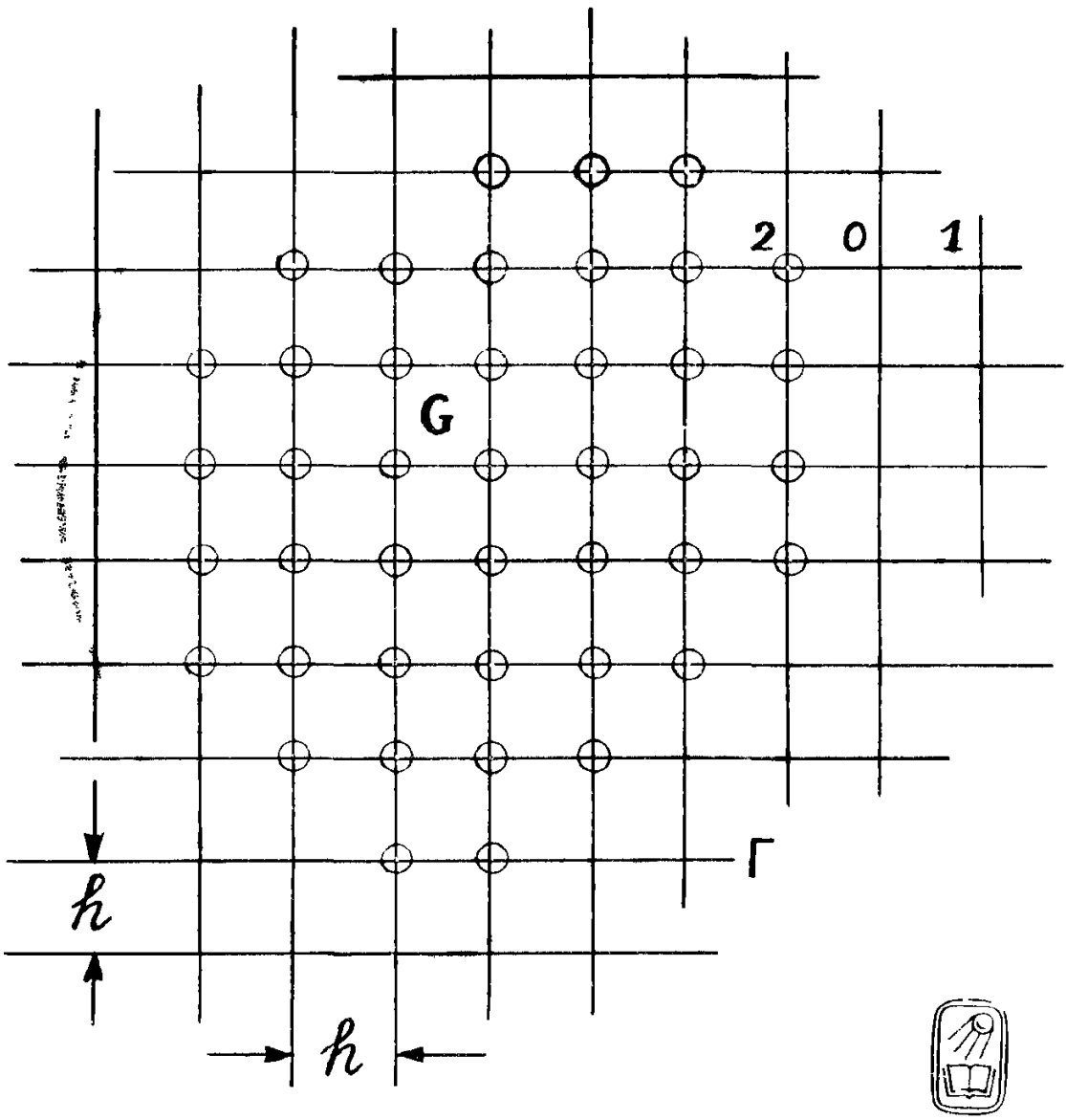
Е. А. ВОЛКОВ

ЧИСЛЕННЫЕ МЕТОДЫ



Е. А. ВОЛКОВ

ЧИСЛЕННЫЕ МЕТОДЫ

ИЗДАНИЕ ВТОРОЕ, ИСПРАВЛЕННОЕ

Допущено Министерством
высшего и среднего специального образования СССР
в качестве учебного пособия
для инженерно-технических специальностей вузов

МОСКВА «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
1987

ББК 22.19
В 67
УДК 519.6(075.8)

Волков Е. А. **Численные методы:** Учеб. пособие для вузов.— 2-е изд., испр. — М.: Наука. Гл. ред. физ.-мат. лит., 1987. — 248 с.

Соответствует разделу численных методов в программе по высшей математике для инженерно-технических специальностей вузов. Тесно примыкает к учебникам по высшей математике С. М. Никольского и Я. С. Бугрова. Книгу отличает сжатость и емкость изложения в сочетании с математической строгостью. Рассмотрены численные методы: линейной алгебры, интегрирования, решения дифференциальных уравнений, а также основные понятия теории приближений.

Первое издание вышло в 1982 г.

Для студентов инженерно-технических специальностей вузов.

Табл. 11. Ил. 27. Библиогр. 23 назв.

Евгений Алексеевич Волков
ЧИСЛЕННЫЕ МЕТОДЫ

Редактор *И. В. Викторенкова*
Художественный редактор *Г. М. Коровина*
Технический редактор *И. Ш. Аксельрод*
Корректоры *Т. С. Вайсберг, Л. С. Сомова*

ИБ № 12823

Сдано в набор 03.11.87. Подписано к печати 24.04.87. Формат 84×108/32. Бумага тип. № 2. Гарнитура литературная. Печать высокая. Усл. печ. л. 13,02. Усл. кр.-отт. 13,34. Уч.-изд. л. 12,62. Тираж 36 000 экз. Заказ № 359. Цена 45 коп.

Ордена Трудового Красного Знамени издательство «Наука»
Главная редакция физико-математической литературы
117071 Москва В-71, Ленинский проспект, 15

Ленинградская типография № 2 головное предприятие ордена Трудового Красного Знамени Ленинградского объединения «Техническая книга» им. Евгении Соколовой Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли.
198052, г. Ленинград, Л-52, Измайловский проспект, 29

В $\frac{1702070000-120}{053(02)-87}$ 57-87

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

В настоящей книге излагаются основы численных методов. Книга содержит материал, предусмотренный программой курса «Высшая математика для инженерно-технических специальностей высших учебных заведений». Для понимания почти всего содержания книги достаточно знания общего курса математики по указанной программе в объеме трех учебников Я. С. Бугрова, С. М. Никольского: Элементы линейной алгебры и аналитической геометрии. — М.: Наука, 1984; Дифференциальное и интегральное исчисление. — М.: Наука, 1984; Дифференциальные уравнения. Кратные интегралы. Ряды. Функции комплексного переменного. — М.: Наука, 1985.

При чтении § 14, 17 потребуются сведения о случайных величинах из элементарного курса теории вероятностей. Некоторые дополнительные понятия, используемые в тексте, разъясняются во введении и по мере необходимости.

Глава 1 посвящена численным методам приближения функций одной переменной. Здесь, кроме многочленов Тейлора, интерполяционных многочленов и многочленов наилучшего равномерного приближения, рассматриваются аппроксимации кубическими сплайнами, значительное место уделяется важному в инженерно-технических приложениях методу наименьших квадратов в непериодическом и периодическом случаях с анализом погрешности самого метода и случайной ошибки, возникающей за счет ошибок наблюдений. В гл. 1 включен также параграф, относящийся к численному дифференцированию, используемому при построении сплайнов, и вводный § 1 о приближенных числах.

В гл. 2 представлены численные методы интегрирования. Наряду с традиционными квадратурными формулами кратко изложен метод Монте-Карло вычисления определенных интегралов. В § 16 обосновывается практическое правило Рунге оценки погрешно-

сти квадратурных формул и метод уточнения результата по Ричардсону. Численным методам интегрирования обыкновенных дифференциальных уравнений (решению задачи Коши) посвящен § 18.

Глава 3 содержит численные методы решения задач линейной алгебры, в частности метод Гаусса, метод итераций решения систем линейных уравнений и методы решения частичных проблем собственных значений матриц. В § 22 рассматривается метод прогонки решения системы с трехдиагональной матрицей (трехточечного разностного уравнения), получивший широкое распространение.

В гл. 4 излагаются основные приближенные численные методы решения нелинейных уравнений и систем нелинейных уравнений.

Глава 5 посвящена приближенным методам решения краевой задачи для линейного обыкновенного дифференциального уравнения второго порядка. В § 28 даны методы минимизации невязки и Галеркина. В § 29 излагается разностный метод. На базе двухточечной краевой задачи вводятся основные понятия и разъясняются основные положения теории разностных схем.

В гл. 6 изучаются разностные схемы для линейных дифференциальных уравнений с частными производными первого и второго порядков. Рассматриваются вопросы аппроксимации, устойчивости, сходимости, а также экономичности разностных схем.

Список литературы рекомендован для более углубленного изучения численных методов.

При написании книги автор опирался на опыт чтения лекций по численным методам в МИФИ и существенно использовал советы академика С. М. Никольского и профессора Я. С. Бугрова, которым он искренне благодарен.

Автор выражает свою глубокую признательность члену-корреспонденту АН СССР Н. С. Бахвалову, профессору В. А. Треногину и доценту Н. А. Потапкову, прочитавшим книгу в рукописи и сделавшим ряд ценных замечаний.

Е. А. Волков

ВВЕДЕНИЕ

На практике в большинстве случаев найти точное решение возникшей математической задачи не удается. Это происходит главным образом не потому, что мы не умеем этого сделать, а поскольку искомое решение обычно не выражается в привычных для нас элементарных или других известных функциях. Поэтому важное значение приобрели численные методы, особенно в связи с возрастанием роли математических методов в различных областях науки и техники и с появлением высокопроизводительных ЭВМ.

Под численными методами подразумеваются методы решения задач, сводящиеся к арифметическим и некоторым логическим действиям над числами, т. е. к тем действиям, которые выполняет ЭВМ. В зависимости от сложности задачи, заданной точности, применяемого метода и т. д. может потребоваться выполнить от нескольких десятков до многих миллиардов действий. Если число действий не превышает тысячи, то с такой задачей обычно может справиться человек, имея в своем распоряжении настольную клавишную счетную машину без программного управления и набор таблиц элементарных функций. Однако без быстродействующей ЭВМ явно не обойтись, если для решения задачи нужно выполнить, скажем, порядка миллиона действий и тем более, когда решение должно быть найдено в сжатые сроки. Например, задачи, связанные с суточным прогнозом погоды, должны быть решены за несколько часов, а при управлении быстро протекающими технологическими процессами требуется находить решение за доли секунды.

Решение, полученное численным методом, обычно является приближенным, т. е. содержит некоторую погрешность. Источниками погрешности приближенного решения являются: 1) несоответствие математической задачи (математической модели) изучаемому реальному явлению; 2) погрешность исходных данных

(входных параметров); 3) погрешность метода решения; 4) погрешности округлений в арифметических и других действиях над числами.

Погрешность в решении, обусловленная первыми двумя источниками, называется *неустранимой*. Эта погрешность может присутствовать, даже если решение поставленной математической задачи найдено точно. Вопрос о том, насколько хорошо описывает математическая модель исследуемое явление, проверяется путем сравнения результатов экспериментов и типичных частных решений при некоторых значениях входных параметров. Влияние погрешности исходных данных часто удается оценить элементарными средствами, например варьируя исходные данные в пределах их погрешностей и фиксируя решения. Если исходных данных много, а их погрешности носят случайный характер, то на помощь могут прийти статистические методы. (В § 14 даются оценки случайной ошибки при аппроксимации функций методом наименьших квадратов.) В некоторых случаях неустранимую погрешность можно рассматривать как погрешность функции, возникающую за счет погрешности аргументов. (Оценка такой погрешности, опирающаяся на формулу Лагранжа, рассматривается в § 1.)

Численные методы в большинстве случаев сами по себе являются приближенными, т. е. даже при отсутствии погрешностей во входных данных и при идеальном выполнении арифметических действий они дают решение исходной задачи с некоторой погрешностью, называемой *погрешностью метода*. Это происходит потому, что численным методом обычно решается некоторая другая, более простая задача, аппроксимирующая (приближающая) исходную задачу. В ряде случаев используемый численный метод строится на базе бесконечного процесса, который в пределе приводит к искомому решению (например, вычисление значений элементарной функции с помощью частичных сумм степенного ряда, в который разлагается эта функция). Однако реально предельный переход обычно не удается осуществить, и процесс, прерванный на некотором шаге, дает приближенное решение.

Исследованию погрешности численных методов уделяется значительное внимание во всех разделах

данной книги. Численный метод обычно зависит от одного или нескольких параметров, которыми можно распоряжаться. В качестве такого параметра служит, например, число итераций при решении систем уравнений или число учитываемых членов при суммировании ряда, а также шаг, с которым используются значения подынтегральной функции при приближенном вычислении определенного интеграла. Погрешность метода или получаемая ее оценка обычно зависит от соответствующего параметра. Иногда удается получить оценку погрешности, выражаемую только через известные величины.

С помощью этой оценки можно определить значения параметра, задающего метод, при которых погрешность метода лежит в требуемых пределах. Чаще же оценка погрешности содержит неизвестные постоянные множители, а параметр метода входит в нее в виде либо степенной, либо показательной функции. По такой оценке судят о скорости убывания погрешности при изменении параметра метода. Скорость убывания погрешности является важной характеристикой метода.

Вопросом, наиболее сложным технически, является учет погрешностей округления в арифметических действиях. Если действий выполняется немного, то *погрешности округления* при ручных вычислениях можно учесть методом, изложенным в § 1, посвященном элементарной теории погрешностей, рассчитанной на небольшое число действий. При решении задач на ЭВМ характерны две ситуации.

Если количество выполняемых арифметических действий невелико, то обычно погрешности округления не проявляются, так как в ЭВМ числа представляются с 10 и более десятичными значащими цифрами, а окончательный результат редко бывает нужен более чем с 5 десятичными значащими цифрами. Если задача сложная (например, сводящаяся к решению уравнений с частными производными) и для ее приближенного решения потребуется, скажем, 10^7 арифметических действий, то в этой ситуации нереально учитывать влияние погрешностей округления в каждом действии. При таком учете получится слишком завышенная оценка погрешности, отвечающая самому неблагоприятному случаю. Погрешности округления

ведут себя достаточно случайно как по величине, так и по знаку. Поэтому имеются предпосылки для их взаимной компенсации. Однако с погрешностями округления при решении сложных задач все равно приходится серьезно считаться.

Для решения одной и той же задачи могут применяться различные приближенные методы. Чувствительность к погрешностям округления существенно зависит от выбранного численного метода. В гл. 5, 6 исследуются разностные методы решения краевых задач для дифференциальных уравнений и выделяются так называемые устойчивые методы (разностные схемы), которые, в частности, мало чувствительны к погрешностям округления. Мало чувствительными к таким погрешностям являются итерационные сходящиеся методы, поскольку возникающие погрешности на следующих итерациях исправляются.

Численный метод может считаться удачно выбранным, если его погрешность в несколько раз меньше неустранимой погрешности, а погрешность, возникающая за счет округлений, называемая *вычислительной погрешностью*, по крайней мере в несколько раз меньше погрешности метода. Если неустранимая погрешность отсутствует, то погрешность метода должна быть несколько меньше заданной точности решения.

К численному методу, кроме требования достижения заданной точности, предъявляется ряд других требований. Предпочтение отдается методу, который реализуется с помощью меньшего числа действий, требует меньшей памяти ЭВМ и, наконец, является логически более простым, что способствует более быстрой его реализации на ЭВМ. Перечисленные условия обычно противоречат друг другу, поэтому часто при выборе численного метода приходится соблюдать компромисс между ними.

Введем некоторые понятия математического и функционального анализа, которые широко распространены в современной литературе по численным методам. С этими понятиями можно знакомиться по мере того, как они будут встречаться в основном тексте.

1. Пусть $\varphi(h)$ — некоторая функция переменной h с конечной областью определения D_φ на полуоси

$h > 0$, причем $h \in D_\varphi$ может принимать сколь угодно малые значения. Тогда, если существуют такие положительные числа h_0, c, k, что при всех $h \in D_\varphi$, удовлетворяющих условию $0 < h \leqslant h_0$, выполняется неравенство

$$|\varphi(h)| \leqslant ch^k,$$

то пишут

$$\varphi(h) = O(h^k)$$

и говорят, что $\varphi(h)$ *есть О большое от* h^k (*при* $h \to 0$).

Согласно данному определению выполняются следующие очевидные свойства. Если $\varphi(h) = O(h^k)$, $\psi(h) = O(h^k)$, причем $D_\varphi = D_\psi$ то

$$\varphi(h) + \psi(h) = O(h^k),$$

т. е.

$$O(h^k) + O(h^k) = O(h^k).$$

Если $k > m > 0$, то $O(h^k)$ в то же время есть $O(h^m)$. Наконец, если $\varphi(h) = O(h^k)$, то $a\varphi(h) = O(h^k)$, где a — постоянная, не зависящая от h.

Пример. $\sin^2 2h = O(h^2)$, так как $\sin^2 2h \leqslant 4h^2$.

Пусть теперь дана функция $\varphi(h, \tau)$ положительных аргументов h, τ, которые могут принимать сколь угодно малые значения. Тогда, если существуют такие положительные числа h_0, τ_0, c, k, m, что при всех допустимых значениях h, τ, удовлетворяющих условиям $0 < h \leqslant h_0$, $0 < \tau \leqslant \tau_0$, выполняется неравенство

$$|\varphi(h, \tau)| \leqslant c(h^k + \tau^m),$$

то пишут

$$\varphi(h, \tau) = O(h^k + \tau^m)$$

и говорят, что $\varphi(h, \tau)$ *есть О большое от* $h^k + \tau^m$ (*при* $h \to 0$, $\tau \to 0$).

Введем еще одно аналогичное понятие. Функция $\Phi(N)$, заданная для всех натуральных $N > N_0 > 0$, *есть О большое от* N^k (*при* $N \to \infty$), т. е. $\Phi(N) = O(N^k)$, если найдется такая постоянная $c > 0$, что при всех натуральных $N > N_0$

$$|\Phi(N)| \leqslant cN^k.$$

2. Будем говорить, что функция $f(x)$ *принадлежит классу* $C_k[a, b]$, и писать $f \in C_k[a, b]$, если функция f определена на отрезке $[a, b]$ и имеет на нем непрерывные производные до порядка k включительно. Это означает, что на некотором интервале (A, B), содержащем отрезок $[a, b]$, существует k раз непрерывно дифференцируемая функция f^*, совпадающая

на $[a, b]$ с f. Под значениями указанных производных функции f на концах отрезка $[a, b]$ подразумеваются значения соответствующих производных функции f^*.

Пример. Пусть на отрезке $[0, 1]$ задана функция $f(x) = x^{5/2}$. Рассмотрим на интервале $(-1, 2)$ функцию $f^*(x) = |x|^{5/2}$. Очевидно, отрезок $[0, 1]$ содержится в интервале $(-1, 2)$, функция f^* дважды непрерывно дифференцируема на интервале $(-1, 2)$, а именно,

$$\frac{df^*(x)}{dx} = \frac{5}{2}|x|^{3/2}\operatorname{sign} x,$$

$$\frac{d^2f^*(x)}{dx^2} = \frac{15}{4}|x|^{1/2}.$$

Кроме того, $f^*(x) = f(x) \forall x \in [0, 1]$. Следовательно, $f \in C_2[0, 1]$, причем $f'(x) = \frac{5}{2}x^{3/2}$, $f''(x) = \frac{15}{4}x^{1/2}$, $x \in [0, 1]$.

Будем при $k = 0$ вместо $C_0[a, b]$ использовать обозначение $C[a, b]$. Запись $f \in C[a, b]$ означает, что *функция f непрерывна на отрезке $[a, b]$*.

3. В курсе дифференциального исчисления даются понятия частных производных функции многих переменных во внутренних точках области ее определения. Для формулировок задач требуется расширить эти понятия на замкнутую область, включая граничные точки.

Пусть функция f задана на замкнутой области $\bar{G}$ n-мерного пространства $\mathbf{R}^n$, $n \geqslant 2$. Скажем, что f *имеет на замкнутой области $\bar{G}$ некоторую непрерывную частную производную*, если на какой-либо открытой области G^*, содержащей замкнутую область $\bar{G}$, существует функция f^*, совпадающая с f на $\bar{G}$ и имеющая на G^* соответствующую непрерывную частную производную. При этом в качестве значений рассматриваемой частной производной функции f в граничных точках области G примем значения соответствующей частной производной функции f^*.

Определение. Будем говорить, что функция f *непрерывно дифференцируема k раз на замкнутой области $\bar{G}$*, т. е. принадлежит классу $C_k(\bar{G})$, и писать $f \in C_k(\bar{G})$, если функция f имеет на $\bar{G}$ все непрерывные частные производные до порядка k включительно.

Пример. Пусть на замкнутом квадрате $\bar{G} = \{(x_1, x_2): 0 \leqslant x_i \leqslant 1, i = 1, 2\}$ дана функция $f(x_1, x_2) = x_2 \cos(x_1 - 2)^{1/2}$.

5. Множество M называется *линейным пространством*, если в нем определены операции сложения и умножения на действительные числа, не выводящие за пределы M и удовлетворяющие условиям:

1) *сложение ассоциативно*: $(f+g)+r=f+(g+r)$;

2) *сложение коммутативно*: $f+g=g+f$;

3) *существует нулевой элемент* $\theta \in M$, *т. е.* $f+\theta=f\ \forall f \in M$;

4) $0 \cdot f=\theta\ \forall f \in M$;

5) $(\alpha+\beta)f=\alpha f+\beta f$;

6) $\alpha(f+g)=\alpha f+\alpha g$;

7) $\alpha(\beta f)=(\alpha\beta)f$;

8) $1 \cdot f=f$.

Здесь α, β — действительные числа.

6. Система элементов $f_1, f_2, \ldots, f_n$ линейного пространства M называется *линейно зависимой*, если существуют такие не равные одновременно нулю числа $\alpha_1, \alpha_2, \ldots, \alpha_n$, что

$$\alpha_1 f_1+\alpha_2 f_2+\ldots+\alpha_n f_n=\theta.$$

Если данное равенство возможно только при $\alpha_1=\alpha_2=\ldots=\alpha_n=0$, то система элементов $f_1, f_2, \ldots, f_n$ называется *линейно независимой*.

7. Множество F называется *линейным нормированным пространством*, если оно линейно и каждому элементу $f \in F$ поставлено в соответствие действительное число $\|f\|$, называемое *нормой* f и удовлетворяющее *аксиомам нормы*:

1) $\|f\| \geqslant 0$, *причем* $\|f\|=0$ *тогда и только тогда, когда* $f=\theta$, *т. е.* f *является нулевым элементом в* F;

2) $\|\alpha f\|=|\alpha|\,\|f\|$ *для любого действительного* α;

3) $\|f+g\| \leqslant \|f\|+\|g\|\ \forall f, g \in F$.

Аксиома 3) называется *неравенством треугольника* для нормы.

Рассмотрим пример нормированного пространства. Класс $C[a, b]$ всех непрерывных функций, заданных на отрезке $[a, b]$, очевидно является линейным пространством, так как сумма любых двух непрерывных функций непрерывна и непрерывная функция, умноженная на любое число, тоже непрерывна. Нулевым элементом в этом пространстве является единственная функция, тождественно равная нулю на $[a, b]$.

Если ввести в классе $C[a, b]$ норму

$$\|f\| = \max_{[a, b]} |f(x)|, \tag{1}$$

то он становится нормированным пространством. Выполнение аксиом 1), 2) для введенной нормы очевидно.

Проверим неравенство треугольника. Имеем

$$\|f + g\| = \max_{[a, b]} |f(x) + g(x)|.$$

Так как $|f(x) + g(x)|$ является непрерывной функцией на отрезке $[a, b]$, то по теореме Вейерштрасса найдется такая точка $x^* \in [a, b]$, что

$$\max_{[a, b]} |f(x) + g(x)| = |f(x^*) + g(x^*)|.$$

Отсюда легко следует неравенство треугольника:

$$\|f + g\| = |f(x^*) + g(x^*)| \leqslant |f(x^*)| + |g(x^*)| \leqslant$$
$$\leqslant \max_{[a, b]} |f(x)| + \max_{[a, b]} |g(x)| = \|f\| + \|g\|.$$

8. *Любое линейное нормированное пространство является одновременно метрическим пространством с расстоянием (метрикой)*

$$\rho(f, g) = \|f - g\|. \tag{2}$$

Действительно, аксиомы 1), 2) для введенного расстояния выполняются в силу аксиомы 1) и аксиомы 2) при $\alpha = -1$ для нормы *), а неравенство треугольника для расстояния (2) следует из неравенства треугольника для нормы:

$$\rho(f, r) = \|f - r\| = \|(f - g) + (g - r)\| \leqslant$$
$$\leqslant \|f - g\| + \|g - r\| = \rho(f, g) + \rho(g, r).$$

9. В одном и том же линейном пространстве норму можно вводить различными способами. Например, в классе непрерывных функций $C[0, 1]$ норму можно задать, кроме способа (1), еще в следующем виде:

$$\|f\| = \left(\int_0^1 f^2(x)\, dx \right)^{1/2}. \tag{3}$$

*) Аксиомы расстояния сформулированы в п. 4, а аксиомы нормы — в п. 7.

Мы не станем останавливаться на проверке аксиом для данной нормы, отложив этот вопрос до § 13.

Для того чтобы различать нормы, мы будем в случае необходимости использовать индексы у нормы. Например, для нормы (1) распространены обозначения $\|f\|_{C[a,\,b]}$ и $\|f\|_C$, а для нормы (3) часто используется знак $\|f\|_{L_2}$.

В зависимости от введенной в одном и том же линейном пространстве нормы могут изменяться различные свойства получаемого нормированного пространства и, в частности, изменяется «физический» смысл метрики, порождаемой нормой. Например, расстояние (2), заданное с помощью нормы (1), т. е.

$$\rho(f,\,g)=\|f-g\|_C=\max_{[a,\,b]}|f(x)-g(x)|, \qquad (4)$$

означает максимальное по модулю уклонение функций f и g на отрезке $[a, b]$, а при использовании нормы (3) расстояние имеет вид

$$\rho(f,\,g)=\|f-g\|_{L_2}=\left(\int_0^1 (f(x)-g(x))^2\,dx\right)^{1/2} \qquad (5)$$

и носит смысл среднеквадратичного уклонения функций f и g.

Выбор нормы часто диктуется условиями конкретной задачи. Например, при рассмотрении равномерных приближений функций (§ 12) нужна норма (1), а при приближении функций методом наименьших квадратов (§ 13) естественно напрашивается использование нормы типа (3).

10. В метрическом и нормированном пространствах вводится понятие сходимости последовательности его элементов аналогично тому, как в n-мерном пространстве $\mathbf{R}^n$.

Определение. Говорят, что последовательность $\{f_n\}$ элементов метрического (нормированного) пространства *сходится* к его элементу f, если $\rho(f, f_n)\to 0$ ($\|f-f_n\|\to 0$) при $n\to\infty$.

Сходимость в метрическом пространстве называется *сходимостью по метрике*, а в нормированном пространстве — *сходимостью по норме*. Сходимость по метрике (4), или, что одно и то же, по норме (1),

является *равномерной*, а сходимость по метрике (5), или по норме (3) носит название *сходимости в среднеквадратичном смысле* или просто *сходимости в среднем*.

Приближенные решения многих задач, а также их точные решения бывает удобно трактовать как элементы некоторого метрического или нормированного пространства. При этом погрешность измеряется расстоянием или нормой разности точного и приближенного решений в соответствующем пространстве.

ГЛАВА 1

ПРИБЛИЖЕНИЕ ФУНКЦИЙ МНОГОЧЛЕНАМИ

Большинство численных методов основано на замене более сложных объектов, уравнений и т. д. более простыми. Одно из центральных понятий в математике — функция. Наиболее удобной в обращении на практике функцией является алгебраический многочлен. Чтобы задать многочлен, нужно задать только конечное число его коэффициентов. Значения многочлена просто вычисляются, его легко продифференцировать, проинтегрировать и т. д. Поэтому алгебраические многочлены нашли широкое применение для приближения (аппроксимации) функций.

В настоящей главе излагаются практические методы приближения функций одной переменной на отрезке. В частности, рассматриваются приближения с помощью алгебраических многочленов четырех видов: 1) многочлены Тейлора; 2) интерполяционные многочлены; 3) многочлены наилучшего равномерного приближения и близкие к ним многочлены; 4) многочлены наилучшего среднеквадратичного приближения (найденные методом наименьших квадратов).

В данную главу включен также § 10, посвященный численному дифференцированию. В нем даны простейшие формулы численного приближенного дифференцирования с использованием значений функции в конечном числе точек и освещается вопрос о приближении производными интерполяционного многочлена соответствующих производных функции на отрезке.

В § 11 рассматриваются приближения кубическими сплайнами (гладкими функциями, составленными из конечного числа алгебраических многочленов третьей степени).

В методе наименьших квадратов (§ 13) наряду с алгебраическими многочленами применяются также тригонометрические многочлены, которые являются

более естественными для приближения периодических функций.

В § 14 рассматривается погрешность этого метода и изучается влияние случайных ошибок в дискретных значениях функции (ошибок наблюдений) на аппроксимирующий многочлен, найденный методом наименьших квадратов. В частности, обсуждается вопрос сглаживания наблюдений.

§ 1. Приближенные числа и действия с ними

Пусть a — точное, вообще говоря, неизвестное числовое значение некоторой величины, a^* — известное приближенное числовое значение этой величины (*приближенное число*). Величина

$$\Delta(a^*)=|a-a^*| \tag{1}$$

называется *абсолютной погрешностью* приближенного числа a^*, а величина

$$\delta(a^*)=\frac{\Delta(a^*)}{|a^*|} \tag{2}$$

называется его *относительной погрешностью*.

Любое число $\overline{\Delta}(a^*)$ $(\overline{\delta}(a^*))$, о котором известно, что

$$\Delta(a^*)\leqslant\overline{\Delta}(a^*) \qquad (\delta(a^*)\leqslant\overline{\delta}(a^*)), \tag{3}$$

называется *предельной абсолютной* (*предельной относительной*) *погрешностью* приближенного числа a^*. Будем для определенности считать, что если $a^*\neq 0$ и известны числа $\overline{\Delta}(a^*)$ и $\overline{\delta}(a^*)$, то они связаны соотношением $\overline{\delta}(a^*)=\overline{\Delta}(a^*)/|a^*|$.

Учет погрешностей в арифметических действиях. Очевидно, что если $c=a+b$, $c^*=a^*+b^*$ или $c=a-b$, $c^*=a^*-b^*$, то

$$\Delta(c^*)\leqslant\Delta(a^*)+\Delta(b^*) \tag{4}$$

и, следовательно, в качестве $\overline{\Delta}(c^*)$ естественно взять

$$\overline{\Delta}(c^*)=\overline{\Delta}(a^*)+\overline{\Delta}(b^*). \tag{5}$$

Таким образом, *при сложении и вычитании двух приближенных чисел их предельные абсолютные погрешности складывают*. Это правило распространяется

на алгебраическое сложение любого конечного числа приближенных чисел.

Пусть $u = ab$, $u^* = a^*b^*$, $v = a/b$, $v^* = a^*/b^*$. Имеем

$$\Delta(u^*) = |u - u^*| = |ab - a^*b^*| =$$
$$= |ab - a^*b + a^*b - a^*b^*| \leqslant$$
$$\leqslant |b||a - a^*| + |a^*||b - b^*| \leqslant$$
$$\leqslant (|b^*| + \Delta(b^*))\Delta(a^*) + |a^*|\Delta(b^*),$$

т. е. абсолютная погрешность произведения двух приближенных чисел a^* и b^* удовлетворяет неравенству

$$\Delta(u^*) \leqslant |b^*|\Delta(a^*) + |a^*|\Delta(b^*) + \Delta(a^*)\Delta(b^*). \qquad (6)$$

Аналогично, при условии, что $|b^*| > \Delta(b^*)$ и, следовательно, $b \neq 0$, получаем

$$\Delta(v^*) = \left|\frac{a}{b} - \frac{a^*}{b^*}\right| = \left|\frac{ab^* - a^*b}{bb^*}\right| =$$
$$= \left|\frac{ab^* - a^*b^* + a^*b^* - a^*b}{bb^*}\right| \leqslant \frac{|b^*|\Delta(a^*) + |a^*|\Delta(b^*)}{(|b^*| - \Delta(b^*))|b^*|}.$$

Отсюда, поскольку $|b^*| - \Delta(b^*) = (1 - \delta(b^*))|b^*| > 0$, приходим к следующему неравенству для абсолютной погрешности частного двух приближенных чисел:

$$\Delta(v^*) \leqslant \frac{|b^*|\Delta(a^*) + |a^*|\Delta(b^*)}{(1 - \delta(b^*))|b^*|^2}. \qquad (7)$$

Пусть $a^* \neq 0$, $|b^*| > \Delta(b^*)$. Разделив неравенство (6) на произведение $|a^*||b^*|$, а неравенство (7) — на отношение $|a^*|/|b^*|$, получим неравенства, которым удовлетворяют относительные погрешности произведения и частного:

$$\delta(u^*) \leqslant \delta(a^*) + \delta(b^*) + \delta(a^*)\delta(b^*), \qquad (8)$$

$$\delta(v^*) \leqslant \frac{\delta(a^*) + \delta(b^*)}{1 - \delta(b^*)}. \qquad (9)$$

Для каждого из неравенств (4), (6)—(9) можно привести примеры, в которых $\Delta(a^*) \neq 0$, $\Delta(b^*) \neq 0$ и соответствующее неравенство обращается в равенство. С другой стороны, может случиться, что $\Delta(a^*) \neq 0$, $\Delta(b^*) \neq 0$, а левая часть какого-либо из неравенств (4), (6)—(9) фактически равна нулю.

В силу неравенств (6)—(9) в качестве предельных абсолютных и предельных относительных погрешно-

стей произведения и частного можно взять следующие величины:

$$\overline{\Delta}(u^*)=|b^*|\overline{\Delta}(a^*)+|a^*|\overline{\Delta}(b^*)+\overline{\Delta}(a^*)\overline{\Delta}(b^*), \quad (10)$$

$$\overline{\Delta}(v^*)=\frac{|b^*|\overline{\Delta}(a^*)+|a^*|\overline{\Delta}(b^*)}{(1-\bar{\delta}(b^*))|b^*|^2}, \quad (11)$$

$$\bar{\delta}(u^*)=\bar{\delta}(a^*)+\bar{\delta}(b^*)+\bar{\delta}(a^*)\bar{\delta}(b^*), \quad (12)$$

$$\bar{\delta}(v^*)=\frac{\bar{\delta}(a^*)+\bar{\delta}(b^*)}{1-\bar{\delta}(b^*)}. \quad (13)$$

В формулах (11) и (13) предполагается, что $\overline{\delta}(b^*)<1$.

Поскольку обычно предельные относительные погрешности $\overline{\delta}(a^*)$, $\overline{\delta}(b^*)$ и, следовательно, относительные погрешности $\delta(a^*)$, $\delta(b^*)$ значительно меньше единицы, то вместо формул (10)—(13) используют приближенные формулы

$$\overline{\Delta}(u^*)\approx|b^*|\overline{\Delta}(a^*)+|a^*|\overline{\Delta}(b^*), \quad (14)$$

$$\overline{\Delta}(v^*)\approx\frac{|b^*|\overline{\Delta}(a^*)+|a^*|\overline{\Delta}(b^*)}{|b^*|^2}, \quad (15)$$

$$\bar{\delta}(u^*)\approx\bar{\delta}(a^*)+\bar{\delta}(b^*), \quad (16)$$

$$\bar{\delta}(v^*)\approx\bar{\delta}(a^*)+\bar{\delta}(b^*). \quad (17)$$

Итак, *при сложении и вычитании приближенных чисел* согласно (5) *складывают предельные абсолютные погрешности, а при умножении и делении приближенных чисел* в соответствии с (16), (17) *складывают их предельные относительные погрешности.*

Погрешность функции. Рассмотрим для определенности функцию двух переменных $f(x, y)$, которая непрерывно дифференцируема в некоторой области G. Допустим, что числа x_0^*, y_0^* являются приближенными значениями координат некоторой точки (x_0, y_0), причем замкнутый прямоугольник

$$\overline{R}=\{(x, y):\ |x-x_0^*|\leqslant\overline{\Delta}(x_0^*),\quad |y-y_0^*|\leqslant\overline{\Delta}(y_0^*)\},$$

содержащий обе точки (x_0^*, y_0^*) и (x_0, y_0), целиком находится в области G.

Обозначим $z_0 = f(x_0, y_0)$, $z_0^* = f(x_0^*, y_0^*)$. По формуле конечных приращений Лагранжа имеем

$$z_0 - z_0^* = f'_x(\xi, \eta)(x_0 - x_0^*) + f'_y(\xi, \eta)(y_0 - y_0^*), \quad (18)$$

где ξ, η — координаты некоторой точки отрезка, соединяющего точки (x_0, y_0) и (x_0^*, y_0^*). Отсюда

$$\Delta(z_0^*) = |z_0 - z_0^*| \leqslant |f'_x(\xi, \eta)| \Delta(x_0^*) + |f'_y(\xi, \eta)| \Delta(y_0^*). \quad (19)$$

Пусть c_1, c_2 — некоторые оценки для $|f'_x|$, $|f'_y|$ на прямоугольнике $\bar{R}$, т. е.

$$\max_{\bar{R}} |f'_x| \leqslant c_1, \quad \max_{\bar{R}} |f'_y| \leqslant c_2. \quad (20)$$

Желательно иметь оценки c_1, c_2 возможно меньшими.

Поскольку $(\xi, \eta) \in \bar{R}$, то с учетом (3), (19), (20) можно принять в качестве *предельной абсолютной погрешности значения функции* $z_0^* = f(x_0^*, y_0^*)$ величину

$$\bar{\Delta}(z_0^*) = c_1 \bar{\Delta}(x_0^*) + c_2 \bar{\Delta}(y_0^*). \quad (21)$$

Иногда поступают нестрого, полагая

$$\bar{\Delta}(z_0^*) \approx |f'_x(x_0^*, y_0^*)| \bar{\Delta}(x_0^*) + |f'_y(x_0^*, y_0^*)| \bar{\Delta}(y_0^*). \quad (22)$$

Часто встречается обратная задача, а именно, задано условие, чтобы абсолютная погрешность $\Delta(z_0^*)$ значения функции не превышала некоторого числа $\varepsilon > 0$, т. е.

$$\Delta(z_0^*) \leqslant \varepsilon. \quad (23)$$

Требуется найти ограничения на погрешности аргументов.

Предположим теперь дополнительно, что область G, в которой задана дифференцируемая функция $f(x, y)$, выпуклая, т. е. любые две точки области G можно соединить прямолинейным отрезком, целиком расположенным в G. Например, область G является открытым кругом или прямоугольником. Пусть также

$$\sup_G |f'_x| \leqslant C_1, \quad \sup_G |f'_y| \leqslant C_2 \quad (24)$$

и, как прежде, x_0^*, y_0^* являются приближенными значениями координат некоторой точки (x_0, y_0), причем

знаков, то прибегают к записи в нормализованном виде.

Например, $a^* = 0{,}390 \cdot 10^5$. Из этой записи понятно, что у числа a^* три верные значащие цифры. В данной ситуации запись вида $a^* = 39000$ недопустима. В нормализованном виде можно записать и рассмотренные выше приближенные числа: $0{,}344 \cdot 10^{-1}$, $0{,}34400 \cdot 10^{-1}$.

Часто употребляют запись вида

$$a = a^* \pm \bar{\Delta}(a^*),$$

означающую, что неизвестная величина a удовлетворяет неравенствам

$$a^* - \bar{\Delta}(a^*) \leqslant a \leqslant a^* + \bar{\Delta}(a^*).$$

При этом *величина* $\bar{\Delta}(a^*)$ *выписывается с одной или двумя значащими цифрами, а младший разряд в* a^* *соответствует младшему разряду в* $\bar{\Delta}(a^*)$.

Например,

$$a = 2{,}730 \pm 0{,}017. \tag{26}$$

Записи

$$a = 2{,}73018 \pm 0{,}017,$$
$$a = 2{,}73018 \pm 0{,}01592 \tag{27}$$

неестественны.

Если по условию известно, что некоторое число точное, например, шаг вычислений $h = 0{,}1$, то верные нули справа не выписываются.

Можно говорить о числе верных значащих цифр у приближенного числа и о числе верных цифр после запятой. Как правило, при реальных вычислениях у приближенных чисел содержатся цифры после запятой, т. е. имеется дробная часть.

Например, приближенное число $a^* = 25{,}030$ имеет 5 верных значащих цифр и 3 верные цифры после запятой, а у числа $b^* = 0{,}00404$, наоборот, 3 верные значащие цифры и 5 верных цифр после запятой.

Очевидно, *абсолютная погрешность приближенного числа вполне характеризуется числом верных цифр после запятой, а относительная погрешность — числом верных значащих цифр.*

Округление чисел. При вычислениях часто возникает необходимость округления чисел, т. е. представления их с меньшим числом разрядов.

Правило округления чисел. *Если в старшем из отбрасываемых разрядов стоит цифра меньше пяти, то содержимое сохраняемых разрядов числа не изменяется. В противном случае в младший сохраняемый разряд добавляется единица с тем же знаком, что и у самого числа.*

Это простое правило округления чисел применяется и в ЭВМ.

Пример. Округлить соответственно с двумя, тремя и четырьмя знаками после запятой следующие числа: 3,14159, —0,0025, 84,009974. Ответ. 3,14, —0,003, 84,0100.

Очевидно, *погрешность, возникающая при округлении, не превышает по абсолютной величине половины единицы младшего оставляемого разряда.*

Повторное округление не рекомендуется, так как оно может привести к увеличению погрешности. Например, если число 18,34461 сначала округлить с тремя знаками после запятой, а затем с двумя знаками, то мы последовательно получим 18,345; 18,35. При округлении сразу до двух десятичных знаков после запятой имеем 18,34. Абсолютная погрешность при повторном округлении получилась равной 0,00539, а при одноразовом округлении абсолютная погрешность равна 0,00461.

При округлении приближенного числа его предельная абсолютная погрешность увеличивается с учетом погрешности округления. Например, при переходе от записи (27) к записи (26), когда число 2,73018 округляется до трех знаков после запятой, его предельная абсолютная погрешность 0,01592 увеличивается до величины $0{,}017 > 0{,}01592 + 0{,}00018 > > 0{,}016$, представляемой тоже с тремя знаками после запятой.

На практике округляют постоянные, известные с большим числом знаков, произведения многозначных чисел, частные от деления и т. д. Например, при умножении двух приближенных чисел, имеющих по шесть верных значащих цифр, результат получается с 11 или 12 значащими цифрами. Относительная погрешность произведения может оказаться приблизительно вдвое больше, чем у сомножителей, так как неравенство (8) может обратиться в равенство. Поэтому в произведении верных значащих цифр

приблизительно шесть, а остальные знаки, как правило, не несут полезной информации. Полученное произведение естественно округлить с шестью значащими цифрами.

Имеются следующие правила арифметических действий с приближенными числами:

При умножении и делении приближенных чисел, вообще говоря, с различным числом верных значащих цифр производится округление результата с числом значащих цифр, совпадающим с минимальным числом верных значащих цифр у исходных чисел.

При сложении и вычитании приближенных чисел, имеющих одинаковое число верных цифр после запятой, округление не производится. При сложении и вычитании приближенных чисел с различным числом верных цифр после запятой результат округляется по минимальному числу верных цифр после запятой у исходных чисел.

Замечания. 1. На практике при ручных вычислениях с целью уменьшения влияния погрешностей округления у приближенных чисел, кроме верных значащих цифр, обычно оставляют еще одну или две запасные цифры и действуют согласно сформулированным выше правилам с учетом запасных цифр. Эти запасные цифры отбрасываются при округлении окончательного результата.

2. При вычитании близких по величине чисел может произойти значительная потеря значащих цифр и, следовательно, точности результата.

Например, пусть требуется вычислить величину $\sqrt{543}-\sqrt{540}$, где числа 543 и 540 точные. Имеем $\sqrt{543}=23{,}30\ldots$, $\sqrt{540}=23{,}23\ldots$ Округлив эти числа с тремя значащими цифрами, приходим к результату только с одной значащей цифрой: $\sqrt{543}-\sqrt{540}\approx 23{,}3-23{,}2=0{,}1$.

Избавимся теперь от вычитания близких приближенных чисел:

$$\sqrt{543}-\sqrt{540}=\frac{3}{\sqrt{543}+\sqrt{540}}\approx\frac{3}{23{,}3+23{,}2}=\frac{3}{46{,}5}=0{,}06451\ldots$$

Округлив полученное число с тремя значащими цифрами, получим $\sqrt{543}-\sqrt{540}\approx 0{,}0645$.

С помощью более точных вычислений можно убедиться, что в последнем результате все три значащие цифры верные, хотя, как и в предыдущем случае, приближенные значения корней $\sqrt{543}$ и $\sqrt{540}$ использовались с тремя значащими цифрами.

Этот пример показывает, что если возможно, то *следует избегать вычитания близких приближенных чисел.* А если избежать этого невозможно, то следует увеличить точность промежуточных вычислений с учетом потери значащих цифр при вычитании.

§ 2. Вычисление значений многочлена. Схема Горнера

Рассмотрим алгебраический многочлен

$$P_n(x)=a_0+a_1x+a_2x^2+\ldots+a_nx^n, \qquad (1)$$

где a_0, a_1, ..., a_n — числовые коэффициенты, n — степень многочлена. Чтобы вычислить значение многочлена (1) при фиксированном $x=a$, можно поступить по-разному.

Например, можно сначала с помощью $n-1$ умножений найти степени a, т. е. a, a^2, ..., a^n. Затем в соответствии с формулой (1), где $x=a$, выполнив еще n умножений и n сложений, получить $P_n(a)$. Этот наиболее естественный на первый взгляд способ требует в общем случае при $n\geqslant 1$ выполнения $3n-1$ арифметических действий.

Однако более экономно значения многочлена находятся, если его записать следующим образом:

$$P_n(x)=$$
$$=a_0+x(a_1+x(a_2+\ldots+x(a_{n-2}+x(a_{n-1}+xa_n))\ldots)). \qquad (2)$$

Согласно формуле (2) вычисление значения $P_n(a)$ сводится к последовательному нахождению следующих величин:

$$\begin{aligned} b_n&=a_n,\\ b_{n-1}&=a_{n-1}+ab_n,\\ b_{n-2}&=a_{n-2}+ab_{n-1},\\ &\ldots\ldots\ldots\\ b_1&=a_1+ab_2,\\ b_0&=a_0+ab_1=P_n(a). \end{aligned} \qquad (3)$$

Способ нахождения значения многочлена по формулам (3) (по формуле (2)) называется *схемой Горнера*, которая реализуется с помощью n умножений и n сложений, т. е. всего за $2n$ арифметических действий. Можно доказать, что в общем случае не существует способа вычисления алгебраического многочлена n-й степени менее чем за $2n$ арифметических действий. Схема Горнера удобна также для реализации на ЭВМ благодаря цикличности вычислений и необходимости сохранять кроме коэффициентов многочлена и значения аргумента только одной промежуточной величины, а именно b_i или ab_i при текущем $i = n, n-1, \ldots, 0$.

Ручные вычисления значения многочлена по схеме Горнера обычно сопровождаются следующей таблицей:

	a_n	a_{n-1}	a_{n-2}	$\ldots$	a_0	$\lvert a$
$+$		ab_n	ab_{n-1}	$\ldots$	ab_1	
	$b_n = a_n$	b_{n-1}	b_{n-2}	$\ldots$	$b_0 = P_n(a)$	

Пример. Вычислить при $x = -1{,}5$ значение многочлена

$$P_5(x) = 1 - 4x + 3x^2 - x^3 + 2x^4 - x^5.$$

Решение.

	-1	2	-1	3	-4	1	$\lvert -1{,}5$
$+$		1,5	$-5{,}25$	9,375	$-18{,}5625$	33,84375	
	-1	3,5	$-6{,}25$	12,375	$-22{,}5625$	$34{,}84375 =$	
						$= P_5(-1{,}5)$	

Если заданный многочлен есть четная функция, т. е. $n = 2k$ и $P_{2k}(x) = a_0 + a_2x^2 + \ldots + a_{2k}x^{2k}$, то его удобно вычислять по формуле

$$P_{2k}(x) =$$

$$= a_0 + x^2(a_2 + x^2(a_4 + \ldots + x^2(a_{2k-4} + x^2(a_{2k-2} + x^2 a_{2k}))\ldots)), \quad (4)$$

а если многочлен является нечетной функцией, т. е. $n = 2k+1$ и $P_{2k+1}(x) = a_1x + a_3x^3 + \ldots + a_{2k+1}x^{2k+1}$, то для вычислений его следует привести к виду

$$P_{2k+1}(x) =$$

$$= x(a_1 + x^2(a_3 + \ldots + x^2(a_{2k-3} + x^2(a_{2k-1} + x^2 a_{2k+1}))\ldots)). \quad (5)$$

В заключение отметим, что при вычислении на ЭВМ многочленов с очень большими коэффициентами по схеме Горнера (по формулам (2), (4) или (5)) может произойти значительная потеря точности за

счет вычитания больших округленных чисел. Избежать потери точности иногда удается применением рекуррентных формул. Соответствующий пример рассматривается в конце § 6.

§ 3. Многочлены Тейлора

Пусть задана $f(x) \in C_{n+1}[a, b]$ (см. п. 2 введения). Напомним, *многочленом Тейлора n-й степени* функции f в точке $x_0 \in [a, b]$ называется многочлен

$$Q_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k. \qquad (1)$$

Многочлен Тейлора (1) обладает тем свойством, что в точке $x = x_0$ все его производные до порядка n включительно совпадают с соответствующими производными функции f, т. е.

$$Q_n^{(k)}(x_0) = f^{(k)}(x_0), \quad k = 0, 1, \ldots, n,$$

в чем легко убедиться, дифференцируя $Q_n(x)$.

Благодаря этому свойству многочлен Тейлора достаточно хорошо приближает функцию $f(x)$ в окрестности точки x_0. Погрешность, возникающая при замене функции f ее многочленом Тейлора, выражается остаточным членом формулы Тейлора *), т. е.

$$f(x) - Q_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}, \qquad (2)$$

где $x \in [a, b]$, ξ — некоторая точка, лежащая строго между x и x_0 при $x \neq x_0$. Данный остаточный член записан в форме Лагранжа, являющейся удобной для получения оценок погрешности. Поскольку производная $f^{(n+1)}$ по предположению непрерывна на отрезке $[a, b]$, то она ограничена на этом отрезке, т. е.

$$M_{n+1} = \max_{[a, b]} |f^{(n+1)}(x)| < \infty. \qquad (3)$$

На основании (2) имеем

$$|f(x) - Q_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}|x - x_0|^{n+1} \qquad (4)$$

*) См. § 4.14 в книге: Бугров Я. С., Никольский С. М. Дифференциальное и интегральное исчисление. — М.: Наука, 1984.

и

$$\max_{[a,\,b]} |f(x) - Q_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}\, l^{n+1}, \qquad (5)$$

где $l = \max\{x_0 - a,\ b - x_0\}$.

Неравенство (4) свидетельствует о том, что погрешность приближения функции $f(x)$ многочленом Тейлора (1) есть $O(|x - x_0|^{n+1})$*), а неравенство (5) служит оценкой максимальной погрешности на всем отрезке $[a, b]$.

Для погрешности аппроксимации функции многочленом Тейлора характерно то, что она достаточно быстро убывает при приближении x к x_0 и резко возрастает у конца отрезка $[a, b]$, который наиболее удален от точки x_0. Во всяком случае это так, когда $f^{(n+1)}(x) = \text{const} \neq 0$, т. е. $f(x)$ есть многочлен степени $n + 1$. При этом $|f^{(n+1)}(x)| = M_{n+1}$, $x \in [a, b]$, и неравенства (4), (5) обращаются в равенства.

Существенно неравномерная на отрезке $[a, b]$ точность аппроксимации функции f является недостатком многочлена Тейлора. Другой недостаток состоит в том, что для построения многочлена Тейлора требуется находить у функции f производные высоких порядков. Тем не менее многочлены Тейлора, в частности отрезки рядов Тейлора, тоже являющиеся многочленами Тейлора, широко используются на практике для аппроксимации функций, у которых достаточно просто вычисляются старшие производные, а остаточный член (правая часть неравенства (5)) стремится к нулю при $n \to \infty$. Сюда прежде всего относятся элементарные функции $\sin x$, $\cos x$, e^x, $\ln(1 + x)$ и др.

Пример. Аппроксимировать функцию $f(x) = e^x$ многочленом Тейлора на отрезке $[0, 1]$ с абсолютной погрешностью, не превышающей $\varepsilon = 10^{-5}$.

Решение. Выбираем $x_0 = 1/2$, т. е. в середине отрезка $[0, 1]$, с тем чтобы минимизировать величину l, входящую в правую часть оценки (5). Тогда $f^{(k)}(x) = e^x$,

$$f^{(k)}(x_0) = e^{1/2}, \qquad M_{n+1} = e, \qquad l = \frac{1}{2},$$

$$Q_n(x) = e^{1/2} \sum_{k=0}^{n} \frac{1}{k!} \left(x - \frac{1}{2}\right)^k.$$

*) Определение $O(|x - x_0|^{n+1})$ см. в п. 1 введения, положив $|x - x_0| = h$.

Согласно (5) имеем

$$\max_{[0,1]} |e^x - Q_n(x)| \leqslant r_n = \frac{e}{(n+1)!\,2^{n+1}}.$$

Для r_n составляем таблицу

n	2	3	4	5	6
r_n	$5{,}7 \cdot 10^{-2}$	$7{,}1 \cdot 10^{-3}$	$7{,}1 \cdot 10^{-4}$	$5{,}9 \cdot 10^{-5}$	$4{,}3 \cdot 10^{-6}$

Таким образом, следует взять $n = 6$.

§ 4. Интерполяционный многочлен Лагранжа

Пусть известны значения некоторой функции f в $n+1$ различных точках $x_0, x_1, \ldots, x_n$, которые обозначим следующим образом:

$$f_i = f(x_i), \quad i = 0, 1, \ldots, n.$$

Например, эти значения получены из эксперимента или найдены с помощью достаточно сложных вычислений.

Возникает задача приближенного восстановления функции f в произвольной точке x. Часто для решения этой задачи строится алгебраический многочлен $L_n(x)$ степени n, который в точках x_i принимает заданные значения, т. е.

$$L_n(x_i) = f_i, \quad i = 0, 1, \ldots, n, \tag{1}$$

и называется *интерполяционным*. Точки x_i, $i=0, 1, \ldots, n$, называются *узлами интерполяции*.

Для удобства изложения под многочленом степени n мы будем подразумевать многочлен степени не выше n. Например, если $f_i = 0$, $i = 0, 1, \ldots, n$, то интерполяционный многочлен $L_n(x) \equiv 0$ фактически имеет нулевую степень, но его тоже будем называть интерполяционным многочленом n-й степени.

Приближенное восстановление функции f по формуле

$$f(x) \approx L_n(x) \tag{2}$$

называется *интерполяцией* функции f (с помощью алгебраического многочлена). Если x расположен вне минимального отрезка, содержащего все узлы

интерполяции $x_0, x_1, \ldots, x_n$, то замену функции f по формуле (2) называют также *экстраполяцией*.

Сначала выясним вопрос существования и единственности интерполяционного многочлена, а затем исследуем погрешность интерполяции, т. е. какова разность между левой и правой частями приближенного равенства (2).

Теорема 1. *Существует единственный интерполяционный многочлен n-й степени, удовлетворяющий условиям* (1).

Доказательство. Существование интерполяционного многочлена установим непосредственно, выписав его. Пусть $n = 1$, тогда

$$L_1(x) = \frac{x - x_1}{x_0 - x_1} f_0 + \frac{x - x_0}{x_1 - x_0} f_1. \tag{3}$$

При $n = 2$

$$L_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} f_0 + \\ + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} f_2 \tag{4}$$

и, наконец, в общем случае при любом натуральном n

$$L_n(x) = \sum_{i=0}^{n} p_{ni}(x) f_i, \tag{5}$$

где

$$p_{ni}(x) = \frac{(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)}{(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)}, \tag{6}$$

$$i = 0, 1, \ldots, n.$$

Действительно, выражение (3) представляет собой линейную функцию, т. е. многочлен первой степени, причем $L_1(x_0) = f_0$, $L_1(x_1) = f_1$. Таким образом, требования (1) при $n = 1$ выполнены. Аналогично, формула (4) задает некоторый многочлен $L_2(x)$ второй степени, удовлетворяющий при $n = 2$ условиям (1). При произвольном натуральном n функции (6), выражающиеся в виде дроби, в числителе которой стоит произведение n линейных множителей, а в знаменателе — некоторое отличное от нуля число, являются алгебраическими многочленами степени n. Следовательно, функция (5) тоже является алгебраическим многочленом степени n, причем, поскольку $p_{ni}(x_i) = 1$,

а $p_{ni}(x_j)=0$ при $j\neq i$, $0\leqslant j\leqslant n$, выполнены требования (1).

Остается доказать единственность интерполяционного многочлена. Допустим, что кроме интерполяционного многочлена (5) имеется еще некоторый алгебраический многочлен $\tilde{L}_n(x)$ n-й степени, удовлетворяющий условиям

$$\tilde{L}_n(x_i)=f_i, \quad i=0,\ 1,\ \ldots,\ n. \tag{7}$$

Тогда согласно (1), (7)

$$\tilde{L}_n(x_i)-L_n(x_i)=0, \quad i=0,\ 1,\ \ldots,\ n. \tag{8}$$

Если $\tilde{L}_n(x)-L_n(x)\not\equiv 0$, то эта разность, будучи алгебраическим многочленом не выше n-й степени, в силу основной теоремы высшей алгебры имеет не более n корней, что противоречит равенствам (8), число которых равно $n+1$. Следовательно, $\tilde{L}_n(x)\equiv L_n(x)$. Теорема полностью доказана.

Интерполяционный многочлен, представленный в виде (5), называется *интерполяционным многочленом Лагранжа*, а функции (многочлены) (6) — *лагранжевыми коэффициентами*.

Имеются и другие формы записи интерполяционного многочлена. Однако по теореме 1 интерполяционный многочлен n-й степени (точнее говоря, степени не выше n), удовлетворяющий условиям (1), единствен. В § 9 будет дан другой способ задания интерполяционного многочлена.

Замечания. 1. Фактическую степень интерполяционного многочлена (5) можно выяснить после раскрытия скобок и приведения подобных членов. Однако если в узлах x_i в качестве f_i берутся значения некоторого алгебраического многочлена $P_k(x)$ степени $k\leqslant n$, то по теореме 1 заведомо $L_n(x)\equiv P_k(x)$, так как $P_k(x)$ есть тоже многочлен степени не выше n, удовлетворяющий условиям (1). В частности, если $f_i=1$, $i=0,\ 1,\ \ldots,\ n$, то $L_n(x)\equiv P_0(x)\equiv 1$. Отсюда и из (5) вытекает тождество $\sum\limits_{i=0}^{n} p_{in}(x)\equiv 1$, которое может служить контролем при вычислении лагранжевых коэффициентов (6).

2. Поскольку интерполяционный многочлен (5) линейно зависит от значений функции f_i, то

интерполяционный многочлен для суммы двух функций равен сумме интерполяционных многочленов для слагаемых.

Пример. Построить интерполяционный многочлен Лагранжа по следующим данным:

i	0	1	2	3
x_i	0	2	3	5
f_i	1	3	2	5

Решение. Согласно (5) при $n = 3$ имеем

$$L_3(x) = \frac{(x-2)(x-3)(x-5)}{(0-2)(0-3)(0-5)} \cdot 1 + \frac{x(x-3)(x-5)}{2(2-3)(2-5)} \cdot 3 + \\ + \frac{x(x-2)(x-5)}{3(3-2)(3-5)} \cdot 2 + \frac{x(x-2)(x-3)}{5(5-2)(5-3)} \cdot 5 = \\ = 1 + \frac{62}{15}x - \frac{13}{6}x^2 + \frac{3}{10}x^3.$$

Погрешность интерполяции. Всегда можно написать равенство

$$f(x) = L_n(x) + R_n(x), \tag{9}$$

где $R_n(x)$ — остаточный член, т. е. погрешность интерполяции. Если относительно функции f ничего не известно, кроме ее значений f_i в узлах интерполяции, то никаких полезных суждений относительно остаточного члена $R_n(x)$ сделать нельзя. Мы получим некоторое выражение остаточного члена в предположении, что $f \in C_{n+1}[a, b]$, где $[a, b]$ — отрезок, содержащий все узлы интерполяции x_i, $i = 0, 1, \ldots, n$, и точку x.

Ищем $R_n(x)$ в следующем виде:

$$R_n(x) = \omega_n(x) r_n(x), \tag{10}$$

где

$$\omega_n(x) = (x - x_0)(x - x_1) \ldots (x - x_n), \tag{11}$$

$r_n(x)$ — некоторая функция, значения которой в узлах интерполяции x_i можно задать какие угодно, ибо $R_n(x_i) = 0$ и $\omega_n(x_i) = 0$, $i = 0, 1, \ldots, n$.

Зафиксируем произвольное $x \in [a, b]$, $x \neq x_i$, $i = 0, 1, \ldots, n$, и рассмотрим следующую функцию от t:

$$\varphi(t) = L_n(t) + \omega_n(t) r_n(x) - f(t). \tag{12}$$

Эта функция на основании (1), (9)—(11) обращается в нуль при $t = x_i$, $i = 0, 1, \ldots, n$, $t = x$, т. е. во всяком случае в $n + 2$-х точках отрезка $[a, b]$, на котором изменяется t.

По теореме Ролля φ' (штрих по t) обращается в нуль по крайней мере в $n + 1$-й точке интервала (a, b), φ'' равна нулю минимум в n точках этого интервала и т. д. Таким образом, найдется хотя бы одна точка $\xi \in (a, b)$, в которой $\varphi^{(n+1)}(\xi) = 0$ Отсюда и из (12), учитывая, что $L_n^{(n+1)}(\xi) = 0$, $\omega_n^{(n+1)}(\xi) = (n+1)!$, получаем

$$(n+1)!\, r_n(x) - f^{(n+1)}(\xi) = 0.$$

Следовательно,

$$r_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

и, в соответствии с (9), (10),

$$R_n(x) = \omega_n(x) \frac{f^{(n+1)}(\xi)}{(n+1)!}, \tag{13}$$

$$f(x) = L_n(x) + \omega_n(x) \frac{f^{(n+1)}(\xi)}{(n+1)!}, \tag{14}$$

$\xi = \xi(x) \in (a, b)$ — некоторая неизвестная точка.

Из равенства (14) вытекает оценка погрешности интерполяции (в частности, экстраполяции) в текущей точке $x \in [a, b]$:

$$|f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} |\omega_n(x)| \tag{15}$$

и оценка максимальной погрешности интерполяции на всем отрезке $[a, b]$:

$$\max_{[a,\,b]} |f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} \max_{[a,\,b]} |\omega_n(x)|, \tag{16}$$

где M_{n+1} — величина (3.3).

Пример. Оценить погрешность приближения функции $f(x) = \sqrt{x}$ в точке $x = 116$ и на всем отрезке $[a, b]$, где $a = 100$, $b = 144$, с помощью интерполяционного многочлена Лагранжа $L_2(x)$ второй степени, построенного с узлами $x_0 = 100$, $x_1 = 121$, $x_2 = 144$.

Решение.

$$f'(x) = \frac{1}{2} x^{-1/2}, \quad f''(x) = -\frac{1}{4} x^{-3/2}, \quad f'''(x) = \frac{3}{8} x^{-5/2},$$

$$M_3 = \max_{[a,\,b]} |f'''(x)| = \frac{3}{8} 100^{-5/2} = \frac{3}{8} 10^{-5}.$$

На основании неравенства (15) получаем

$$|\sqrt{116} - L_2(116)| \leqslant$$

$$\leqslant \frac{3}{8}\,10^{-5}\,\frac{1}{3!}\,|(116-100)(116-121)(116-144)| =$$

$$= \frac{1}{16}\,10^{-5}\cdot 16\cdot 5\cdot 28 = 1{,}4\cdot 10^{-3}.$$

В силу оценки (16)

$$\max_{[a,\,b]} |\sqrt{x} - L_2(x)| \leqslant$$

$$\leqslant \frac{10^{-5}}{16} \max_{[a,\,b]} |(x-100)(x-121)(x-144)| \approx 2{,}5\cdot 10^{-3}.$$

§ 5. Линейная интерполяция

Интерполяция по формуле (4.2) при $n=1$, т. е. с помощью линейной функции (4.3), называется *линейной*. Если ввести обозначения $h = x_1 - x_0$, $q = (x-x_0)/h$, то формула линейной интерполяции может быть записана в следующем виде:

$$f(x) \approx L_1(x) = L_1(x_0 + qh) = (1-q)f_0 + qf_1. \quad (1)$$

Величина q называется *фазой* интерполяции, которая изменяется в пределах от 0 до 1, когда x пробегает значения от x_0 до x_1.

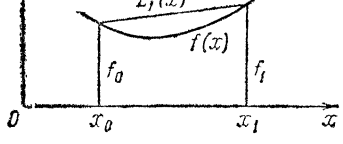


Рис. 1

Геометрически линейная интерполяция означает (рис. 1) замену графика функции на отрезке $[x_0, x_1]$ хордой, соединяющей точки (x_0, f_0), (x_1, f_1).

