Л.Л. Глазырина, М.М. Карчевский

Введение в численные методы

Учебное пособие

Казань

2017

УДК 519.3
ББК 22.311
Г52

Научный редактор
доктор физико-математических наук М.Ф. Павлова
Рецензенты:
доктор физико-математических наук, профессор П.Г. Данилаев,
доктор физико-математических наук, профессор В.С. Желтухин

Глазырина Л.Л., Карчевский М.М.
Г 52 Введение в численные методы: учебное пособие. — Казань: Казан. ун-т, 2017. — 122 с.

Излагаются основные принципы построения и исследования численных методов алгебры, анализа и решения дифференциальных уравнений. Приведены практические задания по методам, предложенным в пособии. Пособие предназначено для студентов, специализирующихся по математическим методам в экономике и информационным технологиям.

УДК 519.3
ББК 22.311

Оглавление

Предисловие

Пособие написано на основе лекций для студентов института вычислительной математики и информационных технологий КФУ, специализирующихся в области математических методов в экономике и информационных технологий. В пособии содержатся также типовые задания по основным разделам численных методов, которые могут быть использованы на практических занятиях. Предполагается, что читатель знаком со стандартными курсами математического анализа, линейной алгебры и аналитической геометрии, обыкновенных дифференциальных уравнений. Многие вопросы, затронутые в пособии, активно обсуждались с сотрудниками кафедры вычислительной математики Казанского федерального университета. Авторы выражают им свою искреннюю благодарность. Рукопись пособия была внимательно прочитана М.Ф.Павловой. Авторы постарались максимально учесть ее замечания. Авторы признательны Е.М.Федотову, оказавшему большую помощь при подготовке пособия к печати.

Глава 1

Численные методы алгебры

§ 1. Прямые методы решения систем линейных уравнений

1. Метод Гаусса. Рассматривается система линейных алгебраических уравнений

$$Ax = b. \tag{1.1}$$

Здесь

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{pmatrix}.$$

Матрица A предполагается невырожденной, то есть $\det A \neq 0$. Поэтому система (1.1) однозначно разрешима при любой правой части, ее решение может быть выписано по формулам Крамера

$$x_i = \frac{\Delta_i}{\Delta}, \quad i = 1, 2, \dots, n,$$

где $\Delta = \det A$ — определитель матрицы A [1)], а Δ_i — это определитель, получающийся из определителя матрицы A заменой i-го столбца столбцом свободных членов системы (1.1).

Казалось бы, формулы Крамера полностью решают задачу построения решения системы линейных уравнений, однако на практике они не используются. Это объясняется следующим. Напомним, что

$$\det A = \sum_{(\alpha_1, \alpha_2, \dots, \alpha_n)} \pm a_{1\alpha_1} a_{2\alpha_2} \cdots a_{n\alpha_n}, \tag{1.2}$$

где $(\alpha_1, \alpha_2, \dots, \alpha_n)$ — перестановка символов $1, 2, \dots, n$. Число слагаемых в сумме равно $n!$, поэтому непосредственное вычисление определителя требует $nn!$ арифметических операций, что уже при $n = 30$ недоступно даже для самых мощных ЭВМ. Кроме того, из-за большого числа сомножителей в слагаемых (1.2) непосредственные вычисления по формуле (1.2) могут приводить к переполнению разрядной

[1)] Определитель матрицы A обозначается также через $|A|$.

сетки ЭВМ или к сильному накоплению погрешностей округления. Поэтому для решения систем уравнений применяют другие, более экономичные и устойчивые по отношению к погрешностям округления методы.

1.1. Начнем с самого простого и наиболее распространенного метода — метода Гаусса. Запишем систему (1.1) в индексной форме:

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1,\\ &a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ &a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n. \end{aligned} \tag{1.3}$$

Предположим, что $a_{11} \neq 0$ и поделим на это число первое уравнение системы. Получим:

$$x_1 + a_{12}^{(1)}x_2 + \ldots + a_{1n}^{(1)}x_n = b_1^{(1)}.$$

Умножим теперь первое уравнение системы на a_{21} и вычтем из второго. Аналогично преобразуем остальные уравнения системы. В результате получим систему уравнений, эквивалентную исходной:

$$\begin{aligned} x_1 + a_{12}^{(1)}x_2 + \ldots + a_{1n}^{(1)}x_n &= b_1^{(1)},\\ a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n &= b_2^{(1)},\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots&\\ a_{n2}^{(1)}x_2 + \ldots + a_{nn}^{(1)}x_n &= b_n^{(1)}, \end{aligned} \tag{1.4}$$

где

$$\begin{aligned} &a_{1j}^{(1)} = a_{1j}/a_{11},\ j = 2, \ldots, n,\ b_1^{(1)} = b_1/a_{11},\ a_{ij}^{(1)} = a_{ij} - a_{1j}^{(1)}a_{i1},\\ &i = 2, \ldots, n, j = 2, \ldots, n,\ b_i^{(1)} = b_i - b_1a_{i1},\ i = 2, \ldots, n. \end{aligned} \tag{1.5}$$

Отбросим первое уравнение системы (1.4), а для оставшихся — выполним преобразования, такие же, как для системы (1.3). Получим:

$$\begin{aligned} x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \ldots + a_{1n}^{(1)}x_n &= b_1^{(1)},\\ x_2 + a_{23}^{(2)}x_3 + \ldots + a_{2n}^{(2)}x_n &= b_2^{(2)},\\ a_{33}^{(2)}x_3 + \ldots + a_{3n}^{(2)}x_n &= b_3^{(2)},\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots&\\ a_{n3}^{(2)}x_3 + \ldots + a_{nn}^{(2)}x_n &= b_n^{(2)}. \end{aligned} \tag{1.6}$$

Проводя далее аналогичные преобразования, придем, наконец, к системе с треугольной матрицей:

$$\begin{aligned} x_1 + a_{12}^{(1)} x_2 + a_{13}^{(1)} x_3 + \ldots + a_{1n}^{(1)} x_n &= b_1^{(1)}, \\ x_2 + a_{23}^{(2)} x_3 + \ldots + a_{2n}^{(2)} x_n &= b_2^{(2)}, \\ x_3 + \ldots + a_{3n}^{(3)} x_n &= b_3^{(3)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ x_{n-1} + a_{n-1,n}^{(n-1)} x_n &= b_{n-1}^{(n-1)}, \\ x_n &= b_n^{(n)}. \end{aligned} \tag{1.7}$$

Решение системы (1.7) вычисляется по формулам:

$$x_n = b_n^{(n)}, \quad x_i = b_i^{(i)} - \sum_{j=i+1}^{n} a_{ij}^{(i)} x_j, \quad i = n-1, n-2, \ldots, 1. \tag{1.8}$$

Описанный способ получения системы (1.7) из системы (1.3) есть прямой ход метода Гаусса. Вычисления по формулам (1.8) составляют обратный ход метода Гаусса.

1.2. Оценка трудоемкости метода Гаусса. Подсчитаем, сколько арифметических операций требуется выполнить для получения решения системы (1.3) по методу Гаусса. Анализируя формулы (1.5) перехода от системы (1.3) к системе (1.4), нетрудно подсчитать, что их реализация требует $n + 2(n-1)^2 + 2(n-1) = 2n^2 - n$ операций. Аналогично, переход от системы (1.4) к системе (1.6) требует $2(n-1)^2 - (n-1)$ операций. Таким образом, для построения системы (1.7) потребуется $2n^2 - n + 2(n-1)^2 - (n-1) + 2(n-2)^2 - (n-2) + \ldots + 1 = 2(1 + 2^2 + \ldots + n^2) - (1 + 2 + \ldots + n) + 1$ операций. Используя хорошо известные формулы:

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

получим, что для реализации прямого хода метода Гаусса требуется $Q = n(n+1)(2n+1)/3 - n(n+1)/2 + 1$ арифметических операций. Ясно, что $Q \approx 2n^3/3$ при больших n. Вычисления по формулам (1.8) требуют выполнения $2(1 + 2 + \ldots + n - 1) = n(n-1)$ арифметических операций. Итак, при больших n для реализации метода Гаусса потребуется примерно $2n^3/3$ операций. Это существенно меньше, чем при использовании формул Крамера.

1.3. Условия применимости метода Гаусса. Описанный метод решения системы линейных уравнений может быть реализован лишь в том случае, когда все числа $a_{11}, a_{22}^{(1)}, \ldots, a_{nn}^{(n-1)}$, называемые ведущими или главными элементами метода Гаусса, отличны от нуля. Выделим класс матриц, для которых это условие выполняется. Пусть

$$A_1 = a_{11},\ A_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix},\ \ldots, A_n = \begin{vmatrix} a_{11} & a_{22} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}$$

есть главные миноры матрицы A.

Теорема 1.1. *Для того, чтобы все ведущие элементы метода Гаусса были отличны от нуля, необходимо и достаточно, чтобы все главные миноры матрицы A были ненулевыми.*

Доказательство. Пусть все главные миноры матрицы A отличны от нуля. Покажем, что тогда все ведущие элементы метода Гаусса не равны нулю. Имеем, в частности, $a_{11} = A_1 \neq 0$. Применяя преобразования, выполнявшиеся при проведении прямого хода метода Гаусса (см. формулы (1.5)), получим

$$A_2 = a_{11} \begin{vmatrix} 1 & \dfrac{a_{12}}{a_{11}} \\ 0 & a_{22} - \dfrac{a_{12}}{a_{11}} a_{21} \end{vmatrix} = a_{11} a_{22}^{(1)},$$

следовательно, $a_{22}^{(1)} \neq 0$. Пусть уже доказано, что $a_{11}, a_{22}^{(1)}, \ldots, a_{k-1,k-1}^{(k-2)}$ не равны нулю. Тогда, приводя минор A_k к треугольному виду при помощи преобразований прямого хода метода Гаусса, получим

$$A_k = a_{11} a_{22}^{(1)} \ldots a_{k-1,k-1}^{(k-2)} \begin{vmatrix} 1 & a_{12}^{(1)} & \ldots & a_{1k}^{(1)} \\ 0 & 1 & \ldots & a_{2k}^{(2)} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & a_{kk}^{(k-1)} \end{vmatrix} =$$

$$= a_{11} a_{22}^{(1)} \ldots a_{k-1,k-1}^{(k-2)} a_{kk}^{(k-1)}, \quad (1.9)$$

следовательно, $a_{kk}^{(k-1)} \neq 0$. Обратное утверждение теоремы есть очевидное следствие соотношения (1.9). $\square$[1)]

[1)] Значком $\square$, как обычно, отмечаем конец доказательства.

Приведем часто встречающиеся в приложениях примеры матриц, для которых метод Гаусса применим.

1) Симметричные положительно определенные матрицы. Напомним, что матрица A называется симметричной, если она совпадает с транспонированной к ней матрицей A^T. Матрица A называется положительно определенной, если $(Ax, x) > 0 \ \forall x \neq 0$. Как обычно, $(x, y) = \sum\limits_{i=1}^{n} x_i y_i$ — скалярное произведение векторов. В соответствии с критерием Сильвестра симметричная матрица положительно определена тогда, и только тогда, когда все ее главные миноры положительны.

2) Матрицы с диагональным преобладанием. Так называют матрицы, элементы которых удовлетворяют условию:

$$\sum_{j=1,\ j\neq i}^{n} |a_{ij}| < |a_{ii}|, \quad i = 1, 2, \ldots, n. \tag{1.10}$$

Теорема 1.2. *Пусть матрица A — матрица с диагональным преобладанием. Тогда все ее главные миноры отличны от нуля.*

Доказательство. Рассмотрим главный минор A_k, $k \geqslant 1$. Достаточно убедиться, что однородная система линейных уравнений

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \ldots + a_{1k}x_k = 0, \\ &a_{21}x_1 + a_{22}x_2 + \ldots + a_{2k}x_k = 0, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kk}x_k = 0 \end{aligned} \tag{1.11}$$

с матрицей, составленной из элементов минора A_k, имеет только тривиальное решение. Предположим противное и пусть $\max\limits_{1\leqslant j\leqslant k} |x_j| = |x_i|$. Ясно, что $x_i \neq 0$. Иначе решение $x_1, \ldots, x_k$ было бы тривиальным. Запишем i-е уравнение системы (1.11):

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{ik}x_k = 0.$$

Отсюда

$$a_{ii}x_i = -a_{i1}x_1 - \ldots - a_{i,i-1}x_{i-1} - a_{i,i+1}x_{i+1} - \ldots - a_{i,k}x_k,$$

следовательно,

$$|a_{ii}||x_i| \leqslant |a_{i1}||x_1| + \ldots + |a_{i,i-1}||x_{i-1}| + |a_{i,i+1}||x_{i+1}| + \ldots + |a_{ik}||x_k| \leqslant$$

$$\leqslant |x_i| \sum_{j=1,\ j\neq i}^{k} |a_{ij}| \leqslant |x_i| \sum_{j=1,\ j\neq i}^{n} |a_{ij}|,$$

поэтому $|a_{ii}| \leqslant \sum\limits_{j=1,\ j\neq i}^{n} |a_{ij}|$, что противоречит условию (1.10). $\square$

1.4. Метод Гаусса с выбором главных элементов. В тех случаях, когда заранее трудно установить, что все ведущие элементы метода Гаусса отличны от нуля, применяются модификации метода Гаусса, позволяющие избежать деления на нуль в случае произвольной невырожденной матрицы A. Опишем вариант такого метода, называемый методом Гаусса с выбором главного элемента по строке.

Найдем максимальный по модулю элемент первой строки матрицы A. Пусть это есть a_{1j}. Ясно, что $a_{1j} \neq 0$, иначе все элементы первой строки матрицы равны нулю и матрица оказывается вопреки нашему предположению вырожденной. Поменяем теперь местами первый и j-й столбцы системы (1.3) и выполним первый шаг прямого хода метода Гаусса. При этом будет выполняться деление на $a_{1j} \neq 0$. Затем проводится поиск максимального по модулю элемента во второй строке системы вида (1.4). Он вновь оказывается ненулевым, поскольку матрица A невырождена. Выполняется перестановка соответствующего столбца системы (1.4) с ее вторым столбцом, затем — второй шаг прямого хода метода Гаусса. Аналогичные вычисления проводятся до приведения системы (1.3) к треугольному виду.

При программной реализации этого метода перестановки столбцов заменяются соответствующими перенумерациями неизвестных.

Ясно, что аналогичный алгоритм можно построить, проводя каждый раз поиск максимального элемента в соответствующем столбце матрицы. Можно было бы проводить поиск максимального элемента и по всей матрице.

Отметим, что выбор главного элемента не только обеспечивает реализуемость метода Гаусса при любой невырожденной матрице A, но и уменьшает влияние погрешностей округления, так как приводит к делению на бо̀льшие числа, чем в основном алгоритме метода Гаусса. Понятно, что это требует дополнительных затрат на поиск максимального элемента и запоминание перенумераций неизвестных. Особенно значительными эти затраты становятся при поиске ведущего элемента по всей матрице. Однако этот вариант метода Гаусса оказывается наиболее устойчивым по отношению к погрешностям округления.

1.5. Метод Гаусса и разложение матрицы на треугольные множители. На практике чаще всего метод Гаусса реализуется в форме разложения матрицы на треугольные множители. Возможность такого разложения гарантирует

Теорема 1.3. *Пусть все главные миноры матрицы A отличны от нуля. Тогда существуют невырожденные нижняя и верхняя треугольные матрицы*

$$L = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{12} & l_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix}, \quad U = \begin{pmatrix} 1 & u_{12} & \dots & u_{1n} \\ 0 & 1 & \dots & u_{n2} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

(все диагональные элементы матрицы U равны единице) такие, что $A = LU$.

Прежде чем доказывать теорему, заметим, что если матрицы L, U построены, то решение системы (1.1) может быть выполнено следующим образом. Запишем систему (1.1) в виде

$$LUx = b.$$

Положим

$$Ux = y. \tag{1.12}$$

Тогда

$$Ly = b. \tag{1.13}$$

Система (1.13) — система с треугольной матрицей. Ее решение может быть построено при помощи обратного хода метода Гаусса. Система (1.12) при уже найденном векторе y также может быть решена при помощи обратного хода метода Гаусса (заметим только, что неизвестные $x_1, \dots, x_n$ находятся при этом начиная с первого). Таким образом, если матрицы L, U построены, то для решения системы (1.1) придется потратить примерно $2n^2$ арифметических операций.

Особенно полезна такая модификация метода Гаусса при необходимости решать множество систем с одной и той же матрицей и различными правыми частями.

На практике такая ситуация возникает довольно часто. Объясняется это тем, что матрица A, обычно, описывает структуру некоторой системы (механической, экономической и т. д.), а вектор b характеризует внешние воздействия на эту систему. Поэтому при изучении реакции системы на различные внешние воздействия приходится решать множество систем с одной и той же матрицей.

Доказательство теоремы 1.3. Используем индукцию по порядку матрицы A. Для матрицы порядка $k = 1$ утверждение теоремы тривиально. Пусть оно верно для некоторого $k \geqslant 1$. Докажем, что тогда оно будет верным и для матрицы порядка $k + 1$. Запишем матрицу A_{k+1} порядка $k + 1$ в блочном виде

$$A_{k+1} = \begin{pmatrix} A_k & a_k \\ b_k & a_{k+1,k+1} \end{pmatrix}. \tag{1.14}$$

Здесь A_k — матрица порядка k, a_k — столбец длины k, b_k — строка длины k. Матрица A_k по условию теоремы невырождена и по индуктивному предположению может быть представлена в виде $A_k = L_k U_k$, где L_k — нижняя треугольная матрица с ненулевыми элементами на диагонали, а U_k — верхняя треугольная матрица с единичными диагональными элементами. Будем искать матрицу A_{k+1} в виде произведения двух блочных матриц:

$$A_{k+1} = \begin{pmatrix} L_k & 0 \\ l_k & l_{k+1,k+1} \end{pmatrix} \begin{pmatrix} U_k & u_k \\ 0 & 1 \end{pmatrix}. \tag{1.15}$$

Здесь l_k — искомая строка длины k, u_k — искомый столбец длины k. Подсчитывая произведение сомножителей в правой части равенства (1.15) и приравнивая результат поблочно матрице A_{k+1} (см. (1.14)), получим

$$L_k U_k = A_k,\ l_k U_k = b_k,\ L_k u_k = a_k,\ l_k u_k + l_{k+1,k+1} = a_{k+1,k+1}.$$

Отсюда можно найти l_k, u_k, решая системы уравнений

$$U_k^T l_k^T = b_k^T, \tag{1.16}$$

$$L_k u_k = a_k, \tag{1.17}$$

и затем вычислить

$$l_{k+1,k+1} = a_{k+1,k+1} - l_k u_k. \tag{1.18}$$

При этом $l_{k+1,k+1}$ не может обратиться в нуль, так как вследствие равенства (1.15) имеем $|A_{k+1}| = |L_k| l_{k+1,k+1}$, а определитель $|A_{k+1}|$ по сделанному нами предположению отличен от нуля. Таким образом, искомое треугольное разложение матрицы A_{k+1} построено. □

Доказательство теоремы 1.3, фактически, дает алгоритм построения матриц L и U: нужно последовательно строить разложения диагональных блоков матрицы порядков $k = 1, 2, \ldots, n$, используя формулы (1.16)–(1.18); k-й шаг такого алгоритма состоит в решении систем с треугольными матрицами и требует примерно $2k^2$ операций.

Таким образом, построение матриц L, U требует примерно $2n^3/3$ операций.

Нетрудно проверить, что матрица U совпадает с матрицей системы (1.7).

Если нет необходимости сохранять матрицу A, то программу вычислений можно организовать так, что элементы матриц L, U будут последовательно замещать соответствующие элементы матрицы A.

Аналогично методу Гаусса с выбором главных элементов можно строить алгоритмы разложения на треугольные множители, применимые для любой невырожденной матрицы.

2. Метод отражений. Вместо разложения матрицы на треугольные множители при решении системы уравнений можно использовать представление матрицы A в виде $A = QU$, где Q — ортогональная матрица, т. е. матрица, удовлетворяющая условию $QQ^T = E$, где E — единичная матрица, а U — верхняя треугольная матрица. Если такое разложение получено, то решение системы (1.1) сводится к последовательному решению систем $Qy = b$, $Ux = y$. При этом ясно, что $y = Q^T b$, а x находится при помощи обратного хода метода Гаусса. Если матрица Q известна, то вычисление вектора y требует $2n^2$ арифметических операций. Таким образом, после получения разложения матрицы A решения системы уравнений требует примерно $3n^2$ операций.

Для построения указанного разложения будем использовать матрицы отражения, т. е. матрицы вида $R = E - 2ww^T$, где w — единичный вектор: $(w, w) = |w|^2 = 1$. Матрица R при любом $|w| = 1$ симметрична и ортогональна. Действительно,

$$R^T = R, \quad R^2 = E - 4ww^T + 4ww^T ww^T = E,$$

так как $w^T w = |w|^2 = 1$. Заметим, далее, что

$$Rw = w - 2ww^T w = -w, \quad Rz = z - 2ww^T z = z, \tag{2.1}$$

если $w^T z = (w, z) = 0$, т. е. векторы w и z ортогональны.

Пусть теперь x — произвольный вектор. По теореме об ортогональном разложении евклидова пространства он однозначно представим в виде $x = \alpha w + z$, где α некоторое число, z — некоторый вектор, ортогональный w. Из равенств (2.1) вытекает, что $Rx = -\alpha w + z$.

Можно сказать, таким образом, что матрица R выполняет отражение вектора x относительно $(n-1)$-мерной гиперплоскости, ортогональной вектору w. Это свойство матрицы R и позволяет называть ее матрицей отражения.

Рассмотрим теперь следующую задачу. Даны ненулевой вектор x и единичный вектор e. Требуется построить матрицу отражения R, такую, что $Rx = \mu e$, где μ — число (ясно, что $|\mu| = |x|$, поскольку матрица R ортогональна).

Нетрудно видеть (сделайте чертеж!), что решение задачи — матрица отражения, определяемая вектором $w = (x - |x|e)/|x - |x|e|$ или вектором $w = (x + |x|e)/|x + |x|e|$.

При вычислениях для минимизации погрешностей округления следует выбрать вектор w с бо̀льшим знаменателем.

Теорема 2.1. *Пусть A — произвольная невырожденная матрица. Тогда существует ортогональная матрица Q такая, что $A = QU$, где U — верхняя треугольная матрица с ненулевыми диагональными элементами.*

Доказательство. Пусть a^1 — первый столбец матрицы A, $e^1 = (1, 0, \ldots, 0)^T$, $w^1 = (a^1 \pm |a^1|e^1)/|a^1 \pm |a^1|e^1||$ (знак выбирается, как указано выше), $R_1 = E - 2w^1 w^{1T}$. Образуем матрицу $A_1 = \left(a_{ij}^{(1)}\right)_{i,j=1}^n = R_1 A$. Ясно, что первый столбец этой матрицы коллинеарен вектору e^1 и, следовательно, имеет вид $(a_{11}^{(1)}, 0, \ldots, 0)^T$, причем $|a_{11}^{(1)}| = |a^1| \neq 0$.

Рассмотрим столбец $\widetilde{a}^2 = (a_{22}^{(1)}, a_{32}^{(1)}, \ldots, a_{n2}^{(1)})^T$ длины $n-1$. Понятно, что $\widetilde{a}^2 \neq 0$, так как в противном случае $|A_1| = 0$, но, с другой стороны, поскольку определитель ортогональной матрицы равен ± 1, то $|A_1| = \pm|A|$, а матрица A по предположению невырождена. Пусть $\widetilde{w} = (\widetilde{a}^2 \pm |\widetilde{a}^2|\widetilde{e}^1)/|\widetilde{a}^2 \pm |\widetilde{a}^2|\widetilde{e}^1||$, где $\widetilde{e}^1 = (1, 0, \ldots, 0)$ — вектор длины $n-1$. Положим $\widetilde{R}_2 = \widetilde{E} - 2\widetilde{w}\widetilde{w}^T$, где $\widetilde{E}$ — единичная матрица размерности $n-1$, и образуем ортогональную матрицу

$$R_2 = \begin{pmatrix} 1 & 0 \\ 0^T & \widetilde{R}_2 \end{pmatrix}.$$

Здесь 0 — нулевая строка длины $n-1$. Пусть $A_2 = \left(a_{ij}^{(2)}\right)_{i,j=1}^n = R_2 A_1$. Нетрудно видеть, что первые столбцы матриц A_1, A_2 совпадают, а второй столбец матрицы A_2 имеет вид $(a_{12}^{(1)}, a_{22}^{(2)}, 0, \ldots, 0)^T$, причем $|a_{22}^{(2)}| = |\widetilde{a}^2| \neq 0$.

Выполняя аналогичные построения, получим ортогональные матрицы $R_3, R_4, \ldots, R_n$ такие, что $R_n R_{n-1} \cdots R_1 A = A_n$, где A_n — верхняя треугольная матрица с ненулевыми элементами $a_{11}^{(1)}, a_{22}^{(2)}, \ldots, a_{nn}^{(n)}$ на диагонали, следовательно, $A = QU$, где $U = A_n, Q = R_1^T R_2^T \cdots R_n^T$ суть ортогональная матрица. $\square$

Доказательство теоремы 2.1, фактически, дает алгоритм построения матриц Q, U. Нетрудно убедиться, что их вычисление требует приблизительно в два раза больше арифметических операций, чем разложение матрицы на треугольные множители. Важным положительным качеством описанного алгоритма является возможность его применения для произвольной невырожденной матрицы без какой-либо перенумерации строк или столбцов.

3. Метод Холесского. В случае, когда матрица системы линейных уравнений симметрична и положительно определена, можно добиться существенного сокращения числа операций и памяти, необходимых для вычисления решения. В основе соответствующего алгоритма лежит

Теорема 3.1. *Пусть матрица A симметрична и положительно определена. Тогда существует нижняя треугольная матрица L с положительными элементами на диагонали такая, что $A = LL^T$.*

Доказательство. Используем индукцию по порядку матрицы A. Для матрицы порядка $k = 1$ имеем тривиальное равенство $A_1 = a_{11} = \sqrt{a_{11}}\sqrt{a_{11}}$. Пусть нужное разложение получено для некоторого $k \geqslant 1$. Покажем, как его построить для матрицы порядка $k+1$. Запишем матрицу A_{k+1} как блочную:

$$A_{k+1} = \begin{pmatrix} A_k & a_k \\ a_k^T & a_{k+1,k+1} \end{pmatrix}.$$

В силу предположения индукции $A_k = L_k L_k^T$, где L_k — нижняя треугольная матрица с положительными элементами на диагонали. Будем искать разложение матрицы A на треугольные множители в виде

$$A_{k+1} = L_{k+1}L_{k+1}^T = \begin{pmatrix} L_k & 0 \\ l_k^T & l_{k+1,k+1} \end{pmatrix} \begin{pmatrix} L_k^T & l_k \\ 0 & l_{k+1,k+1} \end{pmatrix}. \qquad (3.1)$$

Выполняя умножение в правой части последнего равенства и сравнивая поблочно результат с матрицей A_{k+1}, получим систему линейных уравнений

$$L_k l_k = a_k \qquad (3.2)$$

для определения вектора l_k и уравнение $l_k^T l_k + l_{k+1,k+1}^2 = a_{k+1,k+1}$ для элемента $l_{k+1,k+1}$. Можно считать, что $l_{k+1,k+1} > 0$, так как вследствие (3.1) имеем: $|A_{k+1}| = |A_k|\, l_{k+1,k+1}^2$, причем $|A_k|, |A_{k+1}| > 0$, так как по условию матрицы A_k, A_{k+1} положительно определены. Таким образом, для построения матрицы L_{k+1} нужно решить систему уравнений (3.2) с треугольной матрицей, а затем вычислить $l_{k+1,k+1}$ по формуле $l_{k+1,k+1} = \sqrt{a_{k+1,k+1} - l_k^T l_k}$. $\square$

Доказательство теоремы 3.1, фактически, описывает алгоритм разложения на треугольные множители произвольной симметричной положительно определенной матрицы. Нетрудно видеть, что его реализация по затратам памяти и объему вычислений оказывается в два раза более экономичной, чем разложение на треугольные множители произвольной невырожденной матрицы.

После того, как матрица L построена, решение задачи (1.1) сводится к последовательному решению систем уравнений $Ly = b$, $L^T x = y$ с треугольными матрицами.

4. Вычисление определителя и обратной матрицы. Изложенные выше методы решения систем линейных уравнений позволяют решать и эти задачи.

Так, если применяется метод Гаусса, то попутно можно вычислить и $|A| = a_{11}a_{22}^{(1)} \dots a_{nn}^{(n-1)}$ (см. (1.9)). Аналогичные формулы нетрудно написать и в тех случаях, когда строится разложение матрицы A на множители. Надо, однако, иметь в виду, что непосредственные вычисления по этим формулам, как правило, оказываются невозможными: из-за большого числа сомножителей определитель (или результат промежуточных вычислений) зачастую либо слишком велик, либо, наоборот, слишком мал. Приходится писать специальные программы, позволяющие отдельно подсчитывать мантиссу и порядок определителя.

Построение обратной матрицы сводится к решению n систем линейных уравнений с одной и той же матрицей A и различными правыми частями. Действительно, обозначим матрицу A^{-1} через X. Тогда $AX = E$. Осталось записать это равенство подробнее:

$$Ax^k = e^k,\ k = 1, 2, \dots, n.$$

Здесь x^k — k-й столбец матрицы X,

$$e^k = (\underbrace{0, \dots, 0}_{k-1}, 1, \underbrace{0, \dots, 0}_{n-k}).$$

5. Метод прогонки для систем с трехдиагональными матрицами. В приложениях довольно часто возникают системы уравнений с матрицами, большинство элементов которых — нули. Это так называемые разреженные матрицы. Процесс исключения неизвестных в таких системах (или разложение матриц на треугольные множители) во многих практически важных ситуациях удается организовать так, чтобы существенно сократить память и объем вычислений.

Рассмотрим наиболее простой случай, а именно системы с матрицами, ненулевые элементы которых лежат лишь на главной и двух соседних с ней диагоналях. Системы такого вида часто возникают при приближенном решении задач математической физики. Соответствующие матрицы принято называть трехдиагональными.

Произвольную систему с трехдиагональной матрицей можно записать в следующем виде:

$$\begin{aligned} -b_1x_1 + c_1x_2 &= f_1,\\ a_2x_1 - b_2x_2 + c_2x_3 &= f_2,\\ \dots\dots\dots\dots\dots\dots\dots\\ a_ix_{i-1} - b_ix_i + c_ix_{i+1} &= f_i,\\ \dots\dots\dots\dots\dots\dots\dots\\ -a_{n-1}x_{n-1} + b_nx_n &= f_n. \end{aligned} \tag{5.1}$$

Разрешим первое уравнение системы относительно x_1. Получим:

$$x_1 = P_2x_2 + Q_2, \tag{5.2}$$

где

$$P_2 = \frac{c_1}{b_1}, \quad Q_2 = -\frac{f_1}{b_1}. \tag{5.3}$$

Используя соотношение (5.2) и второе уравнение системы (5.1), получим аналогичное выражение для x_2. Вообще, если $x_{i-1} = P_ix_i + Q_i$, то из i-го уравнения системы (5.1) получим

$$x_i = P_{i+1}x_{i+1} + Q_{i+1}, \quad i = 1, 2, \dots, n-1, \tag{5.4}$$

где

$$P_{i+1} = \frac{c_i}{b_i - a_iP_i}, \quad Q_{i+1} = \frac{a_iQ_i - f_i}{b_i - a_iP_i}. \tag{5.5}$$

Это означает, что формулы (5.4) справедливы для $i = 1, 2, \dots, n-1$, формулы (5.5) — для $i = 2, 3, \dots, n-1$.

Используя (5.3) и (5.5), можно найти все $P_i, Q_i,\ i = 2, \dots, n-1$. Записывая теперь соотношение (5.4) при $i{=}n{-}1$ и последнее уравнение системы (5.1), получим

$$\begin{aligned} x_{n-1} &= P_nx_n + Q_n,\\ -a_{n-1}x_{n-1} + b_nx_n &= f_n, \end{aligned}$$

откуда находим, что $x_n = (a_nQ_n - f_n)/(b_n - a_nP_n)$, и, наконец, используя формулы (5.4) для $i = n-1, n-2, \dots, 1$, найдем все остальные компоненты вектора x.

Описанный алгоритм носит название метода прогонки. Понятно, что это — метод Гаусса, записанный применительно к случаю трехдиагональной системы уравнений, причем процесс вычислений P_i, Q_i (прямой ход метода прогонки) соответствует прямому ходу метода Гаусса, а вычисления по формулам (5.4) (обратный ход метода прогонки) соответствуют обратному ходу метода Гаусса.

Нетрудно подсчитать необходимые затраты: требуется примерно $8n$ арифметических операций и не более $6n$ ячеек памяти.

Метод может быть реализован, когда все знаменатели в формулах (5.3), (5.5) отличны от нуля. Учитывая связь метода прогонки с методом Гаусса, можно сказать, что данное условие выполнено, например, когда матрица системы (5.1) — матрица с диагональным преобладанием, т. е. $|c_1| < |b_1|$, $|a_n| < |b_n|$, $|a_i| + |c_i| < |b_i|$, $i = 2, \ldots, n-1$.

