерполяционном квадратурном процессе* (12) *все коэффициенты* A_k^n *неотрицательны, то квадратурный процесс будет сходиться к точному значению интеграла для всякой непрерывной на* $[a, b]$ *функции* f.

Закончим изложение рассмотрением теоремы о сходимости квадратурной формулы наивысшей алгебраической степени точности. В § 4, п. 2 было показано, что если весовая функция $p(x)$ неотрицательна на $[a, b]$, и квадратурное правило имеет наивысшую степень точности $2n-1$, то коэффициенты A_k^n его всегда положительны.

На основании теоремы 9 тогда может быть высказана

Теорема 10. *Если весовая функция* $p(x)$ *неотрицательна на* $[a, b]$, *отрезок* $[a, b]$ *конечный, и квадратурная формула* (12) *при всяких* n *имеет наивысшую степень точности* $2n-1$, *то квадратурный процесс сходится для всякой функции, непрерывной на* $[a, b]$.

Дополнительно заметим, что в условиях теоремы 10 верным является более сильный результат: квадратурный процесс наивысшей степени точности сходится для всякой функции, интегрируемой на $[a, b]$ в смысле Римана.

§ 10. Вычисление неопределенного интеграла

1. Введение. Проблема вычисления неопределенного интеграла

$$y(x) = y_0 + \int_{x_0}^{x} f(t)\,dt \tag{1}$$

будет изучаться в самой простой постановке, когда функция f задана на некотором отрезке $[x_0, X]$ таблицей своих значений в узлах сетки точек на оси аргумента x:

x	y
x_0	y_0
x_1	y_1
...	...
x_n	y_n
x_{n+1}	

$$x_0 < x_1 < x_2 < \ldots < X,$$

и нужно с помощью табличных значений $f(x_k) = f_k$ $(k = 0, 1, \ldots)$ найти значения функции $y(x_k) = y_k$ в узлах той же сетки.

Для простоты ограничимся случаем равноотстоящих точек сетки, когда $x_k = x_0 + kh$, где h есть шаг сетки.

Рассмотрим сначала задачу продолжения уже начатой таблицы. Предположим, что вычисления доведены до узла x_n и составлена таблица, приведенная в тексте. Вычислению подлежит y_{n+1}. Для этого можно воспользоваться любыми найденными раньше значениями y_k $(k \leqslant n)$ и какими угодно известными табличными значениями функции f. Вычислительная формула в достаточно общей форме может быть записана следующим образом:

$$y_{n+1} = \sum_{i=0}^{k} A_i y_{n-i} + h \sum_{j=-m}^{m} B_j f_{n-j}. \tag{2}$$

Множитель h перед второй суммой выделен для того, чтобы коэффициенты B_j, как и A_i, были безразмерными.

Первая из сумм, стоящих справа, может составляться для $n \geqslant k$ и требует знания $k+1$ начальных значений $y_0, y_1, \ldots, y_k$. Вторую сумму можно составлять при $n \geqslant m$.

Поэтому равенство можно применять для $n = l, l+1, \ldots$, где $l = \max(k, m)$, и для применения оно требует составления начала расчетной таблицы, содержащей y_i и f_i $(i = 0, 1, \ldots, l)$.

Формула (2) обычно используется на много шагов — для нахождения многих значений функции $y(x)$. Эта особенность задачи неопределенного интегрирования ставит перед нами новую проблему, которая имеет, как будет выяснено в дальнейшем изложении, широкое значение и важна для многих частей математики — проблему роста погрешности при вычислении на большое число шагов. Оказывается, что для некоторых вычислительных формул рост погрешности может быть настолько быстрым, что после даже малого числа шагов погрешность становится больше принятой для нее границы. Достаточно убедительным здесь будет простой приводимый ниже пример.

Предположим, что на отрезке $[0, 1]$ нужно вычислить в равноотстоящих точках значения интеграла

$$y(x) = \int_0^x e^t\,dt = e^x - 1.$$

Выполним вычисления при помощи двух формул. Сначала интерполируем $y(x)$ по двум значениям y_n и y_{n-1} самой функции и по двум значениям $y'_n = f_n$ и $y'_{n-1} = f_{n-1}$ ее производной и найдем значение интерполирующего многочлена при $x = x_{n+1}$:

$$y_{n+1} = -4y_n + 5y_{n-1} + h(4f_n + 2f_{n-1}). \qquad (3)$$

Формула точна для многочленов третьей степени.

Положим шаг $h = 0,1$ и выполним вычисления, считая известными $y(0) = 0$ и $y(0,1) = 0,10517$. Найденные значения y_n приводятся в таблице.

x_n	y_n	Погрешность	x_n	y_n	Погрешность
0,0	0,00000		0,6	0,81810	+0,00402
0,1	0,10517		0,7	1,03610	−0,02235
0,2	0,22139	+0,00001	0,8	1,11602	+0,10952
0,3	0,34988	−0,00002	0,9	2,01039	−0,55079
0,4	0,49165	+0,00017	1,0	−1,03251	+2,75079
0,5	0,64950	−0,00078			

Как видно из этой таблицы, погрешность на каждом шаге изменяет знак и по абсолютному значению увеличивается приблизительно в пять раз. Причина этого будет объяснена в дальнейшем изложении, но уже сейчас становится очевидным, что избранная вычислительная формула является неудачной, так как при ее применении погрешность вычислений возрастает настолько быстро, что уже через небольшое число шагов ее значение превосходит y_n.