Поскольку согласно (4.11) $\omega_2(x) = (x-x_0)(x-x_1)$ и, следовательно,

$$\max_{[x_0,\,x_1]} |\omega_2(x)| = \max_{[x_0,\,x_1]} |(x-x_0)(x-x_1)| = \frac{h^2}{4},$$

то оценка максимальной погрешности линейной интерполяции на отрезке $[x_0, x_1]$ в соответствии с (4.16) имеет вид

$$\max_{[x_0,\,x_1]} |f(x) - L_1(x)| \leqslant h^2\,\frac{M_2}{8}, \quad (2)$$

где $M_2 = \max\limits_{[x_0,\,x_1]} |f''(x)|$.

Часто задают таблицу большого числа значений некоторой функции f с постоянным шагом h изменения аргумента. Тогда при заданном x выбираются два ближайших к нему узла. Левый узел принимается за x_0, а правый — за x_1, и осуществляется линейная интерполяция по формуле (1). Погрешность интерполяции оценивается по формуле (2).

Пример. Задана таблица функции $\sin x$ с шагом в 1°. Требуется оценить погрешность линейной интерполяции.

Решение. Шаг h таблицы в радианной мере составляет $\pi/180$. Поскольку $M_2 = \max|(\sin x)''| \leqslant 1$, то погрешность линейной интерполяции согласно (2) не превышает

$$\rho_1 = \frac{\pi^2}{180^2} \cdot \frac{1}{8} < 0{,}4 \cdot 10^{-4}.$$

Величина ρ_1 служит оценкой погрешности, возникающей за счет замены на отрезке длины h функции $\sin x$ интерполяционным многочленом $L_1(x)$, в предположении, что табличные значения заданы точно и вычисления по формуле (1) осуществляются без погрешностей.

Допустим теперь, что табличные значения функции $\sin x$ округлены с четырьмя десятичными знаками после запятой. Тогда предельная абсолютная погрешность округленных табличных значений есть $\overline{\Delta}(f) = 0{,}5 \cdot 10^{-4}$. При вычислениях по формуле (1), в которую вместо точных значений f_0, f_1 функции $f(x) = \sin x$ подставлены округленные значения, в значении $L_1(x)$ возникает погрешность, оцениваемая величиной

$$\overline{\Delta}(L_1) = (1 - q)\,\overline{\Delta}(f) + q\overline{\Delta}(f) = \overline{\Delta}(f) = 0{,}5 \cdot 10^{-4},$$

так как $0 \leqslant q \leqslant 1$.

Если найденное таким образом значение интерполяционного многочлена $L_1(x)$ окончательно округлить с четырьмя десятичными знаками после запятой, то возникает еще дополнительная погрешность, ограниченная по модулю величиной $\overline{\Delta}_1(L_1) = 0{,}5 \cdot 10^{-4}$.

Итак, полная погрешность линейной интерполяции функции $\sin x$ при округленных табличных значениях и округлении окончательного результата с четырьмя десятичными знаками после запятой оценивается величиной

$$\rho_1 + \overline{\Delta}(L_1) + \overline{\Delta}_1(L_1) < 1{,}4 \cdot 10^{-4}.$$

§ 6. Минимизация оценки погрешности интерполяции. Многочлены Чебышева

Пусть задана некоторая функция $f \in C_{n+1}[a, b]$. Возникает вопрос, как выбрать на отрезке $[a, b]$ узлы $x_0, x_1, \ldots, x_n$ интерполяционного многочлена (4.5), чтобы максимальная погрешность интерполяции функции f на этом отрезке была минимальной. Эта задача является сложной и ее удается решить только для

частных функций f. Мы ограничимся решением более простой задачи, а именно нахождением такого расположения узлов интерполяции x_i, $i = 0, 1, \ldots, n$, на отрезке $[a, b]$, при котором минимальна величина $\max_{[a, b]} |\omega_n(x)|$ и тем самым минимальна правая часть оценки погрешности (4.16).

Для простоты сначала рассмотрим случай стандартного отрезка $[-1, 1]$. При этом нам потребуются многочлены Чебышева.

Многочлены Чебышева. *Многочлены Чебышева* $T_n(x)$, $n \geqslant 0$, на отрезке $[-1, 1]$ задаются формулой

$$T_n(x) = \cos(n \arccos x). \tag{1}$$

В частности, при $n = 0, 1$ имеем

$$T_0(x) = \cos(0 \cdot \arccos x) = 1, \tag{2}$$

$$T_1(x) = \cos(\arccos x) = x. \tag{3}$$

Далее, из тождества

$$\cos(n+1)\varphi = 2\cos\varphi\cos n\varphi - \cos(n-1)\varphi,$$

полагая $\varphi = \arccos x$, в соответствии с (1) получаем

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \tag{4}$$

где $n = 1, 2, \ldots$ Таким образом, $T_n(x)$ действительно является алгебраическим многочленом степени $n \geqslant 0$.

Полагая $T_0(x) = 1$, $T_1(x) = x$ на всей оси x и распространяя рекуррентную формулу (4) на всю ось x, последовательно по формуле (4) находим

$$T_2(x) = 2x^2 - 1, \qquad T_3(x) = 4x^3 - 3x,$$

$$T_4(x) = 8x^4 - 8x^2 + 1, \quad T_5(x) = 16x^5 - 20x^3 + 5x, \ldots$$

Свойства многочленов Чебышева.

1. *При четном (нечетном) n многочлен $T_n(x)$ содержит только четные (нечетные) степени x, т. е. является четной (нечетной) функцией.*

Это свойство легко вытекает из формул (2)—(4).

2. *Старший коэффициент многочлена $T_n(x)$ при $n \geqslant 1$ равен 2^{n-1}.*

Данное свойство тоже следует из формул (2)—(4).

3. *$T_n(x)$ имеет n действительных корней в интервале $(-1, 1)$, выражаемых формулой*

$$x_i = \cos\frac{(2i+1)\pi}{2n}, \quad i = 0, 1, \ldots, n-1.$$

В самом деле,

$$T_n(x_i) = \cos(n \arccos x_i) = \cos\frac{(2i+1)\pi}{2} = 0,$$

$$i = 0, 1, \ldots, n-1.$$

4. $\max\limits_{[-1,\,1]} |T_n(x)| = 1$, *причем*

$$T_n(x_m) = (-1)^m, \tag{5}$$

где $x_m = \cos(m\pi/n)$, $m = 0, 1, \ldots, n$.

Действительно, согласно (1)

$$T_n(x_m) = \cos m\pi = (-1)^m, \quad |T_n(x)| \leqslant 1, \quad x \in [-1, 1].$$

5. *Многочлен*

$$\bar{T}_n(x) = 2^{1-n}T_n(x), \quad n \geqslant 1, \tag{6}$$

среди всех многочленов n-й степени со старшим коэффициентом, равным единице, имеет на отрезке $[-1, 1]$ наименьшее значение максимума модуля, т. е. не существует такого многочлена $\bar{P}_n(x)$ n-й степени со старшим коэффициентом, равным единице, что

$$\max_{[-1,\,1]} |\bar{P}_n(x)| < \max_{[-1,\,1]} |\bar{T}_n(x)| = 2^{1-n}. \tag{7}$$

Допустим противное: имеется многочлен $\bar{P}_n(x) = a_0 + a_1x + \ldots + a_{n-1}x^{n-1} + x^n$, удовлетворяющий неравенству (7). Тогда, так как по свойству 2 у многочлена $\bar{T}_n(x)$ старший коэффициент тоже равен единице, разность $\bar{T}_n(x) - \bar{P}_n(x)$ является алгебраическим многочленом степени не выше $n-1$, причем в силу (7) $\bar{T}_n(x) - \bar{P}_n(x) \not\equiv 0$. Кроме того, в $n+1$ точках $x_m = \cos(m\pi/n)$, $m = 0, 1, \ldots, n$, на основании (5)—(7) эта разность принимает отличные от нуля значения с чередующимися знаками. Это означает, что алгебраический многочлен $\bar{T}_n(x) - \bar{P}_n(x)$ степени меньше чем n обращается в нуль по крайней мере в n точках, что невозможно.

Замечание. Можно доказать, что если $\bar{P}_n(x) = a_0 + a_1x + \ldots + a_{n-1}x^{n-1} + x^n$, причем $n \geqslant 1$, $\max\limits_{[-1,\,1]} |\bar{P}_n(x)| = 2^{1-n}$, то $\bar{P}_n(x) = 2^{1-n}T_n(x) = \bar{T}_n(x)$.

Благодаря свойству 5 многочлены Чебышева $T_n(x)$ называются *многочленами, наименее уклоняющимися от нуля.*

Узлы, минимизирующие оценку погрешности интерполяции. Возьмем на отрезке $[-1, 1]$ в качестве узлов интерполяции корни многочлена Чебышева $T_{n+1}(x)$, т. е. точки

$$x_i = \cos\frac{(2i+1)\pi}{2n+2}, \quad i = 0, 1, \ldots, n. \tag{8}$$

Тогда многочлен (4.11), у которого старший коэффициент равен единице, будет пропорционален многочлену $T_{n+1}(x)$ и в силу свойства 2 многочленов Чебышева выразится через $T_{n+1}(x)$ следующим образом:

$$\omega_n(x) = 2^{-n}T_{n+1}(x).$$

При этом в соответствии со свойством 4 оценка погрешности интерполяции (4.16) примет вид

$$\max_{[-1,\,1]} |f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!\,2^n}, \tag{9}$$

где $M_{n+1} = \max\limits_{[-1,\,1]} |f^{(n+1)}(x)|$.

В силу свойства 5 многочленов Чебышева оценку (4.16) улучшить на отрезке $[-1, 1]$ по сравнению с оценкой (9) за счет другого выбора узлов интерполяции нельзя. Более того, согласно свойству 5 и замечанию при любом выборе узлов, не совпадающем с (8), соответствующая оценка максимальной погрешности интерполяции на отрезке $[-1, 1]$ будет хуже, т. е. *узлы интерполяции* (8) *являются оптимальными для оценки погрешности* (4.16) *на отрезке* $[-1, 1]$.

В случае интерполирования на произвольном отрезке $[a, b]$ линейной заменой независимой переменной

$$x = \frac{1}{2}((b-a)t + b + a), \quad t = \frac{1}{b-a}(2x - b - a), \tag{10}$$

он переводится в отрезок $[-1, 1]$. При этом корням многочлена Чебышева $T_{n+1}(t)$ отвечают точки

$$x_i = \frac{1}{2}\left((b-a)\cos\frac{(2i+1)\pi}{2n+2} + b + a\right), \quad i = 0, 1, \ldots, n, \tag{11}$$

отрезка $[a, b]$, являющиеся оптимальными узлами для оценки погрешности интерполяции на этом отрезке.

Согласно (8), (10), (11) имеем

$$\omega_n(x) = (x - x_0)(x - x_1) \dots (x - x_n) =$$

$$= \frac{(b-a)^{n+1}}{2^{n+1}} \left(t - \cos\frac{\pi}{2n+2}\right)\left(t - \cos\frac{3\pi}{2n+2}\right) \dots$$

$$\dots \left(t - \cos\frac{(2n+1)\pi}{2n+2}\right) = \frac{(b-a)^{n+1}}{2^{n+1}} \overline{T}_{n+1}(t).$$

Отсюда с учетом (6) и свойства 4 многочленов Чебышева получаем

$$\max_{[a,\,b]} |\omega_n(x)| = \frac{(b-a)^{n+1}}{2^{n+1}} \max_{[-1,\,1]} |\overline{T}_{n+1}(t)| = \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

Следовательно, при узлах (11) оценка погрешности интерполяции (4.16) приобретает вид

$$\max_{[a,\,b]} |f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} \frac{(b-a)^{n+1}}{2^{2n+1}}, \qquad (12)$$

где $M_{n+1} = \max\limits_{[a,\,b]} |f^{(n+1)}(x)|$.

Сравним теперь способы аппроксимации функции $f \in C_{n+1}[a, b]$ многочленом Тейлора $Q_n(x)$ и интерполяционным многочленом $L_n(x)$ с узлами (11). При построении многочлена Тейлора (3.1) целесообразно взять точку x_0 в середине отрезка $[a, b]$, т. е. $x_0 = (a + b)/2$. Тогда в соответствии с (3.5) оценка максимальной погрешности многочлена Тейлора будет следующей:

$$\max_{[a,\,b]} |f(x) - Q_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} \frac{(b-a)^{n+1}}{2^{n+1}}. \qquad (13)$$

Таким образом, оценка погрешности многочлена Тейлора в 2^n раз больше оценки погрешности (12) интерполяционного многочлена Лагранжа $L_n(x)$ с оптимальными узлами (11). Если производная $f^{(n+1)}(x)$ относительно мало изменяется на отрезке $[a, b]$, то обе оценки (12) и (13) мало завышают максимальную погрешность, а если $f^{(n+1)}(x) = \text{const}$, то каждое из неравенств (12) и (13) обращается в равенство.

Погрешность интерполяционного многочлена более равномерно распределена на отрезке $[a, b]$, чем у

многочлена Тейлора, обладающего существенно неравномерной погрешностью (см. § 3). Поэтому не только оценка погрешности, но и обычно фактическая максимальная погрешность интерполяционного многочлена на всем отрезке $[a, b]$ меньше, чем у многочлена Тейлора. Кроме того, для построения интерполяционного многочлена не требуется вычислять производные функции f, а нужны только ее значения.

Итак, следует отдать предпочтение интерполяционному многочлену с оптимальными узлами (11).

Пример. Пусть на отрезке $[-1, 1]$ задана функция $f(x) = e^x$. Ее многочлен Тейлора $Q_5(x)$ в точке $x_0 = 0$ имеет вид $Q_5(x) = \sum_{k=0}^{5} \frac{x^k}{k!}$. Поскольку $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, то

$$\max_{[-1, 1]} |e^x - Q_5(x)| = \max_{[-1, 1]} \left| \sum_{k=6}^{\infty} \frac{x^k}{k!} \right| > \frac{1}{6!} + \frac{1}{7!} > 0{,}15 \cdot 10^{-2}.$$

Что касается интерполяционного многочлена Лагранжа $L_5(x)$ с узлами (8) при $n = 5$, то, так как $M_6 = \max_{[-1, 1]} \left| \frac{d^6}{dx^6} e^x \right| = e$, для него согласно (9) выполняется неравенство

$$\max_{[-1, 1]} |e^x - L_5(x)| \leqslant \frac{e}{6!\, 2^5} < 0{,}12 \cdot 10^{-3}.$$

Таким образом, в рассмотренном примере оценка сверху погрешности интерполяционного многочлена $L_5(x)$ значительно меньше, чем оценка снизу максимальной погрешности многочлена Тейлора $Q_5(x)$. Данный пример подтверждает сделанный выше вывод.

В заключение остановимся на вопросе вычисления значений многочленов Чебышева. Согласно свойствам 2, 4, с одной стороны, значения многочленов $T_n(x)$ при любом n ограничены на отрезке $[-1, 1]$ по модулю единицей. С другой стороны, наблюдается достаточно быстрый рост коэффициентов этих многочленов с ростом n. Например, старший коэффициент у многочлена $T_n(x)$ равен 2^{n-1}, в частности, при $n = 21$ имеем $2^{n-1} > 10^6$. Это приводит к тому, что вычисление значений $T_n(x)$ для $x \in [-1, 1]$ по схеме Горнера (2.2) на реальной ЭВМ, обладающей ограниченным числом цифровых разрядов, происходит при

больших n со значительной потерей точности. Потеря точности вызывается необходимостью вычитания близких больших по модулю округленных чисел, ибо окончательный результат по модулю не превосходит единицу.

Наличие рекуррентной формулы (4) позволяет обойти указанную трудность. При заданном $x \in [-1, 1]$ находятся сначала $T_0(x)$, $T_1(x)$ непосредственно по формулам (2), (3), а затем по формуле (4) вычисляются последовательно значения $T_n(x)$ для возрастающих n, включая заданное n.

Можно доказать, что при этом суммарная погрешность округления в значении $T_n(x)$ увеличивается по сравнению с максимальной погрешностью округления одноразового вычисления по формуле (4) не более чем в $O(n^2)$ раз. Во всяком случае в вычислениях по формуле (4) при $x \in [-1, 1]$ большие величины не участвуют, так как $|T_n(x)| \leqslant 1$, $|T_{n-1}(x)| \leqslant 1$ для любого натурального n.

В вычислениях же по схеме Горнера многочленов $T_n(x)$ влияние погрешностей округления с ростом n значительно сильнее, чем по формуле (4).

§ 7. Интерполяция с равноотстоящими узлами

В § 6 было найдено оптимальное распределение узлов интерполяции, обеспечивающее минимальную оценку погрешности на всем отрезке. Оптимальное распределение узлов является неравномерным. Узлы (см. (6.8), (6.11)) сгущаются к концам отрезка и разрежаются в его середине. Это вызывает неудобства на практике.

Неравномерное распределение узлов можно задать, скажем, если интерполяционный многочлен строится один раз, причем для всего заданного отрезка. Если же отрезок $[a, b]$ большой и требуется высокая точность аппроксимации функции, то одним интерполяционным многочленом приемлемой степени даже с оптимальным распределением узлов обычно не удается обеспечить заданную точность интерполяции на всем отрезке. В таком случае часто пользуются таблицей значений функции в узлах, расположенных с постоянным шагом, число которых может быть достаточно большим.

Когда задается значение аргумента x, то выбирается несколько ближайших к x узлов, используемых для построения интерполяционного многочлена обычно невысокой степени, и производится интерполяция. Ниже выясняется зависимость точности интерполяции от шага, с которым расположены узлы.

Итак, пусть $x_i = x_0 + ih$, $i = 0, 1, \ldots, n$, — узлы интерполяции, $h > 0$ — шаг, $f_i = f(x_i)$ — заданные значения функции $f \in C_{n+1}[a, b]$, причем $[x_0, x_n] \subset [a, b]$.

Введем безразмерную независимую переменную

$$q = \frac{x - x_0}{h} \qquad (x = x_0 + qh). \tag{1}$$

Тогда узлу x_i будет соответствовать

$$q = q_i = \frac{x_i - x_0}{h} = \frac{x_0 + ih - x_0}{h} = i \tag{2}$$

и, кроме того, будут выполняться соотношения

$$x - x_j = h(q - j), \quad x_i - x_j = h(i - j). \tag{3}$$

При этом интерполяционный многочлен (4.3), отвечающий случаю $n = 1$, запишется, как указано в формуле (5.1), а интерполяционный многочлен (4.4) второй степени приобретет вид

$$L_2(x) = L_2(x_0 + qh) =$$
$$= \frac{(q-1)(q-2)}{2} f_0 - q(q-2) f_1 + \frac{q(q-1)}{2} f_2.$$

В общем случае с учетом (1)—(3), (4.6) интерполяционный многочлен (4.5) примет следующий вид:

$$L_n(x) = L_n(x_0 + qh) = \sum_{i=0}^{n} \bar{p}_{ni}(q) f_i, \tag{4}$$

где

$$\bar{p}_{ni}(q) = (-1)^{n-i} \frac{q(q-1)\ldots(q-i+1)(q-i-1)\ldots(q-n)}{i!\,(n-i)!}. \tag{5}$$

Поскольку согласно (3)

$$\omega_n(x) = (x - x_0)(x - x_1) \ldots (x - x_n) = h^{n+1} \bar{\omega}_n(q),$$

где

$$\bar{\omega}_n(q) = q(q-1) \ldots (q-n), \tag{6}$$

то остаточный член (4.13) интерполяционного многочлена (4) может быть представлен в следующем виде:

$$R_n(x) = R_n(x_0 + qh) = h^{n+1}\bar{\omega}_n(q)\frac{f^{(n+1)}(\xi)}{(n+1)!}. \qquad (7)$$

Здесь $f^{(n+1)}$ — производная по x, а промежуточная точка $\xi = \xi(x)$ та же самая, что и в (4.13), (4.14).

Заметим, что согласно (1) изменению переменной x на отрезке $[x_0, x_n]$, где $x_n = x_0 + nh$, отвечает изменение переменной q на отрезке $[0, n]$. Поэтому оценку максимальной погрешности интерполяции на отрезке $[x_0, x_n]$ с учетом (4.9), (7) можно записать в следующем виде:

$$\max_{[x_0, x_n]} |f(x) - L_n(x)| \leqslant h^{n+1}\frac{M_{n+1}^h}{(n+1)!}\Omega_n, \qquad (8)$$

где

$$\Omega_n = \max_{[0, n]} |\bar{\omega}_n(q)|,$$
$$M_{n+1}^h = \max_{[x_0, x_n]} |f^{(n+1)}(x)|. \qquad (9)$$

Величина Ω_n не зависит от h. Ее можно заранее вычислить или оценить. В частности,

$$\Omega_1 = 1/4,\ \Omega_2 = 2/(3\sqrt{3}),\ \Omega_3 = 1,\ \Omega_4 < 3{,}7,\ \Omega_5 < 17. \qquad (10)$$

Величина M_{n+1}^h при малом h мало отличается от $|f^{(n+1)}(x_0)|$. Во всяком случае, если $[x_0, x_n] \subset [a, b]$, то

$$M_{n+1}^h \leqslant M_{n+1}, \qquad (11)$$

где $M_{n+1} = \max_{[a, b]} |f^{(n+1)}(x)|$.

Таким образом, согласно (8), (11) максимальная погрешность интерполяции на отрезке $[x_0, x_n]$, т. е. $\max_{[x_0, x_n]} |f(x) - L_n(x)|$, есть $O(h^{n+1})$. Отметим, например, что так как $M_{n+1}^{h/2} \leqslant M_{n+1}^h$, то при уменьшении шага h вдвое правая часть оценки (8) уменьшится по крайней мере в 2^{n+1} раз.

Исходя из усиленной оценки, получаемой из неравенства (8), в которое вместо M_{n+1}^h в соответствии с (11) подставлено M_{n+1}, выбирают шаг h таблицы значений функции f на отрезке $[a, b]$, с тем чтобы обеспечить заданную точность интерполяции. При

этом имеется еще возможность варьировать в некоторых пределах степень n интерполяционного многочлена. Если функция f достаточно гладкая, то повышение n вначале обычно ведет к увеличению допустимого h, но, с другой стороны, усложняет интерполяцию и усиливает влияние неустранимых погрешностей табличных значений. На практике редко используют интерполяцию с $n \geqslant 5$.

Замечание 1. При заданном x узлы интерполяции $x_0, x_1, \ldots, x_n$, расположенные с шагом h, целесообразно выбирать из совокупности всех узлов заданной таблицы функции так, чтобы точка x оказалась возможно ближе к середине отрезка $[x_0, x_n]$. Это связано с тем, что колебания функции (4.11), соответственно функции (6), вблизи середины указанного отрезка меньше, чем у его концов. На рис. 2 изображен график функции (4.11) при $n = 4$.

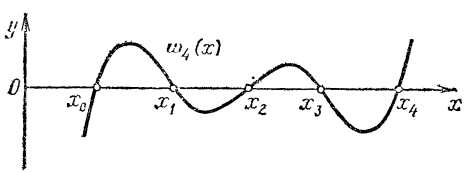


Рис. 2

Если точка x, в которой производится интерполяция, удовлетворяет неравенству

$$\left| x - \frac{x_0 + x_n}{2} \right| \leqslant \frac{h}{2}, \tag{12}$$

то для погрешности интерполяции справедлива оценка

$$|f(x) - L_n(x)| \leqslant h^{n+1} \frac{M_{n+1}^h}{(n+1)!} \Omega_n^0, \tag{13}$$

где $\Omega_n^0 = \max_{|q - n/2| \leqslant 1/2} |\bar{\omega}_n(q)|$. В частности,

$$\Omega_1^0 = \frac{1}{4}, \quad \Omega_2^0 = \frac{3}{8}, \quad \Omega_3^0 = \frac{9}{16}, \quad \Omega_4^0 = \frac{45}{32}, \quad \Omega_5^0 = \frac{225}{64}. \tag{14}$$

Сопоставляя величины (14) и (10), видим, что оценка погрешности интерполяции (13) является более точной, чем (8), для x, удовлетворяющих условию (12).

§ 8. Конечные и разделенные разности

Конечные разности. Пусть $x_k = x_0 + kh$, где k — целое, $h > 0$, $f_k = f(x_k)$. Величина

$$\Delta f_k = f(x_k + h) - f(x_k) = f(x_{k+1}) - f(x_k) = f_{k+1} - f_k \quad (1)$$

называется *конечной разностью первого порядка* функции f в точке x_k (с шагом h), а

$$\Delta^2 f_k = \Delta(\Delta f_k) = \Delta f_{k+1} - \Delta f_k = \\ = (f_{k+2} - f_{k+1}) - (f_{k+1} - f_k) = f_k - 2f_{k+1} + f_{k+2} \quad (2)$$

есть *конечная разность второго порядка* в точке x_k. Вообще, *конечная разность n-го порядка* функции f в точке x_k определяется по рекуррентной формуле

$$\Delta^n f_k = \Delta(\Delta^{n-1} f_k) = \Delta^{n-1} f_{k+1} - \Delta^{n-1} f_k, \quad (3)$$

где $n \geqslant 1$, $\Delta^0 f_k = f_k$.

При вычислениях конечные разности удобно записывать в виде табл. 1.

Таблица 1

x_0	f_0			
		Δf_0		
x_1	f_1		$\Delta^2 f_0$	
		Δf_1		$\Delta^3 f_0$
x_2	f_2		$\Delta^2 f_1$	
		Δf_2		$\Delta^3 f_1$
x_3	f_3		$\Delta^2 f_2$	
		Δf_3		
x_4	f_4			

Лемма 1. *Если* $f \in C_n[x_k, x_{k+n}]$, *то существует такая точка* $\eta \in (x_k, x_{k+n})$, *что*

$$\Delta^n f_k = h^n f^{(n)}(\eta). \quad (4)$$

В силу (1) формула (4) при $n = 1$ совпадает с известной формулой конечных приращений Лагранжа. Докажем ее при $n = 2$. Пусть $\varphi(x) = f(x + h) - f(x)$. Тогда согласно (3)

$$\Delta^2 f_k = \varphi(x_k + h) - \varphi(x_k)$$

и, следовательно, по формуле конечных приращений Лагранжа

$$\Delta^2 f_k = h\varphi'(\eta_1), \quad (5)$$

где $\eta_1 \in (x_k, x_k + h)$ — некоторая точка. Но

$$\varphi'(\eta_1) = f'(\eta_1 + h) - f'(\eta_1).$$

Применяя еще раз формулу конечных приращений Лагранжа к функции f', получаем

$$\varphi'(\eta_1) = hf''(\eta), \tag{6}$$

где $\eta \in (\eta_1, \eta_1 + h) \subset (x_k, x_{k+2})$ — некоторая точка. Из (5), (6) вытекает утверждение леммы при $n = 2$. Для $n > 2$ лемма доказывается аналогично.

Следствие леммы 1. *Конечная разность n-го порядка алгебраического многочлена n-й степени постоянна, т. е. не зависит от k, а конечные разности более высоких порядков равны нулю.*

Остановимся на одном из практических применений конечных разностей. Согласно лемме 1, если $f \in C_{n+1}[x_0, x_{n+1}]$, то величина $\Delta^{n+1} f_0/h^{n+1}$, которую можно вычислить через табличные значения функции f с помощью формулы (3), равна значению производной $f^{(n+1)}$ в некоторой точке $\eta \in (x_0, x_{n+1})$, где $x_{n+1} = x_0 + (n+1)h$. Поэтому, если h мало, то число $|\Delta^{n+1} f_0/h^{n+1}|$ можно приближенно принять за величину (7.9) и использовать в оценке (7.8) погрешности интерполяции. Такой нестрогой оценкой погрешности пользуются, если достаточно сложно вычисляется производная $f^{(n+1)}(x)$ или, вообще, имеются в распоряжении только табличные значения $n+1$ раз дифференцируемой функции.

Разделенные разности. Пусть теперь x_0, x_1, ..., x_k, ... — произвольные точки (узлы) оси x, причем $x_i \neq x_j$ при $i \neq j$.

Значения $f(x_0)$, $f(x_1)$, ... функции f в узлах называются *разделенными разностями нулевого порядка.* Число

$$f(x_0; x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \tag{7}$$

называется *разделенной разностью первого порядка* функции f.

Очевидно,

$$f(x_0; x_1) = f(x_1; x_0) = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0}, \tag{8}$$

т. е. разделенная разность первого порядка является симметрической функцией аргументов x_0 и x_1.

Разделенная разность n-го порядка определяется через разделенные разности $n-1$-го порядка по рекуррентной формуле

$$f(x_0; x_1; \ldots; x_n) = \frac{f(x_1; x_2; \ldots; x_n) - f(x_0; x_1; \ldots; x_{n-1})}{x_n - x_0}. \quad (9)$$

При вычислениях разделенные разности записывают в виде табл. 2.

Таблица 2

x_0	$f(x_0)$			
		$f(x_0; x_1)$		
x_1	$f(x_1)$		$f(x_0; x_1; x_2)$	
		$f(x_1; x_2)$		$f(x_0; x_1; x_2; x_3)$
x_2	$f(x_2)$		$f(x_1; x_2; x_3)$	
		$f(x_2; x_3)$		$f(x_1; x_2; x_3; x_4)$
x_3	$f(x_3)$		$f(x_2; x_3; x_4)$	
		$f(x_3; x_4)$		
x_4	$f(x_4)$			

Лемма 2. *Разделенная разность n-го порядка выражается через узловые значения функции по формуле*

$$f(x_0; x_1; \ldots; x_n) = \sum_{i=0}^{n} \frac{f(x_i)}{(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)}, \quad (10)$$

т. е. является симметрической функцией своих аргументов.

При $n = 1$ сделанное утверждение вытекает из равенства (8). При $n = 2$ в соответствии с (9) имеем

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

$$= \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}. \quad (11)$$

Равенство (10) при $n = 2$ доказано. Доказательство для произвольного n, проводящееся по индукции, опустим.

Итак, согласно лемме 2 значение разделенной разности n-го порядка не зависит от нумерации $n+1$ узлов, по которым она строится. Всего имеется $(n+1)!$ различных вариантов их нумерации целыми числами от 0 до n.

Лемма 3. *Если* $x_k = x_0 + kh$, $k = 0, 1, \ldots$, *т. е. узлы расположены с постоянным шагом* $h > 0$, *то между разделенной разностью n-го порядка и конечной разностью n-го порядка имеется следующая связь:*

$$f(x_0;\ x_1;\ \ldots;\ x_n) = \frac{\Delta^n f_0}{n!\, h^n}. \tag{12}$$

Доказательство. Для $n = 1$ равенство (12) вытекает из (1), (7). При нахождении каждой следующей по порядку конечной разности происходит согласно (3) просто вычитание предыдущих разностей, а при вычислении следующей разделенной разности в соответствии с формулой (9) производится дополнительно к вычитанию деление на величину $x_n - x_0 = nh$. Отсюда и возникает величина $n!h^n$ в знаменателе правой части равенства (12).

Лемма 4. *Пусть* $[\alpha, \beta]$ — *минимальный отрезок, содержащий узлы* $x_0, x_1, \ldots, x_n$, $f \in C_n[\alpha, \beta]$. *Тогда существует такая точка* $\eta \in (\alpha, \beta)$, *что*

$$f(x_0;\ x_1;\ \ldots;\ x_n) = \frac{f^{(n)}(\eta)}{n!}. \tag{13}$$

Для узлов, расположенных с постоянным шагом, равенство (13) следует из лемм 1, 3. Доказательство леммы в общем случае опустим.

Следствие леммы 4. *Разделенная разность n-го порядка от алгебраического многочлена n-й степени принимает постоянное значение, не зависящее от выбора узлов* $x_0, x_1, \ldots, x_n$, *а разделенные разности более высоких порядков равны нулю.*

Конечные и разделенные разности имеют разнообразные применения. В § 9 они используются для построения интерполяционного многочлена.

§ 9. Интерполяционный многочлен Ньютона

Пусть $x_0, x_1, \ldots, x_n$ — произвольные попарно несовпадающие узлы, в которых известны значения функции f.

Лемма 1. *Алгебраический многочлен n-й степени*

$$l_n(x) = f(x_0) + (x - x_0) f(x_0;\ x_1) + \\ + (x - x_0)(x - x_1) f(x_0;\ x_1;\ x_2) + \ldots \\ \ldots + (x - x_0)(x - x_1) \ldots (x - x_{n-1}) f(x_0;\ x_1;\ \ldots;\ x_n) \tag{1}$$

является интерполяционным, т. е.

$$l_n(x_i) = f(x_i), \quad i = 0, 1, \ldots, n. \tag{2}$$

Прежде всего заметим, что так как разделенные разности $f(x_0; x_1)$, $f(x_0; x_1; x_2)$, ..., $f(x_0; x_1; \ldots; x_n)$ являются вполне определенными числами (см. § 8), то функция (1) действительно есть алгебраический многочлен n-й степени.

Докажем равенства (2) при $n = 2$. Имеем

$$l_2(x) = f(x_0) + (x - x_0) f(x_0; x_1) + \\ + (x - x_0)(x - x_1) f(x_0; x_1; x_2). \tag{3}$$

Очевидно, $l_2(x_0) = f(x_0)$. Далее, согласно (3), (8.7)

$$l_2(x_1) = f(x_0) + (x_1 - x_0) f(x_0; x_1) = \\ = f(x_0) + (x_1 - x_0) \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f(x_1).$$

Наконец, учитывая (3), (8.7), (8.11), получаем

$$l_2(x_2) = f(x_0) + (x_2 - x_0) f(x_0; x_1) + \\ + (x_2 - x_0)(x_2 - x_1) f(x_0; x_1; x_2) = f(x_0) + \\ + \frac{x_2 - x_0}{x_1 - x_0} (f(x_1) - f(x_0)) + \frac{(x_2 - x_0)(x_2 - x_1)}{(x_0 - x_1)(x_0 - x_2)} f(x_0) + \\ + \frac{(x_2 - x_0)(x_2 - x_1)}{(x_1 - x_0)(x_1 - x_2)} f(x_1) + f(x_2) = f(x_2).$$

При $n = 1$ равенства (2) устанавливаются аналогично, а для произвольного натурального $n > 2$ они доказываются по индукции.

Замечание. Многочлен (3) называется *интерполяционным многочленом Ньютона для неравных промежутков*. Согласно теореме 4.1 он тождественно совпадает с интерполяционным многочленом Лагранжа (4.5), т. е. $l_n(x) \equiv L_n(x)$.

Итак, мы имеем две различные записи интерполяционного многочлена. Остаточный член интерполяционного многочлена Ньютона тот же, что и у интерполяционного многочлена Лагранжа, т. е. всюду в равенствах (4.9), (4.14) и в неравенствах (4.15), (4.16) можно заменить $L_n(x)$ на $l_n(x)$.

У интерполяционного многочлена Лагранжа (4.5) видна явная его зависимость от каждого значения функции f_i, $i = 0, 1, \ldots, n$. Это во многих случаях

бывает полезно. Однако при изменении n интерполяционный многочлен Лагранжа требуется строить заново. В этом состоит его недостаток.

Интерполяционный многочлен Ньютона (1) выражается не через значения функции f, а через ее разделенные разности. При изменении степени n у интерполяционного многочлена Ньютона требуется только добавить или отбросить соответствующее число стандартных слагаемых. Это удобно на практике.

Случай равноотстоящих узлов. Пусть $x_k = x_0 + kh$, $h > 0$, $k = 0, 1, \ldots, n$, $f_k = f(x_k)$. Тогда, учитывая связь (8.12) разделенной разности с конечной разностью и вводя безразмерную переменную q (7.1), интерполяционный многочлен (1) можно переписать в следующем виде:

$$l_n(x) = l_n(x_0 + qh) = f_0 + q\frac{\Delta f_0}{1!} + \\ + q(q-1)\frac{\Delta^2 f_0}{2!} + \ldots + q(q-1)\ldots(q-n+1)\frac{\Delta^n f_0}{n!}. \tag{4}$$

Этот многочлен называется *интерполяционным многочленом Ньютона для интерполяции вперед.* В нем начало отсчета q расположено в крайнем левом узле x_0, а используемые конечные разности идут в таблице разностей от f_0 вправо вниз (см. табл. 8.1).

Интерполяционный многочлен (4) удобно использовать в начале таблицы и для экстраполяции левее точки x_0, т. е. для $q < 0$.

Рассмотрим пример интерполяции по формуле (4). Пусть дана таблица значений функции $f(x) = \sin x$ и ее конечных разностей:

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
5°	0,087156			
		34 713		
7°	0,121869		—148	
		34 565		—42
9°	0,156434		—190	
		34 375		—43
11°	0,190809		—233	
		34 142		—41
13°	0,224951		—274	
		33 868		
15°	0,258819			

Конечные разности функции в таблице для простоты принято выписывать в числе единиц последнего

десятичного знака, т. е. без явного указания положения запятой.

Допустим, что требуется найти $\sin 6°$. Из таблицы видно, что третьи разности близки к постоянной. Это согласно следствию к лемме 8.1 свидетельствует о том, что функция $f(x) = \sin x$ на рассматриваемом промежутке близка к некоторому алгебраическому многочлену третьей степени.

Полагаем в (4) $n = 3$, $x_0 = 5°$, $h = 7° - 5° = 2°$, $q = (6° - 5°)/2° = 1/2$. Вычисления имеют вид

$$f_0 = 0{,}087156,$$
$$q\,\Delta f_0 = 0{,}5 \cdot 0{,}034713 = 0{,}0173565,$$
$$q(q-1)\frac{\Delta^2 f_0}{2!} = \frac{1}{8} \cdot 0{,}000148 = 0{,}0100185,$$
$$q(q-1)(q-2)\frac{\Delta^3 f_0}{3!} = -\frac{1}{16} \cdot 0{,}000042 = -0{,}0000026$$

$$l_3(6°) = 0{,}104528$$

Здесь промежуточные значения найдены с семью знаками после запятой. Седьмой знак является запасным, в окончательном результате он округлен.

Точное значение $\sin 6°$, округленное с шестью знаками после запятой, равно 0,104528, т. е. все выписанные знаки у $l_3(6°)$ получились верные.

Интерполяционный многочлен с узлами x_0, x_{-1}, ..., x_{-n}, где $x_{-k} = x_0 - kh$, имеет вид

$$l_n(x) = l_n(x_0 + qh) = f_0 + q\frac{\Delta f_{-1}}{1!} + q(q+1)\frac{\Delta^2 f_{-2}}{2!} + \ldots$$
$$\ldots + q(q+1)\ldots(q+n-1)\frac{\Delta^n f_{-n}}{n!} \qquad (5)$$

и называется *интерполяционным многочленом Ньютона для интерполяции назад*. В нем начало отсчета q расположено в крайнем правом узле x_0, а используемые конечные разности идут в таблице от f_0 вправо вверх:

x_{-4}	f_{-4}			
		Δf_{-4}		
x_{-3}	f_{-3}		$\Delta^2 f_{-4}$	
		Δf_{-3}		$\Delta^3 f_{-4}$
x_{-2}	f_{-2}		$\Delta^2 f_{-3}$	
		Δf_{-2}		$\Delta^3 f_{-3}$
x_{-1}	f_{-1}		$\Delta^2 f_{-2}$	
		Δf_{-1}		
x_0	f_0			

Интерполяционный многочлен (5) удобно использовать при интерполяции в конце таблицы и для экстраполяции правее точки x_0, т. е. для $q > 0$.

Если при заданном x в таблице значений функции f с шагом h имеется достаточное число узлов с каждой стороны от x, то согласно замечанию 7.1 целесообразно узлы интерполяции $x_0, x_1, \ldots, x_n$ выбрать так, чтобы точка x оказалась возможно ближе к середине минимального отрезка, содержащего узлы. При этом интерполяционный многочлен можно строить по-разному.

Наиболее естественно задать интерполяционный многочлен в виде (1), где в качестве x_0 берется ближайший к x узел, затем за x_1 принимается ближайший к x узел, расположенный с противоположной от x стороны, чем x_0. Следующие узлы назначаются поочередно с разных сторон от x, расположенные возможно ближе к x. При таком выборе узлов следующие друг за другом слагаемые в выражении (1) обычно убывают, если h мало, а n невелико.

Возможно также в рассматриваемом случае использовать интерполяционные многочлены (4), (5), а также интерполяционный многочлен Лагранжа (7.4).

В заключение укажем, что остаточный член интерполяционного многочлена (4) имеет вид (7.7), а остаточный член интерполяционного многочлена (5) может быть записан в виде

$$R_n(x) = R_n(x_0 + qh) = h^{n+1} q(q+1) \ldots (q+n) \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

где $f^{(n+1)}$ — производная по x, ξ — некоторая точка минимального отрезка, содержащего узлы интерполяции $x_0, x_{-1}, \ldots, x_{-n}$ и точку x.

Согласно (8.4), (7.7) при условии, если h мало, а функция f достаточно гладкая, текущее слагаемое в выражении (4) интерполяционного многочлена Ньютона приблизительно равно погрешности интерполяции многочленом, составленным из всех предшествующих слагаемых. Это замечание относится и к интерполяционному многочлену (5) для интерполяции назад.

§ 10. Численное дифференцирование

В дальнейшем нам потребуется следующая

Лемма 1. *Пусть* *) $f \in C[a, b]$, $\xi_i \in [a, b]$ — *произвольные точки*, $i = 1, 2, \ldots, n$. *Тогда существует такая точка* $\xi \in [a, b]$, *что*

$$\frac{f(\xi_1) + f(\xi_2) + \ldots + f(\xi_n)}{n} = f(\xi).$$

Эта лемма вытекает из очевидных неравенств

$$\min_{[a, b]} f(x) \leqslant \frac{f(\xi_1) + f(\xi_2) + \ldots + f(\xi_n)}{n} \leqslant \max_{[a, b]} f(x)$$

и теоремы о промежуточных значениях непрерывной функции.

Простейшие формулы численного дифференцирования. Допустим, что в некоторой точке x у функции f существует производная

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x},$$

которую точно вычислить либо не удается, либо слишком сложно. Тогда естественно положить

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

Спрашивается, какова же погрешность, т. е. разность между левой и правой частями этого приближенного равенства? Для получения количественных оценок погрешности одного факта существования $f'(x)$ недостаточно. Поэтому при исследовании погрешности приближенных формул численного дифференцирования обычно требуют наличие у функции некоторой производной более высокого порядка, чем искомая производная.

Остановимся на трех простейших наиболее употребляемых формулах численного дифференцирования. Пусть $x_i = x_0 + ih$, $i = 0, \pm 1, \ldots$, $h > 0$ — шаг. Обозначим $f_i = f(x_i)$, $f_i' = f'(x_i)$ и т. д.

---

*) Через $C[a, b]$ обозначен класс непрерывных на отрезке $[a, b]$ функций (см. п. 2 введения).

Допустим, что $f \in C_2[x_0, x_1]$. Тогда существует такая точка ξ, что

$$f'_0 = \frac{f_1 - f_0}{h} - \frac{h}{2} f''(\xi), \quad x_0 < \xi < x_1. \tag{1}$$

Если $f \in C_3[x_{-1}, x_1]$, то, кроме того,

$$f'_0 = \frac{f_1 - f_{-1}}{2h} - \frac{h^2}{6} f'''(\xi), \quad x_{-1} < \xi < x_1. \tag{2}$$

При условии, что $f \in C_4[x_{-1}, x_1]$, имеем

$$f''_0 = \frac{f_{-1} - 2f_0 + f_1}{h^2} - \frac{h^2}{12} f^{(4)}(\xi), \quad x_{-1} < \xi < x_1. \tag{3}$$

Точка ξ в каждой из формул (1)—(3) неизвестна.

Докажем соотношения (1) и (3). Согласно формуле Тейлора имеем

$$f_1 = f_0 + hf'_0 + \frac{h^2}{2} f''(\xi),$$

где ξ — некоторая точка интервала (x_0, x_1), т. е. верно (1). Аналогично, если $f \in C_4[x_{-1}, x_1]$, то

$$f_{\pm 1} = f_0 \pm hf'_0 + \frac{h^2}{2} f''_0 \pm \frac{h^3}{6} f'''_0 + \frac{h^4}{24} f^{(4)}(\xi_\pm), \tag{4}$$

где знак $\pm$ можно заменить либо всюду на $+$, либо всюду на $-$, $x_{-1} < \xi_- < \xi_+ < x_1$.

Подставив в выражение $\dfrac{f_{-1} - 2f_0 + f_1}{h^2}$ значения f_{-1}, f_1 в виде (4) и учтя, что по лемме 1

$$f^{(4)}(\xi_-) + f^{(4)}(\xi_+) = 2f(\xi),$$

где ξ — некоторая точка отрезка $[\xi_-, \xi_+]$, получим

$$\frac{f_{-1} - 2f_0 + f_1}{h^2} = f''_0 + \frac{h^2}{12} f^{(4)}(\xi), \quad x_{-1} < \xi < x_1,$$

т. е. приходим к соотношению (3).

Формулы (1)—(3) называются *формулами численного дифференцирования с остаточными членами*, а формулы

$$f'_0 \approx \frac{f_1 - f_0}{h}, \tag{5}$$

$$f'_0 \approx \frac{f_1 - f_{-1}}{2h}, \tag{6}$$

$$f''_0 \approx \frac{f_{-1} - 2f_0 + f_1}{h^2} \tag{7}$$

называются просто *формулами численного дифференцирования.*

Величины

$$\frac{f_1-f_0}{h}, \quad \frac{f_1-f_{-1}}{2h}, \quad \frac{f_{-1}-2f_0+f_1}{h^2}$$

называются соответственно *разностной производной, центральной разностной производной* и *второй разностной производной.*

Погрешности, например, формул (5), (6) оцениваются с помощью следующих неравенств, вытекающих из соотношений (1), (2):

$$\left|f_0'-\frac{f_1-f_0}{h}\right|\leqslant\frac{h}{2}\max_{[x_0,\,x_1]}|f''(x)|,$$

$$\left|f_0'-\frac{f_1-f_{-1}}{2h}\right|\leqslant\frac{h^2}{6}\max_{[x_{-1},\,x_1]}|f'''(x)|.$$

Говорят, что погрешность формулы (5) имеет *первый порядок относительно* h (или *порядка* h), а погрешность формул (6) и (7) имеет *второй порядок относительно* h (или *порядка* h^2). Также говорят, что формула численного дифференцирования (5) *первого порядка точности* (относительно h), а формулы (6) и (7) имеют *второй порядок точности.*

Применение интерполяционного многочлена Лагранжа. Для нахождения производных любого порядка существуют формулы численного дифференцирования любого порядка точности. Один из универсальных способов построения формул численного дифференцирования состоит в том, что по значениям функции f в некоторых узлах $x_0, x_1, \ldots, x_n$ строят интерполяционный многочлен $L_n(x)$ (4.5) и приближенно полагают

$$f^{(m)}(x)\approx L_n^{(m)}(x), \qquad 0\leqslant m\leqslant n. \tag{8}$$

В ряде случаев наряду с приближенным равенством (8) удается получить точное равенство, содержащее остаточный член, выражающийся через производную $f^{(n+1)}$. Приведем в дополнение к (1)—(3) несколько распространенных формул для первой ($m=1$) и второй ($m=2$) производных в узлах, расположенных с постоянным шагом $h>0$.

$m = 1$, $n = 2$ (три узла):

$$f'_0 = \frac{1}{2h}(-3f_0 + 4f_1 - f_2) + \frac{h^2}{3}f'''(\xi), \tag{9}$$

$$f'_1 = \frac{1}{2h}(f_2 - f_0) - \frac{h^2}{6}f'''(\xi),$$

$$f'_2 = \frac{1}{2h}(f_0 - 4f_1 + 3f_2) + \frac{h^2}{3}f'''(\xi). \tag{10}$$

$m = 2$, $n = 2$ (три узла):

$$f''_0 = \frac{1}{h^2}(f_0 - 2f_1 + f_2) - hf'''(\xi), \tag{11}$$

$$f''_1 = \frac{1}{h^2}(f_0 - 2f_1 + f_2) - \frac{h^2}{12}f^{(4)}(\xi),$$

$$f''_2 = \frac{1}{h^2}(f_0 - 2f_1 + f_2) + hf'''(\xi).$$

$m = 1$, $n = 3$ (четыре узла):

$$f'_0 = \frac{1}{6h}(-11f_0 + 18f_1 - 9f_2 + 2f_3) - \frac{h^3}{4}f^{(4)}(\xi), \tag{12}$$

$$f'_1 = \frac{1}{6h}(-2f_0 - 3f_1 + 6f_2 - f_3) + \frac{h^3}{12}f^{(4)}(\xi),$$

$$f'_2 = \frac{1}{6h}(f_0 - 6f_1 + 3f_2 + 2f_3) - \frac{h^3}{12}f^{(4)}(\xi),$$

$$f'_3 = \frac{1}{6h}(-2f_0 + 9f_1 - 18f_2 + 11f_3) + \frac{h^3}{4}f^{(4)}(\xi). \tag{13}$$

$m = 2$, $n = 3$ (четыре узла):

$$f''_0 = \frac{1}{h^2}(2f_0 - 5f_1 + 4f_2 - f_3) + \frac{11}{12}h^2f^{(4)}(\xi),$$

$$f''_1 = \frac{1}{h^2}(f_0 - 2f_1 + f_2) - \frac{h^2}{12}f^{(4)}(\xi),$$

$$f''_2 = \frac{1}{h^2}(f_1 - 2f_2 + f_3) - \frac{h^2}{12}f^{(4)}(\xi),$$

$$f''_3 = \frac{1}{h^2}(-f_0 + 4f_1 - 5f_2 + 2f_3) + \frac{11}{12}h^2f^{(4)}(\xi). \tag{14}$$

В приведенных формулах ξ есть некоторая неизвестная точка из интервала (x_0, x_n). Остаточные члены этих формул находятся с помощью формулы Тейлора с остаточным членом в интегральной форме. При этом предполагается, что на отрезке $[x_0, x_n]$ у функции f непрерывна производная, через которую выражается остаточный член.

Можно было бы продолжить приведенный список формул численного дифференцирования для возрастающих значений n, m. При этом обнаруживаются следующие закономерности. С ростом n и соответствующей гладкости функции f порядок точности формул увеличивается, а с ростом m, т. е. номера производной, порядок точности относительно h убывает. Выражения производных в узлах, расположенных ближе к середине отрезка $[x_0, x_n]$, более простые, чем у его концов. При четном n в среднем узле для четной производной порядок точности формулы на единицу больше, чем в остальных узлах. Поэтому рекомендуется по возможности использовать формулы численного дифференцирования с узлами, расположенными симметрично относительно той точки, в которой ищется производная.

Выбор оптимального шага. В формулах численного дифференцирования с постоянным шагом h значения функции f делятся на h^m, где m — порядок вычисляемой производной. Поэтому при малом h неустранимые погрешности в значениях функции f оказывают сильное влияние на результат численного дифференцирования. Таким образом, возникает задача выбора оптимального шага h, ибо погрешность собственного метода стремится к нулю при $h \to 0$, а неустранимая погрешность растет.

Допустим, что абсолютная погрешность $\Delta(f_i)$ в каждом значении функции f_i удовлетворяет неравенству

$$\Delta(f_i) \leqslant \bar{\Delta}, \tag{15}$$

т. е. $\bar{\Delta}$ — предельная абсолютная погрешность значений функции. Попытаемся найти оптимальное h в формулах (6) и (7).

Пусть в некоторой окрестности точки x_0 производные, через которые выражаются остаточные члены этих формул (см. (2), (3)), непрерывны и удовлетворяют неравенствам

$$|f'''(x)| \leqslant \bar{M}_3, \qquad |f^{(4)}(x)| \leqslant \bar{M}_4, \tag{16}$$

где $\bar{M}_3$, $\bar{M}_4$ — некоторые числа. Тогда полная погрешность формул (6) и (7) (без учета погрешностей округления) в соответствии с (2), (3) и (15), (16) не

превосходит соответственно величин

$$\varepsilon_1 = \frac{\bar{\Delta} + \bar{\Delta}}{2h} + \frac{h^2}{6}\bar{M}_3, \tag{17}$$

$$\varepsilon_2 = \frac{\bar{\Delta} + 2\bar{\Delta} + \bar{\Delta}}{h^2} + \frac{h^2}{12}\bar{M}_4. \tag{18}$$

Минимизация по h этих величин приводит к следующим значениям h:

$$h = h_1 = \left(\frac{3\bar{\Delta}}{\bar{M}_3}\right)^{1/3}, \qquad h = h_2 = 2\left(\frac{3\bar{\Delta}}{\bar{M}_4}\right)^{1/4}, \tag{19}$$

при этом

$$\varepsilon_1 = \frac{3}{2}\left(\frac{\bar{M}_3\bar{\Delta}^2}{3}\right)^{1/3}, \qquad \varepsilon_2 = 2\left(\frac{\bar{M}_4\bar{\Delta}}{3}\right)^{1/2}. \tag{20}$$

Если при выбранном для какой-либо из формул (6) или (7) значении h отрезок $[x_{-1}, x_1]$, где $x_{\pm 1} = x_0 \pm h$, не выходит за пределы окрестности точки x_0, в которой выполняется соответствующее неравенство (16), то найденное h является *оптимальным* и полная погрешность численного дифференцирования оценивается соответствующей величиной (20). В противном случае h окончательно выбирается так, чтобы отрезок $[x_{-1}, x_1]$ не выходил из указанной окрестности точки x_0, а полная погрешность численного дифференцирования оценивается величиной (17) или (18).

Применение интерполяционных многочленов Ньютона. Все приведенные выше формулы численного дифференцирования выражаются через табличные значения функции. Если продифференцировать интерполяционный многочлен Ньютона (9.4), то получится формула численного дифференцирования, выражающаяся через конечные разности функции.

Принимая во внимание, что согласно (7.1)

$$\frac{d}{dx} = \frac{1}{h}\frac{d}{dq}, \tag{21}$$

из (9.4) находим

$$f'(x) \approx \frac{1}{h}\frac{d}{dq} l_n(x_0 + qh) =$$

$$= \frac{1}{h}\left(\Delta f_0 + (2q-1)\frac{\Delta^2 f_0}{2!} + (3q^2 - 6q + 2)\frac{\Delta^3 f_0}{3!} + \ldots\right).$$

Этой формулой удобно пользоваться в начале таблицы значений функции f с шагом h. Аналогичную формулу численного дифференцирования можно получить из интерполяционного многочлена Ньютона (9.5) для интерполяции назад.

Общая оценка погрешности. Точные равенства вида (1)—(3), (9)—(14) с остаточным членом, выражаемым через $n+1$-ю производную функции f, удается найти для формул численного дифференцирования только в частных случаях, некоторые из которых и приведены выше. Однако возможно получить оценку погрешности общей формулы численного дифференцирования (8), выражаемую в виде неравенства, через максимум модуля производной $f^{(k+1)}$ при любых m, k, n таких, что $0 \leqslant m \leqslant k \leqslant n$. Мы ограничимся рассмотрением случая расположения узлов с постоянным шагом h и сформулируем результат в виде теоремы.

Теорема 1. *Пусть* $x_i = x_0 + ih$, $h > 0$, $i = 0, 1, \ldots, n$, $0 \leqslant k \leqslant n$, $f \in C_{k+1}[x_0, x_n]$. *Тогда существуют такие постоянные* a_{nkm}, *зависящие только от* n, k, m *и не зависящие от шага* h *и функции* f, *что*

$$\max_{[x_0, x_n]} |f^{(m)}(x) - L_n^{(m)}(x)| \leqslant \leqslant h^{k+1-m} a_{nkm} \max_{[x_0, x_n]} |f^{(k+1)}(x)|, \qquad (22)$$

где $L_n(x)$ — *интерполяционный многочлен Лагранжа* (4.5) *для функции* f, $0 \leqslant m \leqslant k \leqslant n$.

Доказательство. Многочлен

$$Q_k(x) = f_0 + (x - x_0)\frac{f_0'}{1!} + \ldots + (x - x_0)^k \frac{f_0^{(k)}}{k!}$$

является многочленом Тейлора степени k для функции f в точке x_0. Если его m раз продифференцировать, $0 \leqslant m \leqslant k$, то получится многочлен

$$Q_k^{(m)}(x) = f_0^{(m)} + (x - x_0)\frac{f_0^{(m+1)}}{1!} + \ldots + (x - x_0)^{k-m}\frac{f_0^{(k)}}{(k-m)!},$$

являющийся многочленом Тейлора степени $k - m$ для производной $f^{(m)}$ в точке x_0.

Поскольку $f^{(m)} \in C_{k+1-m}[x_0, x_n]$, то согласно формуле Тейлора имеем при $x \in [x_0, x_n]$

$$f^{(m)}(x) = Q_k^{(m)}(x) + (x - x_0)^{k+1-m} \frac{f^{(k+1)}(\xi)}{(k+1-m)!}, \qquad (23)$$

где $0 \leqslant m \leqslant k$, $\xi \in [x_0, x]$ зависит от x, m, n.

Если $x \in [x_0, x_n]$, то $|x - x_0| \leqslant nh$. Отсюда и из (23) при $0 \leqslant m \leqslant k$ следует оценка

$$\max_{[x_0, x_n]} |f^{(m)}(x) - Q_k^{(m)}(x)| \leqslant$$

$$\leqslant h^{k+1-m} \frac{n^{k+1-m}}{(k+1-m)!} \max_{[x_0, x_n]} |f^{(k+1)}(x)|. \qquad (24)$$

Положим

$$f(x) = Q_k(x) + \rho(x). \qquad (25)$$

Тогда на основании (23) при $m = 0$ получим

$$|\rho_i| \leqslant h^{k+1} \frac{i^{k+1}}{(k+1)!} \max_{[x_0, x_n]} |f^{(k+1)}(x)|, \qquad (26)$$

где $\rho_i = \rho(x_i)$, $i = 0, 1, \ldots, n$.

Обозначим для удобства через $L_n(x; f)$ интерполяционный многочлен (4.5) именно для функции f. Согласно (25) и замечаниям 4.2, 4.1

$$L_n(x; f) = L_n(x; Q_k) + L_n(x; \rho) = Q_k(x) + L_n(x; \rho). \qquad (27)$$

Введя безразмерную переменную $q = (x - x_0)/h$, аналогично (7.4) получаем

$$L_n(x; \rho) = \sum_{i=0}^{n} \bar{p}_{ni}(q) \rho_i, \qquad (28)$$

где $\bar{p}_{ni}(q)$ — лагранжевы коэффициенты (7.5). Принимая во внимание (21), из (28) находим

$$L_n^{(m)}(x; \rho) = \frac{1}{h^m} \sum_{i=0}^{n} \rho_i \frac{d^m}{dq^m} \bar{p}_{ni}(q), \qquad (29)$$

где m-я производная слева взята по x. Поскольку функции (7.5) не зависят от h, то величины

$$\max_{x \in [x_0, x_n]} \left| \frac{d^m}{dq^m} \bar{p}_{ni}(q) \right| = \max_{q \in [0, n]} \left| \frac{d^m}{dq^m} \bar{p}_{ni}(q) \right| = c_{nmi} \qquad (30)$$

зависят только от n, m, i.

Возвращаясь к первоначальному обозначению $L_n(x)$ интерполяционного многочлена для функции f и учитывая (27), получаем

$$|f^{(m)}(x) - L_n^{(m)}(x)| = |f^{(m)}(x) - Q_k^{(m)}(x) - L_n^{(m)}(x; \rho)| \leqslant$$
$$\leqslant |f^{(m)}(x) - Q_k^{(m)}(x)| + |L_n^{(m)}(x; \rho)|.$$

Отсюда на основании (24), (29), (26), (30) вытекает неравенство (22) с постоянной

$$a_{nkm} = \frac{n^{k+1-m}}{(k+1-m)!} + \frac{1}{(k+1)!} \sum_{i=1}^{n} i^{k+1} c_{nmi}, \qquad (31)$$

не зависящей от шага h и функции f.

Замечания. 1. Данная теорема дает дополнительные сведения и для $m = 0$, т. е. просто для интерполяции. Именно, если строится интерполяционный многочлен n-й степени, а у функции f на отрезке $[x_0, x_n]$ существует непрерывная производная только порядка $k+1$, $k < n$, то полученная ранее оценка погрешности (7.8) непригодна. Согласно же теореме 1 погрешность интерполяции на отрезке $[x_0, x_n]$ в данном случае имеет $k+1$-й порядок относительно h.

2. Оценку (22) с постоянной (31) на практике при численном дифференцировании использовать нецелесообразно, так как эта постоянная сильно завышена. Оценка (22) полезна тем, что она устанавливает скорость убывания погрешности относительно h на всем отрезке $[x_0, x_n]$ при фиксированных параметрах n, k, m, $0 \leqslant m \leqslant k \leqslant n$; шаг h является одним из наиболее гибких параметров, которым распоряжается вычислитель.

§ 11. Сплайны

Пусть отрезок $[a, b]$ разбит на N равных частичных отрезков $[x_i, x_{i+1}]$, где $x_i = a + ih$, $i = 0, 1, \ldots$ $\ldots, N-1$, $x_N = b$, $h = (b-a)/N$.

Сплайном называется функция, которая вместе с несколькими производными непрерывна на всем заданном отрезке $[a, b]$, а на каждом частичном отрезке $[x_i, x_{i+1}]$ в отдельности является некоторым алгебраическим многочленом.

Максимальная по всем частичным отрезкам степень многочленов называется *степенью сплайна*, а разность между степенью сплайна и порядком наивысшей непрерывной на $[a, b]$ производной — *дефектом сплайна*.

Например, непрерывная кусочно линейная функция (ломаная) является сплайном первой степени с дефектом, равным единице, так как непрерывна только сама функция (нулевая производная), а первая производная уже разрывна.

На практике наиболее широкое распространение получили сплайны третьей степени, имеющие на $[a, b]$ непрерывную, по крайней мере, первую производную. Эти сплайны называются *кубическими* и обозначаются через $S_3(x)$. Величина $m_i = S_3'(x_i)$ называется *наклоном сплайна* в точке (узле) x_i.

Нетрудно убедиться, что кубический сплайн $S_3(x)$, принимающий в узлах x_i, x_{i+1} соответственно значения f_i, f_{i+1}, имеет на частичном отрезке $[x_i, x_{i+1}]$ вид

$$S_3(x) = \frac{(x_{i+1}-x)^2(2(x-x_i)+h)}{h^3} f_i + \frac{(x-x_i)^2(2(x_{i+1}-x)+h)}{h^3} f_{i+1} + \frac{(x_{i+1}-x)^2(x-x_i)}{h^2} m_i + \frac{(x-x_i)^2(x-x_{i+1})}{h^2} m_{i+1}. \quad (1)$$

Действительно, легко видеть, что $S_3(x_i) = f_i$, $S_3(x_{i+1}) = f_{i+1}$. Кроме того, простые вычисления показывают, что $S_3'(x_i) = m_i$, $S_3'(x_{i+1}) = m_{i+1}$.

Можно доказать, что любой алгебраический многочлен третьей степени, принимающий в точках x_i, x_{i+1} значения, равные соответственно f_i, f_{i+1}, и имеющий в этих точках производную, соответственно равную m_i, m_{i+1}, тождественно совпадает с многочленом (1).

Итак, чтобы задать кубический сплайн $S_3(x)$ на всем отрезке $[a, b]$, нужно задать в $N+1$ узлах x_i его значения f_i и наклоны m_i, $i = 0, 1, \ldots, N$.

Сплайны используются для различных целей. Кубический сплайн, принимающий в узлах x_i те же значения f_i, что и некоторая функция f, называется *интерполяционным*. Он служит для аппроксимации

функции f на отрезке $[a, b]$ вместе с несколькими производными.

Способы задания наклонов интерполяционного кубического сплайна.

1 (*упрощенный способ*). Полагаем

$$m_i = \frac{f_{i+1} - f_{i-1}}{2h}, \qquad i = 1, 2, \ldots, N-1, \tag{2}$$

$$m_0 = \frac{4f_1 - f_2 - 3f_0}{2h}, \quad m_N = \frac{3f_N + f_{N-2} - 4f_{N-1}}{2h}. \tag{3}$$

Заметим, что формулы (2), (3) согласно (10.2), (10.9), (10.10) являются формулами численного дифференцирования второго порядка точности относительно шага $h = (b-a)/N$.

2. Если известны значения f'_i производной f' в узлах x_i, то полагаем $m_i = f'_i$, $i = 0, 1, \ldots, N$.

Способы 1, 2 называются *локальными*, поскольку с их помощью на каждом частичном отрезке $[x_i, x_{i+1}]$ сплайн строится отдельно (непосредственно по формуле (1)). При этом тем не менее соблюдается непрерывность в узлах x_i производной $S'_3(x)$. Непрерывность же второй производной $S''_3(x)$ в узлах сплайна, построенного локальным способом 1 или 2, не гарантируется. Поэтому дефект такого сплайна обычно равен двум.

3 (*глобальный способ*). Обозначим через $S''_3(x_i + 0)$ значение $S''_3(x)$ в узле x_i справа, найденное непосредственно из выражения (1), а через $S''_3(x_i - 0)$ значение $S''_3(x)$ в узле x_i слева, т. е. найденное из соответствующего выражения $S_3(x)$ на частичном отрезке $[x_{i-1}, x_i]$, которое получается из (1) заменой i на $i-1$.