§ 2. Итерационные методы решения систем линейных уравнений

Рассмотренные выше методы решения систем линейных алгебраических уравнений принято называть прямыми методами. Все они характеризуются тем, что если пренебречь ошибками округления, то решение системы может быть получено за конечное число арифметических операций (зависящее лишь от порядка системы).

При реализации прямых методов важно, чтобы все данные располагались в оперативной (быстрой) памяти компьютера. Если порядок системы настолько велик, что ее матрица может быть размещена только во внешней (медленной) памяти, например, на жестком диске, то время, затрачиваемое на решение системы, существенно увеличивается.

Поэтому для больших систем предпочтительнее оказываются итерационные методы. Основная идея этих методов состоит в построении последовательности векторов x^k, $k = 1, 2, \ldots$, сходящейся к решению системы (1.1). За приближенное решение принимается вектор x^k при достаточно большом k. При реализации итерационных методов, обычно, достаточно уметь вычислять вектор Ax при любом заданном векторе x.

1. Методы Зейделя и Якоби. Будем считать, что все диагональные элементы матрицы A отличны от нуля, и перепишем систему (1.1), разрешая каждое уравнение относительно переменной,

стоящей на диагонали:

$$x_i = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j + \frac{b_i}{a_{ii}}, \quad i = 1, 2, \ldots, n. \tag{1.1}$$

Выберем некоторое начальное приближение $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)^T$ и построим последовательность векторов $x^1, x^2, \ldots$, определяя вектор x^{k+1} по уже найденному вектору x^k при помощи соотношений:

$$x_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^k - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^k + \frac{b_i}{a_{ii}}, \quad i = 1, 2, \ldots, n. \tag{1.2}$$

Формулы (1.2) определяют итерационный метод решения системы (1.1), называемый методом Якоби или методом простой итерации.

Укажем легко проверяемое достаточное условие сходимости этого метода.

Теорема 1.1. *Пусть матрица A — матрица с диагональным преобладанием. Тогда итерационный метод Якоби сходится при любом начальном приближении x^0.*

Доказательство. Заметим, прежде всего, что условие диагонального преобладания (1.10) означает, что

$$\max_{1 \leqslant i \leqslant n} \sum_{j=1,\ j \neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} = q < 1. \tag{1.3}$$

Пусть x — решение системы уравнений (1.1). Здесь и всюду в дальнейшем погрешность метода на k-м шаге итераций, т. е. вектор $x^k - x$, будем обозначать через z^k. Вычитая почленно из равенства (1.2) равенство (1.1), получим

$$z_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} z_j^k - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} z_j^k, \quad i = 1, 2, \ldots, n,$$

следовательно,

$$|z_i^{k+1}| \leqslant \sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|} |z_j^k| + \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|} |z_j^k| \leqslant \left(\sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|} + \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|} \right) \max_{1 \leqslant j \leqslant n} |z_j^k| =$$

$$= q \max_{1 \leqslant j \leqslant n} |z_j^k|,$$

$i = 1, 2, \ldots, n$, откуда вытекает, что

$$\max_{1\leqslant j\leqslant n} |z_j^{k+1}| \leqslant q \max_{1\leqslant j\leqslant n} |z_j^k|$$

для любого $k = 0, 1, \ldots$, поэтому

$$\max_{1\leqslant j\leqslant n} |z_j^k| \leqslant q^k \max_{1\leqslant j\leqslant n} |z_j^0| \to 0 \tag{1.4}$$

при $k \to \infty$, поскольку $0 < q < 1$, а это и означает, что $x^k \to x$. □

Оценка (1.4) показывает, что, чем меньше q, тем быстрее сходится метод простой итерации.

Формулы (1.2) допускают естественную модификацию. Именно, при вычислении x_i^{k+1} будем использовать уже найденные компоненты вектора x^{k+1}, то есть $x_1^{k+1}, x_2^{k+1}, \ldots x_{i-1}^{k+1}$. В результате приходим к итерационному методу Зейделя:

$$x_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{k+1} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^k + \frac{b_i}{a_{ii}}, \quad i = 1, 2, \ldots, n, \tag{1.5}$$

$k = 0, 1, \ldots$

Метод Зейделя позволяет более экономно расходовать память, поскольку в данном случае вновь получаемые компоненты вектора x^{k+1} можно размещать на месте соответствующих компонент вектора x^k, в то время как при реализации метода Якоби все компоненты векторов x^k, x^{k+1} должны одновременно находиться в памяти компьютера.

Достаточное условие сходимости и оценку скорости сходимости метода Зейделя дает

Теорема 1.2. *Пусть матрица A — матрица с диагональным преобладанием. Тогда метод Зейделя сходится при любом начальном приближении x^0; справедлива оценка скорости сходимости:*

$$\max_{1\leqslant j\leqslant n} |z_j^k| \leqslant q^k \max_{1\leqslant j\leqslant n} |z_j^0|. \tag{1.6}$$

Доказательство. Вычитая почленно из равенства (6.2) равенство (1.1), получим

$$z_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} z_j^{k+1} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} z_j^k, \quad i = 1, 2, \ldots, n. \tag{1.7}$$

Пусть $\max\limits_{1\leqslant j\leqslant n} |z_j^{k+1}| = |z_l^{k+1}|$. Из l-го уравнения системы (1.7) вытекает, что

$$|z_l^{k+1}| \leqslant \alpha_l \max_{1\leqslant j\leqslant n} |z_j^{k+1}| + \beta_l \max_{1\leqslant j\leqslant n} |z_j^k|,$$

где

$$\alpha_l = \sum_{j=1}^{l-1} \frac{|a_{lj}|}{|a_{ll}|}, \quad \beta_l = \sum_{j=l+1}^{n} \frac{|a_{lj}|}{|a_{ll}|},$$

следовательно,

$$\max_{1\leqslant j\leqslant n} |z_j^{k+1}| \leqslant \frac{\beta_l}{1-\alpha_l} \max_{1\leqslant j\leqslant n} |z_j^k|.$$

Из условия (1.3) получаем, что $\alpha_l+\beta_l \leqslant q < 1$, но тогда и $q\alpha_l+\beta_l \leqslant q$, таким образом, $\beta_l/(1-\alpha_l) \leqslant q$. $\square$

2. Метод релаксации. Во многих ситуациях существенного ускорения сходимости можно добиться за счет введения так называемого итерационного параметра. Рассмотрим итерационный процесс

$$x_i^{k+1} = (1-\omega)x_i^k + \omega\left(-\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}x_j^{k+1} - \sum_{j=i+1}^{n}\frac{a_{ij}}{a_{ii}}x_j^k + \frac{b_i}{a_{ii}}\right), \qquad (2.1)$$

$i = 1, 2, \ldots, n$, $k = 0, 1, \ldots$ Этот метод называется методом релаксации, число ω — релаксационным параметром. При $\omega = 1$ метод переходит в метод Зейделя.

Ясно, что по затратам памяти и объему вычислений на каждом шаге итераций метод релаксации не отличается от метода Зейделя.

Мы исследуем сходимость метода релаксации в случае, когда матрица A симметрична и положительно определена. С этой целью перепишем его в матричном виде. Обозначим через L нижнюю треугольную матрицу с нулевой главной диагональю; элементы, стоящие под главной диагональю матрицы L, совпадают с соответствующими элементами матрицы A. Через D обозначим диагональную матрицу, на диагонали которой стоят диагональные элементы матрицы A. Понятно, что $A = L + D + L^T$. Нетрудно убедиться, что равенства (2.1) с учетом введенных обозначений принимают вид:

$$Dx^{k+1} = (1-\omega)Dx^k + \omega(-Lx^{k+1} - L^Tx^k + b).$$

После элементарных преобразований получим, что

$$B\frac{x^{k+1}-x^k}{\omega} + Ax^k = b, \qquad (2.2)$$

где $B = D + \omega L$.

Нам потребуется следующая общая теорема, полезная при исследовании многих итерационных методов.

Теорема 2.1. *Пусть матрица A симметрична и положительно определена. Тогда итерационный метод*

$$B\frac{x^{k+1}-x^k}{\tau}+Ax^k=b, k=0,1,\ldots \tag{2.3}$$

где $\tau>0$, сходится при любом начальном приближении x^0, если

$$(Bx,x)>\frac{\tau}{2}(Ax,x)\quad \forall x\neq 0. \tag{2.4}$$

Доказательство. Если x — решение уравнения (1.1), то, очевидно,

$$B\frac{x-x}{\tau}+Ax=b. \tag{2.5}$$

Вычитая почленно из равенства (2.5) равенство (2.3), получим

$$B\frac{z^{k+1}-z^k}{\tau}+Az^k=0. \tag{2.6}$$

Используя тривиальное равенство $z^k=\dfrac{z^{k+1}+z^k}{2}-\dfrac{z^{k+1}-z^k}{2}$, преобразуем (2.6) к виду

$$\left(B-\frac{\tau}{2}A\right)\frac{z^{k+1}-z^k}{\tau}+\frac{1}{2}A(z^{k+1}+z^k)=0. \tag{2.7}$$

Умножим обе части равенства (2.7) скалярно на вектор $2(z^{k+1}-z^k)$. После элементарных преобразований с учетом симметрии матрицы A получим:

$$2\tau\left(\left(B-\frac{\tau}{2}A\right)\frac{z^{k+1}-z^k}{\tau},\frac{z^{k+1}-z^k}{\tau}\right)+ \\ +(Az^{k+1},z^{k+1})-(Az^k,z^k)=0. \tag{2.8}$$

Вследствие условия (2.4) первое слагаемое в правой части (2.8) неотрицательно, поэтому $(Az^{k+1},z^{k+1})\leqslant(Az^k,z^k)$, т. е. числовая последовательность (Az^k,z^k) не возрастает. Кроме того, она ограничена снизу нулем, так как матрица A положительно определена. Таким образом, последовательность (Az^k,z^k) имеет предел. Отсюда вытекает, что $(Az^{k+1},z^{k+1})-(Az^k,z^k)\to 0$ при $k\to\infty$, следовательно, и

$$\left(\left(B-\frac{\tau}{2}A\right)(z^{k+1}-z^k),z^{k+1}-z^k\right)\to 0 \text{ при } k\to\infty,$$

а поскольку матрица $B-(\tau/2)A$ положительно определена (см. условие (2.4)), то $z^{k+1}-z^k \to 0$ при $k \to \infty$. Используя теперь уравнение (2.6) и невырожденность матрицы A, получим, что $z^k \to 0$ при $k \to \infty$. $\square$

С использованием теоремы 2.1 просто доказывается

Теорема 2.2. *Пусть матрица A положительно определена, параметр релаксации удовлетворяет условию: $0 < \omega < 2$. Тогда метод релаксации сходится при любом начальном приближении x^0.*

Доказательство. Покажем, что при сделанных предположениях выполнено условие (2.4) при $\tau = \omega,\ B = D + \omega L$. Действительно,

$$(Bx,x) - \frac{\omega}{2}(Ax,x) = \left(1 - \frac{\omega}{2}\right)(Dx,x) + \frac{\omega}{2}((Lx,x) - (L^Tx,x)),$$

но все диагональные элементы положительно определенной матрицы положительны (докажите!), поэтому $(Dx,x) > 0$ при $x \neq 0$, а $(Lx,x) - (L^Tx,x) = (Lx,x) - (x,Lx) = 0\ \forall x$. $\square$

Естественно, параметр ω следует выбирать так, чтобы метод релаксации сходился наиболее быстро. Решение этой задачи далеко выходит за рамки нашего курса. Отметим только, что чаще всего оптимальное значение ω лежит вблизи $1,8$.

3. Пример решения задачи оптимизации итерационного параметра. Рассмотрим итерационный метод (2.3) при $B = E$. Этот метод называют методом простой итерации с параметром. Будем предполагать, что A — симметричная положительно определенная матрица, и выясним, при каких условиях на τ метод сходится. Найдем также значение τ, обеспечивающее наиболее быструю сходимость.

Напомним нужные в дальнейшем сведения из линейной алгебры. Все характеристические числа симметричной матрицы A вещественны. Будем нумеровать их в порядке неубывания: $\lambda_1 \leqslant \lambda_2, \ldots \leqslant \lambda_n$. Существует полная ортонормированная система собственных векторов $e^1, e^2 \ldots, e^n$ матрицы A:

$$Ae_k = \lambda_k e_k,\ k = 1,2,\ldots,n, (e_k,e_l) = \begin{cases} 0, & k \neq l; \\ 1, & k = l. \end{cases}$$

Очевидно, что $\lambda_k = (Ae_k, e_k)$, поэтому для положительно определенной матрицы $\lambda_1 > 0$.

Представим произвольный вектор x в виде разложения по орто-

нормированному базису собственных векторов матрицы A:

$$x = \sum_{i=1}^{n} c_i e_i. \tag{3.1}$$

Тогда $|x|^2 = \sum\limits_{i=1}^{n} c_i^2$, $(Ax, x) = \sum\limits_{i=1}^{n} \lambda_i c_i^2$, следовательно,

$$\lambda_1 |x|^2 \leqslant (Ax, x) \leqslant \lambda_n |x|^2 \quad \forall x. \tag{3.2}$$

Справедлива

Теорема 3.1. *Пусть $B = E$, матрица A симметрична и положительно определена. Тогда итерационный метод (2.3) сходится при любом начальном приближении, если $\tau \in (0, 2/\lambda_n)$. Наилучшая оценка скорости сходимости достигается при $\tau = \tau_0 = 2/(\lambda_1 + \lambda_n)$. В этом случае $|z^k| \leqslant \rho_0^k |z^0|$, где $\rho_0 = (1 - \xi)/(1 + \xi)$, $\xi = \lambda_1/\lambda_n$.*

Доказательство. Запишем уравнение для погрешности метода:

$$\frac{z^{k+1} - z^k}{\tau} + Az^k = 0, \quad k = 0, 1, \ldots$$

откуда $z^{k+1} = (E - \tau A)z^k$. Представим вектор z^k в виде разложения (3.1). Тогда

$$z^{k+1} = \sum_{i=1}^{n} (1 - \tau\lambda_i) c_i e^i.$$

Вследствие ортонормированности векторов $e_1, \ldots, e_n$ имеем

$$|z^{k+1}|^2 = \sum_{i=1}^{n} (1 - \tau\lambda_i)^2 c_i^2, \quad |z^k|^2 = \sum_{i=1}^{n} c_i^2,$$

откуда

$$|z^{k+1}| \leqslant \max_{1 \leqslant i \leqslant n} |1 - \tau\lambda_i| |z^k|.$$

Понятно, что $\max\limits_{1 \leqslant i \leqslant n} |1 - \tau\lambda_i| \leqslant \max\limits_{\lambda_1 \leqslant \lambda \leqslant \lambda_n} |1 - \tau\lambda|$, причем поскольку функция $1 - \tau\lambda$ линейна по λ, то

$$\max_{\lambda_1 \leqslant \lambda \leqslant \lambda_n} |1 - \tau\lambda| = \max(|1 - \tau\lambda_1|, |1 - \tau\lambda_n|).$$

Анализируя график функции $\varphi(\tau) = \max(|1 - \tau\lambda_1|, |1 - \tau\lambda_n|)$, нетрудно убедиться, что $0 < \varphi(\tau) < 1$ при $0 < \tau < 2/\lambda_n$, а

$$\min_{\tau} \varphi(\tau) = \varphi(\tau_0) = \rho_0. \ \square \tag{3.3}$$

4. Итерационные методы вариационного типа. Отыскание оптимального значения параметра τ в рассмотренном выше методе требует предварительного вычисления минимального и максимального собственных чисел матрицы A. Существуют итерационные методы, позволяющие за счет некоторой дополнительной работы на каждом шаге итераций автоматически настраиваться на оптимальную скорость сходимости. К их числу относятся методы, основанные на замене системы (1.1) эквивалентной задачей минимизации некоторого функционала. В дальнейшем существенную роль играет

Теорема 4.1. *Пусть матрица A симметрична и положительно определена. Тогда задача* (1.1) *эквивалентна задаче отыскания минимума квадратичного функционала $F(x) = (Ax, x) - 2(b, x)$.*

Доказательство. Пусть x^* — решение задачи (1.1), то есть $Ax^* = b$. Используя симметрию матрицы A, получим

$$F(x) = (Ax, x) - 2(Ax^*, x) + (Ax^*, x^*) - (Ax^*, x^*) = \\ = (A(x - x^*), (x - x^*)) - (Ax^*, x^*), \quad (4.1)$$

откуда вследствие положительной определенности матрицы A вытекает, что единственной точкой минимума функционала F является x^*. $\square$

Различные методы минимизации функционала $F(x)$ приводят к различным итерационным процессам для уравнения (1.1).

Рассмотрим сначала метод покоординатного спуска. Выберем некоторое начальное приближение $x^0 = (x_1^0, \ldots, x_n^0)$ и найдем аргумент x_1^1, доставляющий минимальное значение функции одной переменной $F(x_1, x_2^0, \ldots, x_n^0)$. Затем рассмотрим функцию одной переменной $F(x_1^1, x_2, x_3^0, \ldots, x_n^0)$ и найдем точку x_2^2, в которой она достигает минимума. Выполнив n таких шагов, построим вектор $(x_1^1, \ldots, x_n^1)$, примем его за начальное приближение и продолжим описанный процесс.

Используя конкретный вид функционала $F(x)$, найдем явные формулы для вычисления векторов x^k, $k = 1, 2 \ldots$ в полученном итерационном процессе. Компонента x_i^{k+1} вектора x^{k+1} разыскивается как точка минимума функции $F(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \ldots, x_n^k)$. Выпишем необходимое условие экстремума:

$$F'_{x_i}(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \ldots, x_n^k) = 0. \quad (4.2)$$

Вычисляя производную функции $F(x)$ по переменной x_i, получим:

$$F'_{x_i}(x) = 2\sum_{j=1}^{n} a_{ij}x_j - 2b_i, \quad (4.3)$$

следовательно, решая уравнение (1.1) относительно x_i, будем иметь

$$x_i^{k+1} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{k+1} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^k + \frac{b_i}{a_{ii}},$$

и это означает, что метод покоординатного спуска для квадратичного функционала совпадает с методом Зейделя.

Метод релаксации также допускает простую геометрическую интерпретацию. При $0 < \omega < 1$ из точки $(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, x_{i+1}^k, \ldots, x_n^k)$ двигаются в направлении координатной оси x_i, не доходя до точки минимума функционала F на этой прямой, а при $\omega > 1$ проходят несколько дальше, чем точка минимума функционала. Во многих случаях последний способ приводит к ускорению сходимости.

Опишем еще один метод минимизации функционала. Будем двигаться из точки начального приближения x^0 в направлении наибыстрейшего убывания функционала F, т. е. следующее приближение разыскиваем так: $x^1 = x^0 - \tau \operatorname{grad} F(x^0)$. Формула (4.3) показывает, что $\operatorname{grad} F(x^0) = 2(Ax^0 - b)$. Вектор $r^0 = Ax^0 - b$ принято называть невязкой. Для сокращения записей удобно обозначить 2τ вновь через τ. Таким образом, $x^1 = x^0 - \tau r^0$.

Параметр τ выберем так, чтобы значение $F(x^1)$ было минимальным. Проводя элементарные выкладки, получим $F(x^1) = F(x^0 - \tau r^0) = F(x^0) - 2\tau(r^0, r^0) + \tau^2(Ar^0, r^0)$, следовательно, минимум $F(x^1)$ достигается при $\tau = \tau_* = (r^0, r^0)/(Ar^0, r^0)$.

Таким образом, мы пришли к следующему итерационному методу

$$x^{k+1} = x^k - \tau_* r^k,\ r^k = Ax^k - b,\ \tau_* = \frac{(r^k, r^k)}{(Ar^k, r^k)},\ k = 0, 1, \ldots \qquad (4.4)$$

Метод (4.4) называют методом наискорейшего спуска. По сравнению с методом простой итерации этот метод требует на каждом шаге итераций проведения дополнительной работы по вычислению параметра τ_*. Вследствие этого происходит адаптация к оптимальной скорости сходимости.

Теорема 4.2. *Пусть матрица A симметрична и положительно определена. Тогда метод* (4.4) *сходится при любом начальном приближении. Справедлива следующая оценка скорости сходимости:*

$$|z^k| \leqslant \sqrt{\lambda_n/\lambda_1}\, \rho_0^k |z^0|. \qquad (4.5)$$

Замечание 4.1. Оценка (4.5) показывает, что погрешность метода (4.4) убывает с той же скоростью, что и погрешность метода простой итерации при оптимальном выборе параметра τ.

Доказательство теоремы 4.2. Используя (4.4) и вспоминая, что

$$F(x^k - \tau_* r^k) = \min_{\tau} F(x^k - \tau r^k),$$

получим: $F(x^{k+1}) \leqslant F(x^k - \tau_0 r^k)$, где $\tau_0 = 2/(\lambda_1 + \lambda_n)$, откуда вследствие представления (4.1) вытекает, что

$$(A(x^{k+1} - x^*), x^{k+1} - x^*) \leqslant (A(x^k - \tau_0 r^k - x^*), x^k - \tau_0 r^k - x^*).$$

Заметим теперь, что

$$x^k - x^* - \tau_0 r^k = z^k - \tau_0(Ax^k - b) = z^k - \tau_0(Ax^k - Ax^*) = (E - \tau_0 A)z^k,$$

следовательно,

$$(Az^{k+1}, z^{k+1}) \leqslant (A(E - \tau_0 A)z^k, (E - \tau_0 A)z^k).$$

Представляя вектор z^k в виде разложения по ортонормированному базису собственных векторов матрицы A, получим:

$$(A(E - \tau_0 A)z^k, (E - \tau_0 A)z^k) = \sum_{i=1}^{n} \lambda_i (1 - \tau_0 \lambda_i)^2 c_i^2.$$

Как установлено при доказательстве теоремы (3.1),

$$|1 - \tau_0 \lambda_i| \leqslant \rho_0 < 1, \quad i = 1, 2, \ldots, n,$$

следовательно, $(Az^{k+1}, z^{k+1}) \leqslant \rho_0^2 \sum_{i=1}^{n} \lambda_i c_i^2 = \rho_0^2 (Az^k, z^k)$, откуда, очевидно, вытекает неравенство $(Az^k, z^k) \leqslant \rho_0^{2k} (Az^0, z^0)$. Используя теперь оценки (3.2), получим: $\lambda_1 |z^k|^2 \leqslant \lambda_n \rho_0^{2k} |z^0|^2$, а это эквивалентно (4.5). □

§ 3. Методы решения алгебраической проблемы собственных значений

Под алгебраической проблемой собственных значений понимают задачу отыскания собственных чисел и собственных векторов матрицы. Различают полную проблему собственных значений, т. е. нахождение всех собственных чисел и собственных векторов, и частичную проблему собственных значений, т. е. отыскание лишь некоторых собственных чисел и соответствующих им собственных векторов.

Понятно, что методы решения частичной проблемы собственных значений должны быть более простыми. Мы рассмотрим примеры методов обоих классов. При этом ограничимся лишь случаем симметричных матриц.

Собственные числа симметричной матрицы A будем нумеровать в порядке неубывания их модулей:

$$|\lambda_1| \leqslant |\lambda_2| \leqslant \cdots \leqslant |\lambda_{n-1}| \leqslant |\lambda_n|.$$

Через $e^1, e^2, \ldots, e^n$ будем обозначать соответствующие ортонормированные собственные векторы.

1. Метод прямой итерации. Этот метод предназначен для отыскания максимального по модулю собственного числа матрицы и соответствующего ему собственного вектора.

Выберем некоторое нормированное начальное приближение y^0 и образуем последовательность нормированных векторов $y^1, y^2, \ldots$: $x^1 = Ay^0$, $y^1 = x^1/|x^1|$. Вообще, по y^k вычисляем $x^{k+1} = Ay^k$, а затем нормируем: $y^{k+1} = x^{k+1}/|x^{k+1}|$. Строим также последовательность чисел $\lambda^{(k)} = (Ay^k, y^k)$, $k = 1, 2, \ldots$

Теорема 1.1. *Пусть $|\lambda_{n-1}| < |\lambda_n|$, θ_k — угол, образованный векторами y^k и e^n, причем $\theta_0 \neq \pi/2$. Тогда $\operatorname{tg}\theta_k \to 0$, $\lambda^{(k)} \to \lambda_n$ при $k \to \infty$. Справедливы следующие оценки скорости сходимости:*

$$|\operatorname{tg}\theta_k| \leqslant \left|\frac{\lambda_{n-1}}{\lambda_n}\right|^k |\operatorname{tg}\theta_0|, \quad |\lambda^{(k)} - \lambda_n| \leqslant (|\lambda_{n-1}| + |\lambda_n|)\left|\frac{\lambda_{n-1}}{\lambda_n}\right|^{2k}. \quad (1.1)$$

Доказательство. По теореме об ортогональном разложении евклидова пространства вектор y^k однозначно представим в виде

$$y^k = \cos\theta_k\, e^n + \sin\theta_k\, u^k, \quad (1.2)$$

где u^k — единичный вектор, ортогональный e^n. В соответствии с изучаемым алгоритмом

$$y^{k+1} = \alpha(\cos\theta_k\, Ae^n + \sin\theta_k\, Au^k) = \alpha(\cos\theta_k\, \lambda_n e^n + \sin\theta_k\, Au^k).$$

Здесь α — число, выбираемое так, чтобы вектор y^{k+1} имел единичную длину. Перепишем последнее равенство в виде

$$y^{k+1} = \alpha(\cos\theta_k\, \lambda_n e^n + |Au^k| \sin\theta_k\, Au^k/|Au^k|\,) = \\ = \cos\theta_{k+1} e^n + \sin\theta_{k+1} Au^k/|Au^k|. \quad (1.3)$$

Мы учли здесь, что вектор Au^k ортогонален e^n, поскольку

$$(Au^k, e^n) = (u^k, Ae^n) = \lambda_n(u^k, e^n) = 0.$$

Из (1.3) вытекает, что

$$\operatorname{tg}\theta_{k+1} = \frac{|Au^k|}{\lambda_n}\operatorname{tg}\theta_k.$$

Записывая разложение u^k по базису собственных векторов матрицы A и учитывая ортогональность u^k и e^n, получим, что $Au^k = \sum\limits_{i=1}^{n-1} c_i\lambda_i e^i$, следовательно,

$$|Au^k|^2 = \sum_{i=1}^{n-1} c_i^2\lambda_i^2 \leqslant \lambda_{n-1}^2 \sum_{i=1}^{n-1} c_i^2 = \lambda_{n-1}^2|u^k|^2 = \lambda_{n-1}^2. \qquad (1.4)$$

Таким образом,

$$|\operatorname{tg}\theta_{k+1}| \leqslant \frac{|\lambda_{n-1}|}{|\lambda_n|}|\operatorname{tg}\theta_k|,$$

и первая оценка (1) доказана.

Вновь используя представление (1.2) и то, что $(Au^k, e^n) = 0$, получим:

$$\begin{aligned}\lambda_n - \lambda^{(k)} &= \lambda_n - (A(\cos\theta_k\, e^n + \sin\theta_k\, u^k), \cos\theta_k\, e^n + \sin\theta_k\, u^k) = \\ &= \lambda_n - (\cos\theta_k\,\lambda_n e^n + \sin\theta_k\, Au^k, \cos\theta_k\, e^n + \sin\theta_k\, u^k) = \\ &= \lambda_n - (\lambda_n\cos^2\theta_k + (Au^k, u^k)\sin^2\theta_k) = (\lambda_n - (Au^k, u^k))\sin^2\theta_k,\end{aligned}$$

Откуда вследствие (1.4) вытекает, что

$$|\lambda_n - \lambda^{(k)}| \leqslant |\lambda_n + \lambda_{n-1}|\sin^2\theta_k \leqslant |\lambda_n + \lambda_{n-1}|\operatorname{tg}^2\theta_k,$$

Вместе с уже полученной первой оценкой (1.1) это завершает доказательство теоремы. □

Условие $\theta_0 \neq \pi/2$ на практике не слишком обременительно. Если оно нарушается, то при проведении итераций за счет ошибок округления приближения обязательно выйдут из гиперплоскости, ортогональной e^n.

2. Метод обратной итерации. Метод предназначен для отыскания минимального по модулю собственного числа и соответствующего ему собственного вектора. Опишем соответствующий алгоритм.

Выбираем нормированное начальное приближение y^0 и строим последовательность векторов $y^1, y^2, \ldots$ по формулам: $x^{k+1} = A^{-1}y^k$, $y^{k+1} = x^{k+1}/|x^{k+1}|$, а также числа $\lambda^{(k)} = (Ay^k, y^k)$, $k = 0, 1, 2, \ldots$

При реализации метода выгоднее не строить и хранить матрицу A^{-1}, а решать на каждой итерации систему линейных уравнений $Ax^{k+1} = y^k$. Предварительно целесообразно представить матрицу A в виде LU или QU разложения (см. §1).

Относительно сходимости метода справедлива теорема, полностью аналогичная теореме 1.1, но на этот раз скорость сходимости характеризуется отношением $|\lambda_1|/|\lambda_2| < 1$.

2.1. Метод обратной итерации со сдвигом. Рассмотрим обобщение предыдущего метода, а именно переход от вектора y^k к y^{k+1} будем выполнять по формулам: $(A - \sigma E)x^{k+1} = y^k$, $y^{k+1} = x^{k+1}/|x^{k+1}|$. Здесь σ — параметр, называемый сдвигом. Последовательность $\lambda^{(k)}$, $k = 1, 2, \ldots$, по-прежнему, определяется формулой $\lambda^{(k)} = (Ay^k, y^k)$. Сходимость этого метода исследуется по той же схеме, что и в теореме 1.1. При этом оказывается, что $\lambda^{(k)} \to \lambda_j$, где номер j характеризуется условием $|\lambda_j - \sigma| < |\lambda_i - \sigma| \quad \forall\, i \neq j$, а последовательность y^k сходится к соответствующему собственному вектору e^j. Таким образом, метод позволяет находить собственное число матрицы A, ближайшее к заданному числу σ.

3. Метод вращений (Якоби). Этот метод довольно часто используется при решении полной проблемы собственных значений для симметричных матриц не слишком высокого порядка.

Напомним определение матрицы вращения. Ортогональная матрица

$$\begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}$$

порождает преобразование поворота на угол φ в двумерной плоскости. Матрица $Q = (q_{ij})_{i,j=1}^n$, отличающаяся от единичной лишь четырьмя элементами: $q_{k,k} = \cos\varphi$, $q_{ll} = \cos\varphi$, $q_{kl} = \sin\varphi$, $q_{lk} = -\sin\varphi$, где $1 \leqslant k < l \leqslant n$ — заданные целые числа, называется матрицей вращения. Нетрудно проверить, что Q — ортогональная матрица. Она порождает преобразование поворота на угол φ в двумерной плоскости, натянутой на векторы канонического базиса с номерами k, l.

Опишем идею метода Якоби. Пусть A — симметричная матрица. Как и в §1, ее собственные числа будем нумеровать в порядке неубывания. Образуем матрицу T, столбцами которой являются ортонормированные собственные векторы $e^1, e^2, \ldots, e^n$ матрицы A. Нетрудно

убедиться, что $T^T A T = \Lambda$, где Λ — диагональная матрица с элементами $\lambda_1, \lambda_2, \ldots, \lambda_n$ на диагонали. В методе Якоби матрица T строится как предел последовательности ортогональных матриц T_s так, что $\lim\limits_{s\to\infty} T_s^T A T_s = \Lambda$, причем при каждом s матрица T_s конструируется как произведение матриц вращения.

Образуем по матрице A матрицу $\widehat{A} = Q^T A Q$ и попытаемся выбрать параметры матрицы вращения, то есть k, l, φ так, чтобы матрица $\widehat{A}$ была максимально близка к диагональной. Опуская элементарные выкладки, приведем выражение для суммы квадратов внедиагональных элементов матрицы $\widehat{A}$:

$$\sum_{i\neq j} \widehat{a}_{ij}^2 = \sum_{i\neq j} a_{ij}^2 - 2a_{kl}^2 + 2[a_{kl}\cos 2\varphi + \frac{1}{2}(a_{ll} - a_{kk})\sin 2\varphi]^2.$$

Определим теперь числа k, l из условия:

$$|a_{kl}| = \max_{i\neq j} |a_{ij}| \tag{3.1}$$

а затем угол φ так, чтобы

$$a_{kl}\cos 2\varphi + \frac{1}{2}(a_{ll} - a_{kk})\sin 2\varphi = 0,$$

или

$$\operatorname{tg} 2\varphi = \frac{2a_{kl}}{a_{kk} - a_{ll}}. \tag{3.2}$$

При указанном выборе параметров матрицы вращения сумма квадратов внедиагональных элементов матрицы $\widehat{A}$ принимает наименьшее значение.

Теперь можно описать алгоритм метода Якоби. Пусть $A_0 = A$. Образуем последовательность матриц $A_1, A_2, \ldots$, при помощи рекуррентной формулы:

$$A_{p+1} = Q_p^T A_p Q_p, \quad p = 0, 1, \ldots, \tag{3.3}$$

где параметры матрицы вращения Q_p определяются так, чтобы сделать сумму квадратов внедиагональных элементов матрицы A_{p+1} минимально возможной, т. е. по формулам вида (3.1), (3.2).

Итерации проводят до тех пор, пока все внедиагональные элементы матрицы A_p не станут достаточно малыми. Тогда в качестве приближений к собственным числам матрицы A принимают диагональные элементы матрицы A_p, а столбцы матрицы $Q_0 Q_1 \ldots Q_p$ считают приближениями к собственным векторам матрицы A.