Изменим вычислительную формулу. Если в равенстве

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} e^t\,dt$$

к интегралу применить формулу трапеций и отбросить остаточный член, получим

$$y_{n+1} = y_n + \frac{1}{2}h(f_n + f_{n+1}). \qquad (4)$$

Она является точной, когда f есть многочлен первой степени и $y(x)$ — второй. Степень точности ее на единицу ниже степени точности формулы (3), и можно было бы предполагать, что ее применение даст менее точные результаты, чем (3). В таблице даны результаты вычислений.

x_n	y_n	Погрешность	x_n	y_n	Погрешность
0,0	0,00000		0,6	0,82280	—0,00068
0,1	0,10526	—0,00009	0,7	1,01459	84
0,2	0,22159	19	0,8	1,22656	102
0,3	0,35015	29	0,9	1,46082	122
0,4	0,49223	41	1,0	1,71971	143
0,5	0,64926	54			

Несколько первых значений (до y_4) оказались действительно менее точными, чем в случае применения формулы (3), но все последующие значения имеют более высокую точность, чем в предыдущих вычислениях.

Это объясняется тем, что погрешность для формулы (4) растет значительно медленнее, чем для (3).

Чтобы выяснить, какие именно величины, связанные с формулой вычислений, влияют на рост погрешности, получим уравнение, которому удовлетворяет эта погрешность.

Если в квадратурную формулу (2) вместо y_k подставить точное значение $y(x_k)$ функции y, то равенство будет удовлетворяться с некоторой ошибкой, которую мы обозначим r_n:

$$y(x_{n+1}) = \sum_{i=0}^{k} A_i y(x_{n-i}) + h \sum_{j=-m}^{m} B_j f(x_{n-j}) + r_n. \qquad (5)$$

Величину r_n называют *погрешностью уравнения* (2) для приближенных значений y_n.

Рассмотрим погрешность приближенного решения $y(x_n) - y_n = \varepsilon_n$ и получим для нее уравнение. Вычитая из (5) почленно (2), найдем равенство

$$\varepsilon_{n+1} = \sum_{i=0}^{k} A_i \varepsilon_{n-i} + r_n. \qquad (6)$$

Это есть линейное конечно-разностное уравнение порядка $k+1$ относительно ε_n.

Напомним, что при применении формулы (2) должны предварительно находиться начальные значения $y_0, \ldots, y_l$ $[l = \max(k, m)]$. В соответствии с этим нужно считать известными начальные значения погрешностей $\varepsilon_0, \ldots, \varepsilon_l$. Все следующие значения ε_n $(n > l)$ определяются уравнением (6). Как видно из уравнения, следующее значение погрешности ε_{n+1} определяется $k+1$ предыдущими значениями $\varepsilon_n, \varepsilon_{n-1}, \ldots, \varepsilon_{n-k}$.

Если $l > k$, то, отбросив несколько первых начальных значений $\varepsilon_0, \varepsilon_1, \ldots$ и изменив нумерацию ε, мы можем считать, что в (6) индекс n принимает значения $n = k, k+1, \ldots$ Для нахождения погрешности ε_n получится задача решения линейного конечно-разностного уравнения порядка $k+1$ с начальными значениями $\varepsilon_0, \ldots, \varepsilon_k$.

Как сразу же видно, уравнением (6) и начальными значениями $\varepsilon_0, \ldots, \varepsilon_k$ погрешность ε_n определяется единственным образом. В самом деле, при $n = k$ урав-

нение даст явное выражение ε_{k+1} через начальные значения:

$$\varepsilon_{k+1} = \sum_{i=0}^{k} A_i \varepsilon_{k-i} + r_k.$$

При $n = k+1$ уравнение даст выражение ε_{k+2} через $\varepsilon_1, \ldots, \varepsilon_k$ и найденное ε_{k+1} и т. д.

Свойства ε_n зависят, очевидно, от следующих величин: 1) коэффициентов A_i формулы, 2) значений r_n погрешности формулы, 3) начальных значений $\varepsilon_0, \ldots, \varepsilon_k$ погрешности решения. Величины A_i и r_n определяются самой формулой (2), что же касается начальных значений $\varepsilon_0, \ldots, \varepsilon_k$, то их величинами мы имеем право распоряжаться при подготовке вычислений.

Выясним, каким условиям должны удовлетворять все эти величины, чтобы можно было найти значения неопределенного интеграла (1) с любой, как угодно малой, заранее заданной границей погрешности ε_k.

Ограничимся описанием лишь наглядной стороны вопроса и не будем находить явное выражение для погрешности ε_n и ее оценок, которые потребовали бы знания теории линейных разностных уравнений.