Имеем

$$S''_3(x_i + 0) = -\frac{4m_i}{h} - \frac{2m_{i+1}}{h} + 6\frac{f_{i+1} - f_i}{h^2},$$

$$S''_3(x_i - 0) = \frac{2m_{i-1}}{h} + \frac{4m_i}{h} - 6\frac{f_i - f_{i-1}}{h^2}.$$

Требуем непрерывность $S''(x)$ в узлах:

$$S''_3(x_i - 0) = S''_3(x_i + 0), \qquad i = 1, 2, \ldots, N-1,$$

и приходим к следующей системе линейных алгебраических уравнений относительно наклонов:

$$m_{i-1}+4m_i+m_{i+1}=\frac{3(f_{i+1}-f_{i-1})}{h},\quad i=1, 2, \ldots, N-1. \quad (4)$$

Поскольку неизвестных $N+1$, то нужно задать еще два условия, которые называются *краевыми* (они обычно связаны с «крайними» значениями m_0, m_N). Дадим три варианта краевых условий.

а) Если известны $f'_0=f'(a)$, $f'_N=f'(b)$, то задаем

$$m_0=f'_0,\quad m_N=f'_N. \quad (5)$$

б) Производные f'_0, f'_N аппроксимируем формулами численного дифференцирования третьего порядка точности, полученными из (10.12), (10.13) отбрасыванием остаточных членов, и полагаем

$$\begin{aligned} m_0&=\frac{1}{6h}(-11f_0+18f_1-9f_2+2f_3),\\ m_N&=\frac{1}{6h}(11f_N-18f_{N-1}+9f_{N-2}-2f_{N-3}). \end{aligned} \quad (6)$$

в) В некоторых случаях бывают известны значения f'' на концах отрезка $[a, b]$, т. е. величины $f''_0=f''(a)$, $f''_N=f''(b)$. Тогда требования $S''_3(a)=f''_0$, $S''_3(b)=f''_N$ приводят к краевым условиям

$$\begin{aligned} m_0&=-\frac{m_1}{2}+\frac{3}{2}\frac{f_1-f_0}{h}-\frac{h}{4}f''_0,\\ m_N&=-\frac{m_{N-1}}{2}+\frac{3}{2}\frac{f_N-f_{N-1}}{h}+\frac{h}{4}f''_N. \end{aligned} \quad (7)$$

Краевые условия (5)—(7) можно комбинировать, т. е. в левом и правом крайних узлах выбирать их независимо.

Система (4) при всех рассмотренных краевых условиях имеет единственное решение, для нахождения которого могут быть применены методы прогонки и итераций (это будет доказано в § 21, 22).

Решая систему (4) при выбранных краевых условиях, находим наклоны m_i, $i=0, 1, \ldots, N$, во всех узлах. Затем по формуле (1) задаем сплайн на каждом частичном отрезке $[x_i, x_{i+1}]$, $i=0, 1, \ldots, N-1$. Построенный данным глобальным способом сплайн $S_3(x)$ имеет дефект не больше единицы, так как этот

сплайн обладает на отрезке $[a, b]$ непрерывной второй производной $S_3''(x)$.

Погрешность приближения сплайном. Кубический интерполяционный сплайн достаточно хорошо приближает гладкие функции вместе с несколькими производными, о чем свидетельствует следующая теорема, доказываемая аналогично теореме 10.1.

Теорема 1. *Если* $f \in C_{k+1}[a, b]$, $0 \leqslant k \leqslant 3$, *то интерполяционный сплайн* $S_3(x)$ *с наклонами, заданными способом* 2 *или* 3, *удовлетворяет неравенству*

$$\max_{[x_i, x_{i+1}]} |f^{(m)}(x) - S_3^{(m)}(x)| \leqslant ch^{k+1-m} \max_{[a, b]} |f^{(k+1)}(x)|, \quad (8)$$

где $i = 0, 1, \ldots, N-1$, $m = 0, 1, \ldots, k$, c — *не зависящая от* h, i, f *постоянная.*

Таким образом, если $f \in C_4[a, b]$, т. е. $k = 3$, и наклоны сплайна $S_3(x)$ найдены глобальным способом, то на $[a, b]$ максимальные по модулю уклонения $S_3(x)$ от $f(x)$, $S_3'(x)$ от $f'(x)$ и $S_3''(x)$ от $f''(x)$ равны соответственно $O(h^4)$, $O(h^3)$ и $O(h^2)$. При задании наклонов способом 2 имеющиеся в узлах скачки у $S_3''(x)$ не превышают удвоенной правой части неравенства (8), т. е. при $k = 3$ будут $O(h^2)$. Эти скачки у второй производной на графике заметить трудно.

Замечания. 1. Для кубического сплайна $S_3(x)$, наклоны которого задаются упрощенным способом 1, справедлива теорема 1 при условии, что $0 \leqslant k \leqslant 2$.

2. Если функция f непрерывно дифференцируема на всей действительной оси $k+1$ раз, $0 \leqslant k \leqslant 3$, и имеет период, равный $b - a$, то следует положить $m_0 = m_N$ и к системе (4) присоединить уравнение

$$m_{N-1} + 4m_0 + m_1 = \frac{3(f_1 - f_{N-1})}{h},$$

отвечающее значению $i = 0$. Из расширенной системы, имеющей единственное решение, находятся наклоны сплайна $S_3(x) \in C_2[a, b]$; этот сплайн продолжается с отрезка $[a, b]$ с периодом $b - a$ на всю ось x с сохранением непрерывности второй производной. При этом справедливо неравенство (8).

Сплайны являются более удобным средством аппроксимации функций на больших промежутках (при

больших N), чем, скажем, интерполяционные многочлены. Аппроксимация функции на большом промежутке одним многочленом может потребовать для достижения заданной точности значительного увеличения его степени, что на практике неприемлемо.

Разбиение заданного отрезка $[a, b]$ на несколько частей с построением на каждой части самостоятельного интерполяционного многочлена неудобно тем, что на стыках будет терпеть разрыв первая производная двух соседних интерполяционных многочленов. Возможно, даже не совпадут на стыке сами значения интерполяционных многочленов, если точка стыка не является их общим узлом. Кубический же сплайн $S_3(x)$, наклоны которого найдены глобальным способом, дважды непрерывно дифференцируем на всем отрезке $[a, b]$, т. е. имеет непрерывную кривизну. Точность аппроксимации функции f сплайном $S_3(x)$ управляется выбором N, т. е. шагом $h=(b-a)/N$.

§ 12. Равномерные приближения функций

В линейном пространстве непрерывных функций $C[a, b]$ введем норму *)

$$\|f\|_{C[a, b]}=\max_{[a, b]}|f(x)|. \tag{1}$$

Норма разности двух функций f, $g \in C[a, b]$, т. е. величина

$$\|f-g\|_{C[a, b]}=\max_{[a, b]}|f(x)-g(x)|,$$

равна максимальному уклонению этих функций друг от друга на отрезке $[a, b]$.

Пусть некоторая функция f принадлежит $C[a, b]$. Доказано, что среди всех алгебраических многочленов n-й степени существует и притом единственный многочлен $P_n^f(x)$, обладающий тем свойством, что

$$\|f-P_n^f\|_{C[a, b]} \leqslant \|f-P_n\|_{C[a, b]},$$

где $P_n(x)$ — произвольный многочлен n-й степени.

*) Понятия нормы и нормированного пространства даны в п. 7 введения.

Многочлен $P_n^f(x)$ называется *многочленом (n-й степени) наилучшего равномерного приближения* функции f. Число

$$E_n(f) = \|f - P_n^f\|_{C[a,\,b]} = \inf \|f - P_n\|_{C[a,\,b]},$$

где точная нижняя грань берется по всем многочленам n-й степени, называется *наилучшим равномерным приближением* функции f многочленами n-й степени. Очевидно, $E_n(f) \leqslant E_{n-1}(f)$ при любом натуральном n.

Можно доказать, что, какова бы ни была функция $f \in C[a, b]$, последовательность $\{P_n^f(x)\}$ ее многочленов наилучшего равномерного приближения сходится равномерно к f на $[a, b]$, т. е. $E_n(f) \to 0$ при $n \to \infty$.

Пример. Пусть $f(x) = x^{n+1}$. Требуется найти на отрезке $[-1, 1]$ многочлен n-й степени наилучшего равномерного приближения заданной функции.

Решение. Рассмотрим многочлен

$$\overline{T}_{n+1}(x) = 2^{-n} T_{n+1}(x), \tag{2}$$

где $T_{n+1}(x)$ — наименее уклоняющийся от нуля многочлен Чебышева $n+1$-й степени (см. § 6). Многочлен (2) согласно свойствам 1, 2 многочленов Чебышева можно записать в следующем виде:

$$\overline{T}_{n+1}(x) = x^{n+1} - \alpha x^{n-1} - \beta x^{n-3} - \ldots, \tag{3}$$

где α, β, ... — вполне определенные числовые коэффициенты.

По свойству 5 многочленов Чебышева имеем

$$\|\overline{T}_{n+1}\|_{C[-1,\,1]} \leqslant \|\overline{P}_{n+1}\|_{C[-1,\,1]}, \tag{4}$$

где $\overline{P}_{n+1}(x)$ — произвольный многочлен $n+1$-й степени со старшим коэффициентом, равным единице.

Из (3), (4) следует, что многочлен наилучшего равномерного приближения $P_n^f(x)$ для функции $f(x) = x^{n+1}$ на отрезке $[-1, 1]$ имеет вид

$$P_n^f(x) = \alpha x^{n-1} + \beta x^{n-3} + \ldots \tag{5}$$

Действительно,

$$\overline{T}_{n+1}(x) = x^{n+1} - P_n^f(x) = f(x) - P_n^f(x),$$

$$\overline{P}_{n+1}(x) = x^{n+1} - P_n(x) = f(x) - P_n(x),$$

где $P_n(x)$ — произвольный многочлен n-й степени, и согласно (4) выполняется требуемое неравенство

$$\|f - P_n^f\|_{C[-1,\,1]} \leqslant \|f - P_n\|_{C[-1,\,1]}.$$

При этом на основании (2) и свойства 4 многочленов Чебышева получаем

$$E_n(x^{n+1}) = E_n(f) = \|f - P_n^f\|_{C[-1,1]} = \|\overline{T}_{n+1}\|_{C[-1,1]} = \\ = \|2^{-n}T_{n+1}\|_{C[-1,1]} = 2^{-n}\|T_{n+1}\|_{C[-1,1]} = 2^{-n}.$$

Критерий многочлена наилучшего равномерного приближения. Обратим теперь внимание на одно важное свойство многочлена наилучшего равномерного приближения $P_n^f(x)$ в рассмотренном примере. В силу свойства 4 многочленов Чебышева разность $f(x) - P_n^f(x)$, совпадающая с выражением (2) для $\overline{T}_{n+1}(x)$, принимает в $n+2$ точках

$$x_m = \cos\frac{m\pi}{n+1}, \qquad m = 0, 1, \ldots, n+1,$$

отрезка $[-1, 1]$ значения $(-1)^m\|f - P_n^f\|_{C[-1,1]}$. Другими словами, на отрезке $[-1, 1]$ имеются $n+2$ точки, в которых разность $f(x) - P_n^f(x)$ достигает максимального по модулю значения с чередующимися знаками.

Отмеченное свойство является общим и служит критерием многочлена наилучшего равномерного приближения любой непрерывной функции, о чем свидетельствует следующая теорема, которую приведем без доказательства.

Теорема 1 (Чебышева). *Чтобы многочлен $P_n(x)$ был многочленом наилучшего равномерного приближения функции $f \in C[a, b]$, необходимо и достаточно существования на $[a, b]$ по крайней мере $n+2$ точек $x_0 < x_1 < \ldots < x_{n+1}$ таких, что*

$$f(x_i) - P_n(x_i) = \sigma(-1)^i\|f - P_n\|_{C[a,b]},$$

$i = 0, 1, \ldots, n+1$, $\sigma = 1$ или -1 одновременно для всех i.

Точки $x_0, x_1, \ldots, x_{n+1}$, удовлетворяющие условиям теоремы, называются точками *чебышевского альтернанса.*

Пример. Функция $f \in C[a, b]$ приближается многочленом нулевой степени. Пусть

$$\max_{[a,b]} f(x) = f(x_1) = M, \qquad \min_{[a,b]} f(x) = f(x_2) = m.$$

Многочлен $P_0(x) = (M+m)/2$ является многочленом наилучшего равномерного приближения (рис. 3), x_1, x_2 — точками чебышевского альтернанса, причем $E_0(f) = (M-m)/2$.

Пример. Выпуклая дифференцируемая на $[a, b]$ функция f приближается многочленом $P_1(x) = a_0 + a_1x$. Вследствие выпуклости f разность $f(x) - (a_0 + a_1x)$ может иметь только одну точку экстремума на интервале (a, b). Поэтому точки a, b являются точками чебышевского альтернанса.

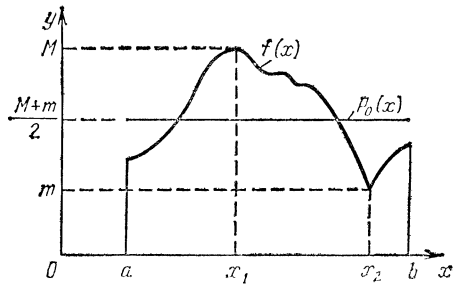


Рис. 3

Пусть $c \in (a, b)$ — третья точка альтернанса. Согласно теореме Чебышева должны выполняться условия

$$f(a) - (a_0 + a_1a) = R, \quad f(c) - (a_0 + a_1c) = -R,$$
$$f(b) - (a_0 + a_1b) = R,$$

где $R = \sigma \| f - P_1^f \|_{C[a, b]}$ неизвестно. Из первого и третьего уравнений получаем

$$a_1 = (f(b) - f(a))/(b - a).$$

Точка c является точкой экстремума разности $f(x) - (a_0 + a_1x)$ и ищется из уравнения $f'(c) - a_1 = 0$. Далее, из первого и второго уравнений определяются a_0 и $E_1(f) = |R|$.

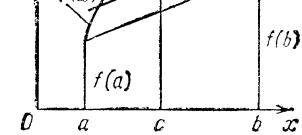


Рис. 4

Геометрически решение данной задачи сводится к следующему (рис. 4). Сначала проводится хорда через точки $(a, f(a))$ и $(b, f(b))$ с наклоном a_1. Затем с этим же наклоном проводится касательная к кривой $y = f(x)$. Наконец, строится секущая посередине между хордой и касательной, дающая среди всех прямых наилучшее равномерное приближение кривой $y = f(x)$ на отрезке $[a, b]$.

Данный геометрический метод очевидным образом распространяется на случай произвольной непрерыв-

ной функции f, график которой расположен по одну сторону от отрезка, соединяющего точки $(a, f(a))$ и $(b, f(b))$.

Построение многочлена наилучшего равномерного приближения n-й степени в общем случае вызывает большие затруднения. Поэтому часто ограничиваются нахождением многочлена, близкого к наилучшему многочлену, или, вообще, некоторого многочлена $P_n(x)$, дающего равномерное приближение заданной функции f с требуемой точностью $\varepsilon > 0$, т. е. удовлетворяющего условию $\|f - P_n\|_{C[a,\,b]} \leqslant \varepsilon$.

Два способа нахождения многочленов, близких к наилучшим.

1. Если $f \in C_{n+1}[a, b]$, причем производная $f^{(n+1)}$ мало изменяется на $[a, b]$, то интерполяционный многочлен $L_n(x)$ (4.5) с чебышевскими узлами (6.11) является близким к наилучшему. Величина уклонения $\|f - L_n\|_{C[a,\,b]}$ оценивается по неравенству (6.12). В частности, если $f^{(n+1)}(x) = \text{const}$, то $L_n(x) \equiv P_n^f(x)$, а если $f(x) = x^{n+1}$ и $[a, b] = [-1, 1]$, то $L_n(x)$ совпадает с многочленом (5).

2. Для функции, разлагающейся в ряд Тейлора, может быть применен метод, который рассмотрим на примере.

Пусть требуется приблизить равномерно функцию $f(x) = \operatorname{arctg} x$ на отрезке $[-\alpha, \alpha]$, где $\alpha = \operatorname{tg}(\pi/8)$, с точностью $0{,}5 \cdot 10^{-5}$. Берем сначала отрезок ряда Тейлора, аппроксимирующий данную функцию с несколько более высокой точностью, чем заданная. Например, полагаем

$$f(x) \approx Q_{11}(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \frac{x^{11}}{11}.$$

При этом

$$\|f - Q_{11}\|_{C[-\alpha,\,\alpha]} < 0{,}82 \cdot 10^{-6}. \tag{6}$$

Подстановкой $x = \alpha t$ переходим к отрезку $[-1, 1]$:

$$S_{11}(t) = Q_{11}(\alpha t) =$$
$$= 0{,}4142136t - 0{,}0236893t^3 + 0{,}0024387t^5 -$$
$$- 0{,}0002989t^7 + 0{,}0000399t^9 - 0{,}0000056t^{11}.$$

Если к данному многочлену прибавить многочлен $0{,}0000056 \cdot 2^{-10} T_{11}(t)$, где $T_{11}(t)$ — многочлен Чебышева:

$$T_{11}(t) = 1024t^{11} - 2816t^9 + 2816t^7 - 1232t^5 + 220t^3 - 11t,$$

то получим многочлен более низкой степени:

$$S_9(t) = S_{11}(t) + 0{,}0000056 \cdot 2^{-10} T_{11}(t) =$$
$$= 0{,}4142135t - 0{,}0236881t^3 + 0{,}0024319t^5 -$$
$$- 0{,}0002835t^7 + 0{,}0000245t^9.$$

Имеем

$$\| S_{11} - S_9 \|_{C[-1,\,1]} = \| 0{,}0000056 \cdot 2^{-10} T_{11} \|_{C[-1,\,1]} =$$
$$= \frac{0{,}0000056}{1024} \| T_{11} \|_{C[-1,\,1]} = \frac{0{,}0000056}{1024} < 0{,}01 \cdot 10^{-6}.$$

Таким образом, многочлен $S_9(t)$ уклоняется на отрезке $[-1, 1]$ от многочлена $S_{11}(t)$ на величину, которая значительно меньше, чем максимальное по модулю значение старшего члена у многочлена $S_{11}(t)$ при $t = \pm 1$. Поэтому данный способ понижения степени многочлена предпочтительнее, чем просто отбрасывание старшего члена.

Поступая аналогично еще два раза, находим

$$S_7(t) = S_9(t) - 0{,}0000245 \cdot 2^{-8} T_9(t) =$$
$$= 0{,}4142127t - 0{,}0236766t^3 +$$
$$+ 0{,}0023906t^5 - 0{,}0002284t^7,$$
$$S_5(t) = S_7(t) + 0{,}0002284 \cdot 2^{-6} T_7(t) =$$
$$= 0{,}4141877t - 0{,}0234768t^3 + 0{,}0019910t^5,$$

где $T_7(t)$, $T_9(t)$ — многочлены Чебышева:

$$T_9(t) = 256t^9 - 576t^7 + 432t^5 - 120t^3 + 9t,$$
$$T_7(t) = 64t^7 - 112t^5 + 56t^3 - 7t.$$

При этом

$$\| S_9 - S_7 \|_{C[-1,\,1]} = 0{,}0000245 \cdot 2^{-8} < 0{,}1 \cdot 10^{-6},$$
$$\| S_7 - S_5 \|_{C[-1,\,1]} = 0{,}0002284 \cdot 2^{-6} < 0{,}36 \cdot 10^{-5}.$$

Возвращаясь к первоначальной переменной x, получаем

$$P_5(x) = S_5(x/\alpha) = \\ = 0{,}9999374x - 0{,}3303433x^3 + 0{,}1632823x^5, \qquad (7)$$

причем

$$\|f - P_5\|_{C[-\alpha, \alpha]} \leqslant \|f - Q_{11}\|_{C[-\alpha, \alpha]} + \\ + \|S_{11} - S_9\|_{C[-1, 1]} + \|S_9 - S_7\|_{C[-1, 1]} + \|S_7 - S_5\|_{C[-1, 1]} < \\ < 0{,}82 \cdot 10^{-6} + 0{,}01 \cdot 10^{-6} + 0{,}1 \cdot 10^{-6} + \\ + 0{,}36 \cdot 10^{-5} < 0{,}46 \cdot 10^{-5}.$$

Итак, найден многочлен (7) пятой степени, аппроксимирующий равномерно функцию $\operatorname{arctg} x$ на отрезке $[-\operatorname{tg}(\pi/8), \operatorname{tg}(\pi/8)]$ с заданной точностью $0{,}5 \cdot 10^{-5}$. Многочлен же Тейлора седьмой степени данную точность не обеспечивает.

Полезна следующая простая теорема, которую приведем без доказательства.

Теорема 2. *Пусть* $f \in C[-1, 1]$,

$$P_{n+1}(x) = a_0 + a_1 x + \ldots + a_{n+1}x^{n+1}, \qquad (8)$$

Тогда

$$E_n(f) \geqslant |a_{n+1}|\, 2^{-n} - \|f - P_{n+1}\|_{C[-1, 1]}. \qquad (9)$$

Из этой теоремы приходим к следующему выводу. Допустим, что найден некоторый многочлен (8), который нас удовлетворяет, т. е. приближает функцию f равномерно на отрезке $[-1, 1]$ с точностью ε, а именно, $\|f - P_{n+1}\|_{C[-1, 1]} \leqslant \varepsilon$. Тогда, если выполняется неравенство

$$|a_{n+1}|\, 2^{-n} - \|f - P_{n+1}\|_{C[-1, 1]} > \varepsilon, \qquad (10)$$

то заведомо не следует пытаться искать многочлен более низкой степени, который обеспечил бы равномерное приближение функции f с точностью ε; он не существует, так как согласно (9), (10) $E_n(f) > \varepsilon$.

Применим теорему 2 в рассмотренном выше примере. Имеем $\varepsilon = 0{,}5 \cdot 10^{-5}$, $n = 4$, $\|f - P_5\|_{C[-\alpha, \alpha]} < 0{,}46 \cdot 10^{-5}$. Старший коэффициент многочлена (7) при переходе к отрезку $[-1, 1]$ совпадает со старшим коэффициентом многочлена S_5, равным 0,0019910. Так как $0{,}0019910 \cdot 2^{-4} - 0{,}46 \cdot 10^{-5} \geqslant 0{,}5 \cdot 10^{-5}$, то неравен-

ство (10) выполнено. Следовательно, даже многочлен наилучшего равномерного приближения четвертой степени не сможет обеспечить заданную точность аппроксимации функции $\operatorname{arctg} x$ на отрезке $[-\alpha, \alpha]$, $\alpha = \operatorname{tg}(\pi/8)$. Таким образом, многочлен (7), удовлетворяющий заданному требованию, имеет минимальную степень.

§ 13. Метод наименьших квадратов

Предварительные сведения. Рассмотрим линейное пространство F действительных функций f, g, ..., заданных на некотором множестве D точек x действительной оси (см. п. 5 введения). Нулевой элемент в F обозначим через θ.

Говорят, что в линейном пространстве F введено *скалярное произведение*, если каждой паре элементов f, $g \in F$ поставлено в соответствие действительное число, обозначаемое (f, g) и называемое скалярным произведением элементов f и g, которое удовлетворяет следующим *аксиомам скалярного произведения*:

а) $(f, g) = (g, f)$;

б) $(f, f) \geqslant 0$, *причем* $(f, f) = 0$ *тогда и только тогда, когда* $f = \theta$, *т. е.* f *является нулевым элементом в* F;

в) $(\alpha f, g) = \alpha (f, g)$, *где* α — *любое действительное число*;

г) $(f_1 + f_2, g) = (f_1, g) + (f_2, g)$.

Линейное пространство F с введенным в нем скалярным произведением (f, g) называется *евклидовым* и обозначается через E.

Для определенности укажем, что нас будут интересовать только два конкретных евклидовых пространства, а именно пространство $E = E_C$ непрерывных на отрезке $[a, b]$ функций ($D = [a, b]$) со скалярным произведением

$$(f, g) = \frac{1}{b-a} \int_a^b f(x) g(x)\, dx \qquad (1)$$

и линейное пространство $E = E_{n+1}$ функций, заданных на конечном (дискретном) множестве точек x_0, x_1, ..., ..., x_n некоторого отрезка $[a, b]$ $\left(D = \{x_i\}_{i=0}^n \subset [a, b]\right)$

со скалярным произведением

$$(f, g)=\frac{1}{n+1}\sum_{i=0}^{n} f(x_i)g(x_i). \qquad (2)$$

Функцию $f\in E_{n+1}$ можно считать $n+1$-мерным вектором $f=(f_0, f_1, \ldots, f_n)$, где $f_i=f(x_i)$. Скалярное произведение (2) отличается от скалярного произведения в $n+1$-мерном евклидовом векторном пространстве, изученном в курсе линейной алгебры, только наличием дополнительного множителя $1/(n+1)$. Поэтому, очевидно, для него выполнены все аксиомы скалярного произведения, проверяемые в линейной алгебре *).

Убедимся, что (1) является скалярным произведением в $C[a, b]$. Остановимся на аксиоме б): для любой функции $f\in C[a, b]$

$$(f, f)=\frac{1}{b-a}\int_a^b f^2(x)\,dx\geqslant 0.$$

Если $f=\theta$, т. е. $f(x)\equiv 0$, то, очевидно, $(f, f)=0$. Предположим теперь, наоборот, что $(f, f)=0$, т. е.

$$\int_a^b f^2(x)\,dx=0. \qquad (3)$$

Допустим, что при этом $f(x)\not\equiv 0$ на $[a, b]$. Тогда в силу непрерывности функции f^2 найдется отрезок $[a', b']\subset[a, b]$, на котором $f^2(x)>0$, и, следовательно,

$$m=\min_{[a', b']} f^2(x)>0,$$

так как непрерывная функция достигает своего минимума на отрезке. Отсюда

$$\int_a^b f^2(x)\,dx\geqslant\int_{a'}^{b'} f^2(x)\,dx\geqslant\int_{a'}^{b'} m\,dx=m(b'-a')>0.$$

Получили противоречие с (3). Поэтому $f=\theta$.

*) См. Бугров Я. С., Никольский С. М. Элементы линейной алгебры и аналитической геометрии. — М.: Наука, 1984. — § 6.

Итак, аксиома б) выполнена. Проверка аксиом в), г) для скалярного произведения (1) легко проводится с помощью известных линейных свойств определенного интеграла, а выполнение аксиомы а) очевидно.

Отметим, что *всякое евклидово пространство одновременно является линейным нормированным пространством* (см. п. 7 введения) *с нормой*

$$\|f\| = (f, f)^{1/2} \tag{4}$$

и, следовательно (см. п. 8 введения), *является метрическим пространством с расстоянием*

$$\rho(f, g) = \|f - g\| = (f - g, f - g)^{1/2}. \tag{5}$$

Для нормы, введенной через скалярное произведение способом (4), выполняются все три аксиомы нормы, сформулированные в п. 7 введения. Аксиома 1) сразу следует из аксиомы б) скалярного произведения. Проверим аксиому 2).

Пусть α — любое действительное число. Имеем

$$\|\alpha f\|^2 = (\alpha f, \alpha f) = \alpha(f, \alpha f) = \alpha(\alpha f, f) = \alpha^2 (f, f) = \alpha^2 \|f\|^2.$$

Отсюда $\|\alpha f\| = |\alpha| \|f\|$, так как $\|\alpha f\| \geqslant 0$. Аксиома 2) выполнена.

Аксиома 3), т. е. неравенство треугольника

$$\|f + g\| \leqslant \|f\| + \|g\|,$$

устанавливается для нормы (4) в произвольном евклидовом пространстве в точности так же, как и неравенство Минковского в n-мерном евклидовом векторном пространстве.

Пусть в евклидовом пространстве E дана система функций $\varphi_0, \varphi_1, \ldots, \varphi_m$.

Определение. Определитель

$$\begin{vmatrix} (\varphi_0, \varphi_0) & (\varphi_1, \varphi_0) & \ldots & (\varphi_m, \varphi_0) \\ (\varphi_0, \varphi_1) & (\varphi_1, \varphi_1) & \ldots & (\varphi_m, \varphi_1) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ (\varphi_0, \varphi_m) & (\varphi_1, \varphi_m) & \ldots & (\varphi_m, \varphi_m) \end{vmatrix}, \tag{6}$$

составленный из скалярных произведений, называется *определителем Грама* системы функций $\varphi_0, \varphi_1, \ldots, \varphi_m$.

Лемма 1. *Определитель Грама* (6) *равен нулю тогда и только тогда, когда система функций* $\varphi_0, \varphi_1, \ldots, \varphi_m$ *линейно зависима* *).

Доказательство. Допустим, что система функций $\varphi_0, \varphi_1, \ldots, \varphi_m$ линейно зависима. Тогда существуют такие не все равные нулю числа $\alpha_0, \alpha_1, \ldots, \alpha_m$, что выполняется равенство

$$\alpha_0\varphi_0 + \alpha_1\varphi_1 + \ldots + \alpha_m\varphi_m = \theta. \tag{7}$$

Умножая это равенство скалярно поочередно на φ_0, φ_1 и т. д. и учитывая, что **) $(\theta, \varphi_j) = 0$, $j=0, 1, \ldots, m$, получаем соотношения

$$\begin{aligned}
&\alpha_0(\varphi_0, \varphi_0) + \alpha_1(\varphi_1, \varphi_0) + \ldots + \alpha_m(\varphi_m, \varphi_0) = 0,\\
&\alpha_0(\varphi_0, \varphi_1) + \alpha_1(\varphi_1, \varphi_1) + \ldots + \alpha_m(\varphi_m, \varphi_1) = 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\alpha_0(\varphi_0, \varphi_m) + \alpha_1(\varphi_1, \varphi_m) + \ldots + \alpha_m(\varphi_m, \varphi_m) = 0,
\end{aligned} \tag{8}$$

которые можно трактовать как однородную систему линейных алгебраических уравнений, имеющую ненулевое решение $\alpha_0, \alpha_1, \ldots, \alpha_m$. Следовательно, определитель системы (8), совпадающий с определителем Грама (6), равен нулю.

Предположим теперь, наоборот, что определитель Грама (6) равен нулю. Докажем, что система функций $\varphi_0, \varphi_1, \ldots, \varphi_m$ линейно зависима. Поскольку определитель системы (8) равен нулю, то она имеет некоторое ненулевое решение $\alpha_0 = \bar{\alpha}_0$, $\alpha_1 = \bar{\alpha}_1, \ldots, \alpha_m = \bar{\alpha}_m$. Подставив в каждое уравнение системы (8) значения $\alpha_j = \bar{\alpha}_j$, $j = 0, 1, \ldots, m$, и пользуясь свойствами (аксиомами) а), в), г) скалярного произведения, получаем следующие равенства:

$$\begin{aligned}
&(\varphi_0, \bar{\alpha}_0\varphi_0 + \bar{\alpha}_1\varphi_1 + \ldots + \bar{\alpha}_m\varphi_m) = 0,\\
&(\varphi_1, \bar{\alpha}_0\varphi_0 + \bar{\alpha}_1\varphi_1 + \ldots + \bar{\alpha}_m\varphi_m) = 0,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&(\varphi_m, \bar{\alpha}_0\varphi_0 + \bar{\alpha}_1\varphi_1 + \ldots + \bar{\alpha}_m\varphi_m) = 0.
\end{aligned} \tag{9}$$

Умножая эти равенства соответственно на $\bar{\alpha}_0, \bar{\alpha}_1, \ldots, \bar{\alpha}_m$ и складывая полученные результаты, на-

*) Определение линейно зависимой и линейно независимой систем функций см. в п. 6 введения.

**) Так как $0 = 0 \cdot f \ \forall f \in E$, то согласно аксиоме в) скалярного произведения имеем $(\theta, f) = (0 \cdot f, f) = 0 \cdot (f, f) = 0 \ \forall f \in E$.

ходим

$$\| \bar{\alpha}_0\varphi_0 + \bar{\alpha}_1\varphi_1 + \ldots + \bar{\alpha}_m\varphi_m \|^2 = 0.$$

Следовательно, по аксиоме 1) нормы имеем

$$\bar{\alpha}_0\varphi_0 + \bar{\alpha}_1\varphi_1 + \ldots + \bar{\alpha}_m\varphi_m = \theta. \qquad (10)$$

Так как не все числа $\bar{\alpha}_j$ равны нулю, то равенство (10) означает, что система функций $\varphi_0, \varphi_1, \ldots, \varphi_n$ линейно зависима. Лемма доказана.

Определение. Система функций $\varphi_0, \varphi_1, \ldots, \varphi_m$ называется *ортогональной*, если

$$(\varphi_j, \varphi_k) = 0, \quad j \neq k, \qquad (\varphi_j, \varphi_j) > 0, \qquad (11)$$

где $0 \leqslant j, k \leqslant m$.

Очевидно, *если система функций* $\varphi_0, \varphi_1, \ldots, \varphi_m$ *ортогональна, то она линейно независима.* Действительно, допустим, что система функций $\varphi_0, \varphi_1, \ldots, \varphi_m$ ортогональна и для нее выполняется равенство (7), где $\alpha_0, \alpha_1, \ldots, \alpha_m$ — некоторые числа. Умножив это равенство скалярно на φ_j, с учетом (11) получим

$$\alpha_j(\varphi_j, \varphi_j) = 0.$$

Отсюда, так как $(\varphi_j, \varphi_j) > 0$. то с необходимостью $\alpha_j = 0$ при любом j, т. е. система $\varphi_0, \varphi_1, \ldots, \varphi_m$ линейно независима.

Многочлен наилучшего среднеквадратичного приближения.

Определение. Функция

$$\Phi_m(x) = c_0\varphi_0(x) + c_1\varphi_1(x) + \ldots + c_m\varphi_m(x), \qquad (12)$$

где $c_0, c_1, \ldots, c_m$ — числовые коэффициенты, называется *обобщенным многочленом* по системе функций $\varphi_0, \varphi_1, \ldots, \varphi_m$.

Пусть произвольная функция f принадлежит E. Ставится задача нахождения такого многочлена (12), что расстояние $\rho(f, \Phi_m)$, называемое также *среднеквадратичным уклонением многочлена* Φ_m от функции f, минимально. Многочлен Φ_m, обладающий указанным свойством, называется *многочленом наилучшего среднеквадратичного приближения* функции f.

Покажем, что *если система функций* $\varphi_0, \varphi_1, \ldots, \varphi_m$ *линейно независима, то для любой функции* $f \in E$ *многочлен наилучшего среднеквадратичного*

приближения существует и притом единственный. Имеем в соответствии с (5) и (12)

$$\rho^2(f, \Phi_m) = \|f - \Phi_m\|^2 = (f - \Phi_m, f - \Phi_m) =$$
$$= (f - c_0\varphi_0 - c_1\varphi_1 - \ldots - c_m\varphi_m, f - c_0\varphi_0 - c_1\varphi_1 - \ldots - c_m\varphi_m) =$$
$$= (f, f) + \sum_{j,k=0}^{m} c_j c_k (\varphi_j, \varphi_k) - 2\sum_{j=0}^{m} c_j (f, \varphi_j). \qquad (13)$$

Таким образом, величина $\rho^2(f, \Phi_m)$ представляет собой квадратичную форму относительно искомых коэффициентов c_j многочлена (12). Поскольку при любых c_j, $j = 0, 1, \ldots, m$, $\rho^2(f, \Phi_m) \geqslant 0$, то из теории квадратичных форм известно, что квадратичная форма (13) достигает своего неотрицательного минимума. Одновременно с $\rho^2(f, \Phi_m)$ минимума достигает и расстояние $\rho(f, \Phi_m) = \sqrt{\rho^2(f, \Phi_m)}$.

Приравняв частные производные формы (13) по c_i, $i = 0, 1, \ldots, m$, к нулю, сократив коэффициент, равный двум, и перенеся вправо свободные члены, приходим к следующей системе линейных алгебраических уравнений:

$$\begin{aligned}
&c_0(\varphi_0, \varphi_0) + c_1(\varphi_1, \varphi_0) + \ldots + c_m(\varphi_m, \varphi_0) = (f, \varphi_0),\\
&c_0(\varphi_0, \varphi_1) + c_1(\varphi_1, \varphi_1) + \ldots + c_m(\varphi_m, \varphi_1) = (f, \varphi_1),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&c_0(\varphi_0, \varphi_m) + c_1(\varphi_1, \varphi_m) + \ldots + c_m(\varphi_m, \varphi_m) = (f, \varphi_m),
\end{aligned} \qquad (14)$$

называемой *нормальной.*

По лемме 1 определитель системы (4), являющийся определителем Грама линейно независимой системы функций $\varphi_0, \varphi_1, \ldots, \varphi_m$, не равен нулю. Поэтому при любой функции $f \in E$ система (14) имеет единственное решение $c_0, c_1, \ldots, c_m$, отвечающее единственной стационарной точке квадратичной формы (13). Эта стационарная точка может быть только точкой минимума, поскольку форма (13) минимума достигает.

Итак, если система $\varphi_0, \varphi_1, \ldots, \varphi_m$ линейно независима, то коэффициенты построенного по ней единственного многочлена наилучшего среднеквадратичного приближения функции $f \in E$ находятся в виде решения нормальной системы линейных алгебраических уравнений (14).

В пространстве E_C непрерывных функций со скалярным произведением (1) расстояние $\rho(f, g)$, называемое *среднеквадратичным расстоянием*, в соответствии с (5) имеет вид

$$\rho(f, g) = \sqrt{\frac{1}{b-a}\int_a^b (f(x) - g(x))^2\,dx}, \qquad (15)$$

а в пространстве E_{n-1} функций, определенных на дискретном множестве $D = \{x_i\}_{i=0}^{n}$, со скалярным произведением (2) среднеквадратичное расстояние задается формулой

$$\rho(f, g) = \sqrt{\frac{1}{n+1}\sum_{i=1}^{n} (f(x_i) - g(x_i))^2}. \qquad (16)$$

Следует иметь в виду, что близость двух непрерывных функций по расстоянию (15), т. е. в смысле среднеквадратичного уклонения, не гарантирует малость их максимального уклонения друг от друга. Например, на рис. 5 изображены график функции $g(x) \equiv 0$, $x \in [a, b]$, и график функции f, отличающийся от графика функции g узким зубцом высоты n и толщины у основания, равной $1/n^3$. Очевидно, в данном случае

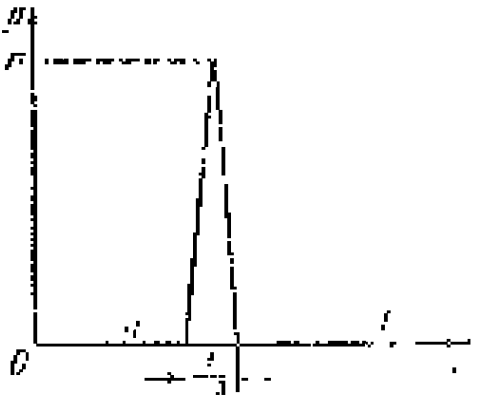


Рис. 5

$$\rho(f, g) = \sqrt{\frac{1}{b-a}\int_a^b (f(x) - g(x))^2\,dx} =$$

$$= \sqrt{\frac{1}{b-a}\int_a^b f^2(x)\,dx} \leqslant \sqrt{\frac{1}{b-a}\,n^2\,\frac{1}{n^3}} = \frac{1}{\sqrt{(b-a)\,n}},$$

т. е. за счет выбора n можно сделать среднеквадратичное расстояние $\rho(f, g)$ сколь угодно малым,

а величину $\max\limits_{[a,\,b]} |f(x)-g(x)| = n$ сколь угодно большой.

Среднеквадратичные приближения алгебраическими многочленами. На практике часто применяются среднеквадратичные приближения функций алгебраическими многочленами, т. е. в качестве системы функций $\varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_m$ берутся степени x: $1, x, x^2, \ldots, x^m$. Система этих функций линейно независима в E_C, т. е. на заданном отрезке $[a, b]$, при любом m. В E_{n+1} она линейно независима, если $m \leqslant n$.

Это следует из того, что при условии $\sum\limits_{j=0}^{m} |\alpha_j| \neq 0$ алгебраический многочлен $\alpha_0 + \alpha_1 x + \ldots + \alpha_m x^m$ может обратиться в нуль не более чем в m точках всей действительной оси. При $m \geqslant n+1$ система $1, x, x^2, \ldots, x^m$ линейно зависима в E_{n+1}, так как многочлен (4.11) $n+1$-й степени со старшим коэффициентом, равным единице, обращается в нуль на всем множестве $D = \{x_i\}_{i=0}^{n}$.

Пример. Построить на отрезке $[0, 1]$ многочлен наилучшего среднеквадратичного приближения $\Phi_1(x) = c_0 + c_1 x$ для функции $f(x) = \sqrt{x}$.

Решение. $\varphi_0(x) = 1$, $\varphi_1(x) = x$, $(\varphi_0, \varphi_0) = \int\limits_0^1 1^2\, dx = 1$,

$$(\varphi_1, \varphi_1) = \int_0^1 x^2\, dx = \frac{1}{3}, \qquad (\varphi_0, \varphi_1) = (\varphi_1, \varphi_0) = \int_0^1 x\, dx = \frac{1}{2},$$

$$(f, \varphi_0) = \int_0^1 \sqrt{x}\, dx = \frac{2}{3}, \quad (f, \varphi_1) = \int_0^1 \sqrt{x}\, x\, dx = \frac{2}{5}.$$

Следовательно, нормальная система уравнений (14) такова:

$$c_0 + \frac{1}{2} c_1 = \frac{2}{3}, \quad \frac{1}{2} c_0 + \frac{1}{3} c_1 = \frac{2}{5}.$$

Отсюда $c_0 = 4/15$, $c_1 = 4/5$, $\Phi_1(x) = 4/15 + (4/5)x$. При этом среднеквадратичное уклонение Φ_1 от f равно

$$\rho(f, \Phi_1) = \sqrt{\int_0^1 \left(\sqrt{x} - \frac{4}{15} - \frac{4}{5} x\right)^2 dx} = \frac{\sqrt{2}}{30}.$$

При нахождении алгебраического многочлена наилучшего среднеквадратичного приближения в E_C для функции f непрерывного аргумента на отрезке $[a, b]$ могут встретиться затруднения в связи с вычислением правых частей уравнений (14), т. е. интегралов

$$(f, \varphi_i) = \frac{1}{b-a} \int_a^b f(x) x^i dx, \quad i = 0, 1, \ldots, m,$$

хотя коэффициенты системы, т. е. скалярные произведения

$$(\varphi_j, \varphi_k) = \frac{1}{b-a} \int_a^b x^{j+k} dx,$$

вычисляются просто.

Поэтому чаще метод наименьших квадратов применяется в дискретном варианте, т. е. в E_{n+1}. Выбирается множество точек $\{x_i\}_{i=0}^n$ на заданном отрезке $[a, b]$, которое может быть предопределено условиями опыта при нахождении значений функции f. Желательно, чтобы число точек было больше степени m многочлена хотя бы в полтора-два раза. Точки на отрезке $[a, b]$ располагаются по возможности равномерно либо несколько сгущаются в той части отрезка, на которой важно получить более точную аппроксимацию функции. В дискретном варианте вычисление коэффициентов и правых частей нормальной системы уравнений (14) с помощью скалярного произведения (2) принципиальных затруднений не вызывает.

Найденный алгебраический многочлен наилучшего среднеквадратичного приближения функции f на дискретном множестве $\{x_i\}_{i=0}^n \subset [a, b]$ в смысле расстояния (16) обычно принимается в качестве некоторого аппроксимирующего многочлена для функции f на всем отрезке $[a, b]$.

Замечание 1. Если $m = n$, то найденный в дискретном варианте методом наименьших квадратов алгебраический многочлен n-й степени совпадает с интерполяционным многочленом, так как уклонение интерполяционного многочлена от заданной функции f на множестве точек $\{x_i\}_{i=0}^n$ в смысле расстояния

(16) в силу (4.1) равно нулю, а меньше быть не может.

Среднеквадратичные приближения функций алгебраическими многочленами используются обычно в тех случаях, когда приближаемая функция не обладает достаточной гладкостью и для нее не удается построить подходящего интерполяционного многочлена (см. § 4—7, 9), сплайна (см. § 11) или многочлена равномерного приближения (см. § 12), а также если значения функции известны в достаточно большом числе точек, но со случайными ошибками. Вопрос влияния случайных ошибок на аппроксимирующий многочлен рассматривается в следующем параграфе.

Применение ортогональных многочленов. Решение нормальной системы уравнений (14) находится наиболее просто, если система функций $\varphi_0, \varphi_1, \ldots, \varphi_m$ ортогональна, т. е. удовлетворяет условиям (11). В этом случае матрица системы (14) становится диагональной и получаемые из нее коэффициенты многочлена (12) наилучшего среднеквадратичного приближения функции f имеют вид

$$c_j = \frac{(f, \varphi_j)}{(\varphi_j, \varphi_j)}, \quad j = 0, 1, \ldots, m. \tag{17}$$

Они называются *коэффициентами Фурье* функции f *по ортогональной системе* $\varphi_0, \varphi_1, \ldots, \varphi_m$.

Из (13) с учетом (4), (11), (17) находим

$$\rho^2(f, \Phi_m) = \|f\|^2 - \sum_{j=0}^{m} c_j^2 \|\varphi_j\|^2, \tag{18}$$

где Φ_m — многочлен наилучшего среднеквадратичного приближения функции f, построенный по ортогональной системе $\varphi_0, \varphi_1, \ldots, \varphi_m$, c_j — его коэффициенты. Отсюда явно видно, что с ростом m, т. е. при добавлении к ортогональной системе $\varphi_0, \varphi_1, \ldots, \varphi_m$ новых функций (без изменения старых), величина $\rho^2(f, \Phi_m)$, вообще говоря, убывает (не возрастает).

В пространстве E_{n+1} ортогональная система $\varphi_0, \varphi_1, \ldots, \varphi_m$ может состоять не более чем из $n+1$-й функции, т. е. $m \leqslant n$. Если $m = n$ и эта система ортогональна, то она, будучи линейно независимой, образует базис в E_{n+1}. Тогда для любой функции $f \in E_{n+1}$ найдется такой многочлен $\Phi_n(x) =$

$= c_0\varphi_0(x) + c_1\varphi_1(x) + \ldots + c_n\varphi_n(x)$ (такая линейная комбинация функций φ_j, $j = 0, 1, \ldots, n$, с коэффициентами, зависящими от f), что $\Phi_n(x) = f(x)$, $x \in \{x_i\}_{i=0}^n$, и, следовательно, $\rho(f, \Phi_m) = 0$. Указанный многочлен, очевидно, совпадает с многочленом наилучшего среднеквадратичного приближения функции $f \in E_{n+1}$ при $m = n$.

На любом отрезке $[a, b]$ существует бесконечная ортогональная в смысле скалярного произведения (1) система алгебраических многочленов. Многочлены, задаваемые выражением

$$X_n(x) = \frac{1}{n!2^n} \frac{d^n}{dx^n}(x^2 - 1)^n, \tag{19}$$

называются *многочленами Лежандра.*

Они обладают свойством ортогональности на стандартном отрезке $[-1, 1]$:

$$(X_j, X_k) = \frac{1}{2} \int_{-1}^{1} X_j(x) X_k(x)\, dx = \begin{cases} 0, & j \neq k, \\ \dfrac{1}{2j+1}, & j = k, \end{cases} \tag{20}$$

которое проверяется интегрированием по частям. Очевидно, $X_n(x)$ является алгебраическим многочленом n-й степени со старшим коэффициентом, не равным нулю, так как при n-кратном дифференцировании многочлена $(x^2 - 1)^n = x^{2n} - nx^{2n-2} + \ldots$ степень понижается в точности на n.

Для многочленов Лежандра справедлива рекуррентная формула

$$(n+1)X_{n+1}(x) - (2n+1)xX_n(x) + nX_{n-1}(x) = 0, \tag{21}$$

по которой с учетом того, что $X_0(x) = 1$, $X_1(x) = x$, может быть найден многочлен Лежандра любой степени. В частности,

$$X_2(x) = \frac{1}{2}(3x^2 - 1), \qquad X_3(x) = \frac{1}{2}(5x^3 - 3x),$$

$$X_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3).$$

Следующая лемма устанавливает важное свойство многочленов Лежандра, которое используется в § 15.

Лемма 2. *Все корни многочлена Лежандра* (19) *действительные, простые и расположены в интервале* $(-1, 1)$.

Доказательство. Прежде всего отметим, что многочлен Лежандра (19) ортогонален любому алгебраическому многочлену $P_k(x)$ степени $k < n$, т. е.

$$\frac{1}{2}\int_{-1}^{1} P_k(x)\,X_n(x)\,dx = 0, \quad k < n. \tag{22}$$

Это утверждение следует из (20) и представления многочлена $P_k(x)$ в виде

$$P_k(x) = \alpha_0 X_0(x) + \alpha_1 X_1(x) + \ldots + \alpha_k X_k(x), \tag{23}$$

где α_0, α_1, ..., α_k — вполне определенные числа. В выражении (23) сначала выбирается α_k, с тем чтобы справа и слева совпали старшие коэффициенты, затем находится α_{k-1} и т. д.

Допустим, что многочлен $X_n(x)$ имеет в интервале $(-1, 1)$ только $k < n$ различных действительных корней нечетной кратности. Обозначим эти корни через $x_1, x_2, \ldots, x_k$ и зададим многочлен k-й степени

$$P_k(x) = \begin{cases} (x - x_1)(x - x_2) \ldots (x - x_k), & k > 0, \\ 1, & k = 0. \end{cases} \tag{24}$$

Очевидно, многочлен $n + k$-й степени, являющийся произведением $P_k(x)X_n(x)$, не изменяет знака в интервале $(-1, 1)$, где у него могут быть корни только четной кратности, и не равен тождественно нулю. Поэтому

$$\int_{-1}^{1} P_k(x)\,X_n(x)\,dx \neq 0,$$

что противоречит равенству (22). Следовательно, у многочлена Лежандра (19) n-й степени в интервале $(-1, 1)$ имеется в точности n простых (однократных) корней.

Пример. Аппроксимировать функцию $f(x) = |x|$ на отрезке $[-1, 1]$ алгебраическим многочленом четвертой степени с помощью метода наименьших квадратов.

Решение. Искомый многочлен $\Phi_4(x)$ представляем через многочлены Лежандра:

$$\Phi_4(x) = c_0 X_0(x) + c_1 X_1(x) + \ldots + c_4 X_4(x).$$

В соответствии с (1), (17), (20) имеем

$$c_j = \frac{2j+1}{2}\int_{-1}^{1} |x|\,X_j(x)\,dx.$$

В частности, $c_0 = 1/2$, $c_1 = 0$, $c_2 = 5/8$, $c_3 = 0$, $c_4 = -3/16$. Таким образом,

$$\Phi_4(x) = \frac{1}{2} + \frac{5}{8\cdot 2}(3x^2 - 1) - \frac{3}{16\cdot 8}(35x^4 - 30x^2 + 3) =$$

$$= \frac{15}{128}(-7x^4 + 14x^2 + 1).$$

Поскольку

$$\|f\|^2 = (f, f) = \frac{1}{2}\int_{-1}^{1} |x|^2\, dx = \frac{1}{3},$$

то согласно (20) и (18), где $\varphi_j = X_j$, получаем следующее значение среднеквадратичного уклонения многочлена $\Phi_4(x)$ от $f(x) = |x|$ на $[-1, 1]$:

$$\rho(f, \Phi_4) = \sqrt{\frac{1}{3} - \left(\frac{1}{2}\right)^2 - \left(\frac{5}{8}\right)^2 \frac{1}{5} - \left(-\frac{3}{16}\right)^2 \frac{1}{9}} =$$

$$= \frac{\sqrt{3}}{48}.$$

Можно доказать, что многочлены Лежандра $X_0(x)$, $X_1(x)$, ..., $X_m(x)$, ... образуют полную систему функций, т. е., какова бы ни была функция $f \in C[-1, 1]$, ее многочлены $\Phi_m(x)$ наилучшего среднеквадратичного приближения, построенные по многочленам Лежандра, сходятся к функции f в среднем, т. е. среднеквадратичное расстояние

$$\rho(f, \Phi_m) = \sqrt{\frac{1}{2}\int_{-1}^{1} (f(x) - \Phi_m(x))^2\, dx}$$

стремится к нулю при $m \to \infty$. Однако сходимость в среднем не гарантирует, что в каждой точке x отрезка $[-1, 1]$ существует равный $f(x)$ предел $\lim\limits_{m\to\infty} \Phi_m(x)$.

Имеются алгебраические многочлены, ортогональные на дискретном множестве точек, т. е. в смысле скалярного произведения (2).

Многочлен $x^{(p)} = x(x-1)\ \ldots\ (x-(p-1))$ называется *факториальным многочленом степени* p, $x^{(0)} = 1$.

Для каждого натурального n определена *система многочленов Чебышева*

$$P_{mn}(x)=\sum_{j=0}^{m}(-1)^j\frac{C_m^jC_{m+j}^j}{n^{(j)}}x^{(j)}, \tag{25}$$

$m=0, 1, \ldots, n$, *ортогональных на дискретном множестве целочисленных точек* $\{x_i\}_{i=0}^n$, $x_i=i$, т. е.

$$(P_{mn}, P_{jn})=\begin{cases}0, & j\neq m,\\ \dfrac{(m+n+1)^{(m+1)}}{n^{(m)}(2m+1)(n+1)}, & j=m,\end{cases} \tag{26}$$

где $(f, g)=\dfrac{1}{n+1}\sum\limits_{i=0}^{n}f(i)g(i)$.

В частности, имеем

$$\begin{aligned}&P_{0n}(x)=1,\\ &P_{1n}(x)=1-2\frac{x}{n},\\ &P_{2n}(x)=1-6\frac{x}{n}+6\frac{x(x-1)}{n(n-1)}.\end{aligned} \tag{27}$$

Среднеквадратичные приближения тригонометрическими многочленами. Функция

$$\Phi_m(x)=a_0+\sum_{p=1}^{m}(a_p\cos px+b_p\sin px), \tag{28}$$

где a_p, b_p — произвольные числовые коэффициенты, называется *тригонометрическим многочленом* порядка m.

Тригонометрическими многочленами естественно приближать периодические функции периода 2π. В теории рядов Фурье устанавливается, что коэффициенты тригонометрического многочлена наилучшего среднеквадратичного приближения непрерывной 2π-периодической функции f, для которого величина

$$\rho(f, \Phi_m)=\sqrt{\frac{1}{2\pi}\int_0^{2\pi}(f(x)-\Phi_m(x))^2dx}$$

принимает минимальное значение, задаются формулами *)

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} f(x)\,dx, \quad a_p = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos px\,dx,$$
$$b_p = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin px\,dx, \quad p > 0, \tag{29}$$

т. е. являются коэффициентами Фурье. Не останавливаясь на этом вопросе подробнее, отметим лишь, что на практике вычисление коэффициентов Фурье (29), задаваемых в виде интегралов, может вызвать значительные затруднения.

Рассмотрим задачу нахождения тригонометрического многочлена наилучшего среднеквадратичного приближения на дискретном множестве точек. Пусть n, m — натуральные, $m \leqslant n/2$,

$$D = \{x_i\}_{i=0}^{n}, \tag{30}$$

где $x_i = \frac{2\pi i}{n+1}$, $i = 0, 1, \ldots, n$. Положим

$$\varphi_0(x) = 1, \quad \varphi_p(x) = \cos px, \quad \psi_p(x) = \sin px, \tag{31}$$
$$p = 1, 2, \ldots, m.$$

Можно доказать, что тригонометрическая система функций (31) ортогональна на дискретном множестве (30) в смысле скалярного произведения (2), а именно,

$$(\varphi_j, \psi_k) = 0, \quad j = 0, 1, \ldots, m, \quad k = 1, 2, \ldots, m,$$
$$(\varphi_r, \varphi_s) = 0, \quad r \neq s, \quad (\psi_t, \psi_q) = 0, \quad t \neq q, \tag{32}$$
$$(\varphi_0, \varphi_0) = 1, \quad (\varphi_p, \varphi_p) = (\psi_p, \psi_p) = 1/2,$$
$$p = 1, 2, \ldots, m \leqslant n/2.$$

Поэтому коэффициенты тригонометрического многочлена наилучшего среднеквадратичного приближе-

*) См. Бугров Я. С., Никольский С. М. Дифференциальные уравнения. Кратные интегралы. Ряды. Функции комплексного переменного. — М.: Наука, 1985. — § 4.3, 4.9.

ния функции f на дискретном множестве (30), имеющего вид

$$\Phi_m(x) = \alpha_0 + \sum_{p=1}^{m} (\alpha_p \cos px + \beta_p \sin px), \qquad (33)$$

находятся согласно (17), (2), (32) по формулам

$$\alpha_0 = \frac{1}{n+1} \sum_{i=0}^{n} f\left(\frac{2\pi i}{n+1}\right), \quad \alpha_p = \frac{2}{n+1} \sum_{i=0}^{n} f\left(\frac{2\pi i}{n+1}\right) \cos p \frac{2\pi i}{n+1},$$

$$\beta_p = \frac{2}{n+1} \sum_{i=0}^{n} f\left(\frac{2\pi i}{n+1}\right) \sin p \frac{2\pi i}{n+1}, \quad p = 1, 2, \ldots, m. \qquad (34)$$

При этом, аналогично (18), получаем

$$\rho(f, \Phi_m) = \sqrt{\frac{1}{n+1} \sum_{i=0}^{n} (f(x_i) - \Phi_m(x_i))^2} =$$

$$= \sqrt{\|f\|^2 - \alpha_0^2 - \frac{1}{2} \sum_{p=1}^{m} (\alpha_p^2 + \beta_p^2)},$$

где $\|f\|^2 = (f, f) = \frac{1}{n+1} \sum_{i=0}^{n} f^2\left(\frac{2\pi i}{n+1}\right)$.

При четном n и при $m = n/2$ ортогональная на D тригонометрическая система (31) состоит из $2m+1 = n+1$ функций и образует базис в E_{n+1}. Поэтому тригонометрический многочлен (33) наилучшего среднеквадратичного приближения функции f в E_{n+1}, коэффициенты которого вычисляются по формулам (34), является при указанных n, m интерполяционным, т. е. $\Phi_{n/2}(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$, и $\rho(f, \Phi_{n/2}) = 0$.

Замечание 2. Как уже отмечалось выше, не всегда полезно, например, когда значения функции известны с погрешностями, аппроксимировать функцию интерполяционным многочленом. Эти погрешности вносят искажения и в многочлен, найденный методом наименьших квадратов, и тем сильнее, чем больше m, т. е. порядок (или степень) многочлена,

а интерполяционному многочлену отвечает максимально возможное m при заданном n. Этот вопрос подробно рассматривается в следующем параграфе.

§ 14. Исследование погрешностей среднеквадратичных приближений. Сглаживание наблюдений

В данном параграфе изучаются погрешности среднеквадратичных приближений, получаемых методом наименьших квадратов в дискретном варианте. Рассматривается влияние случайных ошибок в значениях функций (ошибок результатов наблюдений) и исследуется погрешность метода, возникающая за счет того, что приближаемая функция не принадлежит классу многочленов, которыми осуществляется приближение. Допускается раздельное рассмотрение случайной ошибки и погрешности метода, поскольку оператор построения многочлена наилучшего среднеквадратичного приближения является линейным. В заключение разъясняется процедура сглаживания наблюдений.

Пусть, как и в § 13, E_{n+1} — евклидово пространство функций, заданных на конечном множестве точек $\{x_i\}_{i=0}^{n} \subset [a, b]$, со скалярным произведением (13.2). Предположим, что в E_{n+1} дана ортогональная система функций $\varphi_0, \varphi_1, \ldots, \varphi_m$, т. е. удовлетворяющая условиям (13.11), причем*) $m \leqslant n$. По этой системе функций строится аппроксимирующий многочлен.

Допустим, что вместо значений $f(x_i)$ функции f известны их приближенные значения

$$f_i^* = f(x_i) + \eta_i, \quad i = 0, 1, \ldots, n.$$

Здесь η_i — ошибки наблюдений, являющиеся независимыми случайными величинами с нулевым средним значением и дисперсией, не зависящей от i, т. е., в частности,

$$\mathrm{M}[\eta_i] = 0, \quad i = 0, 1, \ldots, n, \tag{1}$$

$$\mathrm{M}[\eta_i \eta_j] = \begin{cases} 0, & i \neq j, \\ \sigma^2, & i = j, \end{cases} \tag{2}$$

*) Поскольку ортогональные функции линейно независимы, то их число $m+1$ не может превысить размерность пространства E_{n+1}, равную $n+1$.

где $\mathbf{M}[\xi]$ — математическое ожидание случайной величины ξ.

Тогда вместо коэффициентов (13.17) многочлена наилучшего среднеквадратичного приближения функции f, имеющего вид (13.12), будут найдены коэффициенты

$$c_j^* = \frac{(f^*, \varphi_j)}{(\varphi_j, \varphi_j)} = \frac{(f+\eta, \varphi_j)}{(\varphi_j, \varphi_j)} = \frac{(f, \varphi_j)}{(\varphi_j, \varphi_j)} + \frac{(\eta, \varphi_j)}{(\varphi_j, \varphi_j)} = c_j + \gamma_j, \tag{3}$$

где

$$\gamma_j = \frac{(\eta, \varphi_j)}{(\varphi_j, \varphi_j)}, \tag{4}$$

γ_j — случайная ошибка j-го коэффициента.

Отсюда с учетом (13.2), (1) получаем

$$\mathbf{M}[\gamma_j] = \mathbf{M}\left[\frac{(\eta, \varphi_j)}{(\varphi_j, \varphi_j)}\right] = \frac{1}{(\varphi_j, \varphi_j)} \mathbf{M}[(\eta, \varphi_j)] =$$

$$= \frac{1}{(\varphi_j, \varphi_j)} \mathbf{M}\left[\frac{1}{n+1} \sum_{i=0}^{n} \eta_i \varphi_j(x_i)\right] =$$

$$= \frac{1}{(n+1)(\varphi_j, \varphi_j)} \sum_{i=0}^{n} \varphi_j(x_i) \mathbf{M}[\eta_i] = 0. \tag{5}$$

Таким образом, математическое ожидание погрешности каждого коэффициента аппроксимирующего многочлена равно нулю.

Принимая во внимание (4), (13.2), находим

$$\mathbf{M}[\gamma_j \gamma_k] =$$

$$= \frac{1}{(n+1)^2 (\varphi_j, \varphi_j)(\varphi_k, \varphi_k)} \mathbf{M}\left[\sum_{i=0}^{n} \eta_i \varphi_j(x_i) \sum_{q=0}^{n} \eta_q \varphi_k(x_q)\right] =$$

$$= \frac{1}{(n+1)^2 (\varphi_j, \varphi_j)(\varphi_k, \varphi_k)} \sum_{i, q=0}^{n} \varphi_j(x_i) \varphi_k(x_q) \mathbf{M}[\eta_i \eta_q].$$

В силу (2) в двойной сумме могут быть отличными от нуля только слагаемые, отвечающие $q = i$.

Отсюда с учетом (13.11) получаем

$$\mathsf{M}[\gamma_j\gamma_k] = \frac{1}{(n+1)(\varphi_j, \varphi_j)(\varphi_k, \varphi_k)}\left(\frac{1}{n+1}\sum_{i=0}^{n}\varphi_j(x_i)\varphi_k(x_i)\sigma^2\right) = \frac{(\varphi_j, \varphi_k)\sigma^2}{(n+1)(\varphi_j, \varphi_j)(\varphi_k, \varphi_k)} = \begin{cases} 0, & k \neq j, \\ \dfrac{\sigma^2}{(n+1)(\varphi_j, \varphi_j)}, & k = j, \end{cases} \quad (6)$$

т. е. погрешности разных коэффициентов некоррелированы между собой, а дисперсия погрешности j-го коэффициента выражается формулой

$$\mathsf{D}[\gamma_j] = \frac{\sigma^2}{(n+1)(\varphi_j, \varphi_j)}, \quad (7)$$

где σ^2 — дисперсия ошибок наблюдений.

На основании (3) имеем

$$\Phi_m^*(x) = \sum_{j=0}^{m} c_j^*\varphi_j(x) = \sum_{j=0}^{m}(c_j + \gamma_j)\varphi_j(x) = \Phi_m(x) + \Gamma_m(x),$$

где

$$\Gamma_m(x) = \gamma_0\varphi_0(x) + \gamma_1\varphi_1(x) + \ldots + \gamma_m\varphi_m(x), \quad (8)$$

$\Gamma_m(x)$ — случайная ошибка аппроксимирующего многочлена $\Phi_m^*(x)$, найденного методом наименьших квадратов, $\Phi_m(x)$ — многочлен наилучшего среднеквадратичного приближения функции f при отсутствии случайных ошибок.

В (8) аргумент x может принимать значения из конечного множества $\{x_i\}_{i=0}^n$ и любые другие значения, при которых определены функции φ_0, φ_1, ..., φ_m. В частности, если в качестве указанных функций берутся ортогональные на дискретном множестве многочлены Чебышева (13.25), то x может пробегать всю действительную ось.

Принимая во внимание (5)—(8), на основании известных теорем теории вероятностей о том, что математическое ожидание суммы случайных величин равно сумме их математических ожиданий, а дисперсия суммы некоррелированных случайных величин с некоторыми коэффициентами равна сумме их дис-

персий, умноженных на квадраты коэффициентов, получаем

$$\mathbf{M}[\Gamma_m(x)] = \varphi_0(x)\,\mathbf{M}[\gamma_0] + \varphi_1(x)\,\mathbf{M}[\gamma_1] + \ldots$$
$$\ldots + \varphi_m(x)\,\mathbf{M}[\gamma_m] = 0, \quad (9)$$

$$\mathbf{D}[\Gamma_m(x)] = \frac{\sigma^2}{n+1}\sum_{j=0}^{m}\frac{\varphi_j^2(x)}{(\varphi_j,\ \varphi_j)}. \quad (10)$$

Итак, математическое ожидание случайной ошибки $\Gamma_m(x)$ многочлена наилучшего среднеквадратичного приближения равно нулю, а ее дисперсия при любом фиксированном x, вообще говоря, растет с увеличением m, так как при этом в (10) добавляются новые неотрицательные слагаемые.

