МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

Учреждение образования

«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ»

Кафедра информатики

Л.И. Минченко

КРАТКИЙ КУРС ЧИСЛЕННОГО АНАЛИЗА

**Учебное пособие по курсу «Методы численного анализа»
для студентов специальности «Информатика»
для всех форм обучения**

Минск 2006

ДК 681.3 (075.8)
ББК 22.19 я 73
М 57

Р е ц е н з е н т:
член-корр. НАН Беларуси,
д-р физ.-мат. наук, проф. В.В.Гороховик

Минченко Л.И.
М 57 Краткий курс численного анализа. Учебное пособие по курсу «Методы численного анализа» для студ. спец. «Информатика» для всех форм обучения. Л.И. Минченко. – Мн.: БГУИР, 2006. – 92 с.: ил.
ISBN 985 444-931-9

Настоящее пособие содержит краткий курс лекций по методам численного анализа для студентов специальности «Информатика».

УДК 681.3 (075.8)
ББК 22.19 я 73

ISBN 985 444-931-9 (ч.1**)**
ISBN 985 444-930-0

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Изучение курса «Методы численного анализа» предполагает освоение методов решения классических задач вычислительной математики с помощью компьютерной техники. В пособии изложены основные понятия, определения и алгоритмы методов численного анализа, рассмотрены основные методы решения систем линейных уравнений, нахождения собственных значений и собственных векторов, численного решения нелинейных уравнений и систем нелинейных уравнений, а также излагаются основные методы аппроксимации и интерполяции функций, методы численного дифференцирования и интегрирования функций, численного решения задачи Коши и краевых задач для обыкновенных дифференциальных уравнений и дифференциальных уравнений в частных производных. Цель данного пособия: познакомить студентов с основными методами численного анализа, а также научить численному решению типичных задач вычислительной математики, достаточно сложных в вычислительном отношении и требующих применения ЭВМ.

1. ПРИБЛИЖЕННЫЕ ЧИСЛА И ДЕЙСТВИЯ НАД НИМИ

При решении некоторой задачи не всегда удается найти точное решение, т.е. возникает погрешность решения задачи, которая обусловливается следующими причинами:

1) математическое описание задачи является неточным, в частности, неточно заданы исходные данные описания;

2) применяемый для решения метод часто не является точным, так как получение точного решения возникающей математической задачи требует неограниченного или неприемлемо большого числа арифметических операций, поэтому вместо точного решения задачи приходится прибегать к приближенному.

В связи с этим числа, представляющие собой решение задачи, подразделяют на *точные,* которые дают истинное значение некоторой последовательной величины, и *приближенные,* дающие значение величины, близкое к истинному значению или, как говорят, с некоторой погрешностью. Приближенные числа используются, если точное значение некоторой величены неизвестно или использование ее точного значения нецелесообразно.

Существуют различные типы погрешностей:

1) погрешность, источником которой является несовершенство используемой математической модели (что неизбежно, поскольку она является лишь приближением к действительности);
2) погрешность, возникающая из-за неточности исходных данных, получаемых, как правило, экспериментально;
3) погрешность вычислительного метода, которым решают задачу;
4) вычислительная погрешность, появляющаяся при округлении чисел при вычислениях.

Первых два типа погрешности относят к так называемой *неустранимой погрешности,* поскольку численный анализ не располагает средствами для ее оценки или устранения. Два других типа погрешности относятся к предмету изучения численным анализом. Изучением погрешностей вычислений занимается теория погрешностей. Она решает две задачи:

1) *прямую*, когда задана погрешность исходных данных и требуется оценить погрешность результата;
2) *обратную*, когда заранее задана необходимая точность результата, а надо определить требования к точности исходных данных.

Рассмотрим примеры, которые показывают значение правильного учета и оценивания погрешности. Из этих примеров видно как самые малые ошибки в исходных данных могут влиять на численное значение решения и даже на существование самого решения.

Пример. Рассмотрим системы линейных уравнений с близкими значениями коэффициентов.

а) $\begin{cases} 3x - 7{,}0001y = 0{,}9998 \\ 3x - 7y = 1. \end{cases}$

Решение: $x = 5,\ \ y = 2.$

б) $\begin{cases} 3x - 7{,}0001y = 1 \\ 3x - 7y = 1. \end{cases}$

Решение: $x = 1/3,\ \ y = 0.$

в) $\begin{cases} 3x - 7y = 0{,}9998 \\ 3x - 7y = 1. \end{cases}$

Задача не имеет решения.

Округление чисел

Округление – замена одного числа на другое, содержащее меньшее количество цифр. Округление проводится по следующему правилу: если первая из отбрасываемых цифр, считая слева направо, больше или равна пяти, то последняя оставшаяся цифра увеличивается на единицу, в противном случае эта цифра остается без изменения.

Значащей цифрой называют первую слева отличную от нуля цифру и все последующие за ней. Например: в числе 0,07231 цифры 7, 2, 3, 1 - значащие; в числе 1,23 значащими являются все цифры; в числе 0,00720 цифры 7, 2, 0 - значащие.

Значащую цифру называют *верной*, если погрешность записи этого числа не превосходит $\frac{1}{2}$ десятичного разряда, соответствующего этой цифре. Цифры, не являющиеся верными, называются *сомнительными*.

Например: в числе 0,07231 с погрешностью 0,002 верной цифрой является 7 верная, а цифры 2, 3, 1 – сомнительные.

Приближенные числа принято записывать таким образом, чтобы вид записи давал информацию о его абсолютной погрешности, которая не должна превосходить единицы последнего разряда, сохраняемого при записи. Таким образом, записываются только верные цифры, причем верные нули на конце числа не отбрасываются. Так, сама запись числа 0,00720 указывает на погрешность, которая должна быть меньше 0,000005.

Абсолютная и относительная погрешность

Пусть x − точное значение, $\hat{x}$ − приближенное значение. Величину

$\Delta(\hat{x}) = |x - \hat{x}|$ будем называть абсолютной погрешностью приближенного числа $\hat{x}$ (идеальная абсолютная погрешность). Иногда берется значение $\Delta(\hat{x}) \geq |x - \hat{x}|$ (предельная абсолютная погрешность). Наряду с абсолютной погрешностью часто рассматривается относительная погрешность приближенного числа:

$$\delta(\hat{x}) = \frac{\Delta(\hat{x})}{|\hat{x}|} .$$

Промежуток

$$\hat{x} - \Delta(\hat{x}) \leq x \leq \hat{x} + \Delta(\hat{x})$$

принято называть интервалом приближения величины x. Различают также приближение с избытком: $x \leq \hat{x}$, приближение с недостатком: $\hat{x} \leq x$.

Так как

$$\Delta(\hat{x} + \hat{y}) = |(\hat{x} + \hat{y}) - (x + y)| = |(\hat{x} - x) + (\hat{y} - y)| \leq |\hat{x} - x| + |\hat{y} - y| ,$$

то справедливы следующие оценки абсолютной погрешности суммы и разности двух приближенных чисел

$$\Delta(\hat{x} + \hat{y}) \leq \Delta(\hat{x}) + \Delta(\hat{y}) , \quad \Delta(\hat{x} - \hat{y}) \leq \Delta(\hat{x}) + \Delta(\hat{y}).$$

Получим оценку абсолютной погрешности произведения двух приближенных чисел

$$\begin{aligned}
&\Delta(\hat{x}\hat{y}) = |\hat{x}\hat{y} - xy| = |(\hat{x} + \Delta(\hat{x}))(\hat{y} + \Delta(\hat{y})) - \hat{x}\hat{y}| \leq \\
&\leq |(\hat{x}\Delta(\hat{y}) + \hat{y}\Delta(\hat{x}) + \Delta(\hat{y}) + \Delta(\hat{x}) \cdot \Delta(\hat{y})| \leq \\
&\leq |\hat{x}|\Delta(\hat{y}) + |\hat{y}|\Delta(\hat{x}) + \Delta(\hat{x}) \cdot \Delta(\hat{y}),
\end{aligned}$$

откуда

$$\hat{z} - z = f(\hat{x}, \hat{y}) - f(x, y) = f'(x)(\xi, \eta)(\hat{x} - x) + f'(y)(\xi, \eta)(\hat{y} - y),$$

$$\Delta(\hat{x}\hat{y}) \leq |\hat{x}|\Delta(\hat{y}) + |\hat{y}|\Delta(\hat{x}) + \Delta(\hat{x}) \cdot \Delta(\hat{y}).$$

Отметим, что часто пользуются упрощенной оценкой

$$\Delta(\hat{x}\hat{y}) \leq |x|\Delta(\hat{y}) + |\hat{y}|\Delta(\hat{x}),$$

которая, как правило, дает достаточную точность.

Совершенно аналогично получается оценка для абсолютной погрешности частного двух приближенных чисел

$$\Delta\left(\frac{\hat{x}}{\hat{y}}\right) \leq \frac{|\hat{x}|\Delta(\hat{y}) + |\hat{y}|\Delta(\hat{x})}{|\hat{y}|^2 |1 - \delta(\hat{y})|} .$$

Нетрудно получить и оценки для относительной погрешности при выполнении арифметических действий. Это оценки для суммы и разности приближенных чисел

$$\delta(\hat{x}+\hat{y}) \le \frac{\Delta(\hat{x})+\Delta(\hat{y})}{|\hat{x}+\hat{y}|}, \quad \delta(\hat{x}-\hat{y}) \le \frac{\Delta(\hat{x})+\Delta(\hat{y})}{|\hat{x}-\hat{y}|},$$

оценки для их произведения

$$\delta(\hat{x}\hat{y}) \le \delta(\hat{x})+\delta(\hat{y})+\delta(\hat{x})\cdot\delta(\hat{y})$$

или

$$\delta(\hat{x}\hat{y}) \le \delta(\hat{x})+\delta(\hat{y}),$$

а также для частного

$$\delta\left(\frac{\hat{x}}{\hat{y}}\right) \le \frac{\delta(\hat{x})+\delta(\hat{y})}{|1-\delta(\hat{y})|}.$$

Последнюю оценку иногда огрубляют и используют в форме

$$\delta\left(\frac{\hat{x}}{\hat{y}}\right) \le \delta(\hat{x})+\delta(\hat{y}).$$

Рассмотрим теперь абсолютные погрешности вычисления значений функций. Пусть для функции $z = f(x,y)$ имеем x, y − точные значения переменных, а $\hat{x}, \hat{y}$ − их приближенные значения.

Оценим отклонение приближенного значения функции от точного значения. Пусть функция f непрерывно дифференцируема в некотором прямоугольнике *D*, содержащем точки (x,y) и $(\hat{x},\hat{y})$. Так как функция f непрерывно дифференцируема, то она ограничена. Обозначим:

$$C_1 = \max_{(x,y)\in D} |f'_x(x,y)|,$$

$$C_2 = \max_{(x,y)\in D} |f'_y(x,y)|.$$

Тогда, используя формулу конечных приращений Лагранжа, получим

$$\hat{z}-z = f(\hat{x},\hat{y})-f(x,y) = f'(x)(\xi,\eta)(\hat{x}-x)+f'(y)(\xi,\eta)(\hat{y}-y),$$

где $(\xi,\eta) \in D.$

Отсюда

$$|\hat{z}-z| \le C_1|\hat{x}-x|+C_2|\hat{y}-y|$$

или

$$\Delta(\hat{z}) \le C_1\Delta(\hat{x})+C_2\Delta(\hat{y}).$$

2. ПРЯМЫЕ МЕТОДЫ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ

Методы решения систем линейных уравнений делятся на две группы.

Это так называемые *прямые методы*, в которых решение представляется в виде формул или последовательности формул, и *итерационные методы*, когда решение получают в виде сходящейся последовательности приближений. К прямым методам относятся метод исключения Гаусса и его модификации, метод квадратного корня и целый ряд других методов.

Рассмотрим систему линейных уравнений:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2 \\ ..... \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = b_m . \end{cases} \qquad (2.1)$$

Введем обозначения:

$$A=\begin{pmatrix} a_{11} & ... & a_{1n} \\ ... & ... & ... \\ a_{m1} & ... & a_{mn} \end{pmatrix}, \qquad \bar{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \qquad \text{и} \qquad \bar{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

В силу введенных обозначений систему можно переписать в виде

$$A\bar{x} = \bar{b} . \qquad (2.2)$$

Решением системы будем называть любой набор значений неизвестных $x_1 = \alpha_1, x_2 = \alpha_2, ..., x_n = \alpha_n$, обращающих все уравнения системы в тождества.

Система называется *совместной*, если она имеет решения. В противном случае систему называют *несовместной*.

Говорят, что *система однородная*, если свободные члены равны нулю, т. е. $b_1 = b_2 = ... = b_m = 0$.

Критерий совместности:

Система (2.1) совместна тогда и только тогда, когда ранг расширенной матрицы $[A : \bar{b}]$ равен рангу матрицы A.

Известно, что, если матрица A – квадратная и невырожденная, то есть $|A| \neq 0$, то она имеет решение, причем единственное, и его можно найти по правилу Крамера или, вычислив обратную матрицу, записать решение в виде $\bar{x} = A^{-1} \cdot \bar{b}$. К сожалению, данные методы не эффективны для систем большой размерности ввиду трудоемкости вычисления определителей высокого порядка и обращения матриц.

2.1. Метод Гаусса

Метод Гаусса является одним из самых распространенных методов решения систем линейных уравнений. Этот метод, который также называют *методом последовательного исключения неизвестных*, известен в различных вариантах. Он состоит из прямого и обратного ходов.

Вычисления с помощью метода Гаусса на этапе прямого хода заключаются в последовательном исключении неизвестных из системы для преобразования ее к равносильной системе с верхней треугольной (трапециевидной) матрицей. Вычисления значений неизвестных производят на этапе обратного хода.

Пусть система (2.1) содержит n уравнений с n неизвестными. На первом шаге надо исключить неизвестное x_1 из уравнений с номерами $i = 2, 3, \ldots, n$. Пусть элемент $a_{11} \neq 0$. Он называется ведущим элементом первого шага.

Вычтем последовательно из второго, третьего, …, n-го уравнений системы первое уравнение, умноженное соответственно на $\frac{a_{i1}}{a_{11}}$.

Это позволит обратить в нуль коэффициенты при x_1 во всех уравнениях, кроме первого. В результате получим равносильную систему:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ \qquad a'_{22}x_2 + \ldots + a'_{2n}x_n = b'_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad a'_{n2}x_2 + \ldots + a'_{nn}x_n = b'_n, \end{cases} \tag{2.3}$$

где $a'_{ij} = a_{ij} - \frac{a_{i1}}{a_{11}}a_{1j}$, $b'_i = b_i - \frac{a_{i1}}{a_{11}} \cdot b_1$, $i = 2, \ldots, n$, $j = 2, \ldots, n$.

Продолжим дальше, предполагая, что $a'_{22} \neq 0$ и исключим неизвестное x_2 из уравнений, начиная с третьего и так далее.. На k-ом шаге предполагаем, что ведущий элемент k-го шага $a_{kk}^{(k-1)} \neq 0$, и, продолжая процесс, получаем формулы для преобразования элементов матрицы на данном шаге:

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \cdot a_{kj}^{(k-1)},$$

$$b_i^{(k)} = b_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \cdot b_k^{(k-1)},$$

$k+1 \le i \le n$, $k+1 \le j \le n$.

Это так называемый прямой ход метода Гаусса. Если на каком-то шаге получается невыполнимое равенство, это означает, что система не имеет решений. В противном случае после *(n-1)*-го шага исключений можем получить треугольную систему, из последнего уравнения которой мы найдем x_n. Подставляя его в предпоследнее уравнение, найдем x_{n-1}. И так далее. Этот процесс называется обратным ходом метода Гаусса.

Возможно также ситуация, когда в процессе прямого хода получается трапециевидная система, где в последнем оставшемся уравнении имеется более одной переменной. В этом случае придаем всем оставшимся в последнем уравнении неизвестным, кроме первого, произвольные значения $C_1, \ldots, C_l$, затем из последнего уравнения через $C_1, \ldots, C_l$ выражается оставшееся неизвестное и подставляется в предыдущие уравнения. Постепенно все

неизвестные выражаются через параметры $C_1,...,C_l$, которые могут иметь произвольные числовые значения. Таким образом, в последнем случае система имеет бесконечно много решений.

Замечание. При реализации метода Гаусса на каждом шаге производится деление на соответствующий ведущий элемент, поэтому предполагается, что эти элементы не должны быть равными нулю. В противном случае проводится перенумерация уравнений и неизвестных.

2.2. Метод ведущего (разрешающего) элемента

Метод ведущего элемента представляет собой модификацию метода Гаусса.

Рассмотрим систему уравнений (2.1) с одинаковым числом n неизвестных и уравнений:

$$\begin{cases} a_{11}x_1 + ... + a_{1j}x_j + ... + a_{1l}x_l + ... + a_{1n}x_n = b_1 \\ ..... \\ a_{i1}x_1 + ... + a_{ij}x_j + ... + a_{il}x_l + ... + a_{in}x_n = b_i \\ ..... \\ a_{k1}x_1 + ... + a_{kj}x_j + ... + a_{kl}x_l + ... + a_{kn}x_n = b_n \\ ..... \\ a_{n1}x_n + ... + a_{nj}x_j + ... + a_{nl}x_l + ... + a_{nn}x_n = b_n . \end{cases}$$

Выбираем в матрице A любой не равный нулю элемент $a_{kl} \neq 0$, который называют *ведущим элементом*. Строка и столбец, в котором находится этот элемент, называют *рабочими*. То есть рабочими будут l-й столбец и k-ая строка. Исключаем неизвестное x_l из всех уравнений, кроме k-го. В матрице новой системы *l-й* столбец принимает вид:

$$\begin{pmatrix} 0 \\ \vdots \\ 0 \\ a_{kl} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Помечаем рабочие строку и столбец. В последующих действиях они больше не могут являться рабочими.

После исключения неизвестного x_l система принимает вид

$$\left(a_{i1} - \frac{a_{il}}{a_{kl}} \cdot a_{k1}\right)x_i + ... + \left(a_{ij} - \frac{a_{il}}{a_{kl}} \cdot a_{kj}\right)x_j + ... + 0 \cdot x_l + ... + \left(a_{in} - \frac{a_{il}}{a_{kl}} \cdot a_{kn}\right)x_n = b_i - \frac{a_{il}}{a_{kl}} \cdot b_k ,$$

где $i = 1,...,n$ (кроме $i=k$), т. е. коэффициенты пересчитываются по формулам:

$$a_{ij}' = a_{ij} - \frac{a_{il}}{a_{kl}} \cdot a_{kj} = \frac{1}{a_{kl}} \cdot \begin{vmatrix} a_{ij} & a_{il} \\ a_{kj} & a_{kl} \end{vmatrix},$$

$$b_{ij}' = b_i - \frac{a_{il}}{a_{kl}} \cdot b_k = \frac{1}{a_{kl}} \cdot \begin{vmatrix} b_1 & a_{il} \\ b_2 & a_{kl} \end{vmatrix},$$

$$i = 1,...,n; \qquad i \neq k,$$

$$a'_{kj} = a_{kj}, \qquad b'_k = b_k, \qquad j = 1,...,n.$$

Опять ищем ведущий элемент и исключаем соответствующую переменную. Так продолжаем до тех пор, пока не получится уравнение с одним неизвестным. Из этого уравнения находим значение неизвестного и подставляем его в предыдущие уравнения для обратного хода. Продолжаем процесс, пока не получим значения всех неизвестных.

Пример. Пусть

$$\begin{cases} 3x_1 + 2x_2 + x_3 = -1 \\ -2x_1 + 3x_2 - 2x_3 = 1 \\ x_1 - 4x_2 + 4x_3 = 6. \end{cases}$$

Выберем ведущим элементом $a_{12} = 2$. Пересчитываем коэффициенты. Находим:

$$a'_{21} = \frac{1}{2} \cdot \begin{vmatrix} 3 & 2 \\ -2 & 3 \end{vmatrix} = -\frac{13}{2}, \quad a'_{12} = 0, \quad a'_{23} = \frac{1}{2} \cdot \begin{vmatrix} 2 & 1 \\ 3 & -2 \end{vmatrix} = -\frac{7}{2};$$

$$a'_{31} = -\frac{1}{2} \cdot \begin{vmatrix} 3 & 2 \\ 1 & -4 \end{vmatrix} = 7, \quad a'_{32} = 0, \quad a'_{33} = \frac{1}{2} \cdot \begin{vmatrix} 2 & 1 \\ -4 & 4 \end{vmatrix} = 6;$$

$$b'_2 = \frac{1}{2} \cdot \begin{vmatrix} 2 & -1 \\ 3 & 1 \end{vmatrix} = \frac{5}{2}, \quad b'_3 = \frac{1}{2} \cdot \begin{vmatrix} 2 & -1 \\ -4 & 6 \end{vmatrix} = 4.$$

Получаем:

$$\begin{cases} 3x_1 + 2x_2 + x_3 = -1 \\ 13x_1 + \quad + 7x_3 = -5 \\ 7x_1 + \quad + 6x_3 = 4. \end{cases}$$

Выбираем ведущим элементом $a_{33} = 6$ и получаем:

$$\begin{cases} 11x_1 + 12x_2 \quad = -10 \\ 29x_1 \qquad\quad = -58 \\ 7x_1 + \qquad + 6x_3 = 4. \end{cases}$$

Откуда $x_1 = -2, \qquad x_2 = 1, \qquad x_3 = 3.$

2.3. Метод прогонки

Метод прогонки является модификацией метода Гаусса для систем специального вида, так называемых трехдиагональных систем. К необходимости решения такого рода систем приводят, в частности, задачи построения кубических сплайнов и разностные схемы решения краевых задач для дифференциальных уравнений.

Трехдиагональной системой называют систему линейных уравнений с матрицей вида:

$$\begin{pmatrix} b_1 & c_1 & 0 & ... & ... & ... & 0 \\ a_2 & b_2 & c_2 & 0 & ... & ... & 0 \\ 0 & a_3 & b_3 & c_3 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & ... & ... & 0 & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & ... & ... & 0 & 0 & a_n & b_n \end{pmatrix},$$

т. е. матрицей, у которой ненулевыми могут быть только элементы, стоящие на главной и двух смежных с главной диагоналях.

Рассмотрим трехдиагональную систему:

$$\begin{cases} b_1x_1 + c_1x_2 = d_1 \\ a_2x_1 + b_2x_2 + c_2x_3 = d_2 \\ a_3x_2 + b_3x_3 + c_3x_4 = d_3 \\ ..... \\ a_{n-1}x_{n-2} + b_{n-1}x_{n-1} + c_{n-1}x_n = d_{n-1} \\ a_nx_{n-1} + b_nx_n = d_n \end{cases}$$

Суть метода прогонки заключается в построении рекуррентной последовательности для нахождения прогоночных коэффициентов A_i и B_i. При этом каждое неизвестное представляется в виде

$$x_i = A_ix_{i+1} + B_i .$$

Выражаем неизвестное x_1 из 1-го уравнения системы:

$$x_1 = -\frac{c_1}{b_1}x_2 + \frac{d_1}{b_1}.$$

То есть на первом шаге прогоночные коэффициенты имеют вид

$A_1 = -\frac{c_1}{b_1}$, $B_1 = \frac{d_1}{b_1}$, причем $x_1 = A_1x_2 + B_1$.

Далее выражаем x_2 из второго уравнения:

$$a_2(A_1x_1 + B_1) + b_2x_2 + c_2x_3 = d_2,$$

$$x_2 = \frac{-c_2x_3}{a_2A_1 + b_2} + \frac{d_2 - a_2B_1}{a_2A_3 + b_2} \quad ,$$

т. е. прогоночные коэффициенты имеют вид

$$A_2 = \frac{-c_2}{a_2A_1 + b_2}, \qquad B_2 = \frac{d_2 - a_2B_1}{a_2A_1 + b_2}, \quad \text{причем} \quad x_2 = A_2x_3 + B_2 .$$

Продолжаем процесс и находим

$$A_i = \frac{-c_i}{a_i A_{i-1} + b_i}, \qquad B_i = \frac{d_i - a_i B_{i-1}}{a_i A_{i-1} + b_i}$$

для $i = \overline{2, n-1}$.

В итоге получаем

$x_{n-1} = A_{n-1} x_n + B_{n-1}$.

Для нахождения x_n используем данное уравнение и оставшееся последнее уравнение. Получаем систему

$$\begin{cases} a_n x_{n-1} + b_n x_n = d_n \\ x_{n-1} = A_{n-1} x_n + B_{n-1}, \end{cases}$$

решая которую, находим

$$a_n \left(A_{n-1} x_n + B_{n-1} \right) + b_n x_n = d_n,$$
$$x_n = \frac{d_n - a_n B_{n-1}}{a_n A_{n-1} + b_n} = B_n.$$

Для удобства положим, что $a_1 = 0$, $c_n = 0$ и тогда формулы для прогоночных коэффициентов принимают следующий вид:

$$\begin{cases} A_i = \dfrac{-c_i}{a_i A_{i-1} + b_i} \\ B_i = \dfrac{d_i - a_i B_{i-1}}{a_i A_{i-1} + b_i}. \end{cases} \qquad i = \overline{1, n}.$$

Тогда

$$\begin{cases} \mathrm{x}_{\mathrm{i}} = A_i \cdot x_{i+1} + B_i, \quad i = \overline{1, n-1} \\ x_n = B_n. \end{cases}$$

Проводя обратный ход метода прогонки, последовательно найдем значения неизвестных $x_n, x_{n-1}, \ldots, x_1$.

Замечание. Очевидно, чтобы наши действия в выводе метода прогонки были корректны, необходимо чтобы в формулах для вычислении прогоночных коэффициентов знаменатели дробей не обращались в нуль. Можно показать, что метод прогонки будет корректным, если для его коэффициентов выполняется условие преобладания диагональных элементов, т. е. выполняется соотношение

$$|b_i| \geq |a_i| + |c_i|, \qquad \forall i = \overline{1, n},$$

в котором хотя бы одно из неравенств строгое.

2.4. Метод квадратного корня

Этот метод применяется при решении систем вида $A\bar{x} = \bar{f}$ с неособенной симметрической матрицей. Если матрица A не является симметрической, то без предварительного преобразования системы к виду

$A^* A\bar{x} = A^* \bar{f}$

метод применять нельзя. Однако преобразование системы к указанному выше виду связано с выполнением большого числа дополнительных операций умножения и сложения, число которых намного превосходит число аналогичных операций, необходимых при решении системы с симметрической матрицей по методу квадратного корня. Поэтому выполнять указанное преобразование и затем применять к решению системы метод квадратного корня, как правило не целесообразно.

Пусть матрица A симметрическая. Схема метода квадратного корня строится на идее представления матрицы в виде произведения треугольных и диагональных матриц, а именно: находим такую правую треугольную матрицу S и диагональную матрицу D с элементами ± 1 по главной диагонали, чтобы имело место равенство

$A = S^* DS,$

где приняты обозначения

$$S = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ 0 & s_{22} & \dots & s_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & s_{nn} \end{bmatrix}, \quad D = \begin{bmatrix} d_{11} & 0 & \dots & 0 \\ 0 & d_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_{nn} \end{bmatrix}.$$

Предположим, что мы нашли такие матрицы S и D, для которых имеет место равенство $A = S^* DS$.

Тогда решение системы

$A\bar{x} = \bar{f}$

осуществляется по следующему правилу.

Введем следующие обозначения:

$B = S^* D, \quad S\bar{x} = \bar{y}, \quad \bar{y} = (y_1, y_2, \dots, y_n)',$

где B − известная матрица; $\bar{y}$ − неизвестный вектор.

Для определения вектора $\bar{y}$ в силу равенства

$A\bar{x} = S^* DS\bar{x} = (S^* D)S\bar{x} = \bar{f}$

имеем такую систему линейных алгебраических уравнений:

$B\bar{y} = \bar{f}$.

Здесь особенно важно то, что матрица этой системы является левой треугольной, т. е. имеет вид

$$B = \begin{bmatrix} \beta_{11} & 0 & \dots & 0 \\ \beta_{21} & \beta_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \beta_{n1} & \beta_{n2} & \dots & \beta_{nn} \end{bmatrix}.$$

Это позволяет из системы

$B\bar{y} = \bar{f}$

сразу выписать ее решение, выполняя только обратный ход метода Гаусса сверху вниз. В результате, определив вектор $\bar{y}$ по формулам

$$\begin{cases} y_1 = \dfrac{f_1}{\beta_{11}} \\ y_k = \dfrac{f_k - \sum\limits_{s=1}^{k-1} \beta_{ks} y_s}{\beta_{kk}} \quad (k = 2,3,...,n), \end{cases}$$

находим из системы $S\bar{x} = \bar{y}$ искомое решение системы $A\bar{x} = f$.

Для этого нам надо будет в системе

$S\bar{x} = \bar{y},$

выполнить обратный ход метода Гаусса снизу вверх, после чего получим

$$\begin{cases} x_n = \dfrac{y_n}{s_{nn}} \\ x_k = \dfrac{y_k - \sum\limits_{p=k+1}^{n} s_{kp} y_p}{s_{kk}} \quad (k = n-1, n-2,...,1). \end{cases}$$

Как мы видим, для вычисления векторов $\bar{x}$ и $\bar{y}$ требуются простые, негромоздкие вычисления. Теперь, чтобы придать схеме метода окончательный вид, надо указать правило, по которому следует вычислять элементы матриц *S* и *D*. Соотношение $A = S^* DS$ можно рассматривать как систему алгебраических уравнений для определения $n(n+1)/2$ элементов матрицы *S* и *n* элементов матрицы *D*.

Так как матрица *A* − симметрическая, то мы будем располагать $n(n+1)/2$ уравнениями следующего вида:

$$s_{1\,i} d_{11} s_{1\,j} + s_{2\,i} d_{22} s_{2\,j} + ... + s_{i\,i} d_{i\,i} s_{i\,j} = a_{i\,j} \quad (i < j)$$

$$\left|s_{1\,i}\right|^2 d_{11} + \left|s_{2\,i}\right|^2 d_{22} + ... + \left|s_{i\,i}\right|^2 d_{i\,i} = a_{i\,i} \quad (i = j)$$

$(j = 1,\ 2,...,n).$

В предыдущей системе число уравнений меньше числа неизвестных на *n*. Чтобы разложение $A = S^* DS$ было однозначным, определим диагональные элементы s_{ii} так, чтобы они были положительны. Тогда из второго уравнения системы при *i=1* имеем $\left|s_{11}\right|^2 d_{11} = a_{11}$.

Положим $d_{11} = sign\ a_{11}$ и из предыдущего уравнения для s_{11} получим

$s_{11} = \sqrt{|a_{11}|}\ s_{11} = \sqrt{|a_{11}|}$.

Из первого уравнения системы при *i=1* найдем

$$s_{1j} = \frac{a_{1j}}{d_{11} s_{11}} \quad (j = 2,3,...,n).$$

Таким образом, мы сможем определить элементы первой строки матрицы *S*.

Далее, аналогично, из второго уравнения системы и из первого уравнения при *i=2* находим:

$$d_{22} = sign(a_{22} - |s_{12}|^2 d_{11}),\ \ s_{22} = \sqrt{|a_{22} - |s_{12}|^2 d_{11}|},$$

$$s_{2j} = \frac{a_{2j} - s_{12} d_{11} s_{1j}}{d_{22} s_{22}} \quad (j = 3,4,...,n).$$

Эти формулы позволяют вычислить элементы второй строки матрицы *S*. Продолжая этот процесс, мы сможем вычислить все элементы матрицы *S*. Укажем в общем виде формулы, по которым должны вестись вычисления элементов s_{ij}:

$$\left.\begin{aligned}
&d_{11} = sign a_{11}, \quad s_{11} = \sqrt{|a_{11}|}, \quad s_{1j} = \frac{a_{1j}}{d_{11} s_{11}} \\
&d_{ii} = sign(a_{ii} - \sum_{p=1}^{i-1} |s_{pi}|^2 d_{pp}) \\
&s_{ii} = \sqrt{|a_{ii} - \sum_{p=1}^{i-1} |s_{pi}|^2 d_{pp}|} \quad (i > 1) \\
&s_{ij} = \frac{a_{ij} - \sum_{p=1}^{i-1} s_{pi} d_{pp} s_{pj}}{d_{ii} s_{ii}} \quad (i > 1, \quad j = i+1, \quad i+2,..,n).
\end{aligned}\right\} \qquad (2.4)$$

Таким образом, при решении системы $A\bar{x} = \bar{f}$ по методу квадратного корня необходимо:

1) сначала убедиться в том, что A − симметрическая матрица, и затем вычислить элементы матрицы *S;*

2) используя формулы

$$y_1 = \frac{f_1}{\beta_{11}}, \quad y_k = \frac{f_k - \sum_{s=1}^{k-1} \beta_{ks} y_s}{\beta_{kk}} \quad (k = 2,3,..,n)$$, вычислить вектор $\bar{y}$;

3) наконец, по формулам

$$\left.\begin{aligned} x_n &= \frac{y_n}{s_{nn}} \\ x_k &= \frac{y_k - \sum_{p=k+1}^{n} s_{kp} y_p}{s_{kk}} \quad (k = n-1, n-2, ..., 1). \end{aligned}\right\}$$

найти искомое решение исходной системы – вектор $\bar{x}$.

Если матрица A — симметрическая положительно определенная, то ее можно разложить в произведение двух транспонированных друг другу треугольных матриц, а именно:

$$A = S^* S,$$

где S — правая треугольная. В этом случае формулы (2.4) несколько упростятся и будут иметь вид

$$\left.\begin{aligned} & s_{11} = \sqrt{a_{11}}, \quad s_{1j} = \frac{a_{1j}}{s_{11}}, \quad j = 2, ..., n, \\ & s_{ii} = \sqrt{a_{ii} - \sum_{p=1}^{i-1} s_{pi}{}^2} \quad (i > 1), \\ & s_{ij} = \frac{a_{ij} - \sum_{p=1}^{i-1} s_{pi} s_{pj}}{s_{ii}} \quad (i > 1, \quad j = i+1, \quad i+2, .., n). \end{aligned}\right\}$$

Для решения системы линейных алгебраических уравнений с симметрической матрицей порядка n по методу квадратного корня необходимо выполнить:

умножений и делений - $\frac{1}{6} n(n^2 + 9n + 8)$,

извлечений квадратных корней - n.

Отметим в заключение, что метод квадратного корня очень эффективен при решении систем с положительно определенной симметрической матрицей. Такие системы, как правило, возникают при решении задач минимизации положительно определенных квадратичных форм. Главными требованиями к исходной матрице являются симметричность и положительная определенность. Симметричность проверяется простым сравнением соответствующих элементов. Положительная определенность матрицы определяется с помощью критерия Сильвестра.

Критерий Сильвестра. Для того чтобы матрица была положительно определенной, необходимо и достаточно, чтобы все главные диагональные миноры матрицы были строго положительны.

2.5. Вычисление определителей

Для упрощения процесса вычисления определителей пользуются формулами так называемой схемы единственного деления.

Схема единственного деления состоит в следующем.

Пусть необходимо вычислить определитель

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

причем $a_{11} \neq 0$.

Выносим элемент a_{11} из первой строки и получим

$$\Delta = a_{11} \begin{vmatrix} 1 & b_{12} & \dots & b_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}, \quad \text{где } b_{ij} = \frac{a_{ij}}{a_{11}}.$$

Вычитая из каждой строки, начиная со второй, первую строку, умноженную соответственно на $a_{21}, a_{31}, \dots, a_{n1}$, получим

$$\Delta = a_{11} \begin{vmatrix} 1 & b_{12} & \dots & b_{1n} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{vmatrix} = a_{11} \begin{vmatrix} a^{(1)}{}_{22} & a^{(1)}{}_{23} & \dots & a^{(1)}{}_{2n} \\ a^{(1)}{}_{32} & a^{(1)}{}_{33} & \dots & a^{(1)}{}_{3n} \\ \dots & \dots & \dots & \dots \\ a^{(1)}{}_{n2} & a^{(1)}{}_{n3} & \dots & a^{(1)}{}_{nn} \end{vmatrix},$$

где $a^{(1)}{}_{ij} = a_{ij} - a_{ij} b_{ij}$ и соответственно $a^{(m)}{}_{ij} = a^{(m-1)}{}_{ij} - a^{(m-1)}{}_{im} b_{mj}$, $b_{mj} = \dfrac{a^{(m-1)}{}_{mj}}{a^{(m-1)}{}_{mm}}$.

С образовавшимся определителем (*n-1*)-го порядка поступаем таким же образом, если $a^{(1)}{}_{22} \neq 0$. В противном случае перенумеруем и поменяем местами строки или столбцы определителя.

Продолжая процесс, получим, что искомый определитель равен произведению ведущих элементов:

$$\Delta = a_{11} a^{(1)}{}_{22} a^{(2)}{}_{33} \dots a^{(n-1)}{}_{nn}.$$

2.6. Обращение матриц

Часто возникает необходимость обратить квадратную матрицу

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

Поскольку вычисление матрицы A^{-1} обычным способом достаточно трудоемко, поступаем следующим образом. Построим прямоугольную матрицу:

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

которая получается, если к A приписать справа единичную матрицу E.

Применим к этой матрице метод исключения Гаусса, не заботясь о правых столбцах. Когда гауссово исключение закончится, получим

$$\begin{pmatrix} 1 & 0 & \ldots\ldots\ldots & 0 & b_{11} & b_{12} & \ldots\ldots & b_{1n} \\ 0 & 1 & \ldots\ldots\ldots & 0 & b_{21} & b_{22} & \ldots\ldots & b_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots\ldots & \ldots\ldots & \ldots & \ldots & \ldots\ldots & \ldots\ldots \\ 0 & 0 & \ldots\ldots & 0 & 1 & b_{n1} & b_{n2} & \ldots\ldots\ldots b_{nn} \end{pmatrix}.$$

Если матрица A не вырождена, то матрица

$$B=\begin{pmatrix} b_{11} & b_{12} & \ldots\ldots\ldots & b_{1n} \\ b_{21} & b_{22} & \ldots\ldots\ldots & b_{2n} \\ \ldots\ldots & \ldots\ldots & \ldots\ldots\ldots & \ldots\ldots \\ b_{n1} & b_{n2} & \ldots\ldots\ldots & b_{nn} \end{pmatrix}$$

будет обратной к A. Чтобы убедиться в этом, заметим, что каждый шаг процесса исключения эквивалентен умножению слева на некоторую матрицу. Произведение всех этих левых матриц есть некоторая матрица C, умножение на которую слева приводит A к единичной матрице E, то есть $CA=E$, а будучи применено к правым столбцам, это произведение делает из единичной матрицы матрицу B: $CE=B$. В таком случае из существования обратной матрицы A^{-1} следует $C= A^{-1}$ и $C=B$, следовательно, $B= A^{-1}$.

3. ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ УРАВНЕНИЙ

3.1. Матричные нормы

Рассмотрим векторное пространство R^n. Пусть

$\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ – вектор данного пространства.

Нормой вектора называется функция $\|\vec{x}\|$ от вектора $\vec{x}$, для которой выполняются следующие аксиомы:

(H1): $\|\bar{x}\| \geq 0$, $\|\bar{x}\| = 0 \quad \Leftrightarrow \quad \bar{x} = \bar{0}$,

(H2): $\|\alpha \cdot \bar{x}\| = |\alpha| \cdot \|\bar{x}\|$, $\alpha \in R$, $\vec{x} \in R^n$,

(H3): $\|\bar{x} + \bar{y}\| \leq \|\bar{x}\| + \|\bar{y}\|$, $\forall \bar{x}, \bar{y} \in R^n$.

Пример:

$\|\bar{x}\|_1 = \max\limits_{1 \leq j \leq n} \{|x_j|\} = \max\{|x_1|,...,|x_n|\}$ - кубическая норма;

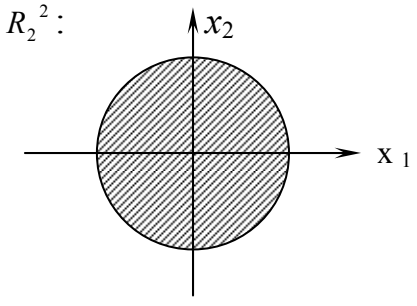
$\|\bar{x}\|_2 = \sqrt{x_1^2 + ... + x_n^2}$ – евклидова норма;

$\|\bar{x}\|_3 = \sum\limits_{j=1}^{n} |x_j| = |x_1| + ... + |x_n|$ - октаэдрическая норма.

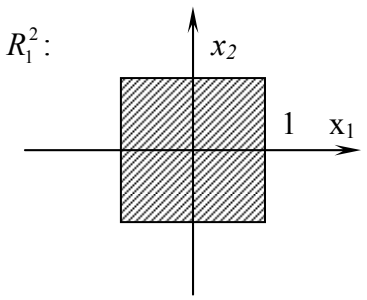
Существуют и другие менее употребляемые векторные нормы.

Векторное пространство вместе с заданной в нем нормой называется *нормированным векторным пространством*.