Ввиду линейности уравнения (2) погрешность ε_n может быть разложена на две части: $\varepsilon_n = \varepsilon_n^1 + \varepsilon_n^2$, первая из которых ε_n^1 зависит только от начальных значений $\varepsilon_0, \ldots, \varepsilon_k$ и не зависит от r_n, т. е. отвечает случаю $r_n \equiv 0$. Эта часть должна быть найдена из однородного уравнения

$$\varepsilon_{n+1}^1 = \sum_{i=0}^{k} A_i \varepsilon_{n-i}^1, \quad n = k, \; k+1, \ldots, \tag{7}$$

при заданных начальных значениях $\varepsilon_i^1 = \varepsilon_i$ $(i = 0, 1, \ldots, k)$.

Часть ε_n^2 зависит только от погрешности r_n формулы и имеет нулевые начальные значения. Она является решением следующей задачи:

$$\varepsilon_{n+1}^2 = \sum_{i=0}^{k} A_i \varepsilon_{n-i}^2 + r_n, \quad \varepsilon_i^2 = 0 \quad (i = 0, 1, \ldots, k). \tag{8}$$

Перед рассмотрением каждой из этих частей отметим, что число точек сетки $x_n = x_0 + nh$, лежащих на от-

резке $[x_0, X]$, где разыскивается интеграл, равно*) $N = \text{ц. ч. } (X - x_0)/h = O(h^{-1})$. При $h \to 0$ число N неограниченно возрастает, и нам необходимо изучить поведение ε_n на интервале $(0, \infty)$.

Начнем с ε_n^1. Уравнение (7) не зависит от h. От h зависит только интервал, на котором необходимо находить функцию ε_n.

Для решения уравнения воспользуемся известными в теории конечно-разностных уравнений результатами. Пусть дано линейное однородное уравнение порядка k с постоянными коэффициентами

$$a_0 y_{n+k} + a_1 y_{n+k-1} + \ldots + a_{k-1} y_{n+1} + a_k y_n = 0. \quad (9)$$

Оно имеет k линейно независимых частных решений, и всякое решение уравнения есть их линейная комбинация. Для нахождения этих решений нужно составить характеристическое уравнение

$$a_0 \lambda^k + a_1 \lambda^{k-1} + \ldots + a_{k-1} \lambda + a_k = 0,$$

найти его корни и определить их кратности. Пусть корни уравнения есть $\lambda_1, \ldots, \lambda_m$ кратностей соответственно $r_1, \ldots, r_m (r_1 + \ldots + r_m = k)$. Если λ есть один из этих корней и имеет кратность r, то ему отвечают r частных решений вида

$$\lambda^n, \quad n\lambda^n, \quad \ldots, \quad n^{r-1}\lambda^n.$$

Всем корням $\lambda_1, \ldots, \lambda_m$ отвечают частные решения

$$\begin{array}{l} \lambda_1^n, \quad n\lambda_1^n, \quad \ldots, \quad n^{r_1-1}\lambda_1^n; \\ \ldots\ldots\ldots\ldots \\ \lambda_m^n, \quad n\lambda_m^n, \quad \ldots, \quad n^{r_m-1}\lambda_m^n. \end{array} \quad (10)$$

Число их равно $r_1 + \ldots + r_m = k$ и они, очевидно, линейно независимы. Всякое решение y_n уравнения есть линейная комбинация их.

Для уравнения (7), из которого должна быть найдена погрешность ε_n^1, характеристическое уравнение будет следующим:

$$\lambda^{k+1} = \sum_{i=0}^{k} A_i \lambda^{k-i}. \quad (11)$$

*) Знаком ц. ч. x обозначена целая часть числа x.

Если $\lambda_1, \ldots, \lambda_m$ есть его корни кратностей $r_1, \ldots, r_m$ $(r_1 + \ldots + r_m = k)$, то всякое решение ε_n^1 уравнения (7) представимо в форме комбинации частных решений вида (10)

$$\varepsilon_n^1 = \sum_{i=1}^{m} \lambda_i^n \left(c_0^i + c_1^i n + \ldots + c_{r_i-1}^i n^{r_i-1}\right); \qquad (12)$$

здесь c_j^i есть постоянные, значения которых находятся из начальных условий. Для определения c_j^i получится линейная система, в которой свободными членами являются начальные значения, и c_j^i будут найдены как линейные функции начальных значений $\varepsilon_0, \ldots, \varepsilon_k$.

Прежде чем получить следствия из представления ε_n^1, заметим, что формула вычислений обычно строится так, чтобы она была точной в том случае, когда $y(x)$ есть многочлен некоторой степени, и она заведомо точна, когда $y(x)$ есть постоянная величина $y \equiv 1$ $(y' = f \equiv 0)$. В этом случае формула (2) даст равенство

$$1 = \sum_{i=0}^{k} A_i.$$

Последнее говорит о том, что единица всегда есть корень характеристического уравнения некоторой кратности, и среди чисел λ_i одно обязательно равно единице.

Определим теперь поведение частных решений вида (10) при неограниченном возрастании n. Возьмем какой-либо корень λ_i. Если $|\lambda_i| > 1$, то частные решения, ему отвечающие, будут при $n \to \infty$ возрастать не медленнее, чем показательная функция λ_i^n. В этом случае среди решений ε_n^1 будут существовать быстро возрастающие по модулю.

