Оглавление

Глава 1. ЧИСЛЕННОЕ РЕШЕНИЕ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ СИСТЕМ (СЛАУ)

В этой главе рассматривается одна из самых важных задач линейной алгебры – решение систем линейных алгебраических уравнений, в которых число уравнений равно числу неизвестных:

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

или в сокращенной записи:

$$\sum_{j=1}^{n} a_{ij}x_j = f_i,\ i = 1,2,\ldots,n.$$

Коэффициенты $a_{i,j}$ при неизвестных x_j образуют матрицу системы (1)

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}. \qquad (2)$$

Всюду на протяжении этой главы мы будем считать определитель матрицы отличным от нуля

$$\Delta = \det A \neq 0. \qquad (3)$$

В этом случае система (1) называется невырожденной. Решение невырожденной системы всегда существует и является единственным. Обсудим методы фактического построения этого решения.

§1. Прямые методы решения СЛАУ.

Прямыми называются методы, которые позволяют получить точное решение невырожденной системы (1) за конечное число операций.

1.1. Формулы Крамера

Формулы Крамера представляют компоненты x_j решения системы (1) в виде отношения двух определителей:

$$x_j = \Delta_j / \Delta,\ j = 1,2,\ldots,n, \qquad (4)$$

где

$$\Delta_j = \det A_j,\ j = 1,2,\ldots,n. \qquad (5)$$

Здесь матрица A_j получается из матрицы A заменой ее j-го столбца столбцом правых частей системы (1)

$$A_j = \begin{bmatrix} a_{11} & \ldots & a_{1,j-1} & f_1 & a_{1,j+1} & \ldots & a_{1n} \\ a_{21} & \ldots & a_{2,j-1} & f_2 & a_{2,j+1} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & \ldots & a_{n,j-1} & f_n & a_{n,j+1} & \ldots & a_{nn} \end{bmatrix} \qquad (6)$$

С теоретической точки зрения формулы Крамера (4) дают исчерпывающее решение проблемы. Чтобы найти решение системы (1), нужно подсчитать $n+1$ определитель. Это можно сделать за конечное число арифметических операций. Однако с точки зрения практики важное значение имеет фактическое число необходимых операций. Здесь нас и поджидает главная трудность. Определитель n-ого порядка – это $n!$ слагаемых, каждое из которых является произведением n чисел. Таким образом, для его вычисления нужно выполнить $(n-1)n!$ умножений и $n!$ сложений – всего $Q_n = n \cdot n!$ арифметических операций. Оценим это число. При $n \gg 1$ число $n!$ можно подсчитать с помощью асимптотической формулы Стирлинга:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n, \text{ так что } Q_n \approx \sqrt{2\pi}\cdot n^{\frac{3}{2}}\left(\frac{n}{e}\right)^n.$$

При умеренном значении $n = 20$ эта формула дает астрономическое число:

$$Q_{20} \approx 5 \cdot 10^{19}.$$

Компьютеру, производительность которого составляет m операций/сек, для вычисления определителя двадцатого порядка понадобится время

$$T_{20} \approx (5 \cdot 10^{19}/m) \text{ сек.}$$

В частности, при $m = 10^{10}$ операций/сек получим

$$T_{20} \approx 5 \cdot 10^9 \text{ сек.} \approx 170 \text{ лет.}$$

Даже увеличение производительности компьютера на два, три порядка не спасает положения.

Такие результаты получены при $n = 20$, в то время, как в современных прикладных задачах приходится решать системы с $n = 10^6$ и более уравнений. Из проведенного анализа ясно, что рассчитывать решение СЛАУ по формулам Крамера с вычислением определителей «в лоб» невозможно, т. е. практическая ценность этих формул невелика.

1.2. Метод Гаусса.

Блестящий конструктивный выход из критической ситуации, описанной выше, дает метод Гаусса. Этот метод удобно условно разделить на два этапа. На первом этапе (прямой ход) система (1) приводится к треугольному виду. Затем на втором этапе (обратный ход) осуществляется последовательное отыскание неизвестных $x_1, \ldots, x_n$ из этой треугольной системы.

Перейдем к подробному описанию метода Гаусса. Не ограничивая общности, будем считать, что коэффициент a_{11}, который называют ведущим элементом первого шага, отличен от нуля (в случае $a_{11} = 0$ поменяем местами уравнения с номерами 1 и i,

при котором $a_{i1} \neq 0$; поскольку система предполагается невырожденной, то такой номер i заведомо найдется).

Разделим все члены первого уравнения на a_{11} и введем в качестве новых коэффициентов $c_{1j}, i = 2,\ldots n$ и правой части y_1 отношения

$$c_{12} = \frac{a_{12}}{a_{11}},\quad c_{13} = \frac{a_{13}}{a_{11}},\quad \ldots\quad c_{1n} = \frac{a_{1n}}{a_{11}},\ y_1 = \frac{f_1}{a_{11}}. \qquad (7)$$

Вычтем из каждого i-го уравнения системы $(i = 2,\ldots,n)$ первое уравнение умноженное на a_{i1}. Проделав это, мы исключим неизвестное x_1 из всех уравнений, кроме первого. Преобразованная таким образом система (1) примет эквивалентный вид:

$$\begin{aligned} &x_1 + c_{12}x_2 + c_{13}x_3 + \ldots + c_{1n}x_n = y_1 \\ &a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = f_2^{(1)} \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n = f_n^{(1)} \end{aligned} \qquad (8)$$

Значения новых коэффициентов и правых частей системы (8) вычисляются по формулам:

$$a_{ij}^{(1)} = a_{ij} - a_{i1}\frac{a_{1j}}{a_{11}},\ f_i^{(1)} = f_i - a_{i1}\frac{f_1}{a_{11}}. \qquad (9)$$

Естественно выделить из (8) «укороченную» систему, содержащую $n-1$ уравнение

$$\begin{aligned} &a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = f_2^{(1)} \\ &a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + \ldots + a_{3n}^{(1)}x_n = f_3^{(1)} \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n = f_n^{(1)}. \end{aligned}$$

Продолжая далее процесс исключения, после $(n-1)$ шага редуцируем исходную систему к виду:

$$\begin{aligned} &x_1 + c_{12}x_2 + c_{13}x_3 + \ldots + c_{1n}x_n = y_1 \\ &x_2 + c_{23}x_3 + \ldots + c_{2n}x_n = y_2 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &x_{n-1} + c_{n-1,n}x_n = y_{n-1} \\ &x_n = y_n \end{aligned} \qquad (10)$$

или в матричной форме

$$C\mathbf{x} = \mathbf{y},$$

где матрица C является верхней треугольной матрицей с единицами на главной диагонали

$$C = \begin{bmatrix} 1 & c_{12} & c_{13} & \ldots & c_{1n} \\ 0 & 1 & c_{23} & \ldots & c_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}. \qquad (11)$$

Построение системы (10) завершает прямой ход метода Гаусса.

Обратный ход состоит в последовательном определении неизвестных из системы (10) в обратном порядке:

$$\begin{aligned} x_n &= y_n \\ x_{n-1} &= y_{n-1} - c_{n-1,n}x_n \\ x_{n-2} &= y_{n-2} - c_{n-2,n-1}x_{n-1} - c_{n-2,n}x_n \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_1 &= y_1 - c_{12}x_2 - c_{13}x_3 - \ldots - c_{1n}x_n \end{aligned} \qquad (12)$$

Подсчитаем число арифметических операций, которое требуется выполнить при решении СЛАУ по методу Гаусса. Первый шаг прямого хода, согласно формулам (7) и (9), требует n делений и $n(n-1)$ сложений и умножений. Мы учитываем деления отдельно, поскольку для компьютера, как и для человека, это более сложная операция. Переходя последовательно от n к $(n-1)$, потом от $(n-1)$ к $(n-2)$ и т.д. подсчитаем общее число арифметических операций на стадии прямого хода. Оно включает делений

$$Q_1 = n + (n-1) + \cdots + 1 = \frac{1}{2}n(n+1), \qquad (13)$$

сложений и умножений

$$Q_2 = n(n-1) + (n-1)(n-2) + \cdots + 2 \cdot 1 = \frac{1}{3}n(n^2-1). \qquad (14)$$

Обратный ход, согласно формулам (12), вообще не требует деления, а необходимое число сложений и умножений подсчитывается по формуле

$$Q_3 = 1 + 2 + \cdots + (n-1) = \frac{1}{2}n(n-1). \qquad (15)$$

Сравнивая (13) и (14) с (15), мы видим, что обратный ход существенно проще прямого. Сумма (14) и (15) дает общее число сложений и умножений, необходимое для решения СЛАУ по методу Гаусса:

$$Q = Q_2 + Q_3 = \frac{1}{3}n(n-1)\left(n + \frac{5}{2}\right) = \frac{1}{3}n^3 + O(n^2). \qquad (16)$$

Оно не идет ни в какое сравнение с числом $n \cdot n!$, которое требуют формулы Крамера при прямом вычислении определителей.

Описанная выше процедура решения системы (1) методом Гаусса может оказаться неустойчивой по отношению к случайным ошибкам, которые неизбежны при компьютерных расчетах в результате округления чисел из-за конечной длины машинного слова. Действительно, предположим, что в процессе приведения системы (1) к треугольному виду (10) у матрицы C (11) образовались большие по модулю

элементы: $|c_{i,j}| > 1$ и даже $|c_{i,j}| \gg 1$. Тогда при вычислении неизвестных по формулам (12) во время обратного хода умножение найденных с ошибками округления чисел x_i на большие по модулю элементы матрицы C приведет к увеличению этих ошибок. Наоборот, если матрица C оказалась такой, что все ее элементы удовлетворяют условию

$$|c_{i,j}| \leq 1, \qquad (17)$$

то роль ошибок округления в процессе вычислений будет нивелироваться.

Опишем, как можно добиться выполнения условия (17). Приступая к первому шагу прямого хода метода Гаусса, рассмотрим элементы $a_{1,j}$ первой строки матрицы A и найдем среди них элемент наибольший по модулю. Пусть он имеет номер j_1. Поменяем в системе (1) первый столбец и столбец с номером j_1 местами, изменив соответствующим образом нумерацию неизвестных. В результате такой процедуры наибольший по модулю элемент первой строки станет ведущим элементом первого шага $a_{1,1}$. Благодаря этому элементы $c_{1,j}$ первой строки матрицы C, которые рассчитываются по формулам (7), будут удовлетворять неравенству (17).

Процедуру выделения наибольшего по модулю элемента в очередной строке и превращения его в ведущий элемент нужно затем повторять во время каждого шага прямого хода метода Гаусса. В этом случае все элементы $c_{i,j}$ треугольной матрицы C (11) будут удовлетворять неравенствам (17), обеспечивая устойчивость метода по отношению к ошибкам округления, Такой способ коррекции называется выбором ведущего элемента по строке.

Поясним важность специального выбора ведущего элемента в каждой строке во время прямого хода метода Гаусса на простом примере. Рассмотрим систему трех уравнений с тремя неизвестными

$$\begin{aligned} 1.2357x_1 + 2.1742x_2 - 5.4834x_3 &= -2.0735 \\ 3.4873x_1 + 6.1365x_2 - 4.7483x_3 &= 4.8755 \\ 6.0696x_1 - 6.2163x_2 - 4.6921x_3 &= -4.8388. \end{aligned} \qquad (18)$$

Легко проверить, что ее решение имеет вид

$$x_1 = x_2 = x_3 = 1. \qquad (19)$$

Решим систему (18) с помощью метода Гаусса, не обращая внимание на величины элементов матрицы. Все результаты расчетов условимся представлять в виде чисел с плавающей запятой с пятью значащими цифрами. Тогда после прямого хода получим систему треугольного вида:

$$\begin{aligned} x_1 + 1.7595x_2 - 4.4375x_3 &= -1.6780 \\ x_2 + 15324x_3 &= 15324 \\ x_3 &= 0.99992. \end{aligned} \qquad (20)$$

Значение $x_3 = 0.99992$ выглядит вполне приемлемым. Однако для двух других неизвестных мы получим следующие значения: $x_2 = 2$, $x_1 = -0.75990$. Причина случившегося заключается в потере точности при вычислении x_2 из-за больших

значений коэффициента c_{23} и правой части y_2 треугольной системы (20), которые вычислены с ошибками вследствие отбрасывания "лишних" значащих цифр.

Теперь воспользуемся процедурой выбора главного элемента по строке. Для этого в данном случае достаточно поменять местами первый и третий столбцы матрицы системы. В результате система примет вид:

$$\begin{aligned} -5.4834x_3 + 2.1742x_2 + 1.2357x_1 &= -2.0735 \\ -4.7483x_3 + 6.1365x_2 + 3.4873x_1 &= 4.8755 \\ -4.6921x_3 - 6.2163x_2 + 6.0696x_1 &= -4.8388 \end{aligned} \qquad (21)$$

При такой ее записи ведущим элементом первого шага становится число -5.4834. Оно является наибольшим по модулю элементом первой строки системы (21). Теперь применение метода Гаусса приводит к следующей системе с треугольной матрицей:

$$\begin{aligned} x_3 - 0.39651x_2 - 0.22535x_1 &= 0.37814 \\ x_2 + 0.56827x_1 &= 1.5682 \\ x_1 &= 0.99995 \end{aligned} \qquad (22)$$

Все ее элементы удовлетворяют неравенству (17). Осуществляя обратный ход, получим решение системы:

$$x_1 = 0.99995,\ x_2 = 0.99996,\ x_3 = 0.99999. \qquad (23)$$

Полученные значения неизвестных x_i хорошо согласуются с ответом (19) в рамках принятой точности вычислений.