Пример: Рассмотрим нормированные векторные пространства с различными нормами в двумерном пространстве R^2 и соответствующие им единичные окрестности начала координат:

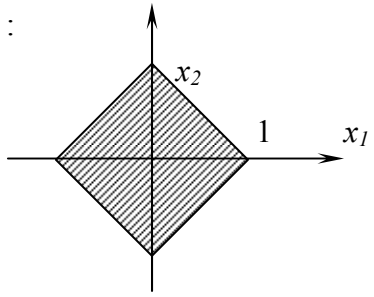


с нормой $\|\bar{x}\|_2 = \sqrt{x_1^2 + x_2^2}$,



с нормой $\|\bar{x}\|_1 = \max\{|x_1|, |x_2|\}$,

R_3^2:



с нормой $\|\bar{x}\|_3 = \sum_{j=1}^{2} |x_j| = |x_1| + |x_2|$.

Рассмотрим квадратную матрицу $A_{n\times n}$.

Нормой матрицы A, индуцированной нормой вектора $\bar{x}$, называется число

$$\|A\| \overset{def}{=} \sup_{\bar{x}\neq\bar{0}} \frac{\|A\bar{x}\|}{\|\bar{x}\|}.$$

Из этого определения непосредственно следует, что для любой индуцированной нормы справедливы следующие свойства:

1) $\|E\| = \sup_{\bar{x}\neq\bar{0}} \frac{\|E\bar{x}\|}{\|\bar{x}\|} = \sup_{\bar{x}\neq\bar{0}} \frac{\|\bar{x}\|}{\|\bar{x}\|} = 1$, где E – единичная матрица;

2) $\|O\| = 0$, где O – нулевая матрица;

3) $\|A\| \geq \frac{\|A\bar{x}\|}{\|\bar{x}\|}$, $\forall \bar{x} \neq \bar{0} \Rightarrow \|A\bar{x}\| \leq \|A\| \cdot \|\bar{x}\|$;

4) $\|AB\| = \sup_{\bar{x}\neq\bar{0}} \frac{\|AB\bar{x}\|}{\|\bar{x}\|} \leq \sup_{\bar{x}\neq\bar{0}} \frac{\|A\| \cdot \|B\bar{x}\|}{\|\bar{x}\|} = \|A\| \sup_{\bar{x}\neq\bar{0}} \frac{\|B\bar{x}\|}{\bar{x}} = \|A\| \cdot \|B\|$,

т.е.

$$\|AB\| \leq \|A\| \cdot \|B\|.$$

Легко видеть, что для индуцированной нормы матрицы выполняются соотношения:

(H1'): $\|A\| \geq 0$, $\|A\| = 0 \Leftrightarrow A = O$;

(H2'): $\|\alpha A\| = |\alpha| \cdot \|A\|$;

(H3'): $\|A + B\| \leq \|A\| + \|B\|$;

(H4'): $\|AB\| \leq \|A\| \cdot \|B\|$.

Нормой матрицы называется функция $||A||$ от матрицы A, для которой выполняются аксиомы (H1') - (H4').

Таким образом, индуцированные матричные нормы представляют собой частный случай матричных норм. Из аксиомы (H4') очевидно следует, что

$$\|A^k\| \leq \|A\|^k, \qquad k = 1,2,... \quad .$$

Найдем матричную норму $\|A\|_1$, индуцированную векторной нормой $\|\vec{x}\|_1$:

$$\|A\|_1 = \sup_{\bar{x}\neq\bar{0}} \frac{\|A\bar{x}\|_1}{\|\bar{x}\|_1} = \sup_{\bar{x}\neq\bar{0}} \frac{1}{\|\bar{x}\|_1} \max_{1\le i\le n} \left\{ \left| \sum_{j=1}^{n} a_{ij} x_j \right| \right\} \le \sup_{\bar{x}\neq\bar{0}} \frac{1}{\|\bar{x}\|_1} \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \cdot |x_j| \le$$

$$\le \sup_{\bar{x}\neq\bar{0}} \frac{1}{\|\bar{x}\|_1} \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \max_j |x_j| = \sup_{\bar{x}\neq\bar{0}} \frac{1}{\|x\|_1} \max_i \sum_{j=1}^{n} |a_{ij}| \cdot \|\bar{x}\|_1 = \max_i \sum_{j=1}^{n} |a_{ij}|.$$

Таким образом $\|A\|_1 \le \max_i \sum_{j=1}^{n} |a_{ij}|$.

Докажем, что $\|A\|_1 = \max_i \sum_{j=1}^{n} |a_{ij}|$. Для этого необходимо показать, что $\|A\|_1 \ge \max_i \sum_{j=1}^{n} |a_{ij}|$.

По определению

$\|A\|_1 = \sup_{\bar{x}\neq\bar{0}} \frac{\|A\bar{x}\|_1}{\|\bar{x}\|_1} \ge \frac{\|A\bar{x}\|_1}{\|\bar{x}\|_1}$ для любого $\bar{x}$.

Предположим, что в $\|A\|_1 = \max_i \sum_{j=1}^{n} |a_{ij}|$ максимум достигается при $i=i_0$, то есть $\max_i \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{i0j}|$.

Возьмем ненулевой вектор $\bar{x}^0$ такой, что $x_j^0 = \operatorname{sgn} a_{i_0 j}$. Тогда $\|\bar{x}^0\|_1 = 1$. Следовательно:

$$\|A\|_1 \ge \frac{\|A\bar{x}^0\|}{\|\bar{x}^0\|_1} \ge \|A\bar{x}^0\|_1 = \max_i \{ |\sum_{j=1}^{n} a_{1j} x_j^0|, \ldots, |\sum_{j=1}^{n} a_{nj} x_j^0| \} \ge |\sum_{j=1}^{n} a_{i_0 j} x_j^0| = \sum_{j=1}^{n} |a_{i_0 j}| = \max_i |\sum_{j=1}^{n} a_{ij}| .$$

То есть, действительно, $\|A\|_1 = \max_i \sum_{j=1}^{n} |a_{ij}|$.

Совершенно аналогично можно показать, что:

$\|A\|_3 = \max_j \sum_{i=1}^{n} |a_{ij}|$.

Сложнее дело обстоит с матричной нормой индуцированной евклидовой векторной нормой.

Спектральным радиусом матрицы A называется число

$\mu(A) = \max \{ |\lambda_1|, \ldots, |\lambda_n| \}$,

где $\lambda_1, \ldots, \lambda_n$ - собственные значения данной матрицы.

Лемма 1. Для любой матричной нормы $\mu(A) \le \|A\|$.

Доказательство. Действительно, пусть λ - собственное значение матрицы A и $\mu(A) = |\lambda|$. Построим матрицу B того же порядка, что и A, у которой первый столбец совпадает с собственным вектором $\bar{x}$ матрицы A, соответствующим собственному значению λ, а остальные столбцы -

нулевые. Тогда очевидно $AB = \lambda B$. Используя аксиомы из определения матричной нормы, получим

$$|\lambda| \, ||B|| \le ||A|| \, ||B||,$$

и поскольку $B \neq O$, и , следовательно, $||B|| \neq 0$, получаем $|\lambda|| \le ||A||$.

По аналогии с векторами в R^n можно определить сходимость последовательности матриц поэлементно, считая что $A^{(k)} \to A$ при $k \to \infty$ тогда и только тогда, когда $a_{ij}^{(k)} \to a_{ij}$ для всех $i,j=1, \ldots, n$. Отметим, что, как и в случае с векторами, для любой матричной нормы из условия сходимости по норме $|| A^{(k)} - A || \to 0$ всегда следует сходимость $A^{(k)} \to A$ при $k \to \infty$.

На практике часто приходится иметь дело с матричной геометрической прогрессией

$$E + A + A^2 + \ldots + A^k + \ldots$$

и встает вопрос о ее сходимости.

Лемма 2. Для того чтобы $A^k \to O$ при $k \to \infty$ необходимо и достаточно, чтобы все собственные значения матрицы A были по модулю меньше единицы.

Доказательство. Докажем лемму для случая симметричной матрицы A. Из линейной алгебры известно, что в этом случае

$$A = T^{-1} \Lambda T,$$

где Λ диагональная матрица с действительными собственными значениями

$$\lambda_1, \ \lambda_2, \ldots, \lambda_n$$

на главной диагонали. Соответственно,

$$A^k = T^{-1} \Lambda^k T,$$

где на главной диагонали диагональной матрицы Λ^k стоят элементы

$$\lambda_1^k, \ \lambda_2^k, \ldots, \lambda_n^k .$$

Таким образом, каждый элемент матрицы A^k является линейной комбинацией $\lambda_1^k, \ \lambda_2^k, \ldots, \lambda_n^k$ с коэффициентами не зависящими от k. Следовательно, если все собственные значения $\lambda_1, \ \lambda_2, \ldots, \lambda_n$ по модулю меньше единицы, то все элементы матрицы A^k стремятся к нулю при $k \to \infty$, т.е. $A^k \to O$.

Обратно,

$$\Lambda^k = T A^k T^{-1},$$

и, следовательно, все $\lambda_1^k, \ \lambda_2^k, \ldots, \lambda_n^k$ стремятся к нулю при $A^k \to O$. Последнее означает, что все числа $\lambda_1, \ \lambda_2, \ldots, \lambda_n$ по модулю меньше единицы.

Лемма 3. Для того чтобы ряд $E + A + A^2 + \ldots + A^k + \ldots$ сходился необходимо и достаточно, все собственные значения матрицы A были по абсолютной величине меньше единицы. В этом случае матрица $E-A$ имеет обратную и

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

Доказательство. 1) Пусть матричный ряд сходится. Это равносильно сходимости n^2 числовых рядов. Тогда в силу необходимого признака сходимости числового ряда каждый элемент матрицы A^k стремится к нулю

при $k \to \infty$, следовательно, $A^k \to O$ при $k \to \infty$. В силу леммы 2 последнее равносильно тому, что все собственные значения матрицы A по модулю меньше единицы.

2) Пусть все собственные значения матрицы A по модулю меньше единицы. Отметим сразу, что матрица $(E - A)$ невырождена, поскольку ее определитель $|E - A| = |A - E|$ не может обращаться в 0 (иначе среди собственных значений матрицы A было бы и число 1). Рассмотрим тождество

$$(E + A + A^2 + ... + A^k)\,(E - A) = E - A^{k+1},$$

откуда следует

$$(E + A + A^2 + ... + A^k) = (E - A)^{-1} - A^{k+1}(E - A)^{-1} .$$

Согласно лемме 2 $A^k \to O$ при $k \to \infty$. Следовательно,

$$(E + A + A^2 + ... + A^k) \to (E - A)^{-1} \text{ при } k \to \infty,$$

т.е. ряд сходится.

Покажем теперь, что норма $||A||$, индуцированная евклидовой нормой вектора, совпадает с $\sqrt{\lambda}$, где λ - наибольшее собственное значение матрицы A^*A, (A^* - матрица, полученная из A транспонированием).

Прежде всего убедимся, что $\lambda \geq 0$. Действительно, поскольку

$$(A^*A)^* = A^*(A^*)^* = A^*A,$$

то матрица A^*A симметрическая. Кроме того,

$$(A\bar{x}\,, A\bar{x}\,) = (\bar{x}\,, A^*Ax) \geq 0$$

для любого вектора $\vec{x}$. То есть A^*A - симметрическая неотрицательно определенная матрица. Как известно из курса линейной алгебры все собственные значения такой матрицы действительны и неотрицательны. Более того, существует ортонормированный базис из собственных векторов $\bar{x}^1, ..., \bar{x}^n$ данной матрицы, соответствующих ее собственным значениям

$\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_n$.

Рассмотрим произвольный вектор $\bar{x}$ единичной евклидовой нормы и разложим его по базису $\bar{x}^1, ..., \bar{x}^n$:

$$\bar{x} = \alpha_1 \bar{x}^1 + \ldots + \alpha_n \bar{x}^n .$$

Тогда

$$\|\bar{x}\|_2^2 = (\bar{x}\,, \bar{x}\,) = \alpha_1^{\,2} + ...+ \alpha_n^{\,2} = 1$$

и, следовательно,

$$||A\bar{x}\,||_2^{\,2} = (A\bar{x}\,, A\bar{x}\,) = (\bar{x}\,, A^*A\bar{x}\,) = (\sum_{i=1}^{n} a_i \bar{x}^i,\ \sum_{i=1}^{n} \lambda_i a_i \bar{x}^i) =$$

$$= \sum_{i=1}^{n} \alpha^2{}_i \lambda_i \ \leq\ \lambda_1 \sum_{i=1}^{n} \alpha^2{}_i = \lambda_1 .$$

Отсюда

$$||A|| \leq \sqrt{\mu(\mathrm{A^*A})} .$$

С другой стороны, для нормы $||A||$, индуцированной евклидовой векторной нормой, справедливо

$||A|| \geq ||A\,\bar{x}^{1}||_2$, где $||A\,\bar{x}^{1}||_2^2 = (A\,\bar{x}^{1}, A\,\bar{x}^{1}) = (\bar{x}^{1}, A^*A\,\bar{x}^{1}) = \lambda_1\,(\bar{x}^{1}, \bar{x}^{1}) = \lambda_1,$

откуда

$||A|| \geq \sqrt{\lambda_1} = \sqrt{\mu(A^*A)}$.

В итоге, действительно матричная норма, индуцированная евклидовой векторной нормой, имеет вид

$||A|| = \sqrt{\mu(A^*A)}$.

Такая норма матрицы A называется спектральной.

В частном случае, когда матрица A симметрическая, $A^*A = A^2$ и поскольку собственные значения этой матрицы совпадают с квадратами $\lambda_1^2,\ \lambda_2^2,\ldots\ldots,\ \lambda_n^2$ собственных значений матрицы A, то спектральная норма матрицы совпадает с наибольшим по абсолютной величине собственным значением матрицы A, то есть равна спектральному радиусу матрицы A :

$||A|| = \mu(A)$.

Спектральная норма матрицы неудобна в практическом плане из-за трудности ее вычисления. Поэтому наряду с нормами $||A||_1$ и $||A||_3$, индуцированными кубической и октаэдрической векторными нормами , мы будем пользоваться евклидовой нормой матрицы

$||A||_2 = \sqrt{Sp(A^*A)}$,

где SpA – след матрицы A, т.е. сумма ее элементов, стоящих на главной диагонали. Как легко проверить, эту норму можно записать также в виде

$$\|A\|_2 = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2}\,.$$

Нетрудно проверить, что для введенной нормы $||A||_2$ выполняются все аксиомы из определения матричной нормы. Кроме того, эта норма является *согласованной с* евклидовой *векторной нормой* в том смысле, что для нее и евклидовой векторной нормы всегда выполняется соотношение

$||A\,\bar{x}|| \leq ||A||\,||\bar{x}||$.

В дальнейшем мы ограничимся рассмотрением матричных норм, согласованные с соответствующими векторными нормами.

3.2. Метод простых итераций

Рассмотрим систему линейных уравнений в векторном виде:

$A\bar{x} = \bar{b}$ (3.1)

Преобразуем ее:

$\bar{x} + A\bar{x} = \bar{b} + \bar{x},$

$\bar{x} = (E - A) \cdot \bar{x} + \bar{b}.$

То есть систему (3.1) можно представить в виде

$\bar{x} = B\bar{x} + \bar{c}$. (3.2)

Существует много других способов представить систему (3.1) в виде (3.2). Так, если матрица A имеет ненулевые диагональные элементы, то i-е уравнение делят на элементы a_{ii}, чтобы получить единичный коэффициент перед x_i. Затем в i-м уравнении в левой части оставляют x_i, а все остальные слагаемые переносят в правую часть. Получаем систему $\bar{x} = B\bar{x} + \bar{c}$, где

$$B = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \dots & -\frac{a_{2n}}{a_{22}} \\ \dots & \dots & \dots & \dots \\ -\frac{a_{n1}}{a_{nn}} & \dots & \dots & 0 \end{pmatrix}, \qquad \bar{c} = \begin{pmatrix} \frac{b_1}{a_{11}} \\ \vdots \\ \frac{b_n}{a_{nn}} \end{pmatrix}.$$

Пусть теперь система записана в виде (3.2). Выбираем произвольным образом $\bar{x}^0$ − начальное приближение и подставляем в правую часть системы. Получим *1*-е приближение

$\bar{x}^1 = B\bar{x}^0 + \bar{c}$.

Повторяя процесс k раз, получаем k-е приближение

$$\bar{x}^k = B\bar{x}^{k-1} + \bar{c}, \qquad (3.3)$$

и так далее.

Таким образом, получается рекуррентная последовательность $\{\bar{x}^k\}$, которую будем также называть итерационной последовательностью. Если последовательность $\{\bar{x}^k\}$ сходится к некоторому вектору $\bar{x}^*$, то очевидно $\bar{x}^*$ является решением системы (3.2). Действительно, переход к пределу в равенстве (3.3) дает $\bar{x}^* = B\bar{x}^* + \bar{c}$.

Часто в качестве начального приближения берут вектор $\bar{c}$, хотя, вообще говоря, начальное приближение может выбираться произвольно. Процесс нахождения последовательности $\{\bar{x}^k\}$ будем называть методом простых итераций.

Выясним условия сходимости последовательности $\{\bar{x}^k\}$ в методе простых итераций к решению системы (3.2).

Теорема 1. Для того чтобы при любом начальном приближении $\bar{x}^0$ итерационная последовательность в методе простых итераций сходилась к

решению системы (3.2), необходимо и достаточно, чтобы все собственные значения матрицы B были по абсолютной величине меньше единицы.

Доказательство. 1) Пусть все собственные значения матрицы B по модулю меньше единицы. Тогда в силу лемм 2 и 3

$$B^k \to O \;\; и \;\; E + B + B^2 + \ldots + B^k \to (E - B)^{-1} \quad \text{при} \quad k \to \infty .$$

Но

$$\bar{x}^k = B\bar{x}^{k-1} + \bar{c} = B(\bar{x}^{k-2} + \bar{c}) + \bar{c} = \ldots .$$
$$= B^k \bar{x}^0 + (E + B + B^2 + \ldots . + B^{k-1})\bar{c},$$

откуда следует, что

$$\bar{x}^k \to (E - B)^{-1}\bar{c} = \bar{x}^* .$$

2) Пусть теперь при любом начальном приближении $\bar{x}^0$ существует предел

$$\bar{x}^k \to \bar{x}^* .$$

Тогда

$$\bar{x}^* = B\bar{x}^* + \bar{c} .$$

Вычитая из этого равенства равенство (3.3), получим

$$\bar{x}^* - \bar{x}^k = B(\bar{x}^* - \bar{x}^{k-1}) = \ldots . = B^k(\bar{x}^* - \bar{x}^0).$$

Перейдем к пределу в равенстве

$$\bar{x}^* - \bar{x}^k = B^k(\bar{x}^* - \bar{x}^0) .$$

Так как вектор $\bar{x}^* - \bar{x}^0$ может быть любым, а $\bar{x}^k \to \bar{x}^*$, то это означает, что $B^k \to O$, а последнее по лемме 2 равносильно тому, что все собственные значения матрицы B по модулю меньше единицы.

Доказанная теорема дает критерий сходимости метода простых итераций. Ее недостаток в трудности проверки полученного критерия, связанного с вычислением спектра матрицы. Поэтому чаще применяются достаточные условия сходимости, проверка которых требует знания только самих элементов матрицы. Некоторые из подобных достаточных условий вытекают из следующей теоремы.

Теорема 2. Пусть $\|B\| < 1$. Тогда система (3.2) имеет решение, причем единственное, и последовательность приближений $\{\bar{x}^k\}$ сходится к этому решению со скоростью геометрической прогрессии (при любом начальном приближении $\bar{x}^0$).

Доказательство. 1) Существование и единственность.

Предположим, что есть решение $\bar{x}^*$ для системы (3.2). Значит, $\bar{x}^* = B\bar{x}^* + \bar{c}$. Тогда $\|\bar{x}^*\| = \|B\bar{x}^* + \bar{c}\| \le \|B\|\|\bar{x}^*\| + \|\bar{c}\|$, откуда получаем оценку

$$\left\|\bar{x}^*\right\| \le \frac{\left\|\bar{c}\right\|}{1-\left\|B\right\|}. \qquad (3.4)$$

Такая оценка справедлива для системы (3.2) с любым вектором $\bar{c}$.Возьмем $\bar{c} = \bar{0}$, тогда система (3.2) является однородной, т. е. $\bar{x} = B\bar{x}$.

Она всегда имеет нулевое решение. Но в силу оценки (3.4) любое ее решение будет нулевым, т. е. однородная система имеет только нулевое решение. Как известно из линейной алгебры, в этом случае неоднородная система (3.2) имеет одно и только одно решение при любом $\bar{c}$.

2) Сходимость.

Пусть $\bar{x}^*$ − решение системы (3.2). Рассмотрим равенства:

$\bar{x}^k = B\bar{x}^{k-1} + \bar{c};$

$\bar{x}^* = B\bar{x}^* + \bar{c}$.

Из них следует:

$$\bar{x}^* - \bar{x}^k = B \cdot \left(\bar{x}^* - \bar{x}^k\right).$$

Обозначим $\bar{r}^k = \bar{x}^* - \bar{x}^k$. Тогда

$\bar{r}^k = B\bar{r}^{k-1} \qquad \forall k = 1,2,\ldots$.

Отсюда следует, что $\bar{r}^k = B^k\bar{r}^0$, где $\bar{r}^0 = \bar{x}^* - \bar{x}^0$, и значит, что

$\left\|\bar{x}^* - \bar{x}^k\right\| = \left\|B^k \cdot \left(\bar{x}^* - \bar{x}^k\right)\right\| \le \left\|B\right\|^k \cdot \left\|\bar{x}^* - \bar{x}^0\right\|$.

Положив $q = \left\|B\right\|$, и можно записать

$\left\|\bar{x}^* - \bar{x}^k\right\| = q^k \cdot \left\|\bar{x}^* - \bar{x}^0\right\|, \qquad k = 1,2,\ldots,$ следовательно, $\bar{x}^k \to \bar{x}^*$.

Теорема доказана.

Следствие 1. Метод простых итераций сходится, если выполнено хотя бы одно из условий:

1) $\max\limits_{1\le i\le n} \sum\limits_{j=1}^{n} \left|b_{ij}\right| < 1,$

2) $\sum\limits_{i,j=1}^{n} b_{ij}^2 < 1,$

3) $\max\limits_{1\le j\le n} \sum\limits_{i=1}^{n} \left|b_{ij}\right| < 1.$

Отметим, что условия 1 − 3 независимы друг от друга. Покажем это.

Пример 1. Рассмотрим матрицу:

$$B = \begin{bmatrix} \frac{3}{5} & -\frac{3}{5} \\ \frac{2}{5} & \frac{1}{5} \end{bmatrix}.$$

Очевидно, $\left\|B\right\|_2 = \sqrt{\left(\frac{9}{25}+\frac{4}{25}+\frac{9}{25}+\frac{1}{25}\right)} < 1$, т. е. метод итераций с матрицей B будет сходиться. С другой стороны,

$\left\|B\right\|_1 = \max\left\{\frac{6}{5};\frac{3}{5}\right\} = \frac{6}{5} > 1$.

Пример 2. Пусть

$$B = \begin{bmatrix} \frac{1}{10} & \frac{4}{5} \\ \frac{3}{5} & \frac{1}{5} \end{bmatrix}.$$

В данном случае

$\|B\|_1 = \max\left\{\frac{9}{10}; \frac{4}{5}\right\} = \frac{9}{10} < 1$, а $\|B\|_2 > 1$.

Итерационная последовательность также будет сходиться.

Оценим погрешность метода на каждой итерации:

$\bar{x}^k = B\bar{x}^{k-1} + \bar{c}$.

Поскольку $\bar{x}^* = B\bar{x}^* + \bar{c}$,

можно записать $\bar{x}^* - \bar{x}^k = B \cdot \left(\bar{x}^* - \bar{x}^{k-1}\right)$

и, следовательно,

$$\bar{x}^* = \bar{x}^k + B \cdot \left(\bar{x}^* - \bar{x}^{k-1}\right).$$

Вычтем из каждой части равенства $\bar{x}^{k-1}$, получим

$$\bar{x}^* - \bar{x}^{k-1} = \left(\bar{x}^k - \bar{x}^{k-1}\right) + B \cdot \left(\bar{x}^* - \bar{x}^{k-1}\right).$$

Отсюда

$$\left\|\bar{x}^* - \bar{x}^{k-1}\right\| \le \left\|\bar{x}^k - \bar{x}^{k-1}\right\| + \left\|B\right\| \cdot \left\|\bar{x}^* - \bar{x}^{k-1}\right\|,$$

и далее

$$\left\|\bar{x}^* - \bar{x}^{k-1}\right\| \le \frac{\left\|\bar{x}^k - \bar{x}^{k-1}\right\|}{1 - \left\|B\right\|}.$$

Так как, $\bar{x}^* - \bar{x}^k = B \cdot \left(\bar{x}^* - \bar{x}^{k-1}\right)$ то, умножая обе части предыдущего неравенства на $\|B\|$ и учитывая оценку

$\left\|\bar{x}^* - \bar{x}^k\right\| \le \|B\| \cdot \left\|\bar{x}^* - \bar{x}^{k-1}\right\|$, получим

$$\left\|\bar{x}^* - \bar{x}^k\right\| \le \frac{\|B\|}{1 - \|B\|} \cdot \left\|\bar{x}^k - \bar{x}^{k-1}\right\|.$$

Таким образом, справедливо следующее.

Следствие 2. Погрешность *k*-го приближения в методе простых итераций оценивается неравенством

$$\left\|\bar{x}^* - \bar{x}^k\right\| \le \frac{\|B\|}{1 - \|B\|} \cdot \left\|\bar{x}^k - \bar{x}^{k-1}\right\|.$$

3.3. Метод Зейделя

Рассмотрим модификацию метода итераций, называемую *методом Зейделя.* Пусть дана система линейных уравнений

$$A\bar{x} = \bar{b},$$

в которой $A_{n\times n}$ – матрица с диагональными элементами $a_{ii} \neq 0, \quad \forall i = \overline{1,n}$.

Каким-либо способом приведем систему к виду (3.2): $\overline{x} = \overline{B}\overline{x} + \overline{c}$. Перепишем систему покоординатно:

$$x_i = \sum_{j=1}^{n} b_{ij} x_j + c_i, \quad i = 1,...,n.$$

Если бы применялся метод простых итераций, итерационная последовательность выглядела бы следующим образом:

$$x_i^k = \sum_{j=1}^{n} b_{ij} x_j^{k-1} + c_i, \qquad (i = 1,2,...,n; \quad k = 1,2,....).$$

При этом алгоритм позволял бы вычислять координаты x_i^k независимо, в любом порядке. Идея метода Зейделя заключается в том, чтобы использовать уже найденные координаты для улучшения значения последующих и проводить вычисления по правилу

$$x_i^k = \sum_{j=1}^{i-1} b_{ij} x_j^k + \sum_{j=i}^{n} b_{ij} x_j^{k-1} + c_i, \qquad i = 1,...,n. \tag{3.5}$$

Метод вычисления решения на основе итерационной последовательности (3.5) называют методом Зейделя.

Можно ожидать, что метод Зейделя будет сходиться быстрее метода простых итераций. Исследуем условия его сходимости. Матрицу *B* разобьем на сумму матриц *H* и *F,* где

$$H = \begin{pmatrix} 0 & 0 & ...0 & 0 \\ b_{21} & 0 & ...0 & 0 \\ b_{31} & b_{32} & ...0 & 0 \\ ... & ... & ... & ... \\ b_{n1} & b_{n2}... & b_{nn-1} & 0 \end{pmatrix}, \qquad F = \begin{pmatrix} b_{11} & b_{12} & ...b_{1n-1} & b_{1n} \\ 0 & b_{22} & ...b_{2n-1} & b_{2n} \\ 0 & 0 & ...b_{3n-1} & b_{3n} \\ ... & ... & ... & ... \\ 0 & 0... & 0 & b_{nn} \end{pmatrix}.$$

Тогда алгоритм метода Зейделя можно переписать в виде

$$\overline{x}^k = H\overline{x}^k + F\overline{x}^{k-1} + \overline{c}, \qquad k = 1,2,....$$

Или

$$(E - H)\overline{x}^k = F\overline{x}^{k-1} + \overline{c}, \qquad k = 1,2,....$$

Поскольку матрица $(E - H)$ не вырождена, то последнее выражение можно записать как

$$\overline{x}^k = (E - H)^{-1} F\overline{x}^{k-1} + (E - H)^{-1}\overline{c}, \qquad k = 1,2,.... \tag{3.6}$$

Таким образом, метод Зейделя эквивалентен методу простых итераций для системы линейных уравнений

$$\bar{x} = (E-H)^{-1}F\bar{x} + (E-H)^{-1}\bar{c},$$

которая, в свою очередь, равносильна исходной системе (3.2). Использование итерационной последовательности (3.6) более трудоемко по сравнению с классической последовательностью (3.5). Однако представление метода Зейделя в форме (3.6) дает возможность выяснить условия сходимости этого метода.

Теорема 3. Для того чтобы метод Зейделя сходился при любом начальном приближении $\bar{x}^0$, необходимо и достаточно, чтобы все корни уравнения

$$| F + \lambda H - \lambda E | = 0$$

были по абсолютной величине меньше единицы.

Доказательство. В силу теоремы 1 для сходимости последовательности (3.6), а значит и метода Зейделя необходимо и достаточно, чтобы все собственные значения матрицы $(E-H)^{-1}F$, то есть корни уравнения

$$| (E-H)^{-1}F - \lambda E | = 0$$

были по модулю меньше единицы. Даже вычисление коэффициентов этого уравнения представляет собой трудную задачу. Поэтому найдем более простое уравнение, корни которого совпадают с корнями данного уравнения. Действительно, поскольку определитель $| E - H |$ равен единице, то

$$| (E-H)^{-1}F - \lambda E | = | (E-H)^{-1}(E-H)[(E-H)^{-1}F - \lambda E] | =$$
$$= | (E-H)^{-1} || F - (E-H)\lambda E | = | F + \lambda H - \lambda E |.$$

И так сходимость метода Зейделя сводится к определению абсолютной величины корней уравнения $| F + \lambda H - \lambda E | = 0$.

Уже непосредственное сравнение этого уравнения с характеристическим уравнением $|B - \lambda E| = 0$ матрицы показывает, что области сходимости метода Зейделя и метода простых итераций, вообще говоря, должны быть различны. Действительно, для случая матрицы

$$B = \begin{pmatrix} 2,5 & 3 \\ 2 & -2,5 \end{pmatrix}$$

уравнение

$$|B - \lambda E| = \lambda^2 - 0,25 = 0$$

имеет корни $\lambda_1 = -0,5$, $\lambda_2 = 0,5$, следовательно, метод простых итераций сходится. Метод Зейделя для той же матрицы B сходиться не будет, так как у уравнения

$$| F + \lambda H - \lambda E | = \lambda^2 + 6\lambda - 6,25 = 0$$

один из корней по модулю больше единицы.

Обратно в случае матрицы

$$B = \begin{pmatrix} 4,2 & -2 \\ 2 & -0,1 \end{pmatrix}$$

сходится метод Зейделя ($\lambda_1 = -0,6,\ \lambda_2 = 0,7$), а метод простых итераций расходится.

Теорема 3 неудобна для практического применения. По аналогии с методом простых итераций можно построить достаточные условия сходимости метода Зейделя. В частности, из теоремы 2 и представления итерационного процесса метода Зейделя в форме (3.6) следует, что для сходимости метода Зейделя достаточно, чтобы $\|(E-H)^{-1}F\|$ была меньше единицы. Однако проверка данного условия тоже достаточно затруднительна.

Получим более простые достаточные условия сходимости метода Зейделя, формулируемые непосредственно через элементы матрицы B.

Лемма 4. Если диагональные элементы матрицы C доминируют по строкам или по столбцам, т.е. если

$$\sum_{j=1,\, j\neq i}^{n} |c_{ij}| < |c_{ii}| \quad (i = 1,\ldots,n)$$

или

$$\sum_{i=1,\, i\neq j}^{n} |c_{ij}| < |c_{jj}| \quad (j = 1,\ldots,n),$$

то определитель матрицы C отличен от нуля.

Доказательство. Докажем лемму для случая доминирования диагональных элементов пол строкам (случай доминирования по столбцам рассматривается совершенно аналогично). Для доказательства утверждения достаточно показать, что однородная линейная система

$$C\bar{x} = \bar{0}$$

имеет только нулевое решение. Предположим противное, т.е. допустим, что система имеет отличное от нулевого решение $\bar{x}^*$. Среди координат вектора $\bar{x}^*$ выберем максимальную по модулю x_i^*. Положим $\bar{x} = \bar{x}^*$ в системе $C\bar{x} = \bar{0}$ и рассмотрим значение левой части i-го уравнения однородной системы:

$$|c_{i1}x^*_1 + c_{i2}x^*_2 + \ldots + c_{ii}x^*_i + \ldots + c_{i1}x^*_1 + \ldots c_{in}x^*_n| \geq$$

$$\geq |c_{ii}||x^*_i| - \sum_{j=1,\, j\neq i}^{n} |c_{ij}||x^*_j| \geq |x_i^*|(|c_{ii}| - \sum_{j=1,\, j\neq i}^{n} |c_{ij}|) > 0,$$

так как $|x^*_i| > 0$ и $\sum_{j=1,\, j\neq i}^{n} |c_{ij}| < |c_{ii}|$.

Полученное противоречие доказывает справедливость утверждения леммы.

Теорема 4. Для того чтобы метод Зейделя сходился, достаточно выполнения одного из следующих условий:

1) $\max\limits_{1\le i\le n}\sum\limits_{j=1}^{n}\left|b_{ij}\right|<1$,

2) $\max\limits_{1\le j\le n}\sum\limits_{i=1}^{n}\left|b_{ij}\right|<1$.

Доказательство. Пусть выполнено первое условие. В силу теоремы 3 достаточно показать, что в этом случае любое число λ^* такое, что $|\lambda^*|\ge 1$, не может быть корнем уравнения $|F+\lambda H-\lambda E|=0$. Действительно, рассматривая сумму абсолютных величин недиагональных элементов любой строки определителя

$$|F+\lambda H-\lambda E|,$$

можно записать:

$$|\lambda^*||b_{i1}|+\ldots.+|\lambda^*||b_{ii-1}|+|b_{ij+1}|+\ldots.+|b_{in}|\le$$

$$\le |\lambda^*|\sum_{i=1,j\ne i}^{n}\left|b_{ij}\right|=|\lambda^*|(\sum_{i=1}^{n}\left|b_{ij}\right|-|b_{ii}|)<|\lambda^*|(1-|b_{ii}|)=$$

$$=|\lambda^*|-|\lambda^*||b_{ii}|\le|\lambda^*|-|b_{ii}|\le|\lambda^*-b_{ii}|=|b_{ii}-\lambda^*|.$$

Полученные неравенства

$$|\lambda^*||b_{i1}|+\ldots.+|\lambda^*||b_{ii-1}|+|b_{ij+1}|+\ldots.+|b_{in}|<|b_{ii}-\lambda^*| \qquad (i=1,\ldots,n)$$

представляют собой как раз условие доминирования диагональных элементов матрицы $F+\lambda^*H-\lambda^*E$.

Тогда по лемме 4 определитель $|F+\lambda^*H-\lambda^*E|$ отличен от нуля и, следовательно, все корни уравнения

$$|F+\lambda^*H-\lambda^*E|=0$$

по модулю меньше единицы. Тогда в силу теоремы 3 метод Зейделя сходится.

Аналогично рассматривается случай доминирования диагональных элементов по столбцам.

Заметим, что, как и в случае метода простых итераций, при выполнении условий теоремы 4 можно получить гарантированные оценки погрешности метода Зейделя. Вообще говоря, эти оценки лучше, чем для метода простых итераций. Однако метод Зейделя не всегда оказывается лучше метода простых итераций. Он даже может расходиться при наличии сходимости метода простых итераций. Области сходимости обоих методов различны,

причем очень многое зависит от способа приведения исходной системы (3.1) к виду (3.2).

Рассмотрим систему (3.1) при условии строгого доминирования диагональных элементов матрицы A. Тогда можно разделить первое уравнение на a_{11}, второе на a_{22} и т. д. и выразить соответственно неизвестные $x_1, x_{2,} \ldots$. Получим систему (3.2), в которой

$$B = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \ldots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \ldots & -\frac{a_{2n}}{a_{22}} \\ \ldots & \ldots & \ldots & \ldots \\ -\frac{a_{n1}}{a_{nn}} & \ldots & \ldots & 0 \end{pmatrix}, \qquad \overline{c} = \begin{pmatrix} \frac{b_1}{a_{11}} \\ \vdots \\ \frac{b_n}{a_{nn}} \end{pmatrix}.$$

В этом случае итерационная последовательность Зейделя имеет вид

$$\begin{cases} x_1^k = b_{11}x_2^{k-1} + \ldots + b_{1n}x_n^{k-1} + c_1 \\ x_2^k = b_{21}x_1^k + b_{23}x_3^{k-1} + \ldots + b_{2n}x_n^{k-1} + c_2 \\ \ldots \\ x_n^k = b_{n1}x_1^k + \ldots + b_{nn-1}x_{n-1}^k + c_n. \end{cases} \qquad (3.7)$$

При этом из условия строгого доминирования диагональных элементов матрицы A по строкам (столбцам) следует выполнение условий теоремы 4. Таким образом, справедлива следующая теорема.

Теорема 5. Если для матрицы A в системе (3.1) выполнено условие строгого доминирования диагональных элементов по строкам или столбцам, то метод Зейделя в форме (3.7) сходится.

4. ПРОБЛЕМА СОБСТВЕННЫХ ЗНАЧЕНИЙ

Вычисление собственных значений и векторов матриц имеет исключительно важное значение для решения широкого круга задач. При этом часто требуется получение всех собственных значений и отвечающих им собственных векторов. Такую задачу принято называть *полной проблемой собственных значений.* В других случаях требуется знание лишь максимальных или минимальных по абсолютной величине собственных значений. Иногда требуется найти два самых больших по абсолютной величине собственных значений или собственные значения, ближайшие к некоторому заданному числу. Такие задачи называют *частичными проблемами собственных значений*.

Проблема собственных значений достаточно проста в теоретическом плане. Из линейной алгебры известно, что собственные значения матрицы A являются корнями λ_1 , λ_2 ,…, λ_n характеристического уравнения

$$| A - \lambda E| = \begin{vmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{vmatrix} = 0 . \qquad (4.1)$$

Данный определитель является многочленом $(-1)^n P_n(\lambda)$ степени n от λ, где $P_n(\lambda) = \lambda^n + p_1\lambda^{n-1} + \dots + p_n$.

Собственные векторы $\bar{x}$, отвечающие собственному значению λ , представляют собой ненулевые решения системы

$$A\bar{x} = \lambda\bar{x} . \qquad (4.2)$$

Таким образом, раскрывая определитель $|A - \lambda E|$ и решая характеристическое уравнение (1), можно найти все собственные значения. Подставляя их последовательно в систему (4.2), находим собственные векторы, отвечающие данным собственным значениям.

Следует отметить, что при раскрытии определителя $|A - \lambda E|$ возникают значительные вычислительные трудности. Поэтому применяются различные методы, позволяющие с помощью конечного числа преобразований привести матрицу А к матрице более простого вида, для которой коэффициенты характеристического уравнения легко вычисляются. При этом, как правило, получаются достаточно простые соотношения и для нахождения собственных векторов.

Применяемые методы решения проблемы собственных значений делятся на *прямые и итерационные*. Прямые методы позволяют найти коэффициенты характеристического уравнения и затем вычислить корни этого уравнения. Прямые методы отличаются простотой и высоким быстродействием. В то же время их недостатком является чувствительность к ошибкам округления результатов промежуточных вычислений в случае высокой размерности матриц. В итерационных

методах коэффициенты характеристического уравнения не вычисляются, но строятся итерационные последовательности для нахождения собственных значений. Итерационные методы более трудоемки, однако менее чувствительны к ошибкам округлений и более надежны.

4.1. Решение частичной проблемы собственных значений

Рассмотрим для простоты случай, когда все собственные значения матрицы A действительны (это заведомо будет, в частности, если матрица A симметрическая). Найдем максимальное по абсолютной величине собственное значение. Для простоты будем предполагать, что

$$|\lambda_1| > |\lambda_2| \geq \ldots\ldots \geq |\lambda_n|$$

и, что существует базис из собственных векторов $\bar{u}^1,....,\bar{u}^n$.

Выберем произвольный вектор $\bar{x}^0$ такой, что

$$\bar{x}^0 = c_1\bar{u}^1 + ... + c_n\bar{u}^n,$$

и построим последовательность $\bar{x}^{k+1} = A\bar{x}^k, \quad k = 0,1,...$.