При $|\lambda_i| < 1$ частные решения, отвечающие такому корню, будут стремиться к нулю при $n \to \infty$.

Наконец, когда $|\lambda_i| = 1$, частные решения вида (10), соответствующие этому корню, будут в случае кратного корня возрастать по модулю как степени n, если же корень λ_i является простым, ему отвечает единственное решение λ_i^n, которое будет ограниченным.

Все изложенные соображения позволяют сказать, что частные решения вида (10) уравнения (7) будут огра-

ничены в том и только в том случае, когда все корни характеристического уравнения (11) будут принадлежать единичному кругу $|\lambda| \leqslant 1$, при этом корни, по модулю равные единице, т. е. лежащие на окружности единичного круга, все должны быть однократными.

Так как всякое решение уравнения (7) есть линейная комбинация решений вида (10), то то же условие будет необходимым и достаточным для ограниченности любого решения ε_n^1 этого уравнения.

Обратим внимание еще на характер оценки решения ε_n^1 в зависимости от оценки начальных значений. Положим $E = \max\limits_{i=0, 1, \ldots, k} |\varepsilon_i|$. Выше мы обращали внимание на то, что при определении c_j^i в представлении (11) по начальным значениям $\varepsilon_0, \ldots, \varepsilon_k$ эти коэффициенты получатся как линейные однородные функции от $\varepsilon_0, \ldots, \ldots, \varepsilon_k$. Поэтому для них верна оценка вида

$$|c_j^i| \leqslant M_0 E.$$

Но тогда, ввиду ограниченности всех частных решений (10), из представления (11) вытекает оценка для ε_n^1, справедливая для $n = 0, 1, 2, \ldots$:

$$|\varepsilon_n^1| \leqslant ME, \tag{13}$$

где M не зависит от начальных значений $\varepsilon_0, \ldots, \varepsilon_k$.

Возвратимся к примеру вычисления неопределенного интеграла посредством формулы (3). Соответствующее ему уравнение для ε_n^1 будет следующим:

$$\varepsilon_{n+1}^1 = -4\varepsilon_n^1 + 5\varepsilon_{n-1}^1.$$

Характеристическое уравнение здесь есть

$$\lambda^2 = -4\lambda + 5.$$

Оно имеет корни $\lambda_1 = 1$ и $\lambda_2 = -5$. Наличие корня -5 привело к тому, что часть погрешности ε_n^1 на каждом шаге изменяла знак и увеличивалась приблизительно в пять раз. Это вызвало быстрый рост погрешности и сделало формулу (3) непригодной для вычислений даже на небольшое число шагов.

Перейдем теперь к рассмотрению второй части ε_n^2 погрешности, происходящей от ошибки вычислительной формулы. Она является решением задачи (8). Изменим немного обозначения. Погрешность формулы раньше обозначалась r_n, и в этой записи отмечалась зависимость погрешности от положения узла на отрезке $[x_0, X]$. Но она зависит не только от n, но и от величины шага h. Такая зависимость ранее подразумевалась, но в явном виде не указывалась. Сейчас полезно ее отметить, заменив r_n на $r_n(h)$.

Введем оценку погрешности, положив $\max\limits_n |r_n(h)| = r(h)$.

Можно просто видеть, что решение задачи (8) со всеми значениями свободного члена $r_n(h)$ $(n = k, k+1, \ldots)$ может быть разложено на сумму решений с отдельными значениями $r_n(h)$ следующим путем. Нужно решить сначала задачу (8), считая в ней значение $r_k(h)$ заданным, а все остальные значения $r_n(h)$ равными нулю: $r_{k+1}(h) = r_{k+2}(h) = \ldots = 0$. Решение такой задачи назовем $\varepsilon_n^{2,\,k}$. Затем надо решить (8), полагая $r_k(h) = 0$, $r_{k+1}(h)$ — заданным и $r_{k+2}(h) = \ldots = 0$. Решение обозначим $\varepsilon_n^{2,\,k+1}$ и т. д. После этого, очевидно, будет $\varepsilon_n^2 = \varepsilon_n^{2,\,k} + \varepsilon_n^{2,\,k+1} + \ldots + \varepsilon_n^{2,\,N-1}$. Но каждая из указанных частных задач есть задача с начальными значениями (одним значением, отличным от нуля). В самом деле, для $\varepsilon_n^{2,\,k}$ зависимость от n определяется так:

при $n \leqslant k$ даны начальные значения $\varepsilon_0^{2,\,k} = 0, \ldots, \ldots, \varepsilon_k^{2,\,k} = 0$;

при $n = k$ уравнение (8) даст

$$\varepsilon_{k+1}^{2,\,k} = \sum_{i=0}^{k} A_i \varepsilon_{n-i}^{2,\,k} + r_k(h) = r_k(h);$$