Замечание 1. Если ошибки наблюдений η_i распределены по нормальному закону, то поскольку согласно (4), (8) ошибка $\Gamma_m(x)$ в аппроксимирующем многочлене зависит линейно от ошибок наблюдений, она тоже при любом фиксированном x распределена нормально.

Периодический случай. Пусть $[a, b] = [0, 2\pi]$, значения 2π-периодической функции f известны в точках $x_i = 2\pi i/(n+1)$, $i = 0, 1, \ldots, n$. Многочлен наилучшего среднеквадратичного приближения функции f на дискретном множестве $\{x_i\}_{i=0}^{n}$, построенный по ортогональной тригонометрической системе функций (13.31), имеет вид (13.33), причем его коэффициенты находятся по формулам (13.34).

Если для построения аппроксимирующего многочлена используются значения функции f с ошибками η_i, являющимися случайными величинами с нулевым средним и обладающими свойством (2), то в соответствии с (10), (13.31), (13.32) имеем при $m \leqslant n/2$

$$\mathbf{D}[\Gamma_m(x)] = \frac{\sigma^2}{n+1}\left(\frac{\varphi_0^2(x)}{(\varphi_0,\ \varphi_0)} + \sum_{p=1}^{m}\left(\frac{\varphi_p^2(x)}{(\varphi_p,\ \varphi_p)} + \frac{\psi_p^2(x)}{(\psi_p,\ \psi_p)}\right)\right) =$$

$$= \frac{\sigma^2}{n+1}\left(1 + 2\sum_{p=1}^{m}(\cos^2 px + \sin^2 px)\right) = \frac{2m+1}{n+1}\,\sigma^2. \quad (11)$$

Таким образом, дисперсия ошибки $\Gamma_m(x)$ аппроксимирующего тригонометрического многочлена по-

рядка m растет линейно от m и не зависит от x, причем согласно (9) $\mathsf{M}[\Gamma_m(x)] = 0$. Среднеквадратичное значение ошибки $\Gamma_m(x)$ обозначим через $\sigma_m(x)$:

$$\sigma_m(x) = \sqrt{\mathsf{D}[\Gamma_m(x)]} = \sigma\sqrt{\frac{2m+1}{n+1}}, \quad m \leqslant \frac{n}{2}, \qquad (12)$$

где σ — среднеквадратичное значение ошибок наблюдений.

Остановимся теперь на искажениях коэффициентов Фурье, возникающих за счет дискретности. Допустим, что непрерывная 2π-периодическая функция f разлагается в равномерно сходящийся ряд Фурье:

$$f(x) = a_0 + \sum_{p=1}^{\infty}(a_p \cos px + b_p \sin px), \qquad (13)$$

где a_0, a_p, b_p — коэффициенты Фурье (13.29).

Тогда можно доказать, что если, например, n — четное, то коэффициенты (13.34) многочлена (13.33) наилучшего среднеквадратичного приближения функции f выражаются через коэффициенты Фурье (13.29) функции f (при отсутствии ошибок наблюдений) следующим образом:

$$\begin{aligned}
\alpha_0 &= a_0 + a_n + a_{2n} + \ldots,\\
\alpha_p &= a_p + a_{n-p} + a_{n+p} + a_{2n-p} + a_{2n+p} + \ldots, \qquad (14)\\
\beta_p &= b_p - b_{n-p} + b_{n+p} - b_{2n-p} + b_{2n+p} - \ldots,\\
&1 \leqslant p \leqslant m \leqslant n/2.
\end{aligned}$$

Если коэффициенты Фурье a_p, b_p убывают быстро, то α_p, β_p при небольших p близки к a_p, b_p, однако с ростом p относительные искажения обычно увеличиваются.

Коэффициенты (13.34) также можно рассматривать как приближения квадратурными формулами прямоугольников (см. § 15) интегралов (13.29) (точнее говоря, равных им интегралов в пределах от π/n до $2\pi + \pi/n$). В предположении, что рассматриваемая периодическая функция f дважды непрерывно дифференцируема на всей действительной оси, из известной формулы $(fg)'' = f''g + 2f'g' + fg''$ легко следуют

неравенства

$$\max_{(-\infty,\ \infty)} |(f(x)\cos px)''| \leqslant M_2 + 2pM_1 + p^2M_0,$$
$$\max_{(-\infty,\ \infty)} |(f(x)\sin px)''| \leqslant M_2 + 2pM_1 + p^2M_0, \tag{15}$$

где $p = 0, 1, \ldots, m$,

$$M_q = \max_{(-\infty,\ \infty)} |f^{(q)}(x)|. \tag{16}$$

На основании (15), (16) и оценки (15.31) погрешности квадратурной формулы прямоугольников следуют оценки

$$|a_0 - \alpha_0| \leqslant \frac{\pi^2}{6n^2} M_2,$$
$$|a_p - \alpha_p| \leqslant \frac{\pi^2}{3n^2}(M_2 + 2pM_1 + p^2M_0), \tag{17}$$
$$|b_p - \beta_p| \leqslant \frac{\pi^2}{3n^2}(M_2 + 2pM_1 + p^2M_0),$$

где $p = 1, 2, \ldots, m \leqslant n/2$.

Остается оценить остаток ряда Фурье, т. е. погрешность, возникающую за счет замены ряда Фурье (13) его частичной суммой порядка m. При условии, что 2π-периодическая функция f непрерывно дифференцируема q раз на действительной оси, применяя q раз интегрирование по частям в интегралах (13.29) и учитывая, что при натуральном p

$$\int_0^{2\pi} |\cos px|\,dx = \int_0^{2\pi} |\sin px|\,dx = 4,$$

получаем оценки

$$|a_p|,\ |b_p| \leqslant \frac{4}{\pi}\frac{M_q}{p^q},\quad p \geqslant 1,$$

где M_q — величина (16).

Отсюда, поскольку $|\cos px| + |\sin px| \leqslant \sqrt{2}$, при $q \geqslant 2$ следует неравенство

$$\left|\sum_{p=m+1}^{\infty} (a_p \cos px + b_p \sin px)\right| \leqslant$$
$$\leqslant \frac{4\sqrt{2}}{\pi} M_q \int_m^{\infty} \frac{dp}{p^q} = \frac{4\sqrt{2}}{\pi}\frac{M_q}{q-1}\frac{1}{m^{q-1}}. \tag{18}$$

Итак, рассмотрены три составляющие погрешности, возникающей при аппроксимации периодической функции тригонометрическим многочленом по методу наименьших квадратов, а именно случайная ошибка, затем погрешность начальных коэффициентов Фурье за счет дискретности и, наконец, остаток ряда Фурье.

Сделаем некоторые выводы. Увеличение n, т. е. числа точек x_i на отрезке $[0, 2\pi]$, влечет уменьшение дисперсии (11) случайной ошибки и способствует уменьшению возникающих за счет дискретности искажений коэффициентов Фурье (во всяком случае правые части оценок (17) стремятся к нулю при $n \to \infty$). Например, если $n = 27$, $m = 3$, то согласно (12) среднеквадратичная погрешность аппроксимирующего многочлена будет в два раза меньше среднеквадратичной ошибки наблюдений. В этом, в частности, состоит сглаживающий эффект метода наименьших квадратов.

Для фиксированного n при выборе $m \leqslant n/2$ следует соблюдать компромисс между случайной ошибкой, растущей согласно (12) при увеличении m, а также погрешностью, вызванной искажениями коэффициентов Фурье, число которых пропорционально m, и погрешностью отбрасывания остатка ряда Фурье, оценка (18) которого улучшается с увеличением m.

Непериодический случай. Рассмотрим среднеквадратичные приближения функций, определенных на множестве $\{i\}_{i=0}^{n}$ целочисленных точек, ортогональными многочленами Чебышева (13.25). Аппроксимирующий многочлен степени $m \leqslant n$ для функции f, найденный методом наименьших квадратов, имеет вид

$$\Phi_m(x) = \sum_{j=0}^{m} c_j P_{jn}(x), \tag{19}$$

где согласно (13.26), (13.17)

$$c_j = \frac{(f,\ P_{jn})}{(P_{jn},\ P_{jn})}, \tag{20}$$

причем $(f,\ g) = \dfrac{1}{n+1} \sum_{i=0}^{n} f(i)\, g(i)$.

Если значения функции f в точках множества $\{i\}_{i=0}^{n}$ известны со случайными ошибками η_i, обла-

дающими свойствами (1), (2), то согласно (9), (10)

$$\mathsf{M}[\Gamma_m(x)]=0, \quad \mathsf{D}[\Gamma_m(x)]=\frac{\sigma^2}{n+1}\sum_{j=0}^{m}\frac{P_{jn}^2(x)}{(P_{jn}, P_{jn})}, \qquad (21)$$

где $\Gamma_m(x)$ — случайная ошибка аппроксимирующего многочлена. В частности, при аппроксимации многочленом первой степени, т. е. при $m=1$, на основании (21), (13.26), (13.27) имеем

$$\mathsf{D}[\Gamma_1(x)]=\frac{\sigma^2}{n+1}\left(1+\frac{12\,(x-n/2)^2}{n\,(n+2)}\right). \qquad (22)$$

Итак, с увеличением m ($m \leqslant n$), т. е. степени аппроксимирующего многочлена (19), при фиксированных n, x дисперсия случайной ошибки, вообще говоря, растет. В отличие от периодического случая, дисперсия случайной ошибки непостоянна, а именно, в середине промежутка наблюдений она значительно меньше, чем ближе к его концам. При $m=1$ это явно видно из формулы (22). Кроме того, для иллюстрации сказанного приведем таблицу значений дисперсии ошибки $\Gamma_m(x)$ при $n=10$, $\sigma^2=1$.

Таблица 3

Значения $\mathsf{D}[\Gamma_m(x)]$

m	0	1	2	3	4	...	10
$x=0;\ 10$	0,09	0,32	0,58	0,79	0,91	...	1,00
$x=5$	0,09	0,09	0,21	0,21	0,33	...	1,00

Замечание 2. Если наблюдения производятся с положительным шагом $h \neq 1$, т. е. в точках $x=0, h, 2h, \ldots, nh$, то всюду многочлены Чебышева $P_{mn}(x)$ заменяются на $P_{mn}(x/h)$ и, следовательно, в формуле (22) справа x заменяется на x/h. Скалярное произведение при этом задается так:

$$(f, g)=\frac{1}{n+1}\sum_{i=0}^{n} f(ih)\,g(ih).$$

Рассмотрим погрешность метода. Пусть ошибок наблюдений нет, но функция f не является алгебраическим многочленом m-й степени. Предположим, что промежуток наблюдений $[0, nh]$ невелик. Тогда, если

$f \in C_{m+1}[0, nh]$, в первом приближении можно считать, что функция f есть алгебраический многочлен степени $m+1$ со старшим коэффициентом, равным $f^{(m+1)}(x_{\text{ср}})/(m+1)!$, где $x_{\text{ср}} = nh/2$.

Допустим для определенности, что функция f является многочленом степени $m+1$ с указанным старшим коэффициентом, причем $m < n$. Тогда метод наименьших квадратов приведет к тождествам

$$f(x) \equiv \Phi_{m+1}(x) \equiv c_{m+1} P_{m+1,\, n}(x/h) + \Phi_m(x), \quad (23)$$

где $\Phi_m(x)$ имеет выражение (19), в котором справа вместо $P_{jn}(x)$ стоит $P_{jn}(x/h)$.

Первое тождество справедливо в силу того, что любой многочлен степени $m+1$ точно восстанавливается методом наименьших квадратов, если степень аппроксимирующего многочлена равна $m+1$, причем $m+1 \leqslant n$. Второе тождество выражает тот факт, что $\Phi_{m+1}(x)$ получается согласно (19) из $\Phi_m(x)$ добавлением одного слагаемого. Таким образом, погрешность аппроксимации многочлена $f(x)$ степени $m+1$ методом наименьших квадратов многочленом степени m имеет выражение

$$f(x) - \Phi_m(x) = c_{m+1} P_{m+1,\, n}(x/h), \quad (24)$$

причем коэффициент c_{m+1} может быть найден как по формуле (20) с учетом замечания 2, так и из условия совпадения старших коэффициентов в левой и правой частях тождеств (23). Принимая во внимание (13.25), вторым способом находим

$$c_{m+1} = (-1)^{m+1} h^{m+1} \frac{(m+1)!}{(2m+2)!}\, n^{(m+1)} f^{(m+1)}(x_{\text{ср}}). \quad (25)$$

В частности, если

$$f(x) = ax^2/2 + bx + c \quad (f'' = a > 0),$$

$m = 1$, то в силу (24), (25), (13.27)

$$f(x) - \Phi_1(x) = ah^2 \frac{n(n-1)}{12} \left(1 - \frac{6}{n}\frac{x}{h} + \right.$$

$$\left. + \frac{6}{n(n-1)} \frac{x}{h}\left(\frac{x}{h} - 1\right)\right),$$

причем при $n \gg 1$

$$f(x) - \Phi_1(x) \approx a\frac{l^2}{12}\left(1 - 6\frac{x}{l}\left(1 - \frac{x}{l}\right)\right),$$

где $l = nh$ — длина промежутка наблюдений. График многочлена, стоящего справа, изображен на рис. 6.

В рассмотренном примере при $m = 1$ погрешность метода, так же как и случайная ошибка, в середине промежутка наблюдений меньше, чем на его краях. Аналогичное явление имеет место и при других m.

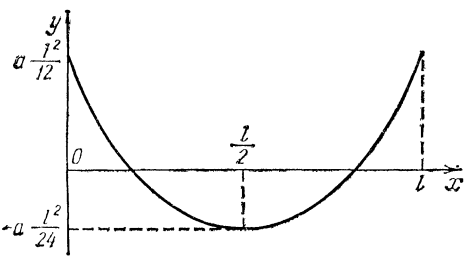


Рис. 6

В целом из рассмотрения случайной ошибки и погрешности метода в непериодическом случае приходим к следующим выводам. Моменты наблюдений (точки измерений) целесообразно выбирать так, чтобы наиболее интересный участок оказался ближе к середине промежутка наблюдений, где меньше как среднеквадратичная случайная ошибка аппроксимирующего многочлена, так и обычно погрешность метода. При постановке наблюдений и выборе степени m аппроксимирующего многочлена следует соблюдать компромисс случайной ошибки и погрешности метода. Среднеквадратичное значение случайной ошибки растет с увеличением m ($m \leqslant n$), но убывает с увеличением n. Погрешность метода обычно растет с увеличением nh, но убывает с увеличением m и уменьшением h при фиксированном n.

Сглаживание наблюдений. Из формул (19), (20) видно, что значение аппроксимирующего многочлена в каждой фиксированной точке x линейно выражается через наблюдаемые значения. Пусть $n = 6$, $m = 3$. Обозначим через $\bar{f}_i$ значение многочлена $\Phi_3(x)$ в средней точке промежутка наблюдений, а че-

рез f_{i-3}, f_{i-2}, f_{i-1}, f_i, f_{i+1}, f_{i+2}, f_{i+3} — используемые наблюдаемые значения, число которых равно $n+1=7$. Тогда

$$\bar{f}_i = \frac{1}{21}(-2f_{i-3}+3f_{i-2}+6f_{i-1}+7f_i+6f_{i+1}+ \\ +3f_{i+2}-2f_{i+3}). \quad (26)$$

Сглаживание состоит в том, что таблица наблюдений f_i с постоянным шагом заменяется на таблицу величин $\bar{f}_i$, вычисленных, например, по формуле (26). Эта таблица предполагается длинной, т. е. содержащей значительно больше, чем $n+1$ значений. На краях таблицы для сглаживания используются другие формулы, получаемые из выражения многочлена $\Phi_3(x)$, который построен по соответствующим крайним узлам.

Для сглаживания, кроме формулы (26), могут быть использованы формулы, получаемые из выражений многочленов (19) при других значениях n, m. Приведем еще несколько формул сглаживания.

$n=2k$, $m=1$:

$$\bar{f}_i = \frac{1}{n+1}\left(f_i + \sum_{j=1}^{k}(f_{i-j}+f_{i+j})\right);$$

$n=4$, $m=3$:

$$\bar{f}_i = \frac{1}{35}(-3f_{i-2}+12f_{i-1}+17f_i+12f_{i+1}-3f_{i+2});$$

$n=8$, $m=5$:

$$\bar{f}_i = \frac{1}{429}(15f_{i-4}-55f_{i-3}+30f_{i-2}+135f_{i-1}+ \\ +179f_i+135f_{i+1}+30f_{i+2}-55f_{i+3}+15f_{i+4}).$$

Кривая, построенная по сглаженным значениям, становится более плавной в силу следующих причин:

1. Если наблюдаемые значения содержат случайные ошибки со свойствами, аналогичными (1), (2), то дисперсия погрешности в сглаженных значениях уменьшается. В частности, при использовании формулы (26) имеем

$$\bar{\sigma}^2 = \frac{1}{3}\sigma^2 \qquad \left(\bar{\sigma} = \frac{1}{\sqrt{3}}\sigma\right).$$

где $\bar{\sigma}$ — среднеквадратичное значение погрешности сглаженных значений, σ — среднеквадратичная ошибка наблюдений.

2. Погрешности в сглаженных значениях $\bar{f}_i$ становятся сильно коррелированными. При применении формулы (26) и при тех же допущениях относительно ошибок наблюдений, что и выше, коэффициенты корреляции сглаженных значений следующие:

$$r_{i,\,i\pm1}=0{,}73, \qquad r_{i,\,i\pm2}=0{,}37, \qquad r_{i,\,i\pm3}=0{,}05,$$
$$r_{i,\,i\pm4}=-0{,}10, \qquad r_{i,\,i\pm5}=-0{,}08, \qquad r_{i,\,i\pm6}=0{,}03.$$

3. Если функция f не является алгебраическим многочленом степени m, то происходит «размазывание» ее графика.

Повторное сглаживание не рекомендуется.

ГЛАВА 2

ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

На практике в редких случаях удается вычислить точно определенный интеграл или проинтегрировать обыкновенное дифференциальное уравнение. Например, в элементарных функциях не выражается интеграл $\int_1^2 \frac{dx}{\ln x}$ и не интегрируется уравнение $u' = e^{-(x^2+u)}$.

В § 15 детально изучаются широко используемые для приближенного вычисления определенных интегралов квадратурные формулы прямоугольников, трапеций и Симпсона, а также рассматриваются квадратурные формулы Гаусса, являющиеся точными для алгебраических многочленов наивысшей степени.

В § 16 обоснованно вводится правило Рунге практической оценки погрешности, в частности, квадратурных формул и приводится метод уточнения по Ричардсону приближенных решений.

Методу Монте-Карло приближенного вычисления определенных интегралов посвящен § 17.

В § 18 излагаются приближенные численные методы решения задачи Коши для обыкновенного дифференциального уравнения $u' = f(x, u)$. Подробно исследуется простейший метод Эйлера, приводятся методы Рунге — Кутта, а также кратко дается метод Адамса. Проводится сравнение методов Рунге — Кутта и Адамса.

§ 15. Квадратурные формулы

Сначала сформулируем теорему интегрального исчисления, которую будем использовать в дальнейшем.

Теорема 1 (обобщенная теорема о среднем). *Пусть* $f, g \in C[a, b]$, *причем* $g(x) \geqslant 0$ *на* $[a, b]$. *Тогда существует такая точка* $\xi \in [a, b]$, *что*

$$\int_a^b f(x)\, g(x)\, dx = f(\xi) \int_a^b g(x)\, dx.$$

Доказательство. Положим

$$M = \max_{[a,\ b]} f(x), \quad m = \min_{[a,\ b]} f(x). \tag{1}$$

Тогда, так как $g(x) \geqslant 0$, то

$$mg(x) \leqslant f(x)\,g(x) \leqslant Mg(x), \qquad x \in [a,\ b],$$

и, следовательно,

$$m \int_a^b g(x)\,dx \leqslant \int_a^b f(x)\,g(x)\,dx \leqslant M \int_a^b g(x)\,dx.$$

Отсюда, поскольку $\int_a^b g(x)\,dx \geqslant 0$, $M \geqslant m$, вытекает существование такого числа c, удовлетворяющего неравенствам

$$m \leqslant c \leqslant M, \tag{2}$$

что

$$\int_a^b f(x)\,g(x)\,dx = c \int_a^b g(x)\,dx. \tag{3}$$

По теореме о промежуточных значениях непрерывной функции в силу (1), (2) найдется точка $\xi \in [a, b]$, в которой $f(\xi) = c$, что вместе с равенством (3) доказывает теорему.

Введем понятие квадратурной формулы. Пусть дан определенный интеграл

$$I = \int_a^b f(x)\,dx \tag{4}$$

от непрерывной на отрезке $[a, b]$ функции f. Приближенное равенство

$$\int_a^b f(x)\,dx \approx \sum_{j=1}^{n} q_j f(x_j), \tag{5}$$

где q_j — некоторые числа, x_j — некоторые точки отрезка $[a, b]$, называется *квадратурной формулой*, определяемой *весами* q_j и *узлами* x_j.

Говорят, что квадратурная формула *точна* для многочленов степени m, если при замене f на произ-

вольный алгебраический многочлен степени m приближенное равенство (5) становится точным.

Рассмотрим сначала наиболее простые квадратурные формулы.

Формула прямоугольников. Допустим, что $f \in C_2[-h/2, h/2]$, $h > 0$. Положим приближенно

$$\int_{-h/2}^{h/2} f(x)\,dx \approx hf_0, \tag{6}$$

где $f_0 = f(0)$, т. е. площадь криволинейной трапеции, ограниченной сверху графиком функции f, аппроксимируется площадью заштрихованного прямоугольника (рис. 7), высота которого равна значению f в средней точке основания трапеции. Найдем остаточный член, т. е. погрешность формулы (6).

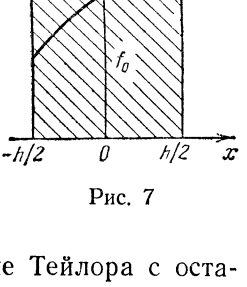


Рис. 7

Пусть

$$F(x) = \int_0^x f(t)\,dt, \tag{7}$$

$F_{\pm 1/2} = F(\pm h/2)$. Так как $F(0) = 0$, $F'(0) = f_0$, $F''(0) = f_0' = f'(0)$, $F'''(x) = f''(x)$, то согласно формуле Тейлора с остаточным членом в форме Лагранжа имеем

$$F_{\pm 1/2} = \pm \frac{h}{2} f_0 + \frac{h^2}{8} f_0' \pm \frac{h^3}{48} f''(\xi_\pm), \tag{8}$$

где ξ_-, ξ_+ — некоторые точки, $-h/2 < \xi_- < \xi_+ < h/2$.

Функция (7) является первообразной для $f(x)$. Поэтому для интеграла, стоящего в левой части приближенного равенства (6), из формулы Ньютона — Лейбница с учетом (8) вытекает следующее выражение:

$$\int_{-h/2}^{h/2} f(x)\,dx = F_{1/2} - F_{-1/2} = hf_0 + \frac{h^3}{24} \frac{f''(\xi_-) + f''(\xi_+)}{2}.$$

Отсюда с помощью леммы 10.1 получаем *формулу прямоугольников с остаточным членом*:

$$\int_{-h/2}^{h/2} f(x)\,dx = hf_0 + \frac{h^3}{24} f''(\xi), \quad |\xi| \leqslant \frac{h}{2}. \tag{9}$$

Формула трапеций. Пусть $f \in C_2[0, h]$. Полагаем

$$\int_0^h f(x)\,dx \approx h\,\frac{f_0 + f_1}{2}, \tag{10}$$

где $f_0 = f(0)$, $f_1 = f(h)$, т. е. интеграл $\int_0^h f(x)\,dx$ приближенно заменяется площадью заштрихованной трапеции, показанной на рис. 8. Выразим f_1 и $F_1 = F(h)$, где F — функция (7), по формуле Тейлора с остаточным членом в интегральной форме:

$$f_1 = f_0 + hf_0' + \int_0^h (h - t) f''(t)\,dt, \tag{11}$$

$$F_1 = F(0) + hF'(0) + \frac{h^2}{2} F''(0) + \frac{1}{2}\int_0^h (h - t)^2 F'''(t)\,dt =$$

$$= hf_0 + \frac{h^2}{2} f_0' + \frac{1}{2}\int_0^h (h - t)^2 f''(t)\,dt. \tag{12}$$

Согласно (11) имеем

$$h\,\frac{f_0}{2} = h\,\frac{f_1}{2} - \frac{h^2}{2} f_0' - \frac{h}{2}\int_0^h (h - t) f''(t)\,dt. \tag{13}$$

Отделив в правой части (12) слагаемое $hf_0/2$ и заменив его выражением (13), с учетом того, что $F_1 = \int_0^h f(x)\,dx$, находим

$$\int_0^h f(x)\,dx = h\,\frac{f_0 + f_1}{2} - \frac{1}{2}\int_0^h (h - t)\,t f''(t)\,dt.$$

Так как $(h - t)\,t \geqslant 0$, $t \in [0, h]$, то по теореме 1

$$\int_0^h (h - t)\,t f''(t)\,dt = f''(\xi)\int_0^h (h - t)\,t\,dt = \frac{h^3}{6} f''(\xi),$$

где $\xi \in [0, h]$ — некоторая точка.

Таким образом, мы пришли к *формуле трапеций с остаточным членом*:

$$\int_0^h f(x)\,dx = h\frac{f_0+f_1}{2} - \frac{h^3}{12}f''(\xi), \quad \xi \in [0,\ h]. \quad (14)$$

Формула Симпсона. Предположим, что $f \in C_4[-h,\ h]$. Интеграл $\int_{-h}^{h} f(x)\,dx$ приближенно заменяем площадью заштрихованной криволинейной трапеции (рис. 9), ограниченной сверху параболой,

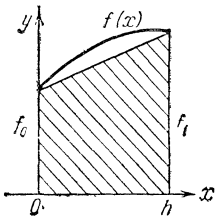


Рис. 8

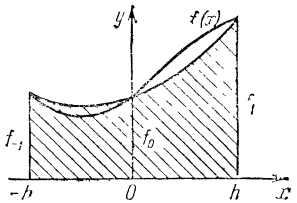


Рис. 9

проходящей через точки $(-h,\ f_{-1})$, $(0,\ f_0)$, $(h,\ f_1)$, где $f_i = f(ih)$. Указанная парабола задается уравнением

$$y = f_0 + \frac{f_1 - f_{-1}}{2h}x + \frac{f_{-1} - 2f_0 + f_1}{2h^2}x^2,$$

в чем нетрудно убедиться, положив поочередно x равным $-h$, 0, h. Отсюда легко находим

$$\int_{-h}^{h} y\,dx = \frac{h}{3}(f_{-1} + 4f_0 + f_1).$$

Таким образом, *формула Симпсона*, называемая также *формулой парабол*, имеет вид

$$\int_{-h}^{h} f(x)\,dx \approx \frac{h}{3}(f_{-1} + 4f_0 + f_1). \quad (15)$$

Положим $F_{\pm 1} = F(\pm h)$, где F — функция (7). Поскольку

$$F(0) = 0,\ F^{(k)}(x) = f^{(k-1)}(x),\ 1 \leqslant k \leqslant 5,$$

то согласно формуле Тейлора с остаточным членом в интегральной форме имеем

$$F_{\pm 1} = \pm h f_0 + \frac{h^2}{2} f_0' \pm \frac{h^3}{6} f_0'' + \frac{h^4}{24} f_0''' \pm$$

$$\pm \frac{1}{24} \int_0^h (h-t)^4 f^{(4)}(\pm t)\, dt,$$

$$f_{\pm 1} = f_0 \pm h f_0' + \frac{h^2}{2} f_0'' \pm \frac{h^3}{6} f_0''' +$$

$$+ \frac{1}{6} \int_0^h (h-t)^3 f^{(4)}(\pm t)\, dt.$$

Отсюда получаем

$$F_1 - F_{-1} - \frac{h}{3}(f_{-1} + 4f_0 + f_1) =$$

$$= -\frac{1}{24} \int_0^h (h-t)^3 \left(\frac{h}{3} + t\right)\left(f^{(4)}(t) + f^{(4)}(-t)\right) dt \qquad (16)$$

(остальные члены взаимно уничтожаются). Поскольку $(h-t)^3(h/3+t) \geqslant 0$, $t \in [0, h]$, то, применяя к интегралу (16) теорему 1, а затем к полученному результату лемму 10.1, находим

$$-\frac{1}{24} \int_0^h (h-t)^3 \left(\frac{h}{3} + t\right)\left(f^{(4)}(t) + f^{(4)}(-t)\right) dt =$$

$$= -\frac{1}{12} \frac{f^{(4)}(\eta) + f^{(4)}(-\eta)}{2} \int_0^h (h-t)^3 \left(\frac{h}{3} + t\right) dt =$$

$$= -\frac{h^5}{90} f^{(4)}(\xi), \qquad (17)$$

где $\eta \in [0, h]$, $\xi \in [-h, h]$ — некоторые точки.

Принимая во внимание, что $F_1 - F_{-1} = \int_{-h}^{h} f(x)\, dx$, из (16), (17) приходим к *формуле Симпсона с остаточным членом*:

$$\int_{-h}^{h} f(x)\, dx = \frac{h}{3}(f_{-1} + 4f_0 + f_1) - \frac{h^5}{90} f^{(4)}(\xi). \qquad (18)$$

Рассмотренные квадратурные формулы прямоугольников (6), трапеций (10) и Симпсона (15) назовем *каноническими.*

Усложненные квадратурные формулы. На практике, если требуется вычислить приближенно интеграл (1), обычно делят заданный отрезок $[a, b]$ на N равных частичных отрезков, на каждом частичном отрезке применяют какую-либо одну каноническую квадратурную формулу и суммируют полученные результаты. Построенная таким путем квадратурная формула на отрезке $[a, b]$ называется *усложненной.* При применении формул прямоугольников и трапеций длину частичных отрезков удобно принять за h, а при использовании формулы Симпсона — за $2h$.

Остановимся подробнее на применении формулы прямоугольников. Пусть $f \in C_2[a, b]$. Обозначим частичные отрезки через $[x_i, x_{i+1}]$, где $x_i = a + ih$, $i = 0, 1, \ldots, N-1$, $x_N = b$, $h = (b-a)/N$. В соответствии с (6) полагаем

$$\int_{x_i}^{x_{i+1}} f(x)\,dx \approx hf_{i+1/2}, \tag{19}$$

где $f_{i+1/2} = f(a + (i + 1/2)h)$ — значение f в середине частичного отрезка $[x_i, x_{i+1}]$. При этом справедливо аналогичное (9) равенство

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = hf_{i+1/2} + \frac{h^3}{24} f''(\xi_i), \tag{20}$$

где $\xi_i \in [x_i, x_{i+1}]$ — некоторая точка.

Суммирование по всем частичным отрезкам приближенного равенства (19) приводит к *усложненной квадратурной формуле прямоугольников*:

$$\int_a^b f(x)\,dx \approx h(f_{1/2} + f_{3/2} + \ldots + f_{N-1/2}), \tag{21}$$

а суммирование равенств (20) с учетом того, что по лемме 10.1

$$\sum_{i=0}^{N-1} f''(\xi_i) = N\left(\frac{1}{N}\sum_{i=0}^{N-1} f''(\xi_i)\right) = Nf''(\xi) = \frac{b-a}{h} f''(\xi),$$

где ξ — некоторая точка отрезка $[a, b]$, дает *усложненную формулу прямоугольников с остаточным членом*:

$$\int_a^b f(x)\,dx = h(f_{1/2} + f_{3/2} + \ldots + f_{N-1/2}) + h^2 \frac{b-a}{24} f''(\xi). \qquad (22)$$

Совершенно так же при условии, что $f \in C_2[a, b]$, с использованием формул (10), (14) получается *усложненная квадратурная формула трапеций*:

$$\int_a^b f(x)\,dx \approx h\left(\frac{f_0}{2} + f_1 + f_2 + \ldots + f_{N-1} + \frac{f_N}{2}\right) \qquad (23)$$

и отвечающая ей формула *с остаточным членом*:

$$\int_a^b f(x)\,dx = h\left(\frac{f_0}{2} + f_1 + f_2 + \ldots + f_{N-1} + \frac{f_N}{2}\right) - h^2 \frac{b-a}{12} f''(\xi), \qquad (24)$$

где $f_i = f(a + ih)$, $h = (b-a)/N$, $\xi \in [a, b]$ — некоторая точка.

Пусть теперь $h = (b-a)/(2N)$ и, как обычно, $x_j = a + jh$, $f_j = f(x_j)$. Перепишем каноническую квадратурную формулу Симпсона (15) применительно к отрезку $[x_{2i}, x_{2i+2}]$ длины $2h$:

$$\int_{x_{2i}}^{x_{2i+2}} f(x)\,dx \approx \frac{h}{3}(f_{2i} + 4f_{2i+1} + f_{2i+2}).$$

Суммируя левую и правую части этого соотношения по i от 0 до $N-1$, получаем *усложненную квадратурную формулу Симпсона*:

$$\int_a^b f(x)\,dx \approx \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \ldots + 4f_{2N-1} + f_{2N}) =$$

$$= \frac{h}{3}\left(f_0 + f_{2N} + 4\sum_{i=1}^{N} f_{2i-1} + 2\sum_{i=1}^{N-1} f_{2i}\right). \qquad (25)$$

Соответствующая ей формула *с остаточным членом*, полученная суммированием по частичным отрезкам $[x_{2i}, x_{2i+2}]$ равенств вида (18), при условии, что $f \in C_4[a, b]$, такова:

$$\int_a^b f(x)\,dx = \frac{h}{3}\left(f_0 + f_{2N} + 4\sum_{i=1}^{N} f_{2i-1} + 2\sum_{i=1}^{N-1} f_{2i}\right) - h^4\,\frac{b-a}{180} f^{(4)}(\xi), \qquad (26)$$

где $f_i = f(a + ih)$, $h = (b-a)/(2N)$, $\xi \in [a, b]$.

Для краткости введем обозначения

$$I_h^{\text{пр}} = h\sum_{i=0}^{N-1} f_{i+1/2}, \quad I_h^{\text{тр}} = h\left(\frac{f_0 + f_N}{2} + \sum_{i=1}^{N-1} f_i\right), \qquad (27)$$

где $h = (b-a)/N$, $f_\mu = f(a + \mu h)$, а также положим

$$I_h^{\text{с}} = \frac{h}{3}\left(f_0 + f_{2N} + 4\sum_{i=1}^{N} f_{2i-1} + 2\sum_{i=1}^{N-1} f_{2i}\right), \qquad (28)$$

где $h = (b-a)/(2N)$, $f_i = f(a + ih)$.

Приближенные равенства

$$I = \int_a^b f(x)\,dx \approx I_h^{\text{пр}}, \quad I = \int_a^b f(x)\,dx \approx I_h^{\text{тр}}, \qquad (29)$$

$$I = \int_a^b f(x)\,dx \approx I_h^{\text{С}} \qquad (30)$$

будем называть соответственно *формулами прямоугольников, трапеций* и *формулой Симпсона*, опуская слова «усложненная квадратурная». В качестве параметра в обозначениях $I_h^{\text{пр}}$, $I_h^{\text{тр}}$, $I_h^{\text{С}}$ выбрано h, а не N, поскольку h, будучи шагом, с которым используются значения функции f в рассматриваемых формулах, более характерно.

Из выражений остаточных членов в (22), (24), (26) видно, что формулы (29) прямоугольников и трапеций точны для многочленов первой степени, т. е. для линейных функций, а формула (30) Симпсона точна для многочленов третьей степени (для них остаточный член равен нулю). Погрешность формул (29) имеет второй порядок относительно h (заведомо

не лучше, если f'' непрерывна на $[a, b]$ и не обращается в нуль), а формула Симпсона при соответствующей гладкости f является формулой четвертого порядка точности. Поэтому для функций класса $C_4[a, b]$ при малом h формула Симпсона обычно дает более высокую точность, чем формулы (29).

Погрешность формулы прямоугольников и формулы Симпсона при вычислении интеграла (4) в силу (22), (26) удовлетворяет неравенствам

$$|I - I_h^{\text{пр}}| \leqslant h^2 \frac{b-a}{24} \max_{[a, b]} |f''(x)|, \tag{31}$$

$$|I - I_h^{\text{С}}| \leqslant h^4 \frac{b-a}{180} \max_{[a, b]} |f^{(4)}(x)|. \tag{32}$$

Аналогичное неравенство имеет место и для погрешности формулы трапеций.

Наряду с оценками погрешности сверху полезны оценки снизу. В частности, для погрешности формулы прямоугольников оценка снизу, вытекающая из (22), такова:

$$|I - I_h^{\text{пр}}| \geqslant h^2 \frac{b-a}{24} \min_{[a, b]} |f''(x)|. \tag{33}$$

В качестве примера исследуем погрешности квадратурных формул для интеграла

$$I = \int_0^{1/2} e^{-x^2} dx,$$

который не выражается через элементарные функции и часто встречается в приложениях.

Имеем

$$f(x) = e^{-x^2}, \quad f''(x) = (4x^2 - 2)e^{-x^2},$$
$$f^{(4)}(x) = 4(4x^4 - 12x^2 + 3)e^{-x^2},$$
$$e^{-1/4} \leqslant |f''(x)| \leqslant 2, \quad |f^{(4)}(x)| \leqslant 12$$

на $[0, 1/2]$. Отсюда при $h = 0{,}05$ согласно (31) — (33) получаем

$$0{,}4 \cdot 10^{-4} \leqslant |I - I_h^{\text{пр}}| \leqslant 0{,}11 \cdot 10^{-3},$$
$$|I - I_h^{\text{С}}| \leqslant 0{,}21 \cdot 10^{-6},$$

т. е. верхняя оценка погрешности формулы Симпсона значительно меньше нижней оценки погрешности формулы прямоугольников.

Формулы прямоугольников и трапеций в отдельности уступают при интегрировании гладких функций формуле Симпсона. Однако в паре они обладают ценным качеством, а именно, если f'' не изменяет знака на $[a, b]$, то формулы (29) дают двусторонние приближения для интеграла (4), так как согласно (22), (24) их остаточные члены имеют противоположные знаки. В рассмотренном примере $f''(x) < 0$, $x \in [0, 1/2]$. Поэтому

$$I_h^{\text{тр}} < I < I_h^{\text{пр}}.$$

В данной ситуации естественно положить $I \approx \dfrac{I_h^{\text{пр}} + I_h^{\text{тр}}}{2}$. Тогда

$$\left| I - \frac{I_h^{\text{пр}} + I_h^{\text{тр}}}{2} \right| < \frac{I_h^{\text{пр}} - I_h^{\text{тр}}}{2},$$

т. е. погрешность оценивается через сами приближенные значения интеграла.

Формулы прямоугольников и трапеций дают двусторонние приближения интеграла (4), не только когда f'' не изменяет знака. К этому вопросу мы вернемся в § 16.

Квадратурная формула Гаусса. Между максимальной степенью многочленов, для которых точна квадратурная формула, и порядком точности квадратурной формулы, скажем по отношению к шагу h, с которым используются значения подынтегральной функции, имеется прямая связь. Например, формулы прямоугольников и трапеций (29) точны для многочленов первой степени и обладают вторым порядком точности относительно h. Формула же Симпсона (25), будучи точной для многочленов третьей степени, соответственно имеет при $f \in C_4[a, b]$ четвертый порядок точности.

В связи со сказанным возникает естественная задача о нахождении среди всех квадратурных формул (5) с n узлами квадратурной формулы с таким расположением узлов x_j на отрезке $[a, b]$ и с такими весами q_j, при которых она точна для алгебраических многочленов максимальной степени. Ясно, что эта

степень меньше, чем $2n$, так как при любом выборе узлов x_j и весов q_j, $j = 1, 2, \ldots, n$, многочлен $P_{2n}(x) = (x-x_1)^2(x-x_2)^2 \ldots (x-x_n)^2$ степени $2n$ обладает тем свойством, что

$$\int_a^b P_{2n}(x)\,dx > 0, \text{ но } \sum_{j=1}^{n} q_j P_{2n}(x_j) = 0.$$

Рассмотрим стандартный отрезок $[a, b] = [-1, 1]$. Пусть пока $x_j \in [-1, 1]$, $j = 1, 2, \ldots, n$, — произвольные попарно непересекающиеся узлы. Тогда, если взять веса

$$q_j = \int_{-1}^{1} p_{n-1,\,j}(x)\,dx, \quad j = 1, 2, \ldots, n, \tag{34}$$

где

$$p_{n-1,\,j}(x) = \frac{(x-x_1)\ldots(x-x_{j-1})(x-x_{j+1})\ldots(x-x_n)}{(x_j-x_1)\ldots(x_j-x_{j-1})(x_j-x_{j+1})\ldots(x_j-x_n)}$$

есть лагранжевы коэффициенты для интерполяционного многочлена степени $n-1$, то квадратурная формула будет точна во всяком случае для многочленов степени $n-1$.

Действительно, любой алгебраический многочлен $P_{n-1}(x)$ степени $n-1$ (см. замечание 4.1) точно восстанавливается построенным по нему интерполяционным многочленом $L_{n-1}(x)$ той же степени. Поэтому

$$\int_{-1}^{1} P_{n-1}(x)\,dx =$$

$$= \int_{-1}^{1} L_{n-1}(x)\,dx = \int_{-1}^{1} \sum_{j=1}^{n} p_{n-1,\,j}(x)\,P_{n-1}(x_j)\,dx =$$

$$= \sum_{j=1}^{n} \left(\int_{-1}^{1} p_{n-1,\,j}(x)\,dx \right) P_{n-1}(x_j) = \sum_{j=1}^{n} q_j P_{n-1}(x_j).$$

Возьмем теперь в качестве узлов x_j, $j = 1, 2, \ldots, n$, корни многочлена Лежандра $X_n(x)$ (13.19), которые согласно лемме 13.2 расположены в интервале $(-1, 1)$, а веса q_j найдем по формуле (34). Покажем,

что квадратурная формула

$$\int_{-1}^{1} f(x)\,dx \approx \sum_{j=1}^{n} q_j f(x_j) \tag{35}$$

с выбранными узлами и весами точна для многочленов степени $2n-1$ и тем самым для многочленов максимальной степени.

Пусть $P_{2n-1}(x)$ — произвольный алгебраический многочлен степени $2n-1$. Представим его в виде

$$P_{2n-1}(x) = U_{n-1}(x)X_n(x) + V_{n-1}(x),$$

где $U_{n-1}(x)$, $V_{n-1}(x)$ — многочлены $n-1$-й степени (частное и остаток от деления $P_{2n-1}(x)$ на многочлен Лежандра $X_n(x)$).

В силу ортогональности многочлена Лежандра $X_n(x)$ любому многочлену степени $n-1$ (см. (13.22)) имеем

$$\int_{-1}^{1} P_{2n-1}(x)\,dx = \int_{-1}^{1} U_{n-1}(x)X_n(x)\,dx + \int_{-1}^{1} V_{n-1}(x)\,dx = \int_{-1}^{1} V_{n-1}(x)\,dx. \tag{36}$$

Поскольку $X_n(x_j) = 0$, $j = 1, 2, \ldots, n$, то

$$\sum_{j=1}^{n} q_j P_{2n-1}(x_j) = \sum_{j=1}^{n} q_j U_{n-1}(x_j)X_n(x_j) + \sum_{j=1}^{n} q_j V_{n-1}(x_j) = \sum_{j=1}^{n} q_j V_{n-1}(x_j). \tag{37}$$

Поскольку квадратурная формула (35) заведомо точна для многочленов степени $n-1$, то

$$\int_{-1}^{1} V_{n-1}(x)\,dx = \sum_{j=1}^{n} q_j V_{n-1}(x_j).$$

Отсюда и из (36), (37) приходим к доказываемому равенству

$$\int_{-1}^{1} P_{2n-1}(x)\,dx = \sum_{j=1}^{n} q_j P_{2n-1}(x_j).$$

Квадратурная формула (35) с n указанными узлами и весами, точная для многочленов степени $2n-1$, называется *формулой Гаусса*. Можно доказать, что ее узлы x_j расположены симметрично относительно точки $x=0$, а веса q_j положительны и в симметричных узлах совпадают при любом n. Приведем численные выражения неотрицательных узлов x_j и весов q_j формулы Гаусса для $n=1, 2, 3, 4$ с десятью десятичными знаками после запятой.

n	x_1 q_1	x_2 q_2	n	x_1 q_1	x_2 q_2
1	0,00000 00000 2,00000 00000		3	0,00000 00000 0,88888 88888	0,77459 66692 0,55555 55556
2	0,57735 02692 1,00000 00000		4	0,33998 10436 0,65214 51549	0,86113 63116 0,34785 48451

Некоторым неудобством формулы Гаусса является иррациональность ее узлов и весов в общем случае. Но это окупается тем, что формула Гаусса точна для многочленов более высокой степени, чем другие квадратурные формулы с n узлами, кроме частного случая $n=1$, когда формула Гаусса совпадает с канонической квадратурной формулой прямоугольников (6), где $h=2$.

С помощью рассмотренной квадратурной формулы Гаусса (35) с n узлами на стандартном отрезке $[-1, 1]$, которую назовем *канонической*, можно построить усложненную квадратурную формулу Гаусса на произвольном отрезке $[a, b]$. Для этой цели разобьем отрезок $[a, b]$ на N равных частичных отрезков $[x_k^*, x_{k+1}^*]$, где $x_k^*=a+k(b-a)/N$, $k=0, 1, \ldots, N-1$, $x_N^*=b$. На каждом частичном отрезке $[x_k^*, x_{k+1}^*]$ зададим n узлов:

$$x_{kj}=\frac{x_k^*+x_{k+1}^*}{2}+x_j\frac{b-a}{2N}, \quad j=1, 2, \ldots, n, \quad (38)$$

где x_j — узлы канонической формулы Гаусса (35).

Расположение узлов x_{kj}, $j=1, 2, \ldots, n$, на частичном отрезке $[x_k^*, x_{k+1}^*]$ геометрически подобно расположению узлов x_j канонической формулы Гаусса

на отрезке $[-1, 1]$. В силу этого квадратурная формула

$$\int_{x_k^*}^{x_{k+1}^*} f(x)\,dx \approx \frac{b-a}{2N} \sum_{j=1}^{n} q_j f(x_{kj}), \tag{39}$$

где q_j — веса канонической формулы Гаусса (35), точна для многочленов степени $2n-1$.

Суммируя соотношение (39) по k от 0 до $N-1$, получаем *усложненную квадратурную формулу Гаусса*:

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2N} \sum_{j=1}^{n} q_j \sum_{k=0}^{N-1} f(x_{kj}), \tag{40}$$

тоже точную для многочленов степени $2n-1$. Ее остаточный член R_n при условии, что $f \in C_{2n}[a, b]$, имеет выражение

$$R_n = \frac{(b-a)^{2n+1}}{N^{2n}} \frac{(n!)^4}{((2n)!)^3 (2n+1)} f^{(2n)}(\xi). \tag{41}$$

В частности,

$$R_2 = \frac{(b-a)^5}{4320N^4} f^{(4)}(\xi), \quad R_3 = \frac{(b-a)^7}{2016000N^6} f^{(6)}(\xi). \tag{42}$$

Сравним усложненную квадратурную формулу Гаусса (40) при $n=2$ и усложненную квадратурную формулу Симпсона (25). Поскольку в формуле Симпсона (25) h и N связаны соотношением $h = (b-a)/(2N)$, то остаточный член формулы (25), который обозначим через R_C, может быть представлен согласно (26) в виде

$$R_C = -h^4 \frac{b-a}{180} f^{(4)}(\xi) = -\frac{(b-a)^5}{2880N^4} f^{(4)}(\xi).$$

Таким образом, остаточный член (42) усложненной формулы Гаусса, отвечающей $n=2$, имеет числовой коэффициент в 1,5 раза меньше по абсолютной величине, чем остаточный член формулы Симпсона. Обе формулы точны для многочленов третьей степени. В формуле Симпсона (25) требуется вычислить $2N+1$ значений функции, а при применении форму-

лы (40) с $n = 2$ число используемых значений функции равно $2N$. Все же предпочтение следует отдать формуле Симпсона (25), у которой узлы расположены с постоянным шагом h на отрезке $[a, b]$.

Квадратурную формулу Гаусса, в том числе и усложненную, целесообразно применять при $n > 2$ для приближенного вычисления интегралов от функций, обладающих высокой гладкостью.

§ 16. Правило Рунге практической оценки погрешности

Учет избыточной гладкости подынтегральной функции. Допустим, что $f \in C_4[a, b]$ и для приближенного вычисления интеграла (15.4) применяется формула прямоугольников (15.21). Ее погрешность, выражающаяся через f'', согласно (15.22) есть $O(h^2)$. Верхнюю оценку погрешности (15.31) формулы прямоугольников, вытекающую из соотношения (15.22), улучшить по порядку относительно h в общем случае за счет существования дополнительных производных f''' и $f^{(4)}$ нельзя. Например, если $\min\limits_{[a, b]} |f''(x)| > 0$, то наряду с (15.31) справедлива нижняя оценка погрешности (15.33), имеющая точный второй порядок относительно h.

Наличие у f «лишних» производных f''' и $f^{(4)}$ при использовании соотношения (15.22) ничего не дает (эти производные в нем не учитываются). Однако если $f \in C_4[a, b]$, то можно получить другое, в некотором смысле более содержательное соотношение.

При условии, что $f \in C_4[-h/2, h/2]$, наряду с (15.8) имеем

$$F_{\pm 1/2} = \pm \frac{h}{2} f_0 + \frac{h^2}{8} f_0' \pm \frac{h^3}{48} f_0'' + \frac{h^4}{384} f_0''' \pm \frac{h^5}{3840} f^{(4)}(\eta_\pm),$$

$-h/2 < \eta_- < \eta_+ < h/2$. Отсюда, аналогично (15.9)

$$\int_{-h/2}^{h/2} f(x)\,dx = F_{1/2} - F_{-1/2} = hf_0 + \frac{h^3}{24} f_0'' + \frac{h^5}{1920} f^{(4)}(\eta), \quad |\eta| \leqslant \frac{h}{2}. \tag{1}$$

Таким образом, справедливо равенство

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = hf_{i+1/2} + \frac{h^3}{24} f''_{i+1/2} + \frac{h^5}{1920} f^{(4)}(\eta_i), \quad (2)$$

где $x_i = a + ih$, $h = (b-a)/N$, $f^{(k)}_{i+1/2} = f^{(k)}(a + (i+1/2)h)$ — значение $f^{(k)}$ в середине частичного отрезка $[x_i, x_{i+1}]$, $\eta_i \in [x_i, x_{i+1}]$ — некоторая точка, $i = 0, 1, \ldots, N-1$. Суммируя равенство (2) по i от 0 до $N-1$, аналогично (15.22) получаем

$$I = I_h^{\text{пр}} + \frac{h^2}{24}\left[h\sum_{i=0}^{N-1} f''_{i+1/2}\right] + h^4 \frac{b-a}{1920} f^{(4)}(\eta), \quad (3)$$

где I — интеграл (15.4), $I_h^{\text{пр}}$ имеет выражение (15.27), $\eta \in [a, b]$ — некоторая точка.

Согласно равенству (15.22), в которое вместо f подставлена $f'' \in C_2[a, b]$, имеем

$$\int_a^b f''(x)\,dx = h\sum_{i=0}^{N-1} f''_{i+1/2} + h^2 \frac{b-a}{24} f^{(4)}(\xi), \quad (4)$$

где $\xi \in [a, b]$ — некоторая точка.

Из (3), (4) находим

$$I = I_h^{\text{пр}} + ch^2 + O(h^4), \quad (5)$$

где

$$c = \frac{1}{24}\int_a^b f''(x)\,dx, \quad (6)$$

c — постоянная, не зависящая от h.

Величина ch^2 в (5) называется *главной частью погрешности* формулы прямоугольников. Может случиться, что $c = 0$. Тогда главная часть погрешности равна нулю и погрешность формулы прямоугольников является величиной порядка h^4. Но обычно $c \neq 0$.

Если $f \in C_4[a, b]$, то справедливо также соотношение

$$I = I_h^{\text{тр}} + c_1 h^2 + O(h^4), \quad (7)$$

где $I_h^{\text{тр}}$ определено в (15.27) и является приближенным значением интеграла (15.4), найденным по фор-

муле трапеций с шагом h, $c_1 = -\frac{1}{12}\int_a^b f''(x)\,dx$ не зависит от h.

Замечание 1. Из соотношений (5), (7), в частности, следует, что если $\int_a^b f''(x)\,dx \neq 0$, причем $f \in C_4[a, b]$, то при достаточно малом h формулы (15.29) прямоугольников и трапеций дают двусторонние приближения интеграла (15.4), даже если f'' не сохраняет знака на $[a, b]$.

При условии, что $f \in C_6[a, b]$, аналогично (5) можно получить следующее соотношение:

$$I = I_h^C + ch^4 + O(h^6), \tag{8}$$

где I — интеграл (15.4), I_h^C — его приближенное значение, найденное по формуле Симпсона, т. е. величина (15.28), c — некоторая не зависящая от h постоянная.

Правило Рунге. Пусть z — неизвестное точное значение некоторой величины, z_h — известное ее приближенное значение (приближенное решение), зависящее от положительного параметра h, который может принимать сколь угодно малые значения. Предположим, что установлена связь (соотношение)

$$z = z_h + ch^k + O(h^{k+m}), \tag{9}$$

где c — неизвестная не зависящая от h постоянная, $k, m > 0$ — известные числа. Тогда

$$z = z_{h/2} + c(h/2)^k + O(h^{k+m}). \tag{10}$$

Вычитая равенства (9) и (10), находим

$$z_{h/2} - z_h = c(h/2)^k(2^k - 1) + O(h^{k+m}).$$

Отсюда

$$c\left(\frac{h}{2}\right)^k = \frac{z_{h/2} - z_h}{2^k - 1} + O(h^{k+m}) \tag{11}$$

и, следовательно, согласно (10) с точностью до $O(h^{k+m})$ имеем

$$z - z_{h/2} \approx \frac{z_{h/2} - z_h}{2^k - 1}, \tag{12}$$

где z_h, $z_{h/2}$ — известные величины.

Если $c \neq 0$, то правая часть (12) в силу (11) имеет в точности k-й порядок относительно h и отличается от главной части погрешности, т. е. от $c(h/2)^k$, на величину более высокого порядка относительно h.

Вычисление приближенной оценки погрешности по формуле (12) при выполнении условия (9) называется *правилом Рунге.*

Замечание 2. На практике подтверждением условия $c \neq 0$ является выполнение неравенства

$$\left| 2^k \frac{z_h - z_{h/2}}{z_{2h} - z_h} - 1 \right| < 0{,}1. \tag{13}$$

И только в этом случае рекомендуется применять правило Рунге. Неравенство (13) может не выполняться по следующим причинам: а) h велико, при этом влияет неучитываемый член $O(h^{k+m})$; б) h слишком мало, тогда могут сказаться погрешности округлений в вычислениях на реальной ЭВМ; в) $c = 0$ или близко к нулю.

Уточнение приближенного решения по Ричардсону. Вычитая из умноженного на 2^k равенства (10) равенство (9), получаем

$$z(2^k - 1) = 2^k z_{h/2} - z_h + O(h^{k+m}).$$

Отсюда

$$z = z_h^* + O(h^{k+m}), \tag{14}$$

где

$$z_h^* = \frac{2^k z_{h/2} - z_h}{2^k - 1}. \tag{15}$$

Число z_h^* называется *уточненным* (или *экстраполированным*) *по Ричардсону* приближенным значением величины z.

Согласно (14) $z - z_h^* = O(h^{k+m})$, в то же время, если в (9) $c \neq 0$, $z - z_{h/2}$ имеет в точности k-й порядок относительно h.

Таким образом, при наличии условия (9), где $c \neq 0$, с помощью приближенных решений z_h и $z_{h/2}$ можно, во-первых, приближенно оценить погрешность $z_{h/2}$ по правилу Рунге, т. е. по формуле (12), и, во-вторых, вычислить по формуле (15) приближенное решение z_h^*, имеющее погрешность более высокого порядка относительно h, чем $z_{h/2}$.

Применение правила Рунге к квадратурным формулам. Обозначим через I_h приближенное значение интеграла I (15.4), найденное по одной из трех формул (15.29), (15.30), и объединим соотношения (5), (7), (8) в одно:

$$I = I_h + ch^k + O(h^{k+2}), \qquad (16)$$

где c не зависит от h; $k = 2$ для формул прямоугольников и трапеций, $k = 4$ для формулы Симпсона. Предполагается, что $f \in C_{k+2}[a, b]$.

Соотношение (16) является соотношением вида (9). Поэтому согласно изложенному, вычислив I_{2h}, I_h, $I_{h/2}$ и убедившись, что $2^k(I_h - I_{h/2})/(I_{2h} - I_h)$ близка к единице, можно приблизительно оценить погрешность $I - I_{h/2}$ по правилу Рунге:

$$I - I_{h/2} \approx \frac{I_{h/2} - I_h}{2^k - 1}.$$

Кроме того, возможно найти уточненное по Ричардсону значение интеграла I:

$$I_h^* = \frac{2^k I_{h/2} - I_h}{2^k - 1}$$

с погрешностью $I - I_h^* = O(h^{k+2})$.

Замечания. 3. Формулы трапеций и Симпсона удобны тем, что при переходе от h к $h/2$ все вычисленные значения функции f могут быть использованы.

4. При условии, что $f \in C_{2n+2}[a, b]$, справедливо соотношение (16), где I_h — приближенное значение интеграла (15.4), найденное по усложненной формуле Гаусса (15.40), $k = 2n$, $h = 1/N$.

Другие применения. Правило Рунге оценки погрешности и уточнение приближенного решения по Ричардсону можно применять при решении других задач приближенными методами, если установлено соотношение (9), где c не зависит от h. В частности, это возможно при численном дифференцировании достаточно гладких функций.

Пусть $f \in C_6[x_0 - h_0, x_0 + h_0]$, $h_0 > 0$. Тогда, используя для представления значений функции $f_{\pm 1} = f(x_0 \pm h)$, $0 < h \leqslant h_0$, формулу Тейлора с остаточным членом, выражающимся через $f^{(6)}$, аналогично (10.3) получаем соотношение

$$f_0'' = f_{0h}'' - \frac{h^2}{12} f_0^{(4)} - \frac{h^4}{360} f^{(6)}(\xi), \qquad (17)$$

где $f_0^{(k)} = f^{(k)}(x_0)$, $|\xi - x_0| < h$, f''_{0h} есть приближенное значение f''_0, равное правой части (10.7). Соотношение (17) имеет вид (9), где $k = m = 2$, $c = -f_0^{(4)}/12$ и не зависит от h. Таким образом, основное условие применимости правила Рунге выполнено.

§ 17. Метод Монте-Карло

Метод Монте-Карло состоит в том, что задается случайная величина ξ, математическое ожидание которой равно искомой величине z, т. е.

$$\mathsf{M}[\xi] = z; \tag{1}$$

осуществляется серия n независимых испытаний случайной величины ξ: $\xi_1, \xi_2, \ldots, \xi_n$ и приближенно полагается

$$\bar{\xi}_n = \frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n} \approx z. \tag{2}$$

В силу (1) при любом натуральном n

$$\mathsf{M}[\bar{\xi}_n] = \mathsf{M}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathsf{M}[\xi_i] = \frac{nz}{n} = z. \tag{3}$$

Если дисперсия $\mathsf{D}[\xi] = \sigma^2$ конечна, то

$$\mathsf{D}[\bar{\xi}_n] = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{D}[\xi_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}, \tag{4}$$

причем в силу центральной предельной теоремы распределение случайной величины $\bar{\xi}_n$ асимптотически нормально. Поэтому при достаточно большом n (практически при $n > 10$) согласно (3), (4) и известному правилу «трех сигм» имеем

$$p\left(|z - \bar{\xi}_n| < 3\frac{\sigma}{\sqrt{n}}\right) \approx 0{,}997,$$

т. е. неравенство

$$|z - \bar{\xi}_n| < 3\frac{\sigma}{\sqrt{n}} \tag{5}$$

выполняется с вероятностью, достаточно близкой к единице ($\approx 0{,}997$).

На практике, если σ неизвестна, но во всяком случае конечна, то ее оценивают при $n > 10$ по формуле

$$\sigma \approx \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \bar{\xi}_n)^2}. \tag{6}$$

Вычисление определенных интегралов. Рассмотрим интеграл

$$I = \int_0^1 f(x)\,dx, \tag{7}$$

который может быть и несобственным интегралом, но таким, что интеграл $\int_0^1 f^2(x)\,dx$ тоже существует.

Пусть η — равномерно распределенная на отрезке $[0, 1]$ случайная величина, т. е. имеющая плотность распределения

$$p_\eta(x) = \begin{cases} 1, & 0 \leqslant x \leqslant 1, \\ 0, & x \notin [0, 1]. \end{cases}$$

Тогда $\xi = f(\eta)$ тоже будет некоторой случайной величиной*), причем по определению математического ожидания

$$\mathbf{M}[\xi] = \int_0^1 f(x)\,p_\eta(x)\,dx = \int_0^1 f(x)\,dx = I.$$

Таким образом,

$$I \approx \bar{\xi}_n = \frac{1}{n}\sum_{i=1}^{n} f(\eta_i), \tag{8}$$

где η_i — независимые реализации случайной величины η. Поскольку наряду с (7) по предположению

*) В точках отрезка $[0, 1]$, в окрестности которых функция f неограничена, формально полагаем $f = 0$. Предполагается, что таких точек на $[0, 1]$ конечное число.

существует интеграл $\int_0^1 f^2(x)\,dx$, то

$$\sigma^2 = \mathsf{D}[f(\eta)] = \mathsf{M}[f^2(\eta)] - (\mathsf{M}[f(\eta)])^2 = \int_0^1 f^2(x)\,dx - \left(\int_0^1 f(x)\,dx\right)^2 < \infty. \quad (9)$$

Величину σ оценивают либо по формуле (6), либо исходя из (9), оценивая $\int_0^1 f^2(x)\,dx$ сверху, а $\left(\int_0^1 f(x)\,dx\right)^2$ снизу, например нулем. Погрешность $I - \overline{\xi}_n$ оценивается по формуле (5) с вероятностью $\approx 0{,}997$.

Кратные интегралы методом Монте-Карло вычисляются аналогично. Например,

$$\int_0^1 \int_0^1 f(x, y)\,dx\,dy \approx \frac{1}{n}\sum_{i=1}^{n} f(\eta_i, \theta_i), \quad (10)$$

где η_i, θ_i — независимые реализации равномерно распределенных на $[0, 1]$ случайных величин η, θ.

Случайные величины (случайные числа), равномерно распределенные на отрезке $[0, 1]$, в современных ЭВМ задаются с помощью специальных физических датчиков или программ. При применении программ эти числа называются *псевдослучайными*. Хотя практически они обладают статистическими характеристиками, свойственными случайным величинам, но, строго говоря, они не являются случайными.

Сравнение с квадратурными формулами. Рассмотрим следующую квадратурную формулу для двойного интеграла:

$$\int_0^1 \int_0^1 f(x, y)\,dx\,dy \approx \sum_{i=0}^{N}\sum_{k=0}^{N} q_{ik} f\left(\frac{i}{N}, \frac{k}{N}\right), \quad (11)$$

где q_{ik} — веса. Допустим, что погрешность этой формулы ведет себя приблизительно как $c_1 N^{-m}$ (здесь и ниже c_1, c_2, ... — некоторые постоянные). Такое поведение погрешности квадратурной формулы является характерным, в чем мы убедились в одномерном случае в § 15, 16.

Пусть $\varepsilon > 0$ — заданная точность вычисления интеграла. Полагая $c_1 N^{-m} \approx \varepsilon$, находим требуемое $N \approx \approx c_2 \varepsilon^{-1/m}$. Следовательно, число N_f вычисляемых значений функции f в квадратурной формуле в ν-мерном случае зависит от ε следующим образом:

$$N_f \approx N^\nu \approx c_3 \frac{1}{\varepsilon^{\nu/m}}.$$

В методе же Монте-Карло должно быть

$$3\frac{\sigma}{\sqrt{n}} \approx \varepsilon,$$

т. е. для достижения той же точности с вероятностью, близкой к единице, метод Монте-Карло требует вычисления

$$n \approx c_4 \frac{1}{\varepsilon^2}$$

значений функции f независимо от размерности ν. Например, при $\nu = 15$, $m = 3$ имеем

$$N_f \approx c_3 \frac{1}{\varepsilon^5}, \quad n \approx c_4 \frac{1}{\varepsilon^2}.$$

Таким образом, в многомерном случае при применении метода Монте-Карло число вычисляемых значений подынтегральной функции растет значительно медленнее относительно $1/\varepsilon$ (ε — точность вычисления интеграла), чем в квадратурных формулах. Это свойство выражает первое преимущество метода Монте-Карло.

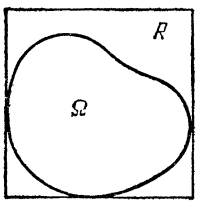


Рис 10

Второе преимущество метода Монте-Карло состоит в том, что его точность не зависит от гладкости подынтегральной функции.

Третье преимущество — простая приспособляемость к форме области интегрирования. Например, полагаем

$$\iint\limits_{\Omega} f(P)\,dP = \iint\limits_{R} f^*(P)\,dP,$$

где R — квадрат, содержащий заданную область Ω (рис. 10),

$$f^*(P)=\begin{cases} f(P), & P\in\Omega, \\ 0, & P\notin\Omega. \end{cases}$$

Далее применяется формула вида (10).