Тогда

$$\bar{x}^k = \sum_{i=1}^{n} c_i A^k \bar{u}^i = \sum_{i=1}^{n} c_i \lambda_i^k \bar{u}^i .$$

Отсюда следует, что

$$\bar{x}^k = c_1\lambda_1^k\bar{u}^1 + O(\lambda_2^k),$$

$$(\bar{x}^k, \bar{x}^k) = (c_1\lambda_1^k\bar{u}^1 + O(\lambda_2^k), c_1\lambda_1^k\bar{u}^1 + O(\lambda_2^k)) = c_1^2\lambda_1^{2k} + O(\lambda_1^k\lambda_2^k),$$

$$(\bar{x}^{k+1}, \bar{x}^k) = (c_1\lambda_1^{k+1}\bar{u}^1 + O(\lambda_2^{k+1}), c_1\lambda_1^k\bar{u}^1 + O(\lambda_2^k)) = \lambda_1 c_1^2\lambda_1^{2k} + O(\lambda_1^k\lambda_2^k).$$

Положим

$$\lambda_1^{(k)} = (\bar{x}^{k+1}, \bar{x}^k)/(\bar{x}^k, \bar{x}^k) .$$

Тогда из последних соотношений получаем (при условии, что c_1 отлично от нуля):

$$\lambda_1^{(k)} = \frac{\lambda_1 c_1^2\lambda_1^{2k} + O(\lambda_1^k\lambda_2^k)}{c_1^2\lambda_1^{2k} + O(\lambda_1^k\lambda_2^k)} = \lambda_1 + O(\frac{\lambda_2^k}{\lambda_1^k}) . \qquad (4.3)$$

Из (4.3) непосредственно следует, что $\lambda_1^{(k)} \to \lambda_1$ при $k \to \infty$. Кроме того,

$$\frac{\bar{x}^k}{\|\bar{x}^k\|} = \frac{c_1\lambda_1^k\bar{u}_1 + c_2\lambda_2^k\bar{u}_2 + ... + c_n\lambda_n^k\bar{u}_n}{|c_1||\lambda_1|^k + O(|\lambda_2|^k)} = \frac{c_1\lambda_1^k\bar{u}_1}{|c_1||\lambda_1|^k} + O)(\frac{|\lambda_2|^k}{|\lambda_1|^k}),$$

то есть,

$$\frac{\bar{x}^k}{\|\bar{x}^k\|}\operatorname{sgn}(c_1\lambda_1^k) \to \bar{u}_1 \text{ при } k \to \infty .$$

Отметим, что требование, чтобы c_1 было отлично от нуля, не является жестким в виду произвольного выбора начального приближения $\bar{x}^0$.

Кроме того, если даже этого и не было вначале, то случайная ошибка сделает слагаемое, содержащее собственный вектор $\bar{u}_1$, ненулевым позже и, в конце концов, оно станет доминирующим. Мысль, что, зная наибольшее собственное значение и соответствующий собственный вектор, мы можем вычитать его на каждом шаге, и, тем самым, дать возможность проявиться второму по модулю собственному значению, очевидна. Это действительно можно сделать, но отнюдь не в точности так, как хотелось бы. На самом деле, можно найти несколько наибольших по модулю собственных значений, затем вычислительный процесс постепенно превратится в шум из-за нарастания погрешности, так что каждое следующее собственное значение будет определяться все с меньшей точностью.

Чтобы тем же методом найти наименьшее (в алгебраическом смысле) собственное значение, достаточно следующего простого наблюдения. Пусть $\bar{x}$ — собственный вектор, т. е. $A\bar{x} = \lambda\bar{x}$. Тогда

$(A - pE)\bar{x} = (\lambda - p)\bar{x}$.

Если уже известна примерная величина наибольшего собственного значения, то можно взять p равным этой величине, и самое маленькое собственное значение станет самым большим (по модулю).

4.2. Метод Данилевского

Метод Данилевского относится к прямым методам и является достаточно простым и экономичным. Известно, что матрицы $S^{-1}AS$, полученные преобразованием подобия из с A, имеют тот же характеристический многочлен, что и A. Известно так же, что любая матрица приводима преобразованием подобия к так называемой канонической форме Фробениуса

$$F=\begin{pmatrix} -p_1 & -p_2 & \ldots & -p_n \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & & & & \\ 0 & \ldots & 0 & 1 & 0 \end{pmatrix},$$

в первой строке которой стоят коэффициенты характеристического многочлена, взятые с обратным знаком. Таким образом, основная задача сводится к нахождению матрицы S такой, что $F= S^{-1}AS$.

Предположим, что элемент a_{nn-1} матрицы A отличен от нуля. Разделим $(n-1)$-й столбец этой матрицы на a_{nn-1} и вычтем его из i–го столбца, умноженного на a_{ni} (для всех $i=1,2,\ldots,n$). Тогда последняя строка примет такой же вид как в матрице F. Непосредственно

проверяется, что проделанная операция равносильна умножению A справа на матрицу

$$M_{n-1}=\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & & & & \dots \\ -\frac{a_{n1}}{a_{nn-1}} & \dots & -\frac{a_{nn-2}}{a_{nn-1}} & \frac{1}{a_{nn-1}} & -\frac{a_{nn}}{a_{nn-1}} \\ 0 & \dots & 0 & 0 & 1 \end{pmatrix}.$$

Непосредственно проверяется также, что M_{n-1} не вырождена и, следовательно, существует

$$M^{-1}{}_{n-1}=\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & & & \dots \\ a_{n1} & \dots & & & a_{nn} \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix}.$$

Очевидно, что умножение AM_{n-1} слева на матрицу $M^{-1}{}_{n-1}$ не меняет последней строки матрицы AM_{n-1}. Таким образом,

$$M^{-1}{}_{n-1}AM_{n-1}=A^{(1)}=\begin{pmatrix} a_{11}^{(1)} & \dots & a_{1n}^{(1)} \\ a_{21}^{(1)} & \dots & a_{2n}^{(1)} \\ \dots & & \dots \\ a_{n-11}^{(1)} & \dots & a_{n-1n-1}^{(1)} & a_{n-1n}^{(1)} \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix}.$$

Заметим, что матрицы M_{n-1} *и* $M^{-1}{}_{n-1}$, умножением на которые мы переходим от матрицы A к матрице $A^{(1)}$, выписываются непосредственно по виду матрицы A. Предположим далее, что элемент $a^{(1)}{}_{n-1n-2}$ тоже отличен от нуля. Делаем второй шаг, полностью аналогичный предыдущему, и приводим вторую снизу строку матрицы к виду необходимому для формы Фробениуса (сохраняя последнюю строку без изменений). Получаем

$$M^{-1}{}_{n-2}\,M^{-1}{}_{n-1}AM_{n-1}M_{n-2}=M^{-1}{}_{n-2}\,A^{(1)}\,M_{n-2}=$$

$$=A^{(2)}=\begin{pmatrix} a_{11}^{(2)} & \dots & a_{1n}^{(2)} \\ a_{21}^{(2)} & \dots & a_{2n}^{(2)} \\ \dots & & \dots \\ a_{n-21}^{(2)} & \dots & a_{n-2n-1}^{(2)} & a_{n-2n}^{(2)} \\ 0 & \dots & 1 & 0 & 0 \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix},$$

где

$$M_{n-2}=\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & & & & \\ -\frac{a^{(1)}_{n-11}}{a^{(1)}_{n-1n-2}} & \dots & -\frac{a^{(1)}_{n-1n-3}}{a^{(1)}_{n-1n-2}} & \frac{1}{a^{(1)}_{n-1n-2}} & -\frac{a^{(1)}_{n-1n-3}}{a^{(1)}_{n-1n-2}} & -\frac{a^{(1)}_{n-1n}}{a^{(1)}_{n-1n-2}} \\ 0 & \dots & 0 & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$M^{-1}_{n-2}=\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & & & & \\ a^{(1)}_{n-11} & \dots & a^{(1)}_{n-1n-1} & a^{(1)}_{n-1n} \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 \end{pmatrix}.$$

Правило построения матриц M_{n-2} и M^{-1}_{n-2} по виду матрицы $A^{(1)}$, как видим, полностью сохраняется. Оно сохраняется и на следующих шагах метода. Таким образом, если имеет место так называемый регулярный случай, когда

$$a_{nn-1} \neq 0,\ a^{(1)}_{n-1n-2} \neq 0,\ a^{(2)}_{n-2n-3} \neq 0,\dots..,\ a^{(n-2)}_{21} \neq 0,$$

то после $(n-1)$ шагов метода Данилевского получим следующий результат

$$A^{(n-1)}= M^{-1}_{1}\dots\ M^{-1}_{n-1}AM_{n-1}\dots M_1=$$

$$=\begin{pmatrix} a^{(n-1)}_{11} & a^{(n-1)}_{12} & \dots & a^{(n-1)}_{1n} \\ 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & & & & \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix}=\begin{pmatrix} -p_1 & -p_2 & \dots & -p_n \\ 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & & & & \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix}=F=S^{-1}\,AS.$$

В таком случае мы можем непосредственно выписать характеристическое уравнение

$$\lambda^n + p_1\lambda^{n-1} + \dots.. + p_n = 0$$

и, решая его, найти собственные значения $\lambda_1, \lambda_2, \dots\dots, \lambda_n$.

Для нахождения собственных векторов в регулярном случае нет необходимости решать систему (4.2). Как уже говорилось выше, матрицы F и A имеют одни и те же собственные значения. Собственные векторы, отвечающие одному и тому же собственному значению, вообще говоря, будут разными. Однако они связаны между собой преобразованием подобия. Так, если $\bar{y}$ собственный вектор матрицы F, отвечающий собственному значению λ, то вектор $S\bar{y}$ будет

собственным вектором матрицы A, отвечающим тому же собственному значению.

Действительно, поскольку $F\bar{y} = \lambda\ \bar{y}$ и $F = S^{-1}AS$, то $S^{-1}AS\bar{y} = \lambda\ \bar{y}$. Умножая это равенство слева на матрицу S, получим $AS\bar{y} = \lambda\ S\bar{y}$. Последнее означает, что $S\bar{y}$ будет собственным вектором матрицы A. Таким образом, собственные векторы матрицы A находятся пересчетом собственных векторов матрицы Фробениуса. Собственные же векторы $\bar{y}$ матрицы Фробениуса определяются из системы

$$\begin{pmatrix} -p_1 & -p_2 & \ldots\ldots\ldots\ldots & -p_n \\ 1 & 0 & \ldots\ldots\ldots\ldots & 0 \\ 0 & 1 & 0\ldots\ldots\ldots\ldots & 0 \\ 0 & 0 & 1\ldots\ldots\ldots\ldots & 0 \\ & & \ldots\ldots\ldots\ldots & \\ 0 & \ldots\ldots\ldots & 0 \quad 1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ \\ y_n \end{pmatrix} = \lambda \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ \\ y_n \end{pmatrix}.$$

Покоординатная запись этой системы имеет вид

$$\begin{aligned} -p_1y_1 - p_2y_2 - \ldots\ldots\ldots - p_ny_n &= \lambda\ y_1, \\ y_1 &= \lambda\ y_2, \\ y_2 &= \lambda\ y_3, \\ &\ldots\ldots\ldots\ldots \\ y_{n-1} &= \lambda\ y_n. \end{aligned}$$

Поскольку собственный вектор определяется с точностью до постоянного множителя, можно принять $y_n = 1$ и вычислить остальные координаты собственного вектора:

$$y_n = 1,\quad y_{n-1} = \lambda\ , \ldots\ldots\ldots\ldots, y_1 = \lambda^{\,n-1}.$$

Равенство же

$$-p_1y_1 - p_2y_2 - \ldots\ldots\ldots - p_ny_n = \lambda\ y_1$$

принимает при этом тривиальный вид

$$\lambda^{\,n} + p_1\lambda^{\,n-1} + \ldots\ldots + p_n = 0$$

и используется для контроля вычислений.

Зная матрицу S, не трудно теперь найти собственные векторы матрицы A.

Отдельно рассмотрим нерегулярный случай метода Данилевского. Пусть выполнено $(n\text{-}k)$ шагов метода и оказалось, что в матрице $A^{(n-k)}$ элемент $a_{kk-1}^{(n-k)} = 0$. Тогда, если левее этого элемента в строке есть отличные от нуля элементы (например в столбце с номером j), то поменяем местами j-й и $(k\text{-}1)$-й столбцы и продолжим процесс. Заметим, что операция замены столбцов местами равносильна умножению матрицы $A^{(n-k)}$ слева и справа на матрицу T, которая

строиться из единичной матрицы E заменой четырех ее элементов. Именно:

$$t_{jj} = t_{k-1k-1} = 0, \quad t_{jk-1} = t_{k-1j} = 1,$$

остальные элементы матрицы T совпадают с соответствующими элементами матрицы E. Таким образом, в цепочке преобразований матрицы на данном шаге добавится дополнительная операция

$$T\, A^{(n-k)} T,$$

после которой процесс пойдет, как и раньше. При этом важно, что дополнительное преобразование $TA^{(n-k)}T$ является преобразованием подобия. Действительно, поскольку двойная перестановка столбцов дает исходную матрицу, то $TT = T^2 = E$, *т.е.* $T^{-1} = T$.

Если левее элемента $a_{kk-1}^{(n-k)} = 0$ в строке матрицы $A^{(n-k)}$ не оказалось ненулевых элементов, то матрица $A^{(n-k)}$ очевидно имеет вид

$$A^{(n-k)} = \begin{pmatrix} B^{(n-k)} & C^{(n-k)} \\ 0 & F^{(n-k)} \end{pmatrix},$$

где $$B^{(n-k)} = \begin{pmatrix} a_{11}^{(n-k)} & \ldots & a_{1k-1}^{(n-k)} \\ a_{21}^{(n-k)} & \ldots & a_{2k-1}^{(n-k)} \\ \ldots & \ldots & \ldots \\ a_{k-11}^{(n-k)} & \ldots & a_{k-1k-1}^{(n-k)} \end{pmatrix},$$

$$F^{(n-k)} = \begin{pmatrix} a_{kk}^{(n-k)} & \ldots & a_{kn}^{(n-k)} \\ 1 \quad 0 & \ldots & 0 \quad 0 \\ \ldots & \ldots & \ldots \\ 0 \quad 0 & \ldots & 1 \quad 0 \end{pmatrix}.$$

Тогда

$$| A^{(n-k)} - \lambda E | = | B^{(n-k)} - \lambda E_{k-1} | \; | F^{(n-k)} - \lambda E_{n-k+1} |$$

и, следовательно, поскольку $F^{(n-k)}$ является матрицей Фробениуса, ее характеристический многочлен можно выписать непосредственно, а к матрице $B^{(n-k)}$ применить снова метод Данилевского. Таким образом, вычислительный процесс даже упрощается.

Подсчетом необходимых арифметических операций можно убедиться, что метод Данилевского является одним из самых экономичных методов решения полной проблемы собственных значений. Однако этот метод очень чувствителен к ошибкам в результатах промежуточных вычислений.

4.3. Метод вращений Якоби для симметрических матриц

Итерационный метод Якоби был предложен еще в середине 19-го века, однако долгое время не находил применения из-за слишком большого по тем временам объема вычислений. В настоящее время известно большое количество его модификаций, основная идея которых однако остается прежней. Из линейной алгебры известно, что всякая симметрическая матрица A может быть приведена к диагональному виду ортогональным преобразованием подобия

$$V^{-1} A V = \Lambda,$$

где Λ – диагональная матрица. При этом для ортогональной матрицы V справедливо условие $V^{-1}=V^{*}$, т.е. ортогональное преобразование подобия можно записать в виде

$$V^{*} A V = \Lambda. \qquad (4.4)$$

Последнее условие дает фактически матричное уравнение, которое можно использовать для вычисления элементов матриц V и Λ. Однако метод Якоби использует итерационный процесс, который приводит исходную симметрическую матрицу A к диагональному виду с помощью последовательности *элементарных ортогональных преобразований* (в дальнейшем называемых *вращениями Якоби* или *плоскими вращениями).* Процедура построена таким образом, что на $(k+1)$-ом шаге осуществляется преобразование вида

$$A^{(k)} \rightarrow A^{(k+1)} = V^{(k)*} A^{(k)} V^{(k)} = V^{(k)*} \ldots V^{(0)*} A^{(0)} V^{(0)} \ldots V^{(k)}, \qquad k=0,1,2\ldots, \qquad (4.5)$$

где $A^{(0)}= A$, $V^{(k)} = V^{(k)}_{ij}(\varphi)$ — ортогональная матрица, отличающаяся от единичной матрицы только элементами

$$v_{ii}= v_{jj} = \cos\varphi \quad v_{ij} = - v_{ji} = -\sin\varphi \quad , \qquad (4.6)$$

значение φ выбирается при этом таким образом, чтобы обратить в 0 наибольший по модулю недиагональный элемент матрицы $A^{(k)}$. Итерационный процесс постепенно приводит к матрице со значениями недиагональных элементов, которыми можно пренебречь, т.е. матрица $A^{(k)}$ все более похожа на диагональную, а диагональная матрица Λ является пределом последовательности $A^{(k)}$ при $k \rightarrow \infty$.

Основное достоинство метода Якоби заключается в том, что при выполнении каждого плоского вращения уменьшается сумма квадратов недиагональных элементов; сходимость этой суммы к нулю по мере увеличения числа шагов гарантирует сходимость процесса диагонализации.

Отметим, что, если разложение (4.4) найдено, то легко указать правило нахождения собственных векторов. Действительно, если λ_i - i-й диагональный элемент матрицы Λ, тогда, как известно из линейной алгебры, координаты собственного вектора матрицы A соответствующего собственному значению λ_i совпадают с элементами i-го столбца матрицы V.

Теперь остается указать способ выбора матрицы $V^{(k)} = V^{(k)}_{ij}(\varphi)$ на k-м шаге и доказать сходимость метода.

Итак пусть есть матрица $A^{(k)}$. Найдем в ней максимальный по модулю недиагональный элемент $a_{ij}^{(k)}$. Поскольку матрица симметрическая, то можно считать, что $i<j$. Найдем значение угла поворота $\varphi = \varphi_k$ из условия равенства нулю элемента $a_{ij}^{(k+1)}$ матрицы

$$A^{(k+1)} = V^{(k)*} A^{(k)} V^{(k)}.$$

Положим $B = A^{(k)} V^{(k)}$. Тогда в виду определения матрицы поворота $V^{(k)} = V^{(k)}{}_{ij}(\varphi)$ элементы всех столбцов матрицы B, кроме i-го и j-го, совпадают с элементами матрицы $A^{(k)}$. Для элементов i-го и j-го столбцов имеем

$$\begin{aligned} b_{si} &= a_{si}^{(k)} \cos\varphi_k + a_{sj}^{(k)} \sin\varphi_k, \\ b_{sj} &= -a_{si}^{(k)} \sin\varphi_k + a_{sj}^{(k)} \cos\varphi_k, \quad s = 1,2,....,n. \end{aligned} \tag{4.7}$$

Аналогично матрица $A^{(k+1)} = V^{(k)*} B$ во всех строках, кроме i-ой и j-ой, имеет те же элементы, что и B. Элементы i-ой и j-ой строк имеют вид

$$\begin{aligned} a_{is}^{(k+1)} &= b_{is} \cos\varphi_k + b_{js} \sin\varphi_k, \\ a_{js}^{(k+1)} &= -b_{is} \sin\varphi_k + b_{js} \cos\varphi_k, \quad s = 1,2,....,n. \end{aligned} \tag{4.8}$$

Обратим внимание, что матрицы $A^{(k+1)}$ и $A^{(k)}$ различаются только суммой

$$[a_{is}^{(k+1)}]^2 + [a_{js}^{(k+1)}]^2 = b_{is}^2 + b_{js}^2 = [a_{is}^{(k)}]^2 + [a_{js}^{(k)}]^2$$

С учетом равенства $a_{ij}^{(k)} = a_{ji}^{(k)}$ из формул (4.7) и (4.8) получим

$$\begin{aligned} a_{ij}^{(k+1)} &= b_{ij} \cos\varphi_k + b_{jj} \sin\varphi_k = \\ &= (-a_{ii}^{(k)} \sin\varphi_k + a_{ij}^{(k)} \cos\varphi_k) \cos\varphi_k + (-a_{ji}^{(k)} \sin\varphi_k + a_{jj}^{(k)} \cos\varphi_k) \sin\varphi_k = \\ &= a_{ij}^{(k)} \cos 2\varphi_k + \frac{1}{2}(a_{jj}^{(k)} - a_{ii}^{(k)}) \sin 2\varphi_k, \end{aligned} \tag{4.9}$$

Полагая в (4.9) $a_{ij}^{(k+1)} = 0$, получим

$$tg\,2\varphi_k = 2a_{ij}^{(k)} / (a_{ii}^{(k)} - a_{jj}^{(k)}) \quad (-\pi/4 < \varphi_k < \pi/4)$$

или

$$\cos\varphi_k = \sqrt{\frac{1}{2}(1 + (1 + p_k^2))^{-1/2}}, \quad \sin\varphi_k = \operatorname{sgn} p_k \sqrt{\frac{1}{2}(1 - (1 + p_k^2))^{-1/2}}, \tag{4.10}$$

где

$$p_k = 2a_{ij}^{(k)} / (a_{ii}^{(k)} - a_{jj}^{(k)}).$$

Обозначим через $t(A)$ сумму квадратов всех недиагональных элементов матрицы A. Тогда

$$\begin{aligned} t(A^{(k+1)}) &= t(A^{(k)}) - 2[a_{ij}^{(k)}]^2 + \frac{1}{2}[(a_{jj}^{(k)} - a_{ii}^{(k)}) \sin 2\varphi_k + 2a_{ij}^{(k)} \cos 2\varphi_k]^2 = \\ &= t(A^{(k)}) - 2[a_{ij}^{(k)}]^2 + \frac{1}{2}[2a_{jj}^{(k+1)}]^2 = t(A^{(k)}) - 2[a_{ij}^{(k)}]^2. \end{aligned} \tag{4.11}$$

Таким образом, значение функции $t(A)$ уменьшается на каждом шаге.

Покажем, что итерационный процесс в методе Якоби сходится. Действительно, в силу выбора элемента $a_{ij}^{(k)}$ справедлива оценка

$$t(A^{(k)}) \le n(n-1)[a_{ij}^{(k)}]^2 ,$$

откуда

$$[a_{ij}^{(k)}]^2 \ge t(A^{(k)}) / n(n-1) .$$

С учетом этого неравенства из формулы (4.11) получаем

$$t(A^{(k+1)}) = t(A^{(k)}) - 2[a_{ij}^{(k)}]^2 \le t(A^{(k)}) - \frac{2t(A^{(k)})}{n(n-1)} = qt(A^{(k)}) ,$$

где

$$q = 1 - \frac{2}{n(n-1)}.$$

Очевидно, что $0<q<1$ при порядке матрицы $n>2$. Таким образом, получаем

$$t(A^{(k)}) \le q^k t(A^{(0)}) \quad k = 1, 2, ...$$

Последнее означает, что

$$\lim_{k\to\infty} t(A^{(k)}) = 0$$

и, следовательно, итерационный процесс сходится.

В итоге получаем следующий алгоритм метода вращений:

1) в матрице $A^{(k)}$ $(k=0,1,2,....)$ среди всех недиагональных элементов выбираем максимальный по абсолютной величине элемент, стоящий выше главной диагонали; определяем его номера i и j строки и столбца, в которых он стоит (если максимальных элементов несколько, можно взять любой из них);

2) по формулам (4.10) вычисляем $\cos\varphi_k$ и $\sin\varphi_k$, далее используя формулы (4.7) и (4.8) находим элементы матрицы $A^{(k+1)}$;

3) итерационный процесс останавливаем, когда в пределах принятой точности величиной $t(A^{(k+1)})$ можно пренебречь;

4) в качестве собственных значений матрицы A берем диагональные элементы матрицы $A^{(k+1)}$, в качестве собственных векторов – соответствующие столбцы матрицы

$$V = V^{(0)} V^{(1)} \ldots V^{(k)}.$$

5. ПРИНЦИП СЖИМАЮЩИХ ОТОБРАЖЕНИЙ

5.1. Полные метрические пространства

Метрическим пространством называется множество X элементов $x,y,\ldots$ произвольной природы, на котором определена так называемая функция расстояния или метрика $\rho = \rho(x,y)$, т. е. функция, для которой выполнены следующие аксиомы:

1) $\rho(x,y) \geq 0 \quad \forall x,y$, причем $\rho(x,y) = 0 \quad \Leftrightarrow \quad x = y$;

2) $\rho(x,y) = \rho(y,x) \quad \forall x,y,z$;

3) $\rho(x,y) \leq \rho(x,z) + \rho(z,y)$ для всех x, y, z.

Пример. Пусть $C_{[a,b]}$ − пространство непрерывных на отрезке [a,b] функций, тогда $\rho(f,g) = \max\limits_{a \leq t \leq b} |f(t) - g(t)|$ − расстояние в этом пространстве.

Очевидно, если X – нормированное пространство с нормой $\| \|$, то можно принять $\rho(x,y) = \|x - y\|$.

Обычно элементы метрического пространства называются точками этого пространства. Введем некоторые определения.

Последовательность $\{x^n\}$ в метрическом пространстве называется сходящейся к x, если $\lim\limits_{n\to\infty} \rho(x^n, x) = 0$. Сходимость последовательности $\{x^n\}$ к x обозначается $x^n \underset{n\to\infty}{\to} x$ или $\lim\limits_{n\to\infty} x^n = x$.

Окрестностью точки x^0 в пространстве X называется множество*:*

$U_\varepsilon(x^0) = \{x \in X \,|\, \rho(x, x^0) < \varepsilon\}$.

Предельной точкой множества $M \subset X$ называется такая точка, в любой окрестности которой находится бесконечно много элементов из множества M.

Замыканием множества $\overline{M}$ называется объединение множества M с множеством всех его предельных точек.

Множество замкнуто, если $M = \overline{M}$, т. е. когда оно совпадает со своим замыканием.

Пусть в метрическом пространстве X дана последовательность $\{x^n\} \subset X$. Эту последовательность будем называть *фундаментальной*, если для любого числа $\varepsilon > 0$ существует число $n_0 = n_0(\varepsilon)$ такое, что $\rho(x^n, y^m) < \varepsilon$ при любых $n,m > n_0$. Легко видеть, что всякая сходящаяся последовательность является фундаментальной.

Метрическое пространство называется *полным* (ПМП), если в нем любая фундаментальная последовательность сходится. Примерами ПМП являются пространства R, R^n, $C_{[a,b]}$.

Очевидно, любое замкнутое подмножество из ПМП в свою очередь тоже является ПМП. Действительно, так как метрика сохраняется и подмножество замкнуто, то любая фундаментальная последовательность сходится в нем.

5.2. Принцип сжимающих отображений

Пусть X – метрическое пространство. Рассмотрим отображение $A: X \to X$ пространства X в себя. Образ элемента x при отображении A обозначается:

$y = A(x)$ *или* $y = Ax$.

Отображение A называется *сжимающим*, если существует такое число α $(0 \le \alpha < 1)$, что

$$\rho(Ax, Ay) \le \alpha\rho(x, y) \quad \forall x, y \in X,$$

иными словами расстояние между образами точек меньше, чем расстояние между самими точками.

Убедимся, что всякое сжимающее отображение непрерывно.

Действительно, по определению отображение A является непрерывным, если для любой сходящейся последовательности $x^k \to x$ выполняется $Ax^k \to Ax$. Пусть $x^k \to x$, т. е. $\rho(x^k, x) \to 0$ при $k \to \infty$.

Тогда $\rho(Ax^k, Ax) \le \alpha\rho(x^k, x)$, откуда получаем $\lim\limits_{k\to\infty} \rho(Ax^k, Ax) \le 0$, что означает $\lim\limits_{x\to\infty} \rho(Ax^k, Ax) = 0$ или $\lim\limits_{x\to\infty} Ax^k = Ax$. Последнее равносильно непрерывности отображения A.

Точку x будем называть *неподвижной точкой* отображения A, если $x = Ax$.

Теорема 1. В ПМП любое сжимающее отображение имеет неподвижную точку, причем единственную.

Доказательство. Пусть x^0 - произвольная точка: $x^0 \in X$. Построим последовательность $\{x^k\}$ такую, что $x^k = Ax^{k-1}$ $k = 1,2,...$ и докажем, что она фундаментальная (а значит сходящаяся). Оценим расстояние

$$\rho(x^m, x^n) = \rho(A^m x^0, A^n x^0) =$$

$$= \rho(A^n(A^{m-n}x^0), A^n x^0) \le \left|\textit{по свойству сжимающего отображения}\right|$$

$$\le \alpha^n \rho(A^{m-n}x^0, x^0) = \alpha^n \rho(x^0, x^{m-n}) \le \left|\textit{вставляем средние точки}\right|$$

$$\le \alpha^n \left[\rho(x^0, x^1) + \rho(x^1, x^2) + ... + \rho(x^{m-n-1}, x^{m-n})\right] = \alpha^n \left[\rho(x^0, x^1) + \rho(Ax^0, Ax^1) + ... +\right.$$

$$\left. + \rho(A^{m-n-1}x^0, A^{m-n-1}x^1)\right] \le \alpha^n \left[\rho(x^0, x^1) + \alpha\rho(x^0, x^1) + ... + \alpha^{m-n-1}\rho(x^0, x^1)\right]$$

$$\le \alpha^n \rho(x^0, x^1)\left[1 + \alpha + ... + \alpha^{m-n-1}\right] \le$$

$$\le \alpha^n \rho(x^0, x^1)\left[1 + \alpha + \alpha^2 + ...\right] \le$$

$$\left|\textit{используем формулу суммы бесконечной геометрической прогрессии}\right|$$

$$\le \frac{\alpha^n}{1-\alpha}\rho(x^0, x^1) < \varepsilon \quad .$$

Решая неравенство

$$\frac{\alpha}{1-\alpha}\rho(x^0, x^1) < \varepsilon,$$

найдем $n_0 = n_0(\varepsilon)$, начиная с которого выполняется данное неравенство и, следовательно, неравенство $p(x^m, x^n) < \varepsilon$. Последнее означает, что последовательность $\{x^n\}$ – фундаментальная. Поскольку X – ПМП, то данная последовательность сходится, т. е. $x^n \to x \in X$.

Убедимся, что x – неподвижная точка отображения A. Действительно, переходя к пределу в равенстве $x^n = Ax^{n-1}$ и используя непрерывность отображения A, получим $x = Ax$.

Методом от противного докажем, что неподвижная точка единственная. Действительно, пусть есть две неподвижные точки. Тогда:

$x = Ax;$ $\qquad y = Ay.$

В силу того, что отображение A сжимающее, получим:

$$\rho(x, y) = \rho(Ax, Ay) \le \alpha\rho(x, y), \qquad \alpha < 1.$$

Откуда $\rho(x, y) = 0.$ Т. е. точки x и y совпадают. Теорема доказана.

Следствие. Теорема дает возможность вычислить неподвижную точку x данного отображения при любом начальном приближении x^0. При этом оценка погрешности на n-м шаге $\rho(x^n, x) \le \frac{\alpha^n}{1-\alpha}\rho(x^0, x^1)$.

Эта оценка получается, если перейти в неравенстве

$$\rho(x^n, x^m) \le \frac{\alpha^n}{1-\alpha}\rho(x^0, x^1)$$

к пределу при $m \to \infty.$

5.3. Приложения принципа сжимающих отображений

1) Решение системы линейных уравнений.

Рассмотрим систему $A\bar{x} = \bar{b}$ в пространстве R^n с некоторой нормой. Преобразуем систему к виду $\bar{x} = B\bar{x} + \bar{c}$, и будем рассматривать отображение $A(\bar{x}) = B\bar{x} + \bar{c}$.

Решить систему – значит найти неподвижную точку $\bar{x} = A\bar{x}$ отображения A. Проверим, когда отображение будет сжимающим:

$$\rho(A\bar{x}, A\bar{y}) = \|A\bar{x} - A\bar{y}\| = \|B\bar{x} + \bar{c} - B\bar{y} - \bar{c}\| = \|B(\bar{x} - \bar{y})\| =$$
(*применяем свойство матричной нормы*) =
$$= \|B\| \cdot \|\bar{x} - \bar{y}\| = \|B\| \cdot \rho(\bar{x}, \bar{y}).$$

Таким образом, A будет сжимающим отображением тогда и только тогда, когда $\|B\| < 1$. Следовательно, в этом случае можно построить сходящуюся итерационную последовательность.

2) Нахождение корней уравнения.

Рассмотрим уравнение:

$f(x) = 0 \Leftrightarrow x = \varphi(x)$, где $\varphi(x) = x - f(x)$ или $\varphi(x) = x + f(x)$.

Очевидно, найти корень уравнения равносильно тому, чтобы найти неподвижную точку отображения $x = \varphi(x)$.

Пусть выполнены условия:

а) $|\varphi(x) - \varphi(y)| \le M|x - y| \quad \forall x, y$, где $M < 1$ (условие Липшица)

(очевидно, данное условие всегда выполняется,
если $|\varphi'(x)| \le M < 1 \quad \forall x \in [a,b]$);

б) $\varphi : [a,b] \to [a,b]$.

Тогда итерационная последовательность $x^n = \varphi(x^{n-1})$ сходится в силу принципа сжимающих отображений (см. рис. 5.1 и 5.2).

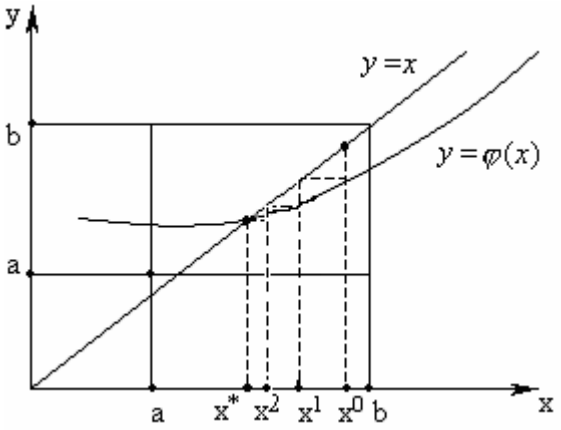


Рис. 5.1.

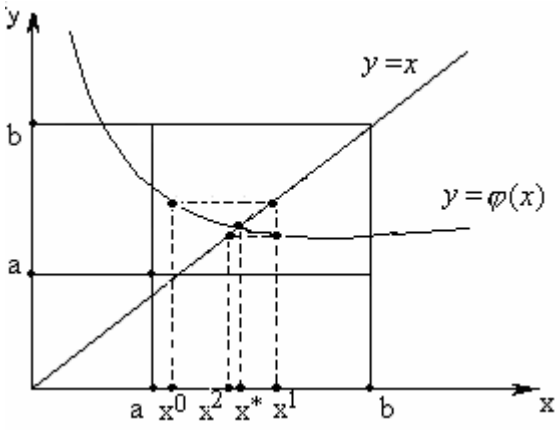


Рис. 5.2.

3) Решение задачи Коши.

Рассмотрим задачу Коши для дифференциального уравнения:

$$\begin{cases} y' = f(x,y) \\ y(x_0) = y_0. \end{cases}$$

Будем предполагать, что $f(x,y)$ – непрерывна по (x,y) и липшицева по y, т. е.

$|f(x,y_1) - f(x,y_2)| \le M|y_1 - y_2| \quad \forall (x,y_1) \; и \; (x,y_2) \in G$,

где G − двумерная область, содержащая точку (x_0, y_0).

Пусть G' − замкнутая ограниченная область, лежащая в G и содержащая точку (x_0, y_0). Тогда $f(x,y)$ ограничена в G', т.е. $|f(x,y)| \le K$ для всех $(x,y) \in G'$.

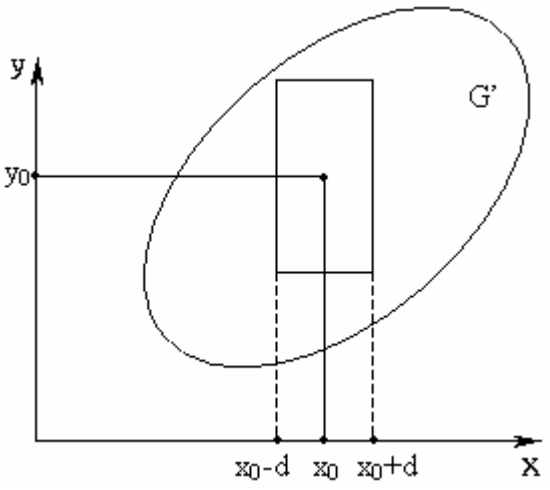


Рис. 5.3.

Выберем число d так, что для прямоугольника

$$P=\left\{(x,y):\left|x-x_0\right|\le d,\ \ \left|y-y_0\right|\le Kd\right\}$$

выполняются условия $P\subset G'$ и $Md<1$.

Рассматриваемая задача Коши равносильна интегральному уравнению:

$$y(x)=y_0+\int_{x_0}^{x}f(t,y(t))dt.$$

Построим отображение $A:C_{[x_0-d,x_0+d]}\to C_{[x_0-d,x_0+d]}$ и положим

$$A(\varphi)=y_0+\int_{x_0}^{x}f(t,\varphi(t))dt,$$

где $\varphi\in C_{[x_0-d,x_0+d]}$.

Покажем, что отображение A – сжимающее в пространстве непрерывных на данном отрезке функций $C^*\subset C_{[x_0-d,x_0+d]}$, которые дополнительно удовлетворяют условию $\left|\varphi(x)-y_0\right|\le Kd$.

Другими словами докажем, что отображение A:

1) не выводит за пределы пространства C^*, $A:C^*\to C^*$;

2)является сжимающим отображением.

Действительно, $\rho(f,g)=\max\limits_{x_0-d\le x\le x_0+d}\left|f(x)-g(x)\right|$ – расстояние в C^*.

Если φ непрерывная функция, то $A\varphi$ – тоже непрерывная функция и для $\forall\varphi\in C^*$

$$\left|A\varphi(x)-y_0\right|=\max\left|y_0+\int_{x_0}^{x}f(t,\varphi(t))dt-y_0\right|\le\max_x\int_{x_0}^{x}\left|f(t,\varphi(t))dt\right|\le\int_{x_0}^{x_0+d}\left|f(t,\varphi(t))dt\right|\le$$

$$\le\int_{x_0}^{x_0+d}Kdt=Kd,$$

т. е. действительно $A\varphi(x)\in C^*$.

Пусть теперь $\varphi,\psi\in C^*$. Тогда

$$\rho(A\varphi, A\psi) = \max_{x_0 \le x \le x_0+d} \left| y_0 + \int_{x_0}^{x} f(t,\varphi(t))\,dt - y_0 - \int_{x_0}^{x} f(t,\psi(t))\,dt \right| \le$$

$$\le \max_{x_0 \le x \le x_0+d} \left| \int_{x_0}^{x} [f(t,\varphi(t)) - f(t,\psi(t))]\,dt \right| \le$$

$$\le \max_{x_0 \le x \le x_0+d} \int_{x_0}^{x} |f(t,\varphi(t)) - f(t,\psi(t))|\,dt \le \int_{x_0}^{x_0+d} M|\varphi(t) - \psi(t)|\,dt \le$$

$$\le M \int_{x_0}^{x_0+d} \max_{x_0 \le x \le x_0+d} |\varphi(t) - \psi(t)|\,dt = Md\,\rho(\varphi,\psi),$$

где $Md<1$.

Следовательно, отображение A – сжимающее и в силу принципа сжимающих отображений при сделанных предположениях существует единственное решение интегрального уравнения

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))dt,$$

а значит и задачи Коши.

6. РЕШЕНИЕ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

Пусть дано уравнение *f(x)* = 0. Ставится задача: найти решение данного уравнения с точностью до некоторой заданной величины ε. Точное решение данного уравнения будем обозначать через x^*, а приближенное через $\hat{x}$.

Методы решения уравнений делятся на прямые и итерационные.
Прямые методы – это методы, позволяющие вычислить решение по формуле. Например, нахождение корней квадратного или кубического уравнения.
Итерационные методы − это методы, в которых задается некоторое начальное приближение и строится сходящаяся последовательность приближений к точному решению, причем каждое последующее приближение вычисляется с использованием предыдущих:

$$x_n = \varphi_n(x_0, x_1, ..., x_{n-1}).$$

Очевидно, что прямые методы могут быть использованы только для решения простейших уравнений (так уже для многочлена 5-й степени не существует общих формул для вычисления корней).

Одним из простейших методов решения нелинейных уравнений является использование теоремы Больцано-Коши. Известно, что если функция f непрерывна на отрезке $[a,b]$ и $f(a)\cdot f(b) < 0,$ то на отрезке $[a,b]$ существует хотя бы один корень уравнения $f(x) = 0$ (рис. 6.1).

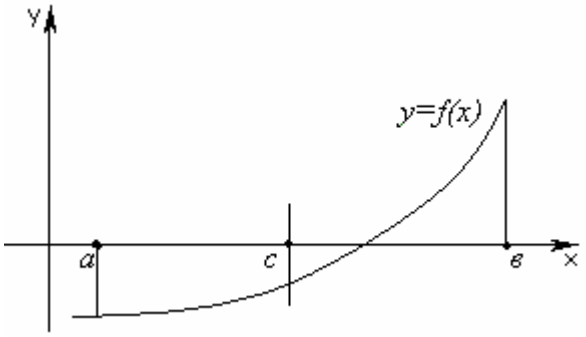


Рис. 6.1

Методом дихотомии (делением отрезка пополам) можно выделить промежуток половинной длины, на концах которого функция принимает значения разных знаков. Продолжая процесс можно найти корень с любой заданной точностью.