при $n > k$ ввиду $r_n(h) = 0$ $(n > k)$ уравнение (8) станет однородным:

$$\varepsilon_n^{2,\,k} = \sum_{i=0}^{k} A_i \varepsilon_{n-i}^{2,\,k}, \qquad n = k+1,\ k+2,\ \ldots \tag{14}$$

Таким образом, для $\varepsilon_n^{2,\,k}$ $(n \geqslant k+1)$ получится однородное уравнение (14) с начальными значениями $\varepsilon_1^{2,\,k} = \ldots$

$\ldots = \varepsilon_k^{2,\,k} = 0$, $\varepsilon_{k+1}^{2,\,k} = r_k(h)$. Эта задача имеет тот же характер, что и (7), но со сдвигом на один узел вправо. Если все собственные значения λ_i принадлежат кругу $|\lambda| \leqslant 1$, и значения λ_i, лежащие на окружности круга, являются простыми, то каждое решение задачи является ограниченным, и для $\varepsilon_n^{2,\,k}$ будет верна оценка (13)

$$|\varepsilon_n^{2,\,k}| \leqslant M |r_k(h)| \leqslant M r(h).$$

Для $\varepsilon_n^{2,\,k+1}$ получится аналогичная однородная задача со сдвигом на два узла вправо и с оценкой

$$|\varepsilon_n^{2,\,k+1}| \leqslant M |r_{k+1}(h)| \leqslant M r(h)$$

и т. д.

Наконец, для ε_n^2 будет верна оценка

$$|\varepsilon_n^2| = |\varepsilon_n^{2,\,k} + \varepsilon_n^{2,\,k+1} + \ldots + \varepsilon_n^{2,\,N-1}| \leqslant M(N-k) r(h) <$$

$$< MNr(h) \leqslant M(b-a)\frac{r(h)}{h} = M_1 \frac{r(h)}{h}. \qquad (15)$$

В приложениях обычно вычислительные схемы (2) выбираются так, чтобы для погрешности r_n схемы имела место оценка

$$|r_n(h)| \leqslant h^s c(h), \qquad (16)$$

где $s > 1$ *), $c(h)$ есть ограниченная функция от h и $c(h) \to c \neq 0$ $(h \to 0)$. Для таких схем верно неравенство

$$|\varepsilon_n^2| \leqslant M_1 h^{s-1} c(h). \qquad (17)$$

Из (13) и (15) для ε_n получается оценка

$$|\varepsilon_n| \leqslant ME + M_1 \frac{r(h)}{h}, \qquad (18)$$

из которой очевидным образом вытекает

Теорема 1. *Пусть для формулы* (2), *для функций* $y(x)$ и $f(x)$ *и для подготовительных вычислений выполняются условия*:

1) $E = \max\limits_{1 \leqslant i \leqslant k} |\varepsilon_i| \to 0$ $(h \to 0)$;

2) $\frac{r(h)}{h} \to 0$ $(h \to 0)$, *где* $r(h) = \max\limits_n |r_n(h)|$.

*) Вычислительные формулы, для которых $s \leqslant 1$, принципиально говоря, возможны, но в практике счета не применяются.

Тогда погрешность ε_n вычислений будет равномерно относительно n стремиться к нулю и приближенная функция y_n $(n = 0, \ldots, N)$ равномерно на $[x_0, X]$ сходиться к неопределенному интегралу $y(x)$.

Остановим еще внимание на понятии устойчивости вычислительной формулы, которая в дальнейшем, например, в теории методов решения дифференциальных уравнений, как обыкновенных, так и, в особенности, с частными производными, будет иметь большое значение. В проблеме неопределенного интегрирования понятие устойчивости является простым, но на нем полезно остановиться как на вводном к другим более сложным случаям.

Выше говорилось о том, что на погрешность ε_n оказывают влияние два фактора: ошибки в вычислении начальных значений $y_0, \ldots, y_k$ и погрешность r_n вычислительного правила. В реальных вычислениях есть еще третий источник погрешностей — ошибки, вызванные округлением чисел. Сейчас этот третий источник мы не рассматриваем, считая, что вычисления выполняются абсолютно точно.

Начнем с части ε_n^1 погрешности, вызываемой ошибками начальных значений $y_0, \ldots, y_k$. Напомним, что ε_n^1 является решением разностного уравнения (7), зависящего только от коэффициентов A_i и не зависящего от h и функции f. Это позволяет формулировать понятие устойчивости для ε_n^1 только с помощью коэффициентов A_i. Определено же это понятие может быть несколькими способами, но все определения либо должны быть равносильны требованию, чтобы при любых начальных значениях погрешности ε_i $(i = 0, 1, \ldots, k)$ величина ε_n^1 при $n = 0, 1, \ldots$ была ограниченной или чтобы это требование содержалось в других как следствие. Приведем одно из определений устойчивости, аналог которого встречается в других вопросах, например в задаче численного решения уравнений с частными производными.