Нетрудно предвидеть, что при бесконтрольном применении метода Гаусса для решения больших систем ($n \gg 1$) возможностей для потери точности становится еще больше, в то время как выполнение процедуры выбора ведущих элементов по строкам снимает эту проблему.

В заключение отметим, что первый этап метода Гаусса может быть использован для вычисления определителя матрицы A. Прямой ход метода Гаусса основан на многократном выполнении операции сложения одной из строк матрицы с другой строкой, взятой с некоторым множителем, что не меняет определителя. Следует лишь учесть, что при делении ведущей строки на ее диагональный элемент определитель также делится на этот элемент. Кроме того, иногда приходится переставлять столбцы при выборе главного элемента по строке. Поскольку определитель приведенной треугольной системы (матрицы C) всегда равен единице, то определитель Δ исходной системы равен

$$\Delta = \det A = (-1)^k a_{11}a_{22}^{(1)} \cdots a_{nn}^{(n-1)},$$

где k – число перестановок столбцов в процессе редукции матрицы A к треугольной матрице C.

1.3. Системы с диагональным преобладанием.

Определение.

Назовем систему (1) системой с диагональным преобладанием по строке, если элементы матрицы A (2) удовлетворяют неравенствам:

$$|a_{i,i}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|,\ 1 \leq i \leq n \qquad (24)$$

Неравенства (24) означают, что в каждой строке матрицы A диагональный элемент выделен: его модуль больше суммы модулей всех остальных элементов той же строки.

Теорема

Система с диагональным преобладанием всегда разрешима и притом единственным образом.

Рассмотрим соответствующую однородную систему:

$$\sum_{j=1}^{n} a_{i,j}x_j = 0,\ 1 \leq j \leq n \qquad (25)$$

Предположим, что она имеет нетривиальное решение $\overline{x}_j$. Пусть наибольшая по модулю компонента этого решения соответствует индексу $j = k$, т. е.

$$|\overline{x}_k| > 0,\ |\overline{x}_k| \geq |\overline{x}_j|,\ 1 \leq j \leq n. \qquad (26)$$

Запишем k-ое уравнение системы (25) в виде

$$a_{k,k}\overline{x}_k = -\sum_{\substack{j=1 \\ j \neq k}}^{n} a_{k,j}\overline{x}_j$$

и возьмем модуль от обеих частей этого равенства. В результате получим:

$$|a_{k,k}||\overline{x}_k| \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{k,j}||\overline{x}_j| \leq |\overline{x}_k|\sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{k,j}|. \qquad (27)$$

Сокращая неравенство (27) на множитель $|\overline{x}_k|$, который, согласно (26), не равен нулю, придем к противоречию с неравенством (24), выражающим диагональное преобладание. Полученное противоречие позволяет последовательно высказать три утверждения:

1. Однородная система (25) с диагональным преобладанием имеет только тривиальное решение.
2. Определитель матрицы A с диагональным преобладанием не равен нулю.
3. Неоднородная система (1) с диагональным преобладанием всегда разрешима и притом единственным образом.

Последнее из них означает, что доказательство теоремы завершено.

1.4. Системы с трехдиагональной матрицей. Метод прогонки.

При решении многих задач приходится иметь дело с системами линейных уравнений вида:

$$A_i x_{i-1} + C_i x_i + B_i x_{i+1} = F_i,\ i = 1,\ldots,n-1, \qquad (28)$$

$$x_0 = q_0,\ x_n = q_n, \qquad (29)$$

где коэффициенты A_i, C_i, B_i, правые части F_i $(i=1,\dots,n-1)$ известны вместе с числами q_0 и q_n. Дополнительные соотношения (29) часто называют краевыми условиями для системы (28). Во многих случаях они могут иметь более сложный вид. Например:

$$x_0 = p_0 x_1 + q_0;\ x_n = p_n x_{n-1} + q_n,$$

где p_0, q_0, p_n, q_n - заданные числа. Однако, чтобы не усложнять изложение, мы ограничимся простейшей формой дополнительных условий (29).

Пользуясь тем, что значения x_0 и x_n заданы, перепишем систему (28) в виде:

$$\begin{aligned} C_1 x_1 + B_1 x_2 &= F_1 - A_1 q_0 \\ A_2 x_1 + C_2 x_2 + B_2 x_3 &= F_2 \\ &\vdots \\ A_{n-1} x_{n-2} + C_{n-1} x_{n-1} &= F_{n-1} - B_{n-1} q_n \end{aligned} \tag{30}$$

Матрица этой системы имеет трёхдиагональную структуру:

$$\begin{bmatrix} C_1 & B_1 & 0 & 0 & \cdots & 0 & 0 \\ A_2 & C_2 & B_2 & 0 & \cdots & 0 & 0 \\ 0 & A_3 & C_3 & B_3 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \dots & A_{n-1} & C_{n-1} \end{bmatrix} \tag{31}$$

Это существенно упрощает решение системы (28) благодаря специальному методу, получившему название метода прогонки.

Метод основан на предположении, что искомые неизвестные x_i и x_{i+1} связаны рекуррентным соотношением

$$x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1},\ 0 \le i \le n-1. \tag{32}$$

Здесь величины α_{i+1}, β_{i+1}, получившие название прогоночных коэффициентов, подлежат определению, исходя из условий задачи (28), (29). Фактически такая процедура означает замену прямого определения неизвестных x_i задачей определения прогоночных коэффициентов с последующим расчетом по ним величин x_i.

Для реализации описанной программы выразим с помощью соотношения (32) x_{i-1} через x_{i+1}:

$$x_{i-1} = \alpha_i x_i + \beta_i = \alpha_i \alpha_{i+1} x_{i+1} + \alpha_i \beta_{i+1} + \beta_i$$

и подставим x_{i-1} и x_i, выраженные через x_{i+1}, в исходные уравнения (28). В результате получим:

$$\begin{gathered} (A_i \alpha_i \alpha_{i+1} + C_i \alpha_{i+1} + B_i) x_{i+1} + A_i \alpha_i \beta_{i+1} + A_i \beta_i + C_i \beta_{i+1} - F_i = 0, \\ i = 1,2,\dots n-1. \end{gathered}$$

Последние соотношения будут заведомо выполняться и притом независимо от решения, если потребовать, чтобы при $i = 1,2,\dots n-1$ имели место равенства:

$$\begin{gathered} A_i \alpha_i \alpha_{i+1} + C_i \alpha_{i+1} + B_i = 0, \\ A_i \alpha_i \beta_{i+1} + A_i \beta_i + C_i \beta_{i+1} - F_i = 0. \end{gathered}$$

Отсюда следуют рекуррентные соотношения для прогоночных коэффициентов:

$$\alpha_{i+1} = \frac{-B_i}{A_i \alpha_i + C_i},\ \beta_{i+1} = \frac{F_i - A_i \beta_i}{A_i \alpha_i + C_i},\ i = 1,2,\dots n-1. \tag{33}$$

Левое граничное условие $x_0 = q_0$ и соотношение $x_0 = \alpha_1 x_1 + \beta_1$ непротиворечивы, если положить

$$\alpha_1 = 0,\quad \beta_1 = q_0. \tag{34}$$

Остальные значения коэффициентов прогонки $\alpha_2,\dots,\alpha_n$ и $\beta_2,\dots,\beta_n$ находим из (33), чем и завершаем этап вычисления прогоночных коэффициентов.

Далее, согласно правому граничному условию

$$x_n = q_n. \tag{35}$$

Отсюда можно найти остальные неизвестные $x_{n-1},\dots x_1$ в процессе обратной прогонки с помощью рекуррентной формулы (32).

Число операций, которое требуется для решения системы общего вида (1) методом Гаусса, растет при увеличении n пропорционально n^3. Метод прогонки сводится к двум циклам: сначала по формулам (33) рассчитываются прогоночные коэффициенты, затем с их помощью по рекуррентным формулам (32) находятся компоненты решения системы x_i. Это означает, что с увеличением размеров системы число арифметических операций будет расти пропорционально n, а не n^3. Таким образом, метод прогонки в пределах сферы своего возможного применения является существенно более экономичным. К этому следует добавить особую простоту его программной реализации на компьютере.

Во многих прикладных задачах, которые приводят к СЛАУ с трехдиагональной матрицей, ее коэффициенты удовлетворяют неравенствам:

$$|C_i| > |A_i| + |B_i|, \tag{36}$$

которые выражают свойство диагонального преобладания. В частности, мы встретим такие системы в третьей и пятой главе.

Согласно теореме предыдущего раздела решение таких систем всегда существует и является единственным. Для них также справедливо утверждение, которое имеет важное значение для фактического расчета решения с помощью метода прогонки.

Лемма

Если для системы с трехдиагональной матрицей выполняется условие диагонального преобладания (36), то прогоночные коэффициенты удовлетворяют неравенствам:

$$|\alpha_i| \le 1. \tag{37}$$

Доказательство проведем по индукции. Согласно (34) $\alpha_1 = 0$, т. е. при $i = 1$ утверждение леммы верно. Допустим теперь, что оно верно для α_i и рассмотрим α_{i+1}:

$$|\alpha_{i+1}| = \left|\frac{B_i}{C_i + A_i \alpha_i}\right| \le \frac{|B_i|}{|C_i| - |A_i|} \le 1. \tag{38}$$

Итак, индукция от i к $i+1$ обоснована, что и завершает доказательство леммы.

Неравенство (37) для прогоночных коэффициентов α_i делает прогонку устойчивой. Действительно, предположим, что компонента решения x_i в результате процедуры округления рассчитана с некоторой ошибкой. Тогда при вычислении

следующей компоненты x_{i-1} по рекуррентной формуле (32) эта ошибка, благодаря неравенству (37), не будет нарастать.

§2. Обусловленность СЛАУ.

Серьезным препятствием при решении систем линейных алгебраических уравнений может оказаться возможность заметного отклонения приближенного решения от точного из-за незначительных возмущений правых частей уравнений, которые неизбежно возникают в приближенных вычислениях. Причиной такого нежелательного эффекта часто оказывается так называемая плохая обусловленность матрицы системы линейных уравнений.

2.1. Норма матрицы.

Рассмотрим линейное вещественное евклидово пространство E_n, элементами которого являются вектора в виде упорядоченной системы n чисел $\mathbf{x} = \{x_1,\dots x_n\}$. В пространстве E_n определены скалярное произведение

$$(\mathbf{x}, \mathbf{y}) = x_1 y_1 + \dots + x_n y_n \tag{39}$$

и евклидова норма

$$\|\mathbf{x}\| = \sqrt{(\mathbf{x}, \mathbf{x})} = \sqrt{x_1^2 + \dots + x_n^2}, \tag{40}$$

удовлетворяющая трем аксиомам нормы:

1. $\|\mathbf{x}\| \ge 0$, $\|\mathbf{x}\| = 0$ тогда и только тогда, когда $\mathbf{x} = 0$;
2. $\|\alpha \mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\| \quad \forall \alpha, \mathbf{x}$;
3. $\|\mathbf{x} + \mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$ (неравенство треугольника).

Для скалярного произведения справедливо неравенство Коши-Буняковского $|(\mathbf{x}, \mathbf{y})| \le \|\mathbf{x}\| \cdot \|\mathbf{y}\|$.

Рассмотрим квадратную матрицу A размером $n \times n$. Она определяет в пространстве E_n линейное преобразование

$$\mathbf{y} = A\mathbf{x} \tag{41}$$

или

$$y_i = \sum_{j=1}^{n} a_{ij} x_j,\ i = 1,\dots,n.$$

Введем величину

$$\|A\| = \sup_{\mathbf{x} \ne 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}, \tag{42}$$

которую принято называть нормой матрицы A, согласованной с нормой вектора $\|\mathbf{x}\|$. Записывая ненулевой вектор $\mathbf{x}$ в виде

$$\mathbf{x} = \|\mathbf{x}\| \mathbf{z},$$

где $\mathbf{z}$ вектор единичной длины: $\|\mathbf{z}\| = 1$, получим представление для нормы, эквивалентное (42)

$$\|A\| = \sup_{\|\mathbf{z}\| = 1} \|A\mathbf{z}\|. \tag{43}$$

Отсюда следует, что в конечномерном пространстве норма матрицы ограничена, причем на единичной сфере всегда найдется такой вектор $\mathbf{z}_0$, что

$$\|A\| = \|A\mathbf{z}_0\|.$$

Наконец, из определения нормы (42) следует, что

$$\|A\mathbf{x}\| \le \|A\| \cdot \|\mathbf{x}\|. \tag{44}$$

Это простое неравенство лежит в основе всех дальнейших оценок.

2.2. Корректность решения СЛАУ.

Следуя Адамару, будем называть математическую задачу корректной, если выполняются три условия:

1. Решение задачи существует.
2. Решение задачи единственное.
3. Решение задачи непрерывно зависит от входных данных.