Очевидным недостатком этого метода является его трудоемкость. Другой менее очевидный, однако более существенный недостаток иллюстрируется рисунком 6.2.

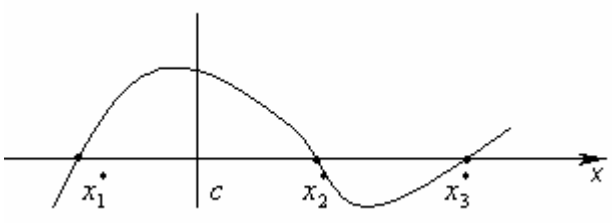


Рис.6.2

Если на отрезке есть несколько корней данного уравнения (см. рис. 6.2), при делении отрезка пополам будет найден только один корень и произойдет потеря других. Чтобы избежать этого, нужно предварительно провести процедуру отделения корней.

Другим простейшим методом решения нелинейных уравнений является использование принципа сжимающих отображений.

Пусть функция f непрерывно дифференцируема на отрезке $[a,b]$ и на концах его принимает значения разных знаков. По уравнению $f(x)=0$ строим уравнение $x=\varphi(x)$, где $\varphi(x)=x-\lambda f(x)$. Множитель λ выбираем таким образом, чтобы на отрезке были выполнены условия φ: $[a,\ b]\to[a,\ b]$ и $|\varphi'(x)|<q<1$ на $[a,b]$. После этого строим итерационную последовательность

$$x_n=\varphi(x_{n-1}),\quad n=1,2,\ldots..$$

которая сходится к искомому решению уравнения.

6.1. Проблема отделения корней

Поставим задачу: найти интервал *(a,b),* на котором для заданной функции *f(x)* выполняется условие $f(a)\cdot f(b)<0$ и который содержит *только один* корень функции *f(x)*.

Если функция на заданном интервале непрерывно дифференцируема, то можно воспользоваться следствием из теоремы Ролля, по которому между парой корней всегда находится по крайней мере одна стационарная точка. Алгоритм решения задачи в данном случае будет следующий:

1)находим производную $f'(x)$,

2) решаем уравнение $f'(x)=0$ для нахождения стационарных точек,

3) разбиваем исходный интервал (*a,b*) на меньшие интервалы с помощью найденных стационарных точек,

4) из полученных интервалов выбираем только те, на концах которых *f(x)* принимает значения разных знаков,

5)уточняем интервалы за счет их сужения.

Очевидным недостатком метода является трудность нахождения стационарных точек (зачастую это более трудная задача, чем решение заданного уравнения). К достоинствам метода можно отнести его принципиальную простоту и то обстоятельство, что часто других более хороших способов нет.

Для отделения корней можно также воспользоваться графиком функции. К достоинствам подобного способа можно отнести его наглядность и простоту, к недостаткам низкую точность и необходимость строить график функции.

Полезным средством для отделения корней является также использование теоремы Штурма.

Пусть *f(x)* – многочлен, и уравнение *f(x)=0* не имеет кратных корней, т.е. нет точек, в которых $f(x)=0$ и $f'(x)=0$ (стационарные точки не являются корнями).

Построим так называемый ряд Штурма: $f_0(x)$, $f_1(x)$, ... , $f_n(x)$, где

$f_0(x)=f(x),$

$f_1(x)=f'(x),$

$f_2(x)$ – остаток от деления $\frac{f_0}{f_1}$, взятый с обратным знаком,

$f_k(x)$ − остаток от деления $\frac{f_{k-2}}{f_{k-1}}$, взятый с обратным знаком,

и так далее, пока не получим постоянную.

Обозначим через *N(a)* – число перемен знаков в ряде Штурма, при *x=a;* через *N(b)* – число перемен знаков в ряде Штурма, при *x=b.*

Теорема Штурма. При сделанных выше предположениях, число корней уравнения *f(x) = 0* на отрезке [a, b] равно *N(a) - N(b).*

Получим простую оценку для погрешности приближения.

Пусть

$f(x)=0,\ a\le x\le b.$

Тогда по формуле конечных приращений

$f(\hat{x})-f(x^{\bullet})=f'(c)\cdot(\hat{x}-x^{*})$, где $c\in(a,b)$.

Так как $x^{\bullet}$ – корень, то $f(x^{\bullet})=0$ и, следовательно,

$f(\hat{x})=f'(c)\cdot(\hat{x}-x^{*})$.

Предполагаем, что в интервале (a,b) корень отделен, а производная не обращается в нуль на $[a,b]$, т.е. стационарных точек нет.

Оценим снизу и сверху абсолютное значение производной:

$m_1\le|f'(x)|\le M_1$. Тогда получаем оценку

$$|\hat{x}-x^{*}|\le\frac{|f(\hat{x})|}{m_1}. \qquad (6.1)$$

6.2. Метод хорд

Пусть дано уравнение $f(x)=0,\ a\le x\le b$, где $f(x)$ − дважды непрерывно дифференцируемая функция.

Пусть выполняется условие $f(a)\cdot f(b)<0$ и проведено отделение корней, т.е. на данном интервале (*a, b*) находится один корень уравнения. При этом, не ограничивая общности, можно считать, что *f(b) > 0* .

Пусть функция *f* выпукла на интервале *(a, b)* (рис. 6.3).

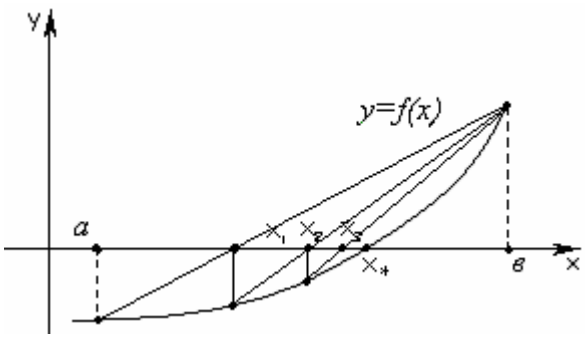


Рис. 6.3

Заменим график функции хордой (прямой), проходящей через точки $M_0(a, f(a))$ и $M_1(b, f(b))$.

Уравнение прямой, проходящей через две заданные точки, можно записать в виде $\frac{y-y_1}{y_2-y_1}=\frac{x-x_1}{x_2-x_1}$. В нашем случае получим $\frac{y-f(a)}{f(b)-f(a)}=\frac{x-a}{b-a}$.

Найдем точку пересечения хорды с осью Ох.

Полагая $y=0$, получаем из предыдущего уравнения:

$$x_1 = a - \frac{f(a)}{f(b)-f(a)} \cdot (b-a).$$

Теперь возьмем интервал *(x1,b)* в качестве исходного и повторим вышеописанную процедуру (рис. 6.3). Получим

$$x_2 = x_1 - \frac{f(x_1)}{f(b)-f(x_1)} \cdot (b-x_1).$$

Продолжим процесс. Каждое последующее приближение вычисляется по рекуррентной формуле

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f(b)-f(x_{n-1})} \cdot (b-x_{n-1}) \quad n = 1,2,... \quad , \tag{6.2}$$

$x_0 = a$.

Если же функция вогнута (см. рис. 6.4),

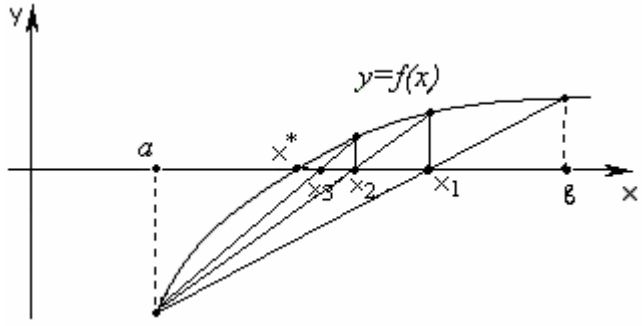


Рис. 6.4

уравнение прямой, соединяющей точки $M_0(a, f(a))$ и $M_1(b, f(b))$ запишем в виде

$$\frac{y-f(b)}{f(a)-f(b)}=\frac{x-b}{a-b}.$$

Найдем точку пересечения хорды с осью Ox:

$$x_1 = b - \frac{f(b)}{f(a) - f(b)} \cdot (a - b).$$

Теперь возьмем интервал (a,x_1) в качестве исходного и найдем точки пересечения хорды, соединяющей точки $(a, f(a))$ и $(x_1, f(x_1))$ с осью абсцисс (рис. 6.4). Получим

$$x_2 = x_1 - \frac{f(x_1)}{f(a) - f(x_1)} \cdot (a - x_1).$$

Повторяя данную процедуру, получаем рекуррентную формулу

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f(a) - f(x_{n-1})} \cdot (a - x_{n-1}) \quad n = 1,2,.. \tag{6.3}$$

$$x_0 = b \ .$$

Описанный выше метод построения рекуррентных последовательностей (6.2) и (6.3) называется методом хорд. Для использования метода хорд нужно было бы предварительно найти точки перегиба и выделить участки, на которых функция не меняет характер выпуклости. Однако на практике поступают проще: в случае $f(b)f''(b) > 0$ для построения рекуррентной последовательности применяются формулы (6.2), а в случае, когда $f(a)f''(a) > 0$, применяют формулы (6.3).

Докажем сходимость метода хорд.

Очевидно, обе последовательности (6.2) и (6.3) могут быть записаны в виде

$$x_n = x_{n-1} - \frac{f(x_{n-1})(\hat{x} - x_{n-1})}{f(\hat{x}) - f(x_{n-1})},$$

где $\hat{x} = a$ или $\hat{x} = b$.

Построим функцию $\varphi(x) = x - \frac{f(x)}{f(\hat{x}) - f(x)}(\hat{x} - x)$,

где $\hat{x} = a$ или $\hat{x} = b$ (в зависимости от характера выпуклости). Тогда метод хорд дает рекуррентную последовательность, определяемую единой формулой

$$x_n = \varphi(x_{n-1}), \quad n = 1,... \ x_0 = \hat{x} \ . \tag{6.4}$$

Покажем, что φ будет сжимающим отображением в случае достаточно малого интервала (a,b). Приведем выражение для $\varphi(x)$ к общему знаменателю:

$$\varphi(x) = \frac{xf(\hat{x}) - xf(x) - \hat{x}f(x) + xf(x)}{f(\hat{x}) - f(x)} = \frac{xf(\hat{x}) - \hat{x}f(x)}{f(\hat{x}) - f(x)}.$$

Найдем производную

$$\varphi'(x) = \frac{f(\hat{x})(f(\hat{x}) - f(x)) - \hat{x}f'(x)(f(\hat{x} - f(x)) + f'(x)(xf(\hat{x}) - \hat{x}f(x))}{(f(\hat{x}) - f(x))^2}.$$

Вычислим значение этой производной в точке x^*, учитывая, что $f(x^*) = 0$:

$$\varphi'(x^*) = \frac{f^2(\hat{x}) - \hat{x}f'(x^*)f(\hat{x}) + f'(x^*)x^*f(\hat{x})}{f^2(\hat{x})} = \frac{f(\hat{x}) - f'(x^*)(\hat{x} - x^*)}{f(\hat{x})}.$$

По формуле Тейлора

$$f(\hat{x}) = f(x^*) + f'(x^*)(\hat{x} - x^*) + \frac{f''(\xi)}{2}(\hat{x} - x^*)^2,$$

где ξ некоторая точка, лежащая между x^* и x. С учетом данного выражения получим

$$\varphi'(x^*) = \frac{f''(\xi)}{f(\hat{x})} \cdot \frac{(\hat{x} - x^*)^2}{2}.$$

Оценим полученное выражение для производной $\varphi'(x^*)$:

$$\left|\varphi'(x^*)\right| = \frac{\left|f''(\xi)\right|}{\left|f(\hat{x})\right|} \cdot \frac{\left|\hat{x} - x^*\right|^2}{2} \le \frac{\left|\hat{x} - x^*\right|}{\left|f(\hat{x})\right|} \cdot \frac{M_2}{2}\left|\hat{x} - x^*\right|,$$

где $M_2 = \max\limits_{a \le x \le b}\left|f''(x)\right|$.

Откуда, учитывая полученную ранее оценку (6.1)

$$\left|\hat{x} - x^*\right| \le \frac{\left|f(\hat{x})\right|}{m_1},$$

имеем

$$\left|\varphi'(x^*)\right| \le \frac{M_2}{2m_1}\left|\hat{x} - x^*\right|.$$

Таким образом, для сходимости итерационной последовательности (6.4) достаточно потребовать выполнения условия

$$\frac{M_2}{2m_1}\left|\hat{x} - x^*\right| < 1 \quad \text{или} \quad \left|\hat{x} - x^*\right| < \frac{2m_1}{M_2}.$$

Очевидно, при значении $\hat{x}$ достаточно близком к точке x^* последнее неравенство всегда выполняется и, значит, в некоторой окрестности $U_\delta(x^*)$ точки x^* выполнено условие $|\varphi'(x)| \le q < 1$ и, следовательно, отображение φ является сжимающим.

Остается показать, что $\varphi : U_\delta(x^*) \to U_\delta(x^*)$. Действительно, для любого x из этой окрестности справедливо

$$|\varphi(x) - x^*| = |\ \varphi(x^*) + \varphi'(\zeta)(x - x^*) - x^*| \le$$

$$\le\ |\ \varphi(x^*) - x^*| + |\varphi'(\zeta)(x - x^*)| \le\ q\,|x - x^*| < \delta,$$

где точка ζ лежит в окрестности $U_\delta(x^*)$. Последнее неравенство означает, что $\varphi : U_\delta(x^*) \to U_\delta(x^*)$.

Таким образом, в силу принципа сжимающих отображений, метод хорд сходится, когда начальное приближение достаточно близко к решению.

6.3. Метод Ньютона (метод касательных)

Пусть дано уравнение $f(x) = 0$, $a \le x \le b$, где $f(x)$ − дважды непрерывно дифференцируемая функция.

Если выполняется условие $f(a) \cdot f(b) < 0$, то на данном интервале содержится корень уравнения. Предположим, что корень отделен, т.е. на данном интервале он только один. Не ограничивая общности, можно считать, что $f(a) < 0,\ f(b) > 0$.

Рассмотрим рис. 6.5.

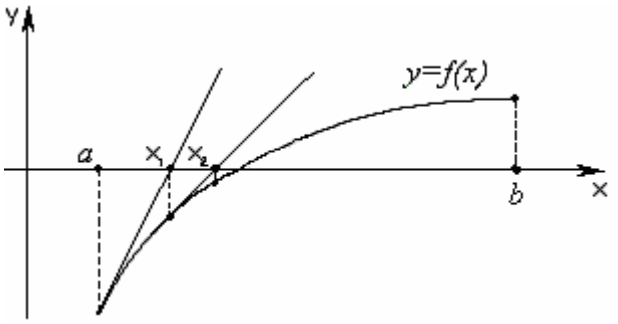


Рис. 6.5

Пусть

$f(x) = 0$, $a \le x \le b$.

Запишем уравнения касательной в точке $(a, f(a))$:

$y - f(a) = f'(a)(x - a)$.

Найдем точку ее пересечения с осью *Ox*. Получаем

$$x_1 = a - \frac{f(a)}{f'(a)}.$$

Построив касательную в точке $(x_1, f(x_1))$, находим точку ее пересечения с осью *Ox*:

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

Продолжая процесс, получим

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \quad , \quad x_0 = a, \quad n = 1,2,\ldots \ .$$

Для рис. 6.5 характерно условие: $f(a) f''(a) > 0$, так как вторая производная и сама функция отрицательны.

Рассмотрим ситуацию с $f(b)f''(b) > 0$ (рис. 6.6).

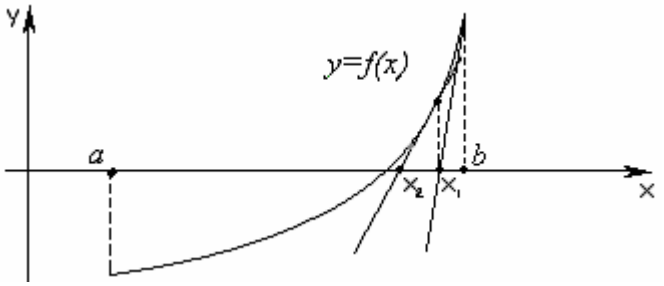


Рис. 6.6

Здесь $x_0 = b$ и $f(b)f''(b) > 0$.

Построив касательную в точке $(b, f(b))$, находим точку ее пересечения с осью *Ox*:

$$x_1 = b - \frac{f(b)}{f'(b)}, \quad \text{т. е.} \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Продолжая процесс, получаем ту же рекуррентную формулу

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}, \qquad n = 1,2,\ldots,$$

но с начальным условием $x_0 = b$.

Таким образом, рекуррентная последовательность Ньютона задается единой формулой

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}, \qquad n = 1,2,\ldots,$$

где $x_0 = a$ (при $f(a)f''(a) > 0$) или $x_0 = b$ (при $f(b)f''(b) > 0$).

Исследуем сходимость метода Ньютона:

$$x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})} \qquad k = 1,2,\ldots,$$

$x_0 = a$ или b.

Причем $x_0 = a$ при $f(a)f''(a) > 0$ и $x_0 = b$ при $f(b)f''(b) > 0$.

Очевидно,

$$f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{f''(\xi)(x^* - x_k)^2}{2},$$

где ξ − некоторая точка между точками *a* и *b*.

Получаем

$$0 = f(x_k) + f'(x_k)(x^* - x_k) + \frac{f''(\xi)(x^* - x_k)^2}{2},$$

откуда

$$\frac{-f(x_k)}{f'(x_k)} = x^* - x_k + \frac{f''(\xi)}{2f'(x_k)}(x^* - x_k)^2,$$

$$x_k - \frac{f(x_k)}{f'(x_k)} = x^* + \frac{f''(\xi)}{2f'(x_k)}(x^* - x_k)^2,$$

или окончательно

$$x_{k+1} - x^* = \frac{f''(\xi)}{2f'(x_k)}(x^* - x_k)^2.$$

Отсюда

$$\left|x^* - x_{k+1}\right| \le \frac{\left|f''(\xi)\right|}{2\left|f'(x_k)\right|}\left|x^* - x_k\right|^2 \le \frac{M_2}{2m_1}\left|x^* - x_k\right|^2,$$

где

$$f(x^k) + J(x^k)(x^k)(x - x^k) = 0 \; M_2 = \max_{a \le x \le b}\left|f''(x)\right|, \qquad m_1 = \min_{a \le x \le b}\left|f'(x)\right|.$$

Таким образом, получаем следующую оценку скорости сходимости:

$$\left|x^* - x_{k+1}\right| \le \frac{M_2}{2m_1}\left|x^* - x_k\right|^2.$$

Из полученной оценки можно получить

$$\left|x^* - x_1\right| \le \frac{M_2}{2m_1}\left|x^* - x_0\right|^2 \le q\left|x^* - x_0\right|,$$

где $\frac{M_2}{2m_1}\left|x^* - x_0\right| = q$, и далее

$$\left|x^* - x_2\right| \le \frac{M_2}{2m_1}\left|x^* - x_1\right|^2 \le \frac{M_2}{2m_1}\cdot q^2 \cdot \left|x^* - x_0\right|^2 \le q^3\left|x^* - x_0\right|,$$

$$\left|x^* - x_3\right| \le \frac{M_2}{2m_1}\left|x^* - x_2\right|^2 \le \frac{M_2}{2m_1}\cdot q^6 \cdot \left|x^* - x_0\right|^2 \le q^7\left|x^* - x_0\right|, \; \ldots \; .$$

Т.е. итерационная последовательность будет заведомо сходиться, если $\frac{M_2}{2m_1}\left|x^* - x_0\right| < 1$, т. е. если x_0 выбрано достаточно близко к x^*.

Кроме того, полученная оценка скорости сходимости свидетельствует об очень быстром характере сходимости. Такая скорость сходимости называется квадратичной. Таким образом, метод Ньютона сходится с квадратичной скоростью.

К некоторым недостаткам метода относится необходимость выбора хорошего начального приближения.

Пример. Вычислить $\sqrt{13}$ с точностью $\varepsilon = 0{,}00001$.

Строим функцию $f(x) = x^2 - 13$ и решаем методом Ньютона уравнение $f(x) = 0$ на отрезке [*3, 4*]. Очевидно $f(4)f''(4) > 0$, следовательно $x_0 = 4$.

Вычисляем

$$x_1 = 4 - \frac{f(4)}{f'(4)} = 4 - \frac{3}{8} = 3\frac{5}{8}, \quad \text{и т.д.}$$

Можно проверить, что уже x_3 дает приближение с необходимой точностью.

6.4. Комбинированный метод хорд и касательных

Пусть для определенности $f(b)f''(b) > 0$ (рис. 6.7).

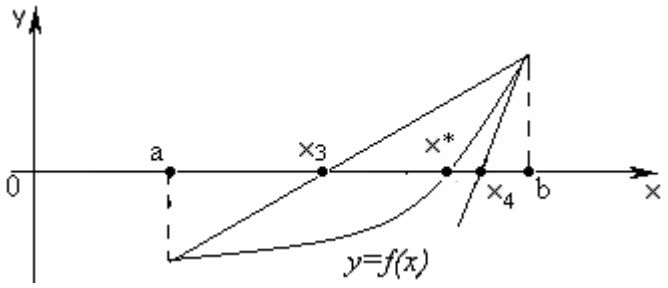


Рис. 6.7.

Тогда, применяя метод хорд при $x_1 = a$, получим

$$x_3 = x_1 - \frac{f(x_1)}{f(b) - f(x_1)}(b - x_1).$$

Применяя метод Ньютона при $x_2 = b$, получим

$$x_4 = x_2 - \frac{f(x_2)}{f'(x_2)}.$$

При этом $x^* \in (x_3, x_4)$.

Продолжая процесс далее, получаем

$$\begin{cases} x_5 = x_3 - \dfrac{f(x_3)}{f(x_4) - f(x_3)}(x_4 - x_3) \\ x_6 = x_4 - \dfrac{f(x_4)}{f'(x_4)}, \end{cases}$$

причем $x^* \in (x_5, x_6)$.

Отсюда

$$x_{2n+1} = x_{2n-1} - \frac{f(x_{2n-1})}{f(x_{2n}) - f(x_{2n-1})}(x_{2n} - x_{2n-1}), \qquad n = 1,\dots .$$

$$x_{2n+2} = x_{2n} - \frac{f(x_{2n})}{f'(x_{2n})}(x_{2n} - x_{2n-1}), \qquad n = 1,\dots ,$$

причем всегда $x^* \in (x_{2n+1}, x_{2n+2})$.

Пусть теперь $f(a)f''(a) > 0$ (рис. 6.8).

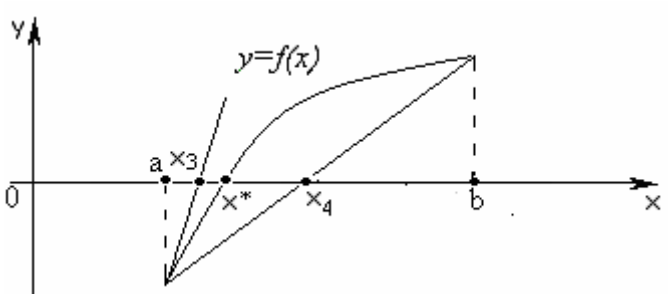


Рис. 6.8.

Применяем метод Ньютона при $x_1 = a$ и метод хорд при $x_2 = b$, и получаем

$$\begin{cases} x_{2n+1} = x_{2n-1} - \dfrac{f(x_{2n-1})}{f'(x_{2n-1})}, \ n = 1, 2... \\ x_{2n+2} = x_{2n} - \dfrac{f(x_{2n})}{f(a) - f(x_{2n})}(a - x_{2n}), \ n = 1, 2..., \end{cases}$$

причем $x^* \in (x_{2n+1}, x_{2n+2})$.

Таким образом, комбинированный метод хорд и касательных удобен тем, что корень уравнения всегда находится в интервале между двумя последовательными приближениями.

6.5. Решение систем нелинейных уравнений

Пусть есть система *n* уравнений с *n* неизвестными:

$$\begin{cases} f_1(x_1, ..., x_n) = 0 \\ ..... \\ f_n(x_1, ..., x_n) = 0. \end{cases} \qquad (6.5)$$

Запишем ее в векторном виде:

$$f(\overline{x}) = \overline{0}, \qquad (6.6)$$

где

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}, \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Метод простых итераций

Система (6.4) преобразуется к виду:

$$x = \varphi(x), . \qquad (6.7)$$

где $\varphi = \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_n \end{pmatrix}$.

Пусть есть начальное приближение x^0, $U_\delta(x^0)$ – δ-окрестность начального приближения и пусть φ - сжимающее отображение на $U_\delta(x^0)$, т. е.

$\|\varphi(x') - \varphi(x^2)\| \le q\|x^1 - x^2\|$, где $0 \le q < 1$,

$\forall x^1, x^2 \in U_\delta(x^0)$.

Последнее условие заведомо выполняется, если

$$\left\|\frac{\partial\varphi}{\partial x}(x)\right\| \le q < 1, \qquad \forall x \in U_\delta(x^0),$$

где

$$\frac{\partial\varphi}{\partial x} = \begin{pmatrix} \frac{\partial\varphi_1}{\partial x_1} & \dots & \frac{\partial\varphi_1}{\partial x_n} \\ \dots & \dots & \dots \\ \frac{\partial\varphi_n}{\partial x_1} & \dots & \frac{\partial\varphi_n}{\partial x_n} \end{pmatrix}$$

- матрица частных производных , называемая также матрицей Якоби функции *φ(x).*

Теорема1. Пусть φ – непрерывно дифференцируемая векторная функция, для которой выполняются следующие условия:

1) $\|\varphi(x^0) - x^0\| \le \delta(1-q)$;

2) $\left\|\frac{\partial\varphi}{\partial x}(x)\right\| \le q, \qquad (q < 1), \qquad \forall x \in U_\delta(x^0)$.

Тогда система (6.7) имеет решение $x^* \in U_\delta(x^0)$, причем единственное, и итерационная последовательность $x^k = \varphi(x^{k-1}), \quad k = 1,2,\dots$, с начальным приближением x^0 сходится к решению x^* системы (6.7) . При этом справедлива следующая оценка скорости сходимости:

$$\|x^k - x^*\| \le \frac{q^n}{1-q}\|x^1 - x^0\|.$$

Доказательство. Из условия (6.7) вытекает сжимаемость этого отображения. Проверим, что

$\varphi : U_\delta(x^0) \to U_\delta(x^0)$.

Разложив $\varphi(x)$ по формуле Тейлора, получаем для $\forall x \in U_\delta(x^0)$

$$\varphi(x) = \varphi(x^0) + \frac{\partial\varphi}{\partial x}(\xi)(x - x^0),$$

где $\xi \in U_\delta(x^0)$. Тогда

$$\|\varphi(x) - x^0\| = \left\|\varphi(x^0) + \frac{\partial\varphi}{\partial x}(\xi)(x - x^0) - x^0\right\| \le$$

$$\le \|\varphi(x^0) - x^0\| + \left\|\frac{\partial\varphi}{\partial x}(\xi)\right\| \cdot \|x - x^0\| < \delta(1-q) + q\delta = \delta,$$

т. е. $\varphi(x) \in U_\delta(x^0)$.

Следовательно, мы можем применить принцип сжимающих отображений, в силу которого получаем результат теоремы. При этом из принципа сжимающих отображений следует, что

$$\rho(x^*, x^k) \le \frac{\alpha^k}{1-\alpha}\rho(x^0, x^1),$$

что равносильно оценке скорости сходимости

$$\left\|x^k - x^*\right\| \le \frac{q^n}{1-q}\left\|x' - x^0\right\|$$

в нашем случае.

Метод Ньютона

Рассмотрим снова систему (6.6):

$f(\bar{x}) = \bar{0}$.

Пусть x^0 – начальное приближение. Предположим, что в окрестности $U_\delta(x^0)$ начального приближения матрица Якоби

$$J(x) = \frac{\partial f}{\partial x} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \dots & \dots & \dots \\ \frac{\partial f_n}{\partial x_1} & \dots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

не вырождена, т. е. $\det J(x) \neq 0$.

Заменим систему (6.6) линеаризованной системой:

$f(x^0) + J(x^0)(x - x^0) = 0$.

Решая данную систему относительно x, получим:

$$x - x^0 = -J^{-1}(x^0) f(x^0),$$

откуда

$x^1 = x^0 - J^{-1}(x^0) f(x^0)$.

Заменим (6.6) на систему вида:

$$f(x^1) + J(x^1)(x^1)(x - x^1) = 0.$$

Отсюда:

$$x^2 = x^1 - J^{-1}(x^1) f(x^1),$$

и продолжая процесс, получим

$$x^k = x^{k-1} - J^{-1}(x^{k-1}) f(x^{k-1}), \qquad k = 1,2,\dots.$$

Полученная рекуррентная последовательность называется *последовательностью Ньютона*. При $n = 1$, из нее получается обычный метод Ньютона.

Так же как и в случае $n=1$, можно показать, что рекуррентная последовательность Ньютона сходится, если начальное приближение выбрано достаточно близко к решению x^*.

Часто метод Ньютона используется не с рекуррентной формулой Ньютона, а на каждой итерации решают систему линейных уравнений

$$f(x^k) + J(x^k)(x^k)(x - x^k) = 0 .$$

Оценим сходимость метода. Очевидно,

$$f(x^*) = f(x^k) + \frac{\partial f(x^k)}{\partial x}(x^* - x^k) + O(\left\|x^* - x^k\right\|^2),$$

где $O(\left\| x^k \right\|^m)$ означает, что $O(\left\| x^k \right\|^m) \le M \left\|x^k\right\|^m$, $M = const > 0$.

Поскольку $f(x^*) = 0$, получим:

$-J^{-1}(x^k) f(x^k) = x^* - x^k + J^{-1}(x^k) O(\left\| x^* - x^k \right\|^2)$, т. е.

$$\left\| x^{k+1} - x^k \right\| = \left\| J^{-1}(x^k) \cdot O(\left\| x^* - x^k \right\|^2) \right\| \le M \left\| x^* - x^k \right\|^2 .$$

Таким образом, метод Ньютона имеет квадратичную сходимость.

Недостатки метода Ньютона: начальное приближение должно быть близким к решению, а матрица Якоби должна быть невырожденной.

Достоинство: быстрая сходимость.

Отметим, что выбор начального приближения является слабым местом итерационных методов.

7. АППРОКСИМАЦИЯ И ИНТЕРПОЛЯЦИЯ ФУНКЦИЙ

Из математического анализа известно, что в окрестности точки x_0 любую n раз непрерывно дифференцируемую функцию можно аппроксимировать (приблизить) ее многочленом Тейлора:

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)(x - x_0)^k}{k!},$$

причем

$$f(x_0) = P_n(x_0),$$
$$f'(x_0) = P_n'(x_0),$$
$$\ldots\ldots\ldots\ldots$$
$$f^{(n)}(x_0) = P_n^{(n)}(x_0).$$

Очевидно, такая аппроксимация во многих отношениях является очень хорошей, но она имеет локальный характер, т.е. хорошо аппроксимирует функцию только вблизи точки x_0 . Это главный недостаток аппроксимации с помощью многочлена Тейлора.

Если речь идет об аппроксимации функции на отрезке, применяются другие методы.

Пусть $f(x) \in C[a,b]$ – непрерывная функция. Рассмотрим задачу аппроксимации (приближения) ее более простой функцией (обычно многочленом).

Известно из математического анализа, что в силу теоремы Вейерштрасса, любую функцию можно с какой угодно точностью приблизить многочленом по норме $\|f(x)\| = \max\limits_{a \le x \le b} |f(x)|$ пространства C[*a, b*], т. е. в смысле равномерной сходимости. Но существуют и другие нормы:

$$\|f(x)\| = \int_a^b |f(x)|\,dx \quad \text{или} \quad \|f(x)\| = \sqrt{\int_b^a |f(x)|^2\,dx}\,.$$

Тогда $\|f(x) - P(x)\| < \varepsilon$ означает, что площадь или усредненная площади фигуры, заключенной между графиками функции *f(x)* и многочлена *P(x)*, должна быть меньше ε (заданной точности).

Возможен и другой подход, когда в качестве аппроксимирующей функции берут многочлен или другую достаточно простую функцию, значения которых совпадают со значениями исходной функции в заданных заранее точках, так называемых узлах. Такого рода приближение функций имеет свое собственное название - интерполяция.

7.1. Интерполяционный многочлен

Пусть *f(x)* – функция, непрерывная на отрезке [*a,b*].

Выберем на этом отрезке точки, называемые *узлами интерполяции*:

$a \le x_0 < x_1 < ... < x_n \le b$.

Предположим, что известны значения функции в узлах интерполяции:
$f(x_k) = y_k, \quad k = 0,1,...,n$.

Ставится задача найти многочлен $P_n(x)$ такой, что

$$P_n(x_k) = y_k, \qquad \forall k = 0,1,...,n. \tag{7.1}$$

Такой многочлен $P_n(x)$ называется *интерполяционным многочленом*, а задача его нахождения – *задачей интерполяции.*

Покажем, что задача интерполяции имеет решение, причем единственное.

Пусть $P_n(x) = \sum_{k=0}^{n} a_k x^{n-k}$.

Тогда для определения коэффициентов многочлена из условия (7.1) получаем систему:

$$\begin{cases} a_0 x_0^n + a_1 x_0^{n-1} + ... + a_n = y_0 \\ a_0 x_1^n + a_1 x_1^{n-1} + ... + a_n = y_1 \\ ..... \\ a_0 x_n^n + a_1 x_n^{n-1} + ... + a_n = y_n . \end{cases}$$

Ее определитель Δ с точностью до знака совпадает с так называемым определителем Вандермонда.

$$W(x_0,...,x_n) = \begin{vmatrix} 1 & 1 & ... & 1 \\ x_0 & x_1 & ... & x_n \\ x_0^2 & x_1^2 & ... & x_n^2 \\ ... & ... & ... & ... \\ x_0^n & x_1^n & ... & x_n^n \end{vmatrix} = \prod_{i<j} (x_j - x_i) \neq 0 .$$

Поскольку все x_i различны, определитель Δ отличен от нуля, и, следовательно, система имеет единственное решение. Отсюда вытекает существование и единственность интерполяционного многочлена.

Погрешность интерполяции.

Обозначим

$R_n(x) = f(x) - P_n(x)$ и будем искать ее оценку.

Пусть $f(x) \in C^{n+1}[a,b]$. Положим $R_n(x) = \omega(x) r(x)$,

где $\omega(x) = (x - x_0)(x - x_1) \cdot ... \cdot (x - x_n)$.

Зафиксируем произвольную точку x, отличную от узлов интерполяции x_i, $i = \overline{0,n}$, и построим вспомогательную функцию:

$$F(t) = P_n(t) + \omega(t) r(x) - f(t), \qquad a \le t \le b \ . \tag{7.2}$$

Очевидно, $F(x) = 0$ и, кроме того $F(x_k) = 0, \qquad k = \overline{0,n}$.

Таким образом, функция $F(t)$ имеет по крайней мере $(n+2)$ нуля на отрезке $[a,b]$. Применим теорему Ролля, по которой между каждой парой нулей функции находится по крайней мере один нуль производной этой функции.

Тогда производная $F'(t)$ имеет по крайней мере $(n+1)$ нулей на данном интервале (a,b). Продолжая рассуждение, получим в итоге, что $F^{(n)}(t)$ имеет, по крайней мере, два нуля, а $F^{(n+1)}(t)$ − один нуль в некоторой точке ξ на (a,b).

Продифференцируем равенство (7.2) $(n+1)$ раз и подставим $t = \xi$. Получим

$$F^{(n+1)}(\xi) = (n+1)! \cdot r(x) - f^{(n+1)}(\xi) = 0.$$

Откуда $r(x) = \dfrac{f^{(n+1)}(\xi)}{(n+1)!}$.

Тогда

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

где $\xi \in [a,b]$ (очевидно формула напоминает остаток формулы Тейлора в форме Лагранжа). В итоге имеем оценку погрешности интерполяции:

$$\left|R_n(x)\right| \le \frac{M_{n+1}}{(n+1)!}\left|\omega(x)\right|, \qquad \text{где} \qquad M_{n+1} = \max_{a \le x \le b}\left|f^{(n+1)}(x)\right|.$$

Интерполяционный многочлен Лагранжа

Пусть даны узлы на отрезке $[a,b]$, $a \le x_0 < x_1 < \ldots < x_n \le b$, и значения функции *F(x)* в узлах

$$f(x_i) = y_i, \qquad i = \overline{0,n}.$$

Пусть $\omega(x) = (x - x_0)(x - x_1) \cdot \ldots \cdot (x - x_n)$,

$\omega_j(x) = (x - x_0) \cdot \ldots \cdot (x - x_{j-1})(x - x_{j+1}) \cdot \ldots \cdot (x - x_n)$,

т. е. $\omega_j(x) = \dfrac{\omega(x)}{x - x_j}$.

Положим $l_j(x) = \dfrac{\omega_j(x)}{\omega_j(x_j)}$,

т. е. $$l_j(x) = \frac{(x - x_0) \cdot \ldots \cdot (x - x_{j-1})(x - x_{j+1}) \cdot \ldots \cdot (x - x_n)}{(x_j - x_0) \cdot \ldots \cdot (x_j - x_{j-1})(x_j - x_{j+1}) \cdot \ldots \cdot (x_j - x_n)}.$$

Очевидно $l_j(x_i) = \begin{cases} 0, & \text{при } i \ne j \\ 1, & \text{при } i = j. \end{cases}$

Построим многочлен $L_n(x) = \sum_{j=0}^{n} l_j(x) y_j$.

Легко видеть, что $L_n(x_i) = l_i(x_i)y_i = 1 \cdot y_i = y_i$, $i = \overline{0,n}$, т.е. это интерполяционный многочлен. Его называют интерполяционным многочленом Лагранжа.

Пример. Рассмотрим задачу интерполяции для функции

$f(x) = \sin\frac{\pi}{2}x$, на $[0,1]$.

Выберем в качестве узлов точки $x_0 = 0$, $x_1 = 1/3$, $x_2 = 1$. Тогда значения функции: $y_0 = 0$, $y_1 = 1/2$, $y_2 = 1$.

Получим

$$L_2(x) = \frac{(x-1/3)\cdot(x-1)}{(-1/3)\cdot(-1)} + \frac{x\cdot(x-1)\cdot\frac{1}{2}}{1/3\cdot(-2/3)} + \frac{(x-1/3)\cdot x}{2/3\cdot 1} = -3/4\cdot x^2 + 7/4\cdot x .$$

Оценим погрешность. Поскольку можно показать, что $|\omega(x)| \le 0{,}079$, то $R_2(x) \le \frac{\pi^3}{3!\cdot 8}\max_{0\le x\le 1}|\varpi(x)| \le \frac{\pi^3}{3!\cdot 8}\cdot 0{,}079$.

Линейная интерполяция

Пусть *n=1*, т. е. даны два узла x_0, x_1 справа и слева от точки x:

$x_0 \le x \le x_1$.

Построим интерполяционный многочлен первой степени по этим узлам.

Значения функции *f(x)* в этих узлах y_0, y_1.

Получаем:

$$L_1(x) = \frac{x-x_1}{x_0-x_1}\cdot y_0 + \frac{x-x_0}{x_1-x_0}\cdot y_1 = y_0 + \frac{y_1-y_0}{x_1-x_0}\cdot(x-x_0).$$

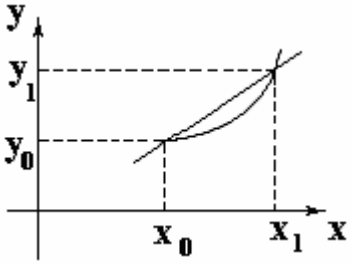


Рис. 7.1.

т. е. графически интерполяционный многочлен представляет собой хорду, соединяющую точки (x_0, y_0) и (x_1, y_1) (рис. 7.1).

Оценим погрешность линейной интерполяции.

Пусть $h = x_1 - x_0$.

Тогда $\max_{x_0\le x\le x_1}|\omega(x)| = \max|(x-x_0)\cdot(x-x_1)| = \frac{h^2}{4}$,

так как функция $|\omega(x)|$ достигает максимума на $[x_0, x_1]$ в точке $x_m = \frac{x_0+x_1}{2}$.

(рис. 7.2).

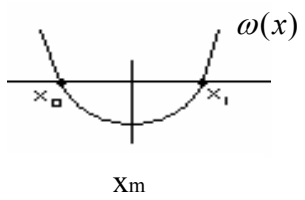


Рис.7.2.

Обозначим $M_2 = \max\limits_{x_0 \le x \le x_1} |f''(x)|$,

тогда $|R_n(x)| \le \dfrac{M_{n+1}}{(n+1)!} \max|\omega(x)| \le M_2 \dfrac{h^2}{8}$,

т. е. $|R_1(x)| \le \dfrac{M_2}{8} h^2$ в случае линейной интерполяции.

Пример. Рассмотрим функцию

$f(x) = \lg x$ на отрезке $[0, 1]$.