Недостатком метода Монте-Карло является вероятностный характер результата, т. е. отсутствие строгих, стремящихся к нулю при $n\to\infty$ оценок погрешности.

§ 18. Численные методы решения задачи Коши для обыкновенных дифференциальных уравнений

Рассмотрим дифференциальное уравнение первого порядка

$$u'=f(x,\ u). \tag{1}$$

В курсе дифференциальных уравнений устанавливается следующая теорема.

Теорема 1. *Если функция $f(x, u)$ непрерывно дифференцируема m раз по переменным x, и на двумерной замкнутой области $\bar{G}$, $f\in C_m(\bar{G})$*), то всякое решение $u(x)$ уравнения* (1), *расположенное в $\bar{G}$, $m+1$ раз непрерывно дифференцируемо по x.*

На основании теоремы 1 судят о гладкости решений уравнения (1).

Пусть для определенности $\bar{G}$ является замкнутым прямоугольником:

$$\bar{G}=\{(x,u)\colon x_0\leqslant x\leqslant x_0+l,\ a\leqslant u\leqslant b\}.$$

Задача Коши. Найти решение $u(x)$ уравнения (1), удовлетворяющее начальному условию

$$u(x_0)=u_0, \tag{2}$$

где u_0 — заданное число, $a<u_0<b$.

Предположим, что $f\in C_1(\bar{G})$ и существует решение $u(x)$ задачи Коши (1), (2), определенное на отрезке $[x_0, x_0+l]$, причем

$$a<u(x)<b,\ \ x\in[x_0, x_0+l]. \tag{3}$$

---

*) См. п. 3 введения.

Единственность решения задачи Коши вытекает из условия $f \in C_1(\bar{G})$. По теореме 1 $u \in C_2[x_0, x_0 + l]$, причем очевидно

$$u'(x) = f(x, u(x)), \tag{4}$$
$$u''(x) = f'_x(x, u(x)) + f'_u(x, u(x)) f(x, u(x)).$$

Приближенное решение задачи Коши (1), (2) будем искать на конечном множестве точек отрезка $[x_0, x_0 + l]$, которое называется *сеткой*. Выберем сетку $\omega_h = \{x_j\}_{j=0}^{N}$, где $x_j = x_0 + jh$, $h = l/N$, N — натуральное.

Функция φ, заданная на сетке ω_h, называется *сеточной функцией*. Обозначим $\varphi_j = \varphi(x_j)$. Введем в линейном пространстве сеточных функций норму

$$\|\varphi\|_h = \max_{0 \leqslant j \leqslant N} |\varphi_j|. \tag{5}$$

Решение $u(x)$ задачи Коши (1), (2) на отрезке $[x_0, x_0 + l]$ определено, в частности, и на сетке ω_h (полагаем $u_j = u(x_j)$). Под $\|u\|_h$ будем подразумевать норму сеточной функции, совпадающей на ω_h с $u(x)$.

Докажем вспомогательную лемму.

Лемма 1. *Пусть* $\alpha > 0$, $\beta \geqslant 0$, $\varepsilon_0 = 0$,

$$|\varepsilon_{j+1}| \leqslant (1 + \alpha)|\varepsilon_j| + \beta, \quad j = 0, 1, \ldots, N - 1. \tag{6}$$

Тогда при $k = 0, 1, \ldots, N$ *справедливо неравенство*

$$|\varepsilon_k| \leqslant \beta(e^{k\alpha} - 1)/\alpha. \tag{7}$$

Доказательство. При $k = 0$ неравенство (7) очевидно. Допустим, что (7) верно при некотором $k = m$, $0 \leqslant m < N$. Тогда согласно (6)

$$|\varepsilon_{m+1}| \leqslant (1 + \alpha)\beta(e^{m\alpha} - 1)/\alpha + \beta =$$
$$= \beta((1 + \alpha)e^{m\alpha} - 1)/\alpha \leqslant \beta(e^{(m+1)/\alpha} - 1)/\alpha,$$

так как $1 + \alpha < e^{\alpha} = 1 + \alpha + \alpha^2/2 + \ldots$ Отсюда по индукции следует (7) для $k = 0, 1, \ldots, N$.

Метод Эйлера. Приближенное решение y задачи Коши (1), (2) вычисляется на сетке ω_h по формулам

$$y_0 = u_0, \tag{8}$$
$$y_{j+1} = y_j + hf(x_j, y_j) + \eta_j, \quad j = 0, 1, \ldots,$$

где η_j — погрешность округлений, в том числе погрешность вычисления значений функции f. Будем предполагать, что приближенное решение y тоже находится в прямоугольнике $\bar{G}$, т. е.

$$a \leqslant y_j \leqslant b,\ j = 0,\ 1,\ \ldots,\ N. \tag{9}$$

Переход от y_j к y_{j+1} геометрически означает при $\eta_j = 0$ перемещение по касательной в точке (x_j, y_j) к интегральной кривой $\tilde{u}_j(x)$ уравнения (1), проходящей через эту точку на шаг h в направлении оси x

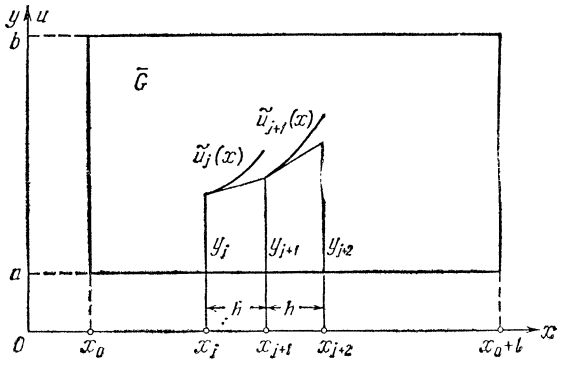


Рис. 11

(рис. 11). На следующем шаге перемещение при отсутствии погрешности округлений происходит по касательной к (вообще говоря) другой интегральной кривой $\tilde{u}_{j+1}(x)$, проходящей через точку (x_{j+1}, y_{j+1}), и т. д. Начальной точкой, с которой начинается построение приближенного решения, является точка (x_0, y_0), где $y_0 = u_0$ — заданное начальное значение неизвестного решения $u(x)$ уравнения (1).

Предположим, что погрешность округлений удовлетворяет условию

$$\max_{0 \leqslant j \leqslant N-1} |\eta_j| \leqslant c_0 h^2, \tag{10}$$

где c_0 не зависит от h. Введем обозначения $\varepsilon_j = u_j - y_j$ (погрешность приближенного решения),

$$M_0 = \max_{\bar{G}} |f(x, u)|, \quad M_1 = \max_{\bar{G}} |f'_x(x, u)|,$$
$$M_2 = \max_{\bar{G}} |f'_u(x, u)|. \tag{11}$$

Будем предполагать, что $M_2 > 0$. Случай $M_2 = 0$, означающий, что $f(x, u)$ не зависит от u, сводится к вычислению интегралов методами, изложенными в § 15.

Для точного решения $u(x)$ задачи Коши (1), (2) согласно формуле Тейлора имеем

$$u_0 = u_0, \quad u_{j+1} = u_j + hf(x_j, u_j) + \frac{h^2}{2} u''(\xi_j), \qquad (12)$$

где $x_j < \xi_j < x_{j+1}$, $j = 0, 1, \ldots, N-1$. Вычитая равенства (12) и (8) и применяя к разности $f(x_j, u_j) - f(x_j, y_j)$ формулу конечных приращений Лагранжа, получаем

$$\varepsilon_0 = 0, \quad \varepsilon_{j+1} = \varepsilon_j + hf'_u(x_j, \theta_j)\varepsilon_j + \frac{h^2}{2} u''(\xi_j) - \eta_j,$$

где θ_j лежит между u_j и y_j, $j = 0, 1, \ldots, N-1$.

Отсюда, учитывая (4), (10), (11), находим $\varepsilon_0 = 0$,

$$|\varepsilon_{j+1}| \leqslant (1 + hM_2)|\varepsilon_j| + h^2 M,$$

где $j = 0, 1, \ldots, N-1$,

$$M = c_0 + (M_1 + M_0 M_2)/2. \qquad (13)$$

Следовательно, по лемме 1

$$\|\varepsilon\|_h \leqslant \frac{h^2 M}{hM_2}(e^{NhM_2} - 1) = h\frac{M}{M_2}(e^{lM_2} - 1). \qquad (14)$$

Итак, получена выражающаяся через известные величины оценка погрешности приближенного решения y задачи Коши (1), (2), найденного методом Эйлера. Из изложенного вытекают два вывода.

1. При отсутствии погрешностей округлений локальная погрешность метода Эйлера, т. е. погрешность на одном шаге h, возникающая за счет перемещения по касательной к интегральной кривой, проходящей через точку (x_j, y_j), а не по самой интегральной кривой, есть $O(h^2)$. Это следует из разложения вида (12) по формуле Тейлора для решения уравнения (1), которое проходит через точку (x_j, y_j). Глобальная погрешность или, точнее, максимальная погрешность решения y на ω_h согласно (14) равна $O(h)$, т. е. на

единицу по порядку хуже *), чем локальная погрешность.

2. Глобальная погрешность не ухудшается по порядку относительно h, если присутствует погрешность округлений того же порядка, что и локальная погрешность.

Основной недостаток метода Эйлера — невысокая относительно h точность. Он является методом первого порядка точности.

Имеется несколько путей построения численных методов решения задачи Коши более высокой по порядку относительно h точности. Один из них основывается на использовании разложения решения по формуле Тейлора (или, как еще говорят, разложения в ряд). Однако на практике предпочтительнее методы, требующие фактического вычисления только значений правой части уравнения (1), а не каких-либо ее производных. Такими методами являются методы Рунге — Кутта и Адамса.

Методы Рунге — Кутта. Приведем три метода Рунге — Кутта.

1. Метод (носящий также название «предиктор — корректор») второго порядка точности:

$$y^*_{j+1} = y_j + hf(x_j, y_j) \quad (y_0 = u_0), \tag{15}$$

$$y_{j+1} = y_j + h\frac{f(x_j, y_j) + f(x_{j+1}, y^*_{j+1})}{2}. \tag{16}$$

Формула (15) предсказывает «грубое» значение y^*_{j+1} по методу Эйлера, а формула (16) уточняет (корректирует) значение приближенного решения в точке x_{j+1}, в чем мы сейчас убедимся.

Предположим, что $f \in C_2(\bar{G})$. Подставив в (16) $x_{j+1} = x_j + h$, а вместо y^*_{j+1} — его выражение (15), получим

$$y_{j+1} = y_j + h\frac{f(x_j, y_j) + f(x_j + h, y_j + hf(x_j, y_j))}{2}.$$

Считая $f(x_j + h, y_j + hf(x_j, y_j))$ функцией только от h (x_j, y_j фиксированы), с помощью формулы

*) Можно доказать, что указанный порядок в общем случае является точным.

Тейлора находим *)

$$f(x_j + h, y_j + hf(x_j, y_j)) = f(x_j, y_j) + \\ + h(f'_x(x_j, y_j) + f'_u(x_j, y_j) f(x_j, y_j)) + O(h^2)$$

и, следовательно,

$$y_{j+1} = y_j + hf(x_j, y_j) + \\ + \frac{h^2}{2}(f'_x(x_j, y_j) + f'_u(x_j, y_j) f(x_j, y_j)) + O(h^3). \quad (17)$$

При условии, что $f \in C_2(\bar{G})$, решение $v(x)$ дифференциального уравнения (1), проходящее через точку (x_j, y_j), согласно теореме 1 трижды непрерывно дифференцируемо по x. Поэтому, учитывая аналогичные (4) выражения для $v'(x)$, $v''(x)$, по формуле Тейлора получаем

$$v_{j+1} = v(x_j) + hv'(x_j) + \frac{h^2}{2} v''(x_j) + O(h^3) = y_j + hf(x_j, y_j) + \\ + \frac{h^2}{2}(f'_x(x_j, y_j) + f'_u(x_j, y_j) f(x_j, y_j)) + O(h^3), \quad (18)$$

где $v_{j+1} = v(x_{j+1})$, причем величины $O(h^3)$ в (17) и (18), вообще говоря, не тождественны.

На основании (17), (18) погрешность ухода y_{j+1} с интегральной кривой, проходящей через точку (x_j, y_j), т. е. разность $v_{j+1} - y_{j+1}$, есть $O(h^3)$. Таким образом, локальная погрешность данного метода имеет третий порядок относительно h и, следовательно, глобальная погрешность $\|u - y\|_h$ равна $O(h^2)$, что доказывается аналогично, как и в методе Эйлера.

2. Метод (называемый также усовершенствованным методом Эйлера) второго порядка точности:

$$y_{j+1/2} = y_j + \frac{h}{2} f(x_j, y_j) \quad (y_0 = u_0), \\ y_{j+1} = y_j + hf\left(x_j + \frac{h}{2}, y_{j+1/2}\right). \quad (19)$$

Его локальная погрешность исследуется так же, как и в предыдущем методе.

---

*) f'_u означает частную производную по второму аргументу функции f, который первоначально в (1) обозначен через u.

3. Метод Рунге—Кутта четвертого порядка точности (наиболее распространенный на практике):

$$y_{j+1} = y_j + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \quad (y_0 = u_0), \quad (20)$$

где

$$k_1 = hf(x_j, y_j), \quad k_2 = hf\left(x_j + \frac{h}{2}, y_j + \frac{k_1}{2}\right),$$

$$k_3 = hf\left(x_j + \frac{h}{2}, y_j + \frac{k_2}{2}\right), \quad k_4 = hf(x_j + h, y_j + k_3).$$

Если $f \in C_4(\bar{G})$, то локальная погрешность этого метода есть $O(h^5)$, а глобальная погрешность, т. е. $\|u - y\|_h$, равна $O(h^4)$.

Метод Адамса. Пусть известны с шагом h приближенные значения $y_{j-m}, \ldots, y_{j-1}, y_j$ решения задачи Коши (1), (2). Обозначим

$$f_j = f(x_j, y_j). \quad (21)$$

Строим интерполяционный многочлен Лагранжа (см. § 4) степени m:

$$L_m(x) = \sum_{i=0}^{m} p_{mi}(x) f_{j-i}, \quad (22)$$

удовлетворяющий условиям

$$L_m(x_{j-i}) = f_{j-i}, \quad i = 0, 1, \ldots, m.$$

Поскольку для точного решения задачи Коши (1), (2) выполняется равенство

$$u_{j+1} = u_j + \int_{x_j}^{x_{j+1}} f(x, u(x))\,dx,$$

где $x_{j+1} = x_j + h$, $u_k = u(x_k)$, то естественно положить

$$y_{j+1} = y_j + \int_{x_j}^{x_{j+1}} L_m(x)\,dx.$$

Тогда с учетом (22) будем иметь следующее выражение для y_{j+1}:

$$y_{j+1} = y_j + h \sum_{i=0}^{m} a_{mi} f_{j-i}, \quad (23)$$

где f_{j-i} — значения правой части уравнения (1), найденные по формуле (21), а

$$\alpha_{mi} = \frac{1}{h} \int_{x_j}^{x_{j+1}} p_{mi}(x)\, dx, \quad i = 0,\ 1,\ \ldots,\ m,$$

являются вполне определенными постоянными, не зависящими ни от h, ни от j. В частности,

$\alpha_{00} = 1$ (метод Эйлера, $m = 0$),

$\alpha_{10} = 3/2, \quad \alpha_{11} = -1/2,$

$\alpha_{20} = 23/12, \quad \alpha_{21} = -4/3, \quad \alpha_{22} = 5/12,$

$\alpha_{30} = 55/24, \quad \alpha_{31} = -59/24, \quad \alpha_{32} = 37/24, \quad \alpha_{33} = -3/8.$

Доказано, что если $f \in C_{m+1}(\bar{G})$ и известны $m+1$ значений $y_0, y_1, \ldots, y_m$ приближенного решения в точках $x_0, x_1, \ldots, x_m = x_0 + mh$, причем

$$|u_j - y_j| \leqslant ch^{m+1}, \quad j = 0,\ 1,\ \ldots,\ m,$$

где $u_j = u(x_j)$ — значения точного решения задачи (1), (2), c не зависит от h, то глобальная погрешность приближенного решения y, найденного в остальных точках сетки ω_h методом Адамса (23), равна $O(h^{m+1})$.

Сравнение методов Рунге — Кутта и Адамса. Сравним методы Рунге — Кутта (20) и Адамса (23) ($m = 3$), обладающие четвертым порядком точности. На каждом шаге h по методу (20) требуется четыре раза вычислять значения функции f, и, следовательно, если функция $f(x, u)$ достаточно сложная, это может потребовать значительного числа действий ЭВМ. В методе (23) на шаге вычисляется только одно новое значение f. Однако метод Рунге — Кутта (20) независим, т. е. им можно начинать считать, имея только начальное значение y_0, и в любой момент возможно изменить шаг h. Метод же Адамса требует знания приближенного решения в нескольких начальных точках сетки и не позволяет изменять шаг. На практике часто комбинируют эти методы. Начальный участок вычисляют по методу Рунге — Кутта, а затем переходят к методу Адамса. Аналогично поступают при необходимости изменения в процессе счета шага h.

Применение правила Рунге. Предположим, что $f \in C_{k+1}(\bar{G})$ и приближенное решение задачи Коши (1), (2) находится методом Рунге — Кутта k-го порядка точности (при $k = 1$ — методом Эйлера). Тогда можно получить следующее соотношение на сетке ω_h:

$$u = y_h + v(x)h^k + r(x, h), \qquad (24)$$

где u — точное решение задачи (1), (2), y_h — приближенное решение (индекс h указывает шаг, с которым оно вычислено методом Рунге — Кутта), $v(x)$ — некоторая не зависящая от h функция от x, $|r(x, h)| \leqslant c'h^{k+1}$, c' не зависит от h.

Поэтому, вычислив приближенное решение три раза с шагами $2h$, h, $h/2$, можно в узлах сетки ω_{2h}, в которых выполнено условие (16.13) для значений приближенных решений y_{2h}, y_h, $y_{h/2}$ (отвечающих одному и тому же x), оценить погрешность по правилу Рунге:

$$u - y_{h/2} \approx \frac{y_{h/2} - y_h}{2^k - 1}.$$

Кроме того, возможно найти уточненное по Ричардсону решение

$$y_h^* = \frac{2^k y_{h/2} - y_h}{2^k - 1},$$

имеющее погрешность $O(h^{k+1})$, а не $O(h^k)$.

Замечания. 1. Соотношение (24) сохранится, если приближенное решение находится по формулам (15), (16), или (19), или (20) не идеально, а с погрешностями округлений (так же, как в методе Эйлера (8)), которые на каждом шаге ограничены величиной ch^{k+2}, где k — порядок глобальной точности метода, c не зависит от h. Проверить непосредственно и выдержать это условие для погрешности округлений на ЭВМ затруднительно. Более реально осуществлять проверку на выполнение условия (16.13) в конечной или некоторой промежуточной точке интегрирования, например в точке $x_0 + l$ или $x_0 + l/2$. Если погрешности округлений начинают превалировать над погрешностями собственно метода, условие (16.13) грубо нарушается. Тогда следует либо увеличить

шаг h, либо перейти на режим счета на ЭВМ с удвоенным, утроенным и т. д. числом разрядов. Однако это не всегда возможно.

2. В случае задачи Коши для нормальной системы дифференциальных уравнений

$$\boldsymbol{u}' = \boldsymbol{f}(x, \boldsymbol{u}), \quad \boldsymbol{u}(x_0) = \boldsymbol{u}_0$$

рассмотренные приближенные методы интегрирования Эйлера, Рунге — Кутта и Адамса формально остаются теми же, только функции u, y, f и в формуле (20) величины k_i заменяются соответственно на вектор-функции $\boldsymbol{u}$, $\boldsymbol{y}$, $\boldsymbol{f}$ и векторы $\boldsymbol{k}_i$. Заданные выше требования к гладкости функции f остаются теми же к компонентам вектор-функции $\boldsymbol{f}$, а гарантируемая точность по порядку относительно h будет та же (что и в случае одного уравнения (1)) для компонент приближенного решения. Правило Рунге применяется для каждой компоненты в отдельности.

Вопросы существования решения исходной задачи. Рассмотрим задачу Коши

$$u' = u^2, \quad u(0) = 1. \tag{25}$$

Если для решения этой задачи применить, например, метод Эйлера или метод Рунге — Кутта (20) с любым шагом $h > 0$, то приближенное решение может быть принципиально вычислено для любого числа шагов, т. е. для сколь угодно больших значений аргумента x. Однако единственным точным решением задачи (25) является функция $u(x) = (1 - x)^{-1}$, определенная для неотрицательных x только на промежутке $[0, 1)$.

Таким образом, наличие приближенного решения задачи Коши на некотором промежутке еще не гарантирует существования решения исходной дифференциальной задачи на этом промежутке.

Приведем две теоремы (доказательства опустим), которые позволяют судить о наличии решения задачи Коши (1), (2) в заданном прямоугольнике $\bar{G}$ по приближенному решению, полученному методом Эйлера. Пусть

$$\bar{c} = \frac{M}{M_2}(e^{lM_2} - 1) + M_0,$$

где M_0, M_2, M определены формулами (11), (13).

Теорема 2. *Если* $f \in C_1(\bar{G})$ *и приближенное решение, найденное по формулам* (8) *при некотором фиксированном* $h = l/N$, *подчиняется неравенствам*

$$a + \bar{c}h < \min_{0 \leqslant j \leqslant N} y_j \leqslant \max_{0 \leqslant j \leqslant N} y_j < b - \bar{c}h, \qquad (26)$$

то решение $u(x)$ *задачи Коши* (1), (2) *существует на отрезке* $[x_0, x_0 + l]$ *и удовлетворяет условиям* (3).

Теорема 3. *Если* $f \in C_1(\bar{G})$ *и существует на отрезке* $[x_0, x_0 + l]$ *решение* $u(x)$ *задачи Коши* (1), (2), *удовлетворяющее условиям* (3), *то найдется такое* $h_0 > 0$, *что при любом* $h = l/N < h_0$ *неравенства* (26) *для приближенного решения, найденного по формулам* (8), *заведомо будут выполнены.*

Согласно теореме 2 для существования решения задачи Коши (1), (2) на отрезке $[x_0, x_0 + l]$, удовлетворяющего условиям (3), достаточно, чтобы приближенное решение, найденное методом Эйлера, удовлетворяло неравенствам (26) только при каком-нибудь одном значении h. Теорема 3 гарантирует, что если указанное решение задачи Коши (1), (2) существует, то неравенства (26) будут выполнены для всех достаточно малых $h = l/N$.

ГЛАВА 3

ЧИСЛЕННЫЕ МЕТОДЫ ЛИНЕЙНОЙ АЛГЕБРЫ

К численным методам линейной алгебры относятся численные методы решения систем линейных алгебраических уравнений, обращения матриц, вычисления определителей и нахождения собственных значений и собственных векторов матриц.

Методы решения систем линейных алгебраических уравнений разбиваются на две группы. К первой группе принадлежат так называемые *точные* или *прямые методы* — алгоритмы, позволяющие получить решение системы за конечное число арифметических действий. Сюда относится известное правило Крамера нахождения решения с помощью определителей, метод Гаусса (метод исключений) и метод прогонки.

Вторую группу составляют *приближенные методы*, в частности итерационные методы решения систем линейных алгебраических уравнений.

Правило Крамера в ЭВМ не применяется, так как оно требует значительно большего числа арифметических действий, чем метод Гаусса. Метод Гаусса используется в ЭВМ при решении систем до порядка 10^3, а итерационные методы — до порядка 10^6. Метод прогонки применяется для решения важного класса специальных систем линейных уравнений с трехдиагональной матрицей, часто возникающих в приложениях.

Метод Гаусса решения систем линейных уравнений, обращения матриц и вычисления определителей излагается в § 19. В § 20 вводится понятие нормы матрицы, используемое в § 21 при рассмотрении метода простых итераций. Важной характеристикой матрицы системы линейных алгебраических уравнений является ее обусловленность — мера чувствительности решения системы к возмущению правой части. Пример плохо обусловленной матрицы дан в § 20, а хорошо обусловленная матрица приведена в § 21. Методу прогонки посвящен § 22, а некоторые проблемы собственных значений, в основном для симметричных матриц, рассматриваются в § 23.

§ 19. Метод Гаусса

Рассмотрим для простоты систему линейных алгебраических уравнений четвертого порядка

$$\begin{aligned} a_{11}^{(0)}x_1 + a_{12}^{(0)}x_2 + a_{13}^{(0)}x_3 + a_{14}^{(0)}x_4 &= a_{15}^{(0)}, \\ a_{21}^{(0)}x_1 + a_{22}^{(0)}x_2 + a_{23}^{(0)}x_3 + a_{24}^{(0)}x_4 &= a_{25}^{(0)}, \\ a_{31}^{(0)}x_1 + a_{32}^{(0)}x_2 + a_{33}^{(0)}x_3 + a_{34}^{(0)}x_4 &= a_{35}^{(0)}, \\ a_{41}^{(0)}x_1 + a_{42}^{(0)}x_2 + a_{43}^{(0)}x_3 + a_{44}^{(0)}x_4 &= a_{45}^{(0)}. \end{aligned} \tag{1}$$

Предположим, что коэффициент $a_{11}^{(0)}$, называемый *ведущим элементом* первой строки, не равен нулю. Разделив первое из уравнений (1) на $a_{11}^{(0)}$, получим новое уравнение

$$x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + a_{14}^{(1)}x_4 = a_{15}^{(1)}, \tag{2}$$

где $a_{1j}^{(1)} = a_{1j}^{(0)} / a_{11}^{(0)}$, $j = 2, 3, 4, 5$.

Исключим неизвестную x_1 из каждого уравнения системы (1), начиная со второго, путем вычитания уравнения (2), умноженного на коэффициент при x_1 в соответствующем уравнении. Преобразованные уравнения имеют вид

$$\begin{aligned} a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + a_{24}^{(1)}x_4 &= a_{25}^{(1)}, \\ a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + a_{34}^{(1)}x_4 &= a_{35}^{(1)}, \\ a_{42}^{(1)}x_2 + a_{43}^{(1)}x_3 + a_{44}^{(1)}x_4 &= a_{45}^{(1)}, \end{aligned} \tag{3}$$

где $a_{ij}^{(1)} = a_{ij}^{(0)} - a_{1j}^{(1)} a_{i1}^{(0)}$, $i = 2, 3, 4$, $j = 2, 3, 4, 5$.

Допустим, что ведущий элемент второй строки, т. е. коэффициент $a_{22}^{(1)}$, тоже отличен от нуля. Тогда, разделив на него первое из уравнений (3), получим уравнение

$$x_2 + a_{23}^{(2)}x_3 + a_{24}^{(2)}x_4 = a_{25}^{(2)}, \tag{4}$$

где $a_{2j}^{(2)} = a_{2j}^{(1)} / a_{22}^{(1)}$, $j = 3, 4, 5$.

Исключив с помощью уравнения (4) неизвестную x_2 из двух последних уравнений в (3), приходим к уравнениям

$$\begin{aligned} a_{33}^{(2)}x_3 + a_{34}^{(2)}x_4 &= a_{35}^{(2)}, \\ a_{43}^{(2)}x_3 + a_{44}^{(2)}x_4 &= a_{45}^{(2)}, \end{aligned} \tag{5}$$

где $a_{ij}^{(2)} = a_{ij}^{(1)} - a_{2j}^{(2)} a_{i2}^{(1)}$, $i = 3, 4$, $j = 3, 4, 5$.

Если ведущий элемент третьей строки $a_{33}^{(2)}$ не равен нулю, то, поделив на него первое из уравнений (5) и вычтя найденное уравнение, умноженное на $a_{43}^{(2)}$, из второго уравнения, получим

$$x_3 + a_{34}^{(3)} x_4 = a_{35}^{(3)}, \tag{6}$$

$$a_{44}^{(3)} x_4 = a_{45}^{(3)}, \tag{7}$$

где $a_{3j}^{(3)} = a_{3j}^{(2)} / a_{33}^{(2)}$, $a_{4j}^{(3)} = a_{4j}^{(2)} - a_{3j}^{(3)} a_{43}^{(2)}$, $j = 4, 5$.

Наконец, если $a_{44}^{(3)} \neq 0$, то, разделив на него уравнение (7), приведем это уравнение к виду

$$x_4 = a_{45}^{(4)}, \tag{8}$$

где $a_{45}^{(4)} = a_{45}^{(3)} / a_{44}^{(3)}$.

Итак, если ведущие элементы $a_{11}^{(0)}$, $a_{22}^{(1)}$, $a_{33}^{(2)}$, $a_{44}^{(3)}$ отличны от нуля, то система (1) эквивалентна следующей системе с треугольной матрицей:

$$\begin{aligned} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 + a_{14}^{(1)} x_4 &= a_{15}^{(1)}, \\ x_2 + a_{23}^{(2)} x_3 + a_{24}^{(2)} x_4 &= a_{25}^{(2)}, \\ x_3 + a_{34}^{(3)} x_4 &= a_{35}^{(3)}, \\ x_4 &= a_{45}^{(4)}, \end{aligned} \tag{9}$$

которая получена объединением уравнений (2), (4), (6), (8). Из системы (9) неизвестные x_1, x_2, x_3, x_4 находятся явно в обратном порядке по формулам

$$\begin{aligned} x_4 &= a_{45}^{(4)}, \\ x_3 &= a_{35}^{(3)} - a_{34}^{(3)} x_4, \\ x_2 &= a_{25}^{(2)} - a_{23}^{(2)} x_3 - a_{24}^{(2)} x_4, \\ x_1 &= a_{15}^{(1)} - a_{12}^{(1)} x_2 - a_{13}^{(1)} x_3 - a_{14}^{(1)} x_4. \end{aligned} \tag{10}$$

Процесс приведения системы (1) к треугольному виду (9) называется *прямым ходом*, а нахождение неизвестных по формулам (10) — *обратным ходом* метода Гаусса.

Решение системы (1) по изложенному методу Гаусса при ручном счете обычно представляется в виде табл. 1. Последний столбец в таблице введен для контроля вычислений, осуществляемого следующим образом. Задаем $a_{i6}^{(0)}$ равным сумме остальных эле-

Таблица 1

$a_{11}^{(0)}$ $a_{21}^{(0)}$ $a_{31}^{(0)}$ $a_{41}^{(0)}$	$a_{12}^{(0)}$ $a_{22}^{(0)}$ $a_{32}^{(0)}$ $a_{42}^{(0)}$	$a_{13}^{(0)}$ $a_{23}^{(0)}$ $a_{33}^{(0)}$ $a_{43}^{(0)}$	$a_{14}^{(0)}$ $a_{24}^{(0)}$ $a_{34}^{(0)}$ $a_{44}^{(0)}$	$a_{15}^{(0)}$ $a_{25}^{(0)}$ $a_{35}^{(0)}$ $a_{45}^{(0)}$	$a_{16}^{(0)}$ $a_{26}^{(0)}$ $a_{36}^{(0)}$ $a_{46}^{(0)}$
1	$a_{12}^{(1)}$ $a_{22}^{(1)}$ $a_{32}^{(1)}$ $a_{42}^{(1)}$	$a_{13}^{(1)}$ $a_{23}^{(1)}$ $a_{33}^{(1)}$ $a_{43}^{(1)}$	$a_{14}^{(1)}$ $a_{24}^{(1)}$ $a_{34}^{(1)}$ $a_{44}^{(1)}$	$a_{15}^{(1)}$ $a_{25}^{(1)}$ $a_{35}^{(1)}$ $a_{45}^{(1)}$	$a_{16}^{(1)}$ $a_{26}^{(1)}$ $a_{36}^{(1)}$ $a_{46}^{(1)}$
	1	$a_{23}^{(2)}$ $a_{33}^{(2)}$ $a_{43}^{(2)}$	$a_{24}^{(2)}$ $a_{34}^{(2)}$ $a_{44}^{(2)}$	$a_{25}^{(2)}$ $a_{35}^{(2)}$ $a_{45}^{(2)}$	$a_{26}^{(2)}$ $a_{36}^{(2)}$ $a_{46}^{(2)}$
		1	$a_{34}^{(3)}$ $a_{44}^{(3)}$	$a_{35}^{(3)}$ $a_{45}^{(3)}$	$a_{36}^{(3)}$ $a_{46}^{(3)}$
			1	x_4	$\bar{x}_4$
		1		x_3	$\bar{x}_3$
	1			x_2	$\bar{x}_2$
1				x_1	$\bar{x}_1$

ментов i-й строки, взятой с обратным знаком, т. е.

$$a_{i6}^{(0)} = -(a_{i1}^{(0)} + a_{i2}^{(0)} + \ldots + a_{i5}^{(0)}), \quad i = 1, 2, 3, 4.$$

Тогда сумма всех элементов каждой из четырех расширенных начальных строк будет равна нулю. Затем в процессе исключения неизвестных производим над элементами последнего столбца те же самые действия, что и над элементами соответствующих строк, а именно, полагаем

$$a_{kj}^{(k)} = a_{kj}^{(k-1)} / a_{kk}^{(k-1)}, \quad a_{ij}^{(k)} = a_{ij}^{(k-1)} - a_{ik}^{(k-1)} a_{kj}^{(k)}, \qquad (11)$$

где $k+1 \leqslant j \leqslant 6$, $k+1 \leqslant i \leqslant 4$, $k = 1, 2, 3, 4$.

При этом сумма элементов вновь получаемых расширенных строк должна оставаться равной нулю, поскольку эти элементы находятся с помощью линейных операций над предыдущими строками. Например,

$$1 + a_{23}^{(2)} + a_{24}^{(2)} + a_{25}^{(2)} + a_{26}^{(2)} = 0. \qquad (12)$$

Если вычисления ведутся с округлениями, то допускается отклонение левой части равенства (12) от нуля в пределах погрешностей округлений. Наличие большого отклонения свидетельствует о наличии грубой ошибки в вычислениях.

Неизвестные x_4, x_3, x_2, x_1 находятся по формулам (10), а величины $\bar{x}_4$, $\bar{x}_3$, $\bar{x}_2$, $\bar{x}_1$ — по тем же формулам, но в которых числа $a_{k5}^{(k)}$ заменены на $a_{k6}^{(k)}$, $k = 4, 3, 2, 1$, x_i — на $\bar{x}_i$. При отсутствии погрешностей округлений сумма элементов в четырех последних строках табл. 1 тоже должна быть равна нулю, т. е. $1 + x_i + \bar{x}_i = 0$, $i = 4, 3, 2, 1$. В табл. 1 элементы $a_{45}^{(4)}$, $a_{46}^{(4)}$ опущены, поскольку они совпадают соответственно с x_4, $\bar{x}_4$.

Пример. Требуется решить систему линейных алгебраических уравнений

$$\begin{aligned} x_1 + 0{,}17x_2 - 0{,}25x_3 + 0{,}54x_4 &= 0{,}3, \\ 0{,}47x_1 + x_2 + 0{,}67x_3 - 0{,}32x_4 &= 0{,}5, \\ -0{,}11x_1 + 0{,}35x_2 + x_3 - 0{,}74x_4 &= 0{,}7, \\ 0{,}55x_1 + 0{,}43x_2 + 0{,}36x_3 + x_4 &= 0{,}9. \end{aligned} \qquad (13)$$

Решение представлено в табл. 2. Вычисления ведутся с двумя запасными десятичными знаками по сравнению с числом десятичных знаков у коэффициентов заданной системы. Сумма элементов каждой строки табл. 2, включая элемент, расположенный в контрольном столбце, отличается от нуля не более чем на две единицы младшего разряда, что вполне допустимо. Найденное решение системы (13), округленное с двумя десятичными знаками после запятой, таково: $x_1 = 0{,}44$, $x_2 = -0{,}36$, $x_3 = 1{,}17$, $x_4 = 0{,}39$.

Аналогично, методом Гаусса решается система линейных алгебраических уравнений любого порядка n. Пусть дана система

$$\sum_{j=1}^{n} a_{ij}^{(0)} x_j = a_{i,\,n+1}^{(0)}, \quad i = 1, 2, \ldots, n. \qquad (1^*)$$

Если $a_{11}^{(0)} \neq 0$, а также ведущие элементы $a_{ii}^{(i-1)}$, $i = 2, 3, \ldots, n$, остальных строк, получаемые в процессе вычислений, отличны от нуля, то система (1*)

Таблица 2

1	0,17	−0,25	0,54	0,3	−1,76
0,47	1	0,67	−0,32	0,5	−2,32
−0,11	0,35	1	−0,74	0,7	−1,20
0,55	0,43	0,36	1	0,9	−3,24
1	0,17	−0,25	0,54	0,3	−1,76
	0,9201	0,7875	−0,5738	0,3590	−1,4928
	0,3687	0,9725	−0,6806	0,7330	−1,3936
	0,3365	0,4975	0,7030	0,7350	−2,2720
	1	0,8559	−0,6236	0,3902	−1,6224
		0,6569	−0,4507	0,5891	−0,7954
		0,2095	0,9128	0,6037	−1,7261
		1	−0,6861	0,8968	−1,2108
			1,0565	0,4158	−1,4724
			1	0,3936	−1,3937
		1		1,1668	−2,1670
	1			−0,3630	−0,6368
1				0,4409	−1,4409

приводится к треугольному виду

$$x_i + \sum_{j=i+1}^{n} a_{ij}^{(i)} x_j = a_{i,\,n+1}^{(i)}, \quad i = 1,\ 2,\ \ldots,\ n. \qquad (9^*)$$

Ведущие элементы $a_{il}^{(i-1)}$ и коэффициенты системы (9*) находятся с помощью формул (11), где $k+1 \leqslant j \leqslant n+1$, $k+1 \leqslant i \leqslant n+1$, $k = 1,\ 2,\ \ldots,\ n$.

Обратный ход, в котором в обратном порядке определяются неизвестные, осуществляется по формулам

$$x_n = a_{n,\,n+1}^{(n)},$$

$$x_i = a_{i,\,n+1}^{(i)} - \sum_{j=i+1}^{n} a_{ij}^{(i)} x_j, \quad i = n-1,\ n-2,\ \ldots,\ 1. \qquad (10^*)$$

Подсчет числа выполняемых арифметических действий. Сначала докажем вспомогательную лемму.

Лемма 1. *Выполняются равенства*

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \quad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}. \quad (14)$$

Доказательство Первое равенство, выражающее сумму арифметической прогрессии, известно. Второе равенство при $n = 1$ очевидно. Допустим, что оно верно при некотором натуральном $n = m$. Тогда

$$\sum_{k=1}^{m+1} k^2 = \sum_{k=1}^{m} k^2 + (m+1)^2 = \frac{m(m+1)(2m+1)}{6} + (m+1)^2 =$$
$$= \frac{(m+1)[(m+1)+1][2(m+1)+1]}{6}.$$

Следовательно, по индукции второе равенство (14) верно для любого натурального n.

Подсчитаем число выполняемых арифметических действий при решении системы порядка n без учета операций контроля. При исключении неизвестной x_1 осуществляются следующие действия:

а) производится всего n делений второго и следующих коэффициентов первого уравнения и его правой части на коэффициент при x_1;

б) расходуется по n умножений и n вычитаний при исключении x_1 из второго и последующих уравнений, число которых равно $n - 1$.

Таким образом, общее число арифметических действий, затрачиваемых на исключение x_1, есть

$$Q_1 = n + 2n(n-1) = 2n^2 - n. \quad (15)$$

При исключении x_2 производятся совершенно аналогичные вычисления, как и при исключении x_1. Разница состоит лишь в том, что осуществляются действия над строками матрицы, размеры которой в каждом направлении на единицу меньше. Поэтому число выполняемых арифметических действий Q_2 при исключении x_2 получается из выражения (15) заменой n на $n - 1$:

$$Q_2 = 2(n-1)^2 - (n-1) = 2(n-2+1)^2 - (n-2+1).$$

Вообще, при исключении неизвестной x_i выполняются Q_i арифметических действий:

$$Q_i = 2(n-i+1)^2 - (n-i+1).$$

Всего в прямом ходе затрачивается число арифметических действий, равное

$$Q_* = \sum_{i=1}^{n} Q_i = \sum_{i=1}^{n} [2(n-i+1)^2 - (n-i+1)].$$

Отсюда, полагая $k = n-i+1$ и опираясь на лемму 1, находим явное выражение для Q_*:

$$Q_* = \sum_{k=n}^{1} (2k^2 - k) = \sum_{k=1}^{n} (2k^2 - k) =$$

$$= \frac{n(n+1)(2n+1)}{3} - \frac{n(n+1)}{2} = \frac{2}{3}n^3 + \frac{n^2}{2} - \frac{n}{6}. \quad (16)$$

В обратном ходе (при $n = 4$, задаваемом формулами (10)) число умножений и вычитаний, вместе взятых, равно удвоенному числу элементов квадратной матрицы порядка n, расположенных под главной диагональю, и, как показывают простые подсчеты, составляет

$$Q^* = n(n-1) = n^2 - n. \quad (17)$$

Итак, число арифметических действий, выполняемых при решении системы линейных алгебраических уравнений порядка n методом Гаусса, в общем случае равно

$$Q = Q_* + Q^* = \frac{2}{3}n^3 + \frac{3}{2}n^2 - \frac{7}{6}n \approx \frac{2}{3}n^3. \quad (18)$$

Дальше нам потребуется знать отдельно число выполняемых арифметических действий над правыми частями уравнений в прямом ходе. При исключении x_1 правая часть первого уравнения делится на коэффициент при x_1, а при нахождении правых частей остальных уравнений производится по одному умножению и одному вычитанию. Таким образом, при исключении x_1 всего над правыми частями выполняется q_1 арифметических действий:

$$q_1 = 1 + 2(n-1) = 2n - 1.$$

Соответственно с правыми частями уравнений при исключении x_i осуществляется q_i арифметических

действий:

$$q_i = 2(n-i+1)-1.$$

Следовательно, в прямом ходе число выполняемых арифметических действий над правыми частями уравнений равно

$$q_* = \sum_{i=1}^{n} [2(n-i+1)-1] = \sum_{k=1}^{n} (2k-1) = n(n+1)-n = n^2. \quad (19)$$

Вычисление определителей. В прямом ходе метода Гаусса матрица $(a_{ij}^{(0)})_{i,\,j=1}^{n}$ заданной системы линейных алгебраических уравнений (1*) при условии, что ведущие элементы всех строк отличны от нуля, приводится к треугольному виду. Это достигается операциями деления строк на их ведущие элементы и вычитания нормированных таким способом строк, умноженных на некоторые числа, из последующих строк (подробно эта процедура описана при $n = 4$).

Определитель матрицы при делении ее строки на ведущий элемент тоже делится на этот элемент, а при вычитании из какой-либо строки матрицы другой строки, умноженной на любое число, как известно, определитель остается тем же. Определитель найденной в прямом ходе треугольной матрицы, у которой на главной диагонали стоят единицы, т. е. определитель системы (9*), совпадающей при $n = 4$ с системой (9), равен единице. Он получен в результате деления определителя матрицы $(a_{ij}^{(0)})_{i,\,j=1}^{n}$, обозначаемого $\det(a_{ij}^{(0)})$, на ведущие элементы $a_{11}^{(0)}$, $a_{22}^{(1)}$, ..., $a_{nn}^{(n-1)}$ всех строк. Следовательно, если $a_{11}^{(0)} \neq 0$, $a_{22}^{(1)} \neq 0$, ..., $a_{nn}^{(n-1)} \neq 0$, то

$$\det(a_{ij}^{(0)}) = a_{11}^{(0)} a_{22}^{(1)} \ldots a_{nn}^{(n-1)}. \quad (20)$$

Таким образом, при решении методом Гаусса системы линейных алгебраических уравнений можно попутно вычислить ее определитель. Если требуется найти только определитель заданной матрицы $(a_{ij}^{(0)})_{i,\,j=1}^{n}$, то вычисления проводятся по схеме, отвечающей прямому ходу метода Гаусса, с той лишь разницей, что отсутствуют действия над столбцом пра-

вых частей. Согласно (16), (19), (20) число арифметических действий, выполняемых при вычислении $\det(a_{ij}^{(0)})$ рассматриваемым методом, равно

$$Q_0 = Q_* - q_* + (n-1) = \frac{2}{3}n^3 + \frac{n^2}{2} - \frac{n}{6} - n^2 + n - 1 = \\ = \frac{2}{3}n^3 - \frac{n^2}{2} + \frac{5}{6}n - 1 \approx \frac{2}{3}n^3, \quad (21)$$

где Q_* — число действий в прямом ходе при решении системы порядка n, q_* — число действий, выполняемых при этом над правыми частями уравнений, $n-1$ — число умножений в формуле (20).

Замечание 1. Если в прямом ходе окажется ведущий элемент какой-либо строки равным нулю, то изложенный простейший вариант метода Гаусса непригоден для решения соответствующей системы или вычисления определителя. Ниже рассматривается более универсальный вариант метода Гаусса с выбором главного элемента.

Решение нескольких систем, отличающихся правыми частями. Если задано несколько систем линейных алгебраических уравнений n-го порядка, отличающихся только правыми частями, то целесообразно прямой ход осуществлять со всеми столбцами правых частей параллельно (так же, как со столбцом правых частей и дополнительным контрольным столбцом в рассмотренном выше примере). Обратный ход для нахождения решения каждой системы проводится последовательно.

Число арифметических действий при решении p указанных систем согласно (17)—(19) составляет

$$Q(p) = Q + (p-1)(q_* + Q^*) = \\ = \frac{2}{3}n^3 + \frac{3}{2}n^2 - \frac{7}{6}n + (p-1)(n^2 + n^2 - n) = \\ = \frac{2}{3}n^3 - \frac{n^2}{2} - \frac{n}{6} + pn(2n-1) \approx \frac{2}{3}n^3 + 2pn^2, \quad (22)$$

где Q — число действий при полном решении одной системы, q_* — число действий в прямом ходе над одним столбцом правых частей, Q^* — число действий в обратном ходе при нахождении одного решения.

Вычисление обратной матрицы. Пусть дана невырожденная матрица $A = (a_{ij}^{(0)})_{i,j=1}^{n}$. Из

курса линейной алгебры известно, что j-й столбец обратной матрицы A^{-1} совпадает со столбцом $(x_{1j}, x_{2j}, \ldots, x_{nj})^*$, где звездочка — знак транспонирования, $(x_{1j}, x_{2j}, \ldots, x_{nj})$ — решение системы *)

$$Ax_j = \delta_j, \tag{23}$$

$\delta_j = (\delta_{1j}, \delta_{2j}, \ldots, \delta_{nj})$, δ_{ij} — символ Кронекера.

Поэтому для нахождения обратной матрицы A^{-1} достаточно решить n систем вида (23) для $j = 1, 2, \ldots, n$, отличающихся только правыми частями. При этом согласно (22) затрачивается

$$Q(n) \approx \frac{8}{3} n^3 \tag{24}$$

арифметических действий.

Схема с выбором главного элемента. Рассмотренный выше простейший вариант метода Гаусса, называемый *схемой единственного деления*, обладает следующими недостатками. Если ведущий элемент какой-либо строки, например, коэффициент $a_{11}^{(0)}$ при x_1 в первом же уравнении окажется равным нулю, то эта схема формально непригодна, хотя заданная система уравнений может иметь единственное решение. Кроме того, если определитель системы не равен нулю, но в процессе вычислений встречаются ведущие элементы, которые достаточно малы по сравнению с другими элементами соответствующих строк, то это обстоятельство способствует усилению отрицательного влияния погрешностей округления на точность результата.

Изложим схему, называемую *схемой с выбором главного элемента*, которая в случае, когда определитель системы отличен от нуля, при отсутствии погрешностей округления всегда приводит к единственному решению и менее чувствительна к погрешностям округления. Эта схема незначительно отличается от схемы, приведенной в начале параграфа.

---

*) Здесь и в следующих параграфах будем придерживаться обозначений, принятых в книге: Бугров Я. С., Никольский С. М. Элементы линейной алгебры и аналитической геометрии. — М.: Наука, 1984. Там векторы неизвестных, решений и правых частей записываются в строку.

Пусть, как и прежде, дана система (1). Сначала добиваемся выполнения условий

$$|a_{11}^{(0)}| \geqslant |a_{ij}^{(0)}|, \quad i, j = 1, 2, 3, 4,$$

путем перестановки в случае необходимости двух уравнений системы (1), а также двух столбцов неизвестных со своими коэффициентами, и соответствующей перенумерации коэффициентов и неизвестных. Найденный максимальный по модулю коэффициент, обозначенный при перенумерации через $a_{11}^{(0)}$, называется *первым главным элементом.*

Затем, разделив на $a_{11}^{(0)}$, первое уравнение приводим к виду (2) и, исключив x_1 из второго, третьего и четвертого уравнений, получаем уравнения (3).

Далее с системой (3) поступаем аналогично, как и со всей системой (1). А именно, осуществив, если нужно, перестановку двух уравнений, а также, возможно, двух столбцов неизвестных с их коэффициентами, и произведя соответствующую перенумерацию, обеспечиваем выполнение неравенств

$$|a_{22}^{(1)}| \geqslant |a_{ij}^{(1)}|, \quad i, j = 2, 3, 4.$$

При этом, если переставлялись столбцы неизвестных, то соответствующая перестановка производится и в уравнении (2). Найденный максимальный по модулю коэффициент, обозначенный через $a_{22}^{(1)}$, называется *вторым главным элементом.*

Разделив на $a_{22}^{(1)}$ уравнение, стоящее теперь на первом месте в (3), получим уравнение (4), а исключив x_2 из второго и третьего уравнений, приходим к уравнениям (5). Если в системе (5) при нахождении главного элемента будут переставляться столбцы, то соответствующие перестановки и перенумерация осуществляются в уже выделенных уравнениях (2), (4), входящих в окончательную систему (9), и т. д.

Если определитель системы (1) отличен от нуля, то в прямом ходе будет получена система вида (9). Обратный ход выполняется по формулам (10). Переходя к первоначальной нумерации теперь уже найденных неизвестных, получаем решение заданной системы (1).

Аналогично, методом Гаусса с выбором главного элемента решается система любого порядка n.

Замечания. 2. Если матрица заданной системы вырожденная, то перед исключением некоторой неизвестной главный элемент, на который должно делиться первое из оставшихся уравнений, окажется равным нулю. Этим самым и обнаружится, что определитель заданной системы равен нулю. Задача решения системы уравнений с вырожденной матрицей рассматривается в курсе линейной алгебры. Более подробно на этом случае останавливаться не будем.

3. Метод Гаусса с выбором главного элемента позволяет вычислить определитель (если последний не равен нулю) по формуле (20), в правой части которой стоит произведение главных элементов и добавлен множитель $(-1)^{\nu}$, где ν — сумма номеров, переставляемых в прямом ходе строк и столбцов на всех шагах приведения матрицы к треугольному виду. Если же определитель равен нулю, то это обстоятельство выяснится при вычислениях, так как на некотором шаге появится равный нулю главный элемент.

4. На практике при приведении системы к треугольному виду фактически действия производятся только над элементами ее матрицы, а найденные значения неизвестных выписываются в обратном ходе, как, например, это иллюстрируется табл. 1, 2.

5. Имеется следующий вариант схемы с выбором главного элемента. Найденный очередной (например, первый) главный элемент оставляется на месте (не переносится в левый верхний угол матрицы путем перестановки ее строк и столбцов). Неизвестная, отвечающая выбранному главному элементу, исключается из всех уравнений, среди коэффициентов которых еще не выделялись главные элементы. Следующий главный (максимальный по модулю) элемент находится среди коэффициентов уравнений, из которых исключена очередная неизвестная, и т. д. Если определитель заданной системы порядка n не равен нулю, то через n шагов будет получена эквивалентная система, которая путем перестановки ее уравнений и перенумерации неизвестных может быть приведена к треугольному виду.

§ 20. Нормы и обусловленность матриц

Введем в линейном пространстве $\mathbf{R}^n$ n-мерных векторов $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ двумя способами норму

$$\|\boldsymbol{x}\| = \begin{cases} \|\boldsymbol{x}\|_1 = \max\limits_{1 \leqslant j \leqslant n} |x_j|, \\ \|\boldsymbol{x}\|_2 = (\boldsymbol{x}, \boldsymbol{x})^{1/2} = \left(\sum\limits_{j=1}^{n} x_j^2\right)^{1/2}. \end{cases} \tag{1}$$

Проверим выполнение аксиом нормы (см. п. 7 введения) для нормы $\|\boldsymbol{x}\|_1$. Для любого $\boldsymbol{x} \in \mathbf{R}^n$ имеем

$$\max_{1 \leqslant j \leqslant n} |x_j| \geqslant 0,$$

т. е.

$$\|\boldsymbol{x}\|_1 \geqslant 0.$$

При этом $\max\limits_{1 \leqslant j \leqslant n} |x_j| = 0$ или $\|\boldsymbol{x}\|_1 = 0$ тогда и только тогда, когда $\boldsymbol{x} = \mathbf{0}$. Кроме того,

$$\max_{1 \leqslant j \leqslant n} |\alpha x_j| = \max_{1 \leqslant j \leqslant n} |\alpha|\,|x_j| = |\alpha| \max_{1 \leqslant j \leqslant n} |x_j|,$$

т. е.

$$\|\alpha \boldsymbol{x}\|_1 = |\alpha|\,\|\boldsymbol{x}\|_1.$$

где α — любое число. Наконец, пусть даны два произвольных вектора $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbf{R}^n$. Тогда найдется такой номер k, что $\max\limits_{1 \leqslant j \leqslant n} |x_j + y_j| = |x_k + y_k|$. Следовательно,

$$\max_{1 \leqslant j \leqslant n} |x_j + y_j| = |x_k + y_k| \leqslant |x_k| + |y_k| \leqslant$$

$$\leqslant \max_{1 \leqslant j \leqslant n} |x_j| + \max_{1 \leqslant j \leqslant n} |y_j|,$$

т. е.

$$\|\boldsymbol{x} + \boldsymbol{y}\|_1 \leqslant \|\boldsymbol{x}\|_1 + \|\boldsymbol{y}\|_1.$$

Аксиомы для нормы $\|\boldsymbol{x}\|_2$, введенной с помощью скалярного произведения, проверяются в курсе линейной алгебры.

Очевидно, первая и вторая нормы любого $\boldsymbol{x} \in \mathbf{R}^n$ удовлетворяют неравенствам

$$\|\boldsymbol{x}\|_1 \leqslant \|\boldsymbol{x}\|_2 \leqslant \sqrt{n}\,\|\boldsymbol{x}\|_1. \tag{2}$$

В линейном пространстве квадратных числовых матриц n-го порядка

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

зададим норму следующим образом:

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}, \tag{3}$$

где $\|x\|$, $\|Ax\|$ — введенные выше нормы соответственно векторов x, Ax. Заданная по формуле (3) норма матрицы называется *согласованной* с нормой вектора.

В частности, если $A = E$, т. е. A является единичной матрицей, то (поскольку $Ex = x$ для любого $x \in \mathbb{R}^n$ и, следовательно, $\|Ax\| = \|Ex\| = \|x\|$) из (3) вытекает, что

$$\|E\| = 1. \tag{4}$$

Через $\|A\|_m$ будем обозначать норму матрицы A в случае, когда для вектора принята норма $\|x\| = \|x\|_m$, $m = 1, 2$.

Можно доказать, что

$$\|A\|_1 = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |a_{ij}|, \tag{5}$$

$$\|A\|_2 \leqslant \left(\sum_{i,j=1}^{n} a_{ij}^2 \right)^{1/2}, \tag{6}$$

причем, если матрица A — симметричная, т. е. $A = A^*$ (A^* — транспонированная к A матрица), то

$$\|A\|_2 = \max_{1 \leqslant i \leqslant n} |\lambda_i(A)|, \quad \|A^{-1}\|_2 = \frac{1}{\min\limits_{1 \leqslant i \leqslant n} |\lambda_i(A)|}, \tag{7}$$

где $\lambda_i(A)$ — собственные значения матрицы A

Пусть A, B — квадратные числовые матрицы n-го порядка, $A + B$ — их сумма, $x \in \mathbb{R}^n$. Поскольку $\|(A+B)x\| = \|Ax + Bx\| \leqslant \|Ax\| + \|Bx\|$, то

$$\|A + B\| = \sup_{x \neq 0} \frac{\|(A+B)x\|}{\|x\|} \leqslant \sup_{x \neq 0} \frac{\|Ax\| + \|Bx\|}{\|x\|} \leqslant$$

$$\leqslant \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} + \sup_{x \neq 0} \frac{\|Bx\|}{\|x\|} = \|A\| + \|B\|.$$

Таким образом, неравенство треугольника для введенной по формуле (3) нормы матрицы выполнено. Справедливость остальных аксиом для нормы (3) очевидна.

Согласно (3) для любого $x \neq 0$ имеем

$$\|Ax\| \leqslant \|A\| \|x\|, \tag{8}$$

причем можно доказать, что существует такой вектор $x \neq 0$, для которого неравенство (8) обращается в равенство. Очевидно, соотношение (8) обращается в равенство и для $x = 0$.

На основании (8) для любого $x \in \mathbf{R}^n$ выполняются неравенства

$$\|ABx\| = \|A(Bx)\| \leqslant \|A\| \|Bx\| \leqslant \|A\| \|B\| \|x\|$$

и, следовательно,

$$\|A^2x\| \leqslant \|A\|^2 \|x\|, \ \ldots, \ \|A^kx\| \leqslant \|A\|^k \|x\|,$$

т. е.

$$\|A^k\| \leqslant \|A\|^k. \tag{9}$$

Рассмотрим систему линейных алгебраических уравнений

$$Ax = b, \tag{10}$$

где $\det A \neq 0$, $b \neq 0$. Система (10) имеет единственное решение $x \neq 0$. На практике при ее решении любым методом, в том числе и методом Гаусса, вычисления производятся с округлением, т. е. неточно.

Погрешности вычислений часто можно интерпретировать как погрешности правой части. Наряду с системой (10) рассмотрим систему

$$A(x + r) = b + \eta, \tag{11}$$

где $\eta \neq 0$ — погрешность (возмущение) правой части, r — погрешность решения. Имеем $Ar = \eta$, $r = A^{-1}\eta$,

$$\frac{\|r\|/\|x\|}{\|\eta\|/\|b\|} = \frac{\|b\|}{\|x\|} \frac{\|r\|}{\|\eta\|} = \frac{\|Ax\|}{\|x\|} \frac{\|A^{-1}\eta\|}{\|\eta\|} \leqslant$$

$$\leqslant \frac{\|A\| \|x\|}{\|x\|} \frac{\|A^{-1}\| \|\eta\|}{\|\eta\|} = \|A\| \|A^{-1}\| = \nu(A),$$

т. е. для отношения относительной погрешности решения $\|r\|/\|x\|$ к относительной погрешности правой

части $\|\boldsymbol{\eta}\|/\|\boldsymbol{b}\|$ выполняется неравенство

$$\frac{\|\boldsymbol{r}\|/\|\boldsymbol{x}\|}{\|\boldsymbol{\eta}\|/\|\boldsymbol{b}\|} \leqslant \nu(A) = \|A\|\|A^{-1}\|, \qquad (12)$$

которое при некоторых $\boldsymbol{b} \neq \boldsymbol{0}$, $\boldsymbol{\eta} \neq \boldsymbol{0}$ обращается в равенство.

Величина

$$\nu(A) = \|A\|\|A^{-1}\| \qquad (13)$$

называется *мерой обусловленности матрицы* A. Она равна максимально возможному коэффициенту усиления относительной погрешности от правой части к решению системы (10). Если матрица A — симметричная и выбрана вторая норма, т. е. $\|A\| = \|A\|_2$, то согласно (7)

$$\nu(A) = \max_{1 \leqslant i \leqslant n} |\lambda_i(A)| / \min_{1 \leqslant i \leqslant n} |\lambda_i(A)|. \qquad (14)$$

Если мера $\nu(A)$ большая, то матрица A (система (10)) называется *плохо обусловленной*, а если мера $\nu(A)$ невелика, то — *хорошо обусловленной*.

Пример плохо обусловленной системы уравнений. Рассмотрим систему уравнений (10), в которой $\boldsymbol{b} = (-1, -1, \ldots, -1, 1)$,

$$A = \begin{bmatrix} 1 & -1 & -1 & \ldots & -1 & -1 \\ 0 & 1 & -1 & \ldots & -1 & -1 \\ 0 & 0 & 1 & \ldots & -1 & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 & -1 \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{bmatrix}, \qquad (15)$$

причем $\det A = 1 \neq 0$. В более подробной записи эта система такова:

$$\begin{aligned} x_1 - x_2 - x_3 - \ldots - x_n &= -1, \\ x_2 - x_3 - \ldots - x_n &= -1, \\ \ldots\ldots\ldots\ldots\ldots & \\ x_{n-1} - x_n &= -1, \\ x_n &= 1. \end{aligned} \qquad (16)$$

Система (16) имеет единственное решение $\boldsymbol{x} = (0, 0, \ldots, 0, 1)$. При нахождении решения системы (16) методом Гаусса прямой ход отсутствует.

Допустим, что в обратном ходе была допущена единственная погрешность, сводящаяся к тому, что

вместо $x_n = 1$ было получено $\tilde{x}_n = x_n + r_n = 1 + \varepsilon$, где $\varepsilon \neq 0$ мало по сравнению с единицей. Тогда вместо решения системы (16) будет найдено решение $\tilde{x} = x + r$ системы (11), где матрица A и вектор b имеют указанный выше вид, $\eta = (0, 0, \ldots, 0, \varepsilon)$.

Погрешность $r = (r_1, r_2, \ldots, r_n)$ удовлетворяет системе уравнений

$$\begin{aligned} r_1 - r_2 - r_3 - \ldots - r_n &= 0, \\ r_2 - r_3 - \ldots - r_n &= 0, \\ \ldots\ldots\ldots\ldots & \\ r_{n-1} - r_n &= 0, \\ r_n &= \varepsilon. \end{aligned}$$

Отсюда получаем

$$\begin{aligned} r_n &= \varepsilon, \\ r_{n-1} &= r_n = \varepsilon, \\ r_{n-2} &= r_n + r_{n-1} = \varepsilon + \varepsilon = 2\varepsilon, \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ r_{n-k} &= r_n + r_{n-1} + \ldots + r_{n-(k-1)} = 2^{k-1}\varepsilon, \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ r_1 &= r_{n-(n-1)} = 2^{(n-1)-1}\varepsilon = 2^{n-2}\varepsilon. \end{aligned}$$

Имеем $\|r\|_1 = 2^{n-2}|\varepsilon|$, $\|x\|_1 = 1$, $\|\eta\|_1 = |\varepsilon|$, $\|b\|_1 = 1$ и согласно (12)

$$\nu(A) = \nu_1(A) = \|A\|_1 \|A^{-1}\|_1 \geqslant \frac{\|r\|_1 / \|x\|_1}{\|\eta\|_1 / \|b\|_1} = 2^{n-2}. \quad (17)$$

В соответствии с (5) находим $\|A\|_1 = n$, где A — матрица (15). Следовательно, наряду с мерой обусловленности $\nu(A)$ велика норма обратной матрицы, т. е. $\|A^{-1}\|_1$, хотя $\det A^{-1} = 1/\det A = 1$. Например, если $n = 102$, то $\|A\|_1 = 102$ и на основании (17) $\nu(A) \geqslant 2^{100} > 10^{30}$, а $\|A^{-1}\|_1 > 10^{27}$. В данном случае $\|r\|_1 = 2^{100}|\varepsilon| > 10^{30}|\varepsilon|$. В частности, если $|\varepsilon| = 10^{-15}$ (т. е. единственная допущенная погрешность в обратном ходе достаточно мала), то тем не менее погрешность найденного решения велика: $\|r\|_1 > 10^{15}$. Данному явлению сопутствует плохая обусловленность рассматриваемой системы уравнений, устанавливаемая неравенством (17).

Пример хорошо обусловленной системы приводится в § 21.

§ 21. Метод простых итераций и метод Зейделя

Метод простых итераций. Рассмотрим систему линейных уравнений вида

$$x = Bx + b, \tag{1}$$

где B — заданная числовая квадратная матрица n-го порядка, $b \in \mathbf{R}^n$ — заданный вектор (свободный член).

Метод простых итераций состоит в следующем. Выбирается произвольный вектор $x^0 \in \mathbf{R}^n$ (начальное приближение) и строится итерационная последовательность векторов по формуле

$$x^k = Bx^{k-1} + b, \quad k = 1, 2, \ldots \tag{2}$$

Приведем теорему, дающую достаточное условие сходимости метода простых итераций.

Теорема 1. *Если* $\|B\| < 1$, *то система уравнений* (1) *имеет единственное решение* x^* *и итерации* (*последовательные приближения*) (2) *сходятся к решению со скоростью геометрической прогрессии.*

Доказательство. Допустим, что x^* — решение системы (1), т. е. выполняется равенство

$$x^* = Bx^* + b. \tag{3}$$

Отсюда с помощью неравенства треугольника для нормы вектора и неравенства (20.8) получаем

$$\|x^*\| \leqslant \|Bx^*\| + \|b\| \leqslant \|B\| \|x^*\| + \|b\|,$$

т. е. $(1 - \|B\|)\|x^*\| \leqslant \|b\|$ или, поскольку $1 - \|B\| > 0$,

$$\|x^*\| \leqslant \frac{\|b\|}{1 - \|B\|}. \tag{4}$$

Из этого неравенства вытекает единственность решения однородной системы $x = Bx$, т. е. при $\|b\| = 0$, а следовательно, существование и единственность решения системы (1) при любом свободном члене b.