Обсудим с точки зрения этого определения задачу решения СЛАУ с неравным нулю определителем

$$A\mathbf{x} = \mathbf{f}, \tag{45}$$

считая матрицу A фиксированной и рассматривая в качестве входных данных вектор правых частей системы $\mathbf{f} = \{f_1, f_2, \dots, f_n\} \in E_n$.

Условие $\Delta \ne 0$ гарантирует существование у матрицы A обратной матрицы A^{-1}, через которую решение системы (45) можно записать в виде

$$\mathbf{x} = A^{-1}\mathbf{f}. \tag{46}$$

Пусть теперь правая часть подверглась возмущению $\delta\mathbf{f}$ и стала равной $\tilde{\mathbf{f}} = \mathbf{f} + \delta\mathbf{f}$. Тогда, согласно (46), решение $\tilde{\mathbf{x}}$ возмущенной системы

$$A\tilde{\mathbf{x}} = \tilde{\mathbf{f}} \tag{47}$$

тоже можно записать через обратную матрицу A^{-1}:

$$\tilde{\mathbf{x}} = A^{-1}\tilde{\mathbf{f}} = A^{-1}\mathbf{f} + A^{-1}\delta\mathbf{f} = \mathbf{x} + \delta\mathbf{x}, \tag{48}$$

где

$$\delta\mathbf{x} = A^{-1}\delta\mathbf{f}. \tag{49}$$

Отсюда получаем

$$\|\delta\mathbf{x}\| \le \|A^{-1}\| \|\delta\mathbf{f}\|. \tag{50}$$

Неравенство (50) доказывает непрерывную зависимость возмущения решения $\delta\mathbf{x}$ от возмущения правой части $\delta\mathbf{f}$:

$$\|\delta\mathbf{x}\| \to 0 \text{ при } \|\delta\mathbf{f}\| \to 0. \tag{51}$$

Это означает, что решение СЛАУ с неравным нулю определителем Δ- корректная математическая задача: для нее выполняются все три требования корректности Адамара.

2.3. Число обусловленности матрицы.

Исходное уравнение (45) позволяет написать неравенство:

$$\|\mathbf{f}\| \le \|A\| \|\mathbf{x}\|. \tag{52}$$

Перемножая его с неравенством того же знака (50), получим:

$$\|\mathbf{f}\| \|\delta\mathbf{x}\| \le \|A\| \|A^{-1}\| \|\mathbf{x}\| \|\delta\mathbf{f}\|. \tag{53}$$

Пусть $\mathbf{f} \ne 0$, тогда, согласно (46), $\mathbf{x} \ne 0$ и неравенство (53) можно переписать в виде:

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le M_A \frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|}, \tag{54}$$

где

$$M_A = \|A\| \cdot \|A^{-1}\|. \tag{55}$$

Число M_A называется числом обусловленности матрицы A. Оно позволяет оценить относительную погрешность решения через относительную погрешность возмущения правой части. Поскольку исходная система (45) линейная, оценка относительной погрешности является более естественной, чем оценка абсолютной погрешности. Чем больше M_A, тем резче реагирует решение на возмущение правой части. Поэтому матрицы с большим числом обусловленности и соответствующие им СЛАУ называют плохо обусловленными. Для оценки роли, которую играет число обусловленности при решении линейных алгебраических систем, разберем задачу.

Задача 1

Рассмотреть систему двух уравнений

$$\begin{gathered} x_1 + 0 \cdot x_2 = 1 \\ x_1 + 0.01 \cdot x_2 = 1 \end{gathered},\ A = \begin{bmatrix} 1 & 0 \\ 1 & 0.01 \end{bmatrix},\ \mathbf{f} = \{1,1\} \tag{56}$$

и соответствующую ей возмущенную систему

$$\begin{gathered} x_1 + 0 \cdot x_2 = 1 \\ x_1 + 0.01 \cdot x_2 = 1.01 \end{gathered},\ A = \begin{bmatrix} 1 & 0 \\ 1 & 0.01 \end{bmatrix},\ \tilde{\mathbf{f}} = \{1,1.01\}. \tag{57}$$

Выписать решения этих систем, подсчитать погрешность возмущения правой части и соответствующую ей погрешность возмущения решения. Найти число обусловленности матрицы A, составить с его помощью теоретическую оценку погрешности (54) и сравнить результат с результатом, полученным непосредственно по известным решениям систем.

В данном случае определитель матрицы A отличен от нуля

$$\Delta = \det A = 0.01,$$

т. е. обе системы невырожденные. Система (57) отличается от системы (56) возмущением правой части

$$\mathbf{f} = \{1,1\},\ \|\mathbf{f}\| = \sqrt{2},\ \tilde{\mathbf{f}} = \{1,1.01\},\ \delta\mathbf{f} = \{0,0.01\},\ \|\delta\mathbf{f}\| = 0.01.$$

Решения систем (56) и (57) имеют вид:

$$\mathbf{x} = \{1,0\},\ \|\mathbf{x}\| = 1,\ \tilde{\mathbf{x}} = \{1,1\}, \delta\mathbf{x} = \{0,1\},\ \|\delta\mathbf{x}\| = 1.$$

При этом

$$\frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|} = \frac{0.01}{\sqrt{2}},\ \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} = 1. \tag{58}$$

Мы видим, что небольшое относительное возмущение правой части привело к сильному возмущению решения: относительная погрешность решения равна единице. Этот результат означает, что исходная система плохо обусловлена. Чтобы убедиться в

этом, подсчитаем число обусловленности матрицы A, напишем с его помощью теоретическую оценку (54) и сравним ее с фактическим результатом (58).

Выпишем линейное преобразование $\mathbf{y} = A\mathbf{x}$ отвечающее матрице системы

$$\begin{aligned} y_1 &= x_1 \\ y_2 &= x_1 + 0.01 x_2, \end{aligned}$$

при этом

$$\|A\mathbf{x}\| = \sqrt{x_1^2 + (x_1 + 0.01 x_2)^2}.$$

Наложим ограничение

$$x_1^2 + x_2^2 = 1,$$

тогда в силу (43)

$$\|A\mathbf{x}\| = \max \sqrt{2x_1^2 + 0.0001 x_2^2 + 0.02 x_1 x_2},\ x_1^2 + x_2^2 = 1.$$

Если положить $x_1 = \cos\varphi$, $x_2 = \sin\varphi$, то задача сведется к отысканию максимума выражения

$$g(\varphi) = \sqrt{2\cos^2\varphi + 0.02 \sin\varphi\cos\varphi + 0.0001 \sin^2\varphi},$$

зависящего только от одной переменной φ, $0 \le \varphi \le 2\pi$.

Переходя к тригонометрическим функциям двойного угла

$$2\cos^2\varphi = 1 + \cos 2\varphi,\ 2\sin^2\varphi = 1 - \cos 2\varphi,\ 2\sin\varphi\cos\varphi = \sin 2\varphi,$$

сведем подрадикальное выражение к виду:

$$1.00005 + 0.01 \sin 2\varphi + 0.99995 \cos 2\varphi$$

Для комбинации

$$B_1 \cos 2\varphi + B_2 \sin 2\varphi = \sqrt{B_1^2 + B_2^2} \cos(2\varphi - \varphi_0),\ 0 \le \varphi \le 2\pi,$$

где

$$\varphi_0 = \operatorname{arctg}\left(B_1 \Big/ B_2\right),\ B_1 = 0.99995,\ B_2 = 0.01,$$

максимальное значение равно

$$\sqrt{B_1^2 + B_2^2} = \sqrt{0.99995^2 + 0.01^2}.$$

Следовательно

$$\|A\| = \sqrt{1.00005 + \sqrt{0.99995^2 + 0.01^2}}.$$

С приемлемой точностью это число равно $\sqrt{2}$: $\|A\| \approx \sqrt{2}$.

Аналогичным образом находится норма обратной матрицы

$$A^{-1} = \begin{bmatrix} 1 & 0 \\ -100 & 100 \end{bmatrix},\ \|A^{-1}\| \approx 100\sqrt{2}.$$

Таким образом, в данном примере

$$M_A = \|A\| \cdot \|A^{-1}\| \approx 200. \tag{59}$$

В результате теоретическая оценка (54) принимает вид:

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le 200 \frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|}$$

Она согласуется с результатом (58), который мы получили, непосредственно решая системы (56) и (57).

В процессе решения задачи мы убедились в том, что подсчет числа обусловленности является сложной задачей, особенно с учетом того, что нужно вычислять норму не только прямой, но и обратной матрицы. Поэтому желательно получить какие-нибудь конструктивные оценки этой важнейшей характеристики системы.

2.4. Оценка числа обусловленности.

Для числа обусловленности матрицы A справедливо неравенство

$$M_A \ge |\lambda_{\max}| / |\lambda_{\min}|, \tag{60}$$

где $\lambda_{\min}$ и $\lambda_{\max}$ соответственно минимальное и максимальное по модулю значения характеристических чисел матрицы A. Соотношение (60) корректно, поскольку в силу невырожденности матрицы $\lambda_{\min} \ne 0$.

В самом деле пусть $\mathbf{y}$ - собственный вектор линейного преобразования, связанного с матрицей A, отвечающий $\lambda_{\max}$:

$$A\mathbf{y} = \lambda_{\max}\mathbf{y},$$

тогда

$$|\lambda_{\max}| \cdot \|\mathbf{y}\| = \|A\mathbf{y}\| \le \|A\| \cdot \|\mathbf{y}\|,$$

и, следовательно, поскольку $\|\mathbf{y}\| \ne 0$

$$|\lambda_{\max}| \le \|A\|.$$

Аналогичным образом для собственного вектора $\mathbf{z}$, связанного с $\lambda_{\min}$, имеем

$$A\mathbf{z} = \lambda_{\min}\mathbf{z}$$

или

$$A^{-1}\mathbf{z} = \frac{1}{\lambda_{\min}}\mathbf{z}.$$

Отсюда следует оценка

$$\frac{1}{|\lambda_{\min}|} \le \|A^{-1}\|.$$

Перемножая два последних неравенства, придем к утверждению (60).

Если матрица симметричная $A = A^*$, то все её характеристические значения вещественны, причем

$$\|A\| = |\lambda_{\max}| \text{ и } \|A^{-1}\| = \frac{1}{|\lambda_{\min}|},$$

поэтому для таких матриц

$$M_A = \left|\frac{\lambda_{\max}}{\lambda_{\min}}\right|. \tag{61}$$

Из полученной оценки для M_A следуют два важных вывода:

1) $M_A \ge 1$;

2) Число обусловленности тем больше, чем больше разброс характеристических чисел матрицы. Поэтому с увеличением размера матрицы, вообще говоря, её обусловленность имеет тенденцию к ухудшению.

Возвращаясь к рассмотренной выше задаче, без труда находим: $\lambda_{\min} = 0.01$, $\lambda_{\max} = 1$ и, следовательно, справедлива оценка снизу

$$M_A \ge \lambda_{\max} / \lambda_{\min} = 100,$$

причем точность этой оценки невысока, но порядок она передает правильно.

В заключение данного параграфа еще раз отметим, что для систем уравнений с большой размерностью "хорошая" обусловленность ($M_A \approx 1$) является скорее исключением, чем правилом и обычно приходится иметь дело с плохо обусловленными матрицами ($M_A \gg 1$), причем получение оценки числа обусловленности вызывает большие трудности.

§3. Итерационные методы.

3.1. Построение итерационных последовательностей.

Мы видели, что процедура решения СЛАУ

$$A\mathbf{x} = \mathbf{f} \tag{62}$$

с плохо обусловленной матрицей A может приводить к существенным отклонениям получаемого ответа от точного решения при незначительных возмущениях правой части. Однако появление таких возмущений неизбежно, например, при преобразовании вектора правых частей в методе Гаусса из-за ошибок округления при выполнении арифметических операций. Чем выше порядок матрицы, тем больше может оказаться результирующая погрешность.

Этого недостатка лишены итерационные методы решения СЛАУ. При их применении ответ получается в процессе построения последовательных приближений (итераций) $\mathbf{x}_k = \{x_1^k, x_2^k, \dots x_n^k\}$, сходящихся к решению системы (62) в пространстве E_n с евклидовой нормой $\|\mathbf{x}\|$

$$\lim_{k \to \infty} \mathbf{x}_k = \mathbf{x} \tag{63}$$

Здесь при записи вектора $\mathbf{x}_k$ через его компоненты x_i^k нижний индекс i означает номер компоненты ($1 \le i \le n$), верхний индекс k - номер итерации. Сходимость последовательности $\mathbf{x}_k$ к решению системы $\mathbf{x}$ означает, что

$$\lim_{k \to \infty} \|\mathbf{x}_k - \mathbf{x}\| = \lim_{k \to \infty} \sqrt{(x_1^k - x_1)^2 + (x_2^k - x_2)^2 + \dots + (x_n^k - x_n)^2} = 0. \tag{64}$$

Необходимым и достаточным условием предельного равенства (64) в конечномерном евклидовом пространстве E_n является покомпонентная сходимость:

$$\lim_{k \to \infty} x_i^k = x_i,\ 1 \le i \le n.$$

Сходимость обеспечивает принципиальную возможность получить в процессе итераций ответ с любой наперед заданной степенью точности.