Пусть $h = 10^{-3}$ − расстояние между узлами. Оценим погрешность линейной интерполяции. Получим

$$M_2 = \max\left| -\frac{1}{x^2} \lg e \right| = \lg e = 0{,}4243,$$

следовательно,

$$|R_1(x)| \le \frac{M_2}{8} h^2 = \frac{0{,}4243}{8} \cdot 10^{-6} \approx 6 \cdot 10^{-8}.$$

Интерполяционный многочлен Ньютона

Пусть $x_0, x_1, ..., x_n$ − набор узлов интерполирования,

$y_0, y_1, ..., y_n$ − значения функции $f(x)$ в узлах.

Величину $\Delta y_k = y_{k+1} - y_k$ называют конечной разностью первого порядка в *к*-м узле.

Аналогично определяются конечные разности высших порядков.

$\Delta^2 y_k = \Delta y_{k+1} - \Delta y_k = y_{k+2} - y_{k+1} - (y_{k+1} - y_k) = y_{k+2} - 2y_{k+1} + y_k$

.....

$$\Delta^i y_k = \Delta^{i-1} y_{k+1} - \Delta^{i-1} y_k = \sum_{i=0}^{n} (-1)^{n-i} C_n^i y_{k+i} \; \Delta^i y_k = \Delta^{i-1} y_{k+1} - \Delta^{i-1} y_k = \sum_{i=0}^{n} (-1)^{n-i} C_n^i y_{k+i}.$$

Конечные разности обычно считают по схеме:

x_i	y_i	$\Delta\, y_i$	$\Delta^2\, y_i$	$\Delta^3\, y_i$
x_0	y_0	$\Delta y_0 = y_1 - y_0$	$\Delta^2 y_0 = \Delta y_1 - \Delta y_0$	$\Delta^3 y_0 = \Delta^2 y_1 - \Delta^2 y_0$
x_1	y_1	$\Delta y_1 = y_2 - y_1$	$\Delta^2 y_1 = \Delta y_2 - \Delta y_1$	
x_2	y_2	$\Delta y_2 = y_3 - y_2$		
x_3	y_3			

Разделенной разностью первого порядка называется выражение

$$f_1(x_k, x_{k+1}) = \frac{y_{k+1} - y_k}{x_{k+1} - x_k} = \frac{\Delta y_k}{\Delta x_k}.$$

Разделенной разностью второго порядка называется выражение

$$f_2(x_k, x_{k+1}, x_{k+2}) = \frac{f_1(x_{k+1}, x_{k+2}) - f_1(x_k, x_{k+1})}{x_{k+2} - x_k}$$ и т. д.

Пусть x – любая точка отрезка, не совпадающая с узлами. Тогда

$$f_1(x, x_0) = \frac{y_0 - f(x)}{x_0 - x},$$

откуда $f(x) = y_0 + f_1(x, x_0)(x - x_0)$. (7.3)

Далее $f_2(x, x_0, x_1) = \frac{f_1(x_0, x_1) - f_1\ (x, x_0)}{x_1 - x}$,

откуда $f_1(x, x_0) = f_1(x_0, x_1) + f_2\ (x, x_0, x_1)(x - x_1)$.

Подставляя в (7.3), получаем

$$f(x) = y_0 + f_1(x_0, x_1)(x - x_0) + f_2\ (x, x_0, x_1)(x - x_0)(x - x_1). \quad (7.4)$$

Далее $f_3(x, x_0, x_1, x_2) = \frac{f_2(x_0, x_1, x_2) - f_2\ (x, x_0, x_1)}{x_2 - x}$,

откуда $f_2(x, x_0, x_1) = f_2(x_0, x_1, x_2) + f_3\ (x, x_0, x_1, x_2)(x - x_2)$.

Подставляя в (4), имеем:

$$f(x) = y_0 + f_1(x_0, x_1)(x - x_0) + f_2\ (x, x_0, x_2)(x - x_0)(x - x_1) + \\ + f_3\ (x, x_0, x_1, x_2)(x - x_0)(x - x_1)(x - x_2). \quad (7.5)$$

Продолжая процесс, получим:

$$f(x) = N_n(x) + f_{n+1}(x, x_0, ..., x_n)(x - x_0)...(x - x_n),$$

где $N_n(x) = y_0 + f_1(x_0, x_1)(x - x_0) + ... + f_n\ (x_0, ..., x_n)(x - x_0)...(x - x_{n-1})$.

Очевидно, при $x = x_i$, $\forall i = \overline{0, n}$, $f(x_i) = N_n(x_i)$, $i = \overline{0, n}$,

т. е. $N_n(x)$ − интерполяционный многочлен. Его называют интерполяционным многочленом Ньютона.

Достоинство интерполяционного многочлена Ньютона: он удобен при расширении интерполяции и добавлении узлов.

Недостаток: в какой-то степени он сложнее в подсчете конечных разностей по сравнению с многочленом Лагранжа.

Интерполяционный многочлен Ньютона - Грегори

Рассмотрим случай задачи интерполяции с равноотстоящими узлами, т. е. пусть

$h = x_{i+1} - x_i$, для всех $i = \overline{0, n}$.

Будем искать интерполяционный многочлен Ньютона в форме

$$N(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + ... + a_n(x - x_0) \cdot ... \cdot (x - x_{n-1}),$$

где коэффициенты многочлена не определены.

Используем условие

$N(x_i) = y_i, \qquad i = \overline{0,n}$.

Получим:

$N(x_0) = y_0 = a_0$

$N(x_1) = y_1 = a_0 + a_1 h$

$N(x_2) = y_2 = a_0 + 2ha_1 + 2h^2 a_2$

..... .

Откуда $a_0 = y_0$,

$$a_1 = \frac{y_1 - y_0}{h} = \frac{\Delta y_0}{h},$$

$$y_2 = y_0 + 2h\frac{y_1 - y_0}{h} + 2h^2 a_2,$$

$$a_2 = \frac{y_2 - y_0 - 2(y_1 - y_0)}{2h^2} = \frac{y_2 - 2y_1 + y_0}{2h^2} = \frac{\Delta^2 y_0}{2h^2}.$$

Продолжая, можем по индукции получить формулу

$$a_k = \frac{\Delta^k y_0}{k!\ h^k}, \qquad k = 1,...,n.$$

В итоге получаем интерполяционный многочлен Ньютона - Грегори:

$$N(x) = y_0 + \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{2!h^2}(x - x_0)(x - x_1) + ... + \frac{\Delta^n y_0}{n!h^n}(x - x_0)(x - x_1)...(x - x_{n-1}).$$

Пример. Пусть требуется найти интерполяционный многочлен для функции $f(x)$, имеющей в узлах $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$ значения $y_0 = 5$, $y_1 = 3$, $y_2 = 2$, $y_3 = 4$, $y_4 = 6$. Вычислим конечные разности:

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\Delta^4 y_i$
0	*5*	*-2*			
1	*3*	*-1*	*1*	*2*	
2	*2*	*2*	*3*	*-3*	*-5*
3	*4*	*2*	*0*		
4	*6*				

Подставляя их значения в формулу для интерполяционного многочлена Ньютона - Грегори, в итоге получаем

$$N(x) = 5 - 2x + 0{,}5x(x-1) + \frac{1}{3}x(x-1)(x-2) - \frac{5}{24}x(x-1)(x-2)(x-3).$$

7.2. Аппроксимация по средне квадратичному отклонению

Пусть есть пространство непрерывных функций $C_{[a,b]}$.
Введем в нем скалярное произведение и новую норму

$$(f,g)=\int_a^b f(x)g(x)dx,$$

$$\|f\|=\sqrt{\int_a^b f^2(x)dx}\quad .$$

Система функций

$$f_1,...,f_n \tag{7.6}$$

называется линейно-независимой, если равенство $\alpha_1 f_1(x)+...+\alpha_n f_n(x)=0$ возможно, тогда и только тогда, когда $\alpha_1=...=\alpha_n=0$. В противном случае система функций называется линейно зависимой.

Известно, что система попарно-ортогональных ненулевых функций всегда линейно независима. Чтобы найти критерий линейной независимости в общем случае, построим определитель, состоящий из скалярных произведений функций:

$$\Gamma(f_{1,\ ...,}f_n)=\begin{vmatrix}(f_{1,}f_1) & (f_{1,}f_2) & ... & (f_{1,}f_n)\\ (f_{2,}f_1) & (f_{2,}f_2) & ... & (f_{2,}f_n)\\ ... & ... & ... & ...\\ (f_{n,}f_1) & (f_{n,}f_2) & ... & (f_{n,}f_n)\end{vmatrix}.$$

Определитель $\Gamma(f_{1,\ ...,}f_n)$ называется определителем Грамма.

Теорема 1: (Критерий линейной независимости). Для того чтобы система функций (7.6) была линейно независима, необходимо и достаточно, чтобы $\Gamma(f_1,...,f_n)\neq 0$.

Доказательство: Докажем утверждение равносильное теореме, т. е. докажем, что система (7.6) линейно зависима тогда и только тогда, когда $\Gamma(f_{1,\ ...,}f_n)=0$.

1) Необходимость. Пусть система линейно зависима, т. е. существуют α_1 , ... , α_n такие, что

$\alpha_1 f_1(x)+...+\alpha_n f_n(x)=0$, и $\alpha_1^2+...+\alpha_n^2>0$.

Будем последовательно умножать это тождество на $f_1,f_2,...,f_n$. Получим систему

$$\begin{cases}\alpha_1(f_{1,}f_1)+...+\alpha_n(f_{1,}f_n)=0\\ \alpha_1(f_{2,}f_1)+...+\alpha_n(f_{2,}f_n)=0\\ .....\\ \alpha_1(f_{n,}f_1)+...+\alpha_n(f_{n,}f_n)=0.\end{cases} \tag{7.7}$$

Это однородная система линейных уравнений относительно неизвестных коэффициентов $a_1, a_2, ..., a_n$ и ее определитель $\Delta = \Gamma(f_1, ..., f_n)$.

Поскольку система (7.7) имеет ненулевые решения, то $\Delta = 0$, т. е. $\Gamma(f_1, ..., f_n) = 0$.

2) Достаточность. Пусть $\Gamma(f_1, ..., f_n) = 0$. Из этого следует, что система (7.7) имеет ненулевые решения $\alpha_1, ..., \alpha_n$. Подставим эти решения в систему (7.7) и получим систему тождеств. Перепишем систему в виде

$$\begin{array}{l|l} (f_1, \alpha_1 f_1 + ... + \alpha_n f_n) = 0 & \alpha_1 \\ (f_2, \alpha_1 f_1 + ... + \alpha_n f_n) = 0 & \alpha_2 \\ ..... & ... \\ (f_n, \alpha_1 f_1 + ... + \alpha_n f_n) = 0 & \alpha_n . \end{array}$$

и умножим равенства последовательно на α_i, а затем просуммируем:

$$(\sum_{i=1}^{n} \alpha_i f_i, \sum_{j=1}^{n} \alpha_j f_j) = 0 .$$

Последнее означает, что $(g(x), g(x)) = 0,$

где $g(x) = \alpha_1 f_1 + ... + \alpha_n f_n$.

Но тогда, поскольку функция g(x) непрерывна,

$$g(x) = \sum_{i=0}^{n} \alpha_i f_i(x) = 0$$

при $\alpha_1^2 + ... + \alpha_n^2 > 0$, т. е. система функций (7.6) линейно зависима.

Теорема доказана.

Рассмотрим функцию *f(x)* на отрезке [*a,b*]. Пусть

$f_1(x), f_2(x), ..., f_n(x)$ − линейно независимые непрерывные функции.

Построим их линейную комбинацию $T_n(x) = \alpha_1 f_1(x) + ... + \alpha_n f_n(x)$, называемую обобщенным многочленом по системе функций $f_1, f_2, ..., f_n$.

Ставится задача: найти такие коэффициенты $\alpha_1, ..., \alpha_n$ обобщенного многочлена, чтобы выполнялось условие:

$$\|f(x) - T_n(x)\| = \min \|f(x) - T_n(x)\|,$$

где минимум берется по всевозможным значениям $\alpha_1, ..., \alpha_n$ и

$$\|f(x) - T_n(x)\| = \sqrt{\int_b^a [f(x) - T_n(x)]^2 dx} .$$

Такой обобщенный многочлен называется многочленом наилучшего средне квадратичного отклонения.

Теорема 2. Решение задачи аппроксимации функции по средне квадратичному отклонению существует и единственно.

Доказательство. Рассмотрим функцию от $\alpha_1, ..., \alpha_n$.

$$Q(\alpha_1, ..., \alpha_n) = \|f(x) - T_n(x)\|^2 =$$

$$= (f - T_n(x), f - T_n(x)) = (f - \sum_{i=1}^{n} \alpha_i f_i, f - \sum_{j=1}^{n} \alpha_j f_j) =$$

$$= (f, f) - 2\sum_{i=0}^{n} \alpha_i (f, f_i) + \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j (f_i, f_j).$$

Очевидно, Q $(\alpha_1, ..., \alpha_n)$ принимает наименьшее значение тогда и только тогда, когда $T_n(x)$ − наилучшее приближение в средне квадратичном для функции *f(x)*. Но для того чтобы Q достигло минимума по $\alpha_1, ..., \alpha_n$, необходимо, чтобы

$$\frac{\partial Q}{\partial \alpha_1} = -2(f, f_1) + 2\sum_{i=1}^{n} \alpha_i (f_i, f_1) = 0,$$

$$\frac{\partial Q}{\partial \alpha_2} = -2(f, f_2) + 2\sum_{i=1}^{n} \alpha_i (f_i, f_2) = 0,$$

.....

$$\frac{\partial Q}{\partial \alpha_n} = -2(f, f_n) + 2\sum_{i=1}^{n} \alpha_i (f_i, f_n) = 0.$$

Перепишем систему в виде следующей системы, называемой нормальной системой:

$$\begin{cases} \alpha_1 (f_1, f_1) + \alpha_2 (f_1, f_2) + ... + \alpha_n (f_1, f_n) = (f, f_1) \\ \alpha_1 (f_2, f_1) + \alpha_2 (f_2, f_2) .. + \alpha_n (f_2, f_n) = (f, f_2) \\ ..... \\ \alpha_1 (f_n, f_1) + \alpha_2 (f_n, f_2) .. + \alpha_n (f_n, f_n) = (f, f_n). \end{cases}$$

Ее определитель $\Delta = \Gamma(f_1, ..., f_n) \neq 0$, т. к. система функций $(f_1, ..., f_n)$ линейно независима. Но тогда нормальная система имеет единственное решение $\alpha_1, ..., \alpha_n$.

Убедимся, что $\frac{\partial^2 Q}{\partial \alpha^2} > 0$, т. е. выполнены достаточные условия минимума. Очевидно,

$$\frac{\partial^2 Q}{\partial \alpha^2} = [\Gamma(f_1, ..., f_n)] = \begin{pmatrix} (f_1, f_1) & (f_1, f_2) & ... & (f_1, f_n) \\ (f_2, f_1) & (f_2, f_2) & ... & (f_2, f_n) \\ ... & ... & ... & ... \\ (f_n, f_1) & (f_n, f_2) & ... & (f_n, f_n) \end{pmatrix} - \text{матрица Грамма.}$$

Матрица положительно определена, когда положительно определена соответствующая ей квадратичная форма.

Квадратичная форма $\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j (f_i, f_j)$, построенная по данной матрице, называется квадратичной формой Грамма.

Но

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j (f_i, f_j) = (\sum_{i=1}^{n} \alpha_i f_i, \sum_{j=1}^{n} \alpha_j f_j) = \left\| \sum_{i=1}^{n} \alpha_i f_i \right\|^2 \geq 0,$$

причем, поскольку функции $f_1,...,f_n$ линейно независимы, квадратичная форма равна нулю только тогда, когда все $\alpha_1,...,\alpha_n$ нулевые.

Следовательно, решение нормальной системы доставляет минимум функции $Q(\alpha_1,...,\alpha_n)$.

Теорема доказана.

Следствие. Чтобы численно решить задачу построения среднеквадратичного многочлена, надо составить и решить нормальную систему, а ее решение взять в качестве коэффициентов обобщенного многочлена.

Пример. Пусть $f(x)=\sqrt{x},\ x\in[0,\ 1]$. Построим многочлен наилучшего средне квадратичного отклонения по системе линейно независимых функций: *1, x.* Обозначим его $T_2(x)=a+b\cdot x.$

Получаем:

$$[\Gamma(1,x)]=\begin{pmatrix}(1,1) & (1,x)\\ (x,1) & (x,x)\end{pmatrix}=\begin{pmatrix}1 & 1/2\\ 1/2 & 1/3\end{pmatrix},$$

$$(\sqrt{x},1)=\int_0^1\sqrt{x}dx=\frac{2}{3}x^{\frac{3}{2}}\Big|_0^1=\frac{2}{3},$$

$$(\sqrt{x},x)=\int_0^1\sqrt{x}\cdot xdx=\frac{2}{5}\ .$$

Записываем нормальную систему:

$$\begin{cases}a+\frac{1}{2}b=\frac{2}{3}\\ \frac{1}{2}a+\frac{1}{3}b=\frac{2}{5},\end{cases}$$

решая ее, находим:

$$a=\frac{4}{15},\qquad b=\frac{4}{5},\qquad T_2(x)=\frac{4}{15}+\frac{4}{5}x\ .$$

7.3. Аппроксимация методом наименьших квадратов

Пусть дана функция $f(x)$ на отрезке [*a, b*].

Разобьем отрезок с помощью узлов

$a\le x_0<x_1<...<x_n\le b$.

Пусть $y_0,y_1,...,y_n$ − значение функции *f(x)* в узлах.

Если n − большое число, то интерполяционный $L_n(x)$ − многочлен высокой степени. Зачастую неудобно использовать многочлены очень высокой степени. Очевидно, мы можем отказаться от использования части узлов и тем самым понизить степень интерполяционного многочлена, но тогда теряется

часть информации. Поэтому вместо интерполяционного многочлена будем искать многочлен $P_m(x)$ меньшей степени $(m<n)$, такой что сумма

$$\sum_{i=0}^{n}[f(x_i)-P_m(x_i)]^2$$

принимает наименьшее значение. Данный многочлен называется многочленом наилучшего приближения по методу наименьших квадратов.

Положим

$$P_m(x)=a_0x^m+\ldots+a_m$$

и будем искать решение задачи

$$S(a_0,\ldots,a_m)=\sum_{i=0}^{n}[a_0x_i^m+\ldots+a_{m-1}x_i+a_m-y_i]^2\to\min.$$

Приравнивая к нулю производные S, получим систему линейных уравнений для определения коэффициентов a_i:

$$\frac{\partial S}{\partial a_0}=2\sum_{i=0}^{n}[a_0x_i^m+\ldots+a_m-y_i]\cdot x_i^m=0$$

$$\frac{\partial S}{\partial a_1}=2\sum_{i=0}^{n}[a_0x_i^m+\ldots+a_m-y_i]\cdot x_i^{m-1}=0$$

.....

$$\frac{\partial S}{\partial a_{m-1}}=2\sum_{i=0}^{n}[a_0x_i^m+\ldots+a_m-y_i]\cdot x_i=0$$

$$\frac{\partial S}{\partial a_m}=2\sum_{i=0}^{n}[a_0x_i^m+\ldots+a_m-y_i]\cdot 1=0\ .$$

Отсюда получается

$$\begin{cases} a_0(\sum_{i=0}^{n}x_i^{2m})+a_1\sum_{i=0}^{n}x_i^{2m-1}+\ldots+a_m\sum_{i=0}^{n}x_i^m=\sum_{i=0}^{n}y_ix_i^m \\ a_0(\sum_{i=0}^{n}x_i^{2m-1})+\ldots+a_m\sum_{i=0}^{n}x_i^{m-1}=\sum_{i=0}^{n}y_ix_i^{m-1} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_0(\sum_{i=0}^{n}x_i^n)+\ldots+a_m\sum_{i=0}^{n}1=\sum_{i=0}^{n}y_i \end{cases}$$

– нормальная система для определения коэффициентов $a_0,a_1,\ldots,a_n$.

Когда $m\le n$, можно показать, что нормальная система имеет единственное решение, которое действительно дает минимальное значение для функции S. Получив решения нормальной системы $a_0,\ldots,a_n$, строим многочлен наилучшего приближения по методу наименьших квадратов.

В частном случае, когда $m=n$, многочлен $P_n(x)$ переходит в интерполяционный многочлен.

Для решения нормальной системы обычно используется следующая таблица:

i	x_i	x_i^2	……	x_i^{2m}	y_i	$y_i x_i$	……	$y_i x_i^m$
0	x_0	x_0^2		x_0^{2m}	y_0	$y_0 x_0$		$y_0 x_0^m$
1	x_1	x_1^2		x_1^{2m}	y_1	$y_1 x_1$		$y_1 x_1^m$
.	.	.		.	.	.		.
.	.	.		.	.	.		.
.	.	.		.	.	.		.
n	x_n	x_n^2		x_n^{2m}	y_n	$y_n x_n$		$y_n x_n^m$
	$\sum_{i=0}^{n} x_i$	$\sum_{i=0}^{n} x_i^2$	…	$\sum_{i=0}^{n} x_i^{2m}$	$\sum_{i=0}^{n} y_i$	$\sum_{i=0}^{n} y_i x_i$	…	$\sum_{i=0}^{n} y_i x_i^m$

7.4. Интерполяция сплайнами

Рассмотрим задачу интерполяции функции *f(x)* на отрезке [*a, b*]. Пусть мы имеем узлы $a = x_0 < x_1 < ... < x_n = b$ и значения функции $y_0, ..., y_n$ в данных узлах. Отрезок разбивается узлами на *n* элементарных отрезков $[x_{i-1}, x_i]$, где $h_i = x_i - x_{i-1}$ − длина элементарного отрезка, $i = \overline{1,n}$.

Сплайном называется функция *S(x)*, которая на каждом элементарном отрезке является многочленом и непрерывна на всем отрезке [*a, b*], вместе со своими производными до некоторого порядка.

Степенью сплайна называется наивысший порядок степени многочлена.

Дефектом сплайна называется разность между его степенью и наивысшим порядком непрерывной на [*a, b*] производной.

Пример. Рассмотрим функцию

$$S(x) = \begin{cases} x^2, & 0 \le x < 1 \\ -x^2 + 4x - 2, & 1 \le x < 2 \\ 2, & 2 \le x < 3 \\ \dfrac{x^3}{27} - x + 4, & 3 \le x < 4. \end{cases}$$

Очевидно, функция $S(x)$ является кубическим сплайном на отрезке [*0, 4*], так как она непрерывна в узловых точках.

Действительно,

$S(1-0) = S(1+0) = 1, \_ S(2-0) = S(2+0) = 2, \_ S(3-0) = S(3+0) = 2.$

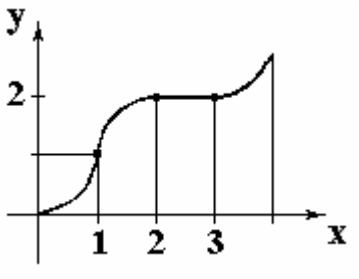


Рис. 7.3.

Найдем дефект сплайна.

$S'(1-0) = S'(1+0) = 2,\ S'(2-0) = S'(2+0) = 0,\ S'(3-0) = S'(3+0) = 0.$

В то же время $S''(2-0) = -2,\ S''(2+0) = 0.$

Таким образом, наибольший порядок непрерывной производной функции S на отрезке $[0,4]$ равен 1 и, следовательно, дефект сплайна равен 2. (См. рис. 7.3).

Отметим, что в общем случае сам сплайн многочленом не является. Чтобы он был многочленом, необходимо и достаточно, чтобы его дефект равнялся нулю.

Будем рассматривать кубические сплайны, у которых непрерывны первая и вторая производные.

Тогда на отрезке $[x_{i-1}, x_i]$ сплайн $S(x)$ имеет вид

$$S(x) = a_i + b_i(x - x_{i-1}) + c_i(x - x_{i-1})^2 + d_i(x - x_{i-1})^3, \quad i = \overline{1,n}.$$

Очевидно, $S(x_i) = y_i$, $i = \overline{0,n}$. Найдем $S(x)$. Для этого требуется определить значения $4n$ неизвестных коэффициентов. Очевидно, для этого необходимо иметь $4n$ уравнений для определения коэффициентов.

Подставим левый конец отрезка (x_{i-1}) в уравнение:

$$S(x_{i-1}) = y_{i-1} = a_i, \qquad i = \overline{1,n}$$

$$S(x_{i+1}) = y_i = a_i + b_i h_i + c_i h_i^2 + d_i h_i^3, \qquad i = \overline{1,n}.$$

В итоге получаем $2n$ уравнений:

$$\begin{cases} y_{i-1} = a_i & i = \overline{1,n} \\ a_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = y_i & i = \overline{1,n}. \end{cases}$$

Далее во всех внутренних узлах должны совпадать первая и вторая производные $S(x)$. Имеем

$$S'(x) = b_i + 2c_i(x - x_{i-1}) + 3d_i(x - x_{i-1})^2,$$

$$S''(x) = 2c_i + 6d_i(x - x_{i-1}), \qquad i = \overline{1,n-1}.$$

Приравниваем во внутренних узлах значения левых и правых производных. Получим:

$$\begin{cases} b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1} & i = \overline{1,n}, \\ c_i + 3d_i h_i = c_{i+1}, \end{cases}$$

т. е. $(2n-2)$ уравнений.

Недостающие два уравнения можно задать разными способами. Обычно берут $S''(x_0) = S''(x_n) = 0$.

Отсюда

$2c_1 = 0$, $\quad 2c_n + 6d_n h_n = 0$.

Для удобства положим еще $c_{n+1} = 0$.

Объединяя все уравнения, получим систему

$$\begin{cases} y_{i-1} = a_i & i = \overline{1,n} \\ a_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = y_i & i = \overline{1,n} \\ b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1} & i = \overline{1,n-1} \\ c_i + 3d_i h_i = c_{i+1} & i = \overline{1,n-1} \\ c_n + 3d_n h_k = 0 \\ c_1 = 0 \\ c_{n+1} = 0. \end{cases}$$

Решая систему, получим

$$\begin{cases} b_i h_i + c_i h_i^2 + d_i h_i^3 = y_i - y_{i-1} & i = \overline{1,n} \\ 2c_i h_i + 3d_i h_i^2 = b_{i+1} - b_i & i = \overline{1,n-1} \\ d =_i \dfrac{c_{i+1} - c_i}{3h_i} \\ c_1 = c_{n+1} = 0, \end{cases}$$

далее

$$\begin{cases} a_i = y_{i-1} & i = \overline{1,n} \\ b_i h_i + c_i h_i^2 + \left(\dfrac{(c_{i+1} - c_i) h_i^2}{3} \right) = y_i - y_{i-1} & i = \overline{1,n} \\ 2c_i h_i + (c_{i+1} - c_i) h_i = b_{i+1} - b_i & i = \overline{1,n-1} \\ d =_i \dfrac{c_{i+1} - c_i}{3h_i} & i = \overline{1,n} \\ c_1 = c_{n+1} = 0. \end{cases}$$

Откуда

$$b_i = \frac{y_i - y_{i-1}}{h_i} - c_i h_i - \frac{(c_{i+1} - c_i) h_i}{3}, \qquad i = \overline{1,n}.$$

$$2c_i h_i + (c_{i+1} - c_i)h_i = \frac{y_{i+1} - y_i}{h_{i+1}} - c_{i+1}h_{i+1} - \frac{(c_{i+2} - c_{i+1})h_{i+1}}{3} - \frac{y_i - y_{i-1}}{h_i} + c_i h_i + \frac{(c_{i+1} - c_i)h_i}{3}.$$

Таким образом, задача определения коэффициентов сплайна свелась к решению системы

$$c_i(\frac{h_i}{3}) + c_{i+1}(\frac{2}{3}h_i + \frac{2}{3}h_{i+1}) + c_{i+2}(\frac{h_{i+1}}{3}) = \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}, \qquad i = \overline{1, n-1}$$

$$c_1 = c_{n+1} = 0.$$

Система трехдиагональна. Будем решать ее методом прогонки. Поскольку для матрицы системы выполнено условие доминирования диагональных элементов

$$\frac{2}{3}h_i + \frac{2}{3}h_{i+1} > \frac{h_i}{3} + \frac{h_{i+1}}{3},$$

то задача имеет решение, причем единственное, и это решение можно найти методом прогонки.

8. ЧИСЛЕННОЕ ДИФФЕРЕНЦИРОВАНИЕ И ИНТЕГРИРОВАНИЕ

8.1. Численное дифференцирование

Пусть требуется найти численные значения y'_k производной функции *f(x)* в узлах $x_0 < x_1 < ... < x_n$ отрезка [*a, b*], в которых известны значения $y_0, y_1, ..., y_n$ функции. Рассмотрим несколько случаев, в зависимости от того, сколько раз дифференцируема исходная функция.

1) Всегда можно воспользоваться простейшей формулой $y'_k \approx \frac{y_{k+1} - y_k}{x_{k+1} - x_k}$.

2) Пусть функция *f(x)* дважды дифференцируема на отрезке [*a, b*] и узлы равноудалены друг от друга: $x_k = x_0 + kh$, $k = 0,...,n$. Разложим $f(x_{k+1})$ в точке x_k по формуле Тейлора

$$y_{k+1} = y_k + y'_k h + \frac{f''(\xi)}{2} h^2$$

и получим

$$y'_k = \frac{y_{k+1} - y_k}{h} - \frac{f''(\xi)}{2} h$$

или

$$y'_k = \frac{y_{k+1} - y_k}{h}.$$

При этом оценка погрешности вычислений имеет вид $R \le \frac{M_2 h}{2}$, где M_2 – максимальное на отрезке [*a, b*] значение второй производной функции *f(x)*. Таким образом, точность метода *O(h)*.

Теорема о среднем. Пусть *f(x)* – непрерывная функция на отрезке [*a, b*]. Тогда для любых точек $x_1 ... x_n$ этого отрезка справедлива формула

$$\frac{f(x_1) + ... + f(x_n)}{n} = f(\xi), \quad \text{где } \xi \in [a,b].$$

Доказательство. Пусть $m = \min_{a \le x \le b} f(x)$ – минимальное и $M = \max_{a \le x \le b} f(x)$ – максимальное значения функции на заданном отрезке. Тогда справедливы неравенства вида $m \le f(x_i) \le M, \quad i = \overline{1,n}$. Просуммируем их и разделим на *n*. Получим следующую оценку:

$$m \le \frac{f(x_1) + ... + f(x_n)}{n} \le M.$$

Тогда в силу теоремы Больцано - Коши о промежуточном значении непрерывной функции найдется такая точка $\xi \in [a,b]$, в которой будет выполняться равенство $\frac{f(x_1) + ... + f(x_n)}{n} = f(\xi)$.

Теорема доказана.

Пусть теперь функция *f(x)* – трижды непрерывно дифференцируема, а отрезок [*a, b*] разбит с шагом *h* точками $x_0, x_1 \ldots x_n$, в которых функция принимает значения $y_0, y_1 \ldots y_n$ соответственно. Возьмем один из внутренних узлов, например x_1, и оценим предыдущее и последующее значения функции:

$$y_0 = y_1 - y_1' h + \frac{y_1'' h^2}{2} - \frac{f'''(\xi_1)}{6} h^3,$$

$$y_2 = y_1 + y_1' h + \frac{y_1'' h^2}{2} + \frac{f'''(\xi_2)}{6} h^3.$$

Вычитая из второго равенства первое, получим

$$\frac{y_2 - y_0}{2h} = y_1' + \frac{f'''(\xi_1) + f'''(\xi_2)}{2} \frac{h^2}{6}$$

и, используя доказанную выше теорему о среднем, можем записать

$$y_1' = \frac{y_2 - y_0}{2h} - \frac{f'''(\xi) h^2}{6},$$

где ξ, ξ_1, ξ_2 - некоторые точки отрезка [*a, b*] .

Таким образом, $y_1' = \frac{y_2 - y_0}{2h}$ или, в общем виде

$y_k' = \frac{y_{k+1} - y_{k-1}}{2h}$, где $k = \overline{1, n-1}$.

При этом оценка погрешности вычисления имеет вид $R \le \frac{M_3}{6} h^2$, т.е. точность метода имеет порядок $O(h^2)$.

Пусть теперь функция *f(x)* – четыре раза непрерывно дифференцируема. Тогда справедливы равенства

$$y_0 = y_1 - y_1' h + y_1'' \frac{h^2}{2} - y_1''' \frac{h^3}{6} + \frac{f^{(4)}(\xi_1)}{24} h^4$$

и

$$y_2 = y_1 + y_1' h + y_1'' \frac{h^2}{2} + y_1''' \frac{h^3}{6} + \frac{f^{(4)}(\xi_1)}{24} h^4 \quad ,$$

из которых следует

$$\frac{y_2 - 2y_1 + y_0}{h^2} = y_1'' + \frac{f^{(4)}(\xi_1) + f^{(4)}(\xi_2)}{2} \frac{h^2}{12} \quad .$$

Применяя теорему о среднем и обозначая среднее арифметическое двух производных 4-го порядка в последнем равенстве через $f^{(4)}(\xi)$, получим

$$y_1'' = \frac{y_2 - 2y_1 + y_0}{h^2} + \frac{f^{(4)}(\xi)}{12} h^2,$$

где ξ, ξ_1, ξ_2 - некоторые точки отрезка [*a, b*] .

В общем виде получается

$$y_k'' \approx \frac{y_{k+1} - 2y_k + y_{k-1}}{h^2}, \text{ где } k = \overline{1, n-1}.$$

При этом погрешность будет составлять $R \le \frac{M_4}{12} h^2$, т.е. точность метода имеет порядок $O(h^2)$.

8.2. Формулы численного интегрирования

Пусть требуется вычислить определенный интеграл $\int_a^b f(x)dx$, где *f(x)* – некоторая заданная на отрезке [*a,b*] непрерывная функция.

Для простоты разобьем промежуток интегрирования точками, равноудаленными друг от друга: $a = x_0 < x_1 < ... < x_n = b$ так, что будет выполняться равенство $x_k = x_0 + kh, \quad k = \overline{0,n.}$, где $h = \frac{b-a}{n}$.

Рассмотрим несколько вариантов решения данной задачи.

1. **Формула прямоугольников.** Аппроксимируем площадь под графиком функции *f(x)* суммой прямоугольников с основанием *h* и высотой *f(ξ)*, где $x_k \le \xi \le x_{k+1}$. Причем, если взять $\xi = x_k \quad k = \overline{0,n-1}$, то получим формулу левых прямоугольников (см. рис.8.1):

$$\int_a^b f(x)dx \approx h(y_0 + ... + y_{n-1}) \approx \frac{b-a}{n}(y_0 + ... + y_{n-1}).$$

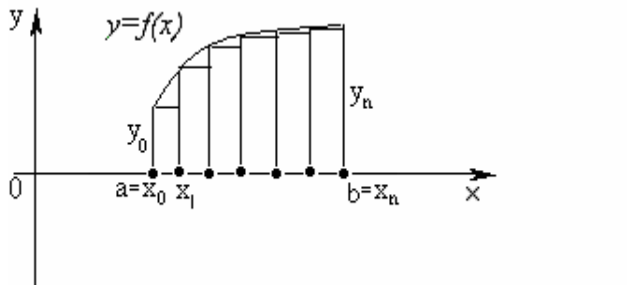


Рис. 8.1

А если взять $\xi = x_k \quad k = \overline{1,n}$, то получим формулу правых прямоугольников (см. рис.8.2):

$$\int_a^b f(x)dx \approx h(y_1 + ... + y_n) \approx \frac{b-a}{n}(y_1 + ... + y_n).$$

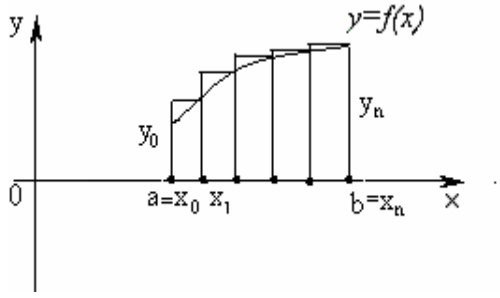


Рис. 8.2

В случае, когда мы берем среднюю точку $\xi = (x_{k-1} + x_k)/2 \quad k = \overline{1,n}$,

получаем формулу средних прямоугольников (рис. 8.3):

$$\int_a^b f(x)dx \approx h\,[f(\overline{x_1})+\ldots+f(\overline{x_n})], \qquad \overline{x_k}=\frac{x_{k-1}+x_k}{2}.$$

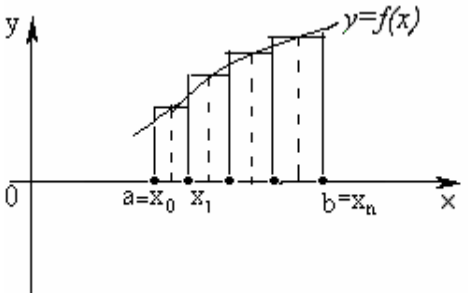


Рис. 8.3.

Оценим точность последней формулы.

Пусть $\int_{x_{K-1}}^{x_K} f(x)dx = F(x_k)-F(x_{k-1})$, где $F(x)=\int_{\overline{x_K}}^{x_K} f(t)dt$, а подынтегральная функция − трижды непрерывно дифференцируема. Тогда

$$F(\overline{x_k})=0,\ F'(\overline{x_k})=f(\overline{x_k}),\ F''(\overline{x_k})=f'(\overline{x_k}).$$

Запишем разложения функции F в точке $\overline{x_k}$ и точке $\overline{x_{k-1}}$:

$$F(x_k)=F(\overline{x_k})+F'(\overline{x_k})\frac{h}{2}+\frac{F''(\overline{x_k})}{2!}\frac{h^2}{4}+\frac{F'''(\xi_1)}{3!}\frac{h^3}{8}=$$

$$=0+\frac{f(\overline{x_k})h}{2}+\frac{f'(\overline{x_k})h^2}{8}+\frac{f''(\xi_1)h^3}{48},$$

$$F(x_{k-1})=-\frac{f(\overline{x_k})h}{2}+\frac{f'(\overline{x_k})h^2}{8}-\frac{f''(\xi_2)h^3}{48},$$

где $\xi_1,\ \xi_2 \in [x_{k-1},\ x_k]$.

Вычтем из первого равенства второе и получим:

$$\int_{x_{K-1}}^{x_K} f(t)dt = h\cdot f(\overline{x_k})+\frac{f''(\xi_1)+f''(\xi_2)}{2}\frac{h^3}{24}.$$

Используя теорему о среднем, можно записать

$$\int_{x_{K-1}}^{x_K} f(t)dt = h\cdot f(\overline{x_k})+\frac{f''(\xi_k)h^3}{24},$$

где ξ_k лежит на отрезке $[x_{k-1},\ x_k]$.

Таким образом, исходный интеграл равен

$$\int_a^b f(x)dx = h\,[f(\overline{x_1})+\ldots+f(\overline{x_n})]+\sum_{k=1}^{n}\frac{f''(\xi_k)h^3}{24}.$$

Отсюда легко получить оценку погрешности:

$$R=\left|\sum_{k=1}^{n}\frac{f''(\xi_k)h^3}{24}\right| \le \sum_{k=1}^{n}\frac{M_2h^3}{24}=\frac{M_2h^3}{24}\cdot n=\frac{M_2(b-a)h^2}{24}, \text{ где } h=\frac{b-a}{n}.$$

Таким образом, точность формулы средних прямоугольников имеет порядок $O(h^2)$.

2. **Формула трапеций.** Поступаем аналогично предыдущему способу, только аппроксимировать площадь под графиком функции *f(x)* будем трапециями. Площадь элементарной криволинейной трапеции приближенно равна $S_k \approx \frac{y_{k-1}+y_k}{2}h$, а интеграл -

$$\int_a^b f(x)dx \approx h\,(\frac{y_0+y_n}{2}+y_1+...+y_{n-1}).$$

Можно показать, что погрешность вычислений составляет $R \le \frac{M_2(b-a)^2}{12h^2} = \frac{M_2 h^2}{12}$, т.е. точность метода имеет порядок $O(h^2)$.

3. **Формула Симпсона или формула парабол.** Теперь аппроксимируем функцию на элементарном отрезке параболой. По сравнению с предыдущими способами вдвое уменьшим расстояние между узлами $h = \frac{b-a}{2n}$, тогда искомый интеграл будет равен $\int_a^b f(x)dx = \sum_{k=1}^{n}\int_{x_{2k-1}}^{x_{2k}} f(x)dx$.

Найдем коэффициенты *a,b,c* параболы, аппроксимирующей $f(x)$ на отрезке $[x_{2k}, x_{2k+2}]$. Для этого решим следующую систему:

$$\begin{cases} ax_{2k}^2 + bx_{2k} + c = y_{2k} \\ ax_{2k+1}^2 + bx_{2k+1} + c = y_{2k+1} \\ ax_{2k+2}^2 + bx_{2k+2} + c = y_{2k+2}. \end{cases} \qquad (8.1)$$

Так как главный определитель системы с точностью до знака совпадает с определителем Вандермонда, т.е.

$$\Delta = \pm\begin{vmatrix} 1 & 1 & 1 \\ x_{2k} & x_{2k+1} & x_{2k+2} \\ x_{2k}^2 & x_{2k+1}^2 & x_{2k+2}^2 \end{vmatrix} \neq 0,$$

то эта задача всегда имеет решение, причем единственное.