При обосновании сходимости метода используется

Лемма 3.1. *Пусть параметры матрицы вращения Q определяются согласно формулам* (3.1), (3.2). *Тогда*

$$\sum_{i\neq j}\widehat{a}_{ij}^2 \leqslant q\sum_{i\neq j}a_{ij}^2, \tag{3.4}$$

где

$$0 < q = 1 - \frac{2}{n(n-1)} < 1$$

при $n > 2$.

Доказательство. Вследствие (3.2) имеем:

$$\sum_{i\neq j}\widehat{a}_{ij}^2 = \sum_{i\neq j}a_{ij}^2 - 2a_{kl}^2, \tag{3.5}$$

а на основании (3.1)

$$\sum_{i\neq j}a_{ij}^2 \leqslant a_{kl}^2 n(n-1). \tag{3.6}$$

Здесь учтено, что матрица порядка n имеет $n^2 - n$ внедиагональных элементов. Из (3.5), (3.6) очевидным образом следует (3.4). □

Будем опираться также на следующий фундаментальный результат теории симметричных матриц, который приведем без доказательства.

Теорема 3.1. *Пусть* A, B *— симметричные матрицы порядка* n, $\lambda_k(A)$, $\lambda_k(B)$, $k = 1, \ldots, n$, *— их собственные числа, занумерованные в порядке неубывания. Тогда*

$$(\lambda_k(A) - \lambda_k(B))^2 \leqslant \sum_{i,j=1}^{n}(a_{ij} - b_{ij})^2,\ \ k = 1, 2, \ldots, n. \tag{3.7}$$

Докажем теперь сходимость метода Якоби. Из рекуррентной формулы (3.3) и леммы **??** вытекает, что

$$\sum_{i\neq j}(a_{ij}^{(p)})^2 \leqslant q^p\sum_{i\neq j}a_{ij}^2 \to 0 \text{ при } p\to\infty.$$

Это значит, что по любому заданному $\varepsilon > 0$ можно указать целое положительное число p такое, что

$$\sum_{i\neq j}(a_{ij}^{(p)})^2 \leqslant \varepsilon^2. \tag{3.8}$$

Обозначим через Λ_p диагональную матрицу, на диагонали которой расположены диагональные элементы матрицы A_p. В соответствии с теоремой 3.1 и оценкой (3.8) можем написать:

$$|\lambda_k(A_p) - \lambda_k^{(p)}| \leqslant \varepsilon, \quad k = 1, 2, \ldots, n,$$

где $\lambda_k^{(p)}$, $k = 1, \ldots, n$, — диагональные элементы матрицы Λ_p, упорядоченные по неубыванию. Вследствие (3.3) имеем: $A_p = T_p^T A T_p$, где $T_p = Q_0 Q_1 \ldots Q_p$, т. е. матрицы A_p и A подобны, а, значит, их собственные числа совпадают, поэтому

$$|\lambda_k(A) - \lambda_k^{(p)}| \leqslant \varepsilon, \quad k = 1, 2, \ldots, n. \tag{3.9}$$

Сходимость метода Якоби доказана.

Подчеркнем, что оценка (3.9) может рассматриваться как апостериорная оценка погрешности метода: добившись в ходе итераций выполнения неравенства (3.8), мы получаем собственные числа с погрешностью, не превосходящей ε.

§ 4. Методы решения нелинейных уравнений

В этом параграфе излагаются наиболее распространенные методы решения нелинейных уравнений вида

$$f(x) = 0, \tag{1}$$

где f — заданная непрерывная функция вещественного переменного Всюду в дальнейшем предполагается, что известен отрезок $[x_0, x_1]$, содержащий корень α уравнения (1). Предполагается, что α — единственный корень уравнения (1) на отрезке $[x_0, x_1]$. Методы отделения корней существенно зависят от конкретного вида функции f и в настоящем пособии не рассматриваются.

1. Метод деления отрезка пополам. Будем считать, что $f(x_0)f(x_1) < 0$, т. е. функция f меняет знак в точке $\alpha \in [x_0, x_1]$. Положим $x_2 = (x_0 + x_1)/2$ и вычислим $f(x_0)f(x_2)$. Если $f(x_0)f(x_2) < 0$, то корень расположен на отрезке $[x_0, x_2]$. В противном случае — на отрезке $[x_2, x_1]$. Выбирая теперь тот из двух отрезков, на котором лежит корень уравнения, применяем к нему описанную процедуру деления пополам. Процесс продолжают до тех пор, пока длина отрезка не станет меньше заданного $\varepsilon > 0$. При этом корень, очевидно, будет найден с точностью $\varepsilon > 0$.

Метод привлекает своей простотой и малыми затратами для реализации каждого тога итерационного процесса (требует лишь вычисления одного значения функции f). Следует подчеркнуть, что метод неприменим для отыскания корней четной кратности, когда функция f обращается в нуль в точке $\alpha \in [x_0, x_1]$, но не меняет знака на отрезке $[x_0, x_1]$.

2. Метод простой итерации. Метод основан на приведении уравнения (1) к эквивалентному уравнению вида

$$x = \varphi(x) \tag{2.1}$$

и построении последовательности приближений x_0, x_1, $\ldots x_n$, $\ldots$ по формуле

$$x_{k+1} = \varphi(x_k), \quad k = 0, 1, 2, \ldots \tag{2.2}$$

Начальное приближение x_0 считается заданным.

Эквивалентное преобразование уравнения (1) к виду (2.1) может быть выполнено различными способами. Например, можно положить $\varphi(x) = x + \psi(x)f(x)$, где ψ — произвольная непрерывная функция, не обращающаяся в нуль в окрестности точки α. Функцию ψ следует конструировать так, чтобы обеспечить сходимость последовательности x_k. Тогда, очевидно, $\lim\limits_{k\to\infty} x_k = \alpha$.

Теорема 2.1. *Пусть в некоторой окрестности*

$$S(\alpha, r) = \{x,\ |x - \alpha| < r\}$$

корня уравнения (2.1) *функция φ удовлетворяет условию Липшица с постоянной меньшей единицы, т. е.*

$$|\varphi(x') - \varphi(x'')| \leqslant q|x' - x''|, \quad x', x'' \in S(\alpha, r),\ 0 < q < 1. \tag{2.3}$$

Тогда при любом $x_0 \in S(\alpha, r)$

$$|x_k - \alpha| \leqslant q^k|x_0 - \alpha| \to 0 \quad \textit{при } k \to \infty. \tag{2.4}$$

Доказательство. Если $x_0 \in S(\alpha, r)$, то

$$|x_1 - \alpha| = |\varphi(x_0) - \alpha| = |\varphi(x_0) - \varphi(\alpha)| \leqslant q|x_0 - \alpha| < r,$$

следовательно, $x_1 \in S(\alpha, r)$. Аналогично, $x_1, x_2, \ldots \in S(\alpha, r)$. Таким образом,

$$|x_k - \alpha| = |\varphi(x_{k-1}) - \varphi(\alpha)| \leqslant q|x_{k-1} - \alpha| \leqslant q^k|x_0 - \alpha| \to 0, \quad k \to \infty. \ \square$$

Следствие 2.1. *Пусть функция φ непрерывно дифференцируема на $S(\alpha, r)$ и*

$$|\varphi'(x)| \leqslant q < 1 \tag{2.5}$$

при $x \in S(\alpha, r)$. Тогда для последовательности x_k, построенной по методу (2.2)*, выполняется оценка* (2.4)*.*

Графическая интерпретация итерационного процесса (2.2) дана на рис. 1, 2. Отметим, что случай $-1 < \varphi'(x) \leqslant 0$ (рис. 2) наиболее

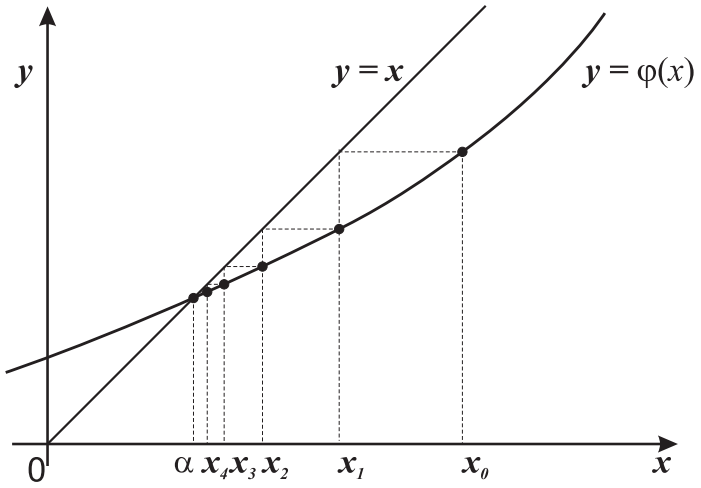

Рис. 1. К итерационному процессу (2.2), $0 \leqslant \varphi'(x) < 1$

выгоден, так как величина $|x_k - x_{k-1}|$ на каждом шаге итерационного процесса оценивает сверху погрешность приближения x_k к α.

Приведем простейший пример конструирования функции φ. Пусть требуется вычислить $\sqrt{a}$, $a > 0$. Соответствующее уравнение имеет вид

$$f(x) \equiv x^2 - a = 0. \tag{2.6}$$

Преобразуем уравнение (2.6):

$$x = \varphi(x), \quad \text{где } \varphi(x) = \frac{a}{x}.$$

В этом случае $\varphi'(x) = -a/x^2$, $\varphi'(\sqrt{a}) = -1$, условие (2.5) в окрестности корня не выполнено, и итерационный процесс может разойтись, даже если x_0 лежит сколь угодно близко к $\sqrt{a}$ (проиллюстрируйте эту ситуацию графически!).

Перепишем теперь уравнение (2.6) в виде

$$x = \frac{1}{2}\left(x + \frac{a}{x}\right)$$

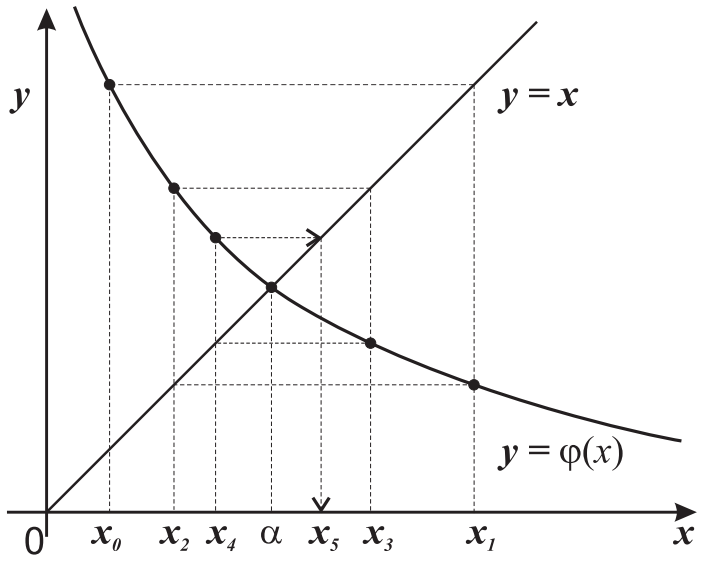

Рис. 2. К итерационному процессу (2.2), $-1 < \varphi'(x) \leqslant 0$

В этом случае

$$\varphi(x) = \frac{1}{2}\left(x + \frac{a}{x}\right), \quad \varphi'(x) = \frac{1}{2}\left(1 - \frac{a}{x^2}\right), \ \varphi'(\sqrt{a}) = 0$$

и, следовательно, $|\varphi'(x)| < 1$ в некоторой окрестности $\sqrt{a}$. Таким образом, итерационный метод, определяемый соотношением

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right), \quad k = 0, 1, \ldots,$$

сходится при любом начальном приближении x_0, достаточно близком к $\sqrt{a}$. Более детальный анализ показывает, что сходимость имеет место при любом $x_0 > 0$.

2.1. Заметим, что если

$$\varphi'(\alpha) = \varphi''(\alpha) = \cdots = \varphi^{(m-1)}(\alpha) = 0, \quad \varphi^{(m)}(\alpha) \neq 0, \quad \text{при } m \geqslant 2,$$

то применяя формулу Тейлора с остаточным членом в форме Лагранжа, получим, что

$$|x_{k+1} - \alpha| \leqslant \frac{1}{m!}|\varphi^{(m)}(\xi)||x_k - \alpha|^m, \tag{2.7}$$

и метод (2.2) сходится вблизи корня очень быстро. Число m принято называть порядком итерацнонного метода (2.2).

3. Метод Ньютона. В этом случае полагают

$$\varphi(x) = x - \frac{f(x)}{f'(x)},$$

то есть

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},\ k = 0, 1, \ldots \tag{3.1}$$

Нетрудно видеть, что в этом случае

$$\varphi'(\alpha) = 0, \quad \varphi''(\alpha) \neq 0, \tag{3.2}$$

если $f(\alpha) = 0$, а $f'(\alpha) \neq 0$, т. е. α — простой корень уравнения (1). Отсюда вытекает, что метод Ньютона (3.1) сходится при любом начальном приближении x_0, достаточно близком к α.

Равенства (3.2) показывают, что метод Ньютона — метод второго порядка и, следовательно, имеет квадратичную сходимость (в оценке (2.7) получаем $m = 2$).

Метод Ньютона имеет простую геометрическую интерпретацию: x_{k+1} — точка пересечения касательной к графику функции f в точке $(x_k, f(x_k))$ с осью x (рис. 3).

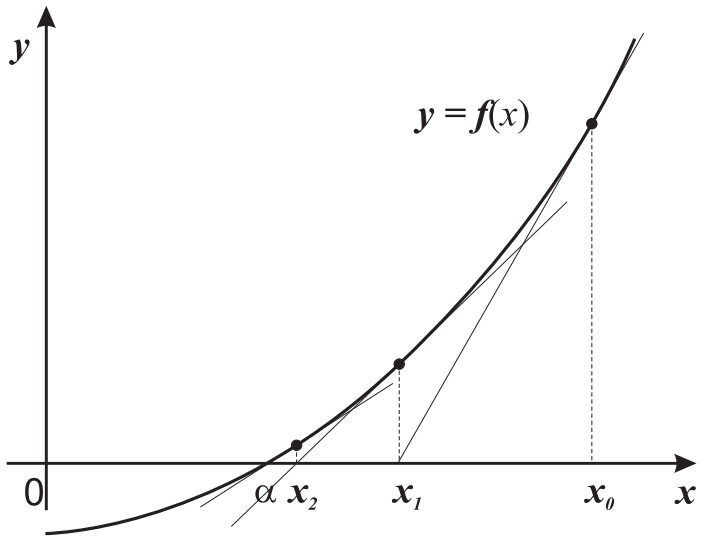

Рис. 3. Метод Ньютона

4. Метод хорд. Недостатком метода Ньютона является необходимость вычисления производной функции f. В некоторых случаях это оказывается слишком трудоемкой задачей.

Если известен интервал (x_0, x_1), на концах которого функция f имеет противоположные знаки, то приближение x_2 к корню можно

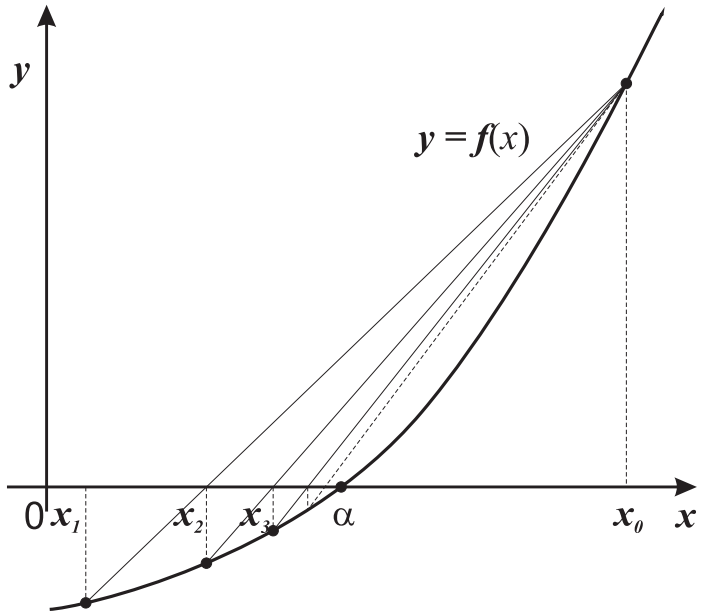

Рис. 4. Метод хорд

определить как точку пересечения прямой, проходящей через точки $(x_0, f(x_0))$ и $(x_1, f(x_1)$, с осью x (рис. 4). Ясно, что

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

Все последующие приближения будем вычислять по формуле

$$x_{k+1} = \frac{x_0 f(x_k) - x_k f(x_0)}{f(x_k) - f(x_0)}, \quad k = 2, 3, \ldots \tag{4.1}$$

Метод (4.1) можно переписать в виде (2.2), где

$$\varphi(x) = \frac{x_0 f(x) - x f(x_0)}{f(x) - f(x_0)}.$$

Используя формулу Тейлора, нетрудно подсчитать, что

$$\varphi'(\alpha) = \frac{(x_0 - \alpha)^2}{2} \frac{f''(\xi)}{f'(\xi_1)}, \quad \xi, \xi_1 \in (x_0, x_1),$$

откуда вытекает, что $|\varphi'(\alpha)| < 1$, если x_0 выбрано достаточно близко к корню. Таким образом, существует окрестность точки α, в которой $|\varphi'(x)| \leqslant q < 1$, и метод (4.1) сходится при любом x_1 из этой окрестности.

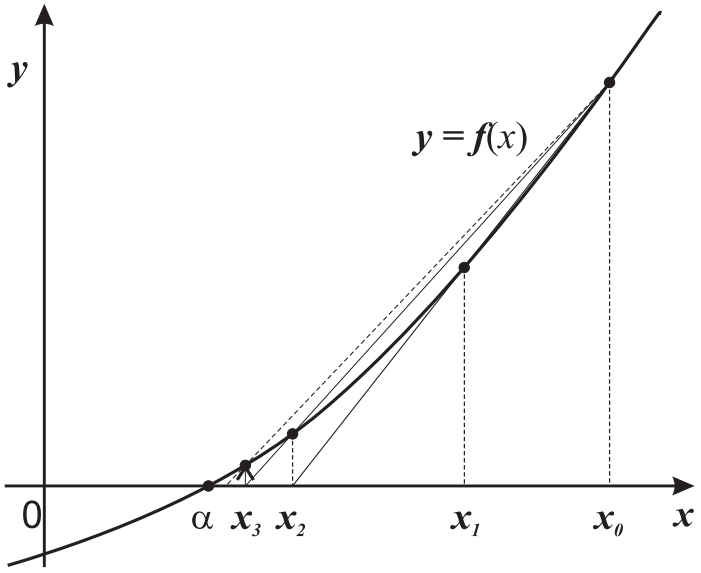

Рис. 5. Метод секущих

5. Метод секущих. Этот метод получается при аппроксимации производной $f'(x_k)$ в методе Ньютона (3.1) разностным отношением

$$f'(x_k) = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$

В результате приходим к следующему итерационному процессу

$$x_{k+1} = x_k - \frac{(x_k - x_{k-1})f(x_k)}{f(x_k) - f(x_{k-1})} \tag{5.1}$$

(рис. 5). Геометрически x_{k+1} — точка пересечения секущей, проходящей через точки $(x_k, f(x_k))$ и $(x_{k-1}, f(x_{k-1}))$ с осью x.

В отличие от ранее рассмотренных методов для вычисления каждого последующего приближения x_{k+1} требуется использовать два предыдущих приближения. Такие методы принято называть двушаговыми.

Вновь используя разложение по формуле Тейлора, можно показать, что справедливо приближенное равенство

$$x_{k+1} - \alpha \approx \frac{1}{2}\frac{f''(\alpha)}{f'(\alpha)}(x_{k-1} - \alpha)(x_k - \alpha),$$

показывающее, что если имеет место сходимость метода (5.1), то его погрешность убывает существенно быстрее, чем в методе хорд, и несколько медленнее, чем в методе Ньютона.

§ 5. Методы решения систем нелинейных уравнений

В этом параграфе мы рассмотрим задачу решения системы нелинейных уравнений

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0, \\ f_2(x_1, \ldots, x_n) &= 0, \\ \ldots\ldots\ldots\ldots\ldots & \\ f_n(x_1, \ldots, x_n) &= 0, \end{aligned} \tag{1}$$

где $f_1, f_2, \ldots, f_n$ — заданные функции n вещественных переменных. Как и в случае одного нелинейного уравнения, мы будем предполагать, что система (1) имеет корень $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, и нам известна его некоторая окрестность. Для уточнения корня применяются различные итерационные процессы. При их изложении и исследовании удобно записывать систему (1) в операторной форме

$$F(x) = 0, \tag{2}$$

где $F(x) = (f_1(x), f_2(x), \ldots, f_n(x))$ — оператор, отображающий пространство R^n векторов $x = (x_1, \ldots, x_n)$ в себя.

1. Метод простой итерации. В этом методе систему (1) предварительно некоторыми эквивалентными преобразованиями приводят к виду

$$x = G(x) \tag{1.1}$$

и, начиная с некоторого начального приближения $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)$, строят последовательность векторов $x^k = (x_1^k, x_2^k, \ldots, x_n^k)$ по следующему правилу

$$x^{k+1} = G(x^k), \quad k = 0, 1, 2, \ldots \tag{1.2}$$

Существуют различные способы приведения системы (2) к эквивалентной системе вида (1.1). Обычно, они естественным образом подсказываются структурой системы (2). Часто поступают следующим образом. Выбирают некоторую невырожденную матрицу B. Систему (2) приводят к виду

$$B(x - x) + F(x) = 0.$$

Последовательность приближений $x^1, x^2, \ldots$ определяют из уравнений

$$B(x^{k+1} - x^k) + F(x^k) = 0, \quad k = 0, 1, 2, \ldots \tag{1.3}$$

Выбор матрицы B, очевидно, должен быть подчинен условиям сходимости итераций. Кроме того, матрица B должна быть такой, чтобы система линейных уравнений $Bx^{k+1} = Bx^k - F(x^k)$ для отыскания x^{k+1} решалась достаточно просто.

2. Метод Ньютона. При наличии достаточно хорошего приближения к решению системы (2) очень часто весьма эффективным оказывается его уточнение по методу Ньютона. Этот метод является непосредственным обобщением метода Ньютона для одного нелинейного уравнения и формально может быть записан следующим образом

$$x^{k+1} = x^k - (F'(x^k))^{-1} F(x^k), \quad k = 0, 1, 2, \ldots \tag{2.1}$$

Здесь

$$F'(x^k) = \left\{ \frac{\partial f_i(x^k)}{\partial x_j} \right\}_{i,j=1}^{n}$$

есть матрица Якоби системы функций $f_1, f_2, \ldots, f_n$, вычисленная в точке x^k.

При практической реализации метода его обычно переписывают в виде

$$F'(x^k)(x^{k+1} - x^k) = -F(x^k),$$

или

$$F'(x^k)\Delta^k = -F(x^k),$$
$$x^{k+1} = x^k + \Delta^k.$$

Таким образом, для построения x^{k+1} по известному x^k нужно решить систему линейных алгебраических уравнений с матрицей $F'(x^k)$.

Во многих случаях формирование матрицы Якоби $F'(x^k)$ оказывается существенно более трудоемкой задачей, чем вычисление вектора $F(x^k)$. Поэтому довольно часто применяют следующую модификацию метода Ньютона

$$F'(x^0)\Delta^k = -F(x^k),$$
$$x^{k+1} = x^k + \Delta^k.$$

Здесь матрица Якоби вычисляется лишь на начальном приближении. Экономия достигается как за счет формирования матрицы системы, так и за счет того, что решение нескольких систем с одинаковыми матрицами существенно проще, чем решение того же количества систем с разными матрицами.

При этом надо, однако, иметь в виду, что если метод (2.1) обладает при достаточно хорошем начальном приближении квадратичной сходимостью, т. е.

$$|x^{k+1} - \alpha| \leqslant \mathrm{const}|x^k - \alpha|^2,$$

то модифицированный метод Ньютона сходится лишь со скоростью геометрической прогрессии.

Глава 2

Численные методы анализа

§ 1. Интерполирование функций

1. Существование и единственность интерполяционного полинома. Под интерполированием понимают следующую задачу. Даны $n+1$ различные точки $x_0, x_1, \ldots, x_n$ на вещественной прямой и значения некоторой вещественной функции $f(x_0), f(x_1), \ldots, f(x_n)$ в этих точках. Даны также вещественные функции $\varphi_0(x), \ldots, \varphi_n(x)$. Требуется построить функцию

$$\varphi(x) = \sum_{k=0}^{n} c_k \varphi_k(x), \tag{1.1}$$

где c_k — искомые вещественные числа, удовлетворяющую условиям:

$$\varphi(x_i) = f(x_i), \quad i = 0, 1, \ldots, n. \tag{1.2}$$

Функцию φ принято называть обобщенным интерполяционным полиномом, функции φ_k — базисными функциями, точки $x_0, x_1, \ldots, x_n$ называют узлами интерполирования. Если φ_k — полиномы, то говорят о задаче алгебраического интерполирования, а функцию φ называют интерполяционным полиномом. Задачей алгебраического интерполирования мы в дальнейшем и будем заниматься.

Если $\varphi_0(x) \equiv 1$, $\varphi_1(x) = x$, $\varphi_2(x) = x^2$, $\ldots$, $\varphi_n(x) = x^n$, то $\varphi(x)$ есть полином степени n.

Теорема 1.1. *Интерполяционный полином степени n однозначно определяется по любым заданным значениям $f(x_0)$, $f(x_1), \ldots,$ $f(x_n)$ в $n+1$ различных точках x_0, $x_1, \ldots,$ x_n.*

Доказательство. Условия (1.2) представляют собой систему из $n+1$ линейных алгебраических уравнений относительно $n+1$ неизвестных $c_0, c_1, \ldots, c_n$, поэтому достаточно установить, что соответствующая однородная система имеет только тривиальное решение. Предполагая противное, получим, что полином степени n имеет $n+1$ различных корней $x_0, x_1, \ldots, x_n$, а это возможно лишь в том случае, когда все его коэффициенты, т. е. $c_0, c_1, \ldots, c_n$, равны нулю. $\square$

Во многих случаях полезно иметь явный вид интерполяционного многочлена. Однако разрешить систему (1.2) в явном виде при

произвольном значении n не удается. Причина — в неудачном выборе формы представления интерполяционного полинома, иными словами, в неудачном выборе базисных функций $\varphi_0(x), \ldots, \varphi_n(x)$.

2. Интерполяционный полином Лагранжа. Понятно, что наиболее простой вид система (1.2) принимает в том случае, когда базисные функции — полиномы степени n, удовлетворяющие условиям:

$$\varphi_i(x_j) = \begin{cases} 0, & i \neq j; \\ 1, & i = j. \end{cases} \tag{2.1}$$

При этом, очевидно, $c_i = f(x_i)$, $i = 0, 1, \ldots, n$. Базисные функции в этом случае легко выписываются в явном виде. Действительно, вследствие условия (2.1) точки $x_0, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ — корни полинома $\varphi_i(x)$, поэтому $\varphi_i(x) = A(x-x_0)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)$, где $A = \text{const}$. Используя условие $\varphi_i(x_i) = 1$, находим

$$A = \frac{1}{(x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)},$$

следовательно,

$$\varphi_i(x) = \frac{(x - x_0)\cdots(x - x_{i-1})(x - x_{i+1})\cdots(x - x_n)}{(x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)}, \tag{2.2}$$

$i = 0, \ldots, n$. Построенный таким образом интерполяционный полином называют интерполяционным полиномом Лагранжа и обозначают через $L_n(x)$:

$$L_n(x) = \sum_{i=0}^{n} f(x_i) \frac{(x - x_0)\cdots(x - x_{i-1})(x - x_{i+1})\cdots(x - x_n)}{(x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)}. \tag{2.3}$$

Полиномы, определяемые формулами (2.2), называют базисными функциями Лагранжа.

3. Интерполяционный полином Ньютона. Определим теперь базисные функции соотношениями:

$$\varphi_0(x) \equiv 1,\ \varphi_1(x) = (x - x_0),\ \varphi_2(x) = (x - x_0)(x - x_1),\ \ldots,$$
$$\varphi_n(x) = (x - x_0)(x - x_1)\cdots(x - x_{n-1}). \tag{3.1}$$

Тогда система (1.2) принимает треугольный вид:

$c_0 = f(x_0)$,

$$
\begin{aligned}
&c_0 + c_1(x_1 - x_0) = f(x_1),\\
&c_0 + c_1(x_2 - x_0) + c_2(x_2 - x_0)(x_2 - x_1) = f(x_2),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&c_0 + c_1(x_n - x_0) + \ldots + c_n(x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1}) = f(x_n).
\end{aligned}
\tag{3.2}
$$

Отсюда последовательно находим

$$
\begin{aligned}
&c_0 = f(x_0),\\
&c_1 = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0},\\
&c_2 = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&c_n = \frac{f(x_0)}{(x_0 - x_1)\cdots(x_0 - x_n)} + \ldots + \frac{f(x_n)}{(x_n - x_0)\cdots(x_n - x_{n-1})}.
\end{aligned}
\tag{3.3}
$$

Часто полученным выражениям для коэффициентов интерполяционного полинома придают несколько иную, рекуррентную, форму.

Заметим, что

$$c_1 = f(x_0, x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Величину $f(x_1, x_0)$ называют разделенной разностью первого порядка функции f. Вообще,

$$f(x_k, x_{k+1}) = \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k}.$$

Далее, легко подсчитать, что

$$c_2 = f(x_0, x_1, x_2) = \frac{f(x_1, x_2) - f(x_1, x_0)}{x_2 - x_0}.$$

Величину

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

называют разделенной разностью второго порядка. Применяя индукцию, нетрудно получить общую формулу:

$$c_k = f(x_0, x_1, \ldots, x_k) = \frac{f(x_1, x_2, \ldots, x_k) - f(x_0, x_1, \ldots, x_{k-1})}{x_k - x_0}. \tag{3.4}$$

Величину $f(x_0, x_1, \ldots, x_k)$ называют разделенной разностью порядка k.

Таким образом, интерполяционный полином приобретает вид

$$P_n(x) = f(x_0) + f(x_0, x_1)(x - x_0) + f(x_0, x_1, x_2)(x - x_0)(x - x_1) + \ldots + f(x_0, x_1, x_2, \ldots, x_n)(x - x_0)(x - x_1)\cdots(x - x_{n-1}). \quad (3.5)$$

Такую форму записи интерполяционного полинома называют интерполяционным полиномом Ньютона.

Процесс вычисления разделенных разностей полезно иллюстрировать таблицей.

$$\begin{array}{lll}
x_0 & \underline{f(x_0)} & \\
 & & \underline{f(x_0, x_1)} \\
x_1 & f(x_1) & \\
 & & f(x_1, x_2) \\
x_2 & f(x_2) & \\
 & & f(x_2, x_3) \\
\vdots & & \\
x_n & f(x_n) &
\end{array}
\qquad
\begin{array}{ll}
 & \\
 & \underline{f(x_0, x_1, x_2)} \\
 & \\
 & f(x_1, x_2, x_3) \\
 & \\
 & \vdots \\
 &
\end{array}
\qquad \ldots \qquad
\underline{f(x_0, x_1, x_2, \ldots, x_n)}$$

Первые два столбца этой таблицы — исходные данные. Остальные — вычисляются последовательно один за другим. Подчеркнутые разделенные разности (они стоят на верхней диагонали таблицы) используются при образовании полинома $P_n(x)$.

4. Оценка погрешности интерполирования.

Теорема 4.1. *Пусть $[a, b]$ — отрезок вещественной оси, содержащий все узлы интерполирования $x_0, x_1, \ldots, x_n$, функция f непрерывно дифференцируема $n + 1$ раз на отрезке $[a, b]$, точка $x \in [a, b]$ и не совпадает ни с одним из узлов интерполирования. Тогда существует точка $\xi \in [a, b]$ такая, что*[1)]

$$f(x) = L_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x). \quad (4.1)$$

Здесь L_n — полином степени n, интерполирующий функцию f по узлам $x_0, x_1, \ldots, x_n$, $\omega_{n+1}(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$.

Доказательство. Введем в рассмотрение функцию

$$g(z) = f(z) - L_n(z) - K\omega_{n+1}(z), \quad z \in [a, b].$$

[1)] Выражение в правой части (4.1) принято называть остаточным членом интерполяционного полинома в форме Лагранжа.