С итерационными последовательностями вы встречались. Каждый следующий член такой последовательности выражается через предыдущие, уже известные. Если, например, формула для вычисления очередного члена последовательности имеет вид:

$$\mathbf{x}_{k+1} = F\left(\mathbf{x}_k, \mathbf{x}_{k-1}, \dots, \mathbf{x}_{k-m+1}\right),$$

то говорят о m-шаговом итерационном алгоритме. В частности, в простейшем случае очередной член последовательности $\mathbf{x}_{k+1}$ может выражается только через предыдущий $\mathbf{x}_k$:

$$\mathbf{x}_{k+1} = F\left(\mathbf{x}_k\right).$$

Такие итерационные алгоритмы называют одношаговыми.

При обсуждении итерационных методов решения СЛАУ мы ограничимся линейными одношаговыми алгоритмами, которые обычно записывают в стандартной канонической форме:

$$B_{k+1} \frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\tau_{k+1}} + A\mathbf{x}_k = \mathbf{f},\ \det B_{k+1} \ne 0,\ \tau_{k+1} > 0. \tag{65}$$

В такой записи процесс характеризуется последовательностью матриц B_{k+1} и числовых параметров τ_{k+1}, которые называют итерационными параметрами. Если матрицы B_{k+1} и параметры τ_{k+1} не меняются в процессе итераций, т. е. не зависят от индекса k, то итерационный процесс называется стационарным.

Перепишем формулу (65) в виде

$$B_{k+1}\mathbf{x}_{k+1} = \mathbf{F}_{k+1}, \tag{66}$$

где

$$\mathbf{F}_{k+1} = \left(B_{k+1} - \tau_{k+1}A_{k+1}\right)\mathbf{x}_k + \tau_{k+1}\mathbf{f}. \tag{67}$$

Мы видим, что построение очередной итерации сводится к решению системы уравнений (66) с правой частью (67), зависящей от предыдущей итерации $\mathbf{x}_k$. Такую задачу приходится решать многократно, поэтому матрицы B_{k+1} следует выбирать достаточно простыми. Если построение отдельных итераций будет соизмеримым по сложности с решением исходной задачи, то метод окажется лишенным практического смысла.

Наиболее прост в реализации итерационный процесс с единичной матрицей: $B_{k+1} = E$. В этом случае формулы (66), (67) дают явное выражение очередной итерации через предыдущую:

$$\mathbf{x}_{k+1} = \left(E - \tau_{k+1}A\right)\mathbf{x}_k + \tau_{k+1}\mathbf{f}. \tag{68}$$

Из неявных итерационных методов выделим сравнительно легко реализуемые методы с диагональными матрицами: $B_{k+1} = D_{k+1}$ и верхними или нижними треугольными матрицами: $B_{k+1} = T_{k+1}$.

3.2. Проблема сходимости итерационного процесса.

Итерационный процесс может быть использован для решения СЛАУ только при условии сходимости. Для исследования его сходимости введем две характеристики. Первая из них – погрешность решения:

$$\mathbf{z}_k = \mathbf{x}_k - \mathbf{x}. \tag{69}$$

Смысл этого вектора ясен. Сходимость итерационного процесса согласно (63) и (64) означает, что

$$\lim_{k\to\infty}\mathbf{z}_k=0,\ \lim_{k\to\infty}z_i^k=0,\ 1\le i\le n. \tag{70}$$

Вторая характеристика – невязка:

$$\boldsymbol{\psi}_k=A\mathbf{x}_k-\mathbf{f}. \tag{71}$$

Она показывает, насколько хорошо или, наоборот, плохо член итерационной последовательности $\mathbf{x}_k$ удовлетворяет исходной системе.

Установим связь между $\mathbf{z}_k$ и $\boldsymbol{\psi}_k$:

$$\boldsymbol{\psi}_k=A\mathbf{x}_k-\mathbf{f}=A(\mathbf{z}_k+\mathbf{x})-\mathbf{f}=A\mathbf{z}_k. \tag{72}$$

Можно также написать обратное соотношение:

$$\mathbf{z}_k=A^{-1}\boldsymbol{\psi}_k. \tag{73}$$

Из формул (72) и (73) вытекают оценки:

$$\|\boldsymbol{\psi}_k\|\le\|A\|\cdot\|\mathbf{z}_k\|,\ \|\mathbf{z}_k\|\le\|A^{-1}\|\cdot\|\boldsymbol{\psi}_k\|. \tag{74}$$

Они показывают, что погрешность решения $\mathbf{z}_k$ стремится к нулю тогда и только тогда, когда стремится к нулю невязка $\boldsymbol{\psi}_k$. Этот результат позволяет судить о сходимости или расходимости итерационного процесса по поведению невязки, которая доступна прямому вычислению и благодаря этому может контролироваться.

При исследовании сходимости итерационных методов большую роль играют свойства матриц A и B_{n+1}, в первую очередь такие как самосопряженность и знакоопределенность. Напомним, что в вещественном евклидовом пространстве E_n для каждого линейного преобразования существует единственное сопряженное к нему линейное преобразование, определяемое тождественным равенством скалярных произведений:

$$(A\mathbf{x},\mathbf{y})=(\mathbf{x},A^*\mathbf{y}),\ \forall\mathbf{x},\mathbf{y}\in E_n. \tag{75}$$

В частности,

$$(A\mathbf{x},\mathbf{x})=(\mathbf{x},A^*\mathbf{x}),\ \forall\mathbf{x}\in E_n.$$

Преобразование называется самосопряженным, если

$$(A\mathbf{x},\mathbf{y})=(\mathbf{x},A\mathbf{y}),\ \forall\mathbf{x},\mathbf{y}\in E_n. \tag{76}$$

Матрицы сопряженных преобразований в ортонормированном базисе связаны простым транспонированием:

$$a_{ij}^*=a_{ji},\ \forall i,j=1,\dots,n.$$

Свойство самосопряженности преобразования равносильно в этом случае выполнению условия совпадения матриц A и A^*:

$$a_{ij}=a_{ji}^*,\ \forall i,j=1,\dots,n,$$

Как известно, любая матрица представима в виде:

$$A=\overline{A}+\tilde{A}, \tag{77}$$

где

$$\overline{A}=\frac{A+A^*}{2}=\overline{A}^*,\ \tilde{A}=\frac{A-A^*}{2}=-\tilde{A}^*. \tag{78}$$

Нетрудно видеть, что

$$\begin{aligned}&(A\mathbf{x},\mathbf{x})=(A^*\mathbf{x},\mathbf{x})=(\overline{A}\mathbf{x},\mathbf{x}),\\&(\tilde{A}\mathbf{x},\mathbf{x})=0.\end{aligned} \tag{79}$$

В дальнейшем мы будем опираться на следующие важные свойства самосопряженных преобразований:

а) все собственные значения самосопряженного линейного преобразования (характеристические числа матрицы A) вещественны;

б) самосопряженное линейное преобразование всегда имеет полный набор линейно независимых собственных векторов, из которых можно образовать ортонормированный базис пространства E_n. В этом базисе матрица линейного преобразования принимает диагональный вид, причем на диагонали стоят все собственные значения этого преобразования с учетом их кратности.

Наконец, матрица линейного преобразования A называется положительно определенной, если для любого, отличного от нуля $\mathbf{x}\in E_n$:

$$(A\mathbf{x},\mathbf{x})>0,\ \sum_{i,j=1}^n a_{ij}x_ix_j>0,\ \forall\mathbf{x}\in E_n,\mathbf{x}\ne0. \tag{80}$$

Для краткости, если это не вызывает недоразумений, будем часто писать $A>0$.

Необходимым и достаточным условием положительной определенности самосопряженной матрицы A является критерий Сильвестра, из которого в частности следует строгая положительность всех диагональных элементов:

$$a_{ij}>0,\ 1\le i\le n. \tag{81}$$

Условимся обозначать собственные векторы линейного преобразования с матрицей A как $\mathbf{e}_i$, её характеристические числа как λ_i, координаты произвольного вектора $\mathbf{x}$ в ортонормированном базисе из собственных векторов $\mathbf{e}_i$ как ξ_i.

Для дальнейшего рассмотрения будут полезны три леммы.

Лемма 1.

Для того, чтобы симметричная $(A=A^*)$ матрица была положительно определенной, необходимо и достаточно, чтобы все её характеристические числа были положительны: $\lambda_i>0$.

Необходимость. Выберем любой собственный вектор $\mathbf{e}_i$ линейного преобразования с матрицей A, тогда

$$(A\mathbf{e}_i,\mathbf{e}_i)=\lambda_i>0.$$

Достаточность. Расположим для определенности все характеристические значения матрицы $A=A^*$ в порядке убывания:

$$\lambda_1\ge\lambda_2\ge\dots\ge\lambda_n>0.$$

Поскольку по условию леммы $\lambda_i>0$, то в ортонормированном базисе из собственных векторов преобразования с матрицей A для любого $\mathbf{x}\ne0$ имеем

$$(A\mathbf{x},\mathbf{x})=\sum_{i=1}^n\lambda_i\xi_i^2>0,\ \forall\{\xi_i\},\left(\sum_{i=1}^n\xi_i^2>0\right).$$

Поэтому, очевидно, что $A>0$.

Лемма 2.

Пусть $A=A^*>0$, и $\lambda_1\ge\dots\ge\lambda_n>0$ - упорядоченный набор характеристических чисел этой матрицы, тогда

$$\lambda_n\|\mathbf{x}\|^2\le(A\mathbf{x},\mathbf{x})\le\lambda_1\|\mathbf{x}\|^2. \tag{82}$$

Доказательство предлагается провести самостоятельно.

Лемма 3.

Если $A>0$, то всегда найдется постоянное число $\delta>0$, такое что

$$(A\mathbf{x},\mathbf{x})\ge\delta\|\mathbf{x}\|^2,\ \forall\mathbf{x}\in E_n \tag{83}$$

Доказательство.

Если $A=A^*$, то достаточно положить $\delta=\lambda_n$. В общем случае напомним, что согласно (79)

$$(A\mathbf{x},\mathbf{x})=(\overline{A}\mathbf{x},\mathbf{x})>0,$$

где $\overline{A}=\overline{A}^*$, поэтому согласно предыдущей лемме

$$(A\mathbf{x},\mathbf{x})\ge\overline{\lambda}_n\|\mathbf{x}\|^2,$$

где $\overline{\lambda}_n>0$ - минимальное характеристическое число матрицы $\overline{A}=(A+A^*)/2$. Полагая, что $\delta=\overline{\lambda}_n$, приходим к требуемому неравенству (83).

3.3. Достаточные условия сходимости итерационного процесса.

В этом разделе мы рассмотрим стационарный итерационный процесс (65), когда матрица B и итерационный параметр τ не зависят от индекса k, и докажем следующую теорему о достаточных условиях его сходимости.

Теорема Самарского

Пусть A - самосопряженная положительно определенная матрица:

$$A=A^*,\ A>0, \tag{84}$$

$B-\frac{\tau}{2}A$ - положительно определенная матрица, τ - положительное число:

$$B-\frac{\tau}{2}A>0,\ \tau>0. \tag{85}$$

Тогда при любом выборе нулевого приближения $\mathbf{x}_0$ итерационный процесс, который определяется рекуррентной формулой (65), сходится к решению исходной системы (62).

Прежде, чем переходить к доказательству теоремы, обсудим более подробно главное ее требование – положительную определенность матрицы $B-\frac{\tau}{2}A$. Это требование можно переписать в виде:

$$(B\mathbf{x},\mathbf{x})>\frac{\tau}{2}(A\mathbf{x},\mathbf{x}),\ \forall\mathbf{x}\in E_n,\ \mathbf{x}\ne0. \tag{86}$$

т. е. оно, в частности, предполагает, что матрица B является положительно определенной. Кроме того, неравенство (86) определяет интервал, в котором может изменяться параметр τ:

$$0<\tau<\tau_0=\inf_{x\ne0}\frac{2(B\mathbf{x},\mathbf{x})}{(A\mathbf{x},\mathbf{x})}. \tag{87}$$

После этих замечаний перейдем к доказательству теоремы. Выразим из соотношения (69) $\mathbf{x}_k$ через $\mathbf{z}_k$:

$$\mathbf{x}_k=\mathbf{z}_k+\mathbf{x}$$

и подставим в рекуррентную формулу для итерационной последовательности (65). В результате получим:

$$B\frac{\mathbf{z}_{k+1}-\mathbf{z}_k}{\tau}+A\mathbf{z}_k=0. \tag{88}$$

Отличие итерационной формулы (88) от (65) заключается в том, что она является однородной.