Пусть $r^k = x^* - x^k$, где x^* — решение системы (1). Вычитая из равенства (3) равенство (2), находим

$$x^* - x^k = B(x^* - x^{k-1}), \tag{5}$$

т. е. $r^k = Br^{k-1}$ и, следовательно, $r^k = B^k r^0$, где B^k — k-я степень матрицы B, r^k — погрешность k-й итера-

ции, r^0 — начальная погрешность. Отсюда на основании (20.8), (20.9) получаем

$$\|r^k\| \leqslant \|B^k\| \|r^0\| \leqslant \|B\|^k \|r^0\|,$$

т. е. норма погрешности $r^k = x^* - x^k$ стремится к нулю при $k \to \infty$ не медленнее геометрической прогрессии со знаменателем $q = \|B\| < 1$.

Замечания. 1. Пусть r_i^k — i-я компонента погрешности r^k k-й итерации (k-го приближения) x^k. Поскольку $|r_i^k| \leqslant \|r^k\|$, $i = 1, 2, \ldots, n$, то все компоненты r_i^k стремятся к нулю при $k \to \infty$ не медленнее, чем $\|B\|^k \|r^0\|$.

2. В силу неравенств (20.2) из сходимости итераций по одной из двух введенных норм векторов вытекает сходимость по другой норме, т. е. если $\|r^k\|_1 = \|x^* - x^k\|_1 \to 0$ при $k \to \infty$, то $\|r^k\|_2 = \|x^* - x^k\|_2 \to 0$ при $k \to \infty$ и наоборот.

Оценка погрешности итераций. Пусть $\|B\| < 1$. Тогда по теореме 1 система (1) имеет единственное решение x^*, для которого выполняется равенство (3). Из равенств (2), (3) вытекает равенство

$$x^* - x^{k-1} = x^k - x^{k-1} + B(x^* - x^{k-1}).$$

Отсюда получаем

$$\|x^* - x^{k-1}\| \leqslant \|x^k - x^{k-1}\| + \|B(x^* - x^{k-1})\| \leqslant$$
$$\leqslant \|x^k - x^{k-1}\| + \|B\| \|x^* - x^{k-1}\|,$$

или

$$(1 - \|B\|) \|x^* - x^{k-1}\| \leqslant \|x^k - x^{k-1}\|,$$

и поскольку $1 - \|B\| > 0$, то

$$\|x^* - x^{k-1}\| \leqslant \frac{1}{1 - \|B\|} \|x^k - x^{k-1}\|. \tag{6}$$

Кроме того, в силу (5) имеем

$$\|x^* - x^k\| \leqslant \|B\| \|x^* - x^{k-1}\|. \tag{7}$$

Из (6), (7) окончательно получаем

$$\|x^* - x^k\| \leqslant \frac{\|B\|}{1 - \|B\|} \|x^k - x^{k-1}\|. \tag{8}$$

Неравенство (8) позволяет оценить норму погрешности k-го приближения через норму разности двух последовательных приближений и норму матрицы B. Эта оценка широко используется на практике. Для

вычисления нормы $\|B\|$ или получения ее оценки используются формулы (20.5)—(20.7).

Примеры:

1) $$B = \begin{bmatrix} -3/5 & 3/5 \\ 2/5 & 1/5 \end{bmatrix},$$

$$\|B\|_1 = \max_{1 \leqslant i \leqslant 2} \sum_{j=1}^{2} |b_{ij}| = \frac{6}{5} > 1,$$

$$\|B\|_2 \leqslant \left(\sum_{i,j=1}^{2} b_{ij}^2 \right)^{1/2} = \frac{\sqrt{23}}{5} < 1;$$

2) $$B = \begin{bmatrix} 1/10 & 4/5 \\ 3/5 & -1/5 \end{bmatrix},$$

$$\|B\|_1 = \frac{9}{10} < 1, \quad \|B\|_2 \leqslant \left(\sum_{i,j=1}^{2} b_{ij}^2 \right) = \sqrt{\frac{21}{20}},$$

т. е. данная оценка для $\|B\|_2$ не представляет интереса, так как правая часть больше 1.

В силу замечания 2 для сходимости итераций (2) достаточно, чтобы хотя бы одна из норм матрицы B была меньше 1.

Приведем без доказательства теорему, дающую необходимое и достаточное условие сходимости метода простых итераций.

Теорема 2. *Пусть система* (1) *имеет единственное решение. Последовательные приближения* (2) *сходятся к решению системы* (1) *при любом начальном приближении* x^0 *тогда и только тогда, когда все собственные значения матрицы* B *по модулю меньше* 1.

Эта теорема дает более общие условия сходимости метода простых итераций, чем теорема 1, так как может случиться, что условия теоремы 2 выполнены, а условия теоремы 1 не выполнены. Однако теоремой 2 в общем случае непросто воспользоваться, так как требуется знание границ собственных значений матрицы B. В частном случае, когда матрица симметричная, в § 23 дан метод нахождения максимального и минимального собственных значений матрицы, позволяющий проверить условия теоремы 2 и, в частности, вычислить вторую норму матрицы по формуле (20.7).

Пример хорошо обусловленной системы уравнений. Система уравнений (1) эквивалентна системе уравнений $x - Bx = b$, или, что то же самое, $Ax = b$, где $A = E - B$, E — единичная матрица. При условии $\|B\| < 1$ согласно теореме 1 $\det A \neq 0$ и, следовательно, $x^* = A^{-1}b$ является решением системы (1). Отсюда

$$\|x^*\| \leqslant \|A^{-1}\| \|b\|, \tag{9}$$

причем существует свободный член $b \neq 0$, при котором неравенство (9) обращается в равенство. Отсюда и из неравенства (4) вытекает, что если $\|B\| < 1$, то

$$\|A^{-1}\| \leqslant \frac{1}{1 - \|B\|}.$$

Принимая во внимание (20.4), получаем

$$\|A\| = \|E - B\| \leqslant \|E\| + \|B\| = 1 + \|B\|.$$

Таким образом, в соответствии с (20.13) мера обусловленности матрицы $A = E - B$, отвечающей системе (1), при условии $\|B\| < 1$ удовлетворяет неравенству

$$\nu(A) = \nu(E - B) = \|A\| \|A^{-1}\| \leqslant \frac{1 + \|B\|}{1 - \|B\|}. \tag{10}$$

Например, пусть элементами матрицы B n-го порядка являются числа

$$b_{ij} = \frac{0{,}8}{n}(-1)^{i+j}, \quad 0 \leqslant i, j \leqslant n.$$

Тогда

$$\|B\|_1 = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |b_{ij}| = \sum_{j=1}^{n} \frac{0{,}8}{n} = 0{,}8,$$

$$\|B\|_2 \leqslant \left(\sum_{i,\, j=1}^{n} b_{ij}^2 \right)^{1/2} = \left(\sum_{i,\, j=1}^{n} \left(\frac{0{,}8}{n}(-1)^{i+j} \right)^2 \right)^{1/2} =$$

$$= \left(n^2 \left(\frac{0{,}8}{n} \right)^2 \right)^{1/2} = 0{,}8.$$

Таким образом, согласно (10)

$$\nu(A) = \nu(E - B) \leqslant \frac{1 + 0{,}8}{1 - 0{,}8} = 9,$$

т. е. мера обусловленности матрицы A невелика при любом n. Соответствующая система (1) является хо-

рошо обусловленной, и, следовательно, ее решение не будет сильно чувствительным к возмущениям свободного члена $\boldsymbol{b}$, в отличие, например, от системы (20.10) с матрицей (20.15).

Метод Зейделя. Пусть дана система линейных уравнений

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{11}$$

где у матрицы $A = (a_{ij})_{i,\,j=1}^{n}$ все диагональные элементы отличны от нуля, т. е. $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$. Если i-е уравнение системы (11) разделить на a_{ii} $i = 1, 2, \ldots, n$, а затем все неизвестные, кроме x_i, перенести вправо, то мы придем к эквивалентной системе вида

$$\boldsymbol{x} = C\boldsymbol{x} + \boldsymbol{d}, \tag{12}$$

где $\boldsymbol{d} = (d_1, d_2, \ldots, d_n)$, $d_i = b_i/a_{ii}$, $C = (c_{ij})_{i,\,j=1}^{n}$,

$$c_{ij} = \begin{cases} -a_{ij}/a_{ii}, & j \neq i, \\ 0, & j = i. \end{cases}$$

Метод Зейделя состоит в том, что итерации производятся по формуле

$$x_i^k = \sum_{j=1}^{i-1} c_{ij}x_j^k + \sum_{j=i+1}^{n} c_{ij}x_j^{k-1} + d_i, \tag{13}$$

где x_i^0 произвольны, $i = 1, 2, \ldots, n$, $k = 1, 2, \ldots$

Итерации (13) по методу Зейделя отличаются от простых итераций (2) тем, что при нахождении i-й компоненты k-го приближения сразу используются уже найденные компоненты k-го приближения с меньшими номерами.

Условия сходимости методов простых итераций и Зейделя не совпадают, но пересекаются. В некоторых случаях метод Зейделя дает более быструю сходимость. Сформулируем теорему о достаточных условиях сходимости метода Зейделя.

Теорема 3. *Для существования единственного решения системы* (11) *и сходимости метода Зейделя достаточно выполнения хотя бы одного из двух условий*:

а) $\sum\limits_{j \neq i} |a_{ij}| < |a_{ii}|$, $i = 1, 2, \ldots, n$;

б) *матрица A — симметричная положительно определенная (все ее собственные значения положительны).*

Замечание 3. Для проверки условия б) может быть использован метод нахождения минимального собственного значения симметричной матрицы, данный в § 23.

§ 22. Метод прогонки

Рассмотрим систему линейных алгебраических уравнений следующего специального вида:

$$A_j z_{j-1} - C_j z_j + B_j z_{j+1} = F_j, \quad j = 1, 2, \ldots, N-1, \quad (1)$$

$$z_0 = \varkappa_0 z_1 + \nu_0, \quad z_N = \varkappa_N z_{N-1} + \nu_N, \quad (2)$$

где $z_0, z_1, \ldots, z_N$ — неизвестные, $A_j, B_j, C_j, F_j, \varkappa_i, \nu_i$ — заданные числа, причем

$$|C_j| \geqslant |A_j| + |B_j| \geqslant |A_j| > 0, \quad |\varkappa_0| < 1, \quad |\varkappa_N| \leqslant 1. \quad (3)$$

Систему (1), (2) можно записать также в векторном виде $Az = b$, где $z = (z_0, z_1, \ldots, z_N)$, $b = (\nu_0, F_1, F_2, \ldots, F_{N-1}, \nu_N)$,

$$A = \begin{bmatrix} 1 & -\varkappa_0 & & & & \\ A_1 & -C_1 & B_1 & & 0 & \\ & A_2 & -C_2 & B_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & A_{N-1} & -C_{N-1} & B_{N-1} \\ & 0 & & & -\varkappa_N & 1 \end{bmatrix}.$$

Таким образом, матрица A системы (1), (2) является *трехдиагональной*, т. е. все ее элементы, не лежащие на главной диагонали и двух соседних диагоналях, равны нулю. Это вызвано тем, что в каждое уравнение системы (1), (2) входят только неизвестные с номерами, отличающимися от номера уравнения не более чем на единицу. При выполнении условий (3) говорят, что рассматриваемая матрица A является матрицей с *доминирующей (преобладающей) главной диагональю.*

Уравнения (1) в совокупности обычно называются *разностным уравнением второго порядка* или *трех-*

точечным разностным уравнением, а уравнения (2) — *краевыми условиями* для разностного уравнения (1).

Задача (1), (2) в целом называется *краевой задачей для трехточечного разностного уравнения* или просто *разностной краевой задачей*.

Разностные краевые задачи (системы линейных алгебраических уравнений) вида (1), (2) являются весьма распространенными. Например, система уравнений, полученная в § 11 при построении сплайна, имеет вид (1), (2) и удовлетворяет условиям (3). Разностные краевые задачи наиболее часто возникают при приближенном решении дифференциальных уравнений (§ 29, 31, 33).

Условия (3) гарантируют существование единственного решения разностной краевой задачи (1), (2) (в чем мы убедимся ниже) и позволяют найти это решение специальным экономичным методом, называемым *методом прогонки*.

Метод прогонки. Подставив из (2) $z_0 = \varkappa_0 z_1 + \nu_0$ в первое уравнение системы (1), получим

$$A_1(\varkappa_0 z_1 + \nu_0) - C_1 z_1 + B_1 z_2 = F_1$$

или

$$z_1 = \varkappa_1 z_2 + \nu_1, \tag{4}$$

где

$$\varkappa_1 = \frac{B_1}{C_1 - A_1\varkappa_0}, \quad \nu_1 = \frac{A_1\nu_0 - F_1}{C_1 - A_1\varkappa_0}.$$

Найденное выражение (4) для z_1 подставим в следующее уравнение системы (1) и получим уравнение, связывающее z_2 и z_3, и т. д.

Допустим, что уже найдено соотношение

$$z_{k-1} = \varkappa_{k-1} z_k + \nu_{k-1}, \quad k < N-1. \tag{5}$$

Подставим z_{k-1} в виде (5) в k-е уравнение системы (1):

$$A_k(\varkappa_{k-1} z_k + \nu_{k-1}) - C_k z_k + B_k z_{k+1} = F_k.$$

Разрешив это уравнение относительно z_k, получим

$$z_k = \varkappa_k z_{k+1} + \nu_k, \tag{6}$$

где

$$\varkappa_k = \frac{B_k}{C_k - A_k\varkappa_{k-1}}, \quad \nu_k = \frac{A_k\nu_{k-1} - F_k}{C_k - A_k\varkappa_{k-1}}. \tag{7}$$

Таким образом, коэффициенты уравнений (6), связывающих соседние значения z_k и z_{k+1}, $k = 1, 2, \ldots, N-1$, можно определить из рекуррентных соотношений (7), поскольку $\varkappa_0$, ν_0 заданы в (2).

Подставив во второе краевое условие (2) выражение z_{N-1}, вытекающее из формулы (6) при $k = N-1$, получим

$$z_N = \varkappa_N(\varkappa_{N-1}z_N + \nu_{N-1}) + \nu_N, \tag{8}$$

где $\varkappa_N$, ν_N — заданные в (2) коэффициенты, $\varkappa_{N-1}$, ν_{N-1} вычислены по формулам (7).

Из уравнения (8) находим неизвестную z_N:

$$z_N = \frac{\nu_N + \varkappa_N\nu_{N-1}}{1 - \varkappa_N\varkappa_{N-1}}. \tag{9}$$

Затем по формуле (6) в обратном порядке находим остальные неизвестные z_{N-1}, z_{N-2}, ..., z_0. Формула (6) при $k = 0$ совпадает с первым краевым условием (2).

Процесс вычисления коэффициентов $\varkappa_k$, ν_k, $k = 1, 2, \ldots, N-1$, по формулам (7) называется *прямым ходом прогонки*, а нахождение неизвестных z_m, $m = N, N-1, \ldots, 0$, по формулам (9), (6) — *обратным ходом прогонки*.

В силу условий (3) вычисления по формулам (7), (9) корректны, т. е. в них знаменатели не обращаются в нуль. Убедимся в этом.

Допустим, что при некотором k, $0 < k < N-1$, выполняется неравенство $|\varkappa_{k-1}| < 1$. Например, это неравенство по условию (3) выполнено при $k = 1$. Поскольку $|A_k| > 0$, то, используя условия (3), получаем

$$|C_k - A_k\varkappa_{k-1}| \geqslant |C_k| - |A_k||\varkappa_{k-1}| > |C_k| - |A_k| \geqslant 0,$$

$$|\varkappa_k| = \frac{|B_k|}{|C_k - A_k\varkappa_{k-1}|} \leqslant \frac{|C_k| - |A_k|}{|C_k| - |A_k||\varkappa_{k-1}|} < 1.$$

Отсюда по индукции вытекает, что

$$|\varkappa_k| < 1, \quad k = 0, 1, \ldots, N-1, \tag{10}$$

и, следовательно, знаменатели в выражениях (7) не равны нулю. Поскольку по условию (3) $|\varkappa_N| \leqslant 1$ и по доказанному $|\varkappa_{N-1}| < 1$, то $1 - \varkappa_N\varkappa_{N-1} > 0$, т. е. знаменатель в формуле (9) тоже не равен нулю.

Лемма 1. *Разностная краевая задача* (1), (2) *при условиях* (3) *имеет единственное решение.*

Доказательство. Получив явные формулы (7), (9), (6) для нахождения решения задачи (1), (2) при условиях (3), мы тем самым установили ее разрешимость. Единственность решения вытекает из разрешимости при любых свободных членах ν_1, ν_N, F_j, $j = 1, 2, \ldots, N-1$.

Действительно, если определитель системы линейных алгебраических уравнений равен нулю, то по теореме Кронекера — Капелли такая система имеет решение не при любых свободных членах. Следовательно, у системы (1), (2) определитель не равен нулю, поэтому она имеет единственное решение. Лемма доказана.

Подсчитаем число арифметических действий, выполняемых при решении разностной краевой задачи (1), (2) методом прогонки. По формулам (7), реализуемым с помощью шести арифметических действий, вычисления производятся $N-1$ раз, $k = 1, 2, \ldots, N-1$, по формуле (9) выполняется пять арифметических действий и, наконец, по формуле (6), требующей всего два действия, вычисления осуществляются N раз, $k = N-1, N-2, \ldots, 0$. Итак, в методе прогонки всего затрачивается

$$Q = 6(N-1) + 5 + 2N = 8(N+1) - 9 \qquad (11)$$

арифметических действий, т. е. число действий растет линейно относительно числа неизвестных, равного $N+1$.

При решении же произвольной системы линейных алгебраических уравнений методом Гаусса число действий, выражаемое формулой (19.18), приблизительно пропорционально кубу числа неизвестных.

Докажем две леммы, дающие оценки решения задачи (1), (2) в двух частных случаях. Эти оценки используются в § 29.

Лемма 2. *Если* $\varkappa_0 = \varkappa_N = \nu_0 = \nu_N = 0$, $A_j = B_j = 1$, $C_j \geqslant 2$, $j = 1, 2, \ldots, N-1$, *то решение* $z = (z_0, z_1, \ldots, z_N)$ *разностной краевой задачи* (1), (2) *удовлетворяет неравенству*

$$\max_{0 \leqslant j \leqslant N} |z_j| \leqslant N^2 \max_{1 \leqslant j \leqslant N-1} |F_j|. \qquad (12)$$

Доказательство. Условия (3) выполнены. Следовательно, по лемме 1 задача (1), (2) имеет единственное решение $z = (z_0, z_1, \ldots, z_N)$, где в силу (2) $z_0 = z_N = 0$. При этом задача (1), (2) может быть решена методом прогонки.

Согласно (10) $|\varkappa_k| < 1$, $k = 0, 1, \ldots, N-1$. Следовательно, при $k = 1, 2, \ldots, N-1$

$$0 < \frac{A_k}{C_k - A_k\varkappa_{k-1}} = \frac{1}{C_k - A_k\varkappa_{k-1}} \leqslant \frac{1}{2 - \varkappa_{k-1}} < 1.$$

Поэтому в соответствии с (7)

$$|v_k| \leqslant |v_{k-1}| + |F_k|, \quad k = 1, 2, \ldots, N-1.$$

Поскольку $v_0 = 0$, то

$$|v_k| \leqslant \sum_{i=1}^{k} |F_i|, \quad k = 1, 2, \ldots, N-1,$$

а значит,

$$\max_{1 \leqslant j \leqslant N-1} |v_j| \leqslant \sum_{i=1}^{N-1} |F_i| \leqslant N \max_{1 \leqslant j \leqslant N-1} |F_j|.$$

Отсюда и из (6), принимая во внимание, что $|\varkappa_k| < 1$, $k = N-1, N-2, \ldots, 1$, $z_0 = z_N = 0$, аналогично получаем неравенство (12).

Лемма 3. *Если* $\varkappa_0 = \varkappa_N = 0$, $F_j = 0$, $A_j = B_j = 1$, $C_j \geqslant 2$, $j = 1, 2, \ldots, N-1$, *то решение разностной краевой задачи* (1), (2) *удовлетворяет соотношению*

$$\max_{0 \leqslant j \leqslant N} |z_j| = \max\{|v_0|, |v_N|\}. \tag{13}$$

Доказательство. Решение $z = (z_0, z_1, \ldots, z_N)$ задачи (1), (2) в силу выполнения условий (3) по лемме 1 существует и единственно. Очевидно, при некотором k справедливо равенство

$$|z_k| = \max_{0 \leqslant j \leqslant N} |z_j|. \tag{14}$$

Если $k = 0$ или N, то отсюда и из (2), где $\varkappa_0 = \varkappa_N = 0$, сразу следует (13).

Допустим, что $0 < k < N$. Тогда согласно (1) при условиях леммы 3 имеем

$$z_{k-1} - C_k z_k + z_{k+1} = 0, \tag{15}$$

где $C_k \geqslant 2$. В силу (14) это равенство возможно при $C_k = 2$, только если $z_{k+1} = z_k = z_{k-1}$, а при $C_k > 2$

из (14) и (15) с необходимостью следует, что $z_{k+1} = z_k = z_{k-1} = 0$. Таким образом, наряду с (14) имеем

$$|z_{k+1}| = \max_{0 \leqslant j \leqslant N} |z_j|.$$

Аналогично убеждаемся, что $z_{k+2} = z_{k+1} = z_k$ и т. д. Через конечное число шагов мы придем к заключению, что $z_N = z_k$. Отсюда и из (14), (2), где $\varkappa_0 = \varkappa_N = 0$, вытекает равенство (13).

Замечание. Свойство решения задачи (1), (2) (при условиях леммы 3), выражаемое равенством (13), называется *принципом максимума*. Согласно принципу максимума решение принимает максимальное по модулю значение хотя бы в одной крайней точке (узле), ибо $z_0 = \nu_0$, $z_N = \nu_N$.

§ 23. Частичные проблемы собственных значений

Задача нахождения всех собственных значений и собственных векторов числовой матрицы называется *полной проблемой собственных значений*. Эта задача в общем случае достаточно сложная. Мы остановимся только на некоторых важных частных задачах.

Нахождение максимального по модулю собственного значения. Предположим, что квадратная матрица A порядка n имеет полную систему нормированных собственных векторов $e_1, e_2, \ldots, e_n$, т. е. в частности,

$$Ae_i = \lambda_i e_i, \tag{1}$$

$$\|e_i\| = \|e_i\|_2 = (e_i, e_i)^{1/2} = 1, \tag{2}$$

где (x, y) — скалярное произведение в n-мерном евклидовом векторном пространстве, λ_i — собственное значение матрицы A, отвечающее собственному вектору e_i, $i = 1, 2, \ldots, n$. Например, полная система собственных векторов существует, если матрица A — симметричная.

Допустим, что

$$|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|. \tag{3}$$

Зададимся произвольным вектором $x^0 \neq 0$. Имеем

$$x^0 = c_1 e_1 + c_2 e_2 + \ldots + c_n e_n,$$

где $c_1, c_2, \ldots, c_n$ — координаты вектора x^0 в базисе $e_1, e_2, \ldots, e_n$.

Предположим, что

$$c_1 \neq 0. \tag{4}$$

Последовательно находим векторы

$$x^k = Ax^{k-1}, \quad k = 1, 2, \ldots \tag{5}$$

Тогда согласно (1)

$$x^1 = Ax^0 = A(c_1e_1 + c_2e_2 + \ldots + c_ne_n) =$$
$$= c_1Ae_1 + c_2Ae_2 + \ldots + c_nAe_n =$$
$$= c_1\lambda_1e_1 + c_2\lambda_2e_2 + \ldots + c_n\lambda_ne_n.$$

И вообще,

$$x^k = c_1\lambda_1^k e_1 + c_2\lambda_2^k e_2 + \ldots + c_n\lambda_n^k e_n = \lambda_1^k (c_1e_1 + \eta^k), \tag{6}$$

где $\eta^k = c_2(\lambda_2/\lambda_1)^k e_2 + \ldots + c_n(\lambda_n/\lambda_1)^k e_n$, причем в силу (3)

$$\|\eta^k\| = O(|\lambda_2/\lambda_1|^k) \to 0, \quad k \to \infty. \tag{7}$$

Здесь и ниже верхний индекс у векторов, например x^k, указывает номер приближения, у скаляров, например λ_1^k, верхний индекс означает степень, а в качестве нормы векторов принята вторая норма, т. е. $\|x\| = \|x\|_2 = (x, x)^{1/2}$.

Принимая во внимание (6), получаем следующее выражение скалярного произведения (x^k, x^{k-1}):

$$(x^k, x^{k-1}) = \lambda_1^{2k-1}(c_1e_1 + \eta^k, c_1e_1 + \eta^{k-1}) =$$
$$= \lambda_1^{2k-1}[c_1^2(e_1, e_1) + c_1(e_1, \eta^{k-1}) +$$
$$+ c_1(\eta^k, e_1) + (\eta^k, \eta^{k-1})]. \tag{8}$$

Поскольку согласно (2) $(e_1, e_1) = \|e_1\|^2 = 1$ и по неравенству Буняковского *)

$$|(e_1, \eta^{k-1})| \leqslant \|e_1\|\|\eta^{k-1}\| = \|\eta^{k-1}\|,$$
$$|(\eta^k, e_1)| \leqslant \|\eta^k\|\|e_1\| = \|\eta^k\|,$$
$$|(\eta^k, \eta^{k-1})| \leqslant \|\eta^k\|\|\eta^{k-1}\|,$$

*) Неравенство Буняковского $|(x, y)| \leqslant \|x\|\|y\|$ доказывается в курсе линейной алгебры.

то из (8) с учетом (7) находим

$$(x^k, x^{k-1}) = \lambda_1^{2k-1}(c_1^2 + O(|\lambda_2/\lambda_1|^{k-1})).$$

Аналогично получаем

$$(x^m, x^m) = \lambda_1^{2m}(c_1^2 + O(|\lambda_2/\lambda_1|^m)).$$

Следовательно,

$$\lambda_{1k} = \frac{(x^k, x^{k-1})}{(x^{k-1}, x^{k-1})} = \lambda_1 + O(|\lambda_2/\lambda_1|^{k-1}), \qquad (9)$$

$$\|x^k\| = (x^k, x^k)^{1/2} = |\lambda_1|^k(|c_1| + O(|\lambda_2/\lambda_1|^k)), \qquad (10)$$

$$e_1^k = x^k/\|x^k\| = (\operatorname{sign}\lambda_1)^k e_1' + r^k, \qquad (11)$$

где $e_1' = (\operatorname{sign} c_1)e_1$, $\|r^k\| = O(|\lambda_2/\lambda_1|^k)$.

Таким образом, при условии (3) итерационный процесс (5) позволяет найти с любой точностью максимальное по модулю собственное значение λ_1 (ибо $\lambda_{1k} \to \lambda_1$ при $k \to \infty$) и соответствующий ему собственный вектор e_1', так как $\|r^k\| \to 0$ при $k \to \infty$.

Скалярные произведения (x^k, x^{k-1}), (x^{k-1}, x^{k-1}), (x^k, x^k), входящие в формулы (9), (10), на практике вычисляются для векторов x^k, x^{k-1}, представленных в исходном ортонормированном базисе, в котором задан оператор A, т. е. по обычной формуле

$$(x, y) = x_1y_1 + x_2y_2 + \ldots + x_ny_n.$$

Замечания. 1. Если $|\lambda_1| > 1$, то $\|x^k\| \to \infty$, а если $|\lambda_1| < 1$, то $\|x^k\| \to 0$ при $k \to \infty$. То и другое явление при счете на ЭВМ нежелательно. В первом случае в ЭВМ может наступить переполнение (выход за допустимый диапазон чисел) и в результате произойдет остановка счета. Во втором случае $\|x^k\|$ может стать машинным нулем (слишком малой величиной) и информация теряется. Поэтому целесообразно итерации вести по формулам

$$e_1^0 = x^0/\|x^0\|,$$

$$x^k = Ae_1^{k-1}, \quad \lambda_{1k} = (x^k, e_1^{k-1}), \quad e_1^k = x^k/\|x^k\|, \qquad (9^*)$$

не имеющим указанного недостатка и дающим тот же результат (те же e_1^k, λ_{1k}), что и формулы (5), (9), (11).

2. Если при выборе x^0 не будет выполнено (4), что маловероятно, то за счет погрешностей округлений через несколько итераций, как правило, появится ненулевая компонента, отвечающая e_1. Тогда, хотя с некоторым запаздыванием, итерационный процесс выйдет на первое собственное значение.

3. Подтверждением того, что λ_1 не является кратным собственным значением и что нет собственного значения, равного $-\lambda_1$, служит сходимость итерационного процесса при выборе различных x^0 к одному и тому же собственному вектору (с точностью до противоположного вектора).

Теперь рассматриваем только случай, когда матрица A — симметричная. Тогда, как известно, заведомо все ее собственные значения действительны и существует ортонормированный базис $e_1, e_2, \ldots, e_n$, составленный из собственных векторов матрицы A.

Пусть E — единичная матрица,

$$P_2(t) = a_0 + a_1 t + a_2 t^2$$

есть некоторый алгебраический многочлен от t второй степени с действительными коэффициентами, B — матрица, выражающаяся через матрицу A следующим образом:

$$B = a_0 E + a_1 A + a_2 A^2.$$

Легко убедиться, что собственные значения матриц A и B связаны соотношением

$$\lambda_i(B) = P_2(\lambda_i(A)), \tag{12}$$

а собственные векторы матрицы A, отвечающие $\lambda_i(A)$, являются собственными векторами матрицы B, отвечающими собственному значению $\lambda_i(B)$.

Действительно, пусть e_i — собственный вектор матрицы A, отвечающий собственному значению $\lambda_i(A)$, т. е. $Ae_i = \lambda_i(A) e_i$. Тогда

$$A^2 e_i = A(Ae_i) = A\lambda_i(A) e_i = \lambda_i(A) Ae_i = \lambda_i^2(A) e_i$$

и, следовательно,

$$\begin{aligned} Be_i = (a_0 E + a_1 A + a_2 A^2) e_i = \\ = a_0 E e_i + a_1 A e_i + a_2 A^2 e_i = \\ = a_0 e_i + a_1 \lambda_i(A) e_i + a_2 \lambda_i^2(A) e_i = \\ = (a_0 + a_1 \lambda_i(A) + a_2 \lambda_i^2(A)) e_i = P_2(\lambda_i(A)) e_i. \end{aligned}$$

Остановимся еще на трех задачах.

Нахождение λ_2, e_2. Пусть

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|,$$

причем λ_1, e_1 известны, $\|e_1\| = 1$.

Задаем итерационный процесс:

x^0 — произвольный вектор $((x^0, e_2) \neq 0)$,

$y^0 = x^0 - (x^0, e_1) e_1$, $e_2^0 = y^0 / \|y^0\|$,

$x^k = A e_2^{k-1}$,

$\lambda_{2k} = (x^k, e_2^{k-1})$,

$y^k = x^k - (x^k, e_1) e_1$ (ортогонализация к e_1),

$e_2^k = y^k / \|y^k\|$ (нормирование), $k = 1, 2, \ldots$

Так же, как и выше, с дополнительным учетом ортонормированности базиса $e_1, e_2, \ldots, e_n$ можно доказать, что

$$\lambda_{2k} = \lambda_2 + O\left(|\lambda_3/\lambda_2|^{2k-1}\right),$$

$$e_2^k = (\operatorname{sign} \lambda_2)^k e_2' + \rho^k,$$

где $e_2' = e_2$ или $-e_2$, $\|\rho^k\| = O\left(|\lambda_3/\lambda_2|^{2k}\right)$.

Нахождение $\max\limits_{1 \leqslant i \leqslant n} \lambda_i(A)$, $\min\limits_{1 \leqslant i \leqslant n} \lambda_i(A)$. Допустим, что максимальное по модулю собственное значение $\bar{\lambda}(A)$ матрицы A известно. Находим максимальное по модулю собственное значение $\bar{\lambda}(B)$ матрицы

$$B = A - \bar{\lambda}(A) E \tag{13}$$

способом, изложенным в начале параграфа.

Если $\bar{\lambda}(A) > 0$, то, очевидно,

$$\max_{1 \leqslant i \leqslant n} \lambda_i(A) = \bar{\lambda}(A). \tag{14}$$

Кроме того, согласно (12), (13)

$$\lambda_i(B) = \lambda_i(A) - \bar{\lambda}(A) \leqslant 0, \quad i = 1, 2, \ldots, n.$$

Поэтому

$$\bar{\lambda}(B) = \min_{1 \leqslant i \leqslant n} (\lambda_i(A) - \bar{\lambda}(A)) = \min_{1 \leqslant i \leqslant n} \lambda_i(A) - \bar{\lambda}(A),$$

т. е.

$$\min_{1 \leqslant i \leqslant n} \lambda_i(A) = \bar{\lambda}(A) + \bar{\lambda}(B). \tag{15}$$

Если $\bar{\lambda}(A) < 0$, то

$$\min_{1 \leqslant i \leqslant n} \lambda_i(A) = \bar{\lambda}(A)$$

и согласно (12), (13)

$$\max_{1 \leqslant i \leqslant n} \lambda_i(A) = \bar{\lambda}(A) + \bar{\lambda}(B). \tag{16}$$

Замечание 4. Заранее обычно неизвестно, имеются ли у симметричной матрицы A максимальные по модулю собственные значения с противоположными знаками или кратное максимальное по модулю собственное значение. Об этих случаях можно судить на основании вычислений следующим образом. Если при различных начальных векторах x^0 значения λ_{1k}, вычисляемые по формулам (9*), сходятся не к одному и тому же значению, то это свидетельствует о наличии максимальных по модулю собственных значений с противоположными знаками. Тогда следует сместить спектр, взяв матрицу $A' = A + cE$, где $c \neq 0$ — число. Если при различных начальных векторах x^0 значения λ_{1k} сходятся к одному и тому же числу, но последовательность векторов e_1^k, получаемых по формулам (9*), приводит к неколлинеарным векторам e_1', то это обстоятельство служит подтверждением того, что максимальное по модулю собственное значение кратно. При этом λ_{1k} сходится к указанному собственному значению, а получаемые векторы e_1' являются собственными.

Нахождение расстояния ρ_0 от заданной точки $\lambda = \lambda_0$ до ближайшего собственного значения матрицы A. Эта задача возникает при изучении явлений типа резонанса. Рассмотрим случай, представляющий наибольший интерес:

$$\min_{1 \leqslant i \leqslant n} \lambda_i(A) < \lambda_0 < \max_{1 \leqslant i \leqslant n} \lambda_i(A).$$

Обозначим

$$\rho_0 = \min_{1 \leqslant i \leqslant n} |\lambda_i(A) - \lambda_0|,$$

$$l = \max\{\max_{1 \leqslant i \leqslant n} \lambda_i(A) - \lambda_0,\ \lambda_0 - \min_{1 \leqslant i \leqslant n} \lambda_i(A)\}. \tag{17}$$

Докажем, что

$$\rho_0 = l\sqrt{1-\bar{\lambda}(B)}, \tag{18}$$

где $\bar{\lambda}(B)$ — максимальное по модулю собственное значение матрицы

$$B = E - \frac{1}{l^2}(A-\lambda_0 E)^2. \tag{19}$$

Согласно (12), (19), (17) имеем

$$\lambda_i(B) = 1 - \frac{1}{l^2}(\lambda_i(A)-\lambda_0)^2 \geqslant 0,\ i = 1,\ 2,\ \ldots,\ n.$$

Отсюда

$$\bar{\lambda}(B) = 1 - \frac{1}{l^2}(\lambda^*(A)-\lambda_0)^2 = 1 - \frac{\rho_0^2}{l^2},$$

где $\lambda^*(A)$ — ближайшее к λ_0 собственное значение матрицы A, т. е. верно (18).

ГЛАВА 4

МЕТОДЫ РЕШЕНИЯ НЕЛИНЕЙНЫХ УРАВНЕНИЙ И СИСТЕМ

Излагаются методы итераций, Ньютона, деления отрезка пополам и наискорейшего (градиентного) спуска.

§ 24. Метод итераций

Пусть дано уравнение с одной неизвестной x:

$$x = \varphi(x), \tag{1}$$

где φ — заданная функция от x. Если не накладывать никаких ограничений на функцию φ, то могут возникнуть различные ситуации, а именно, уравнение (1) может иметь либо одно решение, либо некоторое конечное число решений (больше одного), либо бесконечное множество решений, либо, наконец, уравнение (1) может совсем не иметь решений. Обычно возникают сразу два вопроса: о наличии решений и о том, как найти решения.

Ниже формулируется и доказывается теорема, которая дает достаточные условия существования на некотором отрезке единственного решения уравнения (1). Эта теорема указывает также способ нахождения приближенного решения, называемый *методом итераций*, и обеспечивает оценки погрешности приближенного решения.

Определение. Говорят, что функция φ удовлетворяет на отрезке $[a, b]$ *условию Липшица* с постоянной α, если для любых x_1, $x_2 \in [a, b]$ выполняется неравенство

$$|\varphi(x_1) - \varphi(x_2)| \leqslant \alpha |x_1 - x_2|. \tag{2}$$

Замечание 1. В частности, если функция φ непрерывно дифференцируема на отрезке $[a, b]$, то она удовлетворяет на $[a, b]$ условию Липшица с постоянной

$$\alpha = \max_{[a, b]} |\varphi'(x)|, \tag{3}$$

что легко следует из формулы конечных приращений Лагранжа.

Теорема 1. *Пусть функция* φ *удовлетворяет на отрезке* $[x_0, x_0+r]$ *условию Липшица с постоянной* α, *причем*

$$0<\alpha<1, \tag{4}$$

$$0\leqslant\varphi(x_0)-x_0\leqslant(1-\alpha)r. \tag{5}$$

Тогда уравнение (1) *имеет на отрезке* $[x_0, x_0+r]$ *единственное решение*

$$x_*=\lim_{k\to\infty}x_k, \tag{6}$$

где x_0 — *левый конец отрезка* $[x_0, x_0+r]$,

$$x_k=\varphi(x_{k-1}), \quad k=1, 2, \ldots \tag{7}$$

При этом имеют место оценки

$$|x_*-x_k|\leqslant\rho\alpha^k, \tag{8}$$

$$|x_*-x_k|\leqslant\frac{\alpha}{1-\alpha}|x_k-x_{k-1}|, \tag{9}$$

где $\rho=\dfrac{\varphi(x_0)-x_0}{1-\alpha}\leqslant r, \quad k=1, 2, \ldots$

Доказательство. Установим сначала, что рекуррентная числовая последовательность $\{x_k\}$ действительно может быть найдена по формуле (7) и что эта последовательность целиком расположена на отрезке $[x_0, x_0+\rho]$, принадлежащем заданному отрезку $[x_0, x_0+r]$.

Предположим для простоты, что $x_0=0$. Тогда отрезок $[x_0, x_0+\rho]$ совпадет с отрезком $[0, \rho]$ и с учетом выбора ρ будет выполняться равенство

$$\varphi(0)=(1-\alpha)\rho. \tag{10}$$

В неравенстве (2), которому по условию удовлетворяет функция φ на отрезке $[0, r]$ и, в частности, на отрезке $[0, \rho]$, положим $x_1=x$, $x_2=0$ и заменим полученное таким образом неравенство на следующие два эквивалентных на отрезке $[0, \rho]$ неравенства для функции φ:

$$\varphi(0)-\alpha x\leqslant\varphi(x)\leqslant\varphi(0)+\alpha x. \tag{11}$$

Из равенства (10) и правого неравенства (11) вытекает, что функция φ подчиняется на отрезке $[0, \rho]$

следующему ограничению:

$$\varphi(x) \leqslant \rho. \tag{12}$$

Допустим, что члены последовательности $\{x_k\}$ с номерами $k = 0, 1, \ldots, m-1$ уже найдены с помощью формулы (7) и удовлетворяют условию

$$0 \leqslant x_k \leqslant \rho. \tag{13}$$

Например, при $m = 2$ сделанные предположения заведомо выполнены, так как точки $x_0 = 0$, $x_1 = \varphi(x_0) = \varphi(0)$, являющиеся членами рассматриваемой последовательности, в силу (10) подчиняются неравенствам (13).

Коль скоро $x_{m-1} \in [0, \rho]$, то по формуле (7), где $k = m$, может быть найден следующий член x_m. Покажем, что он тоже удовлетворяет условию (13). Рассмотрим два случая.

1. $0 \leqslant x_{m-1} \leqslant \min\{\rho, \varphi(0)/\alpha\}$. Используя равенство (7) при $k = m$, а также левое неравенство (11), получаем

$$x_m = \varphi(x_{m-1}) \geqslant \varphi(0) - \alpha x_{m-1} \geqslant \varphi(0) - \\ - \alpha \min\left\{\rho, \frac{\varphi(0)}{\alpha}\right\} \geqslant \varphi(0) - \alpha \frac{\varphi(0)}{\alpha} = 0. \tag{14}$$

2. $\varphi(0)/\alpha < x_{m-1} \leqslant \rho$. Вычитая из равенства (7) с $k = m$ равенство (7) с $k = m-1$ и принимая во внимание условие Липшица (2), находим

$$|x_m - x_{m-1}| = |\varphi(x_{m-1}) - \varphi(x_{m-2})| \leqslant \alpha |x_{m-1} - x_{m-2}|.$$

Далее, аналогично получаем

$$|x_{m-1} - x_{m-2}| \leqslant \alpha |x_{m-2} - x_{m-3}|$$

и т. д. Через конечное число шагов мы придем к неравенству

$$|x_m - x_{m-1}| \leqslant \alpha^{m-1} |x_1 - x_0|. \tag{15}$$

Отсюда, учитывая, что

$$|x_1 - x_0| = \varphi(0), \quad x_{m-1} > \varphi(0)/\alpha \geqslant \alpha^{m-1}\varphi(0)$$

(ибо $0 < \alpha < 1$), находим

$$x_m \geqslant x_{m-1} - |x_m - x_{m-1}| > 0. \tag{16}$$

Кроме рассмотренных случаев 1, 2, для x_{m-1}, удовлетворяющего условию (13) при $k = m-1$, других

случаев быть не может. Следовательно, из (14) и (16) вытекает левое неравенство (13) при $k = m$. Правое неравенство (13) при $k = m$ следует из равенства $x_m = \varphi(x_{m-1})$, где $x_{m-1} \in [0, \rho]$, и из неравенства (12).

Из доказанного по индукции следует, что при $x_0 = 0$ бесконечная последовательность $\{x_k\}$ может быть действительно построена по формуле (7) и все ее члены удовлетворяют условию (13). Если $x_0 \neq 0$, то совершенно аналогично доказывается, что последовательность $\{x_k\}$, найденная по формуле (7), целиком расположена на отрезке $[x_0, x_0 + \rho]$.

Установим теперь, что рассматриваемая последовательность $\{x_k\}$ — фундаментальная. Учитывая неравенство (15), которое очевидно справедливо при любом натуральном m и произвольном x_0, и неравенство $|x_1 - x_0| = |\varphi(x_0) - x_0| < r$, вытекающее из условий (4), (5), находим

$$|x_{n+p} - x_n| \leqslant \sum_{m=n+1}^{n+p} |x_m - x_{m-1}| \leqslant \sum_{m=n+1}^{n+p} \alpha^{m-1} |x_1 - x_0| <$$

$$< r \sum_{m=n+1}^{n+p} \alpha^{m-1} = r\alpha^n \frac{1-\alpha^p}{1-\alpha} < \alpha^n \frac{r}{1-\alpha}, \qquad (17)$$

где n, p — любые натуральные числа. Поскольку $\alpha^n \to 0$ при $n \to \infty$, то соотношения (17) показывают, что последовательность $\{x_k\}$ — фундаментальная. Поэтому существует предел (6), причем, так как последовательность $\{x_k\}$ расположена на отрезке $[x_0, x_0 + \rho]$, то $x_* \in [x_0, x_0 + \rho]$.

Заданная функция φ удовлетворяет условию Липшица (2) на отрезке $[x_0, x_0 + \rho]$, которое, в частности, означает, что функция φ непрерывна на этом отрезке. Это позволяет перейти к пределу при $k \to \infty$ в равенстве (7), осуществив предельный переход справа под знаком функции φ. В результате мы получим равенство

$$x_* = \varphi(x_*), \qquad (18)$$

говорящее о том, что $x_* \in [x_0, x_0 + \rho]$ является решением уравнения (1). Существование решения уравнения (1) на отрезке $[x_0, x_0 + \rho]$, а значит, и на отрезке $[x_0, x_0 + r]$, содержащем отрезок $[x_0, x_0 + \rho]$, доказано.

Допустим, что точка $x_{**} \in [x_0,\ x_0 + r]$ тоже является решением уравнения (1), т. е.

$$x_{**} = \varphi(x_{**}). \tag{19}$$

Вычитая из (18) равенство (19) и учитывая условие Липшица (2), получаем

$$|x_* - x_{**}| = |\varphi(x_*) - \varphi(x_{**})| \leqslant \alpha |x_* - x_{**}|,$$

т. е.

$$|x_* - x_{**}| \leqslant \alpha |x_* - x_{**}|.$$

Это неравенство возможно при условии (4), только если $x_{**} = x_*$. Единственность решения уравнения (1) на отрезке $[x_0, x_0 + r]$ тоже доказана.

Остается установить оценки (8), (9) для погрешности $x_* - x_k$ приближенного решения x_k уравнения (1). Используя равенства (18) и (7), а также условие Липшица (2), находим

$$|x_* - x_k| = |\varphi(x_*) - \varphi(x_{k-1})| \leqslant \alpha |x_* - x_{k-1}| =$$
$$= \alpha |\varphi(x_*) - \varphi(x_{k-2})| \leqslant \alpha^2 |x_* - x_{k-2}| \leqslant \ldots \leqslant \alpha^k |x_* - x_0|. \tag{20}$$

Поскольку $x_* \in [x_0,\ x_0 + \rho]$, то $|x_* - x_0| \leqslant \rho$. Отсюда и из (20) следует оценка (8).

Из равенств (7), (18) следует равенство

$$x_* - x_{k-1} = x_k - x_{k-1} + \varphi(x_*) - \varphi(x_{k-1}),$$

из которого с помощью условия Липшица (2) получаем

$$|x_* - x_{k-1}| \leqslant |x_k - x_{k-1}| + \alpha |x_* - x_{k-1}|,$$

или

$$|x_* - x_{k-1}| \leqslant \frac{1}{1-\alpha} |x_k - x_{k-1}|.$$

Отсюда, воспользовавшись первым неравенством в цепочке неравенств (20), приходим к оценке (9). Теорема полностью доказана.

Пример. Рассмотрим уравнение

$$x = \frac{1}{2}\left(x + \frac{a}{x}\right),$$

где a — число, $1/2 \leqslant a < 1$. Это уравнение является уравнением вида (1) с $\varphi(x) = \frac{1}{2}\left(x + \frac{a}{x}\right)$. Оно имеет решение $x_* = \sqrt{a}$, в чем нетрудно убедиться непосредственно.

Попытаемся применить теорему 1. Положим $x_0=a$, $r=1-a$, т. е. выберем $[x_0, x_0+r]=[a, 1]$. Имеем при $1/2 \leqslant a < 1$

$$\max_{[a, 1]} |\varphi'(x)| = \max_{[a, 1]} \left| \frac{1}{2}\left(1-\frac{a}{x^2}\right)\right| = \frac{1-a}{2a} \leqslant \frac{1}{2}.$$

Следовательно, функция φ удовлетворяет на отрезке $[a, 1]$ условию Липшица (2) с $\alpha = 1/2$. Таким образом, условие (4) выполнено.

Проверим условие (5). Находим

$$\varphi(x_0) - x_0 = \frac{1}{2}(a+1) - a = \frac{1}{2}1-a) = (1-\alpha)r.$$

Условие (5) тоже выполнено. Теорема 1 гарантирует, что на отрезке $[a, 1]$ рассматриваемое уравнение имеет единственное решение, и именно то, о котором мы уже догадались, т. е. $x_* = \sqrt{a}$. Для его вычисления может быть применен в соответствии с (7) метод итераций:

$$x_0 = a, \quad x_k = \frac{1}{2}\left(x_{k-1} + \frac{a}{x_{k-1}}\right), \quad k = 1, 2, \ldots$$

Погрешность целесообразно оценивать в процессе вычислений по неравенству (9), которое более точное, чем (8).

Данный способ вычисления квадратного корня, сводящийся к арифметическим действиям, применяется в ЭВМ, в которых отсутствует элементарная операция извлечения квадратного корня.

Замечания. 2. Справедлив другой вариант теоремы 1, когда отрезок $[x_0, x_0+r]$ заменяется на отрезок $[x_0-r, x_0]$, вместо условия (5) фигурирует условие

$$0 \leqslant x_0 - \varphi(x_0) \leqslant (1-\alpha)r, \qquad (5^*)$$

а

$$\rho = \frac{x_0 - \varphi(x_0)}{1-\alpha}.$$

3. Итерации (7) имеют геометрическую интерпретацию. Решение x_* уравнения (1) является абсциссой точки пересечения прямой $y=x$ и кривой $y=\varphi(x)$. Сходящиеся итерации изображены на рис. 12. 13. Геометрически видно, что если в окрестности решения x_* выполняются неравенства $0 < \varphi'(x) \leqslant \alpha < 1$, то последовательность $\{x_k\}$ монотонно сходится к x_*, причем с той стороны, с которой расположено начальное приближение (см. рис. 12). В случае $-1 < -\alpha \leqslant \varphi'(x) < 0$ последовательные приближения расположены поочередно с разных сторон от решения x_* (см. рис. 13). В последнем случае очень просто можно су-

дить по двум последовательным приближениям о достигнутой точности, а именно, уклонение x_k от x_* не превышает $|x_k - x_{k-1}|$. Легко усмотреть также, что сходимость тем быстрее, чем меньше $|\varphi'|$.

4. Если функция φ, входящая в уравнение (1), не удовлетворяет условию Липшица с постоянной $\alpha < 1$, то итерации (7) могут расходиться. Например, рас-

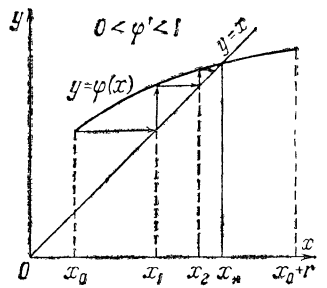


Рис. 12

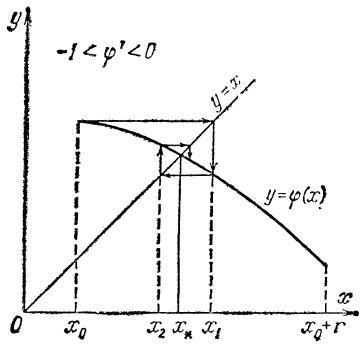


Рис. 13

смотрим уравнение (1), где $\varphi(x) = bx$, $b > 1$, b — число. Очевидно, функция φ удовлетворяет на всей оси x условию Липшица с постоянной $\alpha = b > 1$ и не удовлетворяет условию Липшица ни с какой постоянной меньше единицы на любом отрезке. Рассматриваемое уравнение $x = bx$ имеет единственное решение $x_* = 0$. Однако при любом $x_0 \neq 0$ согласно (7) $x_k = b^k x_0 \to \infty$ при $k \to \infty$.

Решение системы уравнений. Введем в n-мерном векторном пространстве $\mathbf{R}^n$ расстояние*)

$$\rho(x, y) = \|x - y\| \tag{21}$$

с помощью нормы (20.1) (см. п. 8 введения). В соответствии с (20.1), (21) имеем

$$\rho(x, y) = \begin{cases} \rho_1(x, y) = \max\limits_{1 \leqslant j \leqslant n} |x_j - y_j|, \\ \rho_2(x, y) = \left(\sum\limits_{j=1}^{n} (x_j - y_j)^2 \right)^{1/2}. \end{cases} \tag{22}$$

*) При чтении данной главы предполагается знакомство с § 1.5 книги: Бугров Я. С., Никольский С. М. Дифференциальные уравнения. Кратные интегралы. Ряды. Функции комплексного переменного. — М.: Наука, 1985.

Множество $\bar{S}(y^0; r) = \{x: \rho(x, y^0) \leqslant r\}$ называется *замкнутым шаром* радиуса r с центром в точке y^0 (индексы у векторов будем располагать вверху). Опираясь на критерий Коши для числовых последовательностей, можно доказать, что любой замкнутый шар $\bar{S}(y^0; r)$, а также все пространство $\mathbb{R}^n$ является полным метрическим пространством с метрикой (22).

Рассмотрим уравнение (систему n, вообще говоря, нелинейных уравнений с n неизвестными)

$$x = \varphi(x), \tag{23}$$

где $\varphi(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_n(x))$ — заданная вектор-функция переменной $x = (x_1, x_2, \ldots, x_n)$ (переменных $x_1, x_2, \ldots, x_n$), $\varphi_i(x) = \varphi_i(x_1, x_2, \ldots, x_n)$, $i = 1, 2, \ldots, n$. Например, при $n = 2$ система (23) в более подробной записи имеет вид

$$(x_1, x_2) = (\varphi_1(x_1, x_2), \varphi_2(x_1, x_2))$$

или

$$x_1 = \varphi_1(x_1, x_2), \quad x_2 = \varphi_2(x_1, x_2),$$

где $\varphi_1(x_1, x_2)$, $\varphi_2(x_1, x_2)$ — заданные функции переменных x_1, x_2.

Следующая теорема обосновывает метод итераций решения нелинейной системы уравнений (23).

Теорема 2. *Пусть на замкнутом шаре* $\bar{S} = \bar{S}(y^0; r)$ *задана вектор-функция* $\varphi(x)$, *причем для любых* $x, y \in \bar{S}$ *выполняется неравенство*

$$\rho(\varphi(x), \varphi(y)) \leqslant \alpha\rho(x, y) \tag{24}$$

и, кроме того,

$$\rho(\varphi(y^0), y^0) \leqslant (1 - \alpha) r, \tag{25}$$

где α — *число,* $0 \leqslant \alpha < 1$. *Тогда на* $\bar{S}$ *существует единственное решение* x^* *уравнения* (23), *причем*

$$x^* = \lim_{k \to \infty} x^k, \tag{26}$$

где $x^0 \in \bar{S}$ *произвольно,*

$$x^k = \varphi(x^{k-1}), \quad k = 1, 2, \ldots \tag{27}$$

При этом выполняются неравенства

$$\rho(x^*, x^k) \leqslant \alpha^k \rho(x^*, x^0) \leqslant 2\alpha^k r, \tag{28}$$

$$\rho(x^*, x^k) \leqslant \frac{\alpha}{1 - \alpha} \rho(x^k, x^{k-1}). \tag{29}$$

Доказательство. Пусть $x \in \bar{S}(y^0; r)$, т. е. $\rho(x, y^0) \leqslant r$. Тогда, применяя неравенство треугольника для расстояния, на основании (24), (25) получаем

$$\rho(\varphi(x), y^0) \leqslant \rho(\varphi(x), \varphi(y^0)) + \rho(\varphi(y^0), y^0) \leqslant$$
$$\leqslant \alpha\rho(x, y^0) + (1-\alpha) r \leqslant r.$$

Таким образом, если $x \in \bar{S}(y^0; r)$, то и $\varphi(x) \in \bar{S}(y^0; r)$, т. е. оператор $\varphi(x)$ отображает полное метрическое пространство $\bar{S}(y^0; r)$ в себя. Кроме того, по условию (24) он является сжимающим. Следовательно, по принципу сжатых отображений существует на $\bar{S}$ единственная неподвижная точка x^* оператора $\varphi(x)$, являющаяся решением уравнения (системы) (23).

Соотношения (26), (28)', (29), кроме правого неравенства в (28), установлены в § 1.5 указанной выше книги. Неравенство $\alpha^k \rho(x^*, x^0) \leqslant 2\alpha^k r$ вытекает из неравенства треугольника:

$$\rho(x^*, x^0) \leqslant \rho(x^*, y^0) + \rho(y^0, x^0) \leqslant r + r = 2r.$$

Теорема доказана.

Неравенства (28) свидетельствуют о том, что итерации (26) сходятся к искомому решению x^* по геометрической прогрессии со знаменателем $\alpha < 1$. Поскольку известны r, α, то можно предсказать достаточное число итераций k, при котором погрешность $\rho(x^*, x^k)$ будет меньше заданного $\varepsilon > 0$. Достаточно потребовать выполнения неравенства $2\alpha^k r < \varepsilon$ или $\alpha^k < \varepsilon/(2r)$. Отсюда

$$k > \ln \frac{\varepsilon}{2r} \frac{1}{\ln \alpha},$$

так как $\ln \alpha < 0$. Минимальное k находится по формуле

$$k = \max\left\{0, \left[\frac{1}{\ln \alpha} \ln \frac{\varepsilon}{2r}\right] + 1\right\}, \qquad (30)$$

где $[a]$ — целая часть от a.

Для получения оценки погрешности $\rho(x^*, x^k)$ в процессе вычислений отдают предпочтение неравенству (29), правая часть которого тоже выражается через известные величины. Обычно оно дает более точную оценку, чем (28). Это связано с тем, что погрешность

на некоторых промежуточных итерациях может убывать быстрее, чем в α раз. Оценка (28) этого не почувствует, а правая часть в (29) автоматически учтет это обстоятельство.

Замечания. 5. Если оператор $\varphi(x)$ является сжимающим во всем пространстве $\mathbf{R}^n$, т. е. условие (24) с $\alpha < 1$ выполнено для любых $x, y \in \mathbf{R}^n$, то без дополнительного условия (25) из принципа сжатых отображений вытекают существование решения системы (23) и его единственность в $\mathbf{R}^n$. Однако полезно и в этой ситуации ввести условие (25), которое, если задать некоторое y^0 (по возможности ближе к искомому решению x^*) и положить $r = \rho(\varphi(y^0), y^0)/(1 - \alpha)$, будет выполнено. Тогда, еще не решая системы (23), можно утверждать, что единственное в $\mathbf{R}^n$ решение x^* находится на самом деле в замкнутом шаре $\bar{S}(y^0; r)$, и возможно воспользоваться оценками (28), (30).

6. Система линейных алгебраических уравнений (21.1) является частным случаем системы (23) с оператором $\varphi(x) = Bx + b$, где B — заданная матрица, b — заданный вектор. Если $\|B\| < 1$, то данный оператор является сжимающим во всем пространстве $\mathbf{R}^n$. Действительно, пусть произвольные $x, y \in \mathbf{R}^n$. Тогда в силу (21), (20.8)

$$\rho(\varphi(x), \varphi(y)) = \|\varphi(x) - \varphi(y)\| = \|Bx - By\| =$$

$$= \|B(x - y)\| \leqslant \|B\| \|x - y\| = \|B\| \rho(x, y),$$

т. е.

$$\rho(\varphi(x), \varphi(y)) \leqslant \alpha \rho(x, y),$$

где $\alpha = \|B\| < 1$. Следовательно, для последовательных приближений (21.2) справедлива оценка (29), которая при замене $\rho(x^*, x^k)$ на $\|x^* - x^k\|$ и $\rho(x^k, x^{k-1})$ на $\|x^k - x^{k-1}\|$ обращается в неравенство (21.8).

7. Теорема 2 при $n = 1$ не совпадает с теоремой 1. Разница состоит в следующем. В теореме 1 уравнение (1) рассматривается либо только на отрезке $[x_0, x_0 + r]$, либо только на отрезке $[x_0 - r, x_0]$ (см. замечание 2) в зависимости от знака разности $\varphi(x_0) - x_0$. При этом не требуется, чтобы функция φ была определена на другом отрезке. В теореме 2 при $n = 1$ предполагается, что функция φ с нужными

свойствами задана на отрезке $[x_0 - r,\ x_0 + r]$ (здесь y^0 заменено на x_0).

Оценки величины α. Предположим, что вектор-функция $\varphi(x) = (\varphi_1(x),\ \varphi_2(x),\ \ldots,\ \varphi_n(x))$ имеет непрерывные частные производные по $x_1,\ x_2,\ \ldots,\ x_n$ на замкнутом шаре $\bar{S} = \bar{S}(y^0; r)$. Обозначим

$$\alpha_{ij} = \max_{\bar{S}} |\partial\varphi_i/\partial x_j|. \tag{31}$$

Пусть $x,\ y \in \bar{S}$. Согласно формуле конечных приращений Лагранжа имеем

$$\varphi_i(x) - \varphi_i(y) = \sum_{j=1}^{n} \frac{\partial\varphi_i(\xi^i)}{\partial x_j}(x_j - y_j), \tag{32}$$

$$i = 1,\ 2,\ \ldots,\ n,$$

где $\xi^i \in \bar{S}$ — некоторая промежуточная точка.

С помощью матрицы

$$\Phi = \begin{bmatrix} \dfrac{\partial\varphi_1(\xi^1)}{\partial x_1} & \cdots & \dfrac{\partial\varphi_1(\xi^1)}{\partial x_n} \\ \cdots\cdots\cdots\cdots\cdots \\ \dfrac{\partial\varphi_n(\xi^n)}{\partial x_1} & \cdots & \dfrac{\partial\varphi_n(\xi^n)}{\partial x_n} \end{bmatrix} \tag{33}$$

соотношения (32) можно объединить в одно векторное равенство

$$\varphi(x) - \varphi(y) = \Phi(x - y),$$

где справа матрица Φ умножается на вектор $x - y$. Отсюда в соответствии с (21), (20.8) находим

$$\rho(\varphi(x),\ \varphi(y)) = \|\varphi(x) - \varphi(y)\| = \|\Phi(x - y)\| \leqslant$$
$$\leqslant \|\Phi\| \|x - y\| = \|\Phi\| \rho(x,\ y),$$

т. е.

$$\rho_m(\varphi(x),\ \varphi(y)) \leqslant \|\Phi\|_m \rho_m(x,\ y), \quad m = 1,\ 2,$$

где в зависимости от выбора расстояния по формуле (22) первым или вторым способом берется соответствующая норма матрицы (33).

Далее, воспользовавшись соотношениями (20.5), (20.6), с учетом (31) получаем

$$\rho_m(\varphi(x),\ \varphi(y)) \leqslant \alpha_m \rho_m(x,\ y), \quad m = 1,\ 2, \tag{34}$$

где

$$\alpha_1 = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} \alpha_{ij}, \quad \alpha_2 = \left(\sum_{i,\, j=1}^{n} \alpha_{ij}^2 \right)^{1/2}. \qquad (35)$$

Если при $m = 1$ или 2 окажется, что $\alpha_m < 1$, то при соответствующем выборе метрики будет выполнено условие (24) теоремы 2 с $\alpha = \alpha_m$.

Пример. Требуется выяснить существование решения системы уравнений

$$x_1 = \varphi_1(x_1, x_2) = 1{,}1 - \sin \frac{x_2}{3} + \ln\left(1 + \frac{x_1 + x_2}{5}\right),$$
$$x_2 = \varphi_2(x_1, x_2) = 0{,}5 + \cos \frac{x_1 x_2}{6} \qquad (36)$$

в окрестности точки $y^0 = (y_1^0, y_2^0) = (1, 1)$.

Решение. Выбираем метрику $\rho_1(x, y) = \max\{|x_1 - y_1|, |x_2 - y_2|\}$ и полагаем $r = 1$. Тогда

$$\overline{S} = \overline{S}(y^0; r) = \{x: \rho_1(x, y^0) \leqslant 1\} = \{x: 0 \leqslant x_i \leqslant 2,\ i = 1, 2\},$$

т. е. замкнутый шар $\overline{S}$ при выбранной конкретной метрике $\rho_1(x, y)$ и $n = 2$ является в обычном понимании замкнутым квадратом.

Находим частные производные

$$\frac{\partial \varphi_1}{\partial x_1} = \frac{1}{x_1 + x_2 + 5}, \quad \frac{\partial \varphi_1}{\partial x_2} = -\frac{1}{3}\cos\frac{x_2}{3} + \frac{1}{x_1 + x_2 + 5},$$
$$\frac{\partial \varphi_2}{\partial x_1} = -\frac{x_2}{6}\sin\frac{x_1 x_2}{6}, \quad \frac{\partial \varphi_2}{\partial x_2} = -\frac{x_1}{6}\sin\frac{x_1 x_2}{6}.$$

Отсюда, учитывая, что $\sin(2/3) < 2/3$, получаем

$$\max_{\overline{S}} \left| \frac{\partial \varphi_1}{\partial x_1} \right| = \frac{1}{5}, \quad \max_{\overline{S}} \left| \frac{\partial \varphi_1}{\partial x_2} \right| \leqslant \max\left\{ \frac{1}{3} - \frac{1}{9}, \frac{1}{5} \right\} = \frac{2}{9},$$

$$\max_{\overline{S}} \left| \frac{\partial \varphi_2}{\partial x_1} \right| < \frac{2}{6}\frac{2}{3} = \frac{2}{9}, \quad \max_{\overline{S}} \left| \frac{\partial \varphi_2}{\partial x_2} \right| < \frac{2}{9}.$$

В соответствии с (34), (35) в качестве α может быть взята величина

$$\alpha = \max\left\{ \frac{1}{5} + \frac{2}{9}, \frac{2}{9} + \frac{2}{9} \right\} = \frac{4}{9}.$$

Таким образом, условие (24) теоремы 2 выполнено с $\alpha = 4/9 < 1$. Выясним, выполнено ли условие (25). Имеем

$$\rho_1(\varphi(y^0), y^0) = \max\{|\varphi_1(1, 1) - 1|, |\varphi_2(1, 1) - 1|\}. \qquad (37)$$

Поскольку $0 < \sin(1/3) < 1/3$, $0 < \ln(1 + 2/5) < 2/5$, $0 < \cos(1/6) < 1$, то из (36), (37) получаем

$$\rho_1(\varphi(y^0), y^0) < 0{,}5 < (1 - \alpha) r = 5/9.$$

Условие (25) тоже выполнено. По теореме 2 система (36) имеет на рассматриваемом замкнутом квадрате $\bar{S}(y^0; 1)$ единственное решение x^*. Выбор y^0 и r оказался удачным.