Здесь K — постоянная. Определим ее так, чтобы

$$f(x) - L_n(x) - K\omega_{n+1}(x) = 0. \tag{4.2}$$

Это можно сделать, так как $\omega_{n+1}(x) \neq 0$. При указанном выборе постоянной K функция g обращается в нуль в $n+2$ точках, а именно в точке x, а также в точках $x_0, x_1, \ldots, x_n$, так как $L_n(x_i) - f(x_i) = 0$, $\omega_{n+1}(x_i) = 0$ при $i = 0, 1 \ldots, n$. По теореме Ролля первая производная функции g обращается в нуль по крайней мере один раз между каждой парой соседних точек из указанного множества точек. Иными словами, существует $n+1$ различных точек $\xi_1, \xi_2, \ldots, \xi_{n+1} \in [a, b]$ таких, что $g'(\xi_i) = 0,\ i = 1, \ldots, n+1$. Отсюда вытекает, что вторая производная функции g обращается в нуль по крайней мере в n различных точках. Продолжая аналогичные рассуждения, получим, что существует точка $\xi \in [a, b]$ такая, что $g^{(n+1)}(\xi) = 0$. Вычисляя производную порядка $(n+1)$ функции g, получим: $f^{(n+1)}(\xi) - K(n+1)! = 0$, т. е. $K = f^{(n+1)}(\xi)/(n+1)!$, откуда вследствие (4.2) вытекает (4.1). □

Из только что доказанной теоремы вытекает простое, но полезное

Следствие 4.1. *Пусть* $|f^{(n+1)}(x)| \leqslant M_{n+1}\ \forall x \in [a, b]$. *Тогда*

$$|f(x) - L_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)| \quad \forall x \in [a, b]. \tag{4.3}$$

4.1. Связь разделенных разностей с производными. Обобщенная формула Тейлора. Используя представление (4.1) для погрешности интерполяционного полинома, установим связь между разделенными разностями и производными.

Теорема 4.2. *Пусть функция f дифференцируема k раз на отрезке $[\underline{x}, \overline{x}]$, где $\underline{x} = \min(x_0, x_1, \ldots, x_k)$, $\overline{x} = \max(x_0, x_1, \ldots, x_k)$. Тогда существует точка $\xi \in [\underline{x}, \overline{x}]$ такая, что*

$$f(x_0, x_1, \ldots, x_k) = \frac{f^{(k)}(\xi)}{k!}. \tag{4.4}$$

Доказательство. Для упрощения записей будем считать, что узлами, по которым образуется разделенная разность, являются точки $x_0, x_1, \ldots, x_n, x$, и построим интерполяционный полином P_{n+1} степени $n+1$ по значениям функции f в этих точках. Записывая значение полинома P_{n+1} в точке x в форме Ньютона, получим

$$\begin{aligned} P_{n+1}(x) = {} & f(x_0) + (x - x_0)f(x_0, x_1) + \ldots + \\ & + (x - x_0)(x - x_1)\cdots(x - x_{n-1})f(x_0, x_1, \ldots, x_n) + \end{aligned}$$

$$+ (x - x_0)(x - x_1) \cdots (x - x_n) f(x_0, x_1, \ldots, x_n, x) = \\ = P_n(x) + \omega_{n+1}(x) f(x_0, x_1, \ldots, x_n, x) \quad (4.5)$$

(см. (3.5)). С другой стороны, по построению $P_{n+1}(x) = f(x)$, следовательно, $f(x) - P_n(x) = \omega_{n+1}(x) f(x_0, x_1, \ldots, x_n, x)$. Сравнивая это равенство с (4.1), получим утверждение теоремы. $\square$

Выразим разделенную разность в каждом слагаемом равенства (4.5) через значение соответствующей производной. Тогда получим

$$f(x) = f(x_0) + (x - x_0) f'(\xi_1) + (x - x_0)(x - x_1) \frac{f''(\xi_2)}{2!} + \ldots + \\ + (x - x_0)(x - x_1) \cdots (x - x_{n-1}) \frac{f^{(n)}(\xi_n)}{n!} + \omega_{n+1}(x) \frac{f^{(n+1)}(\xi_{n+1})}{(n+1)!}. \quad (4.6)$$

Формулу (4.6) можно назвать обобщенной формулой Тейлора с остаточным членом в форме Лагранжа. В предельном случае, когда все узлы интерполирования стягиваются в узел x_0, формула принимает вид:

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

где ξ — некоторая точка, лежащая между x и x_0, т. е. (4.6) переходит в обычную формулу Тейлора с остаточным членом в форме Лагранжа.

5. Оптимальный выбор узлов интерполирования. Многочлены Чебышева. Из оценки (4.3) следует, что погрешность интерполирования зависит от расположения узлов.

Естественно попытаться решить следующую задачу: при заданном n расположить узлы $x_0, x_1, \ldots, x_n$ так, чтобы минимизировать величину

$$\max_{a \leqslant x \leqslant b} |(x - x_0)(x - x_1) \cdots (x - x_n)|. \quad (5.1)$$

При решении этой задачи будем использовать многочлены Чебышева, определяемые при помощи формул:

$$T_0(x) \equiv 1, \; T_1(x) = x, \quad (5.2)$$

$$T_{k+1} = 2xT_k(x) - T_{k-1}(x), \; k = 1, 2, \ldots \quad (5.3)$$

Здесь k — степень многочлена.

Нетрудно видеть, что $T_k(x) = 2^{k-1}x^k + \ldots$. Многочлены $\widetilde{T}_k(x) = 2^{1-k}T_k(x)$ со старшим коэффициентом, равным единице, называют нормированными многочленами Чебышева.

Построим явную формулу для полиномов Чебышева. Будем разыскивать значение $T_k(x)$ в виде $T_k(x) = \lambda^k$. Используя это представление в рекуррентной формуле (5.3), получим $\lambda^{k+1} = 2x\lambda^k - \lambda^{k-1}$, откуда, предполагая, что $\lambda \neq 0$, приходим к квадратному уравнению $\lambda^2 - 2x\lambda + 1 = 0$ для определения параметра λ. Корни этого уравнения: $\lambda_{1,2} = x \pm \sqrt{x^2-1}$, поэтому функции $T_k^{(1)}(x) = (x + \sqrt{x^2-1})^k$, $T_k^{(2)}(x) = (x - \sqrt{x^2-1})^k$, а следовательно, и функции $T_k(x) = c_1T_k^{(1)}(x) + c_2T_k^{(2)}(x)$, $k = 0, 1, \ldots$, где c_1, c_2 — произвольные постоянные, удовлетворяют рекуррентному соотношению (5.3). Выберем c_1, c_2 так, чтобы были выполнены начальные условия (5.2):

$$c_1 + c_2 = 1,$$

$$(c_1 + c_2)x + (c_1 - c_2)\sqrt{x^2-1} = x.$$

Отсюда $c_1 = c_2 = 1/2$, т. е. полиномы

$$T_k(x) = \frac{1}{2}\left(x + \sqrt{x^2-1}\right)^k + \frac{1}{2}\left(x - \sqrt{x^2-1}\right)^k, \quad k = 0, 1, 2, \ldots$$

удовлетворяют рекуррентному соотношению (5.3) и начальным условиям (5.2). При $|x| \leqslant 1$ полиномам Чебышева можно придать более компактный вид. Положим в этом случае $x = \cos\varphi$. Тогда

$$T_k(x) = \frac{1}{2}\left(\cos\varphi + i\sin\varphi\right)^k + \frac{1}{2}\left(\cos\varphi - i\sin\varphi\right)^k,$$

откуда, используя формулы Муавра, получим $T_k(x) = \cos k\varphi$, или

$$T_k(x) = \cos k \arccos x. \tag{5.4}$$

Представление (5.4) позволяет найти все корни полинома Чебышева. В самом деле, из уравнения $\cos k \arccos x = 0$ сразу получаем, что $k \arccos x = \pi(2j+1)/2$, $j = 0, \pm1, \pm2, \ldots$, следовательно, точки

$$x_j = \cos\frac{\pi(2j+1)}{2k}, \quad j = 0, 1, 2, \ldots, k-1,$$

есть корни полинома $T_k(x)$.

Найдем также точки экстремума полинома $T_k(x)$ на отрезке $[-1, 1]$. Ясно, что — это те точки, в которых $|T_k(x)| = 1$. Они определяются соотношениями

$$\widehat{x}_k = \cos\frac{j\pi}{k}, \quad j = 0, 1, \ldots, k,$$

причем $T_k(\widehat{x}_j) = (-1)^j,\ j = 0, 1, \ldots, k$.

Полиномы Чебышева являются полиномами, наименее уклоняющимися от нуля на отрезке $[-1, 1]$. А именно, справедлива

Теорема 5.1. *Для любого полинома $Q_k(x)$ степени k со старшим коэффициентом, равным единице, справедливо неравенство*

$$\max_{-1\leqslant x\leqslant 1} |Q_k(x)| > \max_{-1\leqslant x\leqslant 1} |\widetilde{T}_k(x)|,$$

где $\widetilde{T}_k(x)$ — нормированный полином Чебышева степени k.

Доказательство. Предположим противное, и пусть $Q_k(x)$ — полином степени k со старшим коэффициентом, равным единице, для которого

$$\max_{-1\leqslant x\leqslant 1} |Q_k(x)| \leqslant \max_{-1\leqslant x\leqslant 1} |\widetilde{T}_k(x)|.$$

Ясно, что полином $R(x) = \widetilde{T}_k(x) - Q_k(x)$ имеет степень не выше $k-1$. Проанализируем значения полинома $R(x)$ в точках экстремума полинома $\widetilde{T}_k(x)$. По определению полинома $\widetilde{T}_k(x)$ имеем $\widetilde{T}_k(\widehat{x}_j) = (-1)^j 2^{1-k}$, а вследствие выдвинутого нами предположения $|Q_k(\widehat{x}_j)| \leqslant 2^{1-k}$, поэтому $R(\widehat{x}_0)\geqslant 0$, $R(\widehat{x}_1)\leqslant 0$, $R(\widehat{x}_2) \geqslant 0, \ldots$ Таким образом, на каждом отрезке $[\widehat{x}_l, \widehat{x}_{l+1}]$, $l = 0, 1, \ldots, k-1$, полином $R(x)$ имеет по крайней мере по одному корню, но это противоречит тому, что $R(x)$ — полином степени не выше, чем $k-1$. □

Используя полученный результат, нетрудно построить полином, наименее уклоняющийся от нуля на произвольном отрезке $[a, b]$. С этой целью выполним линейную замену переменной, переводящую отрезок $-1 \leqslant t \leqslant 1$ в отрезок $a \leqslant x \leqslant b$:

$$x = \frac{a+b}{2} - \frac{a-b}{2}t.$$

При указанной замене полином $T_k(t)$ переходит в полином

$$T_k\left(\frac{2x-(b+a)}{b-a}\right).$$

Коэффициент при x^k у этого полинома равен $2^{2k-1}/(b-a)^k$, а корнями являются точки

$$x_j = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{\pi(2j+1)}{2k}, \quad j = 0, 1, \ldots, k-1.$$

Понятно, что полиномом, наименее уклоняющимся от нуля на отрезке $[a, b]$, среди всех полиномов степени k со старшим коэффициентом, равным единице, является полином

$$\frac{(b-a)^k}{2^{2k-1}} T_k\left(\frac{2x-(b+a)}{b-a}\right).$$

Таким образом, оптимальными узлами интерполирования, минимизирующими величину (5.1), являются точки

$$x_j = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{(2j+1)\pi}{2(n+1)}, \quad j = 0, 1, \ldots, n. \tag{5.5}$$

Решение задачи об оптимальном расположении узлов интерполирования иллюстрирует рис. 1. Сплошной линией здесь изображен график функции $\omega_9(x)$ на отрезке $0 \leqslant x \leqslant 1$ при равномерном распределении узлов интерполирования, график функции $\omega_9(x)$ при

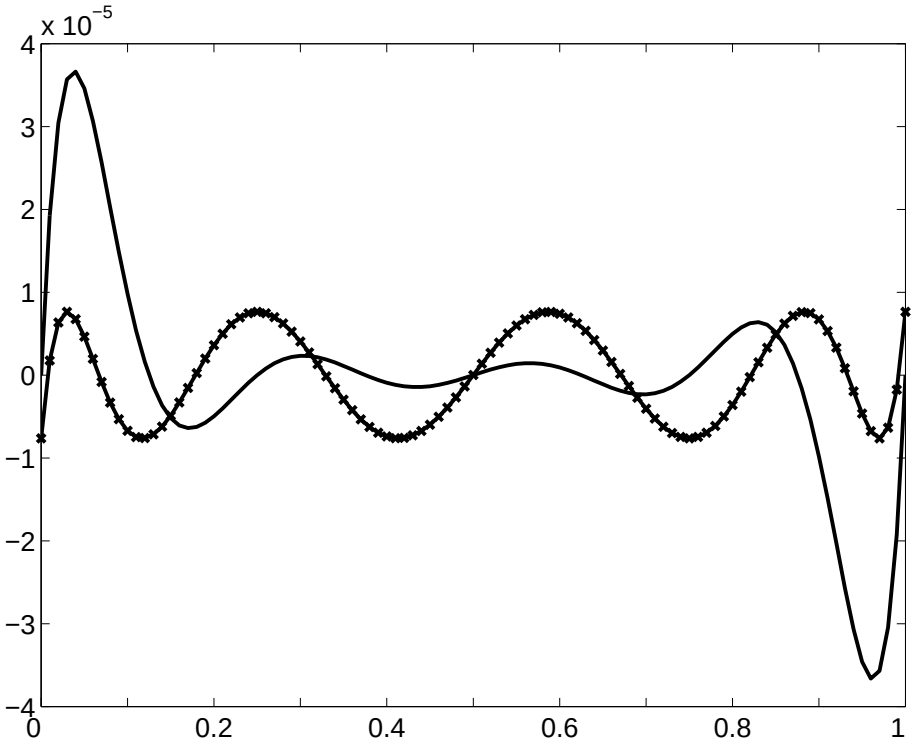

Рис. 1. Погрешность интерполирования.

оптимальном распределении узлов в соответствии с (5.5) помечен значком «×». При равномерном распределении узлов интерполирования погрешность сильно возрастает вблизи границ отрезка.

6. Интерполирование с кратными узлами. Будем предполагать теперь, что наряду со значениями функции f в узлах интерполирования заданы значения производных этой функции вплоть до определенного порядка. Точнее говоря, требуется найти такой полином $P_m(x)$ степени m, что

$$P_m^{(j)}(x_i) = f^{(j)}(x_i), \quad j = 0, 1, \ldots, m_i - 1, \ i = 0, 1, \ldots, n. \tag{6.1}$$

Число $m_i \geqslant 1$ называется кратностью узла x_i. Понятно, что если $m_i = 1$ при $i = 0, 1, \ldots n$, то есть все узлы простые, то получаем обычную задачу интерполирования.

Теорема 6.1. *Пусть*

$$m + 1 = m_0 + m_1 + \ \ldots \ + m_n. \tag{6.2}$$

Тогда полином $P_m(x)$ степени m, удовлетворяющий условиям (6.1), *при любых значениях $f^{(j)}(x_i), \quad j = 0, 1, \ldots, m_i - 1, \ i = 0, 1, \ldots, n$, существует и определяется однозначно.*

Доказательство. Условия (6.1) представляют собой систему линейных алгебраических уравнений относительно коэффициентов полинома $P_m(x)$. Условие (6.2) показывает, что число уравнений и число неизвестных в этой системе одно и то же, поэтому для доказательства теоремы достаточно убедиться, что соответствующая однородная система уравнений

$$P_m^{(j)}(x_i) = 0, \quad j = 0, 1, \ldots, m_i - 1, \ i = 0, 1, \ldots, n \tag{6.3}$$

имеет только тривиальное решение. Равенства же (6.3) означают, что точки x_i — корни полинома $P_m(x)$ с кратностями m_i, то есть полином $P_m(x)$ степени m имеет $m+1$ корень, что может быть выполнено лишь в том случае, когда все его коэффициенты равны нулю. □

Как и в случае интерполирования с простыми узлами, можно выписать явный вид полинома $P_m(x)$ либо в форме, аналогичной интерполяционному полиному Ньютона, либо в форме, аналогичной интерполяционному полиному Лагранжа. В последнем случае приходят к так называемой интерполяционной формуле Эрмита:

$$H_m(x) = \sum_{i=0}^{n} \sum_{j=0}^{m_i - 1} f^{(j)}(x_i) \varphi_{ij}(x). \tag{6.4}$$

Здесь $\varphi_{ij}(x)$ — базисные функции Эрмита, а именно полиномы степени m, удовлетворяющие условиям:

$$\varphi_{ij}^{(k)}(x_l) = \begin{cases} 0, & (l-i)^2 + (j-k)^2 \neq 0; \\ 1, & (l-i)^2 + (j-k)^2 = 0. \end{cases} \tag{6.5}$$

К сожалению, явный вид полиномов $\varphi_{ij}(x)$ в произвольном случае оказывается весьма громоздким. Ограничимся поэтому рассмотрением следующего примера.

В точках $x_0 < x_1 < x_2$ заданы значения $f(x_0), f(x_1), f'(x_1), f(x_2)$. Требуется построить полином третьей степени $P_3(x)$ такой, что

$$P_3(x_0) = f(x_0), P_3(x_1) = f(x_1), P_3'(x_1) = f'(x_1), P_3(x_2) = f(x_2). \quad (6.6)$$

Будем искать полином $P_3(x)$ в форме Эрмита:

$$P_3(x) \equiv H_3(x) = f(x_0)\varphi_{00}(x)+f(x_1)\varphi_{10}(x)+f'(x_1)\varphi_{11}(x)+f(x_2)\varphi_{20}(x).$$

При этом требуется построить базисные функции Эрмита, т. е. полиномы третьей степени, удовлетворяющие условиям:

$$\varphi_{00}(x_0) = 1,\ \varphi_{00}(x_1) = 0,\ \varphi_{00}'(x_1) = 0,\ \varphi_{00}(x_2) = 0, \quad (6.7)$$
$$\varphi_{10}(x_0) = 0,\ \varphi_{10}(x_1) = 1,\ \varphi_{10}'(x_1) = 0,\ \varphi_{10}(x_2) = 0, \quad (6.8)$$
$$\varphi_{11}(x_0) = 0,\ \varphi_{11}(x_1) = 0,\ \varphi_{11}'(x_1) = 1,\ \varphi_{11}(x_2) = 0, \quad (6.9)$$
$$\varphi_{20}(x_0) = 0,\ \varphi_{20}(x_1) = 0,\ \varphi_{20}'(x_1) = 0,\ \varphi_{20}(x_2) = 1. \quad (6.10)$$

Условия (6.7) показывают, что x_1, x_2 — корни полинома φ_{00}, причем x_1 есть корень кратности два, поэтому $\varphi_{00}(x) = A(x - x_1)^2(x - x_2)$, где A — постоянная, которую определим, используя первое из условий (6.7). Таким образом,

$$\varphi_{00}(x) = \frac{(x - x_1)^2(x - x_2)}{(x_0 - x_1)^2(x_0 - x_2)}.$$

Аналогично, используя условия (6.10), (6.9), получаем

$$\varphi_{20}(x) = \frac{(x - x_0)(x - x_1)^2}{(x_2 - x_0)(x_2 - x_1)^2},$$

$$\varphi_{11}(x) = \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_1 - x_2)(x_1 - x_0)}.$$

Условия (6.8) определяют только два корня полинома $\varphi_{10}(x)$, поэтому $\varphi_{10}(x) = (x-x_0)(x-x_2)(\alpha x+\beta)$. Постоянные α, β найдем из условий

$$\varphi_{10}(x_1) = (x_1 - x_0)(x_1 - x_2)(\alpha x_1 + \beta) = 1,$$
$$\varphi_{10}'(x_1) = (x_1 - x_0)(x_1 - x_2)\alpha + (\alpha x_1 + \beta)(2x_1 - x_0 - x_2) = 0.$$

Решая эту систему уравнений, получим

$$\alpha = \frac{2x_1 - x_0 - x_2}{(x_1 - x_0)^2(x_1 - x_2)^2},$$

$$\beta = \frac{1}{(x_1 - x_0)(x_1 - x_2)} \left(1 + \frac{(2x_1 - x_0 - x_2)x_1}{(x_1 - x_0)(x_1 - x_2)}\right).$$

Таким образом, все базисные функции Эрмита, а следовательно, и полином $H_3(x)$, построены.

Оценку погрешности интерполяционного полинома с кратными узлами дает

Теорема 6.2. *Пусть $[a, b]$ — отрезок вещественной оси, содержащий все узлы интерполирования $x_0, x_1, \ldots, x_n$, функция f непрерывно дифференцируема $m + 1$ раз на отрезке $[a, b]$. Пусть, далее, точка x принадлежит $[a, b]$ и не совпадает ни с одним из узлов интерполирования. Тогда существует точка $\xi \in [a, b]$ такая, что*

$$f(x) = P_m(x) + \frac{f^{(m+1)}(\xi)}{(m+1)!}\omega(x), \tag{6.11}$$

где P_m — полином степени $m = m_0 + m_1 + m_n - 1$, определяемый условиями (6.1), $\omega(x) = (x - x_0)^{m_0} \cdots (x - x_n)^{m_n}$.

Доказательство этой теоремы аналогично доказательству теоремы 4.1.

§ 2. Среднеквадратичное приближение функций

1. Элемент наилучшего среднеквадратичного приближения. Говорят, что функции f и φ близки в среднеквадратичном смысле на отрезке $[a, b]$, если мал интеграл

$$\int_a^b p(x)(f(x) - \varphi(x))^2 dx.$$

Здесь p — заданная положительная на отрезке $[a, b]$ функция, называемая весом.

Пусть $\varphi_1(x), \varphi_2(x), \ldots, \varphi_n(x)$ — заданные на отрезке $[a, b]$ функции. Функция $\varphi(x) = c_1\varphi_1(x) + c_2\varphi_2(x) + \ldots + c_n\varphi_n(x)$ называется элементом наилучшего среднеквадратичного приближения к функции $f(x)$, если числа $c_1, \ldots, c_n$ доставляют минимальное значение интегралу

$$\int_a^b p(x)\Big(f(x) - \sum_{i=1}^{n} c_k\varphi_k(x)\Big)^2 dx.$$

При исследовании существования и единственности элемента наилучшего приближения и при описании способов его построения будет полезно ввести некоторые новые понятия и обозначения.

Величину

$$(f,g)=\int_a^b p(x)f(x)g(x)dx$$

называют скалярным произведением функций f,g. Скалярное произведение функций обладает свойствами, аналогичными свойствам скалярного произведения векторов:

1) $(f,g)=(g,f)$,

2) $(f,f)>0$ для любой функции f, не равной тождественно нулю.

3) $(\alpha_1 f_1+\alpha_2 f_2,g)=\alpha_1(f_1,g)+\alpha_2(f_2,g)$ для любых функций f_1, f_2, g и чисел α_1, α_2.

Число $\|f\|=(f,f)^{1/2}$ называется нормой функции f. Норма обладает свойствами, аналогичными свойствам длины вектора.

Напомним определение линейной независимости функций. Функции $\varphi_1(x),\ldots,\varphi_n(x)$ называются линейно независимыми на отрезке $[a,b]$, если равенство

$$c_1\varphi_1(x)+c_1\varphi_2(x)+\ldots+c_n\varphi_n(x)=0\quad\forall x\in[a,b]$$

выполнено лишь при условии, что $c_1=c_2=\ldots=c_n=0$.

Введем в рассмотрение вещественную функцию

$$F(c)=\left\|f-\sum_{i=1}^n c_i\varphi_i\right\|^2$$

n независимых переменных $c_1,\ldots,c_n$. Задача построения элемента наилучшего среднеквадратичного приближения эквивалентна задаче минимизации функции F.

Выпишем необходимые условия минимума функции F. Предварительно, используя свойства симметрии и линейности скалярного произведения, преобразуем ее к виду:

$$F(c)=(f,f)+\sum_{i,j=1}^n c_ic_j(\varphi_i,\varphi_j)-2\sum_{i=1}^n c_i(f,\varphi_i).$$

Вычисляя первые производные, получим

$$\frac{\partial F(c)}{\partial c_i}=2\sum_{j=1}^n c_j(\varphi_j,\varphi_i)-2(f,\varphi_i),\quad i=1,2,\ldots,n.$$

Таким образом, система уравнений для отыскания компонент вектора c, доставляющего минимальное значение функции F, имеет вид

$$\sum_{j=1}^{n} c_j(\varphi_j, \varphi_i) = (f, \varphi_i), \quad i = 1, 2, \ldots, n. \tag{1.1}$$

Матрица этой системы линейных уравнений $A = \{(\varphi_i, \varphi_j)\}_{i,j=1}^{n}$ называется матрицей Грама системы функций $\varphi_1, \ldots, \varphi_n$. Матрица A, очевидно, симметрична.

Теорема 1.1. *Для того, чтобы матрица Грама была положительно определена, необходимо и достаточно, чтобы функции $\varphi_1, \ldots, \varphi_n$ были линейно независимы.*

Доказательство. Выполненные выше преобразования функции F показывают, что $(Ac, c) = \left\| \sum_{i=1}^{n} c_i\varphi_i \right\|^2 \geqslant 0$. Если функции $\varphi_1, \ldots, \varphi_n$ линейно независимы, то равенство в последнем неравенстве достигается лишь при $c_1 = c_2 = \ldots = c_n = 0$, т. е. матрица A положительно определена. Обратное утверждение теоремы доказывается аналогично. □

Из только что доказанной теоремы, очевидно, вытекает, что если функции $\varphi_1, \ldots, \varphi_n$ линейно независимы, то элемент наилучшего приближения для любой функции f существует и определяется единственным образом.

Наиболее просто элемент наилучшего приближения строится в том случае, когда матрица A диагональна, т. е., когда $(\varphi_i, \varphi_j) = 0$ при $i \neq j$. Система функций $\varphi_1, \ldots, \varphi_n$ называется в этом случае ортогональной, а элемент наилучшего приближения принимает вид

$$\varphi(x) = \sum_{i=1}^{n} \frac{(f, \varphi_i)}{(\varphi_i, \varphi_i)} \varphi_i(x).$$

Чаще всего на практике применяются ортогональные системы функций, состоящие из тригонометрических функций или из полиномов.

Например, нетрудно проверить, что

$$\int_0^{2\pi} \sin kx\, dx = 0, \quad \int_0^{2\pi} \cos kx\, dx = 0, \quad k = 1, 2, \ldots$$

$$\int_0^{2\pi} \sin kx \cos lx\, dx = 0,\ k, l = 1, 2, \ldots$$

$$\int_0^{2\pi} \sin kx \sin lx\, dx = \begin{cases} 0, & k \neq l; \\ \pi, & k = l \end{cases};\quad \int_0^{2\pi} \cos kx \cos lx\, dx = \begin{cases} 0, & k \neq l; \\ \pi, & k = l. \end{cases}$$

Поэтому функции $1, \sin x, \cos x, \sin 2x, \cos 2x, \ldots \sin nx, \cos nx$ при любом $n \geqslant 1$ образуют ортогональную систему на отрезке $[0, 2\pi]$.

Элемент наилучшего среднеквадратичного приближения по этой системе функций имеет вид

$$\varphi(x) = a_0 + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx),$$

где

$$a_0 = \frac{1}{2\pi} \int_0^{2\pi} f(x)\, dx,\ a_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos kx\, dx,\ b_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin kx\, dx,$$

$k = 1, 2, \ldots, n$.

Функцию $\varphi(x)$ называют в этом случае отрезком ряда Фурье для функции $f(x)$, числа a_k, b_k — коэффициентами Фурье функции $f(x)$.

2. Ортогональные полиномы. Подробнее рассмотрим ортогональные системы функций, образованные полиномами $P_0(x)$, $P_1(x), \ldots, P_n(x)$.

Теорема 2.1. *Для любого заданного отрезка $[a, b]$ и веса $p(x) > 0$ система ортогональных полиномов существует.*

Доказательство. Построим искомую систему полиномов $P_0(x)$, $P_1(x)$, $P_2(x), \ldots, P_n(x)$, применяя метод ортогонализации Грама — Шмидта к системе степеней независимой переменной: $1, x, x^2, \ldots, x^n$. Положим $P_0(x) \equiv 1$. Полином $P_1(x)$ будем разыскивать в виде $P_1(x) = x - \alpha_{10} P_0(x)$. Постоянную α_{10} определим из условия ортогональности $(P_1, P_0) = 0$. Получим: $\alpha_{10} = (x, P_0)/(P_0, P_0)$. Вообще, если полиномы $P_0(x), P_1(x), \ldots, P_k(x)$ уже построены, то полином $P_{k+1}(x)$ будем разыскивать в виде

$$P_{k+1}(x) = x^{k+1} - \sum_{i=1}^{k} \alpha_{k+1,i} P_i(x),$$

определяя постоянные $\alpha_{k+1,i}$ из условий ортогональности $(P_{k+1},P_i)=0$, $i=1,2,\ldots,k$. В результате получим $\alpha_{k+1,i}=(x^{k+1},P_i)/(P_i,P_i)$. $\square$

Отметим некоторые свойства ортогональных полиномов, непосредственно вытекающие из определения ортогональности.

1. Произвольный полином $R_k(x)$ степени k однозначно представим в виде линейной комбинации ортогональных полиномов:

$$R_k(x)=\sum_{i=0}^{k}\alpha_i P_i(x),$$

где $\alpha_i=(R_k,P_i)/(P_i,P_i)$, $i=0,1,\ldots,n$.

2. Ортогональный полином степени k ортогонален любому полиному q_i меньшей степени: $(P_k,q_i)=0$, $i=0,1,\ldots,k-1$[1)].

3. Задание веса и отрезка ортогональности определяет систему ортогональных полиномов с точностью до множителя, то есть если $Q_0(x)$, $Q_1(x)$, $\ldots$ ортогональны, как и полиномы $P_0(x)$, $P_1(x),\ldots$, с весом $p(x)$ на отрезке $[a,b]$, то $Q_i(x)=\alpha_i P_i(x)$, $i=0,1,\ldots$, где α_i — ненулевые постоянные.

4. Ортогональные полиномы удовлетворяют трехчленному рекуррентному соотношению:

$$(x-\beta_{k+1,k})P_k(x)=\beta_{k+1,k+1}P_{k+1}(x)+\beta_{k+1,k-1}P_{k-1}(x), \qquad (2.1)$$

$k=1,2,\ldots$

Докажем лишь последнее утверждение. Остальные доказываются проще. Представим полином $xP_k(x)$ в виде разложения по ортогональным полиномам:

$$xP_k(x)=\beta_{k+1,k+1}P_{k+1}(x)+\beta_{k+1,k}P_k(x)+\beta_{k+1,k-1}P_{k-1}(x)+\ldots+\\ +\beta_{k+1,0}P_0(x),$$

где $\beta_{k+1,i}=(xP_k,P_i)/(P_i,P_i)$, $i=0,1,\ldots,k+1$. Записывая более подробно числитель последнего выражения, получим

$$(xP_k,P_i)=\int_a^b p(x)xP_k(x)P_i(x)\,dx=(P_k,q_{i+1}),$$

где $q_{i+1}(x)$ — полином степени $i+1$. Отсюда вследствие свойства 2) вытекает, что $\beta_{k+1,i}=0$ при $i<k-1$. $\square$

[1)] Нетрудно убедиться, что это свойство можно принять за определение ортогонального полинома.

Теорема 2.2. *Пусть $P_n(x)$ — полином, ортогональный на отрезке $[a,b]$. Тогда все его корни вещественны, различны и принадлежат интервалу (a,b).*

Доказательство. Обозначим через $x_1, x_2, \ldots x_k$ точки, принадлежащие (a,b), в которых полином $P_n(x)$ меняет знак. Предположение о том, что $k < n$ приводит к противоречию. Действительно, в этом случае $P_n(x) = (x-x_1)(x-x_2)\ldots(x-x_k)Q_{n-k}(x)$, где $Q_{n-k}(x)$ — полином степени $n-k$. Ясно, что корнями полинома $Q_{n-k}(x)$ по построению являются корни полинома $P_n(x)$, принадлежащие (a,b) и имеющие четную кратность, а также корни полинома $P_n(x)$, не принадлежащие (a,b). Поэтому полином $Q_{n-k}(x)$ не меняет знака на отрезке $[a,b]$. Положим $q_k(x) = (x-x_1)(x-x_2)\ldots(x-x_k)$. Тогда

$$(P_n, q_k) = \int_a^b p(x) q_k^2(x) Q_{n-k}(x)\,dx \neq 0,$$

поскольку подынтегральное выражение не меняет знака на интервале (a,b) и лишь в конечном числе точек обращается в нуль. Но, с другой стороны, $(P_n, q_k) = 0$, так как полином $P_n(x)$ ортогонален, а $k < n$. $\square$

Корни ортогональных полиномов обладают весьма важным свойством перемежаемости. Сформулируем его без доказательства.

Теорема 2.3. *Пусть $P_n(x), P_{n+1}(x),\ n \geqslant 1$ — полиномы, ортогональные на отрезке $[a,b]$. Обозначим через $x_1^{(n)} < x_2^{(n)} < \ldots < x_n^{(n)}$, $x_1^{(n+1)} < x_2^{(n+1)} < \ldots < x_{n+1}^{(n+1)}$ их корни. Тогда*

$$a < x_1^{(n+1)} < x_1^{(n)} < x_2^{(n+1)} < x_2^{(n)} < \ldots < x_n^{(n)} < x_{n+1}^{(n+1)} < b.$$

Примеры ортогональных полиномов.

1. Полиномы Чебышева ортогональны на отрезке $[-1,1]$ с весом $p(x) = 1/\sqrt{1-x^2}$. Действительно, используя представление (5.4), получим

$$(T_k, T_l) = \int_{-1}^{1} \frac{\cos(k \arccos x)\cos(l \arccos x)}{\sqrt{1-x^2}}\,dx.$$

Полагая $x = \cos\varphi$, нетрудно подсчитать, что

$$(T_k, T_l) = \int_0^{\pi} \cos k\varphi \cos l\varphi\,d\varphi = \frac{1}{2}\int_0^{\pi} (\cos(k+l)\varphi + \cos(k-l)\varphi)\,d\varphi = 0$$

при $k \neq l$.