Матрица B - положительно определенная. Следовательно она невырожденная и имеет обратную B^{-1}. С ее помощью рекуррентное соотношение (88) можно разрешить относительно $\mathbf{z}_{k+1}$:

$$\mathbf{z}_{k+1}=\mathbf{z}_k-\tau B^{-1}A\mathbf{z}_k=\mathbf{z}_k-\tau\boldsymbol{\omega}_k, \tag{89}$$

где

$$\boldsymbol{\omega}_k=B^{-1}A\mathbf{z}_k,\ \text{так что}\ A\mathbf{z}_k=B\boldsymbol{\omega}_k. \tag{90}$$

Умножая обе части равенства (89) слева на матрицу A, получим еще одно рекуррентное соотношение

$$A\mathbf{z}_{k+1}=A\mathbf{z}_k-\tau A\boldsymbol{\omega}_k. \tag{91}$$

Рассмотрим последовательность положительных функционалов:

$$J_k=(A\mathbf{z}_k,\mathbf{z}_k). \tag{92}$$

Составим аналогичное выражение для J_{k+1} и преобразуем его с помощью рекуррентных формул (89) и (91):

$$\begin{aligned}J_{k+1}&=(A\mathbf{z}_k-\tau A\boldsymbol{\omega}_k,\mathbf{z}_k-\tau\boldsymbol{\omega}_k)=(A\mathbf{z}_k,\mathbf{z}_k)-\tau(A\boldsymbol{\omega}_k,\mathbf{z}_k)-\\&-\tau(A\mathbf{z}_k,\boldsymbol{\omega}_k)+\tau^2(A\boldsymbol{\omega}_k,\boldsymbol{\omega}_k).\end{aligned} \tag{93}$$

Из самосопряженности матрицы A и формулы (90) следует

$$(A\boldsymbol{\omega}_k,\mathbf{z}_k)=(A\mathbf{z}_k,\boldsymbol{\omega}_k)=(B\boldsymbol{\omega}_k,\boldsymbol{\omega}_k).$$

В результате формула (93) принимает вид:

$$J_{k+1}=J_k-2\tau(B\boldsymbol{\omega}_k,\boldsymbol{\omega}_k)+\tau^2(A\boldsymbol{\omega}_k,\boldsymbol{\omega}_k)=J_k-2\tau\left(\left(B-\frac{\tau}{2}A\right)\boldsymbol{\omega}_k,\boldsymbol{\omega}_k\right). \tag{94}$$

Таким образом, последовательность функционалов J_k с учетом условия $B-\frac{\tau}{2}A>0$ образует монотонно невозрастающую последовательность, ограниченную снизу нулем

$$J_k\ge J_{k+1}\ge\dots\ge0. \tag{95}$$

Поэтому она сходится. Далее, согласно лемме 3

$$\left(\left(B-\frac{\tau}{2}A\right)\boldsymbol{\omega}_k,\boldsymbol{\omega}_k\right)\ge\delta\|\boldsymbol{\omega}_k\|^2,$$

где $\delta>0$ - строго положительная константа. В результате, согласно (94) и (95) будем иметь

$$J_{k+1}-J_k=2\tau\left(\left(B-\frac{\tau}{2}A\right)\boldsymbol{\omega}_k,\boldsymbol{\omega}_k\right)\ge2\tau\delta\|\boldsymbol{\omega}_k\|^2. \tag{96}$$

Из этого неравенства и сходимости последовательности функционалов J_k следует, что $\|\boldsymbol{\omega}_k\|\to0$ при $k\to\infty$. В свою очередь $\mathbf{z}_k=A^{-1}\boldsymbol{\omega}_k$, так что

$$\|\mathbf{z}_k\|\le\|A^{-1}\|\cdot\|B\|\cdot\|\boldsymbol{\omega}_k\|\to0$$

Теорема доказана.

3.4. Метод простой итерации.

Такое название получил метод, при котором в качестве матрицы B выбирается единичная матрица: $B=E$, а итерационный параметр τ предполагается независящим от номера итерации k. Иными словами, метод простой итерации – это явный стационарный метод, когда очередная итерация x_{k+1} вычисляется по рекуррентной формуле

$$\mathbf{x}_{k+1}=(E-\tau A)\mathbf{x}_k+\tau\mathbf{f} \tag{97}$$

Будем считать, что матрица A удовлетворяет условию теоремы Самарского, $A=A^*>0$, тогда формула (87), определяющая границу интервала сходимости по итерационному параметру τ, принимает вид

$$\tau_0=\inf_{x\ne0}\frac{2(\mathbf{x},\mathbf{x})}{(A\mathbf{x},\mathbf{x})}=\frac{2}{\sup\limits_{x\ne0}\dfrac{(A\mathbf{x},\mathbf{x})}{(\mathbf{x},\mathbf{x})}}. \tag{98}$$

Пусть $\mathbf{e}_1,\mathbf{e}_2,\dots,\mathbf{e}_n$ - ортонормированный базис собственных векторов оператора, соответствующего матрице A. В силу положительной определенности все его собственные значения положительны. Будем считать их занумерованными в порядке убывания:

$$\lambda_1\ge\lambda_2\ge\dots\ge\lambda_n>0 \tag{99}$$

Разложим вектор $\mathbf{x}\ne0$ по базису собственных векторов

$$\mathbf{x}=\xi_1\mathbf{e}_1+\xi_2\mathbf{e}_2+\dots+\xi_n\mathbf{e}_n,$$

тогда

$$(\mathbf{x},\mathbf{x})=\xi_1^2+\xi_2^2+\dots+\xi_n^2,\ (A\mathbf{x},\mathbf{x})=\lambda_1\xi_1^2+\lambda_2\xi_2^2+\dots+\lambda_n\xi_n^2$$

и

$$\sup_{x\ne0}\frac{(A\mathbf{x},\mathbf{x})}{(\mathbf{x},\mathbf{x})}=\sup_{x\ne0}\frac{\lambda_1\xi_1^2+\lambda_2\xi_2^2+\dots+\lambda_n\xi_n^2}{\xi_1^2+\xi_2^2+\dots+\xi_n^2}=\lambda_1.$$

В результате из формулы (87) следует, что метод простой итерации сходится при любом τ, принадлежащем интервалу

$$0<\tau<\tau_0=\frac{2}{\lambda_1}. \tag{100}$$

Дальнейшее исследование метода простой итерации построим на конкретном анализе рекуррентной формулы (97). Введем матрицу оператора перехода

$$S=E-\tau A,\ S=S^* \tag{101}$$

и перепишем формулу (97) в виде

$$\mathbf{x}_{k+1}=S\mathbf{x}_k+\tau\mathbf{f}. \tag{102}$$

При этом погрешность $\mathbf{z}_k=\mathbf{x}-\mathbf{x}_k$ будет удовлетворять аналогичному рекуррентному соотношению, только однородному

$$\mathbf{z}_{k+1}=S\mathbf{z}_k. \tag{103}$$

Докажем две леммы, которые позволяют более полно исследовать условия сходимости метода простой итерации.

Лемма 1

Пусть оператор, который порождает матрица A, имеет собственный вектор $\mathbf{e}_i$ с собственным значением λ_i, тогда оператор перехода, который порождается матрицей S (101), также имеет собственный вектор $\mathbf{e}_i$, но с собственным значением

$$\mu_i(\tau)=1-\tau\lambda_i. \tag{104}$$

Доказательство элементарно. Оно проводится прямой проверкой

$$S\mathbf{e}_i=(E-\tau A)\mathbf{e}_i=(1-\tau\lambda_i)\mathbf{e}_i=\mu_i\mathbf{e}_i$$

При самосопряженной матрице A матрица S также является самосопряженной (101). Следовательно, ее норма определяется наибольшим по модулю собственным значением $\mu_i(\tau)$ (104):

$$\|S\|=\max_{1\le i\le n}|\mu_i(\tau)|. \tag{105}$$

Лемма 2

Для того, чтобы метод простой итерации сходился к решению системы (62) при любом выборе начального приближения, необходимо и достаточно, чтобы все собственные значения оператора перехода S были по модулю меньше единицы:

$$|\mu_i(\tau)|<1,\ 1\le i\le n \tag{106}$$

Достаточность. Условие (106) означает, что норма матрицы S, согласно (105), будет меньше единицы: $\|S\|<1$. В результате получаем

$$\|\mathbf{z}_{k+1}\|\le\|S\|\cdot\|\mathbf{z}_k\|\le\dots\le\|S\|^k\cdot\|\mathbf{z}_0\|\to0,\ \text{при}\ k\to\infty. \tag{107}$$

Необходимость. Допустим, что среди собственных значений μ_i (104) нашлось хотя бы одно μ_j, которое не удовлетворяет условию леммы (106), т. е.

$$|\mu_j|\ge1.$$

Выберем нулевой член итерационной последовательности в виде $\mathbf{x}_0=\mathbf{x}+\mathbf{e}_j$, где $\mathbf{x}$ решение системы (62), тогда нулевой член последовательности погрешностей совпадет с собственным вектором $\mathbf{e}_j$ оператора перехода S: $\mathbf{z}_0=\mathbf{e}_j$. В результате рекуррентная формула для следующих членов последовательности погрешностей примет вид:

$$\mathbf{z}_k=S^k\mathbf{e}_j=\mu_j^k\mathbf{e}_j,\ \|\mathbf{z}_k\|=|\mu_j|^k\ge1.$$

т. е. $\|\mathbf{z}_k\|\not\to0$. Необходимость выполнения неравенства (106) для всех собственных значений μ_i для сходимости метода простой итерации доказана.

Лемма 2 определяет программу дальнейшего исследования сходимости метода простой итерации: нужно установить диапазон изменения параметра τ при котором все собственные значения удовлетворяют неравенству (106). Это легко сделать. На рис. 1 приведены графики убывающих линейных функций $\mu_i(\tau)$ (104). Все они выходят из одной точки $\tau=0$, $\mu=1$ и идут вниз из-за отрицательных коэффициентов при τ, причем быстрее всех убывает функция $\mu_1(\tau)$. Когда она принимает значение (-1), условие (106) для нее перестает выполняться:

$$\mu_1(\tau)=1-\tau\lambda_1=-1,\ \text{при}\ \tau=\tau_0=2/\lambda_1.$$

Найденное значение τ_0 является границей интервала сходимости метода простой итерации

$$0<\tau<\tau_0=2/\lambda_1. \tag{108}$$

Это неравенство нам уже известно. Оно было получено ранее из теоремы Самарского как достаточное условие сходимости. Дополнительный анализ на основе леммы 2 позволяет уточнить результат. Теперь мы установили, что принадлежность итерационного параметра τ интервалу (108) является необходимым и достаточным условием сходимости метода простой итерации.

Перейдем к исследованию скорости сходимости метода. Оценка погрешности (107) показывает, что она убывает по закону геометрической прогрессии со знаменателем

$$q(\tau)=\|S\|=\max_{1\le i\le n}|\mu_i(\tau)|.$$

Рассмотрим рис. 2, который поможет нам провести анализ этой формулы. Он аналогичен рис.1, только на нем приведены графики не функций $\mu_i(\tau)$, а их модулей. При малых τ все собственные значения $\mu_i(\tau)$ (104) положительны, причем наибольшим из них является $\mu_n(\tau)$, которое убывает с ростом τ с наименьшей скоростью. Однако с переходом через точку $\tau_0/2$ собственное значение $\mu_1(\tau)$, меняя знак, становится отрицательным. В результате теперь его модуль с увеличением τ не убывает, а растет и при $\tau\to\tau_0$ приближается к предельному значению – к единице.

Найдем на отрезке $\left[\frac{1}{2}\tau_0,\tau_0\right]$ точку τ_*, в которой убывающая функция $\mu_n(\tau)$ сравнивается с возрастающей функцией $|\mu_1(\tau)|=-\mu_1(\tau)$. Она определяется уравнением

$$\mu_n(\tau)=1-\tau\lambda_n=-\mu_1(\tau)=\tau\lambda_1-1,$$

которое дает

$$\tau_*=\frac{2}{\lambda_1+\lambda_n}<\tau_0. \tag{109}$$

В результате получаем:

$$\|S\|=\max_{1\le i\le n}|\mu_i(\tau)|=\begin{cases}\mu_n(\tau),0<\tau\le\tau_*\\-\mu_1(\tau),\tau_*\le\tau<\tau_0.\end{cases} \tag{110}$$

Свое наименьшее значение норма матрицы S достигает при $\tau=\tau_*$:

$$\min\|S\|=1-\tau_*\lambda_n=\frac{\lambda_1-\lambda_n}{\lambda_1+\lambda_n}=\frac{M_A-1}{M_A+1}. \tag{111}$$

Формула (111) показывает, что для плохо обусловленной матрицы даже при оптимальном выборе итерационного параметра $\tau=\tau_*$ норма матрицы S близка к единице, так что сходимость метода простой итерации в этом случае оказывается медленной.

В заключение заметим, что формула (108), определяющая границу интервала сходимости τ_0, и формула (109) для оптимального значения итерационного параметра τ_* представляют прежде всего теоретический интерес. Обычно при решении СЛАУ наибольшее и наименьшее характеристические числа матрицы A неизвестны, так что подсчитать величины τ_0 и τ_* заранее невозможно. В результате итерационный параметр τ нередко приходится подбирать прямо в процессе вычислений методом проб и ошибок.

Задача 2.

Рассмотреть систему двух уравнений с двумя неизвестными

$$\begin{cases}x_1+x_2=0,\\x_1+2x_2=1.\end{cases} \tag{112}$$

и построить для нее приближенное решение с помощью метода простой итерации.

Выпишем сразу решение системы (112)

$$x_1=-1,\ x_2=1, \tag{113}$$

чтобы потом иметь возможность сравнивать его с членами итерационной последовательности.