Итерации (27) проводятся по формулам

$$x_1^k = \varphi_1\left(x_1^{k-1}, x_2^{k-1}\right), \quad x_2^k = \varphi_2\left(x_1^{k-1}, x_2^{k-1}\right).$$

В качестве начального приближения естественно взять $x^0 = y^0 = (1, 1)$, т. е. $x_1^0 = x_2^0 = 1$. Точность итераций согласно (29) оценивается неравенством

$$\rho_1(x^*, x^k) \leqslant \frac{4}{5}\rho_1(x^k, x^{k-1}) = \frac{4}{5}\max\left\{\left|x_1^k - x_1^{k-1}\right|, \left|x_2^k - x_2^{k-1}\right|\right\}.$$

Замечание 8. Одним из условий применимости метода итераций является, грубо говоря, малость частных производных вектор-функции $\varphi(x)$. На практике требуется, чтобы хотя бы одна из величин (35) была меньше 1. Но этого для существования решения еще мало. Возникает вопрос о выборе подходящих y^0, r, при которых выполняются условия (24), (25) теоремы 2. Точку y^0 желательно взять по возможности ближе к решению x^*. Варьирование r приводит к тому, что с увеличением r обычно увеличиваются величины (31) или их оценки сверху и, следовательно, увеличивается вычисляемое α, но при слишком малом r может оказаться невыполнимым условие (25). Задача выбора y^0, r практически трудная и в каждом случае требует индивидуального подхода. Успех не гарантируется, даже если решение у системы (23) существует, тем более что заранее может быть не известно о наличии решения. При отсутствии же решения указанные y^0 и r, естественно, нельзя подобрать.

§ 25. Метод Ньютона

Рассмотрим уравнение с одной неизвестной x

$$f(x) = 0. \tag{1}$$

Теорема 1. *Если* $f \in C_2[a, b]$, $f(a)f(b) < 0$, *т. е.* f *принимает на концах отрезка* $[a, b]$ *значения с противоположными знаками, а* f'' *не меняет знака на* $[a, b]$, *то уравнение* (1) *имеет на* $[a, b]$ *единственное решение (корень)* x_*.

Утверждение теоремы достаточно очевидно. При условиях теоремы 1 возможны четыре случая,

изображенные на рис. 14. Существование решения $x_* \in [a, b]$ следует из непрерывности f на $[a, b]$ и предположения $f(a)f(b) < 0$. Неединственность решения при условии $f(a)f(b) < 0$ повлекла бы изменение знака у f''. Это можно доказать формально, но мы будем в основном ограничиваться геометрическими соображениями.

Примем за x_0 конец отрезка $[a, b]$, в котором функция f имеет тот же знак, что и f'' в тех точках

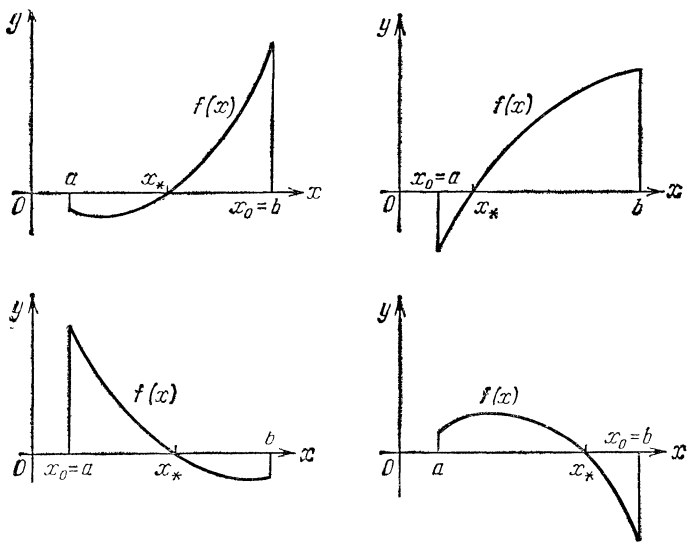


Рис. 14

отрезка $[a, b]$, где $f''(x) \neq 0$ (см. рис. 14). Малоинтересный случай, когда $f''(x) \equiv 0$, исключим.

Обозначим через ω отрезок, концами которого служат точки x_0, x_*. При условиях теоремы 1 функция f' монотонна на ω, причем

$$m_1 = \min_{\omega} |f'(x)| = |f'(x_*)| > 0, \tag{2}$$

так как иначе бы функция f не смогла изменить знака с сохранением знака у f''. Положим

$$M_1 = \max_{\omega} |f'(x)| = |f'(x_0)|, \tag{3}$$

$$M_2 = \max_{[a, b]} |f''(x)| > 0. \tag{4}$$

Метод Ньютона. *Метод Ньютона*, называемый также *методом касательных*, состоит в следующем. Рассмотрим в точке x_0 касательную к кривой $y = f(x)$, задаваемую уравнением

$$Y = f(x_0) + (x - x_0) f'(x_0).$$

Положив $Y = 0$, находим точку x_1 пересечения касательной с осью абсцисс:

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

Построив касательную в точке x_1 (рис. 15), получаем

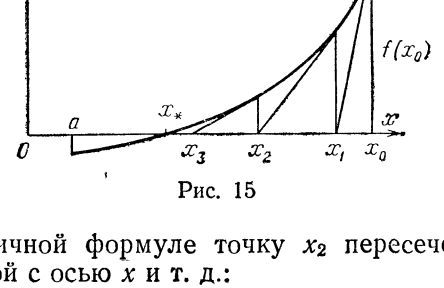


Рис. 15

по аналогичной формуле точку x_2 пересечения этой касательной с осью x и т. д.:

$$x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})}. \tag{5}$$

Из геометрических соображений ясно, что при условиях теоремы 1 итерационная последовательность $\{x_k\}$, полученная по формуле (5), монотонно сходится к искомому решению x_* уравнения (1).

Оценим скорость сходимости. Учитывая, что $f(x_*) = 0$, и применяя формулу конечных приращений Лагранжа, находим

$$|f(x_{k-1})| = |f(x_*) - f(x_{k-1})| = \\ = |f'(\xi)|\,|x_* - x_{k-1}| \geqslant m_1 |x_* - x_{k-1}|,$$

где $\xi \in \omega$ — некоторая точка, m_1 — величина (2). Принимая во внимание это соотношение, а также (3), (5), получаем

$$|x_k - x_{k-1}| = \frac{|f(x_{k-1})|}{|f'(x_{k-1})|} \geqslant \frac{m_1 |x_* - x_{k-1}|}{M_1}. \tag{6}$$

В силу монотонной сходимости последовательности $\{x_k\}$ к x_* имеем

$$|x_{k-1}-x_*|=|x_{k-1}-x_k|+|x_k-x_*|,$$

т. е.

$$|x_*-x_k|=|x_*-x_{k-1}|-|x_k-x_{k-1}|.$$

Отсюда на основании (6) приходим к неравенству

$$|x_*-x_k|\leqslant\alpha|x_*-x_{k-1}|, \tag{7}$$

где $0\leqslant\alpha=1-m_1/M_1<1$.

Неравенство (7) устанавливает, что погрешность x_*-x_k убывает, по крайней мере, по геометрической прогрессии со знаменателем $\alpha<1$. Это характерно для начальных итераций. Затем, когда погрешность становится достаточно малой, скорость сходимости в методе Ньютона увеличивается. Убедимся в этом.

Согласно формуле Тейлора имеем

$$0=f(x_*)=f(x_{k-1})+(x_*-x_{k-1})f'(x_{k-1})+\frac{(x_*-x_{k-1})^2}{2}f''(\xi),$$

где $\xi\in\omega$ — некоторая точка, т. е.

$$x_*=x_{k-1}-\frac{f(x_{k-1})}{f'(x_{k-1})}-(x_*-x_{k-1})^2\frac{f''(\xi)}{2f'(x_{k-1})}. \tag{8}$$

Вычитая из (8) равенство (5) и используя оценку

$$\left|\frac{f''(\xi)}{2f'(x_{k-1})}\right|\leqslant\beta,$$

где $\beta=M_2/(2m_1)$, m_1, M_2 — величины (2), (4) соответственно, приходим к неравенству

$$|x_*-x_k|\leqslant\beta(x_*-x_{k-1})^2.$$

Перепишем это неравенство в виде

$$\beta|x_*-x_k|\leqslant(\beta(x_*-x_{k-1}))^2. \tag{9}$$

Отсюда следует, что, как только при некотором k выполнится неравенство $\beta|x_*-x_k|<1$ (в силу (7) это непременно произойдет), в дальнейшем погрешность, умноженная на β, начнет убывать очень быстро по квадратичному закону. Через n следующих итераций будем иметь

$$|x_*-x_{k+n}|\leqslant\frac{1}{\beta}(\beta(x_*-x_k))^{2^n}. \tag{10}$$

Например, если $\beta|x_*-x_k|=0{,}9$, $n=8$, то $2^n=256$, $(\beta(x_*-x_k))^{256}<10^{-11}$.

Упрощенный метод Ньютона. Если производная $f'(x)$ вычисляется сложно, то вместо формулы (5) используют формулу

$$x_k=x_{k-1}-\frac{f(x_{k-1})}{f'(x_0)}, \quad k=1, 2, \ldots \tag{11}$$

При этом сходимость по геометрической прогрессии, устанавливаемая неравенством (7), сохраняется.

Связь между методом итераций и методом Ньютона. Уравнение (1) при $\lambda\neq 0$ эквивалентно уравнению

$$x=x+\lambda f(x)\equiv\varphi(x).$$

Метод итераций сходится тем быстрее, чем меньше $|\varphi'|$. Потребуем, чтобы $\varphi'(x_{k-1})=0$, т. е. $1+\lambda f'(x_{k-1})=0$ и, следовательно, $\lambda=-1/f'(x_{k-1})$. Тогда в соответствии с (24.7)

$$x_k=\varphi(x_{k-1})=x_{k-1}-\frac{f(x_{k-1})}{f'(x_{k-1})},$$

и мы пришли к методу Ньютона (5).

Решение системы уравнений. Пусть вектор-функция $\boldsymbol{f}(\boldsymbol{x})=(f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x}))$, где $f_i(\boldsymbol{x})=f_i(x_1, x_2, \ldots, x_n)$, дважды непрерывно дифференцируема в некоторой окрестности решения $\boldsymbol{x}^*$ уравнения (системы уравнений)

$$\boldsymbol{f}(\boldsymbol{x})=0. \tag{12}$$

Матрица $F(\boldsymbol{x})$, имеющая вид

$$F(\boldsymbol{x})=\begin{bmatrix} \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \cdots\cdots\cdots & & \\ \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_1} & \cdots & \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_n} \end{bmatrix},$$

называется *матрицей Якоби* системы функций $f_1(\boldsymbol{x})$, $f_2(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x})$ в точке $\boldsymbol{x}$. Обозначим, если $\det F(\boldsymbol{x})\neq 0$, через $F^{-1}(\boldsymbol{x})$ обратную матрицу для матрицы Якоби $F(\boldsymbol{x})$.

Метод Ньютона:

$$\boldsymbol{x}^k=\boldsymbol{x}^{k-1}-F^{-1}(\boldsymbol{x}^{k-1})\boldsymbol{f}(\boldsymbol{x}^{k-1}), \quad k=1, 2, \ldots \tag{13}$$

Упрощенный метод Ньютона:

$$x^k = x^{k-1} - F^{-1}(x^0) f(x^{k-1}). \qquad (14)$$

Упрощение состоит в том, что обратная матрица $F^{-1}(x^0)$ находится один раз, а не в каждой итерации, как в (13).

Если $\det F(x^*) \neq 0$ и начальное приближение x^0 взято достаточно близко к x^*, то итерации (13) и (14) сходятся в метрике (24.22) к x^*. Характер сходимости тот же, что и при $n = 1$, т. е. итерации (13), начиная с некоторого момента, сходятся очень быстро по квадратичному закону, а для итераций (14) гарантируется сходимость только по геометрической прогрессии.

§ 26. Метод деления отрезка пополам

Пусть $f \in C[a, b]$, $f(a)f(b) < 0$ и известно, что уравнение $f(x) = 0$ имеет единственное решение (единственный корень) $x_* \in [a, b]$. Полагаем $a_0 = a$, $b_0 = b$, $c_0 = (a_0 + b_0)/2$, т. е. c_0 — середина отрезка $[a_0, b_0]$. Вычисляем $f(c_0)$. Если $f(c_0) = 0$, то $x_* = c_0$ и вычисления на этом заканчиваются. Если $f(c_0) \neq 0$, то знак $f(c_0)$ совпадает либо со знаком $f(a_0)$, либо со знаком $f(b_0)$, коль скоро $f(a_0)f(b_0) < 0$.

Таким образом, на концах одного из двух отрезков $[a_0, c_0]$ или $[c_0, b_0]$ функция f имеет одинаковые знаки, а на концах другого — противоположные. Сохраняем отрезок, на концах которого f имеет противоположные знаки, а другой отрезок, как не содержащий корень x_*, отбрасываем. Оставленный отрезок обозначим через $[a_1, b_1]$, где

$$a_1 = \begin{cases} c_0, & \operatorname{sign} f(a_0) = \operatorname{sign} f(c_0), \\ a_0, & \operatorname{sign} f(a_0) \neq \operatorname{sign} f(c_0), \end{cases}$$

$$b_1 = \begin{cases} c_0, & \operatorname{sign} f(b_0) = \operatorname{sign} f(c_0), \\ b_0, & \operatorname{sign} f(b_0) \neq \operatorname{sign} f(c_0). \end{cases}$$

Очевидно, $\operatorname{sign} f(a_1) = \operatorname{sign} f(a_0)$ и $\operatorname{sign} f(b_1) = \operatorname{sign} f(b_0)$. Поэтому $f(a_1)f(b_1) < 0$. Искомый корень x_* находится теперь на вдвое меньшем отрезке $[a_1, b_1]$.

Далее поступаем аналогично. Допустим, что уже найден некоторый отрезок $[a_k, b_k] \subset [a, b]$, на концах которого функция f имеет противоположные знаки и, следовательно, он содержит искомый корень x_*. Находим середину отрезка $[a_k, b_k]$:

$$c_k = (a_k + b_k)/2. \tag{1}$$

Вычисляем $f(c_k)$. Если $f(c_k) = 0$, то $x_* = c_k$. Вычисления заканчиваются. Если $f(c_k) \neq 0$, то полагаем

$$a_{k+1} = \begin{cases} c_k, & \operatorname{sign} f(a_k) = \operatorname{sign} f(c_k), \\ a_k, & \operatorname{sign} f(a_k) \neq \operatorname{sign} f(c_k), \end{cases}$$
$$b_{k+1} = \begin{cases} c_k, & \operatorname{sign} f(b_k) = \operatorname{sign} f(c_k), \\ b_k, & \operatorname{sign} f(b_k) \neq \operatorname{sign} f(c_k), \end{cases} \tag{2}$$

и т. д. Этот процесс может быть конечным, если середина отрезка, полученного на некотором шаге, совпадает с искомым корнем x_*, либо этот процесс бесконечный.

На рис. 16 показано несколько начальных шагов. Если вычисления доведены до k-го шага, то в качестве

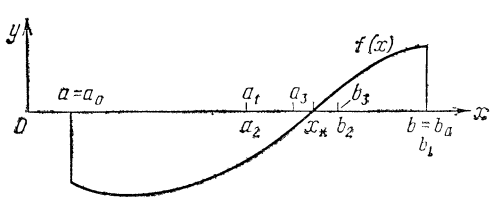


Рис. 16

приближенного значения для искомого корня x_* естественно принять c_k. При этом справедлива очевидная оценка погрешности

$$|x_* - c_k| \leqslant \frac{b-a}{2^{k+1}}. \tag{3}$$

Изложенный метод является типично машинным, так как вычисления по формулам (1), (2) очень простые и цикличные. Он обладает достаточно быстрой сходимостью. На каждом шаге правая часть оценки погрешности (3) убывает вдвое.

§ 27. Метод наискорейшего (градиентного) спуска

Пусть

$$f(x)=(f_1(x),\ f_2(x),\ \ldots,\ f_n(x)),\ x=(x_1,\ x_2,\ \ldots,\ x_n).$$

Система n уравнений, $n \geqslant 2$,

$$f(x)=0 \tag{1}$$

эквивалентна одному уравнению

$$\Psi(x)=0, \tag{2}$$

где $\Psi(x)=f_1^2(x)+f_2^2(x)+\ \ldots\ +f_n^2(x))$. Очевидно, решениями уравнения (2) являются точки нулевых минимумов функции $\Psi(x)=\Psi(x_1,\ x_2,\ \ldots,\ x_n)$.

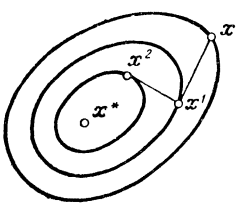


Рис. 17

Допустим, что $\Psi(x)$ дважды непрерывно дифференцируема в области, содержащей изолированное решение x^*, в окрестности которого поверхности уровня функции Ψ имеют вид, изображенный на рис. 17. Задавшись начальным приближением x^0, ищем минимум функции $\Psi(x^0-\lambda\,\mathrm{grad}\,\Psi(x^0))$ одной переменной λ. Фактически находим минимальный неотрицательный корень $\lambda=\lambda_0$ уравнения

$$\frac{d}{d\lambda}\Psi(x^0-\lambda\,\mathrm{grad}\,\Psi(x^0))=0$$

одним из рассмотренных выше способов.

Полагаем

$$x^1=x^0-\lambda_0\,\mathrm{grad}\,\Psi(x^0)$$

и, вообще,

$$x^k=x^{k-1}-\lambda_{k-1}\,\mathrm{grad}\,\Psi(x^{k-1}),\quad k=1,\ 2,\ \ldots, \tag{3}$$

где λ_{k-1} — минимальный неотрицательный корень уравнения

$$\frac{d}{d\lambda}\Psi(x^{k-1}-\lambda\,\mathrm{grad}\,\Psi(x^{k-1}))=0. \tag{4}$$

Сходимость последовательных приближений (3) к решению уравнения (2), вообще говоря, не гарантируется, так как можно попасть в точку относительного минимума.

ГЛАВА 5

МЕТОДЫ РЕШЕНИЯ КРАЕВОЙ ЗАДАЧИ ДЛЯ ЛИНЕЙНОГО ОБЫКНОВЕННОГО ДИФФЕРЕНЦИАЛЬНОГО УРАВНЕНИЯ ВТОРОГО ПОРЯДКА

Данная глава содержит два параграфа. В § 28 освещены приближенные методы решения краевой задачи, созданные еще задолго до появления ЭВМ, но не утратившие своего значения и в настоящее время. Сюда относятся методы коллокации, наименьших квадратов, подобластей, а также эффективный и достаточно универсальный метод Галеркина.

В § 29 детально излагается разностный метод решения краевой задачи для линейного обыкновенного дифференциального уравнения второго порядка. Разностный метод благодаря своей гибкости и большой универсальности, несмотря на значительную трудоемкость, получил широкое признание. Этот метод бурно развивается, чему, безусловно, способствует наличие высокопроизводительных ЭВМ. В § 29 вводятся основные понятия теории разностных схем: аппроксимация, устойчивость, сходимость и доказывается основная теорема о том, что из аппроксимации и устойчивости вытекает сходимость. Эти понятия и теорема иллюстрируются на разностной схеме (двухточечной краевой задаче) для обыкновенного дифференциального уравнения.

§ 28. Методы минимизации невязки и метод Галеркина

Краевая задача состоит в следующем. Требуется найти решение дифференциального уравнения

$$Lu \equiv u'' + p(x)u' + q(x)u = f(x), \quad a \leqslant x \leqslant b, \quad (1)$$

удовлетворяющее двум *краевым условиям*:

$$\begin{aligned} l_0 u &\equiv \alpha_0 u(a) + \beta_0 u'(a) = \gamma_0, \\ l_1 u &\equiv \alpha_1 u(b) + \beta_1 u'(b) = \gamma_1, \end{aligned} \quad (2)$$

где p, q, $f \in C[a, b]$ — заданные функции, α_j, β_j, γ_j — заданные числа, причем $\alpha_j^2 + \beta_j^2 > 0$, $j = 0, 1$.

Краевые условия (2) в общем случае задают линейную связь между значением искомого решения и его производной на концах отрезка $[a, b]$ в отдельности. В частном случае, если $\alpha_j = 1$, $\beta_j = 0$, то на соответствующем конце отрезка задано значение искомого решения. Такое краевое условие называется *условием первого рода*. Если $\alpha_j = 0$, $\beta_j = 1$, то на конце отрезка задано значение производной решения. Это краевое условие является *условием второго рода*. В общем случае, когда $\alpha_j \neq 0$, $\beta_j \neq 0$, краевое условие называется *условием третьего рода*.

В отличие от имеющей всегда единственное решение задачи Коши для уравнения (1), в которой на одном конце отрезка задаются значения решения и его производной, краевая задача (1), (2) может иметь или одно решение, или бесконечное множество решений, или, наконец, может совсем не иметь решений.

Для того чтобы существовало единственное решение неоднородной краевой задачи (1), (2), *необходимо и достаточно, чтобы однородная краевая задача*

$$Lu \equiv u'' + p(x)u' + q(x)u = 0, \quad a \leqslant x \leqslant b, \tag{3}$$

$$\begin{aligned} l_0 u &\equiv \alpha_0 u(a) + \beta_0 u'(a) = 0, \\ l_1 u &\equiv \alpha_1 u(b) + \beta_1 u'(b) = 0 \end{aligned} \tag{4}$$

имела только тривиальное решение $u(x) \equiv 0$.

Это условие не всегда удается проверить. В § 29 будут сформулированы простые достаточные условия однозначной разрешимости краевой задачи, а в данном параграфе мы будем предполагать существование единственного решения заданной краевой задачи, что часто вытекает из физических соображений.

Для нахождения приближенного решения краевой задачи (1), (2) поступаем следующим образом. Задаемся на отрезке $[a, b]$ некоторой линейно независимой системой дважды непрерывно дифференцируемых функций $\varphi_0, \varphi_1, \ldots, \varphi_n, \ldots$ таких, что φ_0 удовлетворяет краевым условиям (2), т. е. $l_0\varphi_0 = \gamma_0$, $l_1\varphi_0 = \gamma_1$, а остальные функции удовлетворяют однородным краевым условиям (4), т. е. $l_0\varphi_i = 0$, $l_1\varphi_i = 0$. $i = 1, 2, \ldots$ Заданная система функций $\varphi_0, \varphi_1, \ldots, \varphi_n, \ldots$ называется *базисной*.

Составляем линейную комбинацию $n+1$ базисных функций

$$y_n(x)=\varphi_0(x)+a_1\varphi_1(x)+\ \dots\ +a_n\varphi_n(x) \qquad (5)$$

с неизвестными пока коэффициентами $a_1, \dots, a_n$. В силу линейности операторов l_0, l_1 функция (5) при любых $a_1, \dots, a_n$ удовлетворяет заданным краевым условиям (2). Действительно,

$$l_j y_n = l_j\left(\varphi_0+\sum_{i=1}^{n} a_i\varphi_i\right)=l_j\varphi_0+\sum_{i=1}^{n} a_i l_j\varphi_i=$$

$$=\gamma_j+0=\gamma_j, \quad j=0,\ 1. \qquad (6)$$

Функция

$$\psi(x;\ a_1,\ \dots,\ a_n)=Ly_n(x)-f(x)=$$

$$=L\varphi_0(x)-f(x)+\sum_{k=1}^{n} a_k L\varphi_k(x) \qquad (7)$$

называется *невязкой*. Невязка равна разности левой и правой частей уравнения (1), когда в левую часть вместо u подставлена функция (5).

Невязка, как видно из (7), линейно зависит от параметров $a_1, \dots, a_n$ и является некоторой характеристикой уклонения функции (5) от неизвестного решения $u(x)$ краевой задачи (1), (2). Во всяком случае, если при некоторых значениях параметров $a_1, \dots, a_n$ невязка тождественно равна нулю по x на $[a, b]$, то функция (5) совпадает с решением краевой задачи, коль скоро удовлетворяются и уравнение (1), и краевые условия (2).

Однако обычно не удается сделать невязку тождественно равной нулю. Поэтому стараются подобрать параметры $a_1, \dots, a_n$ так, чтобы невязка в каком-то смысле стала по возможности меньше. Функцию (5) при выбранных $a_1, \dots, a_n$ принимают за приближенное решение краевой задачи (1), (2). Имеется ряд приближенных методов решения краевой задачи (1), (2), отличающихся способами нахождения параметров $a_1, \dots, a_n$.

Метод коллокации. В интервале (a, b) фиксируется n точек $x_1\ x_2, \dots, x_n$, называемых *точками коллокации*, в которых невязка (7) приравнивается

нулю:

$$\begin{aligned}
&\psi(x_1;\ a_1,\ \ldots,\ a_n)=0,\\
&\psi(x_2;\ a_1,\ \ldots,\ a_n)=0,\\
&\cdots\cdots\cdots\\
&\psi(x_n;\ a_1,\ \ldots,\ a_n)=0.
\end{aligned}$$

Полученная система линейных алгебраических уравнений относительно $a_1, \ldots, a_n$ в более подробной записи имеет вид

$$\begin{aligned}
&a_1L\varphi_1(x_1)+\ \ldots\ +a_nL\varphi_n(x_1)=f(x_1)-L\varphi_0(x_1),\\
&a_1L\varphi_1(x_2)+\ \ldots\ +a_nL\varphi_n(x_2)=f(x_2)-L\varphi_0(x_2),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&a_1L\varphi_1(x_n)+\ \ldots\ +a_nL\varphi_n(x_n)=f(x_n)-L\varphi_0(x_n).
\end{aligned}\tag{8}$$

Если система однозначно разрешима, то найденные из нее коэффициенты $a_1, \ldots, a_n$ подставляются в (5).

Пример. Дана краевая задача

$$\begin{gathered}
Lu\equiv u''+(1+x^2)\,u=-1,\quad -1\leqslant x\leqslant 1,\\
u(-1)=0,\quad u(1)=0.
\end{gathered}\tag{9}$$

Решение. Задаемся базисными функциями

$$\varphi_0(x)=0,\ \varphi_i(x)=x^{2i-2}(1-x^2),\ i=1,\ 2,\ \ldots$$

Очевидно, $\varphi_i(-1)=\varphi_i(1)=0$. т. е. функции φ_i удовлетворяют заданным однородным краевым условиям первого рода. Выбираем точки коллокации $x_0=0$, $x_1=-1/2$, $x_2=1/2$. Хотя точек коллокации три, приближенное решение будем искать в виде

$$y_2(x)=a_1(1-x^2)+a_2x^2(1-x^2),\tag{10}$$

так как и задача (9), и базисные функции, и точки коллокации симметричны относительно точки $x=0$.

Невязка в соответствии с (7), (9) имеет следующий вид:

$$\begin{aligned}
\psi(x;\ a_1,\ a_2)&=y_2''+(1+x^2)\,y_2+1=\\
&=1-a_1(1+x^4)+a_2(2-11x^2-x^6).
\end{aligned}$$

Требования $\psi(0;\ a_1,\ a_2)=0$, $\psi(\pm 1/2;\ a_1,\ a_2)=0$ приводят к системе уравнений

$$\begin{aligned}
&a_1-2a_2=1,\\
&\frac{17}{16}a_1+\frac{49}{64}a_2=1.
\end{aligned}$$

Отсюда находим $a_1=0{,}957$, $a_2=-0{,}022$. Приближенное решение задачи (9) имеет вид

$$y_2(x)=0{,}957\,(1-x^2)-0{,}022\,(x^2-x^4).$$

Интегральный метод наименьших квадратов. На невязку накладывается требование, чтобы интеграл

$$I=\int_a^b \psi^2(x;\ a_1,\ \ldots,\ a_n)\,dx \qquad (11)$$

принимал минимальное значение.

Для минимума интеграла необходимо выполнение следующих условий:

$$\frac{1}{2}\frac{\partial I}{\partial a_i}=\int_a^b \psi\frac{\partial \psi}{\partial a_i}\,dx=0,\quad i=1,\ 2,\ \ldots,\ n.$$

Эти условия с учетом выражения (7) для невязки ψ приводят к следующей системе линейных уравнений относительно $a_1, \ldots, a_n$:

$$\begin{aligned}
&a_1(L\varphi_1,\ L\varphi_1)+\ \ldots\ +a_n(L\varphi_n,\ L\varphi_1)=(f-L\varphi_0,\ L\varphi_1),\\
&a_1(L\varphi_1,\ L\varphi_2)+\ \ldots\ +a_n(L\varphi_n,\ L\varphi_2)=(f-L\varphi_0,\ L\varphi_2), \qquad (12)\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_1(L\varphi_1,\ L\varphi_n)+\ \ldots\ +a_n(L\varphi_n,\ L\varphi_n)=(f-L\varphi_0,\ L\varphi_n),
\end{aligned}$$

где $(f,\ g)=\int_a^b f(x)\,g(x)\,dx$ — скалярное произведение.

Если система функций $L\varphi_1, \ldots, L\varphi_n$ линейно независима на отрезке $[a, b]$, то на основании леммы 13.1 система (12) имеет единственное решение.

Приближенное решение краевой задачи (9), найденное в виде (10) интегральным методом наименьших квадратов, следующее:

$$y_2(x)=0{,}985(1-x^2)-0{,}078(x^2-x^4).$$

Дискретный метод наименьших квадратов. Вместо минимума интеграла (11) ищется минимум конечной суммы

$$\sum_{i=1}^{N}\psi^2(x_i;\ a_1,\ \ldots,\ a_n),$$

где $x_i\in(a, b)$ — некоторые точки, $N\geqslant n$.

Получаемая система уравнений для коэффициентов приближенного решения (5) имеет тот же вид (12),

с той лишь разницей, что используется скалярное произведение

$$(f, g) = \sum_{i=1}^{N} f(x_i) g(x_i).$$

Если $N = n$, то данный метод приводит к методу коллокации.

Приближенное решение краевой задачи (9), полученное в виде (10) дискретным методом наименьших квадратов при $N = 7$, $x_i = -1 + i/4$, имеет вид

$$y_2(x) = 0{,}932(1 - x^2) - 0{,}047(x^2 - x^4).$$

Метод подобластей. Пусть $a = x_0 < x_1 < \ldots$ $\ldots < x_{n-1} < x_n = b$. Коэффициенты приближенного решения (5) находятся из системы уравнений

$$\int_{x_{i-1}}^{x_i} \psi(x;\ a_1, \ldots, a_n)\, dx = 0, \quad i = 1, 2, \ldots, n. \tag{13}$$

Метод Галеркина. В основе метода, предложенного Б. Г. Галеркиным, лежит требование ортогональности базисных функций $\varphi_1, \varphi_2, \ldots, \varphi_n$ к невязке (7), т. е.

$$\int_a^b \psi(x;\ a_1, \ldots, a_n)\, \varphi_i(x)\, dx = 0, \quad i = 1, 2, \ldots, n.$$

Это требование приводит к следующей системе линейных алгебраических уравнений для коэффициентов приближенного решения (5) краевой задачи (1), (2):

$$\begin{aligned}
&a_1(L\varphi_1, \varphi_1) + \ldots + a_n(L\varphi_n, \varphi_1) = (f - L\varphi_0, \varphi_1),\\
&a_1(L\varphi_1, \varphi_2) + \ldots + a_n(L\varphi_n, \varphi_2) = (f - L\varphi_0, \varphi_2),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_1(L\varphi_1, \varphi_n) + \ldots + a_n(L\varphi_n, \varphi_n) = (f - L\varphi_0, \varphi_n),
\end{aligned} \tag{14}$$

где $(f, g) = \int_a^b f(x) g(x)\, dx$.

Пример. Рассмотрим краевую задачу

$$Lu \equiv u'' + u = -x, \quad 0 \leqslant x \leqslant 1, \quad u(0) = 0, \quad u(1) = 0, \tag{15}$$

точным решением которой является $u(x) = \sin x/\sin 1 - x$.

Р е ш е н и е. Выбираем базисные функции

$$\varphi_0(x) = 0, \quad \varphi_i(x) = x^i(1-x), \quad i = 1, 2, \ldots, \tag{16}$$

удовлетворяющие заданным однородным краевым условиям.

Пусть $n = 1$. Тогда система (14) обращается в одно уравнение

$$a_1(\varphi_1'' + \varphi_1, \varphi_1) = (-x, \varphi_1),$$

т. е.

$$a_1 \int_0^1 (-2 + x(1-x))\, x(1-x)\, dx = -\int_0^1 x^2(1-x)\, dx.$$

После вычисления интегралов получим $-(3/10)a_1 = -1/12$ Следовательно, $a_1 = 5/18$ и приближенное решение имеет выражение

$$y_1(x) = \frac{5}{18} x(1-x).$$

Возьмем теперь $n = 2$. Искомые коэффициенты a_1, a_2 удовлетворяют системе уравнений

$$a_1(L\varphi_1, \varphi_1) + a_2(L\varphi_2, \varphi_1) = (f, \varphi_1),$$
$$a_1(L\varphi_1, \varphi_2) + a_2(L\varphi_2, \varphi_2) = (f, \varphi_2),$$

которая в рассматриваемом примере приобретает вид

$$a_1 \int_0^1 (-2 + x(1-x))\, x(1-x)\, dx +$$
$$+ a_2 \int_0^1 (2 - 6x + x^2(1-x))\, x(1-x)\, dx = -\int_0^1 x^2(1-x)\, dx,$$
$$a_1 \int_0^1 (-2 + x(1-x))\, x^2(1-x)\, dx +$$
$$+ a_2 \int_0^1 (2 - 6x + x^2(1-x))\, x^2(1-x)\, dx = -\int_0^1 x^3(1-x)\, dx.$$

Вычисляя интегралы, находим

$$\frac{3}{10}a_1 + \frac{3}{20}a_2 = \frac{1}{12}, \quad \frac{3}{20}a_1 + \frac{13}{105}a_2 = \frac{1}{20}.$$

Отсюда получаем $a_1 = 71/369$, $a_2 = 7/41$. Искомое приближенное решение имеет вид

$$y_2(x) = x(1-x)\left(\frac{71}{369} + \frac{7}{41}x\right).$$

Для сравнения приведем некоторые значения точного решения $u(x)$ краевой задачи (15), а также $y_1(x)$, $y_2(x)$:

x	u	y_1	y_2
0,25	0,044	0,052	0,044
0,50	0,070	0,069	0,069
0,75	0,060	0,052	0,060

Таким образом, у первого приближения $y_1(x)$ погрешность порядка 0,01, а у второго — порядка 0,001.

Можно доказать, что если краевая задача

$$u'' + p(x)u' + q(x)u = f(x), \quad 0 \leqslant x \leqslant 1,$$
$$u(0) = 0, \quad u(1) = 0,$$

где p, q, $f \in C[0, 1]$, имеет единственное решение $u(x)$ (такова, например, задача (15)), то при использовании базисных функций (16) приближенное решение $y_n(x)$ может быть найдено методом Галеркина для всех достаточно больших n. Это приближенное решение сходится равномерно на $[0, 1]$ к $u(x)$ при $n \to \infty$. Более того, имеет место равномерная сходимость $y_n'(x)$ к $u'(x)$ на $[0, 1]$.

§ 29. Разностный метод. Основные понятия теории разностных схем

Рассмотрим краевую задачу

$$Lu \equiv u'' + p(x)u' - q(x)u = f(x), \quad 0 \leqslant x \leqslant 1, \quad (1)$$
$$l_0 u \equiv u(0) = \gamma_0, \quad l_1 u \equiv u(1) = \gamma_1, \quad (2)$$

где p, q, $f \in C_2[0, 1]$ — заданные функции, $q(x) \geqslant 0$ на $[0, 1]$, γ_0, γ_1 — заданные числа. Эта задача является частным случаем краевой задачи (28.1), (28.2).

Теорема 1. *Краевая задача* (1), (2) *имеет единственное решение* $u(x) \in C_4[0, 1]$.

Полное доказательство теоремы 1 выходит за рамки данной книги. Отметим лишь, что указанная гладкость решения вытекает из заданной гладкости функций p, q, f. В курсе дифференциальных уравнений устанавливается, что любое решение уравнения (1) принадлежит классу $C_4[0, 1]$, а значит, этому классу принадлежит и решение краевой задачи (1), (2).

Приступим к изложению разностного метода.

Сетки и сеточные функции. Зададим на отрезке $[0, 1]$ конечное множество точек (узлов) $\omega_h =$

$= \{x_j\}_{j=0}^{N}$, где $x_j = jh$, $h = 1/N$, $N \geqslant 2$ — натуральное. Множество ω_h называется *сеткой*, h — *шагом сетки*. Через ω_h' обозначим подмножество множества ω_h, полученное из ω_h отбрасыванием *крайних* (*граничных*) *узлов* x_0, x_N. Множество узлов ω_h' тоже называется *сеткой*. Говорят, что сетка ω_h' состоит из *внутренних узлов* сетки ω_h. Пусть также $\omega_h^* = \{x_0, x_N\}$ — сетка, состоящая из двух граничных узлов. Таким образом, $\omega_h = \omega_h' \cup \omega_h^*$ (сетка ω_h является объединением сеток ω_h' и ω_h^*).

На рис. 18 изображены сетки: в кружки помещены узлы, принадлежащие сетке ω_h', а в квадратиках расположены граничные узлы, образующие сетку ω_h^*.

Функция y, областью определения которой является какая-либо сетка ω_h, или ω_h', или ω_h^*, называется *сеточной*. Значения сеточной функции в узлах будем

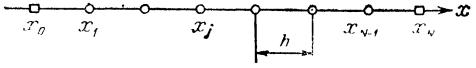


Рис. 18

для краткости обозначать следующим образом: $y_j = y(x_j)$. Функция f, определенная на всем отрезке $[0, 1]$, порождает сеточную функцию, принимающую в узле x_j значение, равное $f(x_j)$, $j = 0, 1, \ldots, N$. Для полученной сеточной функции сохраним символ f, а ее значение в узле x_j, согласно договоренности, будем обозначать через f_j, $j = 0, 1, \ldots, N$.

Обозначим через Y_h, Y_h', Y_h^* множества всех сеточных функций, определенных соответственно на ω_h, ω_h', ω_h^*. Множество Y_h (а также Y_h' и Y_h^*) является линейным пространством, в котором операции умножения функции на число и сложения функций выполняются по обычным правилам. В линейных пространствах Y_h, Y_h', Y_h^* сеточных функций зададим соответственно нормы

$$\|y\|_h = \max_{0 \leqslant j \leqslant N} |y_j|, \quad \|y\|_h' = \max_{1 \leqslant j \leqslant N-1} |y_j|, \qquad \|y\|_h^* = \max\{|y_0|, |y_N|\}. \tag{3}$$

Разностные операторы. Оператор $L^h y$ называется *разностным*, если он каждой сеточной функции $y \in Y_h$ ставит в соответствие некоторую сеточную функцию, принадлежащую Y'_h (или Y^*_h).

Примеры разностных операторов:

$$1)\ (L^h y)_j = y_{j+1}, \quad j = 1, 2, \ldots, N-1;$$
$$2)\ (l^h y)_j = y_j, \quad j = 0, N; \tag{4}$$
$$3)\ (L^h y)_j = \frac{y_{j-1} - 2y_j + y_{j+1}}{h^2}, \quad j = 1, 2, \ldots, N-1;$$
$$4)\ (l^h y)_j = \begin{cases} \dfrac{y_1 - y_0}{h}, & j = 0, \\ \dfrac{y_N - y_{N-1}}{h}, & j = N. \end{cases}$$

Индекс j в левой части равенств указывает номер узла, в котором оператор принимает данное справа значение. Какова бы ни была сеточная функция $y \in Y_h$, в примерах 1), 3) имеем $L^h y \in Y'_h$, а в примерах 2), 4) имеем $l^h y \in Y^*_h$.

Для удобства мы дали определение разностного оператора в широком смысле, не требуя явной его зависимости от разностей первого или более высоких порядков, а только от значений сеточной функции, имея также в виду, что сами значения сеточной функции можно формально считать разностями нулевого порядка.

Рассмотрим дифференциальный оператор

$$Lu = u'' + p(x)u' - q(x)u, \tag{5}$$

задающий левую часть уравнения (1), и введем единый граничный оператор

$$lu = \begin{cases} l_0 u, & x = 0, \\ l_1 u, & x = 1, \end{cases} \tag{6}$$

где в соответствии с заданными краевыми условиями (2) пока будем считать, что $l_0 u = u(0)$, $l_1 u = u(1)$, а в дальнейшем рассмотрим более общий граничный оператор.

С помощью операторов (5), (6) краевую задачу (1), (2) можно записать компактно в следующем

виде:

$$Lu = f, \tag{7}$$

$$lu = \gamma, \tag{8}$$

где

$$\gamma = \begin{cases} \gamma_0, & x = 0, \\ \gamma_1, & x = 1. \end{cases}$$

Нам предстоит аппроксимировать заданную на отрезке $[0, 1]$ дифференциальную краевую задачу (1), (2) (или, что одно и то же, задачу (7), (8)) некоторой разностной краевой задачей, в которой искомой является сеточная функция на сетке ω_h. Обычно при этом действуют по следующему плану. Сначала строят некоторые разностные операторы, аппроксимирующие дифференциальный оператор Lu и граничный оператор lu, и составляют на сетке ω_h разностную краевую задачу, которую принято называть *разностной схемой* для дифференциальной задачи (1), (2). Затем убеждаются, что полученная разностная схема аппроксимирует исходную краевую задачу (1), (2).

После этого исследуют устойчивость разностной схемы и сходимость ее решения к решению дифференциальной задачи при стремлении шага сетки к нулю. К разностной схеме также предъявляется важное требование, чтобы для нахождения ее решения мог быть применен достаточно эффективный на практике численный метод. Понятия аппроксимации, устойчивости и сходимости уточняются ниже.

Аппроксимируем дифференциальный оператор Lu на сетке ω_h' разностным оператором $L^h u$, получаемым из (5) путем замены в узлах сетки ω_h' производных u', u'' соответствующими разностными производными по формулам (10.6), (10.7). Оператор $L^h u$, построенный по указанному правилу, имеет следующий вид:

$$(L^h u)_j = \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} + p_j \frac{u_{j+1} - u_{j-1}}{2h} - q_j u_j, \tag{9}$$

$$j = 1, 2, \ldots, N - 1.$$

Пусть $u \in C_4[0, 1]$ — произвольная функция, в частности решение краевой задачи (1), (2). Тогда на

основании (10.2), (10.3), (3) выполняется неравенство

$$\| Lu - L^h u \|'_h \leqslant c_u h^2, \tag{10}$$

где c_u — некоторая не зависящая от h постоянная.

Действительно, в силу (10.2), (10.3) имеем

$$\left| p(x_j) u'(x_j) - p_j \frac{u_{j+1} - u_{j-1}}{2h} \right| \leqslant \frac{h^2}{6} \| p \|_C \| u''' \|_C,$$

$$\left| u''(x_j) - \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} \right| \leqslant \frac{h^2}{12} \| u^{(4)} \|_C,$$

где $p_j = p(x_j)$, $j = 1, 2, \ldots, N-1$, $\| p \|_C = \max\limits_{[0,\,1]} | p(x) |$.

Отсюда и из (5), (9), согласно определению (3) нормы $\| \cdot \|'_h$, легко следует (10).

Определение. Говорят, что разностный оператор $L^h u$ *аппроксимирует* дифференциальный оператор Lu с k-м порядком относительно шага h в сеточной норме $\| \cdot \|'_h$, если $k > 0$ и для любой достаточно гладкой функции u выполняется неравенство

$$\| Lu - L^h u \|'_h \leqslant c_u h^k, \tag{11}$$

где c_u — некоторая не зависящая от h постоянная.

Рассмотренный выше разностный оператор $L^h u$ (9) аппроксимирует согласно (10) дифференциальный оператор (5) со вторым порядком относительно h.

Граничный оператор (6), с помощью которого задаются краевые условия первого рода (2), т. е. условия (8), устроен очень просто, а именно, на концах отрезка [0, 1] он принимает значения, совпадающие со значениями функции y. Поэтому его естественно аппроксимировать разностным оператором (4), который тоже в узлах x_0, x_N, расположенных на концах отрезка [0, 1], принимает значения, совпадающие со значениями сеточной функции в этих узлах.

Разностные схемы. Теперь можно составить следующую разностную краевую задачу (точнее говоря, семейство разностных краевых задач, зависящее от параметра h):

$$L^h y = f, \tag{12}$$

$$l^h y = g, \tag{13}$$

где $L^h y$ — разностный оператор, определенный по формуле (9), $l^h y$ — разностный оператор (4), $f \in Y'_h$ —

сеточная функция, порождаемая правой частью уравнения (7), т. е. уравнения (1), $g \in Y_h^*$ — заданная сеточная функция со значениями $g_0 = \gamma_0$, $g_N = \gamma_1$.

Разностная краевая задача (12), (13) называется *разностной схемой* для дифференциальной краевой задачи (7), (8). С учетом выражений (9), (4) операторов $L^h y$, $l^h y$ эта разностная схема может быть более подробно записана в следующем виде:

$$\frac{y_{j-1} - 2y_j + y_{j+1}}{h^2} + p_j \frac{y_{j+1} - y_{j-1}}{2h} - q_j y_j = f_j, \quad (14)$$

$$j = 1, 2, \ldots, N-1,$$

$$y_0 = \gamma_0, \quad y_N = \gamma_1. \quad (15)$$

Легко видеть, что разностная краевая задача (14), (15) является краевой задачей вида (22.1), (22.2), в которой

$$A_j = \frac{1}{h^2} - \frac{p_j}{2h}, \quad B_j = \frac{1}{h^2} + \frac{p_j}{2h}, \quad C_j = \frac{2}{h^2} + q_j,$$

$$F_j = f_j, \qquad \varkappa_0 = \varkappa_N = 0, \qquad \nu_0 = \gamma_0, \ \nu_N = \gamma_1.$$

Очевидно, если

$$h < h_0 = \min\left\{1, \frac{2}{\|p\|_C}\right\}, \quad (16)$$

то выполнены и условия (22.3). Таким образом, если шаг h сетки ω_h удовлетворяет неравенству (16), то согласно лемме 22.1 разностная схема (12), (13) имеет единственное решение $y \in Y_h$, для нахождения которого может быть применен весьма эффективный на практике метод прогонки, изложенный в § 22.

Аппроксимация, устойчивость, сходимость. Решение u дифференциальной задачи (7), (8), вообще говоря, не удовлетворяет на сетке ω_h разностным уравнениям (12), (13). Однако всегда можно написать равенства

$$L^h u = f + \psi \quad \text{на} \quad \omega_h', \quad (17)$$

$$l^h u = g + \varphi \quad \text{на} \quad \omega_h^*, \quad (18)$$

где $\psi = L^h u - f$, $\varphi = l^h u - g$ — сеточные функции, называемые *невязками* решения дифференциальной задачи для разностной задачи ($\psi \in Y_h'$ — невязка для

разностного уравнения (12), $\varphi \in Y_h^*$ — невязка для краевых условий (13)).

Определение. Говорят, что разностная схема (12), (13) *аппроксимирует* дифференциальную задачу (7), (8) на ее решении u с k-м порядком относительно h, если $k > 0$ и нормы невязок решения дифференциальной задачи имеют следующие величины:

$$\| \psi \|_h' = O(h^k), \quad \| \varphi \|_1 = O(h^k). \tag{19}$$

Данное определение носит общий характер, т. е. под дифференциальной краевой задачей (7), (8) можно подразумевать не только конкретную краевую задачу (1), (2), а в качестве разностной схемы (12), (13) может фигурировать не только разностная схема (14), (15). Однако всюду будем предполагать, что дифференциальные операторы Lu, lu в задаче (7), (8) и разностные операторы $L^h y$, $l^h y$ в задаче (12), (13) являются линейными. В конце параграфа рассматривается краевая задача с более общими краевыми условиями, чем выше, к которой данное определение подходит.

Убедимся, что разностная схема (14), (15) аппроксимирует краевую задачу (1), (2) на ее решении u со вторым порядком относительно h. Действительно, так как $Lu = f$, то

$$\psi = L^h u - f = L^h u - Lu \quad \text{на} \quad \omega_h'$$

и, следовательно, согласно неравенству (10), где $L^h u$ — оператор (9), имеем

$$\| \psi \|_h' = \| L^h u - Lu \|_h' = \| Lu - L^h u \|_h' = O(h^2).$$

Кроме того, в соответствии с (18), (4), (13), (2) получаем

$$\| \varphi \|_h^* = \max \{ | u_0 - g_0 |, \ | u_N - g_N | \} =$$
$$= \max \{ | u(0) - \gamma_0 |, \ | u(1) - \gamma_1 | \} = 0 \quad \forall h,$$

а постоянная, равная нулю, является величиной $O(h^2)$.

Замечание 1. В случае, если невязка φ равна нулю на ω_h^* при любом h, то говорят, что разностные краевые условия (13) *аппроксимируют точно* заданные краевые условия (8) в исходной задаче.

Следующее важное определение тоже носит общий характер.

Определение. Скажем, что разностная схема (12), (13) *устойчива*, если существует такое $h_0 > 0$, что при любом $h = 1/N < h_0$ и произвольных сеточных функциях $\xi \in Y'_h$, $\eta \in Y^*_h$ разностная задача

$$L^h z = \xi, \tag{20}$$

$$l^h z = \eta \tag{21}$$

имеет единственное решение $z \in Y_h$, причем выполняется неравенство

$$\| z \|_h \leqslant c' \| \xi \|'_h + c^* \| \eta \|^*_h, \tag{22}$$

где c', c^* — некоторые постоянные, не зависящие ни от h, ни от функций ξ, η.

Свойство устойчивости (или же неустойчивости) разностной схемы является внутренним свойством разностной схемы как таковой, т. е. оно никак не связывается с той исходной дифференциальной задачей, для которой построена разностная схема. Очень важно, чтобы разностная схема была устойчива.

Такая разностная схема, в частности, устойчива к погрешностям округлений, которые часто можно трактовать как аддитивные добавки в правые части уравнений (12), (13) в виде некоторых сеточных функций $\xi \in Y'_h$, $\eta \in Y^*_h$. При этом погрешность решения разностной схемы (12), (13) в силу линейности схемы совпадает с решением разностной задачи (20), (21), которое согласно неравенству (22) будет мало, если малы погрешности ξ, η или, точнее говоря, их нормы $\| \xi \|'_h$, $\| \eta \|^*_h$.

Примечательно также, что постоянные c', c^* в неравенстве (22) не зависят от h, т. е., выражаясь языком физики, можно сказать, что чувствительность устойчивой разностной схемы (12), (13) к возмущениям правых частей при измельчении сетки не увеличивается. Примеры неустойчивых разностных схем даны в § 30. Следующая теорема свидетельствует об устойчивости построенной выше разностной схемы (14), (15).

Теорема 2. *Разностная схема* (12), (13), *где* $L^h y$ *— разностный оператор* (9), $l^h y$ *— разностный оператор* (4), *устойчива.*

Доказательство. Выше было установлено, что указанная в теореме разностная схема (12), (13), в более подробной записи имеющая вид (14), (15), при условии (16) однозначно разрешима. Следовательно, разностная краевая задача (20), (21), отвечающая рассматриваемой разностной схеме (12), (13), тоже при условии (16) однозначно разрешима. Действительно, разностная схема (12), (13) и соответствующая краевая задача (20), (21) являются системами линейных алгебраических уравнений, отличающимися только свободными членами. Поэтому они однозначно разрешимы одновременно.

Остается доказать существование постоянных c', c^*, при которых решение задачи (20), (21) удовлетворяет неравенству (22). В силу линейности задачи (20), (21) ее решение z может быть представлено в виде

$$z = \lambda + \mu, \tag{23}$$

где λ, μ — решения разностных задач

$$L^h\lambda = \xi, \quad l^h\lambda = 0, \tag{24}$$

$$L^h\mu = 0, \quad l^h\mu = \eta. \tag{25}$$

Для простоты докажем неравенство (22) только в случае, когда функция p всюду равна нулю. Тогда разностная задача (24) с учетом (14), (15) может быть записана в следующем виде:

$$\lambda_{j-1} - (2 + h^2q_j)\lambda_j + \lambda_{j+1} = h^2\xi_j, \quad j = 1, 2, \ldots, N-1,$$
$$\lambda_0 = 0, \quad \lambda_N = 0.$$

Поскольку $q_j \geqslant 0$, $j = 1, 2, \ldots, N-1$, $Nh = 1$, то на основании леммы 22.2 для решения задачи (24) справедливы оценки

$$\|\lambda\|_h = \max_{0 \leqslant j \leqslant N} |\lambda_j| \leqslant N^2 \max_{1 \leqslant j \leqslant N-1} |h^2\xi_j| =$$
$$= N^2h^2 \max_{1 \leqslant j \leqslant N-1} |\xi_j| = \max_{1 \leqslant j \leqslant N-1} |\xi_j| = \|\xi\|'_h. \tag{26}$$

Аналогично, задача (25) записывается в виде

$$\mu_{j-1} - (2 + h^2q_j)\mu_j + \mu_{j+1} = 0, \quad j = 1, 2, \ldots, N-1,$$
$$\mu_0 = \eta_0, \quad \mu_N = \eta_N.$$

Согласно лемме 22.3 норма решения задачи (25) выражается через $\|\eta\|_h^*$:

$$\|\mu\|_h = \max_{0 \leqslant j \leqslant N} |\mu_j| = \max\{|\eta_0|, |\eta_N|\} = \|\eta\|_h^*. \quad (27)$$

Поскольку согласно (23) $\|z\|_h \leqslant \|\lambda\|_h + \|\mu\|_h$, то на основании (26), (27) решение задачи (20), (21) при $p(x) \equiv 0$ удовлетворяет неравенству (22) с постоянными $c' = c^* = 1$. Можно доказать, что это неравенство справедливо при h, удовлетворяющем неравенству (16), и в случае произвольной функции $p \in C[0, 1]$, но с некоторыми другими постоянными c', c^*.

Дадим еще одно важное определение.

Определение. Говорят, что решение y разностной схемы (12), (13) *сходится* при измельчении сетки к решению u дифференциальной краевой задачи (7), (8) с k-м порядком относительно h, если $k > 0$ и

$$\|u - y\|_h = O(h^k). \quad (28)$$

В этом случае говорят также, что разностная схема имеет *k-й порядок точности*.

Замечания. 2. Данное определение тоже носит общий характер, т. е., как и выше, под разностной схемой (12), (13) подразумевается не только разностная схема (14), (15), а в качестве дифференциальной задачи (7), (8) может фигурировать не только задача (1), (2). Кроме того, сеточная норма $\|\cdot\|_h$, используемая в формуле (28), может быть определена и другим способом, нежели в соответствии с (3). Однако следует иметь в виду, что наличие сходимости в смысле (28) при измельчении сетки зависит от свойств решения u исходной задачи (7), (8), от разностной схемы (12), (13), а также и от выбора норм в Y_h, Y_h', Y_h^*.

3. Говоря о сходимости решения разностной схемы, слова «при измельчении сетки» для краткости будем опускать.

Теперь можно сформулировать *основную теорему теории разностных схем.*

Теорема 3. *Пусть разностная схема* (12), (13) *аппроксимирует дифференциальную краевую задачу* (7), (8) *на ее решении и с k-м порядком относительно h и пусть она устойчива. Тогда решение разностной*

схемы сходится к решению дифференциальной задачи с тем же порядком относительно h.

Доказательство. Поскольку по условию теоремы разностная схема аппроксимирует дифференциальную задачу с k-м порядком, то согласно (19) существуют такие числа $h_1 > 0$, $c_1 > 0$, что при $h < h_1$ выполняются неравенства

$$\| \psi \|'_h \leqslant c_1 h^k, \quad \| \varphi \|^*_h \leqslant c_1 h^k, \tag{29}$$

где ψ, φ — невязки решения дифференциальной задачи (см. (17), (18)). Вычитая из (17), (18) соответственно (12), (13), в силу линейности операторов L^h, l^h получаем

$$L^h (u - y) = \psi, \quad l^h (u - y) = \varphi. \tag{30}$$

Поскольку по предположению разностная схема устойчива, то существуют такие числа h_0, $0 < h_0 \leqslant h_1$, c', c^*, что при $h < h_0$ разностная задача (30) однозначно разрешима. При этом на основании (22), (29) выполняются неравенства

$$\| u - y \|_h \leqslant c' \| \psi \|'_h + c^* \| \varphi \|^*_h \leqslant c' c_1 h^k + c^* c_1 h^k = c h^k,$$

где u — решение дифференциальной краевой задачи (7), (8), y — решение разностной схемы (12), (13), $c = c_1 (c' + c^*)$ не зависит от h, $h < h_0$. Основная теорема доказана.

Замечания. 4. Определения аппроксимации, устойчивости и сходимости решения разностной схемы и доказанная основная теорема, кратко формулируемая так: «аппроксимация плюс устойчивость есть сходимость», носят общий характер и в дальнейшем будут служить рабочим инструментом при рассмотрении разностных схем для других задач. Намеченный путь исследования трудного вопроса сходимости разделением на две самостоятельные части (проверка аппроксимации и исследование устойчивости) является общепринятым. На практике обычно строят различные варианты разностных схем, обладающих свойством аппроксимации, и стремятся выбрать устойчивую разностную схему.

5. Что касается конкретной разностной схемы (14), (15), то, поскольку она аппроксимирует краевую задачу (1), (2) со вторым порядком относительно h (это

было установлено выше) и согласно теореме 2 устойчива, имеет место сходимость решения y разностной краевой задачи (14), (15) при измельчении сетки к решению u задачи (1), (2) со вторым порядком относительно h, т. е.

$$\|u - y\|_h = O(h^2). \tag{31}$$

Другими словами, разностная схема (14), (15) имеет второй порядок точности. Заметим, что согласно определению в (3) нормы $\|\cdot\|_h$ левая часть равенства (31) является максимумом модуля уклонения на сетке ω_h решения y разностной схемы (14), (15) от решения u дифференциальной краевой задачи (1), (2).

Аппроксимация краевых условий второго и третьего рода. Зададим более общие, чем (2), краевые условия

$$\begin{aligned} l_0 u &\equiv u(0) - \beta_0 u'(0) = \gamma_0, \\ l_1 u &\equiv \alpha_1 u(1) + \beta_1 u'(1) = \gamma_1, \end{aligned} \tag{2*}$$

где α_1, β_i, γ_i — заданные числа, причем $\alpha_1 \geqslant 0$, $\beta_i \geqslant 0$, $i = 0, 1$, $\alpha_1 + \beta_1 > 0$. Краевое условие на левом конце отрезка $[0, 1]$ может быть условием первого рода ($\beta_0 = 0$) и условием третьего рода ($\beta_0 > 0$), а краевое условие в точке $x = 1$ может быть также и условием второго рода ($\alpha_1 = 0$, $\beta_1 = 1$).

Краевая задача (1), (2*) при тех же требованиях к функциям p, q, f, что и в задаче (1), (2), тоже имеет единственное решение $u(x) \in C_4[0, 1]$. Краевая задача (1), (2) является частным случаем задачи (1), (2*) ($\alpha_1 = 1$, $\beta_0 = \beta_1 = 0$).

Для аппроксимации краевых условий (2*) поступаем следующим образом. Пусть $u(x)$ — решение краевой задачи (1), (2*). Поскольку

$$u_1 = u_0 + hu_0' + \frac{h^2}{2} u_0'' + O(h^3),$$

где $u_0^{(k)} = u^{(k)}(0)$, то

$$u_0' = \frac{u_1 - u_0}{h} - \frac{h}{2} u_0'' + O(h^2) \tag{32}$$

или, более грубо,

$$u_0' = \frac{u_1 - u_0}{h} + O(h). \tag{33}$$

Принимая во внимание, что u удовлетворяет уравнению (1), можем написать следующее равенство:

$$u_0'' = f_0 - p_0 u_0' + q_0 u_0.$$

Следовательно, согласно (33) имеем

$$u_0'' = f_0 - p_0 \frac{u_1 - u_0}{h} + q_0 u_0 + O(h).$$

Подставив найденное выражение u_0'' в (32), а затем полученное u_0' — в первое краевое условие (2*), получим следующее соотношение:

$$u_0 - \beta_0 \left[\frac{u_1 - u_0}{h} + \frac{h}{2} \left(p_0 \frac{u_1 - u_0}{h} - q_0 u_0 \right) \right] = \gamma_0 - h \frac{\beta_0 f_0}{2} + O(h^2),$$

или, что то же самое,

$$a_0 u_0 - b_0 u_1 = g_0 + O(h^2), \tag{34}$$

где

$$a_0 = 1 + b_0 + \beta_0 q_0 \frac{h}{2}, \quad b_0 = \beta_0 \left(\frac{1}{h} + \frac{p_0}{2} \right), \tag{35}$$

$$g_0 = \gamma_0 - h \frac{\beta_0 f_0}{2}. \tag{36}$$

Аналогично, используя уравнение (1) и второе краевое условие (2*), находим следующую связь между значениями u_N и u_{N-1} решения задачи (1), (2) в крайнем правом и предпоследнем узлах сетки ω_h:

$$a_1 u_N - b_1 u_{N-1} = g_N + O(h^2), \tag{37}$$

где

$$a_1 = \alpha_1 + b_1 + \beta_1 q_N \frac{h}{2}, \quad b_1 = \beta_1 \left(\frac{1}{h} - \frac{p_N}{2} \right), \tag{38}$$

$$g_N = \gamma_1 - h \frac{\beta_1 f_N}{2}. \tag{39}$$

Введем разностный оператор $l^h y$:

$$(l^h y)_j = \begin{cases} a_0 y_0 - b_0 y_1, & j = 0, \\ a_1 y_N - b_1 y_{N-1}, & j = N, \end{cases} \tag{40}$$

где коэффициенты a_0, b_0, a_1, b_1 определены формулами (35), (38).

Будем теперь подразумевать под краевой задачей (7), (8) краевую задачу (1), (2*), считая при этом, что дифференциальный оператор Lu по-прежнему определен формулой (5), а граничный оператор lu

имеет выражение (6), в котором операторы l_0u, l_1u являются дифференциальными операторами, задающими левые части краевых условий (2*).

Дифференциальной краевой задаче (7), (8) поставим в соответствие разностную схему (12), (13), где разностный оператор L^hy по-прежнему определен формулой (9), оператор l^hy имеет вид (40), а сеточная функция $g \in Y_h^*$ задана формулами (36), (39).

Построенная разностная схема (12), (13) аппроксимирует дифференциальную краевую задачу (7), (8) на ее решении u со вторым порядком относительно h. Действительно, рассмотрим для решения задачи (7), (8) равенства (17), (18) с невязками ψ, φ. Поскольку разностное уравнение (12) не изменилось, то, как было установлено ранее, $\|\psi\|_h' = O(h^2)$. Равенство (18) с учетом (40) фактически объединяет два равенства (34) и (37), из которых явно видно, что $\varphi_0 = O(h^2)$, $\varphi_N = O(h^2)$ и, следовательно,

$$\|\varphi\|_h^* = \max\{|\varphi_0|, |\varphi_N|\} = O(h^2).$$

Разностная схема (12), (13) для краевой задачи (1), (2*) в более подробной записи может быть представлена в виде разностного уравнения (14) с краевыми условиями

$$a_0y_0 - b_0y_1 = g_0, \quad a_1y_N - b_1y_{N-1} = g_N, \qquad (15^*)$$

где коэффициенты a_0, b_0, a_1, b_1 и свободные члены g_0, g_N определены формулами (35), (38), (36), (39).

Можно доказать, что рассматриваемая разностная схема (12), (13), т. е. разностная краевая задача (14), (15*), устойчива. Подробно на этом останавливаться не будем. Отметим лишь, что если h удовлетворяет неравенству (16), то разностная краевая задача (14), (15*) имеет единственное решение, для нахождения которого можно применить метод прогонки. Это гарантируется условиями (22.3), которые для коэффициентов разностного уравнения (14), как было проверено выше, выполнены. Для краевых условий (15*) требования (22.3) при указанном h тоже выполнены, ибо согласно (35), (38) $a_0 > b_0 \geqslant 0$, $a_1 \geqslant b_1 \geqslant 0$, $a_1 + b_1 > 0$ и, следовательно, краевые условия (15*) могут быть переписаны в виде (22.2), где $0 \leqslant \varkappa_0 = b_0/a_0 < 1$, $0 \leqslant \varkappa_1 = b_1/a_1 \leqslant 1$.

Таким образом, поскольку разностная схема (14), (15*) аппроксимирует краевую задачу (1), (2*) со вторым порядком относительно h и устойчива, то по основной теореме 3 ее решение y сходится к решению u краевой задачи (1), (2*) со скоростью $O(h^2)$, т. е. справедливо соотношение (31).

Применение правила Рунге. Допустим, что p, q, $f \in C_4[0, 1]$, т. е. функции p, q, f, входящие в уравнение (1), обладают избыточной гладкостью по сравнению с той, при которой установлена оценка (31) погрешности разностного решения краевой задачи (1), (2*). Тогда при h, удовлетворяющем неравенству (16), можно получить следующее соотношение на сетке ω_h:

$$u = y + v(x)h^2 + r, \tag{41}$$

где u — решение дифференциальной краевой задачи (1), (2*), y — решение разностной краевой задачи (14), (15*), $v \in C_2[0, 1]$ — некоторая не зависящая от h функция, r — сеточная функция,

$$\|r\|_h = O(h^{2+m}), \tag{42}$$

$m = 2$ при краевых условиях первого рода (2), $m = 1$ при $\beta_0 + \beta_1 > 0$ в краевых условиях (2*).