2. Полиномы Лежандра. Построим полиномы, ортогональные с весом $p(x) \equiv 1$ на отрезке $[-1, 1]$. Такие полиномы называют полиномами Лежандра. Обозначим искомый полином через $L_n(x), n \geqslant 1$. По определению

$$\int_{-1}^{1} L_n(x) q_{n-1}(x)\, dx = 0 \tag{2.2}$$

для любого полинома $q_{n-1}(x)$ степени $n-1$.

Положим $L_n(x) = Q_{2n}^{(n)}(x)$, где $Q_{2n}(x)$ — некоторый полином степени $2n$, и преобразуем интеграл в правой части равенства (2.2) при помощи формулы интегрирования по частям. Получим

$$\int_{-1}^{1} Q_{2n}^{(n)}(x) q_{n-1}(x)\, dx = -\int_{-1}^{1} Q_{2n}^{(n-1)}(x) q'_{n-1}(x)\, dx + Q_{2n}^{(n-1)}(x) q_{n-1}(x) \Big|_{-1}^{1}.$$

Продолжая аналогичные преобразования и учитывая, что $q_{n-1}^{(n)}(x) \equiv 0$, можем написать:

$$\int_{-1}^{1} Q_{2n}^{(n)}(x) q_{n-1}(x)\, dx = Q_{2n}^{(n-1)}(x) q_{n-1}(x) \Big|_{-1}^{1} - Q_{2n}^{(n-2)}(x) q'_{n-1}(x) \Big|_{-1}^{1} +$$
$$+ Q_{2n}^{(n-3)}(x) q''_{n-1}(x) \Big|_{-1}^{1} - \ldots + (-1)^{n-1} Q_{2n}(x) q_{n-1}^{(n-1)}(x) \Big|_{-1}^{1}.$$

Условие (2.2) будет выполнено, если положить, что

$$Q_{2n}(1) = Q'_{2n}(1) = \ldots = Q_{2n}^{(n-1)}(1) = 0, \tag{2.3}$$
$$Q_{2n}(-1) = Q'_{2n}(-1) = \ldots = Q_{2n}^{(n-1)}(-1) = 0. \tag{2.4}$$

Равенства (2.3), (2.4) означают, что точки $-1, +1$ — корни кратности n полинома $Q_{2n}(x)$, т. е. $Q_{2n}(x) = A(x-1)^n(x+1)^n$, где A — произвольная постоянная. Таким образом, полином Лежандра представим в виде

$$L_n(x) = \frac{d^n}{dx^n}\left(1 - x^2\right)^n. \tag{2.5}$$

Полученную формулу называют формулой Родрига.

§ 3. Приближенное вычисление интегралов

В этом параграфе будут рассмотрены методы приближенного вычисления определенных интегралов

$$\int_a^b f(x)dx.$$

Лишь в исключительных случаях удается найти первообразную $F(x)$ функции $f(x)$ и представить интеграл в виде

$$\int_a^b f(x)dx = F(b) - F(a).$$

Как правило, для вычисления интегралов приходится применять приближенные методы, и чаще всего применяют так называемые формулы механических квадратур, т. е. разыскивают приближенное значение интеграла

$$\int_a^b f(x)dx \approx \sum_{k=1}^{n} c_k f(x_k).$$

Числа c_k называют весами, или коэффициентами квадратурной формулы, x_k — узлами квадратурной формулы.

Конкретные квадратурные формулы отличаются, таким образом, выбором коэффициентов и узлов.

1. Интерполяционные квадратурные формулы. При описании этого класса квадратурных формул интеграл будет удобно записывать в виде

$$I(f) = \int_a^b p(x)f(x)dx,$$

где $p(x)$ — положительная на отрезке $[a, b]$ функция, называемая весом.

Выберем на отрезке $[a, b]$ точки $a \leqslant x_1 < x_2 \ldots < x_n \leqslant b$. Построим полином $L_{n-1}(x)$ степени $n-1$, интерполирующий функцию $f(x)$ по указанным узлам, и примем за приближенное значение интеграла

величину

$$S(f) = \int_a^b p(x) L_{n-1}(x) dx.$$

В дальнейшем будем пользоваться также следующим обозначением для погрешности квадратурной формулы $R(f) = I(f) - S(f)$. Величину $R(f)$ будем называть также остаточным членом квадратурной формулы. Записывая интерполяционный полином $L_{n-1}(x)$ в форме Лагранжа (см. (2.3), § 1): $L_{n-1}(x) = \sum_{k=1}^{n} f(x_k)\varphi_k(x)$, получим

$$S(f) = \sum_{k=1}^{n} c_k f(x_k), \tag{1.1}$$

где

$$c_k = \int_a^b p(x)\varphi_k(x) dx, \quad k = 1, 2, \ldots, n. \tag{1.2}$$

Здесь $\varphi_k(x)$ — базисные функции Лагранжа, определяемые формулами (2.2), § 1.

Квадратурная формула вида (1.1) с коэффициентами, вычисляемыми по формулам (1.2), называется интерполяционной квадратурной формулой.

Следующая теорема дает другое, эквивалентное описание интерполяционных квадратурных формул.

Теорема 1.1. *Для того, чтобы квадратурная формула вида* (1.1) *была интерполяционной, необходимо и достаточно, чтобы она была точной на любом полиноме степени* $n - 1$, *то есть* $R(P_{n-1}) = 0$ *для любого полинома* $P_{n-1}(x)$ *степени* $n - 1$.

Доказательство. Пусть формула (1.1) является интерполяционной. Тогда для любого полинома $P_{n-1}(x)$ имеем $R(P_{n-1}) = I(P_{n-1}) - S(P_{n-1}) = I(P_{n-1}) - I(L_{n-1})$, где $L_{n-1}(x)$ — полином степени $n - 1$, интерполирующий полином P_{n-1} по узлам $x_1, x_2, \ldots, x_n$. Вследствие теоремы 1.1, § 1, $L_{n-1}(x) \equiv P_{n-1}(x)$, т. е. $R(P_{n-1}) = 0$. Докажем обратное утверждение. Поскольку базисная функция Лагранжа $\varphi_k(x)$ при любом $k = 1, 2, \ldots, n$ есть полином степени $n - 1$, то $I(\varphi_k) = S(\varphi_k)$, но $S(\varphi_k) = \sum_{i=1}^{n} c_i \varphi_k(x_i) = c_k$. $\square$

2. Устойчивость квадратурных формул. Важным свойством квадратурных формул является их устойчивость по отношению к ошибкам вычисления подынтегральной функции. При вычислении значений функции f неизбежно возникают ошибки, например, за счет округлений. В результате, вместо значений $f(x_k)$ получаем значения $\widetilde{f}(x_k) = f(x_k) + \varepsilon_k$, и, соответственно, вместо $S(f)$ значение $\widetilde{S}(f) = S(f) + \sum\limits_{k=1}^{n} c_k\varepsilon_k$.

Квадратурная формула называется устойчивой, если существует постоянная C такая, что равномерно по числу узлов $\sum\limits_{k=1}^{n} |c_k| \leqslant C$.

Если предположить, что равномерно по числу узлов квадратурной формулы выполняется оценка $|\varepsilon_k| \leqslant \varepsilon$, то для устойчивой квадратурной формулы $|S(f) - \widetilde{S}(f)| \leqslant C\varepsilon$, т. е. погрешность вычисления $S(f)$ определяется лишь погрешностью вычисления значений подынтегральной функции.

Справедлив следующий простой, но практически очень важный, достаточный признак устойчивости интерполяционных квадратурных формул.

Теорема 2.1. *Если при любом $n \geqslant 1$ все коэффициенты интерполяционной квадратурной формулы неотрицательны, то она устойчива.*

Доказательство. По определению при любом $n \geqslant 1$ интерполяционная квадратурная формула точна на полиноме нулевой степени, следовательно, учитывая неотрицательность коэффициентов c_k, получим

$$\int\limits_a^b p(x)\,dx = \sum_{k=1}^{n} c_k = \sum_{k=1}^{n} |c_k|. \ \square$$

3. Квадратурные формулы Ньютона — Котеса. Так называют интерполяционные квадратурные формулы при $p(x) \equiv 1$ и равномерно расположенных на отрезке $[a, b]$ узлах: $x_i = x_1 + (i-1)h$, $i = 2, \ldots, n$, $h = \text{const}$. Выполняя в формуле (1.2) замену переменной $x = x_1 + th$, нетрудно получить, что

$$c_k = (b-a)\widehat{c}_k, \quad k = 1, 2, \ldots, n,$$

где числа $\widehat{c}_k$ зависят только от k и n.

Приведем примеры простейших квадратурных формул Ньютона — Котеса.

1. Квадратурная формула левых прямоугольников. В этом случае $n = 1$, $x_1 = a$,

$$\int_a^b f(x)\,dx \approx (b-a)f(a). \tag{3.1}$$

2. Квадратурная формула правых прямоугольников, $n = 1$, $x_1 = b$,

$$\int_a^b f(x)\,dx \approx (b-a)f(b). \tag{3.2}$$

3. Квадратурная формула центральных прямоугольников, $n = 1$, $x_1 = c = (b+a)/2$,

$$\int_a^b f(x)\,dx \approx (b-a)f(c). \tag{3.3}$$

4. Квадратурная формула трапеций. В этом случае $n = 2$, $x_1 = a$, $x_2 = b$, $h = b - a$, $\varphi_1(x) = (b-x)/(b-a)$, $\varphi_2(x) = (a-x)/(a-b)$, $c_1 = c_2 = (b-a)/2$,

$$\int_a^b f(x)\,dx \approx \frac{(b-a)}{2}(f(a)+f(b)). \tag{3.4}$$

5. Квадратурная формула Симпсона (парабол). В этом случае $n = 3$, $x_1 = a$, $x_2 = c = (a+b)/2$, $x_3 = b$, $h = (b-a)/2$,

$$\varphi_1(x) = \frac{(c-x)(b-x)}{(c-a)(b-a)},\ \varphi_2(x) = \frac{(a-x)(b-x)}{(a-c)(b-c)},\ \varphi_3(x) = \frac{(a-x)(c-x)}{(a-b)(c-b)},$$

$c_1 = c_3 = (b-a)/6,\ c_2 = 2(b-a)/3$,

$$\int_a^b f(x)\,dx \approx \frac{(b-a)}{6}(f(a)+4f(c)+f(b)). \tag{3.5}$$

Формулы (3.1)–(3.5) имеют простую геометрическую интерпретацию. Применяя формулу (3.1), мы заменяем площадь криволинейной трапеции, т. е.

$$\int_a^b f(x)\,dx,$$

площадью прямоугольника с основанием $b-a$ и высотой, равной $f(a)$, при использовании формулы (3.2) — площадью прямоугольника с тем же основанием и высотой, равной $f(b)$. При использовании формулы (3.3) высота прямоугольника равна $f(c)$. В случае формулы (3.4) площадь криволинейной трапеции заменяется площадью прямолинейной трапеции, ограниченной прямой, проходящей через точки $(a, f(a))$, $(b, f(b))$, и, наконец, при использовании формулы (3.5) вычисляется площадь криволинейной трапеции, ограниченной параболой, проходящей через три точки: $(a, f(a))$, $(c, f(c))$, $(b, f(b))$.

Геометрически очевидно, что при достаточно плавно меняющейся на отрезке $[a, b]$ функции $f(x)$ наиболее точной должна быть формула парабол.

Формулы Ньютона — Котеса при $n > 4$ редко используются на практике. Дело в том, что уже при $n = 8$ среди коэффициентов формулы встречаются отрицательные, и сумма модулей коэффициентов неограниченно возрастает с увеличением n, т. е. формулы Ньютона — Котеса неустойчивы.

4. Оценки точности простейших квадратурных формул Ньютона — Котеса. В этом пункте, предполагая, что функция f дифференцируема на отрезке $[a, b]$ достаточное число раз, получим представления для погрешностей квадратурных формул, построенных в предыдущем пункте.

1. Квадратурные формулы правых и левых прямоугольников. Для формулы левых прямоугольников по построению

$$R(f) = \int_a^b (f(x) - f(a))\, dx.$$

Применяя формулу конечных приращений Лагранжа, получим

$$f(x) - f(a) = f'(\zeta(x))(x - a),$$

Нетрудно видеть, что функция

$$f'(\zeta(x)) = \frac{f(x) - f(a)}{x - a}$$

непрерывна на отрезке $[a, b]$. При $x \neq a$ это следует из непрерывности функции f. Особенность при $x = a$ устранима, поскольку

$$\lim_{x \to a} f'(\zeta(x)) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = f'(a).$$

Таким образом, получаем, что

$$R(f) = \int_a^b f'(\zeta(x))(x-a)\,dx.$$

Функция $x-a$ не меняет знака на отрезке $[a,b]$. Поэтому, применяя теорему о среднем, приходим к представлению остаточного члена для квадратурной формулы левых прямоугольников

$$R(f) = f'(\xi)\int_a^b (x-a)dx = f'(\xi)\frac{(b-a)^2}{2}, \quad \xi \in (a,b). \tag{4.1}$$

Совершенно аналогичные рассуждения приводят к представлению остаточного члена квадратурной формулы правых прямоугольников

$$R(f) = f'(\xi)\int_a^b (x-b)dx = -f'(\xi)\frac{(b-a)^2}{2}, \quad \xi \in (a,b). \tag{4.2}$$

2. Квадратурная формула трапеций. В этом случае

$$R(f) = \int_a^b (f(x) - L_1(x))dx,$$

где $L_1(x)$ — полином первой степени, удовлетворяющий условиям: $L_1(a) = f(a)$, $L_1(b) = f(b)$. Используя формулу (4.1), § 1, для остаточного члена интерполяционного полинома, будем иметь

$$R(f) = \int_a^b \frac{f''(\zeta(x))}{2}(x-a)(x-b)dx,$$

причем функция $(x-a)(x-b)$ не меняет знака на отрезке интегрирования, а функция $f''(\zeta(x))$, как нетрудно проверить при помощи правили Лопиталя, непрерывна на этом отрезке. Вновь применяя теорему о среднем, получим

$$R(f) = \frac{f''(\xi)}{2}\int_a^b (x-a)(x-b)dx = -\frac{f''(\xi)}{12}(b-a)^3, \quad \xi \in (a,b). \tag{4.3}$$

3. Квадратурная формула центральных прямоугольников. Рассуждая сначала, как и в случае квадратурной формулы левых прямоугольников, получим

$$R(f) = \int_a^b f'(\zeta(x))(x-c)\,dx,$$

но функция $x-c$ меняет знак в точке c. Поэтому применить теорему о среднем здесь не удается. В связи с этим введем в рассмотрение полином первой степени $L_1(x) = f(c)+\alpha(x-c)$. При любом значении постоянной α

$$\int_a^b L_1(x)\,dx = f(c)(b-a),$$

следовательно,

$$R(f) = \int_a^b (f(x) - L_1(x))\,dx.$$

Выберем теперь α так, чтобы $L_1'(c) = f'(c)$. Ясно, что для этого достаточно положить $\alpha = f'(c)$. При таком α точка c оказывается узлом интерполирования кратности два, следовательно, применяя формулу (6.11), § 1, можно написать

$$f(x) - L_1(x) = \frac{f''(\zeta(x))}{2}(x-c)^2,$$

причем функция $f''(\zeta(x))$ непрерывна по x на $[a,b]$. Теперь теорема о среднем применима, поэтому

$$R(f) = \frac{f''(\xi)}{2}\int_a^b (x-c)^2\,dx = \frac{f''(\xi)}{24}(b-a)^3, \quad \xi \in (a,b). \tag{4.4}$$

4. Формула Симпсона. В этом случае

$$R(f) = \int_a^b (f(x) - L_2(x))\,dx,$$

где $L_2(x)$ — полином второй степени, удовлетворяющий условиям

$$L_2(a) = f(a),\ L_2(c) = f(c),\ L_2(b) = f(b).$$

На основании формулы (4.1), § 1, имеем

$$R(f) = \int_a^b \frac{f'''(\zeta(x))}{3!}(x-a)(x-c)(x-b)\,dx.$$

Функция $f'''(\zeta(x))$ непрерывна по x, функция $(x-a)(x-c)(x-b)$ меняет знак в точке c. Вновь для того, чтобы получить возможность применить теорему о среднем, повысим порядок интерполяционного полинома, а именно, введем в рассмотрение полином

$$L_3(x) = L_2(x) + \alpha(x-a)(x-c)(x-b).$$

Нетрудно проверить, что функция $(x-a)(x-c)(x-b)$ нечетна относительно точки c, следовательно,

$$\int_a^b (x-a)(x-c)(x-b)\,dx = 0,$$

поэтому

$$R(f) = \int_a^b (f(x) - L_3(x))\,dx.$$

При любом значении α имеем $L_3(a) = f(a), L_3(c) = f(c), L_3(b) = f(b)$. Заметим теперь, что $((x-a)(x-c)(x-b))'|_{x=c} \neq 0$, следовательно, постоянную α можно выбрать так, чтобы $L_3'(c) = f'(c)$. Но тогда на основании формулы (6.11), § 1, получим

$$R(f) = \int_a^b \frac{f^{\mathrm{IV}}(\zeta(x))}{4!}(x-a)(x-c)^2(x-b)\,dx = \\ = -\frac{f^{\mathrm{IV}}(\xi)}{2880}(b-a)^5, \quad \xi \in (a,b). \quad (4.5)$$

5. Составные квадратурные формулы. На практике формулы (3.1)–(3.5) чаще всего используют как составные, т. е. предварительно отрезок $[a,b]$ разбивают на частичные отрезки достаточно малой длины, затем на каждом частичном отрезке применяют ту или иную квадратурную формулу. Введем необходимые обозначения. Положим $a = x_0 < x_1 < \ldots < x_N = b$, $h_i = x_i - x_{i-1}$, $i = 1, 2, \ldots, N$, $h = \max\limits_{1 \leqslant i \leqslant N} h_i$, $x_{i-1/2} = x_i - h_i/2$.

1. Формула левых прямоугольников:

$$\int_a^b f(x)\,dx = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} f(x)\,dx \approx \sum_{i=1}^{N} f(x_{i-1})h_i. \tag{5.1}$$

Используя (4.1), получим

$$R(f) = \sum_{i=1}^{N} \left(\int_{x_{i-1}}^{x_i} f(x)\,dx - f(x_{i-1})h_i \right) = \frac{1}{2} \sum_{i=1}^{N} f'(\xi_i)h_i^2,$$

$\xi_i \in (x_{i-1}, x_i)$, $i = 1, \ldots, N$, откуда

$$|R(f)| \leqslant \frac{M_1}{2} h \sum_{i=1}^{N} h_i = \frac{M_1}{2} h(b-a), \tag{5.2}$$

где $M_1 = \max\limits_{a \leqslant x \leqslant b} |f(x)|$.

2. Формула правых прямоугольников:

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{N} f(x_i)h_i. \tag{5.3}$$

Используя (4.2), получим

$$R(f) = -\frac{1}{2} \sum_{i=1}^{N} f'(\xi_i)h_i^2,$$

$\xi_i \in (x_{i-1}, x_i)$, $i = 1, \ldots, N$, откуда

$$|R(f)| \leqslant \frac{M_1}{2} h(b-a). \tag{5.4}$$

3. Формула центральных прямоугольников:

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{N} f(x_{i-1/2})h_i. \tag{5.5}$$

Используя (4.4), получим

$$R(f) = \frac{1}{24} \sum_{i=1}^{N} f''(\xi_i)h_i^3,$$

$\xi_i \in (x_{i-1}, x_i)$, $i = 1, \ldots, N$, откуда

$$|R(f)| \leqslant \frac{M_2}{24} h^2 (b-a), \tag{5.6}$$

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

4. Формула трапеций:

$$\int\limits_a^b f(x)\,dx \approx \sum_{i=1}^{N} \frac{f(x_{i-1}) + f(x_i)}{2} h_i. \tag{5.7}$$

Используя (4.3), получим

$$R(f) = -\frac{1}{12} \sum_{i=1}^{N} f''(\xi_i) h_i^3,$$

$\xi_i \in (x_{i-1}, x_i)$, $i = 1, \ldots, N$, откуда

$$|R(f)| \leqslant \frac{M_2}{12} h^2 (b-a). \tag{5.8}$$

5. Формула Симпсона:

$$\int\limits_a^b f(x)\,dx \approx \sum_{i=1}^{N} \frac{f(x_{i-1}) + 4f(x_{i-1/2}) + f(x_i)}{6} h_i. \tag{5.9}$$

Используя (4.5), получим

$$R(f) = -\frac{1}{2880} \sum_{i=1}^{N} f^{\mathrm{IV}}(\xi_i) h_i^5,$$

$\xi_i \in (x_{i-1}, x_i)$, $i = 1, \ldots, N$, откуда

$$|R(f)| \leqslant \frac{M_4}{2880} h^4 (b-a), \tag{5.10}$$

где $M_4 = \max\limits_{a \leqslant x \leqslant b} |f^{\mathrm{IV}}(x)|$.

6. Квадратурные формулы типа Гаусса. Интерполяционная квадратурная формула с n узлами при любом расположении узлов точна на любом полиноме степени $n-1$. Однако при удачном расположении узлов точность может повышаться. Так, квадратурная формула центральных прямоугольников оказывается точной на любом полиноме первой степени, а квадратурная формула Симпсона — на любом полиноме третьей степени (см. соответствующие представления остаточных членов (4.4), (4.5)). В связи с этим возникает задача: указать такой способ расположения узлов интерполяционной квадратурной формулы, чтобы сделать ее точной на полиномах возможно более высокой степени. Понятно, что, имея в распоряжении n свободных параметров, а именно $x_1, x_2, \ldots, x_n$, можно попытаться построить формулу, точную на любом полиноме степени $2n-1$. Следующая теорема показывает, что большей точности добиться нельзя.

Теорема 6.1. *Ни при каком расположении узлов квадратурная формула вида* (1.1) *не может быть точной на любом полиноме степени* $2n$.

Доказательство. Пусть $P_{2n}(x) = (x-x_1)^2(x-x_2)^2\cdots(x-x_n)^2$. Тогда

$$\int_a^b p(x)P_{2n}(x)\,dx > 0, \quad \text{а} \quad \sum_{k=1}^{n} c_k P_{2n}(x_k) = 0. \ \square$$

Следующая теорема гарантирует существование квадратурной формулы с n узлами, точной на любом полиноме степени $2n-1$, и, более того, указывает способ ее построения.

Теорема 6.2. *Для того, чтобы квадратурная формула*

$$I(f) = \int_a^b p(x)f(x)\,dx \approx S(f) = \sum_{k=1}^{n} c_k f(x_k) \tag{6.1}$$

была точна на любом полиноме степени $2n-1$, *необходимо и достаточно, чтобы:*

1) квадратурная формула была интерполяционной,

2) узлы квадратурной формулы совпадали с корнями полинома степени n, *ортогонального с весом* $p(x)$ *на отрезке* $[a,b]$.

Доказательство. Пусть выполнены условия 1), 2). Покажем, что квадратурная формула (6.1) точна на любом полиноме $P_{2n-1}(x)$ степени $2n-1$. Представим полином $P_{2n-1}(x)$ в виде

$$P_{2n-1}(x) = \omega_n(x)q_{n-1}(x) + r_{n-1}(x), \tag{6.2}$$

где $\omega_n(x) = (x-x_1)(x-x_2)\cdots(x-x_n)$, а $q_{n-1}(x),\ r_{n-1}(x)$ — полиномы степени $n-1$, однозначно определяемые по полиному $P_{2n-1}(x)$. По построению полином $\omega_n(x)$ ортогонален с весом $p(x)$ на отрезке $[a,b]$, поэтому

$$\int_a^b p(x)P_{2n-1}(x)\,dx = \int_a^b p(x)r_{n-1}(x)\,dx.$$

Формула (6.1) интерполяционная, поэтому она точна на любом полиноме степени $n-1$, следовательно,

$$\int_a^b p(x)r_{n-1}(x)\,dx = \sum_{k=1}^{n} c_k r_{n-1}(x_k),$$

но, учитывая (6.2), имеем $r_{n-1}(x_k) = P_{2n-1}(x_k),\ k = 1, 2, \ldots, n$. Таким образом,

$$\int_a^b p(x)P_{2n-1}(x)\,dx = \sum_{k=1}^{n} c_k P_{2n-1}(x_k).$$

Обратно, если квадратурная формула точна на любом полиноме степени $2n-1$, то она точна и на любом полиноме степени $n-1$. Осталось показать, что узлы такой квадратурной формулы совпадают с корнями полинома степени n, ортогонального с весом $p(x)$ на отрезке $[a,b]$. Действительно, полином $\omega_n(x)q_{n-1}(x)$, где $q_{n-1}(x)$ — произвольный полином степени $n-1$, есть полином степени $2n-1$, следовательно,

$$\int_a^b p(x)\omega_n(x)q_{n-1}(x)\,dx = \sum_{k=1}^{n} c_k\omega_n(x_k)q_{n-1}(x_k) = 0,$$

так как $\omega_n(x_k) = 0,\ k = 1, 2, \ldots, n$. □

Таким образом, построение квадратурной формулы, точной на любом полиноме степени $2n-1$, сводится к следующему.

Сначала вычисляются корни $x_1, x_2, \ldots, x_n$ полинома степени n, ортогонального с весом $p(x)$ на отрезке $[a,b]$. Затем коэффициенты c_k вычисляются по формулам

$$c_k = \int_a^b p(x)\varphi_k(x)\,dx, \quad k = 1, 2, \ldots, n,$$

где $\varphi_k(x)$ — базисные функции Лагранжа, построенные по узлам $x_1, x_2, \ldots, x_n$.

Квадратурные формулы, точные на любом полиноме степени $2n-1$, называют квадратурными формулами наивысшей алгебраической степени точности, или квадратурными формулами типа Гаусса.

Заметим, что из доказательства теоремы 6.2 и свойства единственности ортогональных полиномов вытекает единственность квадратурной формулы наивысшей алгебраической степени точности.

Теорема 6.3. *Коэффициенты квадратурной формулы типа Гаусса положительны.*

Доказательство. Пусть $1 \leqslant k \leqslant n$. Положим

$$\psi_k(x) = (x - x_1)^2 (x - x_2)^2 \cdots (x - x_{k-1})^2 (x - x_{k+1})^2 \cdots (x - x_n)^2.$$

Ясно, что $\psi_k(x) \geqslant 0$ — полином степени $2n - 2$, причем $\psi_k(x_j) = 0$ при $j \neq k$, следовательно,

$$\int_a^b p(x)\psi_k(x)\,dx = \sum_{j=1}^{n} c_j \psi_k(x_j) = c_k \psi_k(x_k),$$

т. е. $c_k > 0$, так как $\psi_k(x_k) > 0$. $\square$

Из этой теоремы и теоремы 2.1 вытекает, что квадратурные формулы типа Гаусса устойчивы.

Получим представление для остаточного члена квадратурной формулы типа Гаусса.

Теорема 6.4. *Пусть функция f непрерывно дифференцируема $2n$ раз на отрезке $[a, b]$. Тогда для остаточного члена квадратурной формулы типа Гаусса справедливо представление*

$$R(f) = \frac{f^{(2n)}(\xi)}{(2n)!} \int_a^b p(x)\omega_n^2(x)\,dx, \quad \xi \in (a, b). \tag{6.3}$$

Доказательство. Построим полином $L_{2n-1}(x)$ степени $2n - 1$, удовлетворяющий условиям:

$$L_{2n-1}(x_1) = f(x_1),\ L_{2n-1}(x_2) = f(x_2), \ldots, L_{2n-1}(x_n) = f(x_n),$$
$$L'_{2n-1}(x_1) = f'(x_1),\ L'_{2n-1}(x_2) = f'(x_2), \ldots, L'_{2n-1}(x_n) = f'(x_n),$$

а также полином $\widetilde{L}_{n-1}$ степени $n - 1$, удовлетворяющий условиям:

$$\widetilde{L}_{n-1}(x_1) = f(x_1),\ \widetilde{L}_{n-1}(x_2) = f(x_2), \ldots, \widetilde{L}_{n-1}(x_n) = f(x_n).$$

По определению

$$R(f) = \int_a^b p(x)(f(x) - \widetilde{L}_{n-1}(x))\,dx.$$

Выполним очевидное преобразование

$$R(f) = \int_a^b p(x)(f(x) - L_{2n-1}(x) + L_{2n-1}(x) - \widetilde{L}_{n-1}(x))\,dx.$$

Вновь используя определение квадратурной формулы типа Гаусса, получим

$$\int_a^b p(x)\widetilde{L}_{n-1}(x) = \sum_{k=1}^{n} c_k f(x_k),$$

а поскольку эта формула точна на любом полиноме степени $2n-1$, то

$$\int_a^b p(x)L_{2n-1}(x)\,dx = \sum_{k=1}^{n} c_k L_{2n-1}(x_k),$$

но $L_{2n-1}(x_k) = f(x_k),\ k = 1, 2, \ldots, n$, поэтому

$$R(f) = \int_a^b p(x)(f(x) - L_{2n-1}(x))\,dx.$$

Используя теперь формулу (6.2), § 1, получим:

$$f(x) - L_{2n-1}(x) = \frac{f^{(2n)}(\zeta(x))}{(2n)!}\omega_n^2(x).$$

Применяя правило Лопиталя, нетрудно убедиться, что функция $f^{(2n)}(\zeta(x))$ непрерывна по x на отрезке $[a, b]$. Для завершения доказательства теоремы достаточно учесть неотрицательность функции $p(x)\omega_n^2(x)$ на отрезке $[a, b]$ и применить теорему о среднем. □

Приведем примеры квадратурных формул типа Гаусса.

1. Формула Эрмита. Так называют формулу типа Гаусса при $a = -1,\ b = 1,\ p(x) = 1/\sqrt{1-x^2}$. Полиномы, ортогональные с указанным весом, — полиномы Чебышева $T_n(x)$. Поэтому узлы квадратурной формулы Эрмита вычисляются по явным формулам:

$$x_k = \cos\frac{(2k-1)\pi}{2n}, \quad k = 1, 2, \ldots, n.$$

Коэффициенты квадратурной формулы Эрмита тоже удается выразить в явном виде: $c_k = \pi/n,\ k = 1, 2, \ldots, n$. Отметим, что квадратурная формула Эрмита — это единственная квадратурная формула наивысшей алгебраической степени точности, у которой при любом n все коэффициенты равны между собой.

2. Квадратурная формула Гаусса (часто называемая также квадратурной формулой Гаусса — Лежандра). В этом случае $a = -1$, $b = 1$, $p(x) \equiv 1$. Соответствующие ортогональные полиномы — полиномы Лежандра $L_n(x)$. Их нули в явном виде удается найти лишь при малых значениях n.

Пользуясь формулой Родрига (2.5), § 1, легко подсчитать, что (с точностью до постоянного множителя):

$$L_1(x) = x,\ L_2(x) = 3x^2 - 1,\ L_3(x) = 5x^3 - 3x.$$

Теперь нетрудно вычислить узлы и коэффициенты квадратурной формулы Гаусса при соответствующих значениях n:

1) $n = 1$, $x_1 = 0$, $c_1 = 2$;

2) $n = 2$, $x_1 = -1/\sqrt{3}$, $x_2 = 1/\sqrt{3}$, $c_1 = c_2 = 1$;

3) $n = 3$, $x_1 = -\sqrt{3/5}$, $x_2 = 0$, $x_3 = \sqrt{3/5}$, $c_1 = c_3 = 5/9$, $c_2 = 8/9$.

При больши́х n узлы и коэффициенты формулы Гаусса вычисляются приближенно. Известны таблицы (см., например, [3]), содержащие соответствующие значения.

Глава 3

Численные методы решения дифференциальных уравнений

§ 1. Численные методы решения задачи Коши

В этом параграфе будут рассмотрены методы приближенного решения задачи Коши для обыкновенного дифференциального уравнения первого порядка

$$u'(x) = f(x,u),\ x > x_0, \tag{1}$$

$$u(x_0) = u_0, \tag{2}$$

При построении приближенных методов вводится сетка, т. е. дискретное множество точек $x_0 < x_1 < x_2 < \dots$ на полуоси $x \geqslant x_0$. В качестве приближенного решения разыскивается сеточная функция $y(x)$, т. е. функция, определяемая совокупностью значений $y(x_0),\ y(x_1), \dots$ Для простоты изложения в дальнейшем будем ограничиваться лишь случаем равномерной сетки: $x_i = x_0 + ih$, h — шаг сетки. Будем также использовать обозначение: $y_k = y(x_k)$, $k = 0, 1, \dots$.