Посчитаем площадь параболической трапеции (рис. 8.4):

$$S = \int_{x_{2k}}^{x_{2k+2}} (ax^2+bx+c)dx.$$

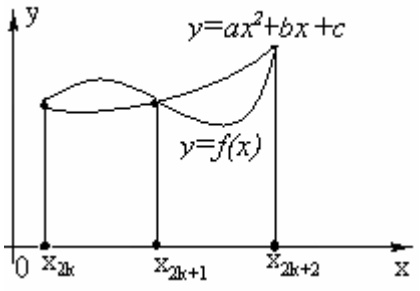


Рис. 8.4.

Возьмем для простоты начальный элементарный интервал $[x_0, x_2]$. Площадь не изменится, если мы сдвинем криволинейную трапецию по оси *Ox* и совместим ее начало с началом координат, т.е. иными словами положим $x_0 = 0,\ x_1 = h,\ x_2 = 2h.$ Тогда

$$S = \int_{x_0}^{x_2} (ax^2 + bx + c)dx = \frac{a8h^3}{3} + b2h^2 + c2h\,.$$

Перепишем систему (8.1) в виде:

$$\begin{cases} c = y_0 \\ ah^2 + bh + c = y_1 \\ 8ah^2 + 2bh + c = y_2, \end{cases}$$

и решим ее. Получим:

$$\begin{cases} c = y_0 \\ a = \dfrac{y_2 - 2y_1 + y_0}{6h^2} \\ b = \dfrac{8y_1 - 7y_0 - y_2}{6h}\,. \end{cases}$$

Тогда получается

$$S = \int_{x_0}^{x_2} (ax^2 + bx + c)dx = \frac{8ah^3}{3} + 2bh^2 + 2ch = 2h\left(\frac{4ah^2}{3} + bh + c\right) =$$

$$= 2h\left(\frac{4(y_2 - 2y_1 + y_0)}{18} + \frac{8y_1 - 7y_0 - y_2}{6} + \frac{6y_0}{6}\right) = \frac{h}{3}(y_0 + 4y_1 + y_2)\,.$$

Очевидно, аналогично

$$\int_{x_{2k}}^{x_{2k+1}} (ax^2 + bx + c)dx = \frac{h}{3}(y_{2k} + 4y_{2k+1} + y_{2k+2}).$$

Таким образом

$$\int_a^b f(x)dx = \sum_{k=1}^{n} \int_{x_{2k}}^{x_{2k+2}} f(x)dx \approx \frac{h}{3}[y_0 + y_{2n} + 2(y_2 + \ldots + y_{2n-2}) + 4(y_1 + \ldots + y_{2n-1})]\,.$$

Данная формула и называется формулой Симпсона. Можно показать, что погрешность формулы Симпсона $R \le \dfrac{M_4(b-a)^5}{2880n^4}$, и ее точность имеет порядок $O(h^4)$.

Таким образом, по сравнению с предыдущими методами формула Симпсона является существенно более точной.

8.3. Интерполяционные квадратурные формулы

Вычислим интеграл $\int_a^b f(x)dx$, заменяя подынтегральную функцию интерполяционным многочленом с узлами $x_k = x_0 + kh,\ k = \overline{0,n}$, где $x_0 = a,\ h = \dfrac{b-a}{n}$. Получим

$$\int_a^b f(x)dx \approx \int_a^b \sum_{k=0}^{n} l_k(x) y_k = \sum_{k=0}^{n} y_k \int_a^b l_k(x)dx = \sum_{k=0}^{n} y_k A_k\ ,$$

где $A_k = \int_a^b \dfrac{\omega_k(x)}{\omega_k(x_k)}dx$

и $\omega_k(x) = (x-x_0)\cdot\ldots\cdot(x-x_{k-1})\cdot(x-x_{k+1})\cdot\ldots\cdot(x-x_n)$.

Формулу

$$\int_a^b f(x)dx = \sum_{k=0}^{n} y_k A_k \qquad (8.2)$$

называют интерполяционной квадратурной формулой Лагранжа.

Заметим, что $\omega_k(x)$ зависит только от самих узлов, на которые разбит промежуток, и не зависит от функции $f(x)$. Следовательно, коэффициенты A_k не зависят от вида функции также и, используя эти коэффициенты, можно считать интегралы от различных функций. При этом, если наша функция является многочленом, то формула (8.2) является точной формулой.

Можно найти A_k при помощи метода неопределенных коэффициентов:

$$\begin{cases} \int_a^b 1\cdot dx = \sum_{k=0}^{n} 1\cdot A_k \\ \int_a^b x\cdot dx = \sum_{k=0}^{n} x_k\cdot A_k \\ \ldots\ldots\ldots\ldots\ldots \\ \int_a^b x^n\cdot dx = \sum_{k=0}^{n} x_k^n\cdot A_k . \end{cases}$$

Из данной системы можно найти $A_k,\ k = \overline{0,n}$.

Пример. Построить интерполяционную квадратурную форму для вычисления интеграла

$$\int_{-1}^{1} f(x)dx,$$

$x_0 = -1,\ x_1 = 0,\ x_2 = 1.$

Строим систему

$$\begin{cases} \int_a^b 1 \cdot dx = A_0 + A_1 + A_2 \\ \int_a^b x \cdot dx = -A_0 + A_2 \\ \int_a^b x^2 \cdot dx = A_0 + A_2. \end{cases}$$

Решая ее, находим

$$A_0 = -1/3, \quad A_1 = 4/3, \quad A_2 = 1/3.$$

Таким образом $\int_{-1}^{1} f(x)dx = -\frac{1}{3} \cdot f(-1) + \frac{4}{3} \cdot f(0) + \frac{1}{3} \cdot f(2).$

Недостатком интерполяционной квадратурной формулы Лагранжа является проблематичность оценки погрешности.

Наряду с интерполяционной квадратурной формулой Лагранжа для вычисления интегралов вида

$$\int_{-1}^{1} f(x)dx$$

часто применяется квадратурная формула Гаусса, в которой узлами разбиения служат корни многочлена

$$X_n(x) = \frac{1}{n!2^n} \cdot \frac{d^n(x^2-1)^{2n}}{dx^n}.$$

Отметим, что формула Гаусса точнее формулы Лагранжа.

Замечания.

1. Вычисление несобственных интегралов: $\int_a^{+\infty} f(x)dx = \lim_{A\to\infty} \int_a^A f(x)dx$.

Если подынтегральная функция разрывная, то аналогично: $\int_a^b f(x)dx = \lim_{\varepsilon\to\infty} \int_a^{b-\varepsilon} f(x)dx$ (b − точка разрыва).

2. Вычисление кратных интегралов. Пусть требуется вычислить $\iint_D f(x,y)dxdy$

по области D. Не ограничивая общности, считаем, что область D является правильной вдоль оси Oy, т.е. имеет вид

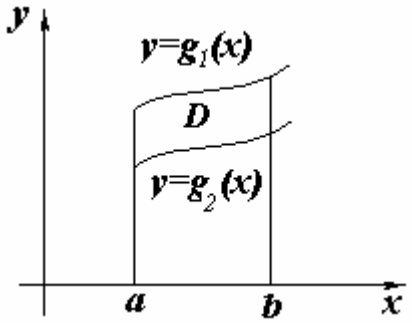


Рис. 8.5

Тогда

$$\iint_D f(x,y)dxdy = \int_a^b \left(\int_{g_1(x)}^{g_2(x)} f(x,y)dy \right) dx = \int_a^b \Phi(x)dx,$$

где $\Phi(x) = \int_{g_1(x)}^{g_2(x)} f(x,y)dy$.

Пусть $a = x_0, x_1, ..., x_n = b$ – точки разбиения. Тогда можно вычислить каждый из интегралов

$$\Phi(x_k) = \int_{g_1(x_K)}^{g_2(x_K)} f(x_k, y)dy,$$

а потом применить формулу Симпсона к интегрированию функции $\Phi(x)$.

9. ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧИ КОШИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Рассмотрим дифференциальное уравнение $y' = f(x,y)$ с начальным условием $y(x_0) = y_0$. Будем предполагать, что *f(x,y)* непрерывная и непрерывно дифференцируемая по y функция в окрестности замкнутой области

$$D = \{(x,y) \mid a \le x \le b, c \le y \le d\},$$

содержащей внутри себя точку (x_0, y_0).

Требуется решить задачу Коши: найти непрерывно дифференцируемую функцию *y=y(x)*, такую что $y'(x) = f(x, y(x))$ при всех $x \in [a,b]$ и $y(x_0) = y_0$.

Разобьем отрезок *[a, b]* с помощью точек разбиения $a = x_0, x_1, ..., x_n = b$ с шагом $h = (b-a)/n$. Тогда узлы разбиения имеют вид $x_k = x_0 + kh,\ k = \overline{0,n}$.

Пусть $y(x_0), y(x_1), ..., y(x_n)$ – значения функции в точках разбиения.

9.1. Метод ломаных Эйлера

Пусть $y = y(x)$ искомое решение задачи Коши. В точке (x_0,y_0) построим касательную (см. рис. 9.1) к графику $y = y(x)$.

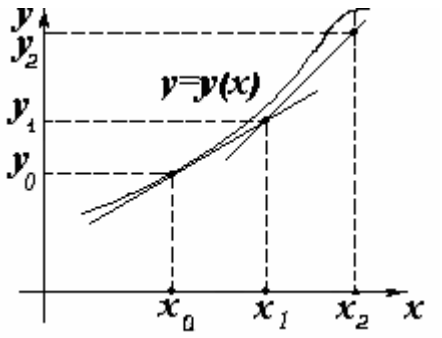


Рис. 9.1

Запишем уравнение касательной:

$$y = y_0 + y'(x_0)(x - x_0) = y_0 + f(x_0, y_0)(x - x_0).$$

и найдем точку пересечения этой касательной с прямой $x = x_1$:

$$y_1 = y_0 + hf(x_0, y_0).$$

Запишем уравнение прямой

$$y = y_1 + f(x_1, y_1)(x - x_1)$$

и найдем точку ее пересечения с прямой с $x = x_2$:

$$y_2 = y_1 + hf(x_1, y_1).$$

Продолжая процесс, получим рекуррентную последовательность:

$$y_{k+1} = y_k + hf(x_k, y_k), \qquad k = 0,1... \tag{9.1}$$

$$y_0 = y(x_0),$$

которую называют последовательностью Эйлера. Соединяя ломаными все точки (x_k, y_k), полученные из рекуррентной последовательности Эйлера, получим ломаную линию, приближающую график решения $y = y(x)$. Функция, график которой совпадает с ломаной Эйлера, принимается за приближенное решение задачи Коши.

Выясним точность метода Эйлера. Сравним значения точного решения y(x) задачи Коши в узловых точках со значениями, полученными методом Эйлера:

$$y(x_{k+1}) = y(x_k) + y'(x_k)h + O(h^2),$$

$$y_{k+1} = y_k + hf(x_k, y_k) + O(h^2).$$

Поскольку

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

то $y(x_{k+1}) - y_{k+1} = O(h^2)$ при условии, что $y_k = y(x_k)$. То есть, точность метода на отдельном отрезке $[x_k, x_{k+1}]$ совпадает с $O(h^2)$. Тогда, очевидно, точность метода Эйлера на всем отрезке [*a, b*] будет *O(h).*

Для повышения точности вычислений иногда используется модифицированный метод Эйлера, в котором рекуррентная последовательность Эйлера вычисляется по формулам

$$y_{k+1} = y_k + hf\left(x_k + \frac{h}{2}, y_k + \frac{h}{2} f(x_k, y_k)\right), \qquad k = 0,1...,n-1. \tag{9.2}$$

Модифицированный метод Эйлера обычно дает более точное приближение решения.

Пример. Пусть требуется решить задачу Коши:

$$\begin{cases} y' = -y, \ x \in [0,1] \\ y(0) = 1. \end{cases}$$

Полагая $h = 0{,}2$ и используя метод Эйлера, получим, как легко убедиться, из формулы Эйлера (9.1)

$$y_{k+1} = y_k + 0.2 \cdot (-y_k) = 0.8 \cdot y_k .$$

С другой стороны, используя модифицированный метод Эйлера, получим в силу формулы (2) рекуррентную последовательность

$$y_{k+1} = y_k + 0.2 \cdot (-y_k) = 0.82 \cdot y_k \ .$$

Поскольку точным решением задачи Коши, как легко проверить, является функция $y = e^{-x}$, можно сравнить точность обоих методов.

	0	1	2	3	4	5
x_k	0	0.2	0.4	0.6	0.8	1
y_k	1	0.8	0.64	0.572	0.4086	0.3277
$y_k^{модиф}$	1	0.82	0.6724	0.5514	0.4521	0.3708
e^{-x}	1	0.8187	0.6703	0.5488	0.4493	0.3679

Общепризнанным недостатком метода Эйлера является его не достаточно высокая точность. Несомненным достоинством метода Эйлера является его простота.

9.2. Методы Рунге-Кутта

1. Метод Рунге-Кутта второго порядка (или метод типа «предиктор-корректор»).

Метод состоит из двух этапов. Сначала находят по методу Эйлера грубое решение:

$$y^*_{\kappa+1} = y_k + hf(x_k, y_k) \ .$$

На следующем шаге это грубое решение сглаживается:

$$y_{k+1} = y_k + h\frac{f(x_k, y_k) + f(x_{k+1}, y^*_{k+1})}{2}, \qquad k = 0,1...,n-1 .$$

Выясним точность метода. Преобразуя y_{k+1}, получаем:

$$y_{k+1} = y_k + \frac{h}{2} f(x_k, y_k) + \frac{h}{2} f(x_k + h, y_k + h \cdot f(x_k, y_k)) =$$

$$= y_k + \frac{h}{2} f(x_k, y_k) + \frac{h}{2} f(x_k, y_k) + \frac{h}{2}[f_x'(x_k, y_k) +$$

$$+ h \cdot f_y'(x_k, y_k) \cdot f(x_k, y_k) + O(h^2)] =$$

$$= y_k + hf(x_k, y_k) + \frac{h^2}{2}[f_x'(x_k, y_k) + + f_y'(x_k, y_k) \cdot f(x_k, y_k)] + O(h^3).$$

С другой стороны, разложим точное решение *y(x)* по формуле Тейлора. Получим

$$y(x_{k+1}) = y(x_k + h) = y(x_k) + y'(x_k)h + y''(x_k)\frac{h^2}{2} + O(h^3) = y(x_k) + hf(x_k, y(x_k)) +$$

$$+ \frac{h^2}{2}[f_x'(x_k, y(x_k)) + f_y'(x_k, y(x_k)) f(x_k, y(x_k)) + O(h^3).$$

Полагая $y(x_k) = y_k$, получаем погрешность на отдельном шаге равную $O(h^3)$. Тогда на всем отрезке погрешность составит $O(h^2)$.

Достоинство метода: его точность превосходит точность метод Эйлера.

2. Метод Рунге-Кутта четвертого порядка.

На каждом шаге производится вычисление коэффициентов K_1, K_2, K_3, K_4:

$K_1 = hf(x_k, y_k)$;

$K_2 = hf(x_k + \frac{h}{2}, y_k + \frac{K_1}{2})$;

$K_3 = hf(x_k + \frac{h}{2}, y_k + \frac{K_2}{2})$;

$K_4 = hf(x_k + h, y_k + K_3)$.

Затем вычисляем

$$y_{k+1} = y_k + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4).$$

Данный метод имеет точность $O(h^4)$ на $[a,b]$.

Рассмотрим пример, который мы использовали для иллюстрации точности метода Эйлера.

Пример. Требуется решить задачу Коши:

$$\begin{cases} y' = -y \\ y(0) = 1 \end{cases}$$ на отрезке [0, 1].

Выберем шаг $h = 0{,}2$. Результат вычислений поместим в таблицу.

	0	1	2	3	4	5
x_k	0	0.2	0.4	0.6	0.8	1
y_k	1	0.8187	0.6703	0.5487	0.4493	0.3678
e^{-x}	1	0.8187	0.6703	0.5488	0.4493	0.3679

Таким образом, метод Рунге-Кутта 4-го порядка отличается очень высокой точностью. К определенным его недостаткам относится большая сложность и трудоемкость (на каждом шаге необходимо четырежды вычислять значения функции f вместо одного раза в методе Эйлера).

Отметим, что на практике выбирают начальную длину шага h таким образом, чтобы $h^4 < \varepsilon$, где ε − заданная точность вычисления решения. Затем шаг выбирают вдвое меньшим и останавливают вычисления, если разность полученных значений y_k со значениями, полученными при начальном выборе шага меньше ε. В противном случае шаг еще раз уменьшают вдвое и т.д.

9.3. Метод Адамса

1. Неявная схема метода Адамса.

Пусть есть дифференциальное уравнение

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

с начальным условием

$$y(x_0) = y_0.$$

Разбиваем отрезок $[a,b]$ с шагом h на n частей. То есть, получаем узлы $x_k = x_0 + kh$, $k = \overline{0,n}$, где $x_0 = a$.

Пусть $y = y(x)$ − решение. Тогда на $[x_k, x_{k+1}]$ справедливо равенство

$$y(x_{k+1}) = y(x_k) + \int_{x_K}^{x_{K+1}} f(x, y(x))dx.$$

Применим формулу левых прямоугольников для вычисления интеграла. Получим

$y_{k+1} = y_k + hf(x_k, y_k)$, то есть формулу Эйлера.

Очевидно это не самый точный метод вычисления интеграла.

Применим формулу трапеций для вычисления интеграла. Получим

$$y_{k+1} = y_k + h\frac{f(x_k, y_k) + f(x_{k+1}, y_{k+1})}{2}, \qquad k = 0,1\ldots.$$

Вычисление более точно, но мы не можем найти y_{k+1} из полученной формулы. Однако в частном случае, когда дифференциальное уравнение линейно, т. е. имеет вид

$$y' = -p(x)y + q(x),$$

мы получим:

$$y_{k+1} = y_k + h\frac{-p_k y_k + q_k - p_{k+1} y_{k+1} + q_{k+1}}{2},$$

где

$p_k = p(x_k)$, $q_k = q(x_k)$.

Отсюда легко находится значение

$$y_{k+1} = \frac{(2 - hp_k)y_k + h(q_k + q_{k+1})}{2 + hp_{k+1}}.$$

2. Явная схема Адамса

Используем интерполяционную квадратурную формулу Лагранжа для вычисления интеграла, т.е.

$\int_{x_K}^{x_{K+1}} f(x, y(x))dx = A_0 f(x_k, y_k) + A_1 f(x_{k-1}, y_{k-1})$, где

$$\int_a^b f(x)dx = \sum_{i=0}^{m} f(x_i)A_i, \quad A_i = \int_a^b l_i(x)dx; \quad l_i(x) = \frac{\varpi_i(x)}{\varpi_i(x_i)}.$$

Найдем коэффициенты A_i методом неопределенных коэффициентов:

$$\int_{x_K}^{x_{K+1}} dx = A_0 + A_1;$$

$$\int_{x_K}^{x_{K+1}} xdx = A_0 x_k + A_1 x_{k-1}.$$

Получаем систему двух уравнений с двумя неизвестными

$$\begin{cases} A_0 = h - A_1 \\ \dfrac{h(x_{k+1} + x_k)}{2} = A_0 x_k + A_1 x_{k-1}. \end{cases}$$

Откуда

$$\frac{h(x_{k+1} + x_k)}{2} = (h - A_1)x_k + A_1 x_{k-1};$$

$$\frac{h(x_{k+1} + x_k)}{2} = hx_k - A_1(x_k - x_{k-1});$$

$$\frac{h(x_{k+1} + x_k)}{2} = hx_k - A_1 h;$$

$$A_1 h = hx_k - \frac{h(x_{k+1} + x_k)}{2} = \frac{h(x_k - x_{k+1})}{2}.$$

В итоге получим:

$$A_1 = -\frac{h}{2};$$

$$A_0 = h - A_1 = \frac{3h}{2}.$$

Откуда

$$\int_{x_K}^{x_{K+1}} f(x, y(x))dx = \frac{3}{2}hf_k - \frac{h}{2}f_{k-1}, \quad \text{где} \quad f_k = f(x_k, y_k).$$

Следовательно, получим

$$y_{k+1} = y_k + h(\frac{3}{2}f(x_k, y_k) - \frac{1}{2}f(x_{k-1}, y_{k-1}) \qquad k = \overline{1, n}..$$

Это формула Адамса второго порядка.

Существенным недостатком метода Адамса второго порядка является то обстоятельство, что для его применения надо знать дополнительно к начальному условию еще

$y_{-1} = y(x_0 - h)$ или $y_1 = y(x_0 + h)$.

Достоинством метода является то, что значение функции f в каждой точке (x_k, y_k) вычисляется только один раз.

Замечания.

1. Если необходимо решить задачу Коши для системы дифференциальных уравнений:

$$\begin{cases} \dfrac{dx}{dt} = f(x,t), & x \in R^n. \\ x(t_0) = x_0, & t \in R^n. \end{cases}$$

можно использовать методы Эйлера или Рунге-Кутта.

2. Если решается задача Коши для уравнений высшего порядка

$y^{(n)} = f(x, y, y^1, ..., y^{(n-1)})$;

$y(x_0) = y_0$;

$y'(x_0) = y_0'$;

……

$y^{(n-1)}(x_0) = y_0^{n-1}$;

то задача сводится к решению задачи Коши для системы дифференциальных уравнений.

То есть, вводим новые переменные

$$\begin{cases} y_1 = y \\ y_2 = y' \\ ..... \\ y_{n-1} = y^{(n-2)} \\ y_n = y^{(n-1)}. \end{cases}$$

Откуда получается система дифференциальных уравнений

$$\begin{cases} y_1' = y_2 \\ y_2' = y_3 \\ ..... \\ y_{n-1}' = y_n \\ y_n' = f(x, y_1, .., y_n). \end{cases}$$

10. РЕШЕНИЕ КРАЕВОЙ ЗАДАЧИ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Будем рассматривать дифференциальное уравнение второго порядка.

$$y'' + p(x)y' + q(x)y = f(x), \qquad (10.1)$$

где $p(x)$, $q(x)$, $f(x)$ − заданные непрерывные на отрезке *[a, b]* функции.

Напомним, что задача Коши для уравнения (1) сводится к нахождению решения $y(x)$, удовлетворяющего начальным условиям:

$$\begin{cases} y(a) = A \\ y'(a) = A_1 . \end{cases}$$

Краевой задачей называется задача нахождения решения $y(x)$, удовлетворяющего граничным условиям:

$$\begin{cases} y(a) = A \\ y(b) = B . \end{cases} \qquad (10.2)$$

Краевая задача отличается от задачи Коши непредсказуемостью. Ее решение может существовать, не существовать, быть единственным, может быть бесконечно много решений.

Часто вместо граничных условий (10.2) используют обобщенные граничные условия:

$$\begin{cases} \alpha_1 y(a) + \beta_1 y'(a) = A \\ \alpha_2 y(b) + \beta_2 y'(b) = B. \end{cases} \qquad (10.3)$$

Граничные условия называются *однородными,* если A=B=0.

Соответственно, краевая задача называется *однородной,* если у нее однородные граничные условия и правая часть уравнения $f(x) \equiv 0$.

Следующая теорема имеет важное теоретическое значение.

Теорема. Краевая задача *(1)* , *(3)* имеет решение, причем единственное тогда и только тогда, когда соответствующая ей однородная краевая имеет только нулевое решение (тривиальное решение однородной краевой задачи).

Способы решения краевой задачи.

Поскольку достаточно хороших аналитических методов нет, то используются приближенные методы.

Система дважды непрерывно дифференцируемых функций $\varphi_0(x), \varphi_1(x), ..., \varphi_n(x)$ называется *базисной системой*, если выполняется:

1) $\varphi_0(x)$ удовлетворяет граничному условию (10.3),

2) функции $\varphi_1(x), ..., \varphi_n(x)$ − линейно независимы на *[a, b]* и удовлетворяют однородным граничным условиям.

По базисным функциям строят приближенное решение:

$$y_n(x) = \varphi_0(x) + a_1\varphi_1(x) + ... + a_n\varphi_n(x) .$$

Задача сводится к выбору коэффициентов $a_1, ..., a_n$ таких, чтобы функция $y_n(x)$ удовлетворяла граничному условию (10.3) и была в некотором смысле близкой к точному решению.

Поступают следующим образом. Выражение

$\psi(x, a_1, ..., a_n) = y_n''(x) + p(x)y_n'(x) + q(x)y_n(x) - f(x)$ называют невязкой.

Легко видеть, что, если бы $\psi(x, a_1, ..., a_n) \equiv 0$, то $y_n(x)$ было бы точным решением. К сожалению, так бывает очень редко. Следовательно, необходимо выбрать коэффициенты таким образом, чтобы невязка была в некотором смысле минимальной.

Метод коллокаций.

На отрезке *[a, b]* выбираются точки $x_1, ..., x_m \in [a, b]$ $(m \geq n)$, которые называются точками коллокации. Точки коллокации последовательно подставляются в невязку. Считая, что невязка должна быть равна нулю в точках коллокации, в итоге получаем систему уравнений для определения коэффициентов $a_1, ..., a_n$.

$$\begin{cases} \psi(x_1, a_1, ..., a_n) = 0 \\ ..... \\ \psi(x_m, a_1, ..., a_n) = 0. \end{cases} \qquad (10.4)$$

Обычно *m=n*. Получается система из *n* линейных уравнений с *n* неизвестными (коэффициентами $a_1, ..., a_n$):

$$\begin{cases} \psi(x_1, a_1, ..., a_n) = 0 \\ ..... \\ \psi(x_n, a_1, ..., a_n) = 0. \end{cases}$$

Решая (10.4), найдем приближенное решение $y_n(x)$. Для повышения точности расширяем базисную систему. Получаем более точное решение.

В значительной степени успех в применении метода зависит от удачного выбора базисной системы.

Пример. Пусть

$y'' + (1 + x^2)y = -1, \qquad -1 \leq x \leq 1,$

$y(-1) = 0, \quad y(1) = 0.$

Выберем базисную систему:

$\varphi_0(x) = 0,$

$\varphi_1(x) = 1 - x^2,$

$\varphi_2(x) = x^2(1 - x^2).$

Поскольку $\frac{\varphi_1}{\varphi_2} = \frac{1}{x^2} \neq const$, следовательно, функции $\varphi_1(x)$ и $\varphi_2(x)$ линейно независимы.

Строим приближенное решение

$y_2(x) = a_1(1 - x^2) + a_2(x^2 - x^4).$

Выберем точки коллокации:

$x_1 = -\frac{1}{2}, \qquad x_2 = 0, \qquad x_3 = \frac{1}{2}.$

Получаем систему уравнений

$$\begin{cases} \psi\left(-\frac{1}{2}, a_1, a_2\right) = \frac{17}{16}a_1 + \frac{49}{64}a_2 - 1 = 0 \\ \psi(0, a_1, a_2) = a_1 - 2a_2 - 1 = 0 \\ \psi\left(\frac{1}{2}, a_1, a_2\right) = \frac{17}{16}a_1 + \frac{49}{64}a_2 - 1 = 0. \end{cases}$$

Решая ее, получим

$y_2(x) = 0{,}957(1-x^2) - 0{,}022(x^2 - x^4)$.

Метод наименьших квадратов.

Как и в методе коллокаций приближенное решение строится по базисной системе. Для нахождения коэффициентов при базисных функциях минимизируется интеграл $I = \int_a^b \psi^2(x, a_1, ..., a_n)dx$.

Строим систему

$$\begin{cases} \frac{\partial I}{\partial a_1} = 2\int_a^b \psi(x, a_1, ..., a_n)\frac{\partial \psi(x, a_1, ..., a_n)}{\partial a_1}dx = 0 \\ ..... \\ \frac{\partial I}{\partial a_n} = 2\int_a^b \psi(x, a_1, ..., a_n)\frac{\partial \psi(x, a_1, ..., a_n)}{\partial a_n}dx = 0. \end{cases}$$

Решая ее, находим $a_1, ..., a_n$.

Синтезом метода наименьших квадратов и метода коллокаций является следующий метод. Выбирают $N>n$ точек и решают задачу

$S = \sum_{i=1}^{N} \psi^2(x_i, a_1, ..., a_n) \to \min$.

Для ее решения строится система

$$\begin{cases} \frac{\partial S}{\partial a_1} = 0 \\ ..... \\ \frac{\partial S}{\partial a_n} = 0. \end{cases}$$

Пример. Рассмотрим краевую задачу

$y'' + (1+x^2)y = -1, \qquad -1 \le x \le 1,$

$y(-1) = 0,$

$y(1) = 0.$

Выберем базисную систему:

$\varphi_0(x) = 0,$

$\varphi_1(x) = 1 - x^2,$

$\varphi_2(x) = x^2(1-x^2).$

Применяя метод наименьших квадратов, можно найти

$y_2(x) = 0{,}985(1-x^2) - 0{,}078(x^2 - x^4).$

Метод Галеркина.

По базисной системе строим приближенное решение

$y_n(x) = \varphi_0(x) + a_1\varphi_1(x) + ... + a_n\varphi_n(x)$.

Рассматриваем невязку $\psi(x, a_1, ..., a_n)$ и для определения коэффициентов при базисных функциях строим систему

$$\begin{cases} \int_a^b \psi(x, a_1, ..., a_n)\varphi_1(x)dx = 0 \\ ..... \\ \int_a^b \psi(x, a_1, ..., a_n)\varphi_n(x)dx = 0. \end{cases}$$

Решая данную систему, находим значение $a_1, ..., a_n$.

Пример. Рассмотрим краевую задачу

$y'' + y = x, \qquad 0 \le x \le 1$,

$y(0) = y(1) = 0$.

Возьмем

$\varphi_0 = 0,$

$\varphi_i(x) = x^i(1-x), \qquad i = 1, 2, ...$

Тогда, применяя метод Галеркина, получим

$$y_1(x) = \frac{5}{18}x(x-1),$$

$$y_2(x) = \frac{71}{369}x(1-x) + \frac{7}{41}x^2(1-x).$$

Сравним значения точного решения $y(x)$ со значениями приближенных решений $y_1(x)$ и $y_2(x)$ в отдельных точках.

x_i	$y(x)$	$y_1(x)$	$y_2(x)$
0,25	0,044	0,052	0,044
0,5	0,07	0,069	0,062
0,75	0,06	0,052	0,06

Разностный метод решения краевых задач.

Рассмотрим краевую задачу

$$\begin{cases} y'' = f(x, y, y'), \ x \in [a, b] \\ y(a) = A, \\ y(b) = B. \end{cases} \qquad (10.5)$$

Разобьем отрезок *[a, b]* на *n* одинаковых частей с шагом $h = \frac{b-a}{n}$ точками:

$a = x_0 < x_1 < ... < x_n = b$.

Заменим

$$y'(x_k) \approx \frac{y_{k+1} - y_k}{2h},$$

$$y''(x_k) \approx \frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} \qquad k = \overline{1, n-1},$$

где $y_k = y(x_k)$.

Получаем для любого внутреннего узла x_k, $k = \overline{1, n-1}$ уравнение

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = f\left(x_k, y_k, \frac{y_{k+1} - y_{k-1}}{2h}\right) \tag{10.6}$$

и для граничных узлов

$y_0 = A, \qquad y_n = B$.

То есть мы имеем систему из *(n+1)* уравнения с *(n+1)* неизвестными y_k. Ее решение дает нам приближенное решение краевой задачи.

Рассмотрим частный случай линейной краевой задачи:

$$y'' - p(x)y = f(x), \qquad p(x) > 0, \qquad a \le x \le b, \tag{10.7}$$

$y(a) = A$, $\qquad y(b) = B$.

В этом случае получаем

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} - p(x_k)y_k = f(x_k), \qquad k = \overline{1, n-1} \tag{10.8}$$

$y_0 = A, \qquad y_n = B$.

То есть получили трехдиагональную систему линейных уравнений

$$y_{k-1} - (2 + h^2 p(x_k))y_k + y_{k+1} = h^2 f(x_k), \qquad k = \overline{1, n-1},$$

в которой выполнено условие преобладания диагональных элементов

$2 + p(x_k) > 1 + 1$.

Такая система легко решается методом прогонки.

11. РАЗНОСТНЫЕ СХЕМЫ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Из предварительного рассмотрения краевых задач в параграфе 10 можно сделать вывод о превосходстве разностных методов численного решения краевых задач. Рассмотрим эти методы более подробно.

11.1. Разностные уравнения первого и второго порядка

Рассмотрим дифференциальное уравнение первого порядка

$$U' + AU = f(x), \qquad x \in [a,b].$$

Пусть его решением является функция *U=U(x)*.

Разобьем отрезок *[a,b]* на *n* равных частей с шагом *h*, то есть построим последовательность $\{x_k\}$:

$x_0 = a$, $x_k = x_0 + kh$, $k = \overline{0,n}$, где $h = \dfrac{b-a}{n}$.

Множество узлов x_k на отрезке *[a,b]* образует *сетку* D_h с шагом *h*.

D_h

При этом, любую функцию g_k $k=0, 1,...,n$, определенную на сетке D_h, будем называть сеточной функцией. В частности, сеточной функцией будет последовательность $U(x_k)$ $k = 0,...,n$, порожденная решением $U(x)$ дифференциального уравнения.

Заменим производные функции во внутренних узлах их разностными аппроксимациями:

$$U'(x_k) \approx \frac{U(x_k + h) - U(x_k)}{h}.$$

Значение функции U в *k*–ом узле обозначим

$$L_h U^{(h)} = \begin{cases} a_k U_{k-1} + b_k U_k + c_k U_{k+1}, & k = \overline{1,n-1} \\ U_0 \\ U_n, \end{cases}$$

$U(x_k) = U_k$, а функции f

$$\left\| U^{(h)} \right\|_{U_h} = \left\| U^{(h)} \right\| = \max_{k=\overline{0,n}} |U_k|$$

соответственно через $f_k = f(x_k)$. В новых обозначениях получим разностное уравнение:

$$\frac{U_{k+1} - U_k}{h} + AU_k = f_k.$$

С большей точностью первую производную можно также представить в виде:

$U'(x_k) \approx \dfrac{U(x_k + h) - U(x_k - h)}{2h}$, тогда уравнение будет иметь вид

$$\frac{U_{k+1} - U_{k-1}}{2h} + AU_k = f_k.$$

Рассмотрим теперь на данном отрезке дифференциальное уравнение второго порядка

$$U'' + AU' + BU = f(x).$$

Аналогично, заменив производные их разностными аппроксимациями, получим:

$$\frac{U_{k+1} - 2U_k + U_{k-1}}{h^2} + A\frac{U_{k+1} - U_{k-1}}{2h} + BU_k = f_k.$$

Преобразуем полученное разностное уравнение:

$$\left(\frac{A}{2h}+\frac{1}{h^2}\right)U_{k+1}+\left(B-\frac{2}{h^2}\right)U_k+\left(\frac{1}{h^2}-\frac{A}{2h}\right)U_{k-1}=f_k.$$

Полученные разностные уравнения будем также называть *разностными схемами* для соответствующих дифференциальных уравнений.

Таким образом, мы перешли от дифференциальных уравнений к разностным уравнениям вида:

$$a_kU_k+b_kU_{k+1}=f_k \qquad (11.1)$$

и

$$a_kU_{k-1}+b_kU_k+c_kU_{k+1}=f_k, \qquad (11.2)$$

где a_k, b_k, c_k − некоторые коэффициенты. Уравнения (11.1) и (11.2) будем называть соответственно *линейными разностными уравнениями первого и второго порядка* (при условии, что $a_k, b_k \neq 0$ в (11.1) и $a_k, c_k \neq 0$ в (11.2)). Заметим, что при переходе от дифференциального уравнения к разностному уравнению, порядок уравнения не обязательно сохраняется.

Рассмотрим свойства линейных разностных уравнений (11.1) и (11.2) , абстрагируясь от дифференциальных уравнений, их породивших. Их решениями будем называть сеточные функции $\{U_k\}$, удовлетворяющие соответствующим уравнениям при *k=0, 1,….,n.*

Очевидно, что для однозначного решения разностного уравнения (11.1) необходимо знать начальное значение сеточной функции U_k в нулевом узле, то есть U_0. Для решения разностного уравнения (11.2) необходимо знать два начальных значения U_0, U_1.

Назовем разностное уравнение (11.1) линейным разностным уравнением первого порядка, а уравнение (11.2) линейным разностным уравнением второго порядка.

Рассмотрим однородное уравнение, соответствующее уравнению (11.1):

$$a_kU_k+b_kU_{k+1}=0. \qquad (11.3)$$

Легко видеть, что если $\{U_k^{(1)}\}$ его решение, то решением также будет и сеточная функция $\{\alpha U_k^{(1)}\}$. Тогда, очевидно, что любое другое решение уравнения (11.3) можно получить при определенном численном значении α. То есть, если известно $\{\alpha U_k^{(1)}\}$, то любое решение $\tilde{U}_k$ представимо в виде $\tilde{U}_k=\alpha_0 U_k^{(1)}$.

Таким образом, $\{\alpha U_k^*\}$ представляет собой общее решение уравнения (11.3), из которого получаются выбором постоянной α все остальные решения. Легко видеть, что общим решением разностного уравнения (11.1) будет сеточная функция $U_k=U_k^*+\alpha U_k^{(1)}$, где U_k^* произвольное решение уравнения (11.1).

Рассмотрим однородное уравнение, соответствующее уравнению (11.2):

$$a_kU_{k-1}+b_kU_k+c_kU_{k+1}=0. \qquad (11.4)$$

Пусть $\{U_k^{(1)}\}$ и $\{U_k^{(2)}\}$ − решения (11.4), причем векторы $(U_0^{(1)}, U_1^{(1)})$ и $(U_0^{(2)}, U_1^{(2)})$ − линейно независимы, т. е.

$$\begin{vmatrix} U_0^{(1)} & U_1^{(1)} \\ U_0^{(2)} & U_1^{(2)} \end{vmatrix} \neq 0 .$$

Подстановка показывает, что при любых α и β сеточная функция $U_k = \alpha U_k^{(1)} + \beta U_k^{(2)}$ является тоже решением. Покажем, что $U_k = \alpha U_k^{(1)} + \beta U_k^{(2)}$ – общее решение, т.е. любое решение $\{U_k^*\}$ можно представить в виде

$U_k^* = \alpha_0 U_k^{(1)} + \beta_0 U_k^{(2)}$.

Действительно, полагая

$\left(U_0^*, U_1^*\right) = \alpha\left(U_0^{(1)}, U_1^{(1)}\right) + \beta\left(U_0^{(2)}, U_1^{(2)}\right),$

получим

$$\begin{cases} \alpha U_1^{(1)} + \beta U_1^{(2)} = U_1^* \\ \alpha U_1^{(1)} + \beta U_1^{(2)} = U_1^* . \end{cases}$$

Поскольку

$\Delta \neq 0$, то решение системы $\alpha = \alpha_0$, $\beta = \beta_0$ существует и является единственным.

Легко убедиться также, что, если U_k^* – частное решение уравнения (11.2), то общим решением уравнения (11.2) будет:

$U_k = U_k^* + \alpha U_k^{(1)} + \beta U_k^{(2)}$.

Теперь рассмотрим линейные однородные разностные уравнения с постоянными коэффициентами:

$$aU_{k-1} + bU_k + cU_{k+1} = 0, \qquad (11.5)$$

где $a, c \neq 0$.

Составим уравнение $a + bq + cq^2 = 0$, которое будем называть характеристическим для (11.5). Найдем его корни q_1 и q_2. Возможны следующие ситуации:

1). Пусть $q_1 \neq q_2$. Тогда подстановкой легко проверить, что $U_k^{(1)} = q_1^k$, $U_k^{(2)} = q_2^k$ – решения. Здесь q_1 и q_2 – действительные числа, отличные от нуля.

Подставив в уравнение (11.5), получим:

$aq_1^k + bq_1^{k+1} + cq_1^{k+2} = 0$ или $a + bq_1 + cq_1^2 = 0$.

Поскольку

$$\begin{vmatrix} 1 & q_1 \\ 1 & q_2 \end{vmatrix} = q_1 - q_2 \neq 0 ,$$

то общее решение имеет вид $U_k = \alpha q_1^k + \beta q_2^k$.

2. Пусть $q_1 = r(\cos\varphi + i\sin\varphi)$, то есть корни комплексные. Тогда

$U_k = q_1^k = r^k(\cos k\varphi + i\sin k\varphi)$.

Это комплексное решение. Отделяя его действительную и мнимую части, получим действительные частные решения

$$\begin{cases} U_k^{(1)} = r^k \cos k\varphi \\ U_k^{(2)} = r^k \sin k\varphi . \end{cases}$$

Значит, общее решение совпадает с линейной комбинацией $U_k = \alpha r^k \cos k\varphi + \beta r^k \sin k\varphi$.

3. Пусть $q_1 = q_2 = q$, то есть корни кратные.

Очевидно $U_k = q^k$ будет решением. Найдем второе решение.