Вычислительная формула (2) *называется устойчивой относительно погрешности начальных значений, если существует число M такое, что из неравенств* $|\varepsilon_i| \leq E$

($i = 0, 1, \ldots, l$) *следует*

$$|\varepsilon_n^1| \leqslant ME \quad (n = 0, 1, \ldots). \tag{19}$$

Теорема 2. *Для устойчивости вычислительной формулы* (2) *относительно погрешности начальных значений необходимо и достаточно, чтобы корни* λ_i *уравнения* $\lambda^{k+1} = \sum_{i=0}^{k} A_i \lambda^{k-i}$ *все лежали в круге* $|\lambda| \leqslant 1$, *при этом корни, имеющие модуль, равный единице, были однократными.*

Доказательство. Необходимость. Из условия устойчивости (19) следует ограниченность всех решений уравнения (7) для ε_n^1. Но это возможно в том и только в том случае, когда все частные решения (10) будут ограниченными, а последнее равносильно условию теоремы 2.

Достаточность: при выполнении условия теоремы 2 все частные решения (10) будут ограничены, а тогда, как было выяснено выше при получении оценки для ε_n^1, верно неравенство (13), совпадающее с (19).

Несколько более сложным является понятие устойчивости (2) относительно погрешности $r_n(h)$, так как эта погрешность зависит не только от коэффициентов A_i и B_j, но также от функции f и шага h. Часть ε_n^2 погрешности ε_n, зависящая от $r_n(h)$, должна быть решением уравнения (8) с нулевыми начальными значениями; при этом величина ε_n^2 должна быть найдена для $n = k$, $k+1, \ldots, N$, где $N =$ ц. ч. $\frac{1}{h}(X - x_0)$. Когда $h \to 0$, множество значений n, для которых находится ε_n^2, неограниченно расширяется.

Условие устойчивости должно, очевидно, содержать требование ограниченности величины ε_n^2 для всех h и всех возможных значений n. Приведем одно из определений устойчивости.

Формула вычислений (2) называется *устойчивой относительно погрешности* $r_n(h)$ этой формулы, если существует такое число M_1, что из неравенства

$$\frac{r(h)}{h} \leqslant G, \quad r(h) = \max_n |r_n(h)| \tag{20}$$

следует оценка

$$|\varepsilon_n^2| < M_1 G \tag{21}$$

при всяких h и $n = k+1, \ldots, N$.

Теорема 3. *Для устойчивости формулы вычислений* (2) *относительно погрешности* $r_n(h)$ *этой формулы достаточно выполнения условий:*

1) *формула устойчива относительно погрешностей начальных значений;*

2) *существует число* G *такое, что при всех значениях* $h > 0$ *выполняется неравенство*

$$\frac{r(h)}{h} \leqslant G, \qquad r(h) = \max_n |r_n(h)|. \tag{22}$$

Доказательство. Выше, при рассмотрении ε_n^2, было показано, что когда выполнено первое условие теоремы, для $\varepsilon_n^2(h)$ верна оценка (15). Если, кроме того, верно неравенство (22), то из (15) следует (21), и это доказывает теорему.

2. Интерполяционная формула вычислений частного вида. Сейчас будет получена вычислительная формула, простейшие случаи которой часто применяются в практике. Предположим, что вычисления доведены до узла $x_n = x_0 + nh$ и составлена таблица, приведенная в начале параграфа. Вычислению подлежит $y(x_{n+1})$. Воспользуемся лишь одним предшествующим значением $y(x_n)$ функции y. Точное значение $y(x_{n+1})$ есть

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(t)\,dt. \tag{23}$$

Оно требует знания f всюду на отрезке $[x_n, x_{n+1}]$. Мы найдем функцию f приближенно, интерполируя ее на $[x_n, x_{n+1}]$ по значениям в узлах сетки. Здесь естественно воспользоваться формулой Ньютона — Бесселя (1.4.6), привлекая одинаковое число узлов с обеих сторон от отрезка $[x_n, x_{n+1}]$. Если взять $2k+2$ узлов, формулу

Ньютона — Бесселя можно записать следующим образом:

$$f(t) = f(x_n + uh) = \frac{1}{2}(f_n + f_{n+1}) + \frac{u-0,5}{1!}\Delta f_n +$$

$$+ \frac{u(u-1)}{2!}\cdot\frac{1}{2}(\Delta^2 f_{n-1} + \Delta^2 f_n) + \frac{(u-0,5)\,u(u-1)}{3!}\Delta^3 f_{n-1} + \ldots$$

$$\ldots + \frac{(u+k-1)\ldots(u-k)}{(2k)!}\cdot\frac{1}{2}(\Delta^{2k} f_{n-k} + \Delta^{2k} f_{n-k+1}) +$$

$$+ \frac{(u-0,5)(u+k-1)\ldots(u-k)}{(2k+1)!}\Delta^{2k+1} f_{n-k} + r(t).$$

Когда функция f имеет на отрезке $[x_n - kh,\ x_n + (k+1)h]$ непрерывную производную порядка $2k+2$, то на этом отрезке существует такая точка ξ, что остаток $r(t)$ можно записать в виде

$$r(t) = h^{2n+2}\frac{(u+k)(u+k-1)\ldots(u-k)}{(2k+2)!} f^{(2k+2)}(\xi).$$

Если приведенное выше представление $f(t)$ внести в интеграл и выполнить почленное интегрирование, получим нужное выражение $y(x_{n+1})$:

$$y(x_{n+1}) = y(x_n) + h\left[\frac{1}{2}(f_n + f_{n+1}) - \frac{1}{12}\cdot\frac{1}{2}(\Delta^2 f_{n-1} + \Delta^2 f_n) +\right.$$

$$+ \frac{11}{720}\cdot\frac{1}{2}(\Delta^4 f_{n-2} + \Delta^4 f_{n-1}) - \frac{191}{60\,480}\cdot\frac{1}{2}(\Delta^6 f_{n-3} + \Delta^6 f_{n-2}) + \ldots$$

$$\left.\ldots + C_k\cdot\frac{1}{2}(\Delta^{2k} f_{n-k} + \Delta^{2k} f_{n-k+1}) + R_{nk},\right.$$

$$C_k = \frac{1}{(2k)!}\int_0^1 (u+k-1)\ldots(u-k)\,du,$$

$$R_{nk} = h\int_0^1 r(x_n + uh)\,du =$$

$$= \frac{h^{2k+3}}{(2k+2)!} f^{(2k+2)}(\eta)\int_0^1 (u+k)(u+k-1)\ldots(u-k-1)\,du,$$

$$x_n - kh \leqslant \eta \leqslant x_n + (k+1)h.$$

При вычислении интеграла от остаточного члена интерполирования, так как ядро интеграла $(u+k)\ldots$

$\dots (u-k-1)$ сохраняет знак на отрезке $0 \leqslant u \leqslant 1$, среднее значение $f^{(2k+2)}$ ввиду непрерывности этой производной можно было вынести за знак интеграла.

Если в равенстве для $y(x_{n+1})$ отбросить неизвестный остаток R_{nk}, получится приближенная вычислительная формула, имеющая погрешность порядка h^{2k+2}. Для применения она требует предварительного нахождения значений $y_1, y_2, \dots, y_k$. Они могут быть вычислены по тому же правилу, если выйти за отрезок $[x_0, X]$ налево на k шагов и найти значения $f(x_n-h)=f_{-1}, \dots$ $\dots, f(x_0-kh)=f_{-k}$. Но можно избежать этой дополнительной затраты труда и получить другие формулы вычислений начала таблицы. Для этого достаточно воспользоваться формулой Ньютона для интерполирования в начале таблицы (1.4.1)

$$f(t)=f(x_0+uh)=$$

$$=f_0+\frac{u}{1!}\Delta f_0+\frac{u(u-1)}{2!}\Delta^2 f_0+\frac{u(u-1)(u-2)}{3!}\Delta^3 f_0+\dots$$

$$\dots+\frac{u(u-1)\dots(u-k+1)}{k!}\Delta^k f_0+r(t), \qquad (24)$$

$$r(t)=h^{k+1}\frac{u(u-1)\dots(u-k)}{(k+1)!}f^{(k+1)}(\xi).$$

Для нахождения $y_1, y_2, \dots$ можно в равенство (23), полагая в нем $n=1, 2, \dots$, вместо $f(t)$ внести выражение (24) и затем выполнить почленное интегрирование. После отбрасывания остатков полученные равенства дадут формулы вычислений начальных значений. Три таких равенства приведены ниже:

$$y(x_1)=y_0+h\Big[\frac{1}{2}(f_0+f_1)-\frac{1}{12}\Delta^2 f_0+\frac{1}{24}\Delta^3 f_0-\frac{19}{720}\Delta^4 f_0+$$

$$+\frac{1}{160}\Delta^5 f_0-\dots+C_k\Delta^k f_0\Big]+R_{nk} \quad (k\geqslant 2),$$

$$C_k=\frac{1}{k!}\int_0^1 u(u-1)\dots(u-k+1)\,du, \qquad (25)$$

$$R_{nk}=h^{k+2}\frac{1}{(k+1)!}f^{(k+1)}(\xi)\int_0^1 u(u-1)\dots(u-k)\,du,$$

$$x_0\leqslant\xi\leqslant x_k;$$

$$y(x_2) = y(x_1) + h\Big[\frac{1}{2}(f_1 + f_2) - \frac{1}{12}\cdot\frac{1}{2}(\Delta^2 f_0 + \Delta^2 f_1) +$$

$$+ \frac{11}{720}\Delta^4 f_0 - \frac{11}{1\,440}\Delta^5 f_0 + \frac{271}{60\,480}\Delta^6 f_0 + \ldots + D_k\Delta^k f_0\Big] + R_{nk}$$

$$(k \geqslant 4), \qquad (26)$$

$$D_k = \frac{1}{k!}\int_0^1 (u+1)\,u\,(u-1)\,\ldots\,(u-k+2)\,du,$$

$$R_{nk} = h^{k+2}\cdot\frac{1}{(k+1)!}\,f^{(k+1)}(\xi)\int_0^1 (u+1)\,u\,\ldots\,(u-k+1)\,du,$$

$$x_0 \leqslant \xi \leqslant x_k;$$

$$y(x_3) = y(x_2) + h\Big[\frac{1}{2}(f_2 + f_3) - \frac{1}{12}\cdot\frac{1}{2}(\Delta^2 f_1 + \Delta^2 f_2) +$$