1. Метод, основанный на формуле Тейлора. Предполагая, что решение задачи (1), (2), т. е. функция $u(x)$, имеет $n+1$ непрерывную производную, можем написать, что

$$u(x_1) = u(x_0+h) = u(x_0) + hu'(x_0) + \frac{h^2}{2!}u''(x_0) + \frac{h^3}{3!}u'''(x_0) + \dots + \frac{h^n}{n!}u^{(n)}(x_0) + \frac{h^{n+1}}{(n+1)!}u^{(n+1)}(\xi),\ \xi \in (x_0, x_1). \tag{1.1}$$

Покажем, что если функция двух переменных $f(x,p)$ дифференцируема достаточное число раз, то значения производных функции u в точке x_0 могут быть вычислены по исходным данным задачи. Действительно, вследствие уравнения (1) и начального условия (2) имеем

$$u'(x_0) = f(x_0, u(x_0)) = f(x_0, u_0) \equiv L_1(x_0, u_0).$$

Далее,

$$u''(x) = (f(x,u))' = f_x(x,u) + f_u(x,u)u' = f_x(x,u) + f_u(x,u)f(x,u),$$

следовательно,

$$u''(x_0) = f_x(x_0, u_0) + f_u(x_0, u_0)L_1(x_0, u_0) \equiv L_2(x_0, u_0).$$

Точно так же

$$u'''(x) = (f_x(x, u) + f_u(x, u)f(x, u))' =$$
$$= f_{xx}(x, u) + 2f_{xu}(x, u)f(x, u) + f_{uu}(x, u)(f(x, u))^2 + f_u(x, u)u''(x),$$

поэтому

$$u'''(x_0) = f_{xx}(x_0, u_0) + 2f_{xu}(x_0, u_0)L_1(x_0, u_0) +$$
$$+ f_{uu}(x_0, u_0)(L_1(x_0, u_0))^2 + f_u(x_0, u_0)L_2(x_0, u_0) \equiv L_3(x_0, u_0).$$

Аналогичные выражения можно получить для производной любого порядка: $u^{(k)}(x_0) = L_k(x_0, u_0)$.

Таким образом,

$$u(x_1) = u(x_0) + hL_1(x_0, u_0) + \frac{h^2}{2!}L_2(x_0, u_0) + \ \ldots \ + \frac{h^n}{n!}L_n(x_0, u_0) +$$
$$+ \frac{h^{n+1}}{(n+1)!}u^{(n+1)}(\xi), \ \xi \in (x_0, x_1).$$

Отбрасывая в этом равенстве остаточный член формулы Тейлора и обозначая получаемое приближение к $u(x_1)$ через $y(x_1)$, можем написать

$$y(x_1) = u(x_0) + hL_1(x_0, u_0) + \frac{h^2}{2!}L_2(x_0, u_0) + \ \ldots \ + \frac{h^n}{n!}L_n(x_0, u_0).$$

Теперь естественно считать, что вообще

$$y_{i+1} = y_i + hF(x_i, y_i), \ i = 0, 1, \ldots, \quad y_0 = u_0. \tag{1.2}$$

Здесь

$$F(x_i, y_i) = L_1(x_i, y_i) + \frac{h}{2!}L_2(x_i, y_i) + \ \ldots \ + \frac{h^{(n-1)}}{n!}L_n(x_i, y_i). \tag{1.3}$$

Таким образом, построен метод вычисления приближенного решения задачи (1), (2). При реализации метода (1.2), (1.3) необходимо вычислять частные производные функции $f(x, p)$, причем с увеличением точности метода, т. е. с увеличением n, порядок требуемых производных возрастает, а формулы, по которым вычисляются выражения $L_j(x_i, y_i)$, становятся все более громоздкими. Поэтому метод,

основанный на формуле Тейлора, довольно редко используется на практике. Правда, в последнее время в связи с появлением эффективных компьютерных программ аналитических вычислений (например, Maple) отношение к этому методу меняется. Тем не менее, наиболее распространены такие приближенные методы решения задачи Коши, которые требуют для своей реализации умения вычислять лишь значения функции $f(x, p)$ по заданным значениям аргументов.

2. Методы типа Рунге — Кутта. Пусть $u(x)$ — решение задачи (1), (2). Проинтегрировав равенство (1) по отрезку $[x_0, x_1]$, получим

$$u(x_1) = u(x_0) + \int_{x_0}^{x_1} f(x, u(x))dx. \tag{2.1}$$

Дальнейшее, фактически, основано на том или ином способе приближенного вычисления участвующего здесь интеграла.

1. Используем сначала самую простую формулу — формулу левых прямоугольников. Получим

$$u(x_1) \approx u_0 + hf(x_0, u_0).$$

Это приводит к следующей расчетной формуле:

$$y_1 = y_0 + hf(x_0, y_0).$$

Вообще,

$$y_{i+1} = y_i + hf(x_i, y_i), \quad i = 0, 1, 2, \ldots, \ y_0 = u_0. \tag{2.2}$$

Метод (2.2) называется явным методом Эйлера (чаще, просто методом Эйлера) или методом ломаных[1)].

Это — самый простой вариант метода Рунге — Кутта. Метод имеет простую и полезную геометрическую интерпретацию. Из точки (x_0, u_0) выпускается прямая по касательной к интегральной кривой уравнения (1), проходящей через точку (x_0, u_0). Из точки (x_1, y_1) прямая идет уже по касательной к интегральной кривой, проходящей через точку (x_1, y_1), и так далее. В результате получается ломаная, аппроксимирующая искомую интегральную кривую. Понятно, что этот метод не может обеспечить высокой точности, если шаг сетки не слишком мал, тем не менее он применяется довольно часто.

2. Неявный метод Эйлера. Используем для приближенного вычисления интеграла формулу правых прямоугольников

$$u(x_1) \approx u_0 + hf(x_1, u_1).$$

[1)] Он совпадает с методом, основанным на формуле Тейлора, при $n = 1$.

В результате приходим к следующему методу:

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}),\ i = 0, 1, \ldots,\ y_0 = u_0.$$

Этот метод называется неявным методом Эйлера. Значение y_{i+1} не может быть явно определено по известному y_i, а разыскивается как решение уравнения

$$y_{i+1} - hf(x_{i+1}, y_{i+1}) = y_i. \tag{2.3}$$

При этом применяются приближенные (итерационные) методы. Как правило, используется метод простой итерации, состоящий в следующем. Выбирается некоторое начальное приближение $y_{i+1}^{(0)}$ к y_{i+1}. Обычно полагают $y_{i+1}^{(0)} = y_i$ или определяют $y_{i+1}^{(0)}$ при помощи явного метода Эйлера по известному y_i. Затем строят последовательность приближений

$$y_{i+1}^{(k+1)} = y_i + hf(x_{i+1}, y_{i+1}^{(k)}), \quad k = 0, 1, \ldots$$

Этот итерационный метод быстро сходится, если h достаточно мало. На практике редко выполняют больше двух-трех итераций.

3. Формула центральных прямоугольников (метод предиктор-корректор). Увеличим точность приближения интеграла. Для этого используем формулу центральных прямоугольников:

$$u(x_1) \approx u(x_0) + hf(x_{1/2}, u(x_{1/2})),$$

где $x_{1/2} = (x_0 + x_1)/2$. Значение $u(x_{1/2})$ приближенно определяется при помощи явного метода Эйлера

$$u(x_{1/2}) \approx u(x_0) + \frac{h}{2} f(x_0, u_0).$$

Расчетные формулы таковы:

$$y_{i+1} = y_i + hf(x_{i+1/2}, y_{i+1/2}),\ i = 1, 2, \ldots, \tag{2.4}$$

где

$$y_{i+1/2} = y_i + \frac{h}{2} f(x_i, y_i), \quad x_{i+1/2} = (x_i + x_{i+1})/2. \tag{2.5}$$

Все вычисления проводятся по явным формулам. Шаг (2.5) называется предиктором, шаг (2.4) — корректором, то есть сначала по формуле (2.5) мы выполняем как бы предсказание (to predict — предсказывать), а затем по формуле (2.4) уточнение (to correct — исправлять, уточнять) значения y_{i+1}. Здесь в отличие от явного метода переход от y_i к y_{i+1} требует вычисления двух значений функции f.

4. Правило трапеций. Используем для приближенного вычисления интеграла формулу трапеций:

$$u(x_1) \approx u(x_0) + \frac{h}{2}\left(f(x_0, u_0) + f(x_1, u(x_1))\right).$$

Отсюда вытекает расчетная формула:

$$y_{i+1} = y_i + \frac{h}{2}\left(f(x_i, y_i) + f(x_{i+1}, y_{i+1})\right),\ i = 0, 1, 2, \ldots,\ y_0 = u_0. \quad (2.6)$$

Это — неявный метод. По поводу его реализации можно сказать то же, что и по поводу реализации неявного метода Эйлера.

Комбинируя правило трапеций с явной формулой Эйлера, получим еще один вариант метода предиктор-корректор:

$$y_{i+1} = y_i + \frac{h}{2}\left(f(x_i, y_i) + f(x_{i+1}, \overline{y}_{i+1})\right),\ i = 0, 1, 2, \ldots,\ y_0 = u_0, \quad (2.7)$$

где

$$\overline{y}_{i+1} = y_i + hf(x_i, y_i). \quad (2.8)$$

Сначала выполняется предсказание по формуле (2.8), а затем коррекция по формуле (2.7).

5. Метод Рунге — Кутта четвертого порядка точности. В основе этого метода лежит квадратурная формула Симпсона. Используются также предсказания по формуле Эйлера. Не останавливаясь на подробном выводе, дадим окончательные расчетные формулы:

$$y_{i+1} = y_i + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right), \quad i = 0, 1, 2, \ldots,\ y_0 = u_0, \quad (2.9)$$

где

$$k_1 = hf(x_i, y_i), \quad k_2 = hf\Big(x_i + \frac{h}{2}, y_i + \frac{k_1}{2}\Big),$$

$$k_3 = hf\Big(x_i + \frac{h}{2}, y_i + \frac{k_2}{2}\Big), \quad k_4 = hf\left(x_i + h, y_i + k_3\right).$$

Этот метод (с небольшими модификациями) на практике применяется, по-видимому, наиболее часто.

3. Элементы теории одношаговых методов решения задачи Коши. Метод, основанный на формуле Тейлора, а также методы типа Рунге — Кутта могут быть представлены в общем виде

$$\frac{y_{i+1} - y_i}{h} = F(x_i, h, y_i, y_{i+1}), \quad i = 0, 1, \ldots, \quad y_0 = u_0. \quad (3.1)$$

Здесь $F(x_i, h, y_i, y_{i+1})$ — некоторая функция, выбор которой и определяет тот или иной метод. Например, для неявного метода типа Рунге — Кутта, основанного на формуле трапеций,

$$F(x_i, h, y_i, y_{i+1}) = \frac{1}{2}\left(f(x_i, y_i) + f(x_i + h, y_{i+1})\right).$$

Понятно, что значения u_i, u_{i+1} точного решения задачи Коши не удовлетворяют равенству (3.1). После их подстановки возникает невязка

$$\psi_i = \frac{u_{i+1} - u_i}{h} - F\left(x_i, h, u_i, u_{i+1}\right), \tag{3.2}$$

которая называется погрешностью аппроксимации.

Интуитивно ясно, что чем меньше погрешность аппроксимации, тем точнее метод. Ниже будет показано, что при определенных условиях это действительно так.

Исследуем поведение погрешности аппроксимации в зависимости от величины шага сетки для построенных выше методов.

1) Метод, основанный на формуле Тейлора. В этом случае и из способа построения метода сразу вытекает, что $\psi_i = O(h^n)$.

2) Явный метод Эйлера. Этот метод есть частный случай метода, основанного на формуле Тейлора, при $n = 1$.

3) Неявный метод Эйлера. В этом случае

$$\psi_i = \frac{u_{i+1} - u_i}{h} - f(x_{i+1}, u_{i+1}).$$

Используя оценку погрешности квадратурной формулы правых прямоугольников, можем написать:

$$hf(x_{i+1}, u_{i+1}) = \int_{x_i}^{x_{i+1}} f(x, u(x))\,dx + O(h^2) = \int_{x_i}^{x_{i+1}} u'(x)\,dx + O(h^2) = \\ = (u(x_{i+1}) - u(x_i)) + O(h^2),$$

следовательно, $\psi_i = O(h)$.

4) Метод предиктор-корректор, основанный на формуле трапеций. В соответствии с формулами (2.4), (2.5)

$$\psi_i = \frac{u_{i+1} - u_i}{h} - f\Big(x_{i+1/2}, u(x_i) + \frac{h}{2} f(x_i, u(x_i))\Big) = \\ = \frac{u_{i+1} - u_i}{h} - f\Big(x_{i+1/2}, u(x_i) + \frac{h}{2} u'(x_i)\Big).$$

По формуле Тейлора $u(x_i) + \dfrac{h}{2}u'(x_i) = u(x_{i+1/2}) + O(h^2)$, поэтому

$$\psi_i = \frac{u_{i+1} - u_i}{h} - f(x_{i+1/2}, u(x_{i+1/2})) + O(h^2).$$

Используя теперь оценку погрешности формулы центральных прямоугольников, получим

$$hf(x_{i+1/2}, u(x_{i+1/2})) = \int\limits_{x_i}^{x_{i+1}} f(x, u(x)\,dx + O(h^3) =$$

$$= (u(x_{i+1}) - u(x_i)) + O(h^3),$$

и, следовательно, $\psi_i = O(h^2)$.

5) Для неявного метода (2.6), основанного на формуле трапеций, и метода предиктор-корректор (2.7), (2.8) совершенно аналогично проверяется, что погрешность аппроксимации есть величина порядка h^2.

6) Отметим, наконец, что, используя формулу Тейлора и оценку погрешности формулы Симпсона, нетрудно доказать, что для метода (2.9) погрешность аппроксимации есть величина порядка h^4.

Укажем теперь общую конструкцию, включающую в себя все рассмотренные выше варианты явного метода Рунге — Кутта как частные случаи. Пусть заданы числа $\alpha_2, \alpha_3, \ldots, \alpha_q$, $p_1, p_2, \ldots, p_q$, β_{ij}, $1 \leqslant j < i \leqslant q$, где q — целое положительное число, определяющее сложность метода. По известному y_i вычислим

$$k_1(x_i, h, y_i) = k_1 = f(x_i, y_i),$$
$$k_2(x_i, h, y_i) = k_2 = f(x_i + \alpha_2 h, y_i + \beta_{21} h k_1),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$k_q(x_i, h, y_i) = k_q = f(x_i + \alpha_q h, y_i + \beta_{q1} h k_1 + \beta_{q2} h k_2 + \ldots \beta_{qq-1} h k_{q-1}),$$

затем

$$F(x_i, h, y_i) = \sum_{l=1}^{q} p_l k_l$$

и положим

$$y_{i+1} = y_i + hF(x_i, h, y_i). \tag{3.3}$$

При построении методов указанного вида параметры $\alpha_i, p_i, \beta_{ij}$ естественно выбирать так, чтобы погрешность аппроксимации ψ_i была величиной возможно более высокого порядка малости относительно шага сетки.

Приведем общий результат о сходимости одношаговых методов вида (3.3). Предположим, что задача (1), (2) решается на отрезке $x_0 \leqslant x \leqslant x_0+l$. Пусть $h = l/N$, N — целое положительное число, $x_i = x_0 + ih$, $i = 0, \ldots, N$. Будем обозначать через y приближенное решение, построенное по методу (3.3), через u — точное решение задачи (1), (2), $z_i = z(x_i) = y_i - u(x_i)$ — погрешность метода.

Теорема 3.1. *Пусть функция $f(x,p)$ непрерывно дифференцируема по p,*

$$|f_p(x,p)| \leqslant M = \text{const} \quad \forall x \in [x_0, x_0+l],\ p \in (-\infty, \infty), \tag{3.4}$$

погрешность аппроксимации метода есть величина порядка h^s.

Тогда

$$|z_i| \leqslant Ch^s, \quad C = \text{const},\ i = 0, 1, \ldots, N. \tag{3.5}$$

Доказательство. Покажем, прежде всего, что выполнение условия (3.4) обеспечивает существование такой постоянной $M_1 > 0$, что для функции $F(x,h,y)$ справедлива оценка

$$|F_y(x,h,y)| \leqslant M_1. \tag{3.6}$$

Для этого установим, что $|\partial k_r(x,h,y)/\partial y| \leqslant c_r$ где $c_r = \text{const}$, $r = 1, 2, \ldots, q$. При $r = 1$ это непосредственно вытекает из условия (3.4). При $r = 2$ имеем

$$\frac{\partial k_2(x,h,y)}{\partial y} = \frac{\partial f(x+\alpha_2 h, y+\beta_{21} h k_1(x,h,y))}{\partial y} = \\ = f_p(x+\alpha_2 h, y+\beta_{21} h k_1(x,h,y))(1+\beta_{21} h k_{1y}(x,h,y)),$$

следовательно, $|\partial k_2(x,h,y)/\partial y| \leqslant M(1+\beta_{21}hc_1) \leqslant M(1+\beta_{21}lc_1) = c_2$. Для $r > 2$ оценки проводятся аналогично. Используя теперь определение погрешности аппроксимации, для любого $i = 0, 1, \ldots, N-1$ можем написать

$$u_{i+1} = u_i + F(x_i, h, u_i) + h\psi_i. \tag{3.7}$$

Вычитая почленно из равенства (3.3) равенство (3.7), получим

$$z_{i+1} = z_i + h(F(x_i,h,y_i) - F(x_i,h,u_i)) - h\psi_i,$$

откуда вследствие формулы конечных приращений Лагранжа и оценки (3.6) имеем $|z_{i+1}| \leqslant (1+M_1h)|z_i| + h|\psi_i|$. Точно так же получаем, что $|z_i| \leqslant (1+M_1h)|z_{i-1}| + h|\psi_{i-1}|$, следовательно,

$$|z_{i+1}| \leqslant (1+M_1h)^2|z_{i-1}| + h(1+M_1h)|\psi_{i-1} + h|\psi_i|.$$

Продолжая аналогичным образом, получим:

$$|z_{i+1}| \leqslant (1 + M_1 h)^{i+1}|z_0| + h \sum_{k=0}^{i} (1 + M_1 h)^{i-k} |\psi_k|.$$

Заметим теперь, что для любого целого r, $0 \leqslant r \leqslant N$,

$$(1+M_1 h)^r = (1+M_1 h)^{\frac{x_r - x_0}{h}} = (1+M_1 h)^{\frac{1}{M_1 h} M_1 (x_r - x_0)} \leqslant (1+M_1 h)^{\frac{1}{M_1 h} M_1 l}.$$

Но, как хорошо известно,

$$(1 + v)^{\frac{1}{v}} \leqslant e \quad \text{при любом} \quad v > 0,$$

поэтому

$$(1 + M_1 h)^{\frac{1}{M_1 h} M_1 l} \leqslant e^{M_1 l},$$

и, следовательно,

$$|z_{i+1}| \leqslant e^{M_1 l}|z_0| + e^{M_1 l} \max_{0 \leqslant k \leqslant N} |\psi_k| h(i+1) \leqslant e^{M_1 l}(|z_0| + l \max_{0 \leqslant k \leqslant N} |\psi_k|).$$

Вспоминая теперь, что $z_0 = y_0 - u_0 = 0$, $|\psi_k| \leqslant ch^s$, $c = \text{const}$, получим, что для любого $i = 0, 1, \ldots, N$ справедлива оценка

$$|z_{i+1}| \leqslant e^{M_1 l} l c h^s. \ \square$$

4. Методы типа Адамса. В отличие от одношаговых методов в методах типа Адамса при вычислении каждого нового значения y_{i+1} используются не только y_i, но и значения приближенного решения в нескольких предыдущих точках сетки. Благодаря этому методы типа Адамса оказываются экономичнее одношаговых методов.

Опишем способ построения методов типа Адамса, основанный на использовании интерполяционных многочленов. Предположим, что нам известны значения решения задачи (1), (2) в точках сетки x_0, $x_1, \ldots, x_{k-1}$. Интегрируя уравнение (1) по отрезку $[x_{k-1}, x_k]$, получим

$$u(x_k) = u(x_{k-1}) + \int_{x_{k-1}}^{x_k} f(x, u(x))\, dx. \tag{4.1}$$

Проинтерполируем функцию $f(x, u(x))$ по ее значениям в точках x_0, $x_1, \ldots, x_{k-1}$, т. е. представим эту функцию в виде

$$f(x, u(x)) = \sum_{l=0}^{k-1} f(x_l, u(x_l))\varphi_l(x) + R_k(x), \tag{4.2}$$

где $\varphi_l(x),\ l = 0, 1, \ldots, k-1$ — базисные функции Лагранжа, $R_k(x)$ — остаточный член интерполяционного полинома. Подставляя выражение (4.2) в равенство (4.1), получим

$$u(x_k) = u(x_{k-1}) + \sum_{l=0}^{k-1} c_l f(x_l, u(x_l)) + \widetilde{R}_k, \tag{4.3}$$

где

$$c_l = \int_{x_{k-1}}^{x_k} \varphi_l(x)\,dx = h\widehat{c}_l, \quad \widetilde{R}_k = \int_{x_{k-1}}^{x_k} R_k(x)\,dx. \tag{4.4}$$

Здесь коэффициенты $\widehat{c}_l,\ l = 0, 1, \ldots, k-1$, зависят только от k и l (ср. с соответствующими выкладками, выполненными при построении квадратурных формул Ньютона — Котеса). Отбрасывая в равенстве (4.3) остаточный член, естественно считать, что приближенные значения решения задачи в точках $x_0,\ x_1, \ldots,\ x_k$, т. е. $y_0,\ y_1, \ldots,\ y_k$ связаны соотношением

$$y_k = y_{k-1} + h \sum_{l=0}^{k-1} \widehat{c}_l f(x_l, y_l)$$

и, вообще,

$$y_i = y_{i-1} + h \sum_{l=0}^{k-1} \widehat{c}_l f(x_{i-k+l}, y_{i-k+l}), \quad i = k, k+1, \ldots \tag{4.5}$$

Если считать значения $y_0, y_1,\ \ldots, y_{k-1}$ известными, то формулу (4.5) можно рассматривать как рекуррентную для определения всех последующих значений приближенного решения. Значения $y_1, y_2,\ \ldots, y_{k-1}$ вычисляют, обычно, при помощи метода Рунге — Кутта.

Метод (4.5) называют явным или экстраполяционным методом Адамса. Последнее название объясняется тем, что интерполяционный полином, построенный для функции $f(x, u(x))$ по узлам $x_0, x_1,\ \ldots, x_{k-1}$, экстраполируется на отрезок $[x_{k-1}, x_k]$.

Приближая функцию $f(x, u(x))$ на отрезке $[x_{k-1}, x_k]$, можно использовать интерполяционный полином, построенный по узлам $x_0, x_1,\ \ldots, x_k$, В результате, вместо (4.5) получаем соотношения

$$y_i = y_{i-1} + h \sum_{l=0}^{k} \widetilde{c}_l f(x_{i-k+l}, y_{i-k+l}), \quad i = k, k+1, \ldots, \tag{4.6}$$

определяющее интерполяционный метод Адамса.

Для отыскания y_i по известным $y_{i-1}, y_{i-2}, \ldots, y_{i-k}$ приходится решать нелинейное, вообще говоря, уравнение

$$y_i - h\widetilde{c}_k f(x_i, y_i) = y_{i-1} + h\sum_{l=0}^{k} \widetilde{c}_l f(x_{i-k+l}, y_{i-k+l}). \qquad (4.7)$$

Поэтому метод (4.6) часто называют неявным методом Адамса.

Проведем сравнение по трудоемкости методов типа Рунге — Кутта с явным методом Адамса. Ясно, что при вычислении значения y_i по методу (4.5) приходится вычислять лишь одно значение функции f, а именно $f(x_{i-1}, y_{i-1})$. Все остальные значения функции f, участвующие в формуле (4.5), уже подсчитаны на предыдущих шагах рекуррентного процесса и могут быть сохранены в памяти компьютера. При применении метода Рунге — Кутта для перехода от y_{i-1} к y_i требуется (в зависимости от порядка метода) вычислить несколько значений функции f. Таким образом, если функция f достаточно сложная, то методы типа Рунге – Кутта существенно уступают в смысле трудоемкости явному методу Адамса.

Трудоемкость неявного метода Адамса определяется количеством итераций, затрачиваемых для решения нелинейного уравнения (4.7). Обычно с этой целью применяют метод простой итерации. При достаточно малом шаге сетки h для достижения приемлемой точности чаще всего оказывается достаточным провести две-три итерации и соответственно два-три раза обратиться к процедуре вычисления функции f. Таким образом, и неявный метод Адамса, как правило, экономичнее методов типа Рунге — Кутта.

Понятие погрешности аппроксимации для методов типа Адамса вводится по аналогии с одношаговыми методами. Например, для явного метода Адамса (4.5) погрешность аппроксимации есть

$$\psi_i = \frac{u_{i+1} - u_i}{h} - \sum_{l=0}^{k-1} \widehat{c}_l f(x_{i-k+l}, u_{i-k+l}).$$

Оценки погрешностей аппроксимации методов (4.5), (4.6) нетрудно получить, используя представление остаточного члена интерполяционного полинома.

Можно показать, что если начальные условия определены с точностью порядка погрешности аппроксимации, то есть $y_i = u_i + O(h^s)$ для $i = 0, 1, \ldots, k-1$, $\psi_i = 0(h^s)$, $i = k,\ k+1, \ldots$, функция $f(x, p)$ удовлетворяет условию (3.4), то $|z_i| = O(h^s)$, $i = 0, 1, \ldots$, иными словами, метод имеет точность порядка h^s.

В заключение приведем примеры методов типа Адамса с указанием порядка погрешности аппроксимации.

1) Явные методы:

$$\frac{y_{i+1}-y_i}{h}=\frac{3}{2}f\left(x_i,y_i\right)-\frac{1}{2}f\left(x_{i-1},y_{i-1}\right),\quad s=2,$$

$$\frac{y_{i+1}-y_i}{h}=\frac{1}{12}\left(23f\left(x_i,y_i\right)-16f\left(x_{i-1},y_{i-1}\right)+5f\left(x_{i-2},y_{i-2}\right)\right),\quad s=3,$$

$$\frac{y_{i+1}-y_i}{h}=\frac{1}{24}\left(55f\left(x_i,y_i\right)-59f\left(x_{i-1},y_{i-1}\right)+\right.$$
$$\left.+37f\left(x_{i-2},y_{i-2}\right)-9f\left(x_{i-3},y_{i-3}\right)\right),\ s=4.$$

2) Неявные методы:

$$\frac{y_{i+1}-y_i}{h}=\frac{1}{2}\left(f\left(x_{i+1},y_{i+1}\right)+f\left(x_i,y_i\right)\right),\quad s=2,$$

$$\frac{y_{i+1}-y_i}{h}=\frac{1}{12}\left(5f\left(x_{i+1},y_{i+1}\right)+8f\left(x_i,y_i\right)-f\left(x_{i-1},y_{i-1}\right)\right),\quad s=3,$$

$$\frac{y_{i+1}-y_i}{h}=\frac{1}{24}\left(9f\left(x_{i+1},y_{i+1}\right)+19f\left(x_i,y_i\right)-\right.$$
$$\left.-5f\left(x_{i-1},y_{i-1}\right)+f\left(x_{i-2},y_{i-2}\right)\right),\quad s=4.$$

§ 2. Методы решения уравнений с частными производными

1. Сеточные методы решения краевых задач для обыкновенных дифференциальных уравнений. Рассматривается обыкновенное дифференциальное уравнение второго порядка

$$-(pu')'+qu=f,\quad 0<x<l, \tag{1.1}$$

с граничными условиями

$$u(0)=u(1)=0, \tag{1.2}$$

где, $p(x),\ q(x),\ f(x)$ — заданные функции, причем

$$p(x)\geqslant c_0=\text{const}>0,\ q(x)\geqslant 0. \tag{1.3}$$

Задача (1.1), (1.2) может быть, например, интерпретирована как задача о стационарном распределении температуры в стержне, на боковой поверхности которого происходит теплообмен по закону Ньютона со средой, имеющей заданную температуру. Наш интерес к этой задаче обусловлен не столько важностью соответствующих приложений, сколько тем, что задача (1.1), (1.2) может рассматриваться как хорошая модель для отработки методов численного решения более сложных двумерных и трехмерных эллиптических уравнений.

Построим на отрезке $[0, l]$ равномерную сетку

$$\omega = \{x_i = ih,\ i = 0, 1, 2, \ldots, N,\ nh = 1\}$$

с шагом h. Если $u(x)$ — дифференцируемая функция, то по определению производной

$$u'(x_i) \approx \frac{u(x_i + h) - u(x_i)}{h} = \frac{u_{i+1} - u_i}{h}$$

при достаточно малом h. Выражение, стоящее в правой части этого приближенного равенства, называется разностным отношением. Понятно, что это не единственный способ приближения производной. Точно так же

$$u'(x_i) \approx \frac{u(x_i) - u(x_i - h)}{h} = \frac{u_i - u_{i-1}}{h}.$$

Приняты следующие наименования и обозначения:

$$\frac{u_{i+1} - u_i}{h} = u_{x,i} \ \text{ — разностное отношение вперед,}$$

или правое разностное отношение,

$$\frac{u_i - u_{i-1}}{h} = u_{\overline{x},i} \ \text{ — разностное отношение назад,}$$

или левое разностное отношение.

Вторую производную $u''(x_i)$ естественно приближать выражением

$$u_{\overline{x}x,i} = (u_{\overline{x}})_{x,i} = \frac{1}{h}\left(\frac{u_{i+1} - u_i}{h} - \frac{u_i - u_{i-1}}{h}\right) = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2};$$

$u_{\overline{x}x,i}$ называется второй разделенной разностью или вторым разностным отношением.

Более сложное дифференциальное выражение $(pu')'$ будем приближать — разностным $(au_{\overline{x}})_x$, где $a(x)$ — сеточная функция, в определенном смысле близкая $p(x)$. Как именно следует выбирать $a(x)$,

будет сказано чуть позже. Таким образом, уравнению (1.1) естественно поставить в соответствие приближенное равенство

$$-(au_{\overline{x}})_x + bu \approx \varphi, \quad x \in \overset{\circ}{\omega},$$

где $\overset{\circ}{\omega} = \{x_i = ih,\ i = 1, 2, \ldots, N-1\}$ — множество внутренних точек сетки, $a(x),\ b(x),\ \varphi(x)$ — сеточные функции, вычисляемые некоторым образом по известным функциям $p(x),\ q(x),\ f(x)$. Чаще всего полагают

$$a(x) = p(x - h/2),\ b(x) = q(x),\ \varphi(x) = f(x). \tag{1.4}$$

Определим теперь сеточную функцию $y(x)$ как решение системы линейных алгебраических уравнений:

$$-(ay_{\overline{x}})_x + by = \varphi, \quad x \in \overset{\circ}{\omega}. \tag{1.5}$$

Приведем более подробную запись этой системы:

$$-\frac{a_{i+1}(y_{i+1} - y_i) - a_i(y_i - y_{i-1})}{h^2} + b_i y_i = \varphi_i, \quad i = 1, 2, \ldots, N-1,$$

или

$$A_i y_{i+1} - B_i y_i + C_i y_{i-1} = -\varphi_i, \quad i = 1, 2, \ldots, N-1, \tag{1.6}$$

где

$$A_i = a_{i+1}/h^2,\ C_i = a_i/h^2,\ B_i = b_i + (a_{i+1} + a_i)/h^2. \tag{1.7}$$

Ясно, что система (1.5) содержит уравнений на два меньше, чем неизвестных. Для того, чтобы пополнить ее, воспользуемся граничными условиями (1.2) и положим

$$y_0 = y_N = 0. \tag{1.8}$$

Совокупность уравнений (1.5), (1.8) образует полную систему линейных алгебраических уравнений относительно $y_0,\ y_1,\ \ldots, y_N$ и называется разностной схемой для задачи (1.1), (1.2). Для решения этой трехдиагональной системы уравнений чаще всего применяют метод прогонки (см. с. 16). Условия (1.3) влекут выполнение неравенств

$$|A_i| + |C_i| < |B_i|, \quad i = 1, 2, \ldots, N-1. \tag{1.9}$$

Именно они обеспечивают реализуемость метода прогонки.

Займемся теперь исследованием погрешности аппроксимации и построением функций $a(x),\ b(x),\ \varphi(x)$. Как и ранее, под погрешностью аппроксимации будем понимать невязку

$$\psi_i = -(au_{\bar{x}})_{x,i} + bu_i - \varphi_i,$$

возникающую при подстановке в уравнение (1.5) точного решения задачи (1.1), (1.2). Более подробно:

$$\psi_i = -\frac{a_{i+1}(u_{i+1} - u_i) - a_i(u_i - u_{i-1})}{h^2} + b_i y_i - \varphi_i.$$

Воспользуемся формулой Тейлора:

$$u_{i+1} = u\left(x_{i+1/2} + \frac{h}{2}\right) = u_{i+1/2} + \frac{h}{2}u'_{i+1/2} + \frac{1}{2!}\left(\frac{h}{2}\right)^2 u''_{i+1/2} + \\ + \frac{1}{3!}\left(\frac{h}{2}\right)^3 u'''_{i+1/2} + \cdots,$$

$$u_i = u\left(x_{i+1/2} - \frac{h}{2}\right) = u_{i+1/2} - \frac{h}{2}u'_{i+1/2} + \frac{1}{2!}\left(\frac{h}{2}\right)^2 u''_{i+1/2} - \\ - \frac{1}{3!}\left(\frac{h}{2}\right)^3 u'''_{i+1/2} + \cdots,$$

следовательно,

$$\frac{u_{i+1} - u_i}{h} = u'_{i+1/2} + \frac{2}{3!}\frac{h^2}{2^3}u'''_{i+1/2} + O(h^4).$$

Аналогично,

$$\frac{u_i - u_{i-1}}{h} = u'_{i-1/2} + \frac{2}{3!}\frac{h^2}{2^3}u'''_{i-1/2} + O(h^4).$$

Отсюда с очевидностью вытекает, что если выбрать a, b, φ в соответствии с (1.4), то, с одной стороны, условия (1.9) будут выполнены, а, с другой — погрешность аппроксимации будет иметь второй порядок малости относительно шага сетки:

$$\psi_i = O(h^2). \tag{1.10}$$

Ясно, конечно, что при этом необходимо, чтобы функции $u(x),\ p(x)$ имели достаточное число непрерывных производных.