Положим $\widetilde{U}_k = y_k q^k$ и подставим в (11.5):

$$ay_{k-1}q^{k-1} + by_k q^k + cy_{k+1}q^{k+1} = 0,$$

откуда

$$ay_{k-1} + by_k q + cy_{k+1}q^2 = 0.$$

По теоремы Виета для корней квадратного уравнения получаем

$$-\frac{b}{c} = 2q \ , \qquad \frac{a}{c} = q^2 \ . \tag{11.6}$$

Поделим квадратное уравнение на *c*:

$$\frac{a}{c}y_{k-1} + \frac{b}{c}ay_k + y_{k+1}q^2 = 0.$$

Тогда, согласно (11.6), получим:

$$q^2 y_{k-1} + 2q^2 y_k + q^2 y_{k+1} = 0$$

или

$$y_{k-1} - y_k = y_k - y_{k+1} \ .$$

Решением такого разностного уравнения будет любая арифметическая прогрессия, в том числе последовательность целых чисел $y_k = k$. То есть

$$\widetilde{U}_k = kq^k \ .$$

Поскольку

$$\begin{vmatrix} U_0 & U_1 \\ \widetilde{U}_0 & \widetilde{U}_1 \end{vmatrix} = \begin{vmatrix} 1 & q \\ 0 & q \end{vmatrix} = q \neq 0,$$

то общим решением будет

$$U_k = \alpha q^k + \beta k q^k .$$

11.2. Разностная краевая задача

Рассмотрим задачу вида:

$$\begin{cases} a_k U_{k-1} + b_k U_k + c_k U_{k+1} = f_k \ , \qquad k = \overline{1, n-1}, \\ U_0 = \varphi, \qquad U_k = \psi, . \end{cases} \tag{11.7}$$

называемую разностной краевой задачей (РКЗ).

Данную задачу можно коротко переписать в виде

$$\begin{cases} LU_k = f_k, \qquad k = \overline{1, n-1}, \\ U_0 = \varphi, \qquad U_k = \psi, \end{cases}$$

где линейный разностный оператор LU_k или $L(U_k)$ имеет вид $LU_k = a_k U_{k-1} + b_k U_k + c_k U_{k+1}$.

Пример 1. Найти решение РКЗ $U_{k-1} - U_k + U_{k+1} = 0, \qquad k = 1,...,299$,

где $U_0 = 0$, $U_{300} = 1$.

Запишем характеристическое уравнение $1 - q + q^2 = 0$. Отсюда

$q = \frac{1}{2} + i\frac{\sqrt{3}}{2}$, и общее решение разностного уравнения имеет вид: $U_k = \alpha \cos\frac{k\pi}{3} + \beta \sin\frac{k\pi}{3}$.

Подставляя значения U_0 и U_n, получаем $\alpha = 0$, $\beta = 1$. То есть рассматриваемая краевая задача не имеет решения.

Возьмем другие граничные условия

$U_0 = 0$, $U_{300} = 0$.

Решая задачу с данными условиями, получим, что решений бесконечно много.

То есть в отличие от задачи Коши для разностного уравнения, в случае РКЗ возникают проблемы:

1) существования решения;

2) единственности решения.

Выделим класс задач РКЗ, которые эффективно решаются.

Пусть $|a_k|, \quad |b_k|, \quad |c_k| < K$, где $K = const > 0$.

Задачу РКЗ будем называть *хорошо обусловленной*, если она имеет решение, причем единственное, при любых значениях $\varphi, \ \psi, \ f_k$, и, кроме того, это решение $\{U_k\}$ удовлетворяет соотношению:

$$|U_k| \le M \max\{|\varphi|, \quad |\psi|, \quad \max_i |f_i|\}, \tag{11.8}$$

где $M = const$, не зависящая от n.

Покажем, что условие (8) фактически означает слабую чувствительность задачи к ошибкам в граничных условиях и в правой части.

Рассмотрим возмущенную задачу:

$$\begin{cases} a_k U_{k-1} + b_k U_k + c_k U_{k+1} = f_k + \Delta f_k \\ U_0 = \varphi + \Delta\varphi, U_n = \psi + \Delta\psi. \end{cases}$$

Возмущенное решение имеет вид $U_k + \Delta U_k$, где ΔU_k − ошибка решения.

Оценим ΔU_k:

$L\ (U_k + \Delta U_k) = f_k + \Delta f_k$,

$L\ (U_k) + L(\Delta U_k) = f_k + \Delta f_k$.

Тогда

$$\begin{cases} L(\Delta U_k) = \Delta f_k, \\ \Delta U_0 = \Delta\varphi, \ \Delta U_n = \Delta\psi. \end{cases}$$

Из условия (11.8) получаем оценку:

$|\Delta U_k| \le M \max\{|\Delta\varphi|, \quad |\Delta\psi|, \quad \max_i |\Delta f_i|\}$.

Следовательно, ошибка в решении ограничена последним соотношением и при уменьшении шага ошибка не увеличивается. Растет лишь число узлов.

Сформулируем теперь достаточный признак хорошей обусловленности.

Теорема1 (достаточный признак хорошей обусловленности). Пусть выполнимо следующее условие:

$$|b_k| \geq |a_k| + |c_k| + \delta, \qquad \delta > 0. \tag{11.9}$$

Тогда задача РКЗ (11.3) хорошо обусловлена, а условие (8) имеет следующий вид:

$$|U_k| \leq \max\{|\varphi|, \quad |\psi|, \quad \frac{1}{\delta}\max_i |f_i|\}. \tag{11.10}$$

Доказательство. Докажем сначалда, что если РКЗ имеет решение, то это решение удовлетворяет (11.10). Пусть есть некоторое решение U_k и пусть $\max\limits_k |U_k| = |U_m|$.

Возможны следующие ситуации:

1) $m=0$ или $m=n$. Тогда (11.10) − тривиально выполняются.

2) $0< m < n$. Тогда

$$b_m U_m = -a_m U_{m-1} - c_m U_{m+1} + f_m,$$

следовательно

$$|b_m| \cdot |U_m| = |b_m U_m| = |-a_m U_{m-1} - c_m U_{m+1} + f_m| \leq |a_m| \cdot |U_m| + |c_m||U_m| + |f_m|.$$

Поскольку

$$|U_{m-1}| \leq |U_m|, \; |U_{m+1}| \leq |U_m|,$$

то, принимая во внимание (11.9), получаем

$$|U_m| \leq \frac{|f_m|}{|b_m| - |a_m| - |c_m|} \leq \frac{|f_m|}{\delta}.$$

То есть $|U_k| \leq |U_m| \leq \frac{1}{\delta}\max\limits_k |f_k|$ для всех $k = \overline{0,n}$. Следовательно, соотношение (11.10) справедливо.

Теперь покажем, что решение задачи РКЗ (11.9) существует и единственно.

При $k = \overline{0,n}$ задача РКЗ сводится к неоднородной линейной системе из *(n+1)* уравнений с *(n+1)* неизвестным. Она имеет единственное решение и то, тогда и только тогда, если соответствующая однородная система имеет только нулевое решение. Соответствующая однородная система имеет вид:

$$\begin{cases} a_k U_{k-1} + b_k U_k + c_k U_{k+1} = 0 \\ U_0 = 0, \qquad U_n = 0. \end{cases}$$

Тогда по условию (11.10) для этой системы $|U_k| \leq 0$. Значит, $U_k \equiv 0$.

То есть однородная система имеет только нулевое решение, и, значит, РКЗ имеет единственное решение.

Теорема доказана.

Теорема 2 (критерий хорошей обусловленности). Пусть $a_k = a, \quad b_k = b, \quad c_k = c$. Тогда задача является хорошо обусловленной тогда и только тогда, когда корни $q_1, \; q_2$ характеристического уравнения удовлетворяют условиям $|q_1| < 1, \; |q_2| > 1$.

Методы решения РКЗ

1. **Метод прогонки**.

Так как РКЗ представляет собой линейную трехдиагональную систему из *(n+1)* уравнений с *(n+1)* неизвестными, то для ее решения можно применить метод прогонки. Он обладает рядом преимуществ.

Достоинства:

а) число арифметических операций при использовании данного метода составляет *O(n)*, т. е. не превосходит Kn , где $K= const$.

б) при выполнении условия преобладания диагональных элементов $|b_k| \geq |a_k| + |c_k| + \delta$ метод прогонки оказывается слабо чувствительным к ошибкам вычислений: вычислительная погрешность не накапливается с ростом числа узлов *n.*

2. Метод стрельбы.

Рассмотрим РКЗ:

$$\begin{cases} a_k U_{k-1} + b_k U_k + c_k U_{k+1} = f_k, \quad k = 0,...,n \\ U_0 = \varphi, \qquad U_n = \psi \end{cases}$$

Вначале находим частное решение, решая задачу Коши с начальными условиями:

$$U_0 = \varphi, \qquad U_1 = 0, \qquad \Rightarrow \left\{ U_k^{(1)} \right\}.$$

Затем находим второе частное решение

$$U_0 = \varphi, \qquad U_1 = 1, \qquad \Rightarrow \left\{ U_k^{(2)} \right\}.$$

Поскольку

$$\begin{vmatrix} \varphi & 0 \\ \varphi & 1 \end{vmatrix} \neq 0 ,$$

общее решение можно записать в виде линейной комбинации этих двух частных решений

$$U_k = \sigma U_k^{(1)} + (1-\sigma) U_k^{(2)} .$$

Нужно выбрать σ так, чтобы выполнялись граничные условия:

$$U_n = \sigma U_k^{(1)} + (1-\sigma) U_k^{(2)} = \psi .$$

Отсюда: $\sigma = \dfrac{\psi - U_k^{(2)}}{U_k^{(1)} - U_k^{(2)}}$.

Достоинством метода стрельбы является его исключительная простота.

Однако он очень неустойчив. Приведем пример.

Пример 2. Рассмотрим РКЗ

$$\begin{cases} 5U_{k-1} - 26U_k + 5U_{k+1} = 0 \\ U_0 = \varphi, \quad U_n = \psi \end{cases} .$$

Данная РКЗ хорошо обусловлена. Это легко проверить, используя критерий хорошей обусловленности. Хорошая обусловленность означает устойчивость к ошибкам в правой части и начальных условиях.

Однако уже при вычислении *1-σ* появляется ошибка округления ε. Можно показать, что данная ошибка имеет экспоненциальный рост при уменьшении шага h, именно $\Delta U_n \sim 5^n \cdot \varepsilon$.

Отсюда видно, что уменьшение шага вычисления не дает положительного результата в виду роста ошибок. То есть, фактически метод стрельбы вследствие его неустойчивости не может иметь широкого применения.

11.3. Сходимость разностных схем

Выявим связь между дифференциальной краевой задачей (ДКЗ)

$$\begin{cases} U'' + a_1(x)U' + a_2(x)U = f(x) \\ U(a) = \varphi, \quad U(b) = \psi. \end{cases} \tag{11.11}$$

и соответствующей ей разностной краевой задачей (РКЗ).

Будем применять в дальнейшем для сокращения записи обозначение

$$LU = f,$$

где

$$LU = \begin{cases} U'' + a_1(x)U' + a_2(x)U, & a \le x \le b \\ U(a) \\ U(b), \end{cases}$$

$$f = \begin{cases} f(x), & a \le x \le b \\ \varphi, & x = a \\ \psi, & x = b. \end{cases}$$

Пример 3. Рассмотрим задачу

$$\begin{cases} U' + AU = 0 \\ U(0) = b \end{cases}, \quad x \in [0,1].$$

Запишем ее в виде:

$LU = f$, где

$$LU = \begin{cases} U' + AU, & x \in [0,1] \\ U(0) \end{cases}$$

$$f = \begin{cases} 0, & x \in [0,1] \\ b, & x = 0 \end{cases}$$

Интервал [0,1] разбиваем на множество узлов с шагом h, то есть строим сетку:

$$D_h = \begin{cases} x_k = x_0 + kh, \quad k = \overline{0,n} \\ x_0 = 0 \\ x_n = 1 \end{cases}.$$

Следует отметить, что дифференциальная краевая задача может не иметь решений или ее решение может быть не единственным. Однако данное обстоятельство относится к структуре этой задачи, а не к численным методам ее решения. Поэтому будем считать, что решение ДКЗ существует ит единственно. Пусть *U(x)* – единственное решение ДКЗ (11.11). По нему можно построить следующую сетчатую функцию из значений функции в узлах:

$[U]_h = (U(x_0),...,U(x_n))$.

Постоим такую последовательность сетчатых функций $U^{(h)} = \{U_0,\ U_1\ ,.....,U_n\}$, которая приближается к $[U]_h$ при убывании шага *h.*

Разностная схема для нашей задачи имеет вид:

$$\begin{cases} a_k U_{k-1} + b_k U_k + c_k U_{k+1} = f_k\ , \qquad k = \overline{1, n-1}\,, \\ U_0 = \varphi, \qquad U_k = \psi,. \end{cases}$$

или

$$L_h U^{(h)} = f^{(h)}, \tag{11.12}$$

где

$$L_h U^{(h)} = \begin{cases} a_k U_{k-1} + b_k U_k + c_k U_{k+1}, & k = \overline{1, n-1} \\ U_0 \\ U_n, \end{cases}$$

$$f^{(h)} = \begin{cases} f_k\ ,\ k = \overline{1, n-1}\,, \\ \varphi, \\ \psi. \end{cases}$$

Отметим, что задача (11.12) представляет собой семейство задач для разных значений шага *h*. Эта задача в свою очередь может не иметь решения или ее решение может быть не единственным. Предположим, что задача (11.12) при любом шаге *h* имеет единственное решение $U^{(h)}$.

Обозначим U_h − линейное пространство всех сетчатых функций на D_h. Введем в этом пространстве норму:

$$\left\|U^{(h)}\right\|_{U_h} = \left\|U^{(h)}\right\| = \max_{k=0,n} |U_k|\,.$$

Следует подчеркнуть, что это не единственный возможный выбор нормы в пространстве U_h. Существует большой набор норм, отличных от введенной выше.

Определение 1. Будем говорить, что сетчатая функция $U^{(h)}$ *сходится* к решению *U(x)* ДКЗ, если выполняется следующее условие:

$\left\|U^{(h)} - [U]_h\right\| \to 0$ при $h \to 0$.

При этом, если выполняется условие $\left\|U^{(h)} - [U]_h\right\| \le C_1 h^m$, где C_1 не зависит от h и n, говорят, что имеет место сходимость порядка h^m или $O(h^m)$.

Отметим, что сходимость – это фундаментальное свойство разностных схем, причем для одной и той же ДКЗ могут существовать разностные схемы

как сходящиеся, так и расходящиеся. Более того, сходимость или расходимость разностной схемы существенно зависит от выбора нормы в пространстве сеточных функций. Примерами других возможных норм в пространстве U_h являются нормы

$$\left\|U^{(h)}\right\|_{U_h} = h \max_{k=\overline{0,n}} |U_k|,$$

$$\left\|U^{(h)}\right\|_{U_h} = \sqrt{h \sum_k U_k^2}$$

и другие.

11.4. Порядок аппроксимации разностной схемы

Рассмотрим РКЗ

$$L_h U^{(h)} = f^{(h)}, \qquad (11.13)$$

где

$$L_h U^{(h)} = \begin{cases} a_{k-1}U_{k-1} + b_k U_k + c_k U_{k-1}, \ k = \overline{1, n-1} \\ U_0, \\ U_n, \end{cases}$$

$$f^{(h)} = \begin{cases} f_k, \ k = \overline{1, n-1} \\ \varphi, \ k = 0 \\ \psi, \ k = n. \end{cases}$$

Предположим она имеет единственное решение и ее решением является сеточная функция $U^{(h)}$.

Пусть сеточная функция $[U]_h$ построена из значений решения $U(x)$ ДКЗ в узлах сетки. Поставим ее в уравнение (11.13) и получим:

$$L_h [U]_h = f^{(h)} + \delta f^{(h)}.$$

Величину $\delta f^{(h)}$ называют невязкой.

Введем F_h − пространство обобщенных правых частей или пространство невязок, элементами которого являются $f^{(h)}$ и $\delta f^{(h)}$.

Введем норму в пространстве F_h следующим способом:

$$\left\|f^{(h)}\right\|_{F_h} = \left\|f^{(h)}\right\| = \max\{|\varphi|, |\psi|, \max_k |f_k|\}.$$

Отметим, что, как и в случае пространства U_h, существует широкий выбор возможной нормы в пространстве невязок, отличный от введенного выше определения нормы.

Определение 2. Будем говорить, что разностная схема (11.13) *аппроксимирует ДКЗ* на решении *U(x)* с порядком h^m, если выполняется следующее условие:

$\left\|\delta f^{(h)}\right\| \le Ch^m$, где C не зависит от h.

Пример 1. Рассмотрим дифференциальную краевую задача

$$\begin{cases} U' + AU = 0 \\ U(0) = b, \end{cases} \quad x \in [0,1].$$

Легко проверить, что ее решением является функция $U(x) = be^{-Ax}$.

Рассмотрим несколько разностных схем, аппроксимирующих данную задачу.
а) Рассмотрим разностную схему, построенную на аппроксимации производной:

$$U'(x) \approx \frac{U(x+h) - U(x-h)}{h}.$$

Запишем РКЗ

$$\begin{cases} \dfrac{U(x_k + h) - U(x_k)}{h} + AU(x_k) = 0 \\ U(0) = b, \end{cases}$$

или

$$\begin{cases} \dfrac{U_{k+1} - U_k}{h} + AU_k = 0 \\ U_0 = b. \end{cases}$$

Отсюда

$$\begin{cases} U_{k+1} + (Ah - 1)U_k = 0 \\ U_0 = b, \end{cases}$$

или

$$\begin{cases} U_{k+1} = (1 - Ah)U_k \\ U_0 = b. \end{cases}$$

Порядок аппроксимации производной равен $O(h)$ и начальные условия выполнены точно, то есть разностное решение аппроксимирует дифференциальную задачу с порядком $O(h)$.

Соответствующая решению ДКЗ сеточная функция будет иметь вид:

$$[U]_h = \{be^{-Ax_k}\}.$$

Выясним порядок сходимости разностной схемы. Для этого используем разложение в ряд:

$$\ln(x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots.$$

Получим следующую цепочку преобразований:

$$U_k = U(x_k) = b(1 - Ah)^k = be^{\ln(1-Ah)^K} =$$
$$= be^{k\ln(1-Ah)} = be^{\left(\frac{x_k}{h}\ln(1-Ah)\right)} = be^{\left(\frac{x_k}{h}(-Ah+O(h^2))\right)} =$$
$$= be^{-Ax_K + O(h)} = be^{-Ax_K}(1 + O(h)) = be^{-Ax_K} + O(h).$$

То есть

$$[U]_h - U_k = O(h),$$

а это значит, что мы имеем сходимость порядка *O(h)*.

б) Возьмем теперь более точную аппроксимацию производной:

$$\begin{cases} U'(x) \approx \dfrac{U(x+h)-U(x-h)}{2h} \\ U_0 = b \end{cases},$$

и получим разностную схему

$$\begin{cases} \dfrac{U_{k+1}-U_{k-1}}{2h} + AU_k = 0 \\ U_0 = b. \end{cases}$$

Преобразуем ее к виду

$$\begin{cases} U_{k+1} + 2AhU_k - U_{k-1} = 0 \\ U_0 = b. \end{cases}$$

Для решения требуется U_1. Найдем его, поступая следующим образом:

$$U'(0) \approx \frac{U(h)-U(0)}{h},$$

значит

$$\frac{U_1 - U_0}{h} = -AU_0 + O(h),$$

и, следовательно

$$U_1 = U_0 - hAU_0 + hO(h) = U_0(1-Ah) + O(h^2).$$

Разностная аппроксимация при подстановке точного решения в схему дает порядок аппроксимации $O(h^2)$.

Выясним порядок сходимости. Запишем соответствующее характеристическое уравнение:

$$q^2 + 2Ahq - 1 = 0.$$

Решая его, имеем

$$q_{1,2} = -Ah \pm \sqrt{A^2h^2+1},$$

откуда общее решение имеет вид

$$U_k = \alpha \cdot q_1^k + \beta \cdot q_2^k.$$

Используя начальные условия, найдем α и β из системы уравнений:

$$b = \alpha + \beta, \quad b \cdot (1-Ah) = \alpha \cdot q_1 + bq_2.$$

Получим

$$\alpha = b + O(h^2), \ \beta = O(h^2).$$

Тогда решение РКЗ имеет вид:

$$U_k = bq_1^k + O(h^2).$$

Так как $(1+x)^m = 1 + mx + \ldots$, получим

$$q_1 = -Ah + \sqrt{A^2h^2+1} = 1 - Ah + \frac{1}{2}A^2h^2 + O(h^4).$$

Тогда

$$q_1^k = q_1^{x_K/h} = e^{x_K/h \ln q_1} = e^{x_K/h \ln(1-Ah+\frac{1}{2}A^2h^2+O(h^4))} = e^{x_K/h \ln(-Ah+\frac{1}{6}A^3h^3+O(h^4))} =$$

$$= e^{-Ax_k} \cdot e^{\frac{1}{6}A^3x_kh^2+O(h^4)} = e^{-Ax_k(1+\frac{1}{6}A^3x_kh^2+O(h^4)}.$$

Следовательно

$$U^{(h)} = be^{-Ax_k} + O(h^2),$$

$$[U]_h = U(x_k) = be^{-Ax_k}.$$

То есть $[U]_h - U_k = O(h^2)$.

В рассмотренном выше примере порядки аппроксимации и сходимости совпадают, то есть, какой порядок аппроксимации разностной схемы, такой и порядок ее сходимости. Следующий пример показывает, что это далеко не всегда верно.

Пример 4. Рассмотрим еще одну разностную схему, основанную на следующей аппроксимации производной

$$\mu \frac{U(x+h)-U(x-h)}{2h} + (1-\mu)\frac{U(x+h)-U(x)}{h} \approx U'(x).$$

Предыдущие схемы были частными случаями данной схемы при $\mu = 0$ и $\mu = 1$. Если взять в разностной схеме $\mu = 4$, то порядок аппроксимации будет *O(h)*, но сходимости, как можно показать, не будет.

Вывод о совпадении порядков аппроксимации и сходимости разной схемы верен только для так называемых устойчивых разностных схем.

11.5. Устойчивость разностных схем

Рассмотрим ДКЗ

$$LU=f \qquad (11.14)$$

и соответствующую ей разностную схему

$$L_h U^{(h)} = f^{(h)}. \qquad (11.15)$$

Пусть $U = U(x)$ - решение задачи (1) и построена сеточная функция $[U]_h$.

Построим невязку

$$\delta f^{(h)} = L_h [U]_h - f^{(h)}.$$

Напомним, что разностная схема аппроксимирует решение $U(x)$ с порядком m, если справедливо соотношение:

$$\left\| \delta f^{(h)} \right\| \le C h^m,$$

где C – постоянная, не зависящая от h.

Рассмотрим возмущенную разностную схему:

$$L_h \cdot Z^{(h)} = f^{(h)} + \varepsilon^{(h)}, \text{ где } \varepsilon^{(h)} \in F_n. \qquad (11.16)$$

Т.е. схема (11.16) получена добавлением к правой части разностной схемы (11.15) возмущения $\varepsilon^{(h)} \in F_n$. Новое решение обозначим через $Z^{(h)}$.

Определение 3. *Разностную схему* (11.15) будем называть *устойчивой*, если существуют числа $h_0 > 0$ и $\delta > 0$ такие, что при всех $0 < h < h_0$ и всех $\varepsilon^{(h)} \in F_h$ таких, что $\left\| \varepsilon^{(h)} \right\| < \delta$, возмущенная разностная схема (11.16) имеет единственное решение, и это решение удовлетворяет оценке

$$\left\| Z^{(h)} - U^{(h)} \right\| \le C \cdot \left\| \varepsilon^{(h)} \right\|, \qquad (11.17)$$

где C=const и не зависит от h.

Необходимо подчеркнуть, что устойчивость не связана с дифференциальной краевой задачей (11.14), а имеет отношение только к разностной краевой задаче (11.15). То есть устойчивость – это внутреннее свойство разностной схемы.

Пусть задачи (11.14) и (11.15) линейны, тогда можно дать равносильное определение устойчивости разностной схемы.

Определение 4. В случае линейной РКЗ *разностная схема* (11.15) называется *устойчивой*, если существует число $h_0 > 0$ такое, что при любом $h < h_0$ разностное задача (11.15) имеет единственное решение при любой правой части $f^{(h)}$, и это решение удовлетворяет соотношению

$$\left\| U^{(h)} \right\| \le C \cdot \left\| f^{(h)} \right\|, \qquad (11.18)$$

где C – независимая от шага h константа.

Убедимся в равносильности этих определений в случае линейной разностной краевой задачи. Пусть задача (11.15) линейна и устойчива по определению 4 . Наряду с задачей (11.15)

$$L_h U^h = f^{(h)},$$

рассмотрим возмущенную задачу (11.16)

$$L_h Z^h = f^{(h)} + \varepsilon^{(h)},$$

которая имеет решение, причем единственное. Вычтем (11.15) из (11.16), используя линейность разностного оператора. Получим

$$L_h(Z^h - U^{(h)}) = \varepsilon^{(h)},$$

$$L_h W^{(h)} = \varepsilon^{(h)}, \qquad (11.19)$$

где $W^{(h)} = Z^{(h)} - U^{(h)}$.

Разностная задача (11.19) имеет единственное решение по определению 4, причем выполнено условие (11.18):

$$\left\|W^{(h)}\right\| \le C \cdot \left\|\varepsilon^{(h)}\right\|.$$

С учетом введенных обозначений выполнено также условие **(**11.17**),** то есть

$$\left\|Z^{(h)} - U^{(h)}\right\| \le C \cdot \left\|\varepsilon^{(h)}\right\|.$$

То есть из определения 4 следует определение 3. Нетрудно показать, что справедливо и обратное.

Замечания.

1) Понятие устойчивости зависит от определения нормы. За счет выбора подходящей нормы можно в известных пределах добиться устойчивости разностной схемы, неустойчивой в другой норме.

2) Понятие хорошей обусловленности представляет собой частный случай устойчивости, когда разностное уравнение линейно, второго порядка и выбор нормы зафиксирован (как выше).

Пример 5. Рассмотрим ДКЗ

$$\begin{cases} U'' - (1+x^2)U = \sqrt{x+1}, \quad 0 \le x \le 1 \\ U(0) = 2 \\ U(1) = 1. \end{cases}$$

Для данной задачи построим разностную схему:

$$U(x_i + h) - 2U(x_i) + U(x_i - h) - (1 - x^2)U(x_i) = \sqrt{x_i + 1}\ , \qquad i=0,\ldots,n.$$

Она аппроксимирует задачу с порядком аппроксимации $O(h^2)$.

Тогда имеем РКЗ:

$$\begin{cases} \dfrac{U_{i+1} - 2U_i + U_{i-1}}{h^2} - (1 + x_i^2)U_i = \sqrt{x_i + 1} \\ U_0 = 2, \qquad U_n = 1. \end{cases}$$

Покажем, что задача хорошо обусловлена.

$$|b_i| = \left|\frac{2}{h^2} + (1 + x_i)\right|,$$

$$|a_i| + |c_i| = \frac{1}{h^2} + \frac{1}{h^2}, \quad |b_i| \ge |a_i| + |c_i| + \delta.$$

При $\delta = 1$ задача является хорошо обусловленной. Из этого следует устойчивость задачи.

Теорема 3. Если разностная схема (11.15) аппроксимирует дифференциальную краевую задачу (11.14) на решении *U(x)* с порядком аппроксимации $O(h^m)$ и разностная схема (11.15) устойчива, то имеет место сходимость разностной схемы (11.15) на решении *U(x)* с порядком сходимости $O(h^m)$.

Доказательство. Рассмотрим решение *U(x)* задачи (11.14) и сеточную функцию $[U]_h$. Запишем невязку:

$\delta \cdot f^{(h)} = L_h[U]_h - f^{(h)}$.

Преобразуем это выражение в

$$L_h[U]_h = f^{(h)} + \delta \cdot f^{(h)}.$$

Мы получили возмущенную задачу, где $Z^{(h)} = [U]_h$ и $\delta \cdot f^{(h)}$ − возмущение.

Так как разностная схема устойчива, то

$$\left\| U^{(h)} - [U]_h \right\| \le C \cdot \left\| \delta \cdot f^{(h)} \right\| \le CC_1 h^m .$$

Последнее означает, что имеет место сходимость порядка *m*.

Теорема доказана.

Возвращаясь к примеру 3, можем сделать вывод, что порядок сходимости разностной схемы в данном примере равен $O(h^2)$.

Следующий пример показывает применение теоремы 1 для доказательства устойчивости разностных схем.

Пример 6. Рассмотрим разностную краевую задачу

$$\begin{cases} \dfrac{du}{dx} - G(x,u) = g(x) & 0 \le x \le 1 \\ U(0) = \varphi. \end{cases}$$

Будем предполагать функции *G* и *g* непрерывными по совокупности переменных, а функцию *G* дополнительно непрерывно дифференцируемой по *u*. Тогда в некоторой замкнутой ограниченной области на плоскости *Oxu,* содержащей внутри себя точку (0, φ) , будет выполнено неравенство

$$\left| \frac{\partial G}{\partial u}(x,u) \right| \le M ,$$

где *M*=const.

Запишем разностную схему:

$$\frac{U(x_i + h) - U(x_i)}{h} - G(x_i, U(x_i)) = g(x_i)$$

или

$$\begin{cases} \dfrac{U_{i+1} - U_i}{h} - G(x_i, U_i) = g_i \\ U_0 = U. \end{cases} \qquad (11.20)$$

Очевидно, разностная схема (11.20) представляет собой известную схему Эйлера.

Рассмотрим возмущенную задачу:

$$\begin{cases} \dfrac{Z_{i+1}-Z_i}{h} - G(x_i, Z_i) = g_i + \varepsilon_i \\ Z_0 = \varphi + \varepsilon, \end{cases} \qquad (11.21)$$

или $L_h Z^{(h)} = g^{(h)} + \varepsilon^{(h)}$.

Обозначим $\omega_i = Z_i - U_i$ и, вычитая из (11.21) равенство (11.20) и применяя формулу Лагранжа, получим задачу

$$\begin{cases} \dfrac{\omega_{i+1}-\omega_i}{h} - G_U^{'}(x_i, \xi_i)\omega_i = \varepsilon_i \\ \omega_0 = \varepsilon. \end{cases}$$

Отсюда

$\omega_{i+1} = \omega_i + hM_i\omega_i + \varepsilon_i h$, где $M_i = G_U^{'}(x_i, \xi_i)$, ξ_i - некоторое промежуточное значение между U_i и Z_i. Очевидно $|M_i| \le M$, и, следовательно,

$$\omega_{i+1} \le (1+hM)\cdot|\omega_i| + h|\varepsilon_i| \le (1+hM)\cdot|\omega_i| + h\left\|\varepsilon^{(h)}\right\|,$$

где

$$\left\|\varepsilon^{(h)}\right\| = \max\left\{|\varepsilon|, \max_{i=0,n}|\varepsilon_i|\right\}.$$

Тогда $|\omega_i| \le (1+hM)\cdot|\omega_{i-1}| + h\left\|\varepsilon^{(h)}\right\|$,

и

$$|\omega_{i+1}| \le (1+hM)^2|\omega_{i-1}| + h(1+hM)\left\|\varepsilon^{(h)}\right\| + h\left\|\varepsilon^{(h)}\right\| \le (1+hM)^2|\omega_{i-1}| + 2(1+hM)\left\|\varepsilon^{(h)}\right\|,$$

$$|\omega_{i+1}| \le (1+hM)^{i+1}|\omega_0| + (i+1)(1+hM)^i\left\|\varepsilon^{(h)}\right\| \le (1+hM)^h\left\|\varepsilon^{(h)}\right\| + (1+Mh)^h\left\|\varepsilon^{(h)}\right\|,$$

$$|\omega_{i+1}| \le 2(1+hM)^h\left\|\varepsilon^{(h)}\right\|, \quad \forall \quad i = \overline{0, n-1}.$$

Поскольку

$$(1+hM)^h = (1+\frac{M}{h})^h \le e^M,$$

то

$$|\omega_{i+1}| \le 2e^M\left\|\varepsilon^{(h)}\right\|,$$

и значит

$$\left\|\omega^{(h)}\right\| \le 2e^M\left\|\varepsilon^{(h)}\right\|.$$

Таким образом, выполнено условие (11.17), и, по определению 3, разностная схема устойчива. Тогда, согласно теореме 1, схема Эйлера сходится с порядком сходимости *O(h)*.

12. РАЗНОСТНЫЕ СХЕМЫ РЕШЕНИЯ КРАЕВЫХ ЗАДАЧ ДЛЯ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ В ЧАСТНЫХ ПРОИЗВОДНЫХ

12.1. Разностные аппроксимации дифференциальных краевых задач

Рассмотрим дважды непрерывно дифференцируемую функцию двух переменных $U=U(x,y)$ в области D (см. рис.12.1), лежащей на плоскости Oxy. Необходимо аппроксимировать ее частные производные

$$\frac{\partial U}{\partial x}, \frac{\partial U}{\partial y}, \frac{\partial^2 U}{\partial x^2}, \frac{\partial^2 U}{\partial x \partial y}, \frac{\partial^2 U}{\partial y^2}$$

с помощью разностных производных.

Разобьем область D с помощью вертикальных и горизонтальных прямых с шагом h и l соответственно. То есть, проведем прямые $x = x_i$, $y = y_j$, где

$$x_i = x_0 + ih, \qquad i = \overline{0,n}$$

$$y_j = y_0 + jl, \qquad j = \overline{0,m}.$$

Точки их пересечения образуют сетку на плоскости.

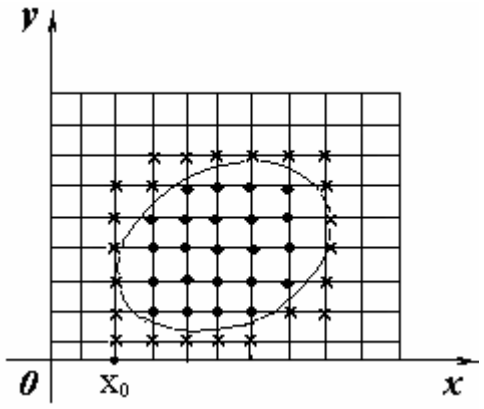


Рис. 12.1

Положим $U_{i,j} = U(x_i, y_j)$. Очевидно, возможны различные аппроксимации частных производных первого порядка:

$$\frac{\partial U}{\partial x}(x_i, y_j) \approx \frac{U(x_i + h, y_j) - U(x_i - h, y_j)}{2h} = \frac{U_{i+1,j} - U_{i-1,j}}{2h},$$

$$\frac{\partial U}{\partial x}(x_i, y_j) \approx \frac{U(x_i + h, y_j) - U(x_i, y_j)}{h} = \frac{U_{i+1,j} - U_{i,j}}{h},$$

$$\frac{\partial U}{\partial x}(x_i, y_j) \approx \frac{U(x_i, y_j + l) - U(x_i, y_j - l)}{2h} = \frac{U_{i,j+1} - U_{i,j-1}}{2h}.$$

Их называют соответственно левой, правой и центральной разностными производными.

Аналогично получаются аппроксимации для производной $\frac{\partial U}{\partial y}$. Для вторых производных можно ввести аппроксимации

$$\frac{\partial^2 U}{\partial x^2}(x_i, y_j) \approx \frac{U(x_i + h, y_j) - 2U(x_i, y_j) + U(x_i - h_i, y_j)}{h^2} = \frac{U_{i+1,j} - 2U_{ij} + U_{i-1,j}}{h^2},$$

$$\frac{\partial^2 U}{\partial y^2} \approx \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{l^2},$$

$$\frac{\partial^2 U}{\partial x \partial y}(x_i, y_j) = \left(\frac{\partial U}{\partial x}\right)'_y \approx \left(\frac{U(x_i + h, y_j) - U(x_i - h_i, y_j)}{2h}\right)'_y \approx$$

$$\approx \frac{U(x_i + h, y_j + l) - U(x_i - h_i, y_j + l)}{2h} - \frac{U(x_i + h, y_j - l) - U(x_i - h_i, y_j - l)}{2h} =$$

$$= \frac{U_{i+1,j+1} - U_{i-1,j+1} - U_{i+1,j-1} + U_{i-1,j-1}}{4hl}.$$

Пусть *U=U(x,y)*. Рассмотрим в области *D* уравнение в частных производных вида:

$$a_{11}\frac{\partial^2 U}{\partial x^2} + 2a_{12}\frac{\partial^2 U}{\partial x \partial y} + a_{22}\frac{\partial^2 U}{\partial y^2} + a_{31}\frac{\partial U}{\partial x} + a_{32}\frac{\partial U}{\partial y} + a_{33}U = g(x, y) \qquad (12.1)$$

где a_{11}, a_{12}, a_{22}, a_{31}, a_{32}, a_{33} − некоторые числовые коэффициенты, *g(x,y)* − непрерывная в области *D* функция. Будем рассматривать уравнение (12.1) совместно с граничным условием

$$U|_\Gamma = \varphi(x, y), \qquad (12.2)$$

где Г − контур, ограничивающий область *D*.

Задачу (12.1), (12.2) будем называть дифференциальной краевой задачей и записывать коротко в виде

$$LU = f, \qquad (12.3)$$

где

$$LU = \begin{cases} a_{11}\dfrac{\partial^2 U}{\partial x^2} + 2a_{12}\dfrac{\partial^2 U}{\partial x \partial y} + a_{22}\dfrac{\partial^2 U}{\partial y^2} + a_{13}\dfrac{\partial U}{\partial x} + a_{23}\dfrac{\partial U}{\partial y} + a_{33}U & \text{при } (x, y) \in D \\ U|_\Gamma & \text{при } (x, y) \in \Gamma, \end{cases}$$

$$f = \begin{cases} g(x, y) & \text{при } (x, y) \in D \\ \varphi(x, y) & \text{при } (x, y) \in \Gamma, \end{cases}$$

$U = U(x, y)$ − неизвестная функция двух переменных.

Заменяя частные производные во внутренних узлах сетки их разностными аппроксимациями, получим следующую разностную схему дифференциального уравнения (12.1)

$$a_{11}\frac{U_{i+1,j}-2U_{i,j}+U_{i-1,j}}{h^2}+a_{12}\frac{U_{i+1,j+1}-U_{i-1,j+1}-U_{i+1,j-1}+U_{i-1,j-1}}{2lh}+$$
$$+a_{22}\frac{U_{i,j+1}-2U_{i,j}+U_{i,j-1}}{h^2}+a_{31}\frac{U_{i+1,j}-U_{i-1,j}}{2h}+a_{32}\frac{U_{i,j+1}-U_{i,j-1}}{2l}+a_{33}U_{i,j}=g_{i.j} \qquad (12.4)$$

где $U_{i,j}=U(x_i,y_j)$, $g_{i,j}=g(x_i,y_j)$.

Каждый узел (x_i,y_j), лежащий внутри области, будем называть внутренним и множество внутренних узлов обозначим D^*. Узлы, лежащие на контуре Γ (если они есть) и узлы, окаймляющие контур Γ извне, будем называть граничными и обозначим их совокупность Γ^* (см. рис. 12.1).

Аппроксимируем граничные условия (2), полагая

$$U_{ij}=\varphi_{ij}^* \text{ для } (x_i,y_j)\in\Gamma^*, \qquad (12.5)$$

где φ_{ij}^* – значение функции $\varphi(x,y)$ в точке на контуре Γ, ближайшей к узлу $(x_i,y_j)\in\Gamma^*$.

Таким образом, мы получили разностную краевую задачу (12.4), (12.5), соответствующую дифференциальной краевой задаче (12.1), (12.2). Задачу (12.4), (12.5) называют также разностной схемой задачи (12.1), (12.2). В ней неизвестными являются всевозможные значения U_{ij}, соответствующие внутренним узлам. Отметим, что по существу задача (12.4), (12.5) представляет собой систему линейных уравнений относительно неизвестных U_{ij}.

Обычно выбирают $l=r(h)$, где r – некоторая функция, или $l=rh$, где r – постоянная. Совокупность всех узлов обозначим

$$D_h=D^*\cup\Gamma^*$$

и будем называть сеткой в области D.

Функцию $U^{(h)}$, определенную на сетке D_h, будем называть сеточной. Т.е. это функция, которая ставит в соответствие каждому узлу $(x_i,y_j)\in D_h$ число $U(x_i,y_j)=U_{ij}$.

Обозначим через U_h пространство всех сеточных функций на сетке D_h и введем в нем норму сеточной функции $U^{(h)}$, полагая

$$\left\|U^{(h)}\right\|_{U_h}=\max_{i,j}\left|U_{ij}^{(h)}\right|,$$

где максимум берется по всем узлам сетки D_h.