Перейдем к решению системы методом простой итерации. Матрица системы имеет вид

$$A=\begin{bmatrix}1&1\\1&2\end{bmatrix}.$$

Она самосопряженная и положительно определенная, поскольку

$$(A\mathbf{x},\mathbf{x})=(x_1+x_2)x_1+(x_1+2x_2)x_2=(x_1+x_2)^2+x_2^2>0.$$

Составим характеристическое уравнение для матрицы A и найдем его корни:

$$\begin{vmatrix}1-\lambda&1\\1&2-\lambda\end{vmatrix}=\lambda^2-3\lambda+1=0,$$

$$\lambda_1=\frac{3+\sqrt{5}}{2}\approx2.618,\ \lambda_2=\frac{3-\sqrt{5}}{2}\approx0.382$$

С их помощью можно определить границу интервала сходимости τ_0 и оптимальное значение итерационного параметра τ_*:

$$\tau_0=\frac{2}{\lambda_1}\approx0.764,\ \tau_*=\frac{2}{\lambda_1+\lambda_2}\approx0.745.$$

Для построения итерационной последовательности выберем какое-нибудь значение итерационного параметра на интервале сходимости, например, $\tau=1/2$. В этом случае рекуррентная формула для членов итерационной последовательности (102) принимает вид:

$$\mathbf{x}_{k+1}=S\mathbf{x}_k+\frac{1}{2}\mathbf{f},\ \text{где}\ S=\begin{bmatrix}1/2&-1/2\\-1/2&0\end{bmatrix}$$

Возьмем простейшее начальное приближение $\mathbf{x}_0=0$ и выпишем несколько первых членов итерационной последовательности $\mathbf{x}_k$, подсчитывая для каждого из них невязку $\boldsymbol{\psi}_k$ (71). В результате получим:

$$\mathbf{x}_1=\left\{0,\frac{1}{2}\right\},\ \boldsymbol{\psi}_1=\left\{\frac{1}{2},0\right\},\ \|\boldsymbol{\psi}_1\|=\frac{1}{2},$$

$$\mathbf{x}_2=\left\{-\frac{1}{4},\frac{1}{2}\right\},\ \boldsymbol{\psi}_2=\left\{\frac{1}{4},-\frac{1}{4}\right\},\ \|\boldsymbol{\psi}_2\|=\frac{1}{2\sqrt{2}},$$

$$\mathbf{x}_3=\left\{-\frac{3}{8},\frac{5}{8}\right\},\ \boldsymbol{\psi}_3=\left\{\frac{1}{4},-\frac{1}{8}\right\},\ \|\boldsymbol{\psi}_3\|=\frac{\sqrt{5}}{8},$$

$$\mathbf{x}_4=\left\{-\frac{1}{2},\frac{11}{16}\right\},\ \boldsymbol{\psi}_4=\left\{\frac{3}{16},-\frac{1}{8}\right\},\ \|\boldsymbol{\psi}_4\|=\frac{\sqrt{10}}{16}.$$

Норма невязок, хотя и медленно, но убывает, что говорит о сходимости процесса. Это же видно из сравнения членов итерационной последовательности $\mathbf{x}_k$ с решением системы (113). Медленная сходимость связана с плохой обусловленностью матрицы A:

$$M_A=\frac{\lambda_1}{\lambda_2}\approx6.854.$$

3.5. Неявные итерационные методы. Метод Зейделя.

Вернемся к общей записи итерационного стационарного процесса в канонической форме (65).

Рассмотрим произвольную квадратную матрицу:

$$A=\begin{bmatrix}a_{11}&a_{12}&\dots&a_{1m}\\a_{21}&a_{22}&\dots&a_{2m}\\\dots&\dots&\dots&\dots\\a_{m1}&a_{m2}&\dots&a_{mm}\end{bmatrix}.$$

Разложим её на сумму трех матриц

$$A=D+T_{\text{Л}}+T_{\text{П}}, \tag{114}$$

где D - диагональная часть матрицы A, которая содержит элементы a_{ii}, стоящие на главной диагонали:

$$D_{ij} = a_{ij}\delta_{ij} = \begin{cases} 0, & i \neq j \\ a_{ii}, & i = j \end{cases},$$

T_H - нижняя треугольная матрица

$$(T_H)_{ij} = \begin{cases} a_{ij}, & i > j \\ 0, & i \le j \end{cases},$$

T_B - верхняя треугольная матрица.

$$(T_B)_{ij} = \begin{cases} 0, & i \ge j \\ a_{ij}, & i < j \end{cases}.$$

В классическом методе Зейделя, записанном в канонической форме, полагают

$$B = D + T_H, \qquad \tau = 1. \tag{115}$$

В результате формула (65) принимает вид:

$$\left(D + T_H\right)\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right) + A\mathbf{x}_k = \mathbf{f},$$

или

$$\left(D + T_H\right)\mathbf{x}_{k+1} + T_B\mathbf{x}_k = \mathbf{f}. \tag{116}$$

Перейдем от векторной формы записи рекуррентной формулы (116) к построчной:

$$\begin{aligned} a_{11}x_1^{k+1} + a_{12}x_2^k + a_{13}x_3^k + \cdots + a_{1n}x_n^k &= f_1 \\ a_{21}x_1^{k+1} + a_{22}x_2^{k+1} + a_{23}x_3^k + \cdots + a_{2n}x_n^k &= f_2 \\ \vdots \qquad\qquad & \vdots \\ a_{n1}x_1^{k+1} + a_{n2}x_2^{k+1} + a_{n3}x_3^{k+1} + \cdots + a_{nn}x_n^{k+1} &= f_n. \end{aligned} \tag{117}$$

Уравнения (117) позволяют последовательно рассчитать компоненты вектора $(k+1)$ - ой итерации подобно тому, как это делалось во время обратного хода в методе Гаусса:

$$x_i^{k+1} = \frac{1}{a_{ii}}\left[f_i - \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}x_j^k\right], \; i = 1,\ldots,n. \tag{118}$$

Формула (118) предполагает, что $a_{ii} \neq 0$, $1 \le i \le n$. Если матрица A удовлетворяет условиям теоремы Самарского (84): $A = A^* > 0$, то, согласно неравенству (81), все ее диагональные элементы должны быть строго положительными и, тем самым, не могут обращаться в ноль.

Алгоритм в методе Зейделя прост и удобен для вычислений. Он не требует никаких действий с матрицей A. Ранее вычисленные на текущей итерации компоненты $x_j^{k+1}(j<i)$ сразу же участвуют в расчетах наряду с компонентами $x_j^k(j>i)$ и, таким образом, не требуют дополнительного резерва памяти, что существенно при решении больших систем.

Сходимость метода Зейделя в случае, когда матрица A удовлетворяет условию теоремы Самарского, т.е. является самосопряженной и положительно определенной, будет доказана в следующем разделе. К этому утверждению добавим без доказательства еще один результат: метод Зейделя сходится для любой системы (62), в которой матрица A обладает свойством диагонального преобладания.

Задача 3.
Рассмотреть систему (112) (см. задачу 2) и построить для нее приближенное решение с помощью метода Зейделя.

В рассматриваемом случае рекуррентные формулы (118) для построения $(k+1)$ - ой итерации по k -ой итерации принимают вид:

$$\begin{aligned} x_1^{k+1} &= -x_2^k \\ x_2^{k+1} &= \frac{1}{2}\left(1 - x_1^{k+1}\right). \end{aligned} \tag{119}$$

Принимая, как и при решении задачи 2, за начальное приближение нулевой вектор, подсчитаем по формулам (119) несколько первых итераций, сопровождая этот процесс подсчетом невязки:

$$\mathbf{x}_1 = \left\{0, \frac{1}{2}\right\}, \; \boldsymbol{\psi}_1 = \left\{\frac{1}{2}, 0\right\}, \; \|\boldsymbol{\psi}_1\| = \frac{1}{2},$$

$$\mathbf{x}_2 = \left\{-\frac{1}{2}, \frac{3}{4}\right\}, \; \boldsymbol{\psi}_2 = \left\{\frac{1}{4}, 0\right\}, \; \|\boldsymbol{\psi}_2\| = \frac{1}{4},$$

$$\mathbf{x}_3 = \left\{-\frac{3}{4}, \frac{7}{8}\right\}, \; \boldsymbol{\psi}_3 = \left\{\frac{1}{8}, 0\right\}, \; \|\boldsymbol{\psi}_3\| = \frac{1}{8}.$$

Обсудим полученные результаты. Начнем с невязки. Ее вторая компонента все время остается равной нулю, поскольку второе уравнение системы на каждой итерации выполняется, как видно из (119), точно. Первые компоненты невязки и норма убывают по закону геометрической прогрессии с знаменателем $1/2$, т.е. гораздо быстрее, чем в методе простой итерации. Хорошая сходимость процесса видна также из прямого сравнения членов итерационной последовательности $\mathbf{x}_k$ с точным решением системы $\mathbf{x} = \{-1, 1\}$.

3.6. Метод верхней релаксации

Модифицируем метод Зейделя. С этой целью введем параметр ω и запишем рекуррентное соотношение (65) в виде

$$\left(D + \omega T_H\right)\frac{\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right)}{\omega} + A\mathbf{x}_k = \mathbf{f}. \tag{120}$$

В данном случае

$$B = D + \omega T_H, \; \tau = \omega > 0. \tag{121}$$

При $\omega = 1$ мы возвращаемся к методу Зейделя.

Соотношению (120) можно придать вид

$$\left(\frac{1}{\omega}D + T_H\right)\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right) + A\mathbf{x}_k = \mathbf{f}. \tag{122}$$

Такая форма записи показывает, что параметр ω влияет на диагональ матрицы B.

Для построения алгоритма вычисления очередной итерации нужно разделить в левой части рекуррентной формулы (122) члены, содержащие $\mathbf{x}_k$ и $\mathbf{x}_{k+1}$, и придать ей форму, аналогичную (116):

$$\left(\frac{1}{\omega}D + T_H\right)\mathbf{x}_{k+1} + \left[\left(1 - \frac{1}{\omega}\right)D + T_B\right]\mathbf{x}_k = \mathbf{f}. \tag{123}$$

Если перейти от векторной записи к записи типа (117) в виде отдельных уравнений, то можно получить для компонент x_i^{k+1} очередной итерации формулы, структурно похожие на (118):

$$x_i^{k+1} = x_i^k + \frac{\omega}{a_{ii}}\left(f_i - \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i}^{n} a_{ij}x_j^k\right), \; i = 1,\ldots,n. \tag{124}$$

Исследуем условия сходимости метода верхней релаксации при дополнительном предположении, что матрица A удовлетворяет условиям теоремы Самарского (84). Самосопряженность матрицы A означает, что $T_H^* = T_B$, $T_B^* = T_H$. Отсюда следует

$$\left(T_H\mathbf{x}, \mathbf{x}\right) = \left(T_B^*\mathbf{x}, \mathbf{x}\right) = \left(T_B\mathbf{x}, \mathbf{x}\right). \tag{125}$$

Составим для рассматриваемого случая матрицу $B - \frac{\tau}{2}A$. Согласно (121)

$$B - \frac{\tau}{2}A = \left(D + \omega T_H\right) - \frac{\omega}{2}\left(D + T_H + T_B\right) = \left(1 - \frac{\omega}{2}\right)D + \frac{\omega}{2}\left(T_H - T_B\right). \tag{126}$$

Запишем условие ее положительной определенности

$$\left(\left(B - \frac{\tau}{2}A\right)\mathbf{x}, \mathbf{x}\right) = \left(1 - \frac{\omega}{2}\right)\left(D\mathbf{x}, \mathbf{x}\right) > 0. \tag{127}$$

Второе слагаемое в выражении (126) не дает вклада в квадратичную форму (127) в силу соотношения (125).

Матрица A является, по предположению, положительно определенной. Следовательно, все ее диагональные элементы строго положительны: $a_{ii} > 0$, $1 \le i \le n$. Это означает положительную определенность матрицы D: $\left(D\mathbf{x}, \mathbf{x}\right) > 0$. В результате знак выражения (127) определяется знаком первого множителя, так что достаточное условие для сходимости итерационной последовательности метода верхней релаксации принимает вид:

$$0 < \omega < 2 \tag{128}$$

Метод Зейделя, соответствующий случаю $\omega = 1$, удовлетворяет этому условию.

Можно поставить вопрос об оптимальном выборе параметра $\omega = \omega_*$, при котором метод сходится быстрее всего. Теоретическое исследование, на котором мы не будем останавливаться, показывает, что такое значение существует и может быть выражено через наибольшее и наименьшее собственные значение матрицы A. Однако на практике его приходится подбирать экспериментально методом проб и ошибок, поскольку найти $\lambda_{\min}$ и $\lambda_{\max}$ с достаточной точностью удается в редких случаях.

Задача 4
Построить приближенное решение системы (112) методом верхней релаксации, полагая $\omega = 4/3$.