Нужно отметить, что коэффициенты разностной схемы можно вычислять и другими способами. Например, часто используют соотношения

$$a_i = \frac{p_i + p_{i-1}}{2}, \text{ или } a_i = \left[\frac{1}{h} \int\limits_{x_{i-1}}^{x_i} \frac{1}{p(\xi)} d\xi \right]^{-1}.$$

При этом оценка (1.10) сохраняется.

2. Вариационные методы. Наряду с разностными методами для решения граничных задач для эллиптических уравнений широко применяются вариационные методы. Между разностными и вариационными методами существует довольно тесная связь, наиболее ярко проявляющаяся в методе конечных элементов. Вновь мы ограничимся простейшей эллиптической граничной задачей (1.1), (1.2). Рассмотрим наряду с этой задачей так называемый энергетический функционал (функционал Лагранжа)

$$F(u) = \frac{1}{2} \int\limits_0^l (pu'^2 + qu^2) dx - \int\limits_0^l f u dx. \tag{2.1}$$

Отметим, что если интерпретировать задачу (1.1), (1.2) как задачу о равновесии струны ($q(x)$ можно при этом трактовать как коэффициент жесткости упругого основания (постели), на котором находится струна), то $F(u)$ — потенциальная энергия системы струна — внешние силы.

Теорема 2.1. *Пусть функция $u(x)$ доставляет минимальное значение функционалу F на множестве функций, удовлетворяющих граничным условиям* (1.2). *Тогда $u(x)$ — решение задачи* (1.1), (1.2).

Доказательство. Рассмотрим наряду с функцией $u(x)$ функцию $u(x)+t\eta(x)$, где $\eta(x)$ удовлетворяет граничным условиям (1.2), а t — вещественное число. Ясно, что поскольку на функции $u(x)$ функционал достигает минимального значения, то

$$F(u + t\eta) \geqslant F(u), \tag{2.2}$$

значит, функция вещественного переменного

$$\varphi(t) = F(u + t\eta)$$

достигает минимального значения при $t = 0$, следовательно,

$$\varphi'(0) = 0. \tag{2.3}$$

Подсчитаем $\varphi'(0)$. Для этого запишем $F(u+t\eta)$ более подробно. Имеем:

$$F(u+t\eta) = \frac{1}{2}\int_0^l \left(pu'^2 + qu^2\right) dx - \int_0^l fu dx + t\int_0^l (pu'\eta' + qu\eta)\, dx -$$

$$-t\int_0^l f\eta dx + \frac{t^2}{2}\int_0^l \left(p\eta'^2 + q\eta^2\right),$$

следовательно,

$$\varphi'(0) = \int_0^l (pu'\eta' + qu\eta)\, dx - \int_0^l f\eta dx,$$

и условие (2.3) принимает вид:

$$\int_0^l (pu'\eta' + qu\eta)\, dx = \int_0^l f\eta dx. \tag{2.4}$$

Уравнение (2.4) часто называют вариационным уравнением, соответствующим функционалу $F(u)$. Оно представляет собой необходимое условие минимума функционала.

Покажем, что уравнение (1.1) вытекает из (2.4). Действительно, используя формулу интегрирования по частям и учитывая граничные условия для функции $\eta(x)$, получим

$$\int_0^l pu'\eta' dx = -\int_0^l (pu')'\,\eta dx + pu'\eta\Big|_0^l = -\int_0^l (pu')'\, dx,$$

то есть

$$\int_0^l \left(-(pu')' + qu - f\right)\eta dx = 0.$$

Используя теперь то, что функция η в равенстве (2.4) произвольна, покажем, что функция $\zeta(x) \equiv -(pu')' + qu - f \equiv 0$ на интервале $(0, l)$. Действительно, если предположить противное, то можно указать такую точку $x_0 \in (0, l)$, что $\zeta(x_0) \neq 0$. Для определенности можно считать, что $\zeta(x_0) > 0$. Случай противоположного знака исследуется

точно так же. Вследствие непрерывности $\zeta(x)$ найдется окрестность $(x_0-\varepsilon, x_0+\varepsilon) \subset (0,l)$, на которой функция $\zeta(x)$ сохраняет знак. Выберем теперь функцию $\eta(x)$ так, чтобы она была положительной внутри указанной окрестности и равна тождественно нулю в остальных точках интервала $(0,l)$. При таком выборе функции $\eta(x)$ интеграл в левой части равенства есть

$$\int\limits_{x_0-\varepsilon}^{x_0+\varepsilon} \zeta(x)\eta(x)dx > 0,$$

что невозможно. Остается принять, что $\zeta(x) \equiv 0$, т. е. уравнение (1.1) выполнено. $\square$

Уравнение (1.1), возникающее как необходимое условие минимума функционала $F(u)$, часто называют уравнением Эйлера.

Доказанная нами теорема позволяет заменить задачу (1.1), (1.2) задачей минимизации функционала F на множестве функций, удовлетворяющих граничным условиям (1.2). Основанные на такой замене приближенные методы решения задачи (1.1), (1.2) называют вариационными.

3. Метод Ритца. Это — исторически первый и весьма распространенный вариационный метод решения эллиптических уравнений. Опишем его, по-прежнему используя как пример задачу (1.1), (1.2).

Пусть заданы функции $\varphi_1(x)$, $\varphi_2(x)$, $\ldots$, $\varphi_n(x)$, такие, что $\varphi_i(0) = \varphi_i(l) = 0$, $i = 1, 2, \ldots, n$. Эти функции называют координатными или базисными функциями метода Ритца.

Приближенное решение задачи о минимуме функционала будем искать в виде линейной комбинации

$$u_n(x) = \sum_{i=1}^{n} c_k \varphi_i(x),$$

коэффициенты которой и подлежат определению. А именно, они находятся так, чтобы функция n вещественных переменных

$$\Phi(c_1, c_2, \ldots, c_n) = F\left(\sum_{i=1}^{n} c_i \varphi_i\right)$$

принимала минимальное значение. Запишем необходимые условия минимума:

$$\frac{\partial F\big(\sum\limits_{i=1}^{n} c_i\varphi_i\big)}{\partial c_k} = 0, \quad k = 1, 2, \ldots, n. \tag{3.1}$$

Система уравнений (3.1) называется системой уравнений метода Ритца.

Видно, что сама формулировка метода Ритца никак не привязана к конкретному виду функционала F, и, действительно, область применения этого метода чрезвычайно широка.

В рассматриваемом нами случае систему уравнений (3.1) нетрудно записать более подробно. Имеем:

$$F\left(\sum_{i=1}^{n} c_i\varphi_i\right) = \frac{1}{2}\int_0^l \left[p\left(\sum_{i=1}^{n} c_i\varphi_i'\right)^2 + q\left(\sum_{i=1}^{n} c_i\varphi_i\right)^2\right] dx - \int_0^l f\sum_{i=1}^{n} c_i\varphi_i dx.$$

Легко находится, что

$$\frac{\partial F(u_n)}{\partial c_k} = \int_0^l \left[p\sum_{i=1}^{n} c_i\varphi_i'\varphi_k' + q\sum_{i=1}^{n} c_i\varphi_i\varphi_k\right] dx - \int_0^l f\varphi_k dx.$$

Следовательно, система (3.1) принимает вид:

$$\sum_{i=1}^{n} a_{ki}c_i = b_k, \quad k = 1, 2, \ldots, n, \tag{3.2}$$

где

$$a_{ki} = \int_0^l \left[p\varphi_i'\varphi_k' + q\varphi_i\varphi_k\right] dx, \tag{3.3}$$

$$b_k = \int_0^l f\varphi_k dx. \tag{3.4}$$

Понятно, что фактическое построение системы требует умения вычислять интегралы (3.3), (3.4). На практике для этого, обычно, применяют квадратурные формулы.

Свойства системы (3.2) и построенного с ее помощью приближенного решения u_n, конечно, определяются выбором координатной системы $\varphi_1,\ \varphi_2,\ \ldots,\ \varphi_n$. Если координатная система выбрана удачно, то с увеличением n точность приближенного метода улучшается. Можно показать, что если координатная система линейно независима, то матрица системы линейных уравнений (3.2) положительно определена. Отсюда, в частности, вытекает однозначная разрешимость системы метода Ритца.

В качестве примеров координатных систем метода Ритца укажем:
а) полиномиальную систему

$$\varphi_0(x) = x(l-x),\ \varphi_1(x) = x(l-x)x,\ \varphi_2(x) = x(l-x)x^2,\ \ldots \\ \varphi_n(x) = x(l-x)x^n,\ \ldots;$$

б) тригонометрическую систему

$$\varphi_k(x) = \sin\frac{\pi k x}{l}, \quad k = 1, 2,\ \ldots, n,\ \ldots$$

При использовании указанных, или аналогичных, координатных систем матрица метода Ритца оказывается заполненной, т. е. все ее элементы отличны от нуля. Это принципиально отличает метод Ритца от разностного метода, при использовании которого матрица системы линейных уравнений — разреженная матрица. Правда, при удачном выборе координатной системы метода Ритца можно ограничиться небольшими значениями n, т. е. система будет иметь небольшие размеры и ее решение не вызывает затруднений.

4. Метод конечных элементов. В 1943 г. Р. Курантом было замечено, что при специальном выборе базисных функций метод Ритца приводит к системам линейных уравнений, по свойствам весьма близким к разностным уравнениям. В простейших случаях метод совпадает с разностным. Впоследствии (в 50-х годах) метод, предложенный Курантом, был переоткрыт инженерами, существенно развит и обобщен. В настоящее время этот метод, называемый методом конечных элементов (МКЭ), принадлежит к числу наиболее распространенных методов решения эллиптических уравнений и систем, в частности, возникающих при математическом моделировании упругих конструкций. Мы опишем простейший вариант метода конечных элементов на примере задачи (1.1), (1.2).

Построим на отрезке $[0, l]$ сетку (вообще говоря, неравномерную)

$$\omega = \{x_0 = 0 < x_1 < x_2 < \cdots < x_N = l\}.$$

Свяжем с каждой внутренней точкой сетки x_k непрерывную функцию $\varphi_k(x)$, линейную на каждом интервале (x_{i-1}, x_i), $i = 1, 2, \ldots, N$, и такую, что

$$\varphi_k(x_i) = \begin{cases} 1, & i = k, \\ 0, & i \neq k, \quad i = 0, 1, \ldots, N. \end{cases}$$

Легко написать аналитическое выражение для $\varphi_k(x)$, но в дальнейшем оно нам не потребуется. Функции $\varphi_k(x)$ принято называть базисными функциями с локальным носителем, поскольку множество

точек, где $\varphi_k(x) \neq 0$, (носитель функции $\varphi_k(x)$) — малый интервал (в рассматриваемом случае — два шага сетки), содержащий точку x_k. Функции $\varphi_k(x)$ иногда называют функциями Куранта.

Будем искать решение задачи на минимум функционала F в виде

$$y(x) = \sum_{k=1}^{N-1} c_k \varphi_k(x).$$

Нетрудно убедиться, что $y(x)$ — кусочно линейная функция, равная нулю на концах отрезка $[0, l]$, причем $c_k = y(x_k) = y_k$, $k = 1, 2, \ldots, N-1$, то есть можно написать

$$y(x) = \sum_{k=1}^{N-1} y_k \varphi_k(x).$$

Система метода Ритца, по-прежнему, имеет вид (3.2). Элементы a_{ki} матрицы этой системы, очевидно, отличны от нуля лишь при условии, что носители функций $\varphi_k(x), \varphi_i(x)$ пересекаются, то есть $|k - i| < 2$. Это означает, что матрица системы трехдиагональна, как и в случае разностного метода.

Рассмотрим самый простой частный случай. Пусть $p(x) \equiv 1$, $q(x) \equiv 0$, сетка равномерна. Элементы матрицы системы метода Ритца будут при этом равны

$$a_{ki} = \int_0^l \varphi_k'(x), \varphi_i'(x) dx.$$

Фиксируем некоторое k, $2 \leqslant k \leqslant N-2$. Ненулевыми коэффициентами в k-м уравнении системы метода Ритца будут лишь

$$a_{kk-1} = \int_0^l \varphi_k'(x), \varphi_{k-1}'(x) dx, \; a_{kk} = \int_0^l \varphi'^2_k(x) dx,$$

$$a_{kk+1} = \int_0^l \varphi_k'(x), \varphi_{k+1}'(x) dx.$$

Подсчитаем их значения. Ясно, что

$$a_{kk-1} = \int_{x_{k-1}}^{x_k} \varphi_k'(x), \varphi_{k-1}'(x) dx,$$

причем

$$\varphi_k'(x) = \frac{1}{h} \text{ при } x \in (x_{k-1}, x_k),\ \varphi_{k-1}'(x) = -\frac{1}{h} \text{ при } x \in (x_{k-1}, x_k).$$

Отметим, что для вычисления производных нет нужды выписывать аналитические выражения базисных функций. Достаточно заметить, что они линейны на рассматриваемом интервале и, следовательно, их производные совпадают с разностными отношениями, вычисленными, например, по точкам $x_{k-1},\ x_k$. Таким образом, очевидно, $a_{kk-1} = -1/h$. Точно так же $a_{kk+1} = -1/h$. Далее

$$a_{kk} = \int\limits_{x_{k-1}}^{x_{k+1}} \varphi_k'^2(x)dx = \int\limits_{x_{k-1}}^{x_k} \frac{1}{h^2}dx + \int\limits_{x_k}^{x_{k+1}} \frac{1}{h^2}dx = \frac{2}{h},$$

$$b_k = \int\limits_0^l f\varphi_k(x)dx = \int\limits_{x_{k-1}}^{x_{k+1}} f\varphi_k(x)dx.$$

Вычислим последний интеграл приближенно, полагая $f(x) \approx f(x_k)$. Тогда

$$b_k = f(x_k) \int\limits_{x_{k-1}}^{x_{k+1}} \varphi_k(x)dx = hf(x_k),$$

и, следовательно, k-е уравнение системы метода Ритца принимает вид

$$-\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = f_k,$$

то есть совпадает с разностной аппроксимацией уравнения (1.1).

5. Метод конечных элементов для эллиптических уравнений. Опишем этот метод на примере задачи Дирихле для уравнения Пуассона:

$$-\Delta u = f, \quad (x, y) \in \Omega, \tag{5.1}$$

$$u(x, y) = 0, \quad (x, y) \in \Gamma. \tag{5.2}$$

Для простоты предполагается, что область Ω — единичный квадрат: $\Omega = [0 < x, y < 1]$, Γ — граница Ω. Метод конечных элементов (МКЭ) основан на вариационной формулировке задачи (5.1), (5.2). Рассмотрим функционал

$$F(u) = \frac{1}{2} \int\limits_\Omega \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] dxdy - \int\limits_\Omega fudxdy.$$

Замечание 5.1. Если задачу (5.1), (5.2) интерпретировать как задачу о равновесии мембраны, то $F(u)$ — потенциальная энергия системы мембрана — внешние силы на перемещении u.

Можно показать, что если u — решение задачи (5.1), (5.2), то u доставляет минимальное значение функционалу F на множестве дважды непрерывных функций, удовлетворяющих граничному условию (5.2), и наоборот, функция, доставляющая минимальное значение функционалу F на множестве функций, удовлетворяющих граничному условию (5.2), — решение задачи (5.1), (5.2). При этом используются, по существу, точно такие же рассуждения, какие были нами проведены для обыкновенного дифференциального уравнения второго порядка.

Таким образом, можно строить приближенное решение, отправляясь от задачи минимизации функционала F. Чаще всего для этого используется метод Ритца: приближенное решение разыскивается в виде

$$u_n = \sum_{k=1}^{n} c_k \varphi_k(x, y),$$

где $\varphi_k(x, y)$ — некоторые заданные функции, удовлетоворяющие граничному условию (5.2) (их называют — базисными), а коэффициенты c_k подлежат определению. При этом их подбирают так, чтобы доставить функционалу $F(u_n)$ минимальное значение. Получим подробное выражение для $F(u_n)$. Имеем:

$$F(u_n) = \frac{1}{2} \int\limits_{\Omega} \left[\left(\sum_{k=1}^{n} c_k \frac{\partial \varphi_k}{\partial x} \right)^2 + \left(\sum_{k=1}^{n} c_k \frac{\partial \varphi_k}{\partial y} \right)^2 \right] dxdy - \int\limits_{\Omega} \sum_{k=1}^{n} c_k \varphi_k f dxdy =$$

$$= \frac{1}{2} \int\limits_{\Omega} \sum_{k,l=1}^{n} c_k c_l \left(\frac{\partial \varphi_k}{\partial x} \frac{\partial \varphi_l}{\partial x} + \frac{\partial \varphi_k}{\partial y} \frac{\partial \varphi_l}{\partial y} \right) dxdy - \int\limits_{\Omega} \sum_{k=1}^{n} c_k \varphi_k f dxdy =$$

$$= \frac{1}{2}\sum_{k,l=1}^{n} a_{kl}c_k c_l - \sum_{k=1}^{n} c_k f_k \equiv \Phi(c_1, c_2 \ldots, c_n),$$

где

$$a_{kl} = \int\limits_{\Omega} \left(\frac{\partial \varphi_k}{\partial x}\frac{\partial \varphi_l}{\partial x} + \frac{\partial \varphi_k}{\partial y}\frac{\partial \varphi_l}{\partial y} \right) dxdy, \quad f_k = \int\limits_{\Omega} \varphi_k f dxdy. \tag{5.3}$$

Запишем необходимые условия минимума функции Φ:

$$\frac{\partial \Phi}{\partial c_k} = \sum_{l=1}^{n} a_{kl}c_l - f_k = 0, \ k = 1, 2, \ldots, n.$$

В результате получаем систему линейных алгебраических уравнений для определения коэффициентов $c_k, \ k = 1, 2, \ldots, n$:

$$\sum_{l=1}^{n} a_{kl}c_l = f_k, \quad k = 1, 2, \ldots, n. \tag{5.4}$$

Построение системы (5.4) связано с вычислением интегралов, определяющих $a_{kl}, \ f_k$. Эта процедура оказывается более или менее трудоемкой в зависимости от выбора функций φ_k.

Классические примеры выбора базисных функций:

$$1) \quad \varphi_{kl}(x, y) = x(1-x)y(1-y)x^k y^l, \ k, l = 1, 2, \ldots, n$$

(первые множители здесь служат для удовлетворения граничным условиям),

$$2) \quad \varphi_{kl}(x, y) = \sin \pi k x \sin \pi l y, \ k, l = 1, 2, \ldots, n.$$

Отметим, что для удобства здесь принята двойная нумерация базисных функций.

Метод конечных элементов основан на использовании специальных базисных функций. При этом матрица системы метода Ритца оказывается разреженной, т. е. большинство ее элементов — нули. Аналогичное имеет место и для существенно более общих эллиптических уравнений, чем уравнение Пуассона.

Построим на области Ω квадратную сетку с шагом h. Каждую ячейку сетки разделим на два треугольника диагональю, параллельной биссектрисе первого координатного угла. Получим разбиение области Ω на треугольники, т. е. триангуляцию области.

Рассмотрим произвольную внутреннюю точку сетки $(x_k, y_l) = (kh, lh)$. Обозначим через Ω_{kl} объединение всех треугольников триангуляции с вершиной в этой точке (их — шесть). Обозначим далее через $\varphi_{kl}(x, y)$ функцию, равную нулю вне области Ω_{kl}, равную единице в точке (x_k, y_l), непрерывную на Ω_{kl} и линейную на каждом треугольнике триангуляции. Такая функция называется функцией Куранта («шапочкой Куранта»). Используем эти функции при построении приближенного решения по методу Ритца:

$$u^h(x, y) = \sum_{k,l=1}^{N-1} c_{kl}\varphi_{kl}(x, y).$$

Заметим, что

$$u^h(x_i, y_j) = \sum_{k,l=1}^{N-1} c_{kl}\varphi_{kl}(x_i, y_j) = c_{ij},$$

то есть

$$u^h(x, y) = \sum_{k,l=1}^{N-1} u^h(x_k, y_l)\varphi_{kl}(x, y).$$

Это значит, что коэффициентами в разложении приближенного решения служат значения приближенного решения в точках сетки. Кроме того отметим, что приближенное решение — линейная функция на каждом треугольнике триангуляции.

Построим систему уравнений для определения значений $u^h(x_k, y_l)$. Для этого нужно только подсчитать коэффициенты и правую часть в системе метода Ритца, а именно,

$$a_{klk'l'} = \int\limits_\Omega \left(\frac{\partial\varphi_{kl}}{\partial x}\frac{\partial\varphi_{k'l'}}{\partial x} + \frac{\partial\varphi_{kl}}{\partial y}\frac{\partial\varphi_{k'l'}}{\partial y} \right) dxdy, \quad f_{kl} = \int\limits_\Omega \varphi_{kl} f dxdy.$$

Понятно, что $a_{klk'l'}$ может быть не нулем только в том случае, когда области Ω_{kl}, $\Omega_{k'l'}$ имеют общие точки. Для заданных k, l таких областей $\Omega_{k'l'}$ шесть, а именно:

$$\Omega_{k-1,l},\ \Omega_{k-1,l-1},\ \Omega_{k,l-1},\ \Omega_{k+1,l},\ \Omega_{k+1,l+1},\ \Omega_{k,l+1}.$$

При вычислении конкретных значений коэффициентов $a_{klk'l'}$ потребуются значения производных функции φ_{kl}. Пронумеруем треугольники, принадлежащие Ω_{kl}. Поскольку функция φ_{kl} линейна на каждом из указанных треугольников, то ее первые производные постоянны

на каждом из этих треугольников, причем производная по x равна разностному отношению, вычисленному по любым двум точкам, лежащим на прямой, параллельной оси x; аналогично вычисляется производная по y.

Сведем результаты этих очевидных вычислений в таблицу (n — номер треугольника).

n	$\frac{\partial\varphi_{kl}}{\partial x}$	$\frac{\partial\varphi_{kl}}{\partial y}$
1	0	$-1/h$
2	$-1/h$	$1/h$
3	$-1/h$	0
4	0	$-1/h$
5	$1/h$	$-1/h$
6	$1/h$	0

Теперь совершенно очевидно, что

$$a_{k,l,k-1,l-1} = \int\limits_{\Delta_6 \bigcap \Delta_1} \left(\frac{\partial\varphi_{kl}}{\partial x}\frac{\partial\varphi_{k-1,l-1}}{\partial x} + \frac{\partial\varphi_{kl}}{\partial y}\frac{\partial\varphi_{k-1,l-1}}{\partial y} \right) dxdy =$$

$$= \frac{h^2}{2}\left(\frac{1}{h}\cdot 0 + 0\cdot\frac{-1}{h}\right) + \frac{h^2}{2}\left(0\cdot\frac{-1}{h} + \frac{-1}{h}\cdot 0\right) = 0.$$

Поясним, что $h^2/2$ — площадь треугольника Δ_k; значения производных функции $\varphi_{k-1,l-1}$ на Δ_6 совпадают со значениями производных $\varphi_{k,l}$ на Δ_4; значения производных функции $\varphi_{k-1,l-1}$ на Δ_1 совпадают со значениями производных $\varphi_{k,l}$ на Δ_3.

Точно так же получаем $a_{k,l,k+1,l+1} = 0$,

$$a_{k,l,k,l-1} = \int\limits_{\Delta_1 \bigcap \Delta_2} \left(\frac{\partial\varphi_{kl}}{\partial x}\frac{\partial\varphi_{k,l-1}}{\partial x} + \frac{\partial\varphi_{kl}}{\partial y}\frac{\partial\varphi_{k,l-1}}{\partial y} \right) dxdy = -1,$$

$a_{k,l,k-,l} = -1,\ a_{k,l,k,l+1} = -1,\ a_{k,l,k+1,l} = -1,$

$$a_{k,l,k,l} = \int\limits_{\Omega_{kl}} \left(\frac{\partial\varphi_{kl}}{\partial x}\frac{\partial\varphi_{k,l}}{\partial x} + \frac{\partial\varphi_{kl}}{\partial y}\frac{\partial\varphi_{k,l}}{\partial y} \right) dxdy = 4.$$

Это означает, что уравнение с номером k, l системы Ритца записывается в виде

$$\frac{4u^h_{k,l} - u^h_{k-1,l} - u^h_{k+1,l} - u^h_{k,l-1} - u^h_{k,l+1}}{h^2} = g^h_{k,l}, \tag{5.5}$$

где $$g^h_{k,l} = \frac{1}{h^2}\int\limits_{\Omega_{kl}} f(x,y)\varphi_{kl}(x,y)dxdy.$$

Заметим, что при малом h справедливо приближенное равенство

$$\int\limits_{\Omega_{kl}} f(x,y)\varphi_{kl}(x,y)dxdy \approx f(x_k,y_l)\int\limits_{\Omega_{kl}} \varphi_{kl}(x,y)dxdy = f(x_k,y_l)h^2.$$

Поясним, что для вычисления интеграла от $\varphi_{kl}(x,y)$ по Ω_{kl} достаточно заметить, что он равен объему пирамиды с основанием Ω_{kl}, площадь которого равна $6h^2/2 = 3h^2$, и высотой, равной единице. Таким образом, уравнение (5.5) приближенно представляется в виде

$$\frac{4u^h_{k,l} - u^h_{k-1,l} - u^h_{k+1,l} - u^h_{k,l-1} - u^h_{k,l+1}}{h^2} = f_{k,l}.$$

Очевидно, точно такое же уравнение получается при замене в уравнении Пуассона (5.1) производных разделенными разностями

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{u(x_{k-1,l}) - 2u(x_{k,l}) + u(x_{k+1,l})}{h^2},$$

$$\frac{\partial^2 u}{\partial y^2} \approx \frac{u(x_{k,l-1}) - 2u(x_{k,l}) + u(x_{k,l+1})}{h^2}.$$

Записывая уравнения (5.5) во внутренних точках сетки, т. е. при $k,l = 1,2,\ldots,N-1$, и присоединяя к ним граничные условия, соответствующие (5.2):

$$\begin{array}{l} u^h(0,y_l) = 0, \quad u^h(1,y_l) = 0, \quad l = 1,2,\ldots,N-1, \\ u^h(x_k,0) = 0, \quad u^h(x_k,1) = 0, \quad k = 1,2,\ldots,N-1, \end{array} \tag{5.6}$$

получим полную систему линейных алгебраических уравнений для отыскания приближенного решения в точках сетки.

Замечание 5.2. Совершенно аналогично строится система метода конечных уравнений для произвольных областей: сначала область аппроксимируется многоугольником, затем производят триангуляцию, т. е. разбиение многоугольника на достаточно малые треугольники, далее определяются базисные функции φ_{kl}, при этом, конечно, области Ω_{kl} могут иметь достаточно сложное строение.

6. Итерационные методы решения сеточных уравнений. Для решения системы уравнений МКЭ (сеточных уравнений) при умеренных значениях числа неизвестных чаще всего применяют прямые методы, обычно метод Гаусса и его модификации. При очень большом числе неизвестных более эффективными становятся итерационные методы. Опишем простейшие из них, используя в качестве примера систему (5.5).

6.1. Метод Якоби (метод простой итерации). Для сокращения записей используем обозначения $y_{kl} = u^h(x_k, y_l), b_{kl} = g^h_{kl}$. Зададимся некоторым начальным приближением $y^0_{kl},\ k, l = 0, 1, \ldots, N$, удовлетворяющим граничным условиям (5.6). Чаще всего полагают $y^0_{kl} = 0$. Отправляясь от y^0_{kl}, строим последовательность сеточных функций $y^s_{kl},\ s = 1, 2, \ldots$, по реккурентным формулам:

$$y^{s+1}_{kl} = \frac{1}{4}\left(y^s_{k+1,l} + y^s_{k-1,l} + y^s_{k,l+1} + y^s_{k,l-1} + h^2 b_{kl}\right), \qquad k, l = 1, 2, \ldots, N-1, \quad (6.1)$$

$$\begin{aligned} y^{s+1}_{0l} &= 0, \quad y^{s+1}_{Nl} = 0, \quad l = 1, 2, \ldots, N-1, \\ y^{s+1}_{k0} &= 0, \quad y^{s+1}_{kN} = 0, \quad k = 1, 2, \ldots, N-1, \end{aligned}$$

$s = 0, 1, 2, \ldots$

В ходе вычислений контролируется невязка

$$\frac{4y^s_{k,l} - y^s_{k-1,l} - y^s_{k+1,l} - y^s_{k,l-1} - y^s_{k,l+1}}{h^2} - b_{k,l}.$$

Вычисления останавливают, если при некотором s значения невязки при всех $k, l = 1, 2, \ldots, N-1$ по модулю не превышают заданной величины ε, определяющей точность вычислений.

6.2. Метод Зейделя. При вычислениях по формуле (6.1) порядок перебора точек сетки, по крайней мере теоретически, может быть произвольным. Условимся теперь перебирать точки в порядке возрастания индексов k, l и модифицируем формулы (6.1) следующим образом:

$$y^{s+1}_{kl} = \frac{1}{4}\left(y^s_{k+1,l} + y^{s+1}_{k-1,l} + y^s_{k,l+1} + y^{s+1}_{k,l-1} + h^2 b_{kl}\right), \qquad k, l = 1, 2, \ldots, N-1, \quad (6.2)$$

$$\begin{aligned} y^{s+1}_{0l} &= 0, \quad y^{s+1}_{Nl} = 0, \quad l = 1, 2, \ldots, N-1, \\ y^{s+1}_{k0} &= 0, \quad y^{s+1}_{kN} = 0, \quad k = 1, 2, \ldots, N-1, \end{aligned}$$

$s = 0, 1, 2, \ldots$ Этот метод сходится не медленнее, чем метод Якоби (на практике, как правило, быстрее).

6.3. Метод верхней релаксации. Значительного ускорения сходимости метода Зейделя можно добиться, несколько модифицируя формулы (6.2) введением итерационного параметра:

$$y_{kl}^{s+1} = \\ = (1-\omega)y_{kl}^{s} + \omega\tfrac{1}{4}\left(y_{k+1,l}^{s} + y_{k-1,l}^{s+1} + y_{k,l+1}^{s} + y_{k,l-1}^{s+1} + h^2 b_{kl}\right), \tag{6.3}$$

$k, l = 1, 2, \ldots, N-1,$

$$y_{0l}^{s+1} = 0, \quad y_{Nl}^{s+1} = 0, \quad l = 1, 2, \ldots, N-1,$$
$$y_{k0}^{s+1} = 0, \quad y_{kN}^{s+1} = 0, \quad k = 1, 2, \ldots, N-1,$$

$s = 0, 1, 2, \ldots$ Понятно, что при $\omega = 1$ метод (6.3) превращается в метод Зейделя. Оптимальное значение параметра ω зависит от сетки. На практике, обычно, ω полагают близким к $1,8$.

7. Разностные методы решения нестационарных задач математической физики. Рассмотрим первую краевую задачу для уравнения теплопроводности стержня:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(p\frac{\partial u}{\partial x}\right) - qu + f(x,t), \quad 0 < x < l, \ t > 0, \tag{7.1}$$

$$u(x,0) = u_0(x), \quad 0 < x < l, \tag{7.2}$$

$$u(0,t) = u(l,t) = 0, \quad t > 0. \tag{7.3}$$

7.1. Явная схема. Построим на области $\Omega = [0 < x < l, \ t > 0]$ сетку ω с шагами h по оси x, τ — по оси t. Значение сеточной функции $y = y(x_i, t_j)$ в точке $(x_i, t_j) = (ih, tj)$, $0 \leqslant i \leqslant N, Nh = l, \ j \geqslant 0$ обозначим через y_i^j. Как и ранее, пусть

$$y_{x,i}^j = \frac{y(x_{i+1}, t_j) - y(x_i, t_j)}{h} \text{ — разность вперед по } x,$$

$$y_{\overline{x},i}^j = \frac{y(x_i, t_j) - y(x_{i-1}, t_j)}{h} \text{ — разность назад по } x.$$

Положим еще

$$y_{t,i}^j = \frac{y(x_i, t_{j+1}) - y(x_i, t_j)}{h} \text{ — разность вперед по } t.$$

Действуя теперь по аналогии с обыкновенным дифференциальным уравнением, заменим в точках сетки (x_i, t_j) уравнение (7.1) — разностным

$$y_{t,i}^j = (a y_{\overline{x}}^j)_{x_i} - b y_i^j + \varphi_i^j, \quad 1 \leqslant i \leqslant N-1, \ j \geqslant 0. \tag{7.4}$$

В результате, получим систему линейных алгебраических уравнений. Эта система, однако, неполная. Нужно еще учесть начальные и граничные условия:

$$y(x_i, 0) = u_0(x_i), \quad 0 \leqslant i \leqslant N, \tag{7.5}$$

$$y(0, t_j) = y(x_N, t_j) = 0, \quad j \geqslant 0. \tag{7.6}$$

Совокупность линейных алгебраических уравнений (7.4)–(7.6) называется разностной схемой для задачи (7.1)–(7.3). Это — явная схема. Вычисления по указанной схеме проводятся так. Значения y_i^0 при всех i, $0 \leqslant i \leqslant N$, известны, так как задано начальное условие (7.5). Предположим, что уже найдены значения y_i^j, $0 \leqslant i \leqslant N$. Совокупность точек сетки (x_i, t_j) при фиксированном j называется j-м временным слоем. Напишем уравнение (7.4) во всех внутренних точках j-го временного слоя чуть более подробно:

$$\frac{y_i^{j+1} - y_i^j}{\tau} = (a y_{\overline{x}}^j)_{x_i} - b y_i^j + \varphi_i^j, \quad 1 \leqslant i \leqslant N-1,$$

или

$$y_i^{j+1} = y_i^j + \tau \left[(a y_{\overline{x}}^j)_{x_i} - b y_i^j + \varphi_i^j \right], \quad 1 \leqslant i \leqslant N-1. \tag{7.7}$$

В правой части последнего уравнения только известные величины. Поэтому значения y во всех точках $j+1$ временного слоя вычисляются по явным формулам.