Для краткости будем записывать разностную схему (12.4), (12.5) дифференциальной краевой задачи (12.1), (12.2) в виде операторного уравнения

$$L_hU^{(h)}=f^{(h)}, \qquad (12.6)$$

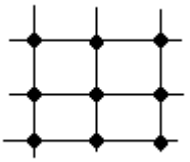
определенного на сетке D_h, где

$$L_h U^{(h)} = \begin{cases} a_{11}\dfrac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} + a_{12}\dfrac{U_{i+1,j+1} - U_{i-1,j+1} - U_{i+1,j-1} + U_{i-1,j-1}}{2lh} + \\ + a_{22}\dfrac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2} + a_{31}\dfrac{U_{i+1,j} - U_{i-1,j}}{2h} + a_{32}\dfrac{U_{i,j+1} - U_{i,j-1}}{2l} + a_{33}U_{i,j} \\ при \quad (x_i, y_j) \in D^* \\ \\ U^{(h)}|_{\Gamma^*}, \quad при \quad (x_i, y_j) \in \Gamma^*, \end{cases}$$

$$f^{(h)} = \begin{cases} g_{ij} = g(x_i, y_j), \quad при \quad (x_i, y_j) \in D^* \\ \varphi^*_{ij}, \quad при \quad (x_i, y_j) \in \Gamma^*, \end{cases}$$

φ^*_{ij} – значение функции $\varphi(x, y)$ в точке на контуре Γ, ближайшей к узлу $(x_i, y_j) \in \Gamma^*$.

Для классификации разностных схем используется понятие шаблона. Шаблоном разностной схемы называется геометрическое место узлов сетки, участвующих в разностном уравнении (4) (см. рис.12.2). В зависимости от разностного уравнения шаблон может быть полным и неполным.



– полный шаблон

Рис. 12.2.

Возникают естественные вопросы: Существует ли решение краевой дифференциальной задачи (12.1), (12.2) и единственно ли оно, в случае, если существует? Существует ли решение разностной краевой разностной задачи (12.4), (12.5)? Если существует, будет ли оно единственным? Если разностная краевая задача имеет решение, будет ли оно сходится при h и l стремящихся к нулю? Если оно сходится, то сходится ли к решению дифференциальной краевой задачи (12.1), (12.2)? Если оно сходится, то с какой скоростью?

Первый вопрос относится к теории дифференциальных уравнений. Ответы на остальные вопросы дает численный анализ.

Рассмотрим отдельные примеры дифференциальных краевых задач в частных производных.

12.2. Уравнение теплопроводности

Рассмотрим одномерный однородный стержень длины *L*. Пусть $U(x,t)$ − температура в точке стержня с абсциссой *x* в момент времени *t*. Из математической физики известно, что распределение температуры в точках стержня в зависимости от времени описывается дифференциальным уравнением

$$\frac{\partial U}{\partial t} = a^2 \frac{\partial^2 U}{\partial x^2}. \qquad (12.7)$$

Добавим к уравнению естественные граничные условия:

$$U(0,t) = \varphi(t),$$
$$U(L,t) = \psi(t), \qquad (12.8)$$
$$U(x,0) = g(x),$$

которые описывают температуру, измеряемую на концах стержня и начальное распределение температуры в точках стержня.

Построим сетку $x_i = ih,\ i = \overline{0,n},\ t_j = jl,\ j = \overline{0,m}.$ в области *D* (см. рис.12.3)

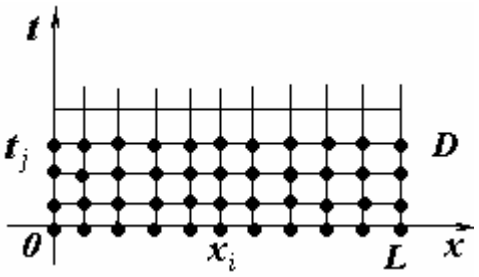


Рис. 12.3.

Заменим в дифференциальном уравнении значения частных производных во внутренних узлах их разностными аппроксимациями:

$$\frac{\partial U}{\partial t} \approx \frac{U_{i,j+1} - U_{i,j}}{l},$$
$$\frac{\partial^2 U}{\partial x^2}(x_i, y_j) \approx \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2}.$$

Получим разностное уравнение соответствующее исходному дифференциальному уравнению:

$$\frac{U_{i,j+1} - U_{i,j}}{l} = a^2 \frac{U_{i+1} - 2U_{i,j} + U_{i-1,j}}{h^2},$$

или

$$U_{i,j+1} = U_{i,j} + a^2 \frac{l}{h^2}(U_{i+1,j} - 2U_{i,j} + U_{i-1,j}). \qquad (12.9)$$

Положим

$$g_i = g(x_i), \quad \varphi_j = \varphi(t_j), \quad \psi_j = \psi(t_j).$$

Тогда граничные условия аппроксимируются следующим образом

$$U_{i0} = g_i, \qquad U_{0,j} = \varphi_j, \qquad U_{nj} = \psi_j. \qquad (12.10)$$

Зная значения U_{ij} на нижнем (нулевом) слое и на границе слева и справа, вычисляем $U_{i,1}$ для $i=\overline{1,n-1}$:

$$U_{i,1}=U_{i,0}+\frac{a^2l}{h^2}(U_{i+1,0}-2U_{i,0}+U_{i-1,0}).$$

Таким образом, мы получили разностную краевую задачу (12.9), (12.10), которая очевидно имеет решение, причем единственное.

Рассматривая уравнение (12.9), построим его шаблон (см. рис. 12.4):

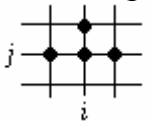


Рис. 12.4.

Вопрос сходимости к решению дифференциальной задачи зависит от соотношения между l и h в уравнении (12.7). Известно, что если

$$\frac{a^2l}{h^2}\le\frac{1}{2},\quad \text{т. е.}\quad l\le\frac{a^2l}{h^2},$$

то имеет место сходимость.

Рассмотрим подробнее ситуацию:

$$\frac{a^2l}{h^2}=\frac{1}{2}.$$

В этом случае уравнение (12.9) существенно упрощается и принимает вид:

$$U_{i,j+1}=\frac{1}{2}(U_{i+1,j}+U_{i-1,j}).$$

На рис. 12.5 показан его шаблон.

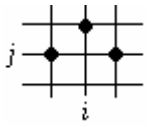


Рис. 12.5.

Пусть $U_{h,l}(x,t)$ - функция двух переменных, график которой построен на основе аппроксимации плоскостями найденных при решении задачи (12.9), (12.10) значений U_{ij}. Можно показать, что при выполнении условия $\frac{a^2l}{h^2}\le\frac{1}{2}$ функция $U_{h,l}(x,t)$ будет сходиться к решению дифференциальной краевой задачи (7), (8) и скорость сходимости будет:

$$\left|(U_{h,l}(x,t)-U(x,y)\right|=O(h^2).$$

12.3. Волновое уравнение

Волновое уравнение − дифференциальное уравнение в частных производных 2-го порядка, описывающее процесс распространения колебаний в некоторой среде. Мы будем рассматривать малые колебания натянутой струны, закрепленной в двух точках на оси Ох.

Пусть *U(x, t)* – отклонение точки струны с абсциссой *x* в момент времени *t* от положения равновесия. Тогда величина *U(x, t)* описывается уравнением

$$\frac{\partial^2 U}{\partial t^2} = a^2 \frac{\partial^2 U}{\partial x^2} \qquad (12.11)$$

Добавим к уравнению граничные условия:

$$U(0,t) = 0\,,\ \ U(L,t) = 0\,,\ \ U(x,0) = \varphi(x)\,,\ \ \frac{\partial U}{\partial t}(x,0) = \psi(x)\,, \qquad (12.12)$$

которые отражают тот факт, что концы струны закреплены и дают величину начального отклонения точек струны от положения равновесия и их начальные скорости.

Построим сетку:

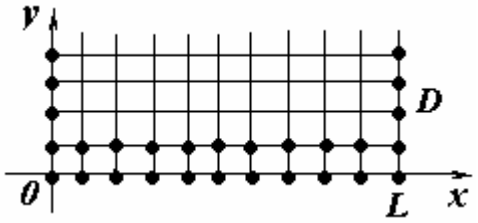


Рис. 12.6.

$$x_i = ih, \qquad i = \overline{0,n}\ , \qquad t_j = jl, \qquad j = \overline{0,m}\,.$$

Заменим значения производных во внутренних узлах сетки их разностными аппроксимациями:

$$\frac{\partial^2 U}{\partial x^2}(x_i,t_j) \approx \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2},$$

$$\frac{\partial^2 U}{\partial t^2}(x_i,t_j) \approx \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{l^2}.$$

Получим разностное уравнение, отвечающее уравнению

$$\frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{l^2} = a^2 \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2}$$

или

$$U_{i,j+1} = 2U_{i,j} - U_{i,j-1} + \lambda(U_{i+1,j} - 2U_{i,j} + U_{i-1,j})\,,$$

где $\lambda = \dfrac{a^2 l^2}{h^2}$.

Запишем его окончательно в виде:

$$U_{i,j+1} = 2(1-\lambda)U_{i,j} - U_{i,j-1} + \lambda(U_{i+1,j} + U_{i-1,j})\,. \qquad (12.13)$$

Данное разностное уравнение имеет шаблон, изображенный на рисунке 12.7.

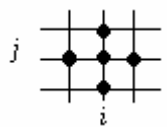


Рис. 12.7.

Получим аппроксимацию граничных условий (12.12):

$$U_{0,j} = 0, \qquad U_{n,j} = 0, \qquad \forall j = \overline{0,m} \tag{12.14}$$

$$U_{i,0} = \varphi_i, \qquad \varphi_i = \varphi(x_0), \qquad \psi_i = \varphi(x_i), \tag{12.15}$$

а также

$$\frac{U_{i,1} - U_{i,0}}{l} = \psi_i$$

или

$$U_{i,1} = U_{i,0} - l\psi_i. \tag{12.16}$$

Как легко видеть разностная схема (12.13)-(12.16) имеет единственное решение и легко реализуется последовательным вычислением значений сеточной функции, начиная с нулевого и первого слоя. Можно показать, что приближенное решение $U_{h,l}(x,t)$, полученное из решения разностной схемы (12.13)-(12.16), будет сходиться к решению исходной дифференциальной краевой задачи, если

$$\frac{a^2 l^2}{h^2} \le \frac{1}{2},$$

причем со скоростью $\left|(U_{h,l}(x,t) - U(x,y)\right| = O(h^2)$.

12.4. Задача Дирихле для уравнения Лапласа

Пусть *D* область на плоскости *Oxy* (см. рис.12.1), ограниченная замкнутым контуром *Г*. Рассмотрим в этой области однородное уравнение Лапласа

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = 0$$

с граничными условиями

$U\big|_{\Gamma} = \varphi(x,y)$.

Данную краевую задачу называют задачей Дирихле для уравнения Лапласа. Поскольку переменные x и y имеет одинаковый характер, разобьем область с помощью вертикальных и горизонтальных прямых с одинаковым шагом *l=h.*

Для каждого внутреннего узла (x_i, y_j) составим разностное уравнение:

$$\frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,}}{h^2} + \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2} = 0, \tag{12.17}$$

заменяя значения частных производных в дифференциальном уравнении их разностными аппроксимациями.

Аппроксимируем граничные условия, полагая

$U_{ij} = \varphi^*_{ij}$ для $(x_i, y_j) \in \Gamma^*$,

где Γ^* - совокупность граничных узлов (напомним, что это узлы, лежащие на контуре Γ и узлы, окаймляющие контур Γ извне, см.рис.12.1), φ^*_{ij} – значение функции $\varphi(x, y)$ в точке на контуре Γ, ближайшей к узлу $(x_i, y_j) \in \Gamma^*$.

Приводя подобные, получаем разностную схему

$$U_{i,j} = \frac{1}{4}(U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1}), \qquad (12.18)$$

$U_{ij} = \varphi^*_{ij}$ (в граничных узлах). (12.19)

Полученную разностную схему будем рассматривать как систему линейных уравнений относительно неизвестных U_{ij} во внутренних узлах. При этом линейная система является неоднородной, поскольку некоторые уравнения будут содержать кроме неизвестных еще и определенные значения $U_{ij} = \varphi^*_{ij}$ в граничных узлах. Чтобы система система линейных уравнений (12.18), (12.19) имела единственное решение, необходимо и достаточно, чтобы соответствующая ей однородная система имела только нулевое решение. Но соответствующая однородная система линейных уравнений имеет вид

$$U_{i,j} = \frac{1}{4}(U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1}), \qquad (12.20)$$

$U_{i,j} = 0$ (в граничных узлах).

Допустим, что система (12.20) имеет ненулевое решение. Предположим, что хотя бы в некоторых узлах это решение принимает положительные значения, и положим

$\overline{U_{i,j}} = \max U_{i,j} = U_{i_0,j_0}$,

где максимум берется по всем внутренним узлам.

Если подставить $i=i_0$, $j= j_0$ в уравнение (12.20), то максимальное значение U_{i_0,j_0} равно среднему арифметическому значений нашего решения в соседних узлах. Но это возможно только в случае, когда значения в соседних узлах совпадают с максимальным значением U_{i_0,j_0}. Подставим теперь в уравнение (12.20)

$i=i_0+1, j= j_0$ $(i=i_0, j= j_0+1; i=i_0 -1, j= j_0; i=i_0, j= j_0 -1)$

и совершенно аналогично получим, что в соседних с данными узлами узлах решение U_{ij} должно принимать также максимальное значение. Поскольку через конечное число шагов мы придем к уравнению, в котором участвуют граничные узлы, получается, что и в граничных узлах значение решения уравнения (12.20) максимально и, следовательно, положительно. Последнее очевидно невозможно в виду нулевых граничных условий. Аналогично рассматривается и случай, когда решение уравнения (12.20) не

положительно во всех узлах. Полученное противоречие говорит о том, что решение (12.20) должно быть нулевым.

Следовательно, система (12.18),(12.19) имеет решение, причем единственное.

12.5. Сходимость разностных аппроксимаций

В области D , ограниченной контуром Г, рассмотрим дифференциальную краевую задачу

$$LU = f, \qquad (12.21)$$

где

$$LU = \begin{cases} a_{11}\dfrac{\partial^2 U}{\partial x^2} + 2a_{12}\dfrac{\partial^2 U}{\partial x \partial y} + a_{22}\dfrac{\partial^2 U}{\partial y^2} + a_{13}\dfrac{\partial U}{\partial x} + a_{23}\dfrac{\partial U}{\partial y} + a_{33}U & при \quad (x,y) \in D \\ U|_{\Gamma} & при \quad (x,y) \in \Gamma, \end{cases}$$

$$f = \begin{cases} g(x,y) & при \quad (x,y) \in D \\ \varphi(x,y) & при \quad (x,y) \in \Gamma, \end{cases}$$

$U = U(x,y)$ − неизвестная функция двух переменных.

Разобьем область D (см. рис. 12.1) с помощью вертикальных и горизонтальных прямых с шагом h и l соответственно. То есть, проведем прямые $x = x_i$, $y = y_j$, где

$$x_i = x_0 + ih, \qquad i = \overline{0,n}$$

$$y_j = y_0 + jl, \qquad j = \overline{0,m}.$$

Положим $l = r(h)$, где r − некоторая функция, или $l = rh$, где r − постоянная. Определим сетку D_h (совокупность внутренних и граничных узлов) и пространство сеточных функций на сетке D_h с нормой сеточной функции $U^{(h)}$

$$\left\| U^{(h)} \right\|_{U_h} = \max_{i,j} \left| U_{ij}^{(h)} \right|,$$

где максимум берется по всем узлам сетки D_h .

Для дифференциальной краевой задачи построим некоторую разностную схему

$$L_h U^{(h)} = f^{(h)}, \qquad (12.22)$$

где

$$L_h U^{(h)} = \begin{cases} a_{11}\dfrac{U_{i+1,j}-2U_{i,j}+U_{i-1,j}}{h^2} + a_{12}\dfrac{U_{i+1,j+1}-U_{i-1,j+1}-U_{i+1,j-1}+U_{i-1,j-1}}{2lh} + \\ +a_{22}\dfrac{U_{i,j+1}-2U_{i,j}+U_{i,j-1}}{h^2} + a_{31}\dfrac{U_{i+1,j}-U_{i-1,j}}{2h} + a_{32}\dfrac{U_{i,j+1}-U_{i,j-1}}{2l} + a_{33}U_{i,j} \\ при \quad (x_i, y_j) \in D^* \\ \\ U^{(h)}|_{\Gamma^*}, \quad при \quad (x_i, y_j) \in \Gamma^*, \end{cases}$$

$$f^{(h)} = \begin{cases} g_{ij} = g(x_i, y_j), \quad при \quad (x_i, y_j) \in D^* \\ \varphi^*_{ij}, \quad при \quad (x_i, y_j) \in \Gamma^*, \end{cases}$$

φ^*_{ij} − значение функции $\varphi(x, y)$ в точке на контуре Γ, ближайшей к узлу $(x_i, y_j) \in \Gamma^*$.

Будем предполагать далее, что

1) дифференциальная краевая задача (12.21) имеет решение $U = U(x, y)$, причем единственное;
2) разностная краевая задача (12.22) имеет единственное решение при любом выборе шага h меньшего некоторого значения h_0.

Определим сеточную функцию $[U]_h$, значения которой совпадают на сетке D_h со значениями решения $U(x, y)$ дифференциальной краевой задачи (12.21) в узлах сетки D_h.

Определение 1. будем говорить, что разностная схема (12.22) сходится на решении $U(x, y)$ дифференциальной краевой задачи (12.21), если

$$\left\| U^{(h)} - [U]_h \right\|_{U_h} \to 0 \text{ при } h \to 0,$$

где $U^{(h)}$ − решение разностной краевой задачи (12.22). При этом, если

$$\left\| U^{(h)} - [U]_h \right\|_{U_h} \le C_1 h^m,$$

где $C_1 = const$ не зависящая от h, то имеет место порядок сходимости $O(h^m)$.

Введем пространство сеточных функций F_h, элементами которого являются всевозможные сеточные функции $f^{(h)}$, определенные на сетке D_h. Норму в пространстве F_h определим как

$$\left\| f^{(h)} \right\|_{F_h} = \max_{(x_i, y_j) \in \Gamma^*} \left| \varphi_{ij} \right| + \max_{(x_i, y_j) \in D^*} \left| g_{ij} \right|.$$

Рассмотрим невязку

$$\delta f^{(h)} = L_h [U]_h - f^{(h)},$$

соответствующую решению $U = U(x, y)$ дифференциальной краевой задачи (12.21).

Определение 2. Будем говорить, что разностная схема (12.22) аппроксимирует дифференциальную краевую задачу (21) на ее решении $U = U(x, y)$, если

$$\left\| \delta f^{(h)} \right\|_{F_h} \to 0 \text{ при } h \to 0.$$

При этом будем говорить, что имеет место порядок аппроксимации $O(h^m)$, если

$$\left\| \delta f^{(h)} \right\|_{F_h} \le C_2 h^m,$$

где $C_2 = const$ не зависит от h.

По аналогии с параграфом 11 введем понятие устойчивости разностной схемы. (Отметим, что здесь мы имеем дело только с линейными дифференциальными операторами LU).

Определение 3. Разностная схема (12.22) называется устойчивой, если существует число $h_0 > 0$ такое, что для любого $h < h_0$ разностная краевая задача (12.22) имеет единственное решение при любой правой части $f^{(h)}$, и это решение удовлетворяет условию

$$\left\| U^{(h)} \right\| \le C \left\| f^{(h)} \right\|, \qquad (12.23)$$

где $C = const$ не зависит от h и $f^{(h)}$.

Теорема 1. Пусть разностная схема (12.22) аппроксимирует дифференциальную краевую задачу (12.21) с порядком аппроксимации $O(h^m)$ и разностная схема (12.22) устойчива. Тогда разностная схема (12.22) сходится на решении $U(x, y)$ дифференциальной краевой задачи (12.21) с порядком сходимости $O(h^m)$.

Доказательство теоремы проводится аналогично доказательству теоремы 1 предыдущего параграфа.

Теорема 1 представляет собой эффективное средство исследования сходимости разностных схем. Покажем это на примере задачи теплопроводности.

Пример. Рассмотрим задачу теплопроводности

$$\frac{\partial U}{\partial t} - a^2 \frac{\partial^2 U}{\partial x^2} = 0, \; x \in [o, L], \; t \in [0, T],$$

с граничными условиями

$$U(x,0) = g(x), \; U(0,t) = \varphi(t), \; U(L,t) = \psi(t).$$

Введем сетку D_h, определенную прямыми

$$x_i = ih, \quad i = \overline{0,n}; \; t_j = jl, \quad j = \overline{0,m},$$

где $n = L/h, \quad m = T/l, \quad l = \dfrac{h^2}{2a^2}$.

Составим разностную схему, соответствующую рассматриваемой дифференциальной краевой задаче в узлах сетки D_h. Получим

$$\frac{U(x_i, t_{j+1}) - U(x_i, t_j)}{l} - a^2 \frac{U(x_{i+1}, t_j) - 2U(x_i, t_j) + U(x_{i-1}, t_j)}{h^2} = 0$$

или, полагая $U_{i,j} = U(x_i, t_j)$,

$$\frac{U_{i,j+1}-u_{i,j}}{h^2}-\frac{1}{2}\frac{(U_{i+1,j}-2U_{i,j}+U_{i-1,j})}{h^2}=0$$

откуда

$$\frac{U_{i,j+1}-\frac{1}{2}U_{i+1,j}-\frac{1}{2}U_{i-1,j}}{h^2}=0.$$

Граничные условия будут иметь вид

$U_{i.,0}=g_i$, $U_{o,j}=\varphi_j$, $U_{n,j}=\psi_j$,

где $g_i=g(x_i)$, $\varphi_j=\varphi(t_j)$, $\psi_j=\psi(t_j)$.

Таким образом, мы имеем разностную схему

$L_hU^{(h)}=f^{(h)}$,

где

$$L_hU^{(h)}=\begin{cases}\dfrac{U_{i,j+1}-\frac{1}{2}U_{i+1,j}-\frac{1}{2}U_{i-1,j}}{h^2} & \textit{для } (i,j),\ \textit{соответствующих внутренним узлам}\\ U_{i,0} \quad \textit{при } i=\overline{0,n}\\ U_{0,j} \quad \textit{при } j=\overline{0,m}\\ u_{n,j} \quad \textit{при } j=\overline{0,m},\end{cases}$$

$$f^{(h)}=\begin{cases}0, & \textit{для } (i,j),\ \textit{соответствующих внутренним узлам сетки}\\ g_i, & i=\overline{0,n}\\ \varphi_j, & j=\overline{0,m}\\ \psi_j, & j=\overline{0,m}.\end{cases}$$

Покажем, что разностная схема сходится на решении $U(x,t)$ дифференциальной краевой задачи.

Порядок аппроксимации определить легко. Точность разностной аппроксимации производной $\frac{\partial U}{\partial t}$ определяется $O(l)=O(h^2)$, производной $\frac{\partial^2 U}{\partial x^2}$ – $O(h^2)$, граничные условия аппроксимируются точно. В итоге, разностная схема имеет порядок аппроксимации $O(h^2)$.

Покажем, что построенная разностная схема устойчива. Рассмотрим возмущенную краевую задачу

$L_hz^{(h)}=f^{(h)}$

с произвольной правой частью

$$f^{(h)}=\begin{cases}f_{ij}, & \textit{для всех } (i,j) \textit{ соответствующих внутренним узлам } D_h\\ \overline{g}_i, & i=\overline{0,n}\\ \overline{\varphi}_j, & j=\overline{0,m}\\ \overline{\psi}_j, & j=\overline{0,m},\end{cases}$$

где f_{ij}, $\overline{g}_i$, $\overline{\varphi}_j$, $\overline{\psi}_j$ – произвольные сеточные функции, определенные на внутренних и граничных узлах сетки.

Надо показать, что возмущенная задача имеет

1) единственное решение $z^{(h)}$;

2) это решение удовлетворяет оценке

$$\left\|z^{(h)}\right\|_{U_h} \le C\left\|f^{(h)}\right\|_{F_h}, \qquad (12.23)$$

где

$$\left\|z^{(h)}\right\|_{U_h} = \max_{i,j}\left|z_{ij}\right|, \ \left\|f^{(h)}\right\|_{F_h} = \max\{\max_i\left|\overline{g}_i\right|, \max_j\left|\overline{\varphi}_j\right|, \max_j\left|\overline{\psi}_j\right|, \max_{i,j}\left|f_{ij}\right|\}.$$

Существование решения и его единственность вытекают из того, что значения $z^{(h)}$ на нулевом слое (при $j=0$) определяются граничными условиями, в первом слое ($j=1$) однозначно определяются из разностной схемы и т.д.

С другой стороны

$$\left|z_{i,0}\right| \le \left|\overline{g}_i\right| \le \left\|f^{(h)}\right\|, \quad i=\overline{0,n}.$$

Далее

$$\left|z_{i,1}\right| \le \frac{1}{2}\left|z_{i+1,0}\right| + \frac{1}{2}\left|z_{i-1,0}\right| + h^2\left|f_{i1}\right| \le \frac{1}{2}\left|\overline{g}_i\right| + \frac{1}{2}\left|\overline{g}_i\right| + h^2\left|f_{i1}\right| \le \left\|f^{(h)}\right\| + h^2\left|f_{i1}\right|, \quad i=\overline{0,n}.$$

Отсюда

$$\left|z_{i,1}\right| \le (1+h^2)\left\|f^{(h)}\right\|, \quad i=\overline{0,n}.$$

Аналогично

$$\left|z_{i,2}\right| \le (1+h^2)\left\|f^{(h)}\right\| + h^2\left|f_{i,2}\right| \le (1+2h^2)\left\|f^{(h)}\right\|, \quad i=\overline{0,n},$$

… … … … … … … … … … … … … … … … …

$$\left|z_{i,m}\right| \le (1+mh^2)\left\|f^{(h)}\right\|, \quad i=\overline{0,n}.$$

Поскольку

$$mh^2 = \frac{T}{l}h^2 = \frac{2a^2\cdot T\cdot h^2}{h^2} = 2Ta^2 = const,$$

то

$$\left\|z^{(h)}\right\| \le C\left\|f^{(h)}\right\|.$$

Таким образом, разностная схема устойчива и, следовательно, в силу теоремы 1 она сходится на решении $U(x,t)$ дифференциальной краевой задачи с порядком сходимости $O(h^2)$.

13. СВОЙСТВА РАЗНОСТНЫХ СХЕМ ДЛЯ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ

Пусть D − область на плоскости $0xy$, ограниченная контуром Γ. Рассмотрим дифференциальную краевую задачу

$$LU = f, \tag{13.1}$$

где

$$LU = \begin{cases} a_{11}\dfrac{\partial^2 U}{\partial x^2} + a_{12}\dfrac{\partial^2 U}{\partial x \partial y} + a_{22}\dfrac{\partial^2 U}{\partial y^2} + a_{13}\dfrac{\partial U}{\partial x} + a_{23}\dfrac{\partial U}{\partial y} + a_{33}U \quad при \quad (x,y) \in D \\ U|_{\Gamma} \quad при \quad (x,y) \in \Gamma, \end{cases}$$

$$f = \begin{cases} g(x,y) \quad при \quad (x,y) \in D \\ \varphi(x,y) \quad при \quad (x,y) \in \Gamma, \end{cases}$$

$U = U(x,y)$ − неизвестная функция двух переменных.

Разобьем область D с помощью прямых $x_i = x_0 + ih \;\; i = \overline{0,n}$, $y_j = y_0 + jl \;\; j = \overline{0,m}$. Каждый узел (x_i, y_j), лежащий внутри области, будем называть внутренним и множество внутренних узлов обозначим D^*. Узлы, лежащие на контуре Γ (если они есть) и узлы, окаймляющие контур Γ извне, будем называть граничными и обозначим их совокупность Γ^* (см. рисунок 13.1).

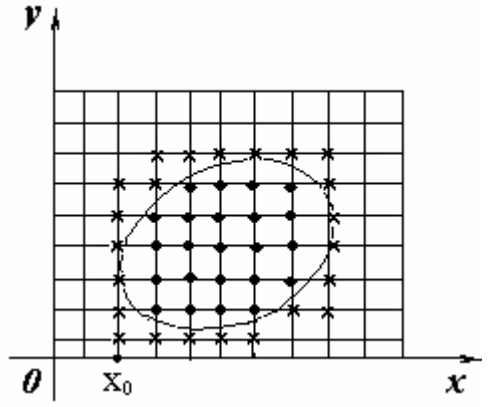


Рис. 13.1

Обычно выбирают $l = r(h)$, где r − некоторая функция, или $l = rh$, где r − постоянная. Совокупность всех узлов обозначают

$$D_h = D^* \cup \Gamma^*$$

и называют сеткой.

Функцию $U^{(h)}$, определенную на сетке D_h, будем называть сеточной. Т.е. это функция, которая ставит в соответствие каждому узлу $(x_i, y_j) \in D_h$ число $U(x_i, y_j) = u_{ij}$.

Обозначим через U_h пространство всех сеточных функций на сетке D_h и введем в нем норму сеточной функции $U^{(h)}$, положим

$$\left\|U^{(h)}\right\|_{U_h} = \max_{i,j}\left|u_{ij}^{(h)}\right|,$$

где максимум берется по всем узлам сетки D_h.

Построим для дифференциальной краевой задачи (13.1) разностную схему

$$L_h U^{(h)} = f^{(h)}, \qquad (13.2)$$

определенную на сетке D_h, где

$$L_h U^{(h)} = \begin{cases} \textit{Разностная схема, соответствующая левой части} \\ \textit{дифференциального уравнения из (1), при } (x_i, y_j) \in D^* \\ U^{(h)}|_{\Gamma^*}, \quad \textit{при } (x_i, y_j) \in \Gamma^*, \end{cases}$$

$$f^{(h)} = \begin{cases} g_{ij} = g(x_i, y_j), \quad \textit{при } (x_i, y_j) \in D^* \\ \varphi_{ij}^*, \quad \textit{при } (x_i, y_j) \in \Gamma^*, \end{cases}$$

где φ_{ij}^* – значение функции $\varphi(x, y)$ в точке на контуре Γ, ближайшей к узлу $(x_i, y_j) \in \Gamma^*$.

Будем предполагать далее, что

1) дифференциальная краевая задача (13.1) имеет решение $U = U(x, y)$, причем единственное;
2) разностная краевая задача (13.2) имеет единственное решение при любом выборе шага h меньшего некоторого значения h_0.

Определим сеточную функцию $[U]_h$, значения которой совпадают на сетке D_h со значениями решения $U(x, y)$ дифференциальной краевой задачи (13.1) в узлах сетки D_h.

Определение 1. Будем говорить, что разностная схема (13.2) сходится на решении $U(x, y)$ дифференциальной краевой задачи (13.1), если

$$\left\|U^{(h)} - [U]_h\right\|_{U_h} \to 0 \text{ при } h \to 0,$$

где $U^{(h)}$ – решение разностной краевой задачи (2). При этом, если

$$\left\|U^{(h)} - [U]_h\right\|_{U_h} \le C_1 h^m,$$

где $C_1 = const$ не зависящая от h, то имеет место порядок сходимости $O(h^m)$.

Введем пространство сеточных функций F_h, элементами которого являются всевозможные сеточные функции $f^{(h)}$, определенные на сетке D_h. Норму в пространстве F_h определим как

$$\left\|f^{(h)}\right\|_{F_h} = \max_{(x_i, y_j) \in \Gamma^*}\left|\varphi_{ij}\right| + \max_{(x_i, y_j) \in D^*}\left|g_{ij}\right|.$$

Введем невязку

$$\delta f^{(h)} = L_h[U]_h - f^{(h)},$$

соответствующую решению $U = U(x, y)$ дифференциальной краевой задачи (13.1).

Определение 2. будем говорить, что разностная схема (13.2) аппроксимирует дифференциальную краевую задачу (13.1) на ее решении $U = U(x, y)$, если

$$\left\| \delta f^{(h)} \right\|_{F_h} \to 0 \text{ при } h \to 0.$$

При этом будем говорить, что имеет место порядок аппроксимации $O(h^m)$, если

$$\left\| \delta f^{(h)} \right\|_{F_h} \le C_2 h^m,$$

где $C_2 = const$ не зависит от h.

По аналогии с §12 введем понятие устойчивости разностной схемы. (Отметим, что в данном параграфе мы имеем дело с линейными дифференциальными операторами LU).

Определение 3. Разностная схема (13.2) называется устойчивой, если существует число $h_0 > 0$ такое, что для любого $h < h_0$ разностная краевая задача (2) имеет единственное решение при любой правой части $f^{(h)}$, и это условие удовлетворяет условию

$$\left\| U^{(h)} \right\| \le C \left\| f^{(h)} \right\|,$$

где $C = const$ не зависит от h и $f^{(h)}$.

Теорема 1. Пусть разностная схема (13.2) аппроксимирует дифференциальную краевую задачу (13.1) с порядком аппроксимации $O(h^m)$ и разностная схема (13.2) устойчива. Тогда разностная схема (13.2) сходится на решении $U(x, y)$ дифференциальной краевой задачи (13.1) с порядком сходимости $O(h^m)$.

Пример. Рассмотрим задачу теплопроводности

$$\frac{\partial U}{\partial t} - a^2 \frac{\partial^2 U}{\partial x^2} = 0, \; x \in [o, L], \; t \in [0, T],$$

с граничными условиями

$$U(x,0) = g(x), \; U(0,t) = \varphi(t), \; U(L,t) = \psi(t).$$

Введем сетку D_h, определенную прямыми

$$x_i = ih, \quad i = \overline{0,n}; \; t_j = jl, \quad j = \overline{0,m},$$

где $n = L/h, \quad m = T/l, \quad l = \dfrac{h^2}{2a^2}$.

Составим разностную схему, соответствующую рассматриваемой дифференциальной краевой задаче в узлах сетки D_h. Получим

$$\frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{l} - a^2 \frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2} = 0$$

или, полагая $u_{i,j} = u(x_i, t_j)$,

$$\frac{u_{i,j+1} - u_{i,j}}{h^2} - \frac{1}{2} \frac{(u_{i+1,j} - 2u_{i,j} + u_{i-1,j})}{h^2} = 0$$

откуда

$$\frac{u_{i,j+1} - \frac{1}{2}u_{i+1,j} - \frac{1}{2}u_{i-1,j}}{h^2} = 0.$$

Граничные условия будут иметь вид

$u_{i.,0} = g_i$, $u_{o,j} = \varphi_j$, $u_{n,j} = \psi_j$,

где $g_i = g(x_i)$, $\varphi_j = \varphi(t_j)$, $\psi_j = \psi(t_j)$.

Таким образом, мы имеем разностную схему

$L_h U^{(h)} = f^{(h)}$,

где

$$L_h U^{(h)} = \begin{cases} \dfrac{u_{i,j+1} - \frac{1}{2}u_{i+1,j} - \frac{1}{2}u_{i-1,j}}{h^2} & \textit{для } (i,j), \textit{ соответствующих внутренним узлам} \\ u_{i,0} \quad \textit{при} \quad i = \overline{0,n} \\ u_{0,j} \quad \textit{при} \quad j = \overline{0,m} \\ u_{n,j} \quad \textit{при} \quad j = \overline{0,m}, \end{cases}$$

$$f^{(h)} = \begin{cases} 0, \quad \textit{для } (i,j), \textit{ соответствующих внутренним узлам сетки} \\ g_i, \quad i = \overline{0,n} \\ \varphi_j, \quad j = \overline{0,m} \\ \psi_j, \quad j = \overline{0,m}. \end{cases}$$

Покажем, что разностная схема сходится на решении $u(x,t)$ дифференциальной краевой задачи.

Порядок аппроксимации определить легко. Точность разностной аппроксимации производной $\frac{\partial U}{\partial t}$ определяется $O(l) = O(h^2)$, производной $\frac{\partial^2 U}{\partial x^2}$ – $O(h^2)$, граничные условия аппроксимируются точно. В итоге, разностная схема имеет порядок аппроксимации $O(h^2)$.

Покажем, что построенная разностная схема устойчива. Рассмотрим возмущенную краевую задачу

$L_h z^{(h)} = f^{(h)}$

с произвольной правой частью

$$f^{(h)} = \begin{cases} f_{ij}, \quad \textit{для всех } (i,j) \textit{ соответствующих внутренним узлам } D_h \\ \overline{g}_i, \quad i = \overline{0,n} \\ \overline{\varphi}_j, \quad j = \overline{0,m} \\ \overline{\psi}_j, \quad j = \overline{0,m}, \end{cases}$$

где f_{ij}, $\overline{g}_i$, $\overline{\varphi}_j$, $\overline{\psi}_j$ – произвольные сеточные функции, определенные на внутренних и граничных узлах сетки.

Надо доказать, что возмущенная задача имеет

3) единственное решение $z^{(h)}$;

4) это решение удовлетворяет оценке

$$\left\|z^{(h)}\right\|_{U_h} \le C\left\|f^{(h)}\right\|_{F_h},$$

где $\left\|z^{(h)}\right\|_{U_h} = \max\limits_{i,j}\left|z_{ij}\right|$, $\left\|f^{(h)}\right\|_{F_h} = \max\{\max\limits_{i}\left|\overline{g}_i\right|, \max\limits_{j}\left|\overline{\varphi}_j\right|, \max\limits_{j}\left|\overline{\psi}_j\right|, \max\limits_{i,j}\left|f_{ij}\right|\}$.

Существование решения и его единственность вытекают из того, что значения $z^{(h)}$ на нулевом слое (при $j=0$) определяются граничными условиями, в первом слое ($j=1$) однозначно определяются из разностной схемы и т.д.

С другой стороны

$$\left|z_{i,0}\right| \le \left|\overline{g}_i\right| \le \left\|f^{(h)}\right\|, \quad i = \overline{0,n}.$$

Далее

$$\left|z_{i,1}\right| \le \frac{1}{2}\left|z_{i+1,0}\right| + \frac{1}{2}\left|z_{i-1,0}\right| + h^2\left|f_{i1}\right| \le \frac{1}{2}\left|\overline{g}_i\right| + \frac{1}{2}\left|\overline{g}_i\right| + h^2\left|f_{i1}\right| \le \left\|f^{(h)}\right\| + h^2\left|f_{i1}\right|, \quad i = \overline{0,n}.$$

Отсюда

$$\left|z_{i,1}\right| \le (1+h^2)\left\|f^{(h)}\right\|, \quad i = \overline{0,n}.$$

Аналогично

$$\left|z_{i,2}\right| \le (1+h^2)\left\|f^{(h)}\right\| + h^2\left|f_{i,2}\right| \le (1+2h^2)\left\|f^{(h)}\right\|, \quad i = \overline{0,n},$$

… … … … … … … … … … … … … … … … …

$$\left|z_{i,m}\right| \le (1+mh^2)\left\|f^{(h)}\right\|, \quad i = \overline{0,n}.$$

Поскольку $mh^2 = \frac{T}{l}h^2 = \frac{2a^2 \cdot T \cdot h^2}{h^2} = 2Ta^2 = const$, то

$$\left\|z^{(h)}\right\| \le C\left\|f^{(h)}\right\|.$$

Таким образом, разностная схема устойчива и, следовательно, в силу теоремы 1 она сходится на решении $u(x,t)$ дифференциальной краевой задачи с порядком сходимости $O(h^2)$.

ЛИТЕРАТУРА

1. Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. Численные методы. – М: БИНОМ, 2004. – 636с.
2. Мысовских И.П. Лекции по методам вычислений. М.: Наука, 1993. – 496с.
3. Калитин Н.Н. Численные методы. – М : Наука, 1978. – 612с.
4. Бахвалов Н.С. Численные методы. – М : Наука, 1976. – 632с.
5. Демидович В.П. и др. Численные методы анализа. –М: Физматгиз, 1963. – 400с.
6. Волков Е.А. Численные методы. – М: Наука, 1982. – 255с.
7. Крылов В.И. и др. Вычислительные методы высшей математики. Т.1. - Мн : Выш шк., 1972. – 684с.
8. Крылов В.И. и др. Вычислительные методы высшей математики. Т.2. – Мн : Выш. шк., 1976. – 672с.
9. Форсайт Дж. и др. Машинные методы математических вычислений. – М : Мир, 1980. – 280с.
10. Шуп Т. Решение инженерных задач на ЭВМ. – М: Мир, 1982. – 238с.
11. Сборник задач по методам вычислений / Под ред. Монастырного П.И. – М: Наука, 1994. – 318с.
12. Самарский А.А. Введение в численные методы. – М: Наука, 1987. – 288с.
13. Березин И.С. , Жидков Н.П. Методы вычислений. Т.1. - М: Фитматгиз, 1962. – 464с.

Св. план 2005,поз.52

Учебное издание

Минченко Леонид Иванович

КРАТКИЙ КУРС ЧИСЛЕННОГО АНАЛИЗА

Учебное пособие по курсу «Методы численного анализа» для студентов специальности «Информатика» для всех форм обучения

В 2-х частях
Часть1

Редактор Т.Н.Крюкова
Корректор

Подписано в печать	Формат 60x84 1/16	Бумага
Гарнитура «Таймс»	Печать офсетная	Усл.-печ.л.
Уч.-изд.л.	Тираж 150 экз.	Заказ

Издатель и полиграфическое исполнение: Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники»
Лицензия на осуществление издательской деятельности№ 02330/0056964 от 01.04.2004.
Лицензия на осуществление полиграфической деятельности№ 02330/0131518 от 30.04.2004.
220013, Минск, П.Бровки, 6