Выпишем для рассматриваемого случая матрицы $\frac{1}{\omega}D + T_H$ и $\left(1 - \frac{1}{\omega}\right)D + T_B$, определяющие итерационный процесс:

$$\frac{3}{4}D + T_H = \begin{bmatrix} 3/4 & 0 \\ 1 & 3/2 \end{bmatrix}, \; \frac{1}{4}D + T_B = \begin{bmatrix} 1/4 & 1 \\ 0 & 1/2 \end{bmatrix}.$$

С их помощью рекуррентное соотношение (123), записанное покомпонентно, принимает вид:

$$\begin{aligned} \frac{3}{4}x_1^{k+1} + \frac{1}{4}x_1^k + x_2^k &= 0, \\ x_1^{k+1} + \frac{3}{2}x_2^{k+1} + \frac{1}{2}x_2^k &= 1. \end{aligned}$$

Выражая из первого соотношения x_1^{k+1}, из второго x_2^{k+1}, получим окончательные расчетные формулы для компонент очередной итерации:

$$\begin{aligned} x_1^{k+1} &= -\frac{1}{3}x_1^k - \frac{4}{3}x_2^k, \\ x_2^{k+1} &= \frac{2}{3} - \frac{2}{3}x_1^{k+1} - \frac{1}{3}x_2^k. \end{aligned}$$

Примем, как и в предыдущих случаях, за начальное приближение нулевой вектор и сделаем три итерации. При этом для каждой из них подсчитаем невязку (71), позволяющую следить за сходимостью процесса

$$\mathbf{x}_1 = \left\{0, \frac{2}{3}\right\}, \; \boldsymbol{\psi}_1 = \left\{\frac{2}{3}, \frac{1}{3}\right\}, \; \|\boldsymbol{\psi}_1\| = \frac{\sqrt{5}}{3} \approx 0.745,$$

$$\mathbf{x}_2 = \left\{-\frac{8}{9}, \frac{28}{27}\right\}, \; \boldsymbol{\psi}_2 = \left\{\frac{4}{27}, \frac{5}{27}\right\}, \; \|\boldsymbol{\psi}_2\| = \frac{\sqrt{41}}{27} \approx 0.237,$$

$$\mathbf{x}_3 = \left\{-\frac{88}{81}, \frac{256}{243}\right\}, \; \boldsymbol{\psi}_3 = \left\{-\frac{8}{243}, \frac{5}{243}\right\}, \; \|\boldsymbol{\psi}_3\| = \frac{\sqrt{89}}{243} \approx 0.039.$$

Поведение невязок, а также сравнение членов итерационной последовательности $\mathbf{x}_k$ с точным решением системы $\mathbf{x} = \{-1, 1\}$ показывают сходимость процесса, более быструю, чем в методе Зейделя. Выбранное значение параметра $\omega = 4/3$ оказалось близким к оптимальному $\omega = \omega_*$.

Глава 2. ПРИБЛИЖЕНИЕ ФУНКЦИЙ.

Пусть на отрезке $[a,b]$ определена некоторая функция $y = f(x)$, однако полная информация о ней недоступна. Известны лишь ее значения в конечном числе точек $x_0, x_1, \ldots x_n$, этого отрезка, которые мы будем считать занумерованными в порядке возрастания:

$$a \le x_0 < x_1 < \ldots < x_i < x_{i+1} < \ldots x_n \le b. \tag{1}$$

Требуется по известным значениям

$$y_i = f(x_i), \; i = 0,1,\ldots,n \tag{2}$$

«восстановить», хотя бы приближенно, исходную функцию $y = f(x)$, то есть построить на отрезке $[a,b]$ функцию $F(x)$, достаточно близкую к $f(x)$. Функцию $F(x)$ принято называть интерполирующей функцией, точки $x = x_0, x = x_1, \ldots x = x_n$ - узлами интерполяции.

Подобные задачи часто возникают на практике, например, при обработке экспериментальных данных, когда значения переменной y, зависящей от x, измеряется в конечном числе точек x_i: $y_i = f(x_i)$, $(i = 0,1,\ldots,n)$ или при работе с табличными функциями, если требуется вычислить $y = f(x)$, при значениях аргумента , не совпадающего ни с одним из табличных x_i.

Поставленный выше в общей форме вопрос о приближении функций является достаточно сложным. Существует не один подход к его решению. Мы ограничимся изложением трех наиболее распространенных методов.

§1. Интерполирование.

1.1. Классическая постановка задачи интерполирования.

Выберем некоторую систему функций $\varphi_0(x), \varphi_1(x), \ldots \varphi_n(x)$, заданных на отрезке $[a,b]$, и будем строить $F(x)$ как их линейную комбинацию:

$$F(x) = \sum_{i=0}^{n} c_i\varphi_i(x), \tag{3}$$

где числовые коэффициенты c_i $(i = 0,1,\ldots n)$ подлежат определению, согласно условиям:

$$F(x_j) = f(x_j), \; j = 0,1,\ldots,n. \tag{4}$$

Равенства (4) представляют собой систему линейных алгебраических уравнений относительно коэффициентов c_i:

$$\sum_{i=0}^{n} c_i\varphi_i(x_j) = f(x_j), \; j = 0,1,\ldots,n$$

или в развернутом виде:

$$\begin{cases} c_0\varphi_0(x_0) + c_1\varphi_1(x_0) + \ldots + c_n\varphi_n(x_0) = f(x_0) \\ c_0\varphi_0(x_1) + c_1\varphi_1(x_1) + \ldots + c_n\varphi_n(x_1) = f(x_1) \\ \vdots \\ c_0\varphi_0(x_n) + c_1\varphi_1(x_n) + \ldots + c_n\varphi_n(x_n) = f(x_n) \end{cases}. \tag{5}$$

Для того, чтобы коэффициенты c_i $(i = 0,1,\ldots n)$ можно было определить и притом единственным образом, необходимо и достаточно, чтобы определитель полученной системы линейных уравнений был отличен от нуля:

$$\Delta = \begin{vmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \ldots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_n(x_1) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \ldots & \varphi_n(x_n) \end{vmatrix} \neq 0. \tag{6}$$

Определение.
Система функций $\varphi_i(x)$ $(i = 0,1,\ldots n)$, *удовлетворяющая при фиксированных значениях* x_j $(j = 0,1,\ldots n)$ *условию (6), называется Чебышевской.*

Очевидно, что для однозначной разрешимости задачи интерполирования в классической постановке необходимо и достаточно, чтобы система функций $\varphi_i(x)$ $(i = 0,1,\ldots n)$ была Чебышевской. Только такие системы функций мы и будем использовать в этой главе. Необходимым условием принадлежности системы функций $\varphi_i(x)(i = 0,1,\ldots n)$ к Чебышевской является их линейная независимость. Однако это условие не является достаточным. Например, для системы из двух линейно независимых функций $\varphi_0(x) = \sin x$, $\varphi_1(x) = \cos x$, с узлами интерполяции $x_0 = 0$, $x_1 = \pi$, определитель

$$\Delta = \begin{vmatrix} \sin x_0 & \cos x_0 \\ \sin x_1 & \cos x_1 \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 0 & -1 \end{vmatrix} = 0$$

и данная система функций при выбранных значениях x_0 и x_1 не является Чебышевской.

1.2. Интерполирование полиномами.

При построении интерполирующей функции $F(x)$ в виде (3) функции $\varphi_i(x)$, естественно, выбираются такими, чтобы их вычисление было простым. В частности, широкое распространение получило интерполирование с помощью степенных функций:

$$\varphi_0(x) = 1; \;\; \varphi_1(x) = x; \;\; \varphi_2(x) = x^2, \;\; \ldots \;\; \varphi_n(x) = x^n.$$

В этом случае интерполирующая функция представляет собой полином степени n:

$$F(x) = P_n(x) = \sum_{i=0}^{n} c_i x^i \tag{7}$$

с неизвестными коэффициентами c_i $(i = 0,1,\ldots n)$.

Согласно рассмотренной выше общей схеме построения интерполирующей функции, следует потребовать, чтобы коэффициенты c_i с учетом (7) удовлетворяли системе линейных уравнений:

$$\sum_{i=0}^{n} c_i x_j^i = f(x_j), \; j = 0,1,\ldots,n. \tag{8}$$

Определителем этой системы является определитель Ван-дер- Монда:

$$\Delta = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix} = \prod_{i>j} (x_i - x_j).$$

В нашем случае этот определитель отличен от нуля, поскольку, согласно (1), все узлы интерполирования различны между собой. Итак, интерполирование с помощью полиномов при сделанных в начале главы предположениях всегда осуществимо и притом единственным образом.

Задача 1.
Построить линейный полином

$$P_1(x) = c_0 + c_1 x$$

по заданным узлам интерполяции $x_1 < x_1$ *и соответствующим им значениям функции* $y_0 = f(x_0)$ *и* $y_1 = f(x_1)$.

Линейная система уравнений для определения c_0 и c_1 в данном случае имеет вид:

$$\begin{aligned} c_0 + c_1 x_0 &= f(x_0), \\ c_0 + c_1 x_1 &= f(x_1). \end{aligned}$$

Определитель этой системы равен $\Delta = x_1 - x_0 \neq 0$. Решив систему, получим:

$$c_0 = \frac{x_1 f(x_0) - x_0 f(x_1)}{x_1 - x_0}; \quad c_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Следовательно,

$$P_1(x) = \frac{x_1 f(x_0) - x_0 f(x_1)}{x_1 - x_0} + \frac{f(x_1) - f(x_0)}{x_1 - x_0}x. \tag{9}$$

Перепишем этот полином в несколько другой форме, выделяя $f(x_0)$ и $f(x_1)$ в качестве множителей

$$P_1(x) = f(x_0)\frac{x - x_1}{x_0 - x_1} + f(x_1)\frac{x - x_0}{x_1 - x_0}. \tag{10}$$

Геометрический образ интерполирующей функции $P_1(x)$ - прямая, проходящая на плоскости (x,y) через точки с координатами (x_0, y_0) и (x_1, y_1). Уравнение этой прямой, наряду с (9) и (10), можно переписать в виде:

$$y = P_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0). \tag{11}$$

Из данного примера видно, что всегда существуют различные эквивалентные между собой формы записи интерполяционного полинома, удобные в различных ситуациях.

1.3. Построение интерполяционного полинома в форме Лагранжа.

Интерполяционный полином первой степени (9) мы построили, решая напрямую систему двух уравнений с двумя неизвестными - коэффициентами c_0 и c_1. Однако решить таким же образом систему (8) при произвольном n технически очень сложно. Проще сделать это с помощью специальных методов, учитывающих особенности рассматриваемой задачи. Один из таких методов, принадлежащих Лагранжу, мы и рассмотрим в этом разделе.

Представим искомый полином $P_n(x)$ в виде:

$$P_n(x) = \sum_{i=0}^{n} f(x_i)Q_{n,i}(x), \tag{12}$$

где $Q_{n,i}(x)$ полиномы степени n, «ориентированные» на точки x_i в том смысле, что

$$Q_{n,i}(x) = \begin{cases} 0, & x = x_j \;\; \forall j \neq i, \\ 1, & x = x_i. \end{cases} \tag{13}$$

Такие полиномы легко построить:

$$Q_{n,i}(x) = \prod_{\substack{j=0 \\ j \neq i}}^{j=n} \frac{(x - x_j)}{(x_i - x_j)} \tag{14}$$

или в развернутом виде:

$$\begin{aligned} Q_{n,0}(x) &= \frac{(x - x_1)(x - x_2)\ldots(x - x_n)}{(x_0 - x_1)(x_0 - x_2)\ldots(x_0 - x_n)}, \\ Q_{n,i}(x) &= \frac{(x - x_0)\ldots(x - x_{i-1})(x - x_{i+1})\ldots(x - x_n)}{(x_i - x_0)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)}, \\ Q_{n,n}(x) &= \frac{(x - x_0)(x - x_1)\ldots(x - x_{n-1})}{(x_n - x_0)(x_n - x_1)\ldots(x_n - x_{n-1})}. \end{aligned} \tag{15}$$

Иногда нам будет удобно записывать $Q_{n,i}(x)$ в виде:

$$Q_{n,i}(x) = \frac{(x - x_0)\ldots[i]\ldots(x - x_n)}{(x_i - x_0)\ldots[i]\ldots(x_i - x_n)}.$$

Из выражения (12) и формул (13) очевидно, что построенный полином $P_n(x)$ действительно является интерполяционным полиномом для функции $y = f(x)$ на сетке с узлами $x_0, x_1, \ldots x_n$. Его принято называть интерполяционным полиномом в форме Лагранжа. Этим подчеркивается, что возможны и другие эквивалентные представления интерполяционного полинома $P_n(x)$. С одним из них мы познакомимся в следующем разделе.

В заключение отметим, что из трех различных представлений интерполяционного полинома первой степени (9)- (11) формула (10) дает его запись в форме Лагранжа.

Задача 2.

Написать интерполяционный полином второй степени для функции $y = \sin x$ *по ее значениям в трех точках:* $x_0 = 0$, $x_1 = \pi/6$, $x_2 = \pi/2$. *Вычислить с помощью этого полинома приближенное значение синуса в точке* $x = \pi/4$, *сравнить полученный результат с точным значением синуса и подсчитать погрешность* $R_2\left(\frac{\pi}{4}\right)$.

Воспользуемся для записи полинома формулой Лагранжа (12). В рассматриваемом случае $y_0 = \sin x_0 = 0$, так что в формуле останется только два слагаемых соответствующих точкам x_1 и x_2. В результате получим:

$$P_2(x) = \frac{1}{2}\frac{x\left(x - \frac{\pi}{2}\right)}{\frac{\pi}{6}\left(-\frac{\pi}{3}\right)} + \frac{x\left(x - \frac{\pi}{6}\right)}{\frac{\pi}{2}\left(\frac{\pi}{3}\right)} = \frac{x}{\pi^2}\left(\frac{7\pi}{2} - 3x\right) \tag{16}$$

Перейдем к выполнению второй части задания. Вычислим с помощью интерполяционного полинома (16) приближенное значения синуса в точке $x = \pi/4$ и подсчитаем погрешность:

$$P_2\left(\frac{\pi}{4}\right) = \frac{11}{16}, \; R_2\left(\frac{\pi}{4}\right) = \frac{1}{\sqrt{2}} - \frac{11}{16} = 0.0197 < 0.02. \tag{17}$$

На рис. 1 приведены для сравнения графики функций $\sin x$ (сплошная линия) и $P_2(x)$ (пунктир).