Наличие соотношения (41) позволяет так же, как и в случае задачи Коши (см. § 18), применить в общих узлах сеток ω_h и $\omega_{h/2}$ правило Рунге для получения приближенной оценки погрешности и осуществить уточнение разностного решения по Ричардсону. Кроме того, из соотношения (41) может быть получена следующая формула:

$$u'(x_j) = \frac{y_{j+1} - y_{j-1}}{2h} + O(h^2), \tag{43}$$

говорящая о том, что с помощью разностного решения y можно в узлах сетки ω_h' найти производную неизвестного решения краевой задачи (1), (2*) с точностью $O(h^2)$.

Действительно, согласно (41) имеем

$$\frac{u_{j+1} - u_{j-1}}{2h} = \frac{y_{j+1} - y_{j-1}}{2h} + \frac{v_{j+1} - v_{j-1}}{2h} h^2 + \frac{r_{j+1} - r_{j-1}}{2h}. \tag{44}$$

Поскольку $u \in C_3[0, 1]$, то на основании (10.2) получаем

$$\frac{u_{j+1} - u_{j-1}}{2h} = u'(x_j) + O(h^2). \tag{45}$$

Поскольку $v \in C_1[0, 1]$, то согласно формуле конечных приращений Лагранжа

$$\frac{v_{j+1} - v_{j-1}}{2h} = v'(\xi_j), \tag{46}$$

где $\xi_j \in (x_{j-1}, x_{j+1})$ — некоторая точка. Наконец, в силу (42)

$$\frac{r_{j+1} - r_{j-1}}{2h} = O(h^2). \tag{47}$$

Из (44)—(47) следует формула (43).

Замечание 6. Из уравнения (1) можно найти в узлах сетки ω_h с точностью $O(h^2)$ вторую производную решения краевой задачи (1), (2*), если в это уравнение подставить приближенные значения u, u', полученные с точностью $O(h^2)$.

Устойчивость по правой части. Наряду со сформулированным выше определением устойчивости «в целом» целесообразно дать еще одно определение.

Определение. Скажем, что разностная схема (12), (13) *устойчива по правой части*, если существуют такие числа $h_0 > 0$, $c' > 0$, что при любом $h = 1/N < h_0$ и любой сеточной функции $\xi \in Y'_h$ разностная краевая задача

$$L^h z = \xi, \quad l^h z = 0 \tag{20*}$$

имеет единственное решение $z \in Y_h$, причем

$$\|z\|_h \leqslant c' \|\xi\|'_h. \tag{22*}$$

Дело в том, что во многих случаях краевые условия (и другие дополнительные условия, о которых речь пойдет в гл. 6) аппроксимируются точно, т. е. $\varphi = l^h u - g = 0$ (невязка φ равна нулю как элемент Y^*_h). Тогда погрешность $u - y$ на сетке ω_h является решением краевой задачи

$$L^h(u - y) = \psi, \quad l^h(u - y) = 0$$

и из доказательства основной теоремы 3 легко усмотреть, что для сходимости (при наличии аппроксимации уравнения (7) разностным уравнением (12)) достаточно, чтобы разностная схема была устойчива только по правой части.

Например, если в краевых условиях (2*) $\beta_0 = \beta_1 = 0$, то эти краевые условия, будучи условиями первого рода, в разностной схеме (14), (15*) аппроксимируются точно.

ГЛАВА 6

РАЗНОСТНЫЕ СХЕМЫ ДЛЯ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ

Изучаются разностные схемы для линейных уравнений с частными производными первого и второго порядка. Главное внимание уделяется вопросам аппроксимации и устойчивости, а также алгоритмам нахождения решений разностных схем.

§ 30. Линейное уравнение с частными производными первого порядка

Постановка задачи Коши. Найти непрерывную в замкнутой полосе $-\infty < x < \infty$, $0 \leqslant t \leqslant T$ функцию $u(x, t)$, удовлетворяющую при $0 < t \leqslant T$ дифференциальному уравнению

$$Lu \equiv \frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0, \tag{1}$$

а при $t = 0$ начальному условию

$$lu \equiv u(x, 0) = g(x), \tag{2}$$

где $a > 0$, $T > 0$ — числа, g — заданная дважды непрерывно дифференцируемая на всей оси x функция, причем

$$\sup_{-\infty < x < \infty} |g''(x)| = G_2 < \infty. \tag{3}$$

Уравнение (1) называется *линейным дифференциальным уравнением с частными производными первого порядка.*

Точным решением задачи Коши (1), (2) является функция

$$u(x, t) = g(x - at). \tag{4}$$

Действительно, $u'_x(x, t) = g'(x - at)$, $u'_t(x, t) = -ag'(x - at)$. Следовательно, функция (4) удовлетворяет уравнению (1). Выполнение начального условия (2) очевидно.

Хотя решение задачи Коши (1), (2) известно, и поэтому необходимости в приближенных методах для нее нет, но эта задача является удобной моделью, на которой будем изучать разностные схемы для уравнений с частными производными.

Сетки и нормы. Построим в полосе, в которой поставлена задача Коши, сетки (рис. 19). Пусть

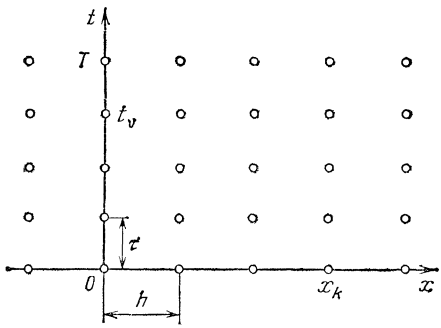


Рис. 19

$h > 0$ — шаг по x, $\tau = T/M$ (M — натуральное) — шаг по времени t, $x_k = kh$, $t_\nu = \nu\tau$, $u_k^\nu = u(x_k, t_\nu)$.

Сеткой называется каждое из следующих трех множеств точек (узлов):

$$\omega_h = \{(x_k, t_\nu): k = 0, \pm 1, \ldots; \nu = 0, 1, \ldots, M\},$$

$$\omega_h' = \{(x_k, t_\nu): k = 0, \pm 1, \ldots; \nu = 1, 2, \ldots, M\},$$

$$\omega_h^* = \{(x_k, 0): k = 0, \pm 1, \ldots\}.$$

Сетка ω_h^* состоит из узлов, расположенных на оси абсцисс, где задано начальное условие (2). Сетка ω_h является объединением сеток ω_h' и ω_h^*.

Для сеточных функций, заданных на сетках ω_h и ω_h', введем соответственно нормы

$$\|y\|_h = \sup_{\omega_h} |y_k^\nu|, \quad \|y\|_h' = \sup_{\omega_h'} |y_k^\nu|. \tag{5}$$

Разностная схема. Заменив в уравнении (1) частные производные u_t', u_x' разностными производными (разделенными разностями) по t и x и переобозначив u на y, приходим к разностной схеме

(к разностной задаче Коши)

$$(L^h y)_k^{\nu} \equiv \frac{y_k^{\nu} - y_k^{\nu-1}}{\tau} + a\frac{y_k^{\nu-1} - y_{k-1}^{\nu-1}}{h} = 0, \qquad (6)$$

$$(l^h y)_k \equiv y_k^0 = g_k. \qquad (7)$$

Здесь и ниже индекс k пробегает все целочисленные значения, $g_k = g(x_k)$, а $\nu = 1, 2, \ldots, M$.

Разностная схема (6), (7) кратко может быть записана в виде

$$L^h y = 0 \quad \text{на} \quad \omega_h', \quad l^h y = g \quad \text{на} \quad \omega_h^*. \qquad (8)$$

Хотя разностная схема зависит от двух параметров h и τ, для обозначения разностных операторов $L^h y$, $l^h y$, сеток и норм ради простоты будем использовать только один индекс h.

В разностное уравнение (6) при фиксированных k и ν входят только три неизвестные $y_{k-1}^{\nu-1}$, $y_k^{\nu-1}$, y_k^{ν} относящиеся соответственно к узлам $(x_{k-1}, t_{\nu-1})$, $(x_k, t_{\nu-1})$, (x_k, t_ν).

Геометрическое место точек (узлов), в которых используются значения функции в разностном уравнении при фиксированных k, ν, называется *шаблоном* разностного уравнения. Шаблон разностного уравнения (6), состоящий из трех узлов, изображен на рис. 20.

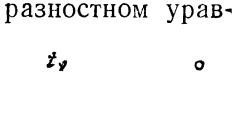


Рис. 20

Аппроксимация. Прежде всего заметим, что решение (4) задачи Коши (1), (2) дважды непрерывно дифференцируемо по переменным x, t на всей замкнутой полосе $-\infty < x < \infty$, $0 \leqslant t \leqslant T$. Поэтому согласно формуле Тейлора имеем

$$u(x_k, t_\nu) = u(x_k, t_{\nu-1}) + \tau u_t'(x_k, t_{\nu-1}) + \frac{\tau^2}{2} u_{tt}''(x_k, \theta_{k\nu}),$$

$$u(x_{k-1}, t_{\nu-1}) = u(x_k, t_{\nu-1}) - h u_x'(x_k, t_{\nu-1}) + \frac{h^2}{2} u_{xx}''(\eta_{k\nu}, t_{\nu-1}),$$

т. е.

$$\frac{u_k^{\nu} - u_k^{\nu-1}}{\tau} = u_t'(x_k, t_{\nu-1}) + \frac{\tau}{2} u_{tt}''(x_k, \theta_{k\nu}),$$

$$\frac{u_k^{\nu-1} - u_{k-1}^{\nu-1}}{h} = u_x'(x_k, t_{\nu-1}) - \frac{h}{2} u_{xx}''(\eta_{k\nu}, t_{\nu-1}),$$

где u — решение задачи Коши (1), (2), $t_{\nu-1} < \theta_{k\nu} < t_\nu$, $x_{k-1} < \eta_{k\nu} < x_k$. Отсюда

$$(L^h u)_k^\nu \equiv \frac{u_k^\nu - u_k^{\nu-1}}{\tau} + a\frac{u_k^{\nu-1} - u_{k-1}^{\nu-1}}{h} =$$
$$= u'_t(x_k, t_{\nu-1}) + au'_x(x_k, t_{\nu-1}) +$$
$$+ [\tau u''_{tt}(x_k, \theta_{k\nu}) - ahu''_{xx}(\eta_{k\nu}, t_{\nu-1})]/2. \qquad (9)$$

Поскольку частные производные u'_t, u'_x решения задачи (1), (2) непрерывны на замкнутой полосе $-\infty < x < \infty$, $0 \leqslant t \leqslant T$, то по непрерывности это решение удовлетворяет уравнению (1) и при $t = 0$. Значит, выполняется равенство

$$u'_t(x_k, t_{\nu-1}) + au'_x(x_k, t_{\nu-1}) = 0 \qquad (10)$$

для всех k и всех ν, в том числе и для $\nu = 1$, отвечающего $t = 0$.

Согласно определению (7) оператора $l^h y$ и в силу (2) решение u задачи (1), (2) удовлетворяет на сетке ω_h^* начальному условию

$$(l^h u)_k = u_k^0 = u(x_k, 0) = g(x_k) = g_k. \qquad (11)$$

Итак, на основании (9)—(11) для решения u дифференциальной задачи Коши (1), (2) выполняются на сетке $\omega_h = \omega'_h \cup \omega_h^*$ следующие равенства:

$$(L^h u)_k^\nu = \psi_k^\nu \quad \text{на} \quad \omega'_h, \qquad (12)$$

$$(l^h u)_k = g_k \quad \text{на} \quad \omega_h^*, \qquad (13)$$

где

$$\psi_k^\nu = [\tau u''_{tt}(x_k, \theta_{k\nu}) - ahu''_{xx}(\eta_{k\nu}, t_{\nu-1})]/2 \qquad (14)$$

есть невязка указанного решения для разностного уравнения (6).

В силу (3)—(5) и (14) невязка ψ удовлетворяет неравенству

$$\|\psi\|'_h \leqslant c_0(h + \tau), \qquad (15)$$

где $c_0 = a(1 + a)G_2/2$ — постоянная, не зависящая ни от h, ни от τ. Начальное же условие (2), как видно из (7), (13), аппроксимируется на ω_h^* точно. Таким образом, разностная схема (6), (7) аппроксимирует

задачу Коши (1), (2) на ее решении (4) с первым порядком относительно h и τ.

Вычислительный алгоритм. Разрешив уравнение (6) относительно y_k^ν, получим

$$y_k^\nu = \left(1 - a\frac{\tau}{h}\right) y_k^{\nu-1} + a\frac{\tau}{h} y_{k-1}^{\nu-1}. \tag{16}$$

Поскольку y_k^0 известны из начального условия (7), то по формуле (16) можно вычислить «слой за слоем» сначала y_k^1 для $k = 0, \pm 1, \ldots$, затем y_k^2 для всех k и т. д.

Устойчивость. Рассмотрим разностную задачу Коши

$$L^h z = \xi, \quad l^h z = 0, \tag{17}$$

где ξ — произвольная сеточная функция, заданная на ω_h', причем $\|\xi\|_h' < \infty$. Задача (17), аналогично (16), может быть записана в следующем виде:

$$z_k^\nu = \left(1 - a\frac{\tau}{h}\right) z_k^{\nu-1} + a\frac{\tau}{h} z_{k-1}^{\nu-1} + \tau\xi_k^\nu, \tag{18}$$

$$z_k^0 = 0. \tag{19}$$

Предположим, что

$$0 < a\frac{\tau}{h} \leqslant 1. \tag{20}$$

Тогда, очевидно,

$$\left|1 - a\frac{\tau}{h}\right| + \left|a\frac{\tau}{h}\right| = 1$$

и, следовательно, в соответствии с (18) имеем

$$\sup_k |z_k^\nu| \leqslant \sup_k |z_k^{\nu-1}| + \tau\|\xi\|_h'. \tag{21}$$

Поскольку согласно (19) $\sup\limits_k |z_k^0| = 0$, то из (21) получаем

$$\sup_k |z_k^\nu| \leqslant \nu\tau\|\xi\|_h'.$$

Отсюда для решения разностной задачи Коши (17) находим оценку

$$\|z\|_h = \sup_{\omega_h} |z_k^\nu| = \max_{0 \leqslant \nu \leqslant M} \sup_k |z_k^\nu| \leqslant$$

$$\leqslant \max_{0 \leqslant \nu \leqslant M} \nu\tau\|\xi\|_h' = M\tau\|\xi\|_h' = T\|\xi\|_h',$$

где постоянная T не зависит от h, τ и сеточной функции ξ. Эта оценка и означает устойчивость разностной схемы (8), т. е. разностной задачи Коши (6), (7), по правой части (при условии (20)).

Сходимость. Поскольку начальное условие (2) аппроксимируется на сетке ω_h^* точно, разностное уравнение (6) согласно (12), (15) аппроксимирует дифференциальное уравнение (1) с первым порядком относительно h и τ и, наконец, разностная схема (6), (7) устойчива по правой части, то в силу основной теоремы 29.3 с учетом замечания 29.4 и разъяснения к теореме в конце § 29 имеем при условии (20)

$$\|u - y\|_h = O(h + \tau),$$

где u — решение задачи Коши (1), (2), y — решение разностной схемы (6), (7).

Другими словами, решение разностной схемы (6), (7) сходится при измельчении сетки к решению задачи Коши (1), (2) с первым порядком относительно h и τ, если $h \to 0$, $\tau \to 0$ с соблюдением требования (20).

Исследование устойчивости методом возмущения. Покажем, что условие (20) является существенным и нарушение его может привести к отсутствию устойчивости. Предположим, что $h \to 0$, $\tau \to 0$, причем

$$a \frac{\tau}{h} = \text{const} = \lambda > 1. \tag{22}$$

При этом, очевидно, аппроксимация, т. е. неравенство (15), сохраняется, так как при выводе неравенства (15) никаких ограничений на соотношение h и τ не накладывалось.

Рассмотрим разностную задачу Коши (17), или, что одно и то же, задачу (18), (19), где

$$\xi_0^1 = \tau^m \varepsilon_0, \quad \xi_k^\nu = 0 \quad \text{при} \quad (\nu - 1)^2 + k^2 \neq 0, \tag{23}$$

$\varepsilon_0 > 0$, $m \geqslant 0$ — любые фиксированные числа. Таким образом, функция ξ отлична от нуля только в одной точке, т. е. возмущение задано только в одном узле $(0, \tau)$, причем это возмущение достаточно быстро затухает при $\tau \to 0$, если m достаточно велико. Согласно (23)

$$\|\xi\|'_h = \sup_{\omega'_h} |\xi_k^\nu| = |\xi_0^1| = \tau^m \varepsilon_0. \tag{24}$$

Обозначим $\varepsilon = \tau^{m+1}\varepsilon_0$. Тогда на основании (18), (19), (23) и (22) имеем (рис. 21)

$$z_0^1 = \tau\xi_0^1 = \tau^{m+1}\varepsilon_0 = \varepsilon, \qquad z_k^1 = 0, \quad k > 0,$$

$$z_1^2 = \left(1 - a\frac{\tau}{h}\right)\cdot 0 + \lambda z_0^1 = \lambda\varepsilon, \quad z_k^2 = 0, \quad k > 1,$$

$$z_2^3 = \left(1 - a\frac{\tau}{h}\right)\cdot 0 + \lambda z_1^2 = \lambda^2\varepsilon, \quad z_k^3 = 0, \quad k > 2,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$z_{M-1}^M = \lambda^{M-1}\varepsilon = \lambda^{M-1}\tau^{m+1}\varepsilon_0 = \frac{\lambda^{M-1}}{M^{m+1}}T^{m+1}\varepsilon_0 \to \infty$$

при $\tau = 1/M \to 0$ $(M \to \infty)$, каковы бы ни были фиксированные $\varepsilon_0 > 0$, $m \geqslant 0$.

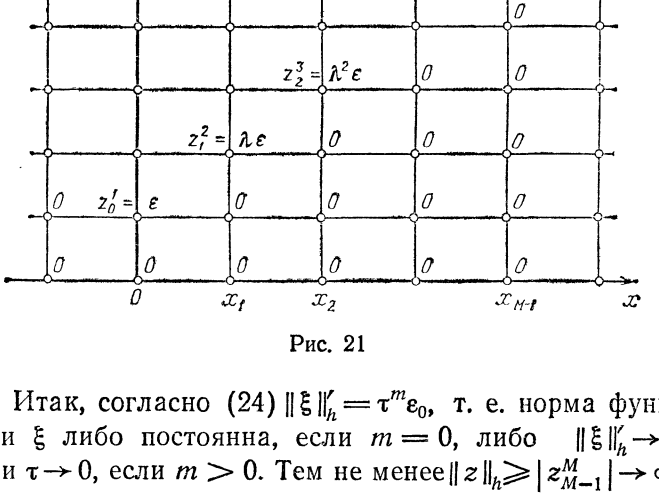


Рис. 21

Итак, согласно (24) $\|\xi\|_h' = \tau^m\varepsilon_0$, т. е. норма функции ξ либо постоянна, если $m = 0$, либо $\|\xi\|_h' \to 0$ при $\tau \to 0$, если $m > 0$. Тем не менее $\|z\|_h \geqslant |z_{M-1}^M| \to \infty$ при $\tau \to 0$. Поэтому неравенство вида $\|z\|_h \leqslant c\|\xi\|_h'$ для решения задачи (17), означающее устойчивость разностной схемы (8) по правой части, ни при какой постоянной c, не зависящей от h и τ, невозможно. *При условии* (22) *разностная схема* (6), (7) *неустойчива, хотя аппроксимация сохранилась.*

Неустойчивые разностные схемы непригодны для практических целей хотя бы потому, что они чрезвы-

чайно чувствительны к погрешностям округлений. В рассмотренном примере функцию (23), умноженную на τ, можно трактовать как единственную погрешность округлений, допущенную при вычислении y_0^1 по формуле (16), z — как погрешность разностного решения, являющуюся следствием указанной погрешности округлений.

Мы видели, что если погрешность округлений убывает со скоростью τ^{m+1} при $\tau \to 0$, то при любом m возникающая в разностном решении погрешность z может расти при условии (22) по показательному закону относительно $M = T/\tau$, т. е. очень быстро. Требовать же от погрешности округлений, чтобы она убывала быстрее, чем по степенному закону относительно τ, неразумно, так как погрешность аппроксимации носит степенной характер относительно τ и h.

Разностная схема, неустойчивая при любом отношении шагов. Если в уравнении (1) производную u'_x заменить иначе, чем при построении разностной схемы (6), (7), а именно, положить

$$u'_x \approx \frac{u_{k+1}^{\nu-1} - u_k^{\nu-1}}{h}, \tag{25}$$

то, очевидно, аппроксимация порядка $O(h+\tau)$, т. е. неравенство (15), сохранится. Однако при этом вместо формулы (16) будет получена формула

$$y_k^\nu = \left(1 + a\frac{\tau}{h}\right) y_k^{\nu-1} - a\frac{\tau}{h} y_{k+1}^{\nu-1}. \tag{26}$$

Если $h \to 0$, $\tau \to 0$, причем отношение τ/h постоянно, т. е. $\tau/h = \mu$, то поскольку первый коэффициент в формуле (26) больше единицы ($\lambda = 1 + a\tau/h = 1 + a\mu > 1$), описанным выше методом возмущения легко устанавливается неустойчивость рассматриваемой разностной схемы при любом μ.

Замечание. Аппроксимация производной u'_x в точке $(x_k, t_{\nu-1})$ по формуле (25) на первый взгляд является более естественной, чем это сделано в (6) с помощью разделенной разности «назад». Однако, как мы видим, более естественная аппроксимация может оказаться менее удачной с точки зрения устойчивости.

§ 31. Смешанная задача для уравнения теплопроводности

Постановка смешанной задачи. Пусть

$$\bar{D} = \{(x, t):\ 0 \leqslant x \leqslant 1,\ 0 \leqslant t \leqslant T\},$$
$$D' = \{(x, t):\ 0 < x < 1,\ 0 < t \leqslant T\},$$

т. е. $\bar{D}$ — замкнутый прямоугольник, D' — полуоткрытый прямоугольник. Требуется найти непрерывную на замкнутом прямоугольнике $\bar{D}$ функцию $u(x, t)$, которая на D' удовлетворяет уравнению теплопроводности

$$Lu \equiv \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x, t), \tag{1}$$

при $t = 0$ удовлетворяет начальному условию

$$u(x, 0) = s(x), \tag{2}$$

а при $x = 0$ и $x = 1$ подчиняется краевым условиям

$$u(0, t) = p(t),\ u(1, t) = q(t), \tag{3}$$

где $f(x, t)$, $s(x)$, $p(t)$, $q(t)$ — заданные достаточно гладкие функции, причем $s(0) = p(0)$, $s(1) = q(0)$.

Задача (1)—(3) называется *смешанной*, поскольку она содержит как *начальное условие*, так и *краевые условия*. Известно, что у поставленной задачи существует единственное решение $u(x, t)$. Мы будем предполагать, что это решение имеет на замкнутом прямоугольнике $\bar{D}$ непрерывные частные производные $\partial u/\partial t$, $\partial^2 u/\partial t^2$, $\partial^2 u/\partial x^2$, $\partial^4 u/\partial x^4$.

Сетки и нормы. Пусть $h = 1/N$, $\tau = T/M$ — шаги по x и t, где N, M — натуральные, $x_k = kh$, $t_\nu = \nu\tau$, $u_k^\nu = u(x_k, t_\nu)$. Построим сетки (рис. 22)

$$\omega_h = \{(x_k, t_\nu):\ k = 0, 1, \ldots, N,\ \nu = 0, 1, \ldots, M\},$$
$$\omega_h' = \{(x_k, t_\nu):\ k = 1, 2, \ldots, N-1,\ \nu = 1, 2, \ldots, M\},$$
$$\omega_h^* = \omega_h \setminus \omega_h'.$$

Сетка ω_h^* состоит из узлов сетки ω_h, обозначенных на рис. 22 крестиками. Эти узлы расположены на трех сторонах прямоугольника $\bar{D}$, на которых заданы начальное и краевые условия. Сетка ω_h' состоит из

остальных узлов сетки ω_h. Зададим для сеточных функций, определенных на ω_h или на ω_h', следующие нормы:

$$\|y\|_h = \max_{\omega_h} |y_k^\nu|, \qquad \|y\|_h' = \max_{\omega_h'} |y_k^\nu|. \tag{4}$$

Разностные схемы. Введем разностный оператор Λ:

$$\Lambda y_k^\nu = -\frac{y_{k-1}^\nu - 2y_k^\nu + y_{k+1}^\nu}{h^2}. \tag{5}$$

Здесь под выражением Λy_k^ν подразумевается значение сеточной функции Λy в точке (x_k, t_ν), т. е. $(\Lambda y)_k^\nu$.

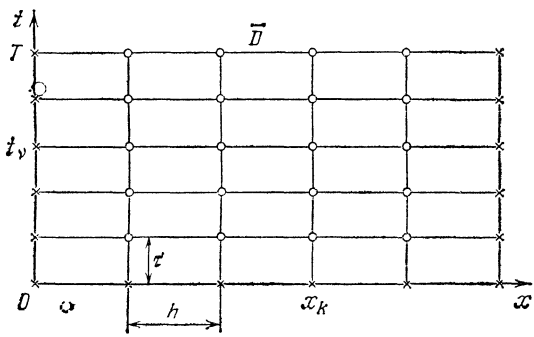


Рис. 22

Скобки опущены для упрощения записи. Аналогичные упрощения в записи будем допускать и при введении других операторов.

Зададим на сетке ω_h^* тождественный оператор

$$l^h y \equiv y \tag{6}$$

и сеточную функцию

$$g = \begin{cases} s(x_k), & x = x_k, \quad t = 0, \quad k = 1, 2, \ldots, N-1, \\ p(t_\nu), & x = 0, \quad t = t_\nu, \quad \nu = 0, 1, \ldots, M, \\ q(t_\nu), & x = 1, \quad t = t_\nu, \quad \nu = 0, 1, \ldots, M. \end{cases} \tag{7}$$

Рассмотрим две разностные схемы

$$L_1^h y_k^\nu \equiv \frac{y_k^\nu - y_k^{\nu-1}}{\tau} + \Lambda y_k^{\nu-1} = f_k^{\nu-1}, \tag{8}$$

$$l^h y = g; \tag{9}$$

$$L_2^h y_k^\nu \equiv \frac{y_k^\nu - y_k^{\nu-1}}{\tau} + \Lambda y_k^\nu = f_k^\nu, \tag{10}$$

$$l^h y = g. \tag{11}$$

Здесь и далее $k = 1, 2, \ldots, N-1$, а $\nu = 1, 2, \ldots, M$.

Шаблоны разностных уравнений (8), (10) представлены соответственно на рис. 23, 24. Обе разностные схемы (8), (9) и (10), (11) называются *двухслойными*, так как шаблоны разностных уравнений

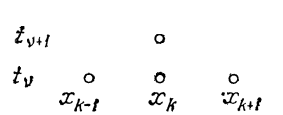


Рис. 23

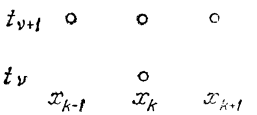


Рис. 24

(8) и (10) содержат узлы, лежащие только на двух *слоях* — подмножествах сетки ω_h, отвечающих значениям времени $t = t_{\nu-1}$ и $t = t_\nu$.

Слой, находящийся на горизонтальной прямой $t = t_{\nu-1}$, называется *нижним*, а другой — *верхним*. Разностные схемы (8), (9) и (10), (11) отличаются тем, что в уравнении (8) оператор Λ действует на нижнем слое, а в уравнении (10) оператор Λ вынесен на верхний слой и, кроме того, значения правой части $f_k^{\nu-1} = f(x_k, t_{\nu-1})$ и $f_k^\nu = f(x_k, t_\nu)$ берутся на разных слоях. Ограничимся пока сделанным формальным описанием двух разностных схем. Их качественное различие выяснится ниже.

Аппроксимация. Сопоставляя, с одной стороны, дифференциальное уравнение (1), а с другой стороны, разностные уравнения (8) и (10), видим, что частной производной u_t' отвечает разностная производная $(y_k^\nu - y_k^{\nu-1})/\tau$, а частной производной $-u_{xx}''$ соответствует разностная производная второго порядка в направлении x, образуемая с противоположным знаком с помощью оператора Λ (см. (5)).

Пусть $u(x, t)$ — решение задачи (1)—(3). Поскольку его частные производные $\partial^2 u/\partial t^2$, $\partial^4 u/\partial x^4$ по предположению непрерывны и, следовательно, ограничены на замкнутом прямоугольнике $\bar{D}$, то согласно (5), (10.3), (10.1)

$$\Lambda u_k^{\nu-1} = -u''_{xx}(x_k, t_{\nu-1}) + r_k^\nu, \tag{12}$$

$$\frac{u_k^\nu - u_k^{\nu-1}}{\tau} = u'_t(x_k, t_{\nu-1}) + \rho_k^\nu, \tag{13}$$

где $k = 1, 2, \ldots, N-1$, $\nu = 1, 2, \ldots, M$,

$$|r_k^\nu| \leqslant c_1 h^2, \quad |\rho_k^\nu| \leqslant c_2 \tau, \tag{14}$$

c_1, c_2 — некоторые постоянные, не зависящие от h, τ, k, ν.

В силу непрерывности частных производных u'_t, u''_{xx} на $\bar{D}$ решение задачи (1)—(3) удовлетворяет уравнению (1) на замкнутом прямоугольнике $\bar{D}$. Следовательно, выполняется равенство

$$u'_t(x_k, t_{\nu-1}) - u''_{xx}(x_k, t_{\nu-1}) = f_k^{\nu-1} \tag{15}$$

для $k = 1, 2, \ldots, N-1$, $\nu = 1, 2, \ldots, M$, т. е., в частности, и для $t_{\nu-1} = 0$.

Согласно (12), (13), (15) невязка ψ_1 решения u задачи (1)—(3) для разностного уравнения (8) имеет следующее выражение:

$$\psi_{1k}^\nu = L_1^h u_k^\nu - f_k^{\nu-1} = \frac{u_k^\nu - u_k^{\nu-1}}{\tau} + \Lambda u_k^{\nu-1} - f_k^{\nu-1} =$$

$$= u'_t(x_k, t_{\nu-1}) + \rho_k^\nu - u''_{xx}(x_k, t_{\nu-1}) + r_k^\nu - f_k^{\nu-1} = r_k^\nu + \rho_k^\nu.$$

Отсюда с учетом (14) получаем

$$\|\psi_1\|'_h = \max_{\omega_h} |\psi_{1k}^\nu| = \max_{1 \leqslant \nu \leqslant M} \max_{1 \leqslant k \leqslant N-1} |r_k^\nu + \rho_k^\nu| =$$

$$= O(h^2 + \tau). \tag{16}$$

Аналогично находим

$$\|\psi_2\|'_h = O(h^2 + \tau), \tag{17}$$

где ψ_2 — невязка решения u задачи (1)—(3) для разностного уравнения (10).

Таким образом, оба разностных уравнения (8) и (10) аппроксимируют дифференциальное уравнение (1) на решении u задачи (1)—(3) со вторым порядком по h и с первым порядком по τ.

Дополнительные условия, т. е. начальное условие (2) и краевые условия (3), аппроксимируются на сетке ω_h^* с помощью тождественного оператора l^h условиями (9) или соответственно условиями (11) точно, т. е. невязка решения u задачи (1)—(3) для условий (9), а также для (11) равна нулю на сетке ω_h^*.

Итак, обе разностные схемы (8), (9) и (10), (11) с точки зрения аппроксимации задачи (1)—(3) обладают по порядку относительно h и τ одинаковой гарантируемой точностью.

Вычислительные алгоритмы. Разрешив разностное уравнение (8) относительно y_k^ν, получим

$$y_k^\nu = \frac{\tau}{h^2} y_{k-1}^{\nu-1} + \left(1 - \frac{2\tau}{h^2}\right) y_k^{\nu-1} + \frac{\tau}{h^2} y_{k+1}^{\nu-1} + \tau f_k^{\nu-1}. \quad (18)$$

Поскольку y_k^0, y_0^ν, y_N^ν, $k=1, 2, \ldots, N-1$, $\nu=0, 1, \ldots, M$, известны (они задаются на ω_h^* условием (9)), решение разностной схемы (8), (9) находится по формуле (18) явно, слой за слоем. Разностная схема (8), (9) называется поэтому *явной*.

Разностное уравнение (10) с учетом (5) может быть записано в виде

$$\frac{\tau}{h^2} y_{k-1}^\nu - \left(1 + \frac{2\tau}{h^2}\right) y_k^\nu + \frac{\tau}{h^2} y_{k+1}^\nu = -y_k^{\nu-1} - \tau f_x^\nu. \quad (19)$$

Согласно (6), (7), (11) имеем также

$$y_0^\nu = p_\nu, \quad y_N^k = q_\nu. \quad (20)$$

Таким образом, если $y_k^{\nu-1}$, $k=1, 2, \ldots, N-1$, известны (в частности, y_k^0, $k=1, 2, \ldots, N-1$, заданы условием (11)), то для нахождения решения разностной схемы (10), (11) на следующем ν-м слое нужно решить трехточечное разностное уравнение (19) с краевыми условиями первого рода (20), т. е. разностную краевую задачу вида (22.1), (22.2). Поэтому разностная схема (10), (11) называется *неявной*.

Для нахождения разностного решения на ν-м слое может быть применен метод прогонки, поскольку для задачи (19), (20) условия (22.3) выполнены (проверьте, положив $k=j$, $y_k^\nu=z_j$, $y_{k\pm1}^\nu=z_{j\pm1}$, $-y_k^{\nu-1}-\tau f_k^\nu=F_j$). При этом число выполняемых арифметических действий для нахождения разностного решения на одном слое согласно (22.11) есть $O(N)$, т. е. по порядку относительно N не больше, чем при применении явной формулы (18) для схемы (8), (9).

Устойчивость и сходимость. Поскольку дополнительные условия (2), (3) аппроксимируются в разностных схемах (8), (9) и (10), (11) на сетке ω_h^* точно, то нам будет достаточно исследовать устойчивость только по правой части. Остановимся сначала на разностной схеме (8), (9).

Для исследования ее устойчивости по правой части нужно рассмотреть решение z вспомогательной разностной задачи

$$L_1^h z_k^\nu \equiv \frac{z_k^\nu - z_k^{\nu-1}}{\tau} + \Lambda z_k^{\nu-1} = \xi_k^\nu, \tag{21}$$

$$l^h z = 0, \tag{22}$$

где ξ — произвольная заданная на ω_h' сеточная функция. Разрешив разностное уравнение (21) относительно z_k^ν, аналогично (18) получим

$$z_k^\nu = \frac{\tau}{h^2} z_{k-1}^{\nu-1} + \left(1 - \frac{2\tau}{h^2}\right) z_k^{\nu-1} + \frac{\tau}{h^2} z_{k+1}^{\nu-1} + \tau\xi_k^\nu, \tag{23}$$

$$k=1,\ 2,\ \ldots,\ N-1, \qquad \nu=1,\ 2,\ \ldots,\ M.$$

Кроме того, в соответствии с (22) имеем

$$\begin{aligned} &z_k^0 = 0, \quad k=1,\ 2,\ \ldots,\ N-1; \\ &z_0^\nu = z_N^\nu = 0, \quad \nu=0,\ 1,\ \ldots,\ M. \end{aligned} \tag{24}$$

Предположим, что τ и h удовлетворяют следующему условию:

$$\tau/h^2 \leqslant 1/2. \tag{25}$$

Тогда, очевидно,

$$\frac{\tau}{h^2} + \left|1 - \frac{2\tau}{h^2}\right| + \frac{\tau}{h^2} = 1.$$

Отсюда и из (23), (24) вытекает неравенство

$$\max_{0\leqslant k\leqslant N}\left|z_k^{\nu}\right|\leqslant \max_{0\leqslant k\leqslant N}\left|z_k^{\nu-1}\right|+\tau\max_{1\leqslant k\leqslant N-1}\left|\xi_k^{\nu}\right|, \quad (26)$$

и поскольку $\max\limits_{0\leqslant k\leqslant N}\left|z_k^0\right|=0$, то

$$\max_{0\leqslant k\leqslant N}\left|z_k^{\nu}\right|\leqslant \nu\tau\,\|\xi\|_h'.$$

Следовательно,

$$\|z\|_h=\max_{0\leqslant \nu\leqslant M}\max_{0\leqslant k\leqslant N}\left|z_k^{\nu}\right|\leqslant M\tau\,\|\xi\|_h'=T\,\|\xi\|_h',$$

или, окончательно,

$$\|z\|_h\leqslant T\,\|\xi\|_h'. \quad (27)$$

Полученное неравенство для решения задачи (21), (22), в котором постоянная T не зависит от h, τ, а также от функции ξ, и означает устойчивость разностной схемы (8), (9) по правой части при условии (25). Можно доказать, что нарушение условия (25) может привести к нарушению устойчивости разностной схемы (8), (9). В частности, если $h\to 0$, $\tau\to 0$, $\tau/h^2\geqslant$ $\geqslant \text{const}>1/2$, то разностная схема (8), (9) неустойчива.

Для исследования устойчивости разностной схемы (10), (11) зададим на ω_h' произвольную сеточную функцию ξ и рассмотрим разностную задачу

$$L_2^h z_k^{\nu}\equiv\frac{z_k^{\nu}-z_k^{\nu-1}}{\tau}+\Lambda z_k^{\nu}=\xi_k^{\nu}, \quad (28)$$

$$l^h z=0, \quad (29)$$

причем не накладывая никаких ограничений на соотношение шагов τ и h. Задачу (28), (29) можно аналогично (19), (20) записать в следующем виде:

$$\frac{\tau}{h^2}z_{k-1}^{\nu}-\left(1+\frac{2\tau}{h^2}\right)z_k^{\nu}+\frac{\tau}{h^2}z_{k-1}^{\nu}=-z_k^{\nu-1}-\tau\xi_k^{\nu}, \quad (30)$$

$$z_0^{\nu}=0,\quad z_N^{\nu}=0. \quad (31)$$

Если $z_k^{\nu-1}$, $k=1, 2, \ldots, N-1$, известны (в частности, по условию (29) $z_k^0=0$, $k=0, 1, \ldots, N$), то, как отмечалось выше, для разностной задачи (30), (31), где ν фиксировано, выполнены условия (22.3).

Следовательно, по лемме 22.1 эта задача однозначно разрешима на ν-м слое.

Очевидно, имеется такое k', $0 < k' < N$, что

$$|z^{\nu}_{k'}| = \max_{0 \leqslant k \leqslant N} |z^{\nu}_{k}|. \tag{32}$$

Поскольку $|z^{\nu}_{k'-1}| \leqslant |z^{\nu}_{k'}|$, $|z^{\nu}_{k'+1}| \leqslant |z^{\nu}_{k'}|$, то

$$|z^{\nu}_{k'}| \leqslant \left| z^{\nu}_{k'} \left(1 + \frac{2\tau}{h^2}\right) - \frac{\tau}{h^2} (z^{\nu}_{k'-1} + z^{\nu}_{k'+1}) \right|$$

и, следовательно, согласно (30)

$$|z^{\nu}_{k'}| \leqslant |z^{\nu-1}_{k'}| + \tau |\xi^{\nu}_{k'}|.$$

Из полученного неравенства с учетом (32) вытекает неравенство (26) и, в конечном счете, оценка (27), что и означает устойчивость по правой части разностной схемы (10), (11) при любом соотношении шагов τ и h.

Итак, поскольку дополнительные условия (2), (3) аппроксимируются на ω_h^* точно, то из аппроксимации (см. (16), (17)) и установленной устойчивости по правой части в силу основной теоремы теории разностных схем (см. § 29) вытекает сходимость решений разностных схем (8), (9) и (10), (11) к решению задачи (1)—(3) со вторым порядком по h и с первым порядком по τ, т. е.

$$\|u - y\|_h = O(h^2 + \tau). \tag{33}$$

При этом в случае явной схемы (8), (9) предполагается выполнение ограничения (25).

Определение. Разностная схема, устойчивая при любом соотношении шагов τ и h, называется *абсолютно устойчивой*, а устойчивая при ограничениях на τ и h — *условно устойчивой*.

Недостатком разностной схемы (8), (9) является ее условная устойчивость (ограничение (25) является жестким для шага τ по времени). Преимущество — простота счета по явной формуле (18) и возможность распространения на задачу Коши (когда условие (2) задано на всей оси x, а краевые условия (3) отсутствуют). В случае смешанной задачи (1)—(3) предпочтение отдают неявной абсолютно устойчивой разностной схеме (10), (11). Разностная краевая задача (19), (20) при переходе на каждый следующий слой решается методом прогонки весьма эффективно.

§ 32. Волновое уравнение

Рассмотрим смешанную задачу для волнового уравнения

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = f(x, t), \quad 0 < x < 1, \quad 0 < t \leqslant T, \tag{1}$$

$$u(x, 0) = p(x), \quad u'_t(x, 0) = q(x), \quad 0 \leqslant x \leqslant 1, \tag{2}$$

$$u(0, t) = 0, \quad u(1, t) = 0, \quad 0 \leqslant t \leqslant T, \tag{3}$$

где $f(x, t)$, $p(x)$, $q(x)$ — заданные достаточно гладкие функции, причем $p(0) = p(1) = q(0) = q(1) = 0$.

Будем предполагать, что задача (1)—(3) имеет решение $u(x, t) \in C_4(\bar{D})$, $\bar{D} = \{(x, t)\colon 0 \leqslant x \leqslant 1, 0 \leqslant t \leqslant T\}$ — замкнутый прямоугольник. Это решение единственно.

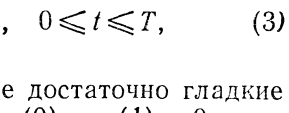


Рис. 25

Разностная схема. Будем использовать сетки, построенные на замкнутом прямоугольнике $\bar{D}$ в § 31, и соответствующие обозначения сеточных функций. Заменяем в уравнении (1) частную производную u''_{tt} приближенно второй разностной производной в направлении t, частную производную $-u''_{xx}$ аппроксимируем с помощью разностного оператора (31.5) и, переобозначив u на y, приходим к разностному уравнению

$$\frac{y_k^{\nu+1} - 2y_k^\nu + y_k^{\nu-1}}{\tau^2} + \Lambda y_k^\nu = f_k^\nu, \tag{4}$$

$$k = 1, 2, \ldots, N-1, \quad \nu = 1, 2, \ldots, M-1.$$

Шаблон разностного уравнения (4) показан на рис. 25. Это уравнение можно разрешить явно относительно $y_k^{\nu+1}$. Но для того чтобы находить значения разностного решения на $\nu+1$-м слое, требуется иметь уже вычисленные значения искомого решения на двух предыдущих слоях. Поэтому нужно получить разностное решение сначала отдельно на слоях, отвечающих значениям $\nu = 0$ и $\nu = 1$. В этом нам помогут начальные условия (2).

Прежде всего, используя первое начальное условие (2), задаем

$$y_k^0 = p_k, \quad k = 1, 2, \ldots, N-1. \tag{5}$$

Кроме того, полагаем при $k = 1, 2, \ldots, N-1$

$$y_k^1 = p_k + \tau q_k + \frac{\tau^2}{2}(f_k^0 - \Lambda p_k). \tag{6}$$

Правая часть формулы (6) аппроксимирует многочлен Тейлора

$$u(x_k, 0) + \tau u_t'(x_k, 0) + \frac{\tau^2}{2} u_{tt}''(x_k, 0),$$

поскольку согласно (2) $u(x_k, 0) = p_k$, $u_t'(x_k, 0) = q_k$, а из уравнения (1) для частных производных решения задачи (1)—(3) вытекает связь

$$u_{tt}''(x_k, 0) = f(x_k, 0) + u_{xx}''(x_k, 0).$$

Для аппроксимации $-u_{xx}''(x_k, 0) = -p''(x_k)$ используется оператор (31.5).

Наконец, согласно краевым условиям (3) имеем

$$y_0^\nu = 0, \quad y_N^\nu = 0, \quad \nu = 0, 1, \ldots, M. \tag{7}$$

Теперь разностная схема (4)—(7) полностью определена. Эта схема явная трехслойная (см. шаблон на рис. 25), условно устойчивая в некоторых естественных нормах.

Если $h \to 0$, $\tau \to 0$, причем $\tau/h \leqslant c < 1$, $c = \text{const}$, то решение y разностной схемы (4)—(7) сходится к рассматриваемому решению u задачи (1)—(3) в следующем смысле:

$$\|u - y\|_h = O(h^2 + \tau^2), \tag{8}$$

где

$$\|u - y\|_h = \max_{0 \leqslant \nu \leqslant M} \left(h \sum_{k=1}^{N-1} (u_k^\nu - y_k^\nu)^2 \right)^{1/2}.$$

Схема (4)—(7) имеет второй порядок точности и по h, и по τ.

Понятие о методе прямых. Если в задаче (1)—(3) ввести дискретность только по x, то мы придем к системе линейных обыкновенных дифференци-

альных уравнений

$$\frac{d^2y_k}{dt^2} - \frac{y_{k-1} - 2y_k + y_{k+1}}{h^2} = f(x_k, t), \qquad (9)$$

$$k = 1, 2, \ldots, N-1,$$

с начальными условиями

$$y_k(0) = p_k, \quad y_k'(0) = q_k, \qquad (10)$$

причем $y_0(t) \equiv y_N(t) \equiv 0$.

При сделанном предположении относительно гладкости решения u задачи (1)—(3) имеем

$$\|u - y\|_h = O(h^2), \qquad (11)$$

где

$$\|u - y\|_h = \max_{0 \leqslant t \leqslant T} \left(h \sum_{k=1}^{N-1} (u(x_k, t) - y_k(t))^2 \right)^{1/2},$$

$y_1(t), y_2(t), \ldots, y_{N-1}(t)$ — решение задачи Коши (9), (10).

Данный метод называется *методом прямых*, поскольку приближенное решение задачи (1)—(3) ищется на прямых $x = x_k$, $k = 1, 2, \ldots, N-1$, расположенных в плоскости x, t.

Разностный же метод часто называется *методом сеток*.

§ 33. Уравнение теплопроводности с двумя пространственными переменными

Пусть

$$\bar{R} = \{(x, y, t)\colon 0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 1,\ 0 \leqslant t \leqslant T\},$$

$$R' = \{(x, y, t)\colon 0 < x < 1,\ 0 < y < 1,\ 0 < t \leqslant T\},$$

$$S = \bar{R} \setminus R',$$

т. е. $\bar{R}$ — замкнутый прямоугольный параллелепипед, R' — полуоткрытый параллелепипед, S — множество граничных точек прямоугольного параллелепипеда $\bar{R}$, не принадлежащих R', $f(x, y, t)$ — заданная на $\bar{R}$ достаточно гладкая функция.

Требуется найти непрерывную на $\bar{R}$ функцию $u(x, y, t)$, удовлетворяющую на R' уравнению тепло-

проводности

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y, t) \tag{1}$$

и, кроме того, подчиняющуюся на S дополнительному условию

$$u = 0. \tag{2}$$

Условие (2) включает в себя как начальное условие $u(x, y, 0) = 0$ при $t = 0$, так и однородные краевые условия первого рода при $x = 0$, $x = 1$, $y = 0$, $y = 1$, т. е. на боковых гранях параллелепипеда $\bar{R}$.

Смешанная задача (1), (2) имеет единственное решение $u(x, y, t)$.

Положим $h = 1/N$, $\tau = T/M$, $x_k = kh$, $y_m = mh$, $t_\nu = \nu\tau$. $f^\nu_{km} = f(x_k, y_m, t_\nu)$,

$$\Lambda_1 v^\nu_{km} = -\frac{v^\nu_{k-1,\,m} - 2v^\nu_{km} + v^\nu_{k+1,\,m}}{h^2}, \tag{3}$$

$$\Lambda_2 v^\nu_{km} = -\frac{v^\nu_{k,\,m-1} - 2v^\nu_{km} + v^\nu_{k,\,m+1}}{h^2}, \tag{4}$$

$$\Lambda v^\nu_{km} = \Lambda_1 v^\nu_{km} + \Lambda_2 v^\nu_{km}. \tag{5}$$

Явная разностная схема

$$\frac{v^\nu_{km} - v^{\nu-1}_{km}}{\tau} + \Lambda v^{\nu-1}_{km} = f^{\nu-1}_{km}, \tag{6}$$

$$k, m = 1, 2, \ldots, N-1, \quad \nu = 1, 2, \ldots, M,$$

устойчива при $\tau/h^2 \leqslant 1/4$ в равномерных нормах, аналогичных тем, которые введены в § 31 в случае одной пространственной переменной.

Неявная разностная схема

$$\frac{v^\nu_{km} - v^{\nu-1}_{km}}{\tau} + \Lambda v^\nu_{km} = f^\nu_{km}, \tag{7}$$

$$k, m = 1, 2, \ldots, N-1, \quad \nu = 1, 2, \ldots, M,$$

абсолютно устойчива, но на каждом слое по времени требуется решить систему $(N-1)^2$ уравнений.

В (6), (7) и ниже подразумевается, что разностное решение v в точках, расположенных на S, в соответствии с (2) равно нулю.

Метод переменных направлений или дробных шагов. Обе приведенные выше разностные схемы (6) и (7) обладают существенными недостатками, а именно, в одной из них имеется жесткое ограничение на шаг по времени τ, а при применении другой схемы требуется на каждом слое по времени решить систему с $(N-1)^2$ неизвестными.

Свободной от указанных недостатков является следующая разностная схема, называемая схемой *переменных направлений* или *дробных шагов*:

$$\frac{v_{km}^{\nu-1/2}-v_{km}^{\nu-1}}{\tau/2}+\Lambda_1 v_{km}^{\nu-1/2}+\Lambda_2 v_{km}^{\nu-1}=f_{km}^{\nu-1/2}, \tag{8}$$

$$\frac{v_{km}^{\nu}-v_{km}^{\nu-1/2}}{\tau/2}+\Lambda_1 v_{km}^{\nu-1/2}+\Lambda_2 v_{km}^{\nu}=f_{km}^{\nu-1/2}, \tag{9}$$

$$k,\ m=1,\ 2,\ \ldots,\ N-1,\quad \nu=1,\ 2,\ \ldots,\ M,$$

где $f_{km}^{\nu-1/2}=f(x_k,\ y_m,\ t_\nu-\tau/2)$.

В разностной схеме (8), (9) шаг τ по времени делится на два полушага. Разностное уравнение (8) отвечает первому полушагу, в нем величины $v_{km}^{\nu-1}$ и $\Lambda_2 v_{km}^{\nu-1}$ считаются уже известными (в частности, $v_{km}^{0}=0$, $k,\ m=0,\ 1,\ \ldots,\ N$), а неизвестные имеют верхний индекс $\nu-1/2$ (кроме правой части $f_{km}^{\nu-1/2}$, которая задана). Перепишем разностное уравнение (8), предварительно умножив его на $-\tau/2$, следующим образом:

$$\frac{\tau}{2h^2}v_{k-1,\,m}^{\nu-1/2}-\left(1+\frac{\tau}{h^2}\right)v_{km}^{\nu-1/2}+\frac{\tau}{2h^2}v_{k+1,\,m}^{\nu-1/2}=F_{km}^{\nu-1/2}, \tag{10}$$

где

$$F_{km}^{\nu-1/2}=\frac{\tau}{2}\left(\Lambda_2 v_{km}^{\nu-1}-f_{km}^{\nu-1/2}\right)-v_{km}^{\nu-1}$$

известно, и присоединим к разностному уравнению (10) краевые условия

$$v_{0m}^{\nu-1/2}=0,\quad v_{Nm}^{\nu-1/2}=0 \tag{11}$$

в соответствии с условием (2).

Видим, что разностная задача (10), (11) распадается на $N-1$ независимых трехточечных разност-

ных краевых задач, отвечающих каждому фиксированному m, $1 \leqslant m \leqslant N-1$. Разностная краевая задача (10), (11) решается методом прогонки (см. § 22) при каждом m отдельно. Прогонка осуществляется по индексу k, т. е. в направлении оси x. На решение разностной задачи (10), (11) при одном значении m затрачивается согласно (22.11) $O(N)$ арифметических действий, а значит, всего на $N-1$ задач расходуется $O(N^2)$ арифметических действий.

После того как найдены все неизвестные $v_{km}^{\nu-1/2}$ на промежуточном слое с номером $\nu-1/2$, переносим их в разностном уравнении (9), соответствующем второму полушагу, вправо. Это разностное уравнение переписываем в виде

$$\frac{\tau}{2h^2} v_{k,\,m-1}^{\nu} - \left(1 + \frac{\tau}{h^2}\right) v_{km}^{\nu} + \frac{\tau}{2h^2} v_{k,\,m+1}^{\nu} = F_{km}^{\nu}, \quad (12)$$

где

$$F_{km}^{k} = \frac{\tau}{2}\left(\Lambda_1 v_{km}^{\nu-1/2} - f_{km}^{\nu-1/2}\right) - v_{km}^{\nu-1/2}$$

известно, и присоединяем к уравнению (12) в соответствии с условием (2) краевые условия

$$v_{k0}^{\nu} = 0, \quad v_{kN}^{\nu} = 0. \quad (13)$$

Задача (12), (13) тоже распадается на $N-1$ трехточечных разностных краевых задач, отвечающих различным фиксированным k, $1 \leqslant k \leqslant N-1$. Каждая такая задача решается методом прогонки. Прогонка производится теперь уже по индексу m, т. е. в направлении оси y. На решение задач (12), (13) при всех k расходуется $O(N^2)$ арифметических действий.

Таким образом, при переходе от $\nu-1$-го слоя к ν-му слою по схеме (8), (9) затрачивается число арифметических действий порядка числа искомых неизвестных на одном слое, т. е. $O(N^2)$. Такая разностная схема называется *экономичной*.

Разностная схема (8), (9) является абсолютно устойчивой в некоторых естественных нормах. Если решение $u(x, y, t)$ задачи (1), (2) достаточно гладкое на замкнутом параллелепипеде $\bar{R}$, то решение v разностной схемы (8), (9) сходится к u в следующем

смысле:

$$\| u - v \|_h = O(h^2 + \tau^2), \tag{14}$$

где $\| u - v \|_h = \max\limits_{0 \leqslant \nu \leqslant M} \left(h^2 \sum\limits_{k,\, m=1}^{N-1} (u^\nu_{km} - v^\nu_{km})^2 \right)^{1/2}$.

§ 34. Задача Дирихле для уравнения Пуассона

Пусть $D = \{(x, y)\colon\ 0 < x < 1,\ 0 < y < 1\}$ — открытый квадрат, Γ — его граница, $\bar{D} = D \cup \Gamma$ — замкнутый квадрат, $f(x, y)$ — заданная на $\bar{D}$ достаточно гладкая функция.

Задача Дирихле состоит в следующем. Требуется найти непрерывную на $\bar{D}$ функцию $u(x, y)$, удовлетворяющую на открытом квадрате D уравнению Пуассона

$$-\Delta u \equiv -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) = f(x, y) \tag{1}$$

и обращающуюся на границе квадрата в нуль, т. е.

$$u = 0 \quad \text{на } \Gamma. \tag{2}$$

Задача Дирихле (1), (2) имеет единственное решение $u(x, y)$.

Положим $h = 1/N$, $x_k = kh$, $y_m = mh$, $f_{km} = f(x_k, y_m)$. Построим сетки

$$\omega_h = \{(x_k, y_m)\colon\ k, m = 0, 1, \ldots, N\},$$

$$\omega_h' = \{(x_k, y_m)\colon\ k, m = 1, 2, \ldots, N-1\},$$

$\omega_h^* = \omega_h \setminus \omega_h'$ (ω_h^* — множество узлов, лежащих на Γ).

Зададим нормы

$$\| v \|_h = \max_{\omega_h} | v_{km} |, \quad \| v \|_h' = \max_{\omega_h'} | v_{km} |.$$

Разностная схема:

$$\Lambda v_{km} = f_{km}, \quad k, m = 1, 2, \ldots, N-1, \tag{3}$$

$$v_{km} = 0 \quad \text{на } \omega_h^*, \tag{4}$$

где Λ — оператор (33.5).

Разностное уравнение (3) в более подробной записи имеет вид

$$-\frac{v_{k-1,\,m}-2v_{km}+v_{k+1,\,m}}{h^2}-\frac{v_{k,\,m-1}-2v_{km}+v_{k,\,m+1}}{h^2}=f_{km}. \tag{3*}$$

Его шаблон изображен на рис. 26.

Решение v разностной задачи Дирихле (3), (4) находится методом последовательных приближений по схеме переменных направлений (33.8), (33.9), где $f_{km}^{\nu-1/2}=f(x_k, y_m)$, v_{km}^0 — произвольные. Можно доказать, что

$$\lim_{\nu\to\infty} v_{km}^{\nu}=v_{km},$$

$k, m=1, 2, \ldots, N-1,$

при любых начальных приближениях v_{km}^0, причем наибольшая скорость сходимости достигается при $\tau\approx h/\pi$. Здесь положена в основу идея о стабилизации при $t\to+\infty$ решения уравнения теплопроводности к решению уравнения Пуассона, если f не зависит от t.

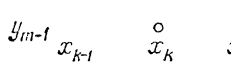


Рис. 26

Разностная схема (3), (4) устойчива по правой части, т. е. разностная задача

$$\Lambda z_{km}=\xi_{km}, \quad k, m=1, 2, \ldots, N-1,$$
$$z_{km}=0 \quad \text{на } \omega_h^*$$

при любом $h=1/N$, $N\geqslant 2$, имеет единственное решение z, и это решение удовлетворяет неравенству

$$\|z\|_h\leqslant c\|\xi\|_h', \tag{5}$$

где c — некоторая постоянная, не зависящая от h и сеточной функции ξ.

Предположим, что решение задачи Дирихле (1), (2) достаточно гладкое на замкнутом квадрате $\bar{D}$, а именно, $u(x, y)\in C_4(\bar{D})$. Тогда разностное уравнение (3) аппроксимирует дифференциальное уравнение (2) на решении u задачи (1), (2) со вторым порядком относительно h, т. е.

$$\|\psi\|_h'=O(h^2), \tag{6}$$

где

$$\psi_{km} = \Lambda u_{km} - f_{km} \tag{7}$$

есть невязка u для разностного уравнения.

При получении оценки (6) используется тот факт, что частным производным u''_{xx}, u''_{yy}, входящим в уравнение (1), в разностном уравнении (3*) отвечают вторые разностные производные, аппроксимирующие на основании (10.3) указанные частные производные с точностью $O(h^2)$. Более подробно аналогичные оценки невязок проводятся в § 30, 31.

Поскольку краевое условие (2) аппроксимируется на сетке ω_h^* согласно (4) точно, то из (6) и устойчивости разностной схемы (3), (4) по правой части вытекает сходимость ее решения v к решению $u \in C_4(\bar{D})$ задачи (1), (2) со вторым порядком относительно h, т. е.

$$\|u - v\|_h = O(h^2). \tag{8}$$

Действительно, из уравнения (3), равенства (7) и условий (2), (4) вытекает, что погрешность $r = u - v$ на сетке ω_h является решением разностной задачи

$$\Lambda r_{km} = \psi_{km}, \quad k, m = 1, 2, \ldots, N-1,$$

$$r_{km} = 0 \quad \text{на } \omega_h^*.$$

Отсюда и из (5), (6) следует (8).

Разностная схема (3), (4) обладает вторым порядком точности.

Случай произвольной области. Рассмотрим задачу Дирихле

$$-\Delta u = f(x, y) \quad \text{на } G, \tag{9}$$

$$u = \varphi \quad \text{на } \Gamma, \tag{10}$$

где G — некоторая конечная область (рис. 27), Γ — граница области G, $f(x, y)$ — заданная на области G функция, φ — заданная на границе Γ функция.

Строится, как и выше, квадратная сетка с шагом h. Во всех расположенных в области G узлах сетки, которые можно соединить с четырьмя ближайшими узлами отрезками прямых, не пересекая гра-

ницу Γ, разностное уравнение задается в следующем виде:

$$\Lambda v_{km} = f_{km}, \tag{11}$$

где Λ — оператор (33.5). Указанные узлы обозначены на рис. 27 кружками. Шаблон разностного уравнения (11) показан на рис. 26.

В узлах, находящихся в области G вблизи ее границы Γ (отмеченных на рис. 27 треугольниками), для

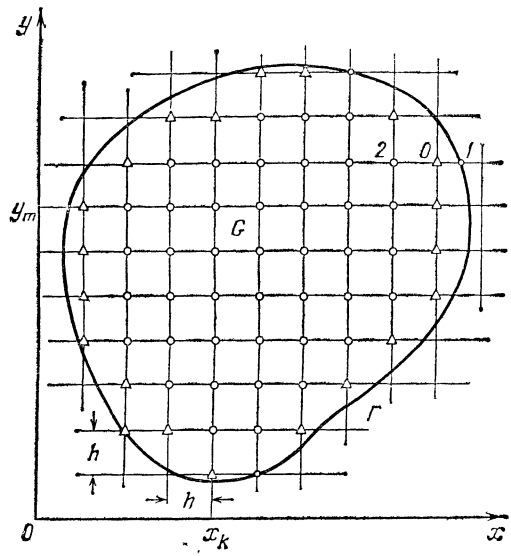


Рис. 27

задания разностных уравнений применяется линейная интерполяция в направлении оси x или оси y. Например, в точке с номером 0 уравнение имеет вид

$$v_0 = \frac{\rho_1}{\rho_1 + \rho_2} v_2 + \frac{\rho_2}{\rho_1 + \rho_2} \varphi_1, \tag{12}$$

где ρ_1 — расстояние от точки 0 до точки 1 на границе Γ, в которой берется заданное значение функции φ, обозначенное через φ_1; v_0, v_2 — неизвестные в точках 0, 2; $\rho_2 = h$ — расстояние между этими точками. Здесь для простоты используется один индекс.

Формула (12) означает линейную интерполяцию между точками 1, 2 в точку 0 (см. § 5). Аналогично разностные уравнения задаются в остальных узлах, обозначенных треугольниками. При этом расстояния от точки, в которую производится интерполяция, до обеих крайних точек не должны превышать h и одна или обе крайние точки должны лежать на границе Γ

Уравнения (11), имеющие более подробную запись (3*), разрешим относительно v_{km}:

$$v_{km} = \frac{v_{k-1,\,m} + v_{k+1,\,m} + v_{k,\,m-1} + v_{k,\,m+1}}{4} + \frac{h^2}{4} f_{km}. \quad (13)$$

Итак, в каждом узле, обозначенном кружком, задано уравнение (13), а в каждом узле, отмеченном треугольником, уравнение имеет вид (12). Общее число уравнений совпадает с числом неизвестных. Полученная система линейных алгебраических уравнений имеет единственное решение v, для нахождения которого могут быть применены методы простых итераций и Зейделя, изложенные в § 21.

Если решение задачи Дирихле (9), (10) $u(x, y) \in C_4(\bar{G})$, то справедлива оценка

$$\max_{G_h} |u - v| = O(h^2), \quad (14)$$

где G_h — множество всех узлов, обозначенных кружками и треугольниками. Решение $u(x, y)$ принадлежит классу $C_4(\bar{G})$, например, если граница Γ обладает трижды непрерывно дифференцируемой кривизной, функция φ длины s дуги границы Γ имеет ограниченную пятую производную, а $f(x, y) \in C_3(\bar{G})$.

СПИСОК ЛИТЕРАТУРЫ

1. Бахвалов Н. С. Численные методы. Т. 1. — М.: Наука, 1975.

2. Березин И. С., Жидков Н. П. Методы вычислений. Т. 1. — М.: Наука, 1966. — Т. 2. — М.: Физматгиз, 1962.

3. Воеводин В. В. Вычислительные основы линейной алгебры. — М.: Наука, 1977.

4. Годунов С. К., Рябенький В. С. Разностные схемы. — М.: Наука, 1977.

5. Дьяченко В. Ф. Основные понятия вычислительной математики. — М.: Наука, 1977.

6. Калиткин Н. Н. Численные методы. — М.: Наука, 1978.

7. Канторович Л. В., Крылов В. И. Приближенные методы высшего анализа. — М. — Л.: Физматгиз, 1962.

8. Крылов В. И., Бобков В. В., Монастырный П. И. Вычислительные методы. Т. 1. — М.: Наука, 1976. — Т. 2. — М.: Наука, 1977.

9. Ляшко И. И., Макаров В. Л., Скоробогатько А. А. Методы вычислений. — Киев: Вища школа, 1977.

10. Марчук Г. И. Методы вычислительной математики. — М.: Наука, 1980.

11. Марчук Г. И., Агошков В. И. Введение в проекционно-сеточные методы. — М.: Наука, 1981.

12. Марчук Г. И., Шайдуров В. В. Повышение точности решений разностных схем. — М.: Наука, 1979.

13. Никольский С. М. Квадратурные формулы. — М.: Наука, 1979.

14. Самарский А. А. Теория разностных схем. — М.: Наука, 1983.

15. Самарский А. А. Введение в численные методы. — М.: Наука, 1987.

16. Самарский А. А., Андреев В. Б. Разностные методы для эллиптических уравнений. — М.: Наука, 1976.

17. Самарский А. А., Гулин А. В. Устойчивость разностных схем. — М.: Наука, 1973.

18. Самарский А. А., Николаев Е. С. Методы решения сеточных уравнений. — М.: Наука, 1978.

19. Саульев В. К. Интегрирование уравнений параболического типа методом сеток. — М.: Физматгиз, 1960.

20. Соболь И. М. Численные методы Монте-Карло. — М.: Наука, 1985.

21. Стечкин С. Б., Субботин Ю. Н. Сплайны в вычислительной математике. — М.: Наука, 1976.

22. Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры. — М. — Л.: Физматгиз, 1963.

23. Яненко Н. Н. Метод дробных шагов решения многомерных задач математической физики. — Новосибирск: Наука, 1967.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