$$+ \frac{11}{720}\cdot\frac{1}{2}(\Delta^4 f_0 + \Delta^4 f_1) - \frac{191}{60\,480}\Delta^5 f_0 +$$

$$+ \frac{191}{120\,960}\Delta^7 f_0 + \ldots + E_k\Delta^k f_0\Big] + R_{nk} \qquad (k \geqslant 6), \qquad (27)$$

$$E_k = \frac{1}{k!}\int_0^1 (u+2)(u+1)\,\ldots\,(u-k+3)\,du,$$

$$R_{nk} = h^{k+2}\cdot\frac{1}{(k+1)!}\,f^{(k+1)}(\xi)\int_0^1 (u+2)(u+1)\,\ldots$$

$$\ldots\,(u-k+2)\,du, \qquad x_0 \leqslant \xi \leqslant x_k.$$

Сходные формулы могут быть получены для вычисления вблизи точки x_N, являющейся конечным узлом таблицы, для чего достаточно воспользоваться формулой Ньютона, предназначенной для интерполирования

в конце таблицы:

$$y(x_N) = y(x_{N-1}) +$$

$$+h\left[\frac{1}{2}(f_N + f_{N-1}) - \frac{1}{12}\Delta^2 f_{N-2} - \frac{1}{24}\Delta^3 f_{N-3} - \frac{19}{720}\Delta^4 f_{N-4} -\right.$$

$$\left. - \frac{1}{160}\Delta^5 f_{N-5} - \ldots - (-1)^{k-1} C_k \Delta^k f_{N-k}\right] + R'_{nk} \quad (k \geqslant 2),$$

$$y(x_{N-1}) = y(x_{N-2}) + h\left[\frac{1}{2}(f_{N-2} + f_{N-1}) -\right.$$

$$\left. - \frac{1}{12}\cdot\frac{1}{2}(\Delta^2 f_{N-2} + \Delta^2 f_{N-3}) + \frac{11}{720}\Delta^4 f_{N-4} - \frac{11}{1440}\Delta^5 f_{N-5} +\right.$$

$$\left. + \frac{271}{60\,480}\Delta^6 f_{N-6} - \ldots + (-1)^k D_k \Delta^k f_{N-k}\right] + R'_{nk} \quad (k \geqslant 4),$$

$$y(x_{N-2}) = y(x_{N-3}) +$$

$$+h\left[\frac{1}{2}(f_{N-3} + f_{N-2}) - \frac{1}{12}\cdot\frac{1}{2}(\Delta^2 f_{N-4} + \Delta^2 f_{N-3}) +\right.$$

$$\left. + \frac{11}{720}\cdot\frac{1}{2}(\Delta^4 f_{N-5} + \Delta^4 f_{N-4}) - \frac{191}{60\,480}\Delta^6 f_{N-6} -\right.$$

$$\left. - \frac{191}{120\,960}\Delta^7 f_{N-7} - \ldots - (-1)^{k-1} E_k \Delta^k f_{N-k}\right] + R'_{nk} \quad (k \geqslant 6).$$

Представления погрешностей R'_{nk} трех последних формул получаются из выражений погрешностей R_{nk} соответствующих формул (25) — (27) при помощи умножения на $(-1)^{k+1}$ и замены условия $x_0 \leqslant \xi \leqslant x_k$ на условие $x_{N-k} \leqslant \xi \leqslant x_N$.

ЛИТЕРАТУРА

1. Бахвалов Н. С., Численные методы, I, «Наука», М., 1973.
2. Крылов В. И., Приближенное вычисление интегралов, «Наука», М., 1967.
3. Крылов В. И., Шульгина Л. Т., Справочная книга по численному интегрированию, «Наука», М., 1966.
4. Stroud A. H., Don Sekrest, Gaussian quadrature formulas, USA, Prentice-hall, Series in automatic computation.
5. Мак-Кракен Д., Дорн У., Численные методы и программирование на Фортране, «Мир», М., 1969.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Владимир Иванович Крылов,
Владимир Васильевич Бобков,
Петр Ильич Монастырный

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ
ТОМ I

М., 1976 г., 304 стр. с илл.

Редакторы *Н. Н. Калиткин, Н. П. Рябенькая*
Техн. редактор *Л. В. Лихачева*
Корректор *Е. В. Сидоркина*

Сдано в набор 16/XII 1975 г. Подписано к печати 28/V 1976 г. Бумага 84×108¹/₃₂ тип. № 3. Физ. печ. л. 9,5. Условн. печ. л. 15,96. Уч.-изд. л. 15,85. Тираж 44000 экз. Цена книги 71 коп. Заказ № 953.

Издательство «Наука»
Главная редакция
физико-математической литературы,
117071, Москва, В-71, Ленинский проспект, 15.

Ордена Трудового Красного Знамени
Ленинградская типография № 2
имени Евгении Соколовой Союзполиграфпрома
при государственном комитете Совета
Министров СССР по делам издательств,
полиграфии и книжной торговли
198052, Ленинград, Л-52,
Измайловский проспект, 29.