7.2. Неявная схема. По аналогии с (7.4) можно написать разностное уравнение

$$y_{t,i}^j = (a y_{\overline{x}}^{j+1})_{x_i} - b y_i^{j+1} + \varphi_i^j, \quad 1 \leqslant i \leqslant N-1, \ j \geqslant 0. \tag{7.8}$$

Присоединяя начальные и граничные условия

$$y(x_i, 0) = u_0(x_i), \quad 0 \leqslant i \leqslant N, \tag{7.9}$$

$$y(0, t_j) = y(x_N, t_j) = 0, \quad j \geqslant 0, \tag{7.10}$$

вновь получим полную систему линейных алгебраических уравнений. Эта система называется чисто неявной (чаще, просто, неявной) разностной схемой для уравнения теплопроводности. Построение решения этой системы кардинально отличается от случая явной схемы. Запишем разностное уравнение (7.8) подробнее в точках j-го временного слоя:

$$\frac{y_i^{j+1} - y_i^j}{\tau} = (a y_{\overline{x}}^{j+1})_{x_i} - b y_i^{j+1} + \varphi_i^j, \quad 1 \leqslant i \leqslant N-1,$$

или

$$y_i^{j+1} - \tau(ay_{\overline{x}}^{j+1})_{x_i} + \tau b y_i^{j+1} = y_i^j + \tau\varphi_i^j, \quad 1 \leqslant i \leqslant N-1.$$

В силу граничных условий

$$y_0^{j+1} = y_N^{j+1} = 0.$$

Более подробная запись полученной системы уравнений такова:

$$A_i y_{i+1}^{j+1} - B_i y_i^{j+1} + C_i y_{i-1}^{j+1} = -F_i^j, \tag{7.11}$$
$$y_0^{j+1} = y_N^{j+1} = 0,$$

где

$$A_i = \frac{\tau a_{i+1}}{h^2},\ C_i = \frac{\tau a_i}{h^2},\ B_i = 1 + A_i + B_i + b_i,\ F_i = (y_i^j + \tau\varphi_i^j). \tag{7.12}$$

Систему (7.11) можно решать методом прогонки. Условие реализуемости метода прогонки выполнено, так как, очевидно,

$$A_i + C_i < B_i, \quad 1 \leqslant i \leqslant N-1.$$

7.3. Схемы с весами. По аналогии с явной и неявной схемами можно написать целое семейство схем (это так называемые схемы с весами):

$$y_{t,i}^j = (ay_{\overline{x}}^{(\sigma)})_{x_i} - by_i^{(\sigma)} + \varphi_i^j, \quad 1 \leqslant i \leqslant N-1,\ j \geqslant 0. \tag{7.13}$$

$$y(x_i, 0) = u_0(x_i), \quad 0 \leqslant i \leqslant N, \tag{7.14}$$

$$y(0, t_j) = y(x_N, t_j) = 0, \quad j \geqslant 0, \tag{7.15}$$

где $y_i^{(\sigma)} = \sigma y_i^{j+1} + (1-\sigma)y_i^j$, σ — число, называемое весом слоя.

При $\sigma = 0$ как частный случай мы получаем явную схему, а при $\sigma = 1$ — чисто неявную схему. Все схемы при $\sigma \neq 0$ неявные. Их решения строятся при помощи метода прогонки. Среди схем с весами особо выделяется схема с $\sigma = 1/2$. Эта схема называется симметричной схемой с весами, или схемой Кранка — Никольсон. Если провести вполне уместную аналогию с обыкновенными дифференциальными уравнениями, то явная схема аналогична явному методу Эйлера, неявная схема — неявному методу Эйлера, а схема Кранка — Никольсон — методу трапеций.

Погрешность аппроксимации схемы с весами. Основной характеристикой схемы является погрешность аппроксимации, т. е. сеточная функция

$$\psi_i^j = u_{t,i}^j - (au_{\overline{x}}^{(\sigma)})_{x_i} + bu^{(\sigma)} - \varphi_i^j,$$

где u — точное решение задачи (7.1)–(7.3). Понятно, что она зависит от σ и способа выбора функций a, b, φ. Покажем, что если a, b выбраны в соответствии с (1.4), а $\varphi_i^j = f(x_i, t_j + \tau/2)$, то

$$\psi_i^j = \begin{cases} O(h^2 + \tau), & \sigma \neq 1/2, \\ O(h^2 + \tau^2), & \sigma = 1/2, \end{cases} \tag{7.16}$$

т. е. среди всех схем с весами наилучшую оценку погрешности аппроксимации имеет схема Кранка — Никольсона.

Используя тождество

$$u^{(\sigma)} = \sigma u_i^{j+1} + (1-\sigma)u_i^j = \frac{u_i^{j+1} + u_i^j}{2} - \left(\sigma - \frac{1}{2}\right)\tau \frac{u_i^{j+1} - u_i^j}{\tau} = \\ = \frac{u_i^{j+1} + u_i^j}{2} - \left(\sigma - \frac{1}{2}\right)\tau u_{t,i}^j.$$

представим погрешность аппроксимации в виде $\psi_i^j = \psi_i^{j(0)} + \psi_i^{j(1)}$, где

$$\psi_i^{j(0)} = u_{t,i}^j - (a u_{\overline{x}}^{(1/2)})_{x_i} + b u_i^{(1/2)} - \varphi_i^j,$$

$$\psi_i^{j(1)} = \tau\left(\sigma - \frac{1}{2}\right)\left[(a u_{t\overline{x}})_x - b u_t\right].$$

Очевидно,

$$\psi_i^{j(1)} = \begin{cases} O(\tau), & \sigma \neq 1/2, \\ 0, & \sigma = 1/2. \end{cases}$$

Оценим $\psi_i^{j(0)}$. По формуле Тейлора имеем:

$$u_i^{j+1} = u_i^{j+1/2} + \frac{\tau}{2}\left(\frac{\partial u}{\partial t}\right)_i^{j+1/2} + \frac{1}{2}\left(\frac{\tau}{2}\right)^2\left(\frac{\partial^2 u}{\partial t^2}\right)_i^{j+1/2} + \cdots, \tag{7.17}$$

$$u_i^{j} = u_i^{j+1/2} - \frac{\tau}{2}\left(\frac{\partial u}{\partial t}\right)_i^{j+1/2} + \frac{1}{2}\left(\frac{\tau}{2}\right)^2\left(\frac{\partial^2 u}{\partial t^2}\right)_i^{j+1/2} - \cdots, \tag{7.18}$$

то есть

$$u^{(1/2)} = u^{j+1/2} + O(\tau^2).$$

Используя еще раз разложения (7.17)–(7.18), нетрудно проверить, что

$$u_t = \frac{u^{j+1} - u^j}{\tau} = \left(\frac{\partial u}{\partial t}\right)_i^{j+1/2} + O(\tau^2). \tag{7.19}$$

Повторяя теперь выкладки, проведенные на с. 89, получим:

$$-(au_{\overline{x}}^{(1/2)})_{x_i} + bu_i^{(1/2)} = -(au_{\overline{x}}^{j+1/2})_{x_i} + bu_i^{j+1/2} + O(\tau^2) =$$
$$= -(pu')'(t_{j+1/2}, x_i) + bu'(t_{j+1/2}, x_i) + O(h^2) + O(\tau^2),$$

откуда, очевидно, вытекает, что $\psi_i^{j(0)} = O(h^2 + \tau^2)$. Таким образом, оценка (7.16) доказана.

Малость погрешности аппроксимации еще не гарантирует близости точного и приближенного решения. Нужно еще, чтобы разностная схема была устойчива. Исследование устойчивости разностных схем для уравнения теплопроводности проведем лишь в частных случаях. Будем предполагать, что $q(x) \equiv 0$ (это влечет выполнения условия $b(x) \equiv 0$), и рассмотрим только явную ($\sigma = 0$) и чисто неявную ($\sigma = 1$) разностные схемы.

1. Исследование устойчивости явной разностной схемы. Воспользуемся равенством (7.7). Запишем его более подробно:

$$y_i^{j+1} = \left(1 - \frac{\tau}{h^2}(a_{i+1} + a_i)\right) y_i^j + \frac{\tau}{h^2} a_{i+1} y_{i+1}^j + \frac{\tau}{h^2} a_i y_{i-1}^j + \tau\varphi_i^j,$$

$i = 1, 2, \ldots, N-1$, откуда

$$|y_i^{j+1}| \leqslant \left|1 - \frac{\tau}{h^2}(a_{i+1} + a_i)\right| |y_i^j| + \frac{\tau}{h^2} a_{i+1} |y_{i+1}^j| +$$
$$+ \frac{\tau}{h^2} a_i |y_{i-1}^j| + \tau|\varphi_i^j|, \quad i = 1, 2, \ldots, N-1. \quad (7.20)$$

Предположим, что

$$1 - \frac{\tau}{h^2}(a_{i+1} + a_i) \geqslant 0, \quad i = 1, 2, \ldots, N-1.$$

Это условие, очевидно, будет выполнено, если

$$\tau \leqslant \frac{h^2}{2 \max\limits_{0 \leqslant x \leqslant l} p(x)}. \quad (7.21)$$

В этом случае знак модуля в первом множителе первого слагаемого справа в (7.20) можно убрать, и, следовательно,

$$|y_i^{j+1}| \leqslant \max_{0 \leqslant i \leqslant N} |y_i^j| + \tau \max_{1 \leqslant i \leqslant N-1} |\varphi_i^{j+1}|. \quad (7.22)$$

Выражение в правой части (7.22) не зависит от i, поэтому

$$\max_{0 \leqslant i \leqslant N} |y_i^{j+1}| \leqslant \max_{0 \leqslant i \leqslant N} |y_i^j| + \tau \max_{1 \leqslant i \leqslant N-1} |\varphi_i^{j+1}|. \quad (7.23)$$

Неравенство (7.23) выполнено для любого $j \geqslant 0$, следовательно,

$$\max_{0\leqslant i\leqslant N} |y_i^j| \leqslant \max_{0\leqslant i\leqslant N} |y_i^0| + \tau \sum_{k=0}^{j-1} \max_{1\leqslant i\leqslant N-1} |\varphi_i^k|.$$

Усиливая последнее неравенство, можно написать:

$$\max_{0\leqslant i\leqslant N} |y_i^j| \leqslant \max_{0\leqslant i\leqslant N} |y_i^0| + t_{j-1} \max_{0\leqslant k\leqslant j-1} \max_{1\leqslant i\leqslant N-1} |\varphi_i^k|. \tag{7.24}$$

Неравенство (7.24) означает устойчивость явной разностной схемы (на конечном отрезке времени). Отсюда сразу вытекает сходимость разностной схемы. Действительно, если $z = y - u$ — погрешность разностной схемы, то, подставляя $y = z + u$ в уравнения (7.4)—(7.6), получим разностную схему для погрешности.

$$z_t = (az_{\overline{x}})_x + \psi,$$

$$z_i^0 = 0,$$

$$z_0^j = z_N^j = 0,$$

откуда при выполнении условия (7.21) сразу получим

$$\max_{0\leqslant i\leqslant N} |z_i^j| \leqslant t_{j-1} \max_{0\leqslant k\leqslant j-1} \max_{1\leqslant i\leqslant N-1} |\psi_i^k| = O(\tau + h^2).$$

2. Исследование устойчивости неявной схемы. Обратимся к уравнению (7.11). Вследствие конечности числа точек сетки на временном слое найдется такой номер $i_0,\ 1 \leqslant i_0 \leqslant N-1$, что

$$|y_{i_0}^{j+1}| = \max_{0\leqslant i\leqslant N} |y_i^{j+1}|. \tag{7.25}$$

Запишем уравнение (7.11) при $i = i_0$. При этом слева оставим только слагаемое, содержащее $y_{i_0}^{j+1}$, а остальные — перенесем в левую часть:

$$B_{i_0} y_{i_0}^{j+1} = A_{i_0} y_{i_0+1}^{j+1} + C_{i_0} y_{i_0-1}^{j+1} + \left(y_{i_0}^j + \tau\varphi_{i_0}^j\right),$$

откуда

$$B_{i_0}|y_{i_0}^{j+1}| \leqslant A_{i_0}|y_{i_0+1}^{j+1}| + C_{i_0}|y_{i_0-1}^{j+1}| + |y_{i_0}^j| + \tau|\varphi_{i_0}^j|,$$

следовательно (см. (7.25)),

$$B_{i_0} \max_{0\leqslant i\leqslant N} |y_i^{j+1}| \leqslant (A_{i_0} + C_{i_0}) \max_{0\leqslant i\leqslant N} |y_i^{j+1}| + \max_{0\leqslant i\leqslant N} |y_i^j| + \tau \max_{0\leqslant i\leqslant N} |\varphi_i^j|.$$

Используя теперь равенство $B_{i_0} - A_{i_0} - C_{i_0} = 1 + b_{i_0}$ (см. (7.12)) и условие $b_i \geqslant 0$, получим

$$\max_{0\leqslant i\leqslant N} |y_i^{j+1}| \leqslant \max_{0\leqslant i\leqslant N} |y_i^j| + \tau \max_{0\leqslant i\leqslant N} |\varphi_i^j|,$$

откуда точно так же, как и в случае явной разностной схемы вытекает неравенство устойчивости неявной разностной схемы и ее сходимость со скоростью $O(\tau + h^2)$.

1. При отсутствии внешних источников тепла ($f(x,t) \equiv 0$) неравенство (7.3) принимает вид

$$\max_{0\leqslant i\leqslant N} |y_i^{j+1}| \leqslant \max_{0\leqslant i\leqslant N} |y_i^j|.$$

Его можно трактовать как принцип максимума для разностной схемы: температура стержня при отсутствии источников тепла не возрастает с ростом времени.

2. Устойчивость явной разностной схемы была нами доказана лишь при условии, что шаги сетки удовлетворяют неравенству (7.21). Это условие на шаги сетки, как можно проверить, является необходимым. При его нарушении решение явной разностной схемы быстро «разбалтывается». Условие (7.21) на практике оказывается особенно обременительным, если коэффициент p быстро меняется. Тогда шаг по времени приходится брать слишком маленьким, возможно существенно меньшим, чем это диктуется соображениями точности. Это ведет к неоправданному увеличению вычислительной работы. Неявная схема устойчива, как мы показали, при любых шагах сетки (абсолютно устойчива). При ее использовании шаги сетки можно выбирать лишь из соображений точности.

3. Применяемая нами методика (основанная на дискретном принципе максимума) не может быть использована для анализа устойчивости схем с весами при $\sigma \neq 0,\ \sigma \neq 1$. Другими, более сложными методами можно показать, что все схемы при $\sigma \geqslant 1/2$ абсолютно устойчивы. В частности, абсолютно устойчива и схема Кранка — Никольсона. Все схемы при $\sigma < 1/2$ условно устойчивы, то есть устойчивы, если шаг по времени достаточно мал по сравнению с h^2.

Глава 4

Практикум по численным методам

§ 1. Системы линейных уравнений

Решить систему линейных алгебраических уравнений

$$\begin{aligned}
(a_1 + a_2 + h^2 g_1) y_1 - a_2 y_2 &= f_1 h^2, \\
&\dots\dots\dots\dots\dots\dots \\
-a_i y_{i-1} + (a_i + a_{i+1} + h^2 g_i) y_i - a_{i+1} y_{i+1} &= f_i h^2, \\
&\dots\dots\dots\dots\dots\dots \\
(a_{n-1} + a_n + h^2 g_{n-1}) y_{n-1} - a_{n-1} y_{n-2} &= f_{n-1} h^2.
\end{aligned}$$

Здесь $a_i = p(ih)$, $g_i = q(ih)$, $f_i = f(ih)$, $f(x) = -(p(x)u'(x))' + q(x)u(x)$, $h = 1/n$, p, q, u — заданные функции.

Данную систему решить методом прогонки и итерационными методами:

1) Якоби,

2) Зейделя,

3) релаксации,

4) наискорейшего спуска.

Во всех итерационных методах вычисления продолжать до выполнения условия

$$\max_{1 \le i \le n-1} | r_i^k | \le \varepsilon,$$

r — вектор невязки, ε — заданное число.

1. Исходные данные. Во всех вариантах $n = 40$, $\varepsilon = h^3$. $u(x) = x^\alpha (1-x)^\beta$, $p(x) = 1 + x^\gamma$, $g(x) = x + 1$.

2. Отчет должен содержать:

- постановку задачи и исходные данные;
- описание методов решения и расчетные формулы;
- проверку условий устойчивости метода прогонки и условий сходимости итерационных методов;

- таблицы и графики значений y_i и y_i^k с указанием числа итераций, потребовавшихся для достижения заданной точности;
- определение оптимального параметра ω для метода релаксации (графики зависимости числа итераций от ω);
- графики убывания погрешностей в зависимости от числа итераций и применяемого метода на одном рисунке;
- листинг программы.

3. Варианты заданий.

1. $\alpha = 1,\ \beta = 1,\ \gamma = 1.$
2. $\alpha = 2,\ \beta = 1,\ \gamma = 1.$
3. $\alpha = 1,\ \beta = 2,\ \gamma = 1.$
4. $\alpha = 1,\ \beta = 1,\ \gamma = 2.$
5. $\alpha = 2, \beta = 2,\ \gamma = 2.$
6. $\alpha = 3,\ \beta = 1,\ \gamma = 1.$
7. $\alpha = 3,\ \beta = 2,\ \gamma = 1.$
8. $\alpha = 3,\ \beta = 1,\ \gamma = 2.$
9. $\alpha = 3, \beta = 2,\ \gamma = 2.$
10. $\alpha = 4,\ \beta = 1,\ \gamma = 1.$

§ 2. Нелинейные уравнения

На отрезке $[a, b]$ задана функция вида

$$f(x) = \sum_{n=0}^{\infty} a_n(x).$$

1. Построить таблицу обратной к $f(x)$ функции, $F = f^{-1}$

$$\begin{array}{ccccc} F_0 & F_1 & F_2 & \dots & F_n \\ z_0 & z_1 & z_2 & \dots & z_n \end{array}$$

решая уравнения

$$f(z) = F_i, \quad F_i = f(x_0) + i \cdot \frac{f(x_n) - f(x_0)}{n}, \quad i = 0, 1, \dots, n,$$

здесь $x_i = a + i \cdot h, \quad h = \dfrac{b-a}{n}$.

Нелинейные уравнения решить итерационными методами:

1. касательных,
2. хорд,
3. секущих.

Во всех итерационных методах в качестве начального приближения к точке z_i взять x_i и вычисления продолжать до выполнения условия

$$| r^k | \le \varepsilon,$$

r — невязка, ε — заданное число.

2. Отчет должен содержать:

- постановку задачи и исходные данные;
- описание методов решения;
- графики функций $f(x)$ и $f^{-1}(x)$ на одном рисунке;
- листинг программы.

3. Варианты заданий.

1. $\sum\limits_{n=0}^{\infty}(-1)^n \dfrac{(x^2/4)^n}{(n!)^2}, \ a = 0, \ b = 3.$
2. $\dfrac{2}{\sqrt{\pi}} \sum\limits_{n=0}^{\infty}(-1)^n \dfrac{x^{2n+1}}{(n!)(2n+1)}, \ a = 0, \ b = 2.$
3. $\sum\limits_{n=0}^{\infty}(-1)^n \dfrac{x^{2n+1}}{(2n+1)(2n+1)!}, \quad a = 0, \ b = 4.$

4. $\sum\limits_{n=1}^{\infty}(-1)^n\dfrac{x^{2n}}{2n\,(2n)!},\ \ a=0.4,\ b=4.$

5. $\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{(\pi/2)^{2n}\,x^{4n+1}}{(2n)!\,(4n+1)},\ \ a=0,\ b=1.5.$

6. $\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{(\pi/2)^{2n+1}x^{4n+3}}{(2n+1)!(4n+1)},\ \ a=0,\ b=1.2.$

7. $\sum\limits_{n=1}^{\infty}(-1)^n\dfrac{(x-1)^n}{n^2},\ \ a=0,\ b=2.$

8. $\dfrac{x}{2}\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{(x^2/4)^n}{n!\,(n+1)!},\ \ a=0,\ b=3.$

9. $\left(\dfrac{x}{2}\right)^2\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{(x^2/4)^n}{n!\,(n+2)!},\ \ a=0,\ b=4.$

10. $\left(\dfrac{x}{2}\right)^3\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{(x^2/4)^n}{n!\,(n+3)!},\ \ a=2,\ b=6.$

11. $\left(\dfrac{x}{2}\right)^4\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{(x^2/4)^n}{n!\,(n+4)!},\ \ a=3,\ b=7.$

4. Замечание. При вычислении ряда $\sum\limits_{n=0}^{\infty}a_n(x)$ учесть, что каждый последующий член ряда a_{n+1} получается из предыдущего члена a_n умножением на величину q_n, то есть $a_{n+1}=a_nq_n$. Это позволит избежать переполнения при вычислении факториалов.

§ 3. Интерполирование функций

На отрезке $[a, b]$ задана функция вида

$$f(x) = \sum_{n=0}^{\infty} a_n(x).$$

1. Вычислить значения данной функции и ее призводной с помощью интерполяционного полинома Лагранжа $L_n(x)$. В качестве узлов интерполяции взять:

1) равномерно распределенные точки на отрезке $[a, b]$,

2) чебышевский набор узлов на отрезке $[a, b]$.

При табулировании функции вычислять ряд с точностью до 10^{-6}.

2. Вычислить погрешность интерполирования

$$\varepsilon_1(x) = |f(x) - L_n(x)|,\ \varepsilon_{1n} = \max_{x\in(a,b)} \varepsilon_1(x),$$

$$\varepsilon_2(x) = |f'(x) - L'_n(x)|,\ \varepsilon_{2n} = \max_{x\in(a,b)} \varepsilon_2(x).$$

3. Исследовать зависимость погрешности ε_{in} от числа узлов интерполяции.

4. Отчет должен содержать:

- постановку задачи и исходные данные;
- описание методов решения;
- графики функций $f(x),\ L_n(x),\ f'(x),\ L'_n(x),\ \varepsilon_{in}(x)$;
- графики зависимости ε_{in} от числа узлов интерполяции;
- листинг программы.

5. Варианты заданий.

1. $\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{(x^2/4)^n}{(n!)^2},\ a = 0,\ b = 3.$

2. $\dfrac{2}{\sqrt{\pi}}\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{x^{2n+1}}{(n!)(2n+1)},\ a = 0,\ b = 2.$

3. $\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{x^{2n+1}}{(2n+1)(2n+1)!},\ \ a=0,\ b=4.$

4. $\sum\limits_{n=1}^{\infty}(-1)^n\dfrac{x^{2n}}{2n\,(2n)!},\ \ a=0.4,\ b=4.$

5. $\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{(\pi/2)^{2n}\,x^{4n+1}}{(2n)!\,(4n+1)},\ \ a=0,\ b=1.5.$

6. $\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{(\pi/2)^{2n+1}x^{4n+3}}{(2n+1)!(4n+1)},\ a=0,\ b=1.2.$

7. $\sum\limits_{n=1}^{\infty}(-1)^n\dfrac{(x-1)^n}{n^2},\ \ a=0,\ b=2.$

8 $\dfrac{x}{2}\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{(x^2/4)^n}{n!\,(n+1)!},\ \ a=0,\ b=3.$

9. $\left(\dfrac{x}{2}\right)^2\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{(x^2/4)^n}{n!\,(n+2)!},\ \ a=0,\ b=4.$

10. $\left(\dfrac{x}{2}\right)^3\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{x^2/4)^n}{n!\,(n+3)!},\ \ a=2,\ b=6.$

11. $\left(\dfrac{x}{2}\right)^4\sum\limits_{n=0}^{\infty}(-1)^n\dfrac{(x^2/4)^n}{n!\,(n+4)!},\ \ a=3,\ b=7.$

6. Замечание. При вычислении ряда $\sum\limits_{n=0}^{\infty}a_n(x)$ учесть, что каждый последующий член ряда a_{n+1}получается из предыдущего члена a_n умножением на величину q_n, то есть $a_{n+1}=a_nq_n$. Это позволит избежать переполнения при вычислении факториалов.

§ 4. Численное интегрирование

Вычислить значения интегралов вида

$$I(x) = \int_a^x f(t)dt, \quad I(x) = \int_a^b f(x,t)dt$$

в точках $x_i = a + i \cdot h,\ i = 0, 1, \ldots, n$, где $h = (b-a)/n$, используя составные квадратурные формулы:

1. левых прямоугольников,
2. центральных прямоугольников,
3. трапеции,
5. Симпсона,
6. Гаусса с двумя узлами.

Интеграл вычислить с точностью $\varepsilon = 10^{-6}$. Точность вычисления интеграла определяется сравнением результатов при различном числе разбиения отрезка интегрирования. Именно, точность ε считается достигнутой, если

$$|S^N(f) - S^{2N}(f)| \le \varepsilon,$$

здесь S^N— значение составной квадратурной формулы при разбиении отрезка интегрирования на N частей.

1. Отчет должен содержать:

- постановку задачи и исходные данные;
- описание методов решения и расчетные формулы;
- таблицы значений интегралов с указанием числа разбиений, потребовавшихся для достижения заданной точности;
- листинг программы.

2. Варианты заданий.

1. $\dfrac{1}{\pi}\displaystyle\int_0^{\pi} \cos(x\cos t)\,dt,\ \ a = 0,\ \ b = 3.$

2. $\dfrac{2}{\sqrt{\pi}} \int\limits_0^x e^{-t^2}\, dt,\ a = 0, b = 2.$

3. $\int\limits_0^x \dfrac{\sin t}{t}\, dt,\ a = 0,\ b = 4.$

4. $\int\limits_0^x \dfrac{\cos t - 1}{t}\, dt,\ a = 0.4,\ b = 4.$

5. $\int\limits_0^x \cos(\dfrac{\pi t^2}{2})\, dt,\ a = 0,\ b = 1.5.$

6. $\int\limits_0^x \sin(\dfrac{\pi t^2}{2})\, dt,\ a = 0,\ b = 1.2.$

7. $-\int\limits_0^x \dfrac{\ln t}{1-t}\, dt, a = 0, b = 2.$

8. $\dfrac{1}{\pi} \int\limits_0^\pi \cos(x \sin t - t)\, dt, a = 0, b = 3.$

9. $\dfrac{1}{\pi} \int\limits_0^\pi \cos(x \sin t - 2t)\, dt, a = 0, b = 4.$

10. $\dfrac{1}{\pi} \int\limits_0^\pi \cos(x \sin t - 3t)\, dt, a = 2, b = 6.$

11. $\dfrac{1}{\pi} \int\limits_0^\pi \cos(x \sin t - 4t)\, dt, a = 3, b = 7.$

§ 5. Задача Коши для системы обыкновенных дифференциальных уравнений

Решить задачу Коши для системы из двух дифференциальных уравнений первого порядка вида

$$y' = f(t, y), \quad y(0) = y_0, \quad y(t) \in R^2,$$

на отрезке $[a, b]$, используя методы Рунге-Кутта с постоянным шагом h:

1. 2-го порядка точности

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + h, y_n + hk_1), \\ y_{n+1} &= y_n + h(k_1 + k_2)/2; \end{aligned}$$

2. 3-го порядка точности

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + h/3, y_n + h/3k_1), \\ k_3 &= f(t_n + 2/3h, y_n + 2/3hk_2), \\ y_{n+1} &= y_n + h(k_1 + 3k_3)/4; \end{aligned}$$

3. 4-го порядка точности

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + h/4, y_n + h/4k_1), \\ k_3 &= f(t_n + h/2, y_n + h/2k_2), \\ k_4 &= f(t_n + h, y_n + hk_1 - 2hk_2 + 2hk_3), \\ y_{n+1} &= y_n + h(k_1 + 4k_3 + k_4)/6; \end{aligned}$$

4. 5-го порядка точности

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + h/3, y_n + hk_1/3), \\ k_3 &= f(t_n + h/3, y_n + hk_1/6 + hk_2/6), \\ k_4 &= f(t_n + h/2, y_n + h/8k_1 + 3/8hk_3), \\ k_5 &= f(t_n + h, y_n + h/2k_1 - 3/2hk_3 + 2hk_4), \\ y_{n+1} &= y_n + h(k_1 + 4k_4 + k_5)/6. \end{aligned}$$

1. Отчет должен содержать:

- постановку задачи и исходные данные;

- описание методов решения;
- графики зависимости максимальной погрешности решения e и e/h^k от выбранного шага h для каждого метода (k – порядок точности метода);
- сделать сравнительный анализ методов.

2. Варианты заданий.

1.
$$y_1' = -y_2 + y_1(y_1^2 + y_2^2 - 1),$$
$$y_2' = y_1 + y_2(y_1^2 + y_2^2 - 1),$$
$a = 0,\ b = 5$, точное решение:
$$y_1 = \cos(t)/(1 + e^{2t})^{1/2}, \quad y_2 = \sin(t)/(1 + e^{2t})^{1/2}.$$

2.
$$y_1' = -\alpha y_1 - \beta y_2 + (\alpha + \beta - 1)e^{-t},$$
$$y_2' = \beta y_1 - \alpha y_2 + (\alpha + \beta - 1)e^{-t},$$
$a = 0,\ b = 4$, точное решение:
$$y_1 = y_2 = e^{-t}, \quad \alpha = 2, \quad \beta = 3.$$

3.
$$y_1' = y_2, \quad y_2' = 2y_1^2(1 - 4t^2 y_1),$$
$a = 0,\ b = 5$, точное решение: (проверьте!)
$$y_1 = 1/(1 + t^2), \quad y_2 = -2t/(1 + t^2)^2.$$

4.
$$y_1' = -\sin(t)/(1 + e^{2t})^{1/2} + y_1(y_1^2 + y_2^2 - 1),$$
$$y_2' = \cos(t)/(1 + e^{2t})^{1/2} + y_2(y_1^2 + y_2^2 - 1),$$
$a = 0, b = 5$, точное решение:
$$y_1 = \cos(t)/(1 + e^{2t})^{1/2}, \quad y_2 = \sin(t)/(1 + e^{2t})^{1/2}.$$

5.
$$y_1' = -\sin(t)/(1 + e^{2t})^{1/2} + y_1(y_1^2 + y_2^2 - 1),$$
$$y_2' = \cos(t)/(1 + e^{2t})^{1/2} + y_2(y_1^2 + y_2^2 - 1),$$
$a = 0, b = 5$, точное решение:
$$y_1 = \cos(t)/(1 + e^{2t})^{1/2}, \quad y_2 = \sin(t)/(1 + e^{2t})^{1/2}.$$

6.
$$y_1' = y_1/(2+2t) - 2ty_2, \quad y_2' = y_2/(2+2t) + 2ty_1,$$
$a = 0,\ b = 2$, точное решение :
$$y_1 = \cos(t^2)\sqrt{1+t}, \quad y_2 = \sin(t^2)\sqrt{1+t}.$$

7.
$$y_1' = -y_2 + t^2 + 6t + 1,$$
$$y_2' = y_1 - 3t^2 + 3t + 1,$$
$a = 0,\ b = 3$, точное решение:
$$y_1 = 3t^2 - t - 1 + \cos(t) + \sin(t), \quad y_2 = t^2 + 2 - \cos(t) + \sin(t).$$

8.
$$y_1' = y_2, \quad y_2' = 2y_1^2(1 - 4t^2y_1),$$
$a = 0,\ b = 5$, точное решение:
$$y_1 = 1/(1+t^2), \quad y_2 = -2t/(1+t^2)^2.$$

9.
$$y_1' = -\sin(t)/(1+e^{2t})^{1/2} + y_1(y_1^2 + y_2^2 - 1),$$
$$y_2' = \cos(t)/(1+e^{2t})^{1/2} + y_2(y_1^2 + y_2^2 - 1),$$
$a = 0, b = 5$, точное решение:
$$y_1 = \cos(t)/(1+e^{2t})^{1/2}, \quad y_2 = \sin(t)/(1+e^{2t})^{1/2}.$$

10.
$$y_1' = -\sin(t)/(1+e^{2t})^{1/2} + y_1(y_1^2 + y_2^2 - 1),$$
$$y_2' = \cos(t)/(1+e^{2t})^{1/2} + y_2(y_1^2 + y_2^2 - 1),$$
$a = 0,\ b = 5$, точное решение:
$$y_1 = \cos(t)/(1+e^{2t})^{1/2}, \quad y_2 = \sin(t)/(1+e^{2t})^{1/2}.$$

Литература

1. **Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.** Численные методы. — Москва: БИНОМ. Лаб. знаний, 2007.

2. **Самарский А.А., Гулин А.В.** Численные методы. — М.: Наука, 1989.

3. **Справочник** по специальным функциям с формулами, графиками и математическими таблицами/Под ред. М. Абрамовица и И. Стиган. — М.: Наука, 1979.