1.4. Интерполяционный полином в форме Ньютона.

Интерполяционный полином в форме Лагранжа, несмотря на своё изящество, неудобен для вычислений тем, что при увеличении числа узлов интерполяции приходится перестраивать весь полином заново.

Перепишем интерполяционный полином Лагранжа в иной, эквивалентной форме

$$P_n(x) = P_0(x) + \sum_{i=1}^{n}(P_i(x) - P_{i-1}(x)), \tag{18}$$

где $P_i(x)$ - полиномы Лагранжа степени $i \le n$, соответствующие узлам интерполирования $x_0, x_1, \ldots x_i$. В частности, $P_0(x) = f(x_0)$ - полином нулевой степени.

Полином

$$Q_i(x) = P_i(x) - P_{i-1}(x) \tag{19}$$

имеет степень i и по построению обращается в ноль при $x = x_0, x = x_1, \ldots x = x_{i-1}$, поэтому его можно представить в виде

$$Q_i(x) = A_i(x - x_0)(x - x_1)\ldots(x - x_{i-1}), \tag{20}$$

где A_i - числовой коэффициент при x^i. Поскольку $P_{i-1}(x)$ не содержит степени i, то A_i просто совпадает с коэффициентом при x^i в полиноме $P_i(x)$. Согласно (12) и (15) его можно записать в виде

$$A_i = \sum_{k=0}^{i} \frac{f(x_k)}{\omega_{k,i}}, \tag{21}$$

где

$$\omega_{k,i} = (x_k - x_0)\ldots(x_k - x_{k-1})(x_k - x_{k+1})\ldots(x_k - x_i). \quad (22)$$

При этом

$$A_0 = f(x_0). \quad (23)$$

Формулы (19) и (21) позволяют написать рекуррентное соотношение для полинома $P_n(x)$:

$$P_n(x) = P_{n-1}(x) + A_n(x - x_0)\ldots(x - x_{n-1}). \quad (24)$$

Выражая аналогичным образом по индукции $P_{n-1}(x)$ через $P_{n-2}(x)$, $P_{n-2}(x)$ через $P_{n-3}(x)$ и т. д., получим окончательную формулу для полинома $P_n(x)$:

$$P_n(x) = A_0 + A_1(x - x_0) + A_2(x - x_0)(x - x_1) + \ldots + A_n(x - x_0)\ldots(x - x_{n-1}). \quad (25)$$

Представление (25) удобно для вычислителя, поскольку увеличение n на единицу требует только добавления к «старому» многочлену одного дополнительного слагаемого. Такое представление интерполяционного полинома $P_n(x)$ называют интерполяционным полиномом в форме Ньютона.

Из трех эквивалентных представлений интерполяционного полинома первой степени (9) - (11) формула (11) дает его запись в форме Ньютона.

Задача 3.

Написать интерполяционный полином второй степени в форме Ньютона для функции $y = \sin x$ по ее значениям в трех точках: $x_0 = 0$, $x_1 = \pi/6$, $x_2 = \pi/2$ (см. задачу 2).

Согласно формуле (25)

$$P_2(x) = A_0 + A_1 x + A_2 x\left(x - \frac{\pi}{6}\right). \quad (26)$$

Коэффициенты в этом разложении вычисляются по формулам (21) и(23):

$$A_0 = 0,\ A_1 = \frac{1}{2}\left(\frac{6}{\pi}\right),\ A_2 = -\frac{18}{2\pi^2} + \frac{6}{\pi^2} = -\frac{3}{\pi^2}. \quad (27)$$

Подставляя найденные значения коэффициентов в формулу (26), получим

$$P_2(x) = \frac{3}{\pi}x - \frac{3x}{\pi^2}\left(x - \frac{\pi}{6}\right) = \frac{x}{\pi^2}\left(\frac{7\pi}{2} - 3x\right). \quad (28)$$

Первоначальные выражения для интерполяционного полинома в форме Лагранжа и Ньютона различны, но окончательные ответы, естественно, совпадают.

1.5. Погрешность интерполирования.

Поставим вопрос о том, насколько хорошо интерполяционный полином $P_n(x)$ приближает функцию $f(x)$ на отрезке $[a,b]$, то есть попытаемся оценить погрешность (остаточный член)

$$R_n(x) = f(x) - P_n(x),\ x \in [a,b]. \quad (29)$$

Сразу же отметим, что по определению интерполяционного полинома

$$R_n(x_i) = 0 \text{ при } i = 0,1,\ldots,n, \quad (30)$$

поэтому речь идет об оценке $R_n(x)$ при значениях $x \neq x_i$.

Для того, чтобы это сделать, следует ввести дополнительно предположение о гладкости функции $f(x)$. Предположим, что $f(x)$ имеет $(n+1)$ непрерывную производную на отрезке $[a,b]$.

В силу (30) $R_n(x)$ можно представить в виде:

$$R_n(x) = \omega_{n+1}(x)r_n(x), \quad (31)$$

где $\omega_{n+1}(x)$ - полином степени $(n+1)$:

$$\omega_{n+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_n). \quad (32)$$

Зафиксируем произвольное значение $x \in [a,b]$ и рассмотрим вспомогательную функцию от переменной t:

$$g(t) = f(t) - P_n(t) - \omega_{n+1}(t)r_n(x),$$

заданную на отрезке $[a,b]$ и содержащую переменную x в качестве параметра. В силу своего определения функция $g(t)$ обязана обращаться в нуль в узлах интерполирования при $t = x_i$ и кроме того при $t = x$, т. е. как функция аргумента t она имеет $(n+2)$ нуля:

$$g(x_i) = 0,\ i = 0,1,\ldots,n,\ g(x) = 0. \quad (33)$$

Если $x \in [x_0, x_n]$, то все ее нули также лежат на отрезке $[x_0, x_n]$. Если $x < x_0$, то эти нули, вообще говоря, принадлежат отрезку $[x, x_n]$, а если $x > x_n$, то они находятся на отрезке $[x_0, x]$. Объединяя эти три случая, скажем, что указанные нули функции $g(t)$ принадлежат отрезку $[\alpha,\beta]$, где $\alpha = \min(x_0, x) \geq a$, $\beta = \max(x_n, x) \leq b$.

Согласно известной теореме Ролля можно утверждать, что производная $g'(t)$ имеет по крайней мере $(n+1)$ нуль на отрезке $[\alpha,\beta]$ (эти нули перемежаются с нулями самой функции $g(t)$). Повторяя это рассуждение, заключаем, что $g''(t)$ имеет по крайней мере n нулей на отрезке $[\alpha,\beta]$, $g''(t)$ - $(n-1)$ нуль и, наконец, $g^{(n+1)}(t)$ обращается хотя бы один раз в нуль в некоторой точке $t = \xi \in [\alpha,\beta]$, то есть

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

Учитывая, что $(n+1)$ производная полинома степени n тождественно равна нулю, получаем, что

$$r_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!},\ \xi \in [\alpha,\beta] \quad (34)$$

и соответственно

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x). \quad (35)$$

Формула (35) не позволяет вычислить погрешность, поскольку точное значение аргумента ξ нам неизвестно. Однако с ее помощью погрешность можно оценить:

$$|R_n(x)| \leq \frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)|, \quad (36)$$

где

$$M_{n+1} = \max_{x\in[\alpha,\beta]}|f^{(n+1)}(x)| \leq \max_{x\in[a,b]}|f^{(n+1)}(x)|. \quad (37)$$

Обсудим роль полинома $\omega_{n+1}(x)$ (32) в оценке (36). На отрезке $[x_0, x_n]$ он имеет $(n+1)$ нуль, а его значения между этими нулями сравнительно невелики, но, когда точка x выходит за пределы отрезка $[x_0, x_n]$ и удаляется от точки x_0 влево или от точки x_n вправо, оценка (36) ухудшается из-за быстрого роста функции $|\omega_{n+1}(x)|$. Это хорошо видно на рис. 2, где в качестве примера приведен график функции $\omega_4(x)$ с корнями $x_0 = -3/2$, $x_1 = -1/2$, $x_2 = 1/2$, $x_3 = 3/2$:

$$\omega_4(x) = \left(x^2 - \frac{9}{4}\right)\left(x^2 - \frac{1}{4}\right).$$

Ее наибольшее по модулю значение на отрезке $\left[-\frac{3}{2}, \frac{3}{2}\right]$ равно единице. Однако уже в точках $x = \pm 2$ за пределами отрезка полином $\omega_4(x)$ принимает значение

$$\omega_4(\pm 2) = \frac{105}{16} = 6.5625.$$

Из сказанного можно сделать следующий вывод. Если $x \in [x_0, x_n]$, то множитель $|\omega_{n+1}(x)|$ не обесценивает оценку (36). Такой случай называют собственно интерполяцией $f(x)$. Противоположный случай, когда точка x лежит вне отрезка называют экстраполяцией функции $f(x)$. Отмеченная выше особенность поведения полинома $\omega_{n+1}(x)$ резко ухудшает оценку (36) при экстраполяции. Поэтому на практике экстраполяции избегают или ограничиваются многочленами невысокой степени $(n = 1,2)$, когда рост функции $|\omega_{n+1}(x)|$ не настолько критичен.

Задача 4.

Написать мажорантную оценку для погрешности (36) при вычислении приближенного значения $\sin x$ в точке $x = \pi/4$ с помощью интерполяционного полинома второй степени $P_2(x)$ (16). Сравнить ее с погрешностью (17), подсчитанной непосредственно.

Формула для погрешности (35) принимает в данном случае вид:

$$R_2\left(\frac{\pi}{4}\right) = \frac{1}{6}(-\cos\xi)\omega_3\left(\frac{\pi}{4}\right) = \cos\xi\frac{\pi^3}{1152},\ 0 \leq \xi \leq \frac{\pi}{2}.$$

Она правильно определяет знак погрешности, но не позволяет вычислить ее величину, поскольку значение аргумента ξ неизвестно. Чтобы получить мажорантную оценку погрешности (36), нужно заменить $\cos\xi$ на его наибольшее значение – единицу. В результате будем иметь:

$$R_2\left(\frac{\pi}{4}\right) \leq \frac{\pi^3}{1152} < 0.027.$$

Эта оценка согласуется с величиной погрешности (17), вычисленной «в лоб».

1.6. О сходимости интерполяционного процесса.

Поставим вопрос, будут ли сходится интерполяционные полиномы $P_n(x)$ к интерполируемой функции $f(x)$ на отрезке $[a,b]$ при неограниченном возрастании числа узлов n.

Упорядоченное множество точек x_i, $i = 0,1,\ldots n$ (1) назовем сеткой на отрезке $[a,b]$ и обозначим для краткости Ω_n. Рассмотрим последовательность сеток с возрастающим числом узлов:

$$\Omega_0 = \{x_0^{(0)}\},\ \Omega_1 = \{x_0^{(1)}, x_1^{(1)}\}, \ldots, \Omega_n = \{x_0^{(n)}, x_1^{(n)} \ldots x_n^{(n)}\}, \ldots$$

и отвечающую ей последовательность интерполяционных полиномов $P_n(x)$, построенных для фиксированной непрерывной на отрезке $[a,b]$ функции $f(x)$.

Интерполяционный процесс для функции сходится в точке $x_* \in [a,b]$, если существует предел

$$\lim_{n\to\infty} P_n(x_*) = f(x_*).$$

Наряду с обычной сходимостью часто рассматривается сходимость в различных нормах. Так, равномерная сходимость на отрезке $[a,b]$ означает, что

$$\max_{x\in[a,b]}|f(x) - P_n(x)| \to 0 \text{ при } n \to \infty.$$

Сходимость или расходимость интерполяционного процесса зависит как от выбора последовательности сеток, так и от гладкости функции $f(x)$. Если $f(x)$ - целая аналитическая функция, то при произвольном расположении узлов на отрезке $[a,b]$ интерполяционный многочлен $P_n(x)$ равномерно сходится к $f(x)$ при $n \to \infty$.

Положение резко меняется, если производные функции разрывны или не существуют в отдельных точках. Например для функции $f(x) = |x|$ на отрезке $[-1,1]$, покрытом равномерной сеткой узлов, значения $P_n(x)$ между узлами интерполяции неограниченно возрастают при $n \to \infty$. Вместе с тем, для заданной непрерывной функции $f(x)$ за счет выбора сеток можно добиться сходимости и притом равномерной на $[a,b]$. Однако построение таких сеток довольно сложно и, главное, такие сетки «индивидуальны» для каждой конкретной функции.

Если заметить дополнительно, что объем вычислений при построении интерполяционного полинома быстро нарастает с ростом n, то становится понятно, что на практике вычислители избегают пользоваться интерполяционными полиномами высокой степени. Вместо этого, в случае необходимости, при больших значениях n используется кусочно-полиномиальная интерполяция, которую мы обсудим в следующем параграфе.

1.7. Интерполяционный полином Эрмита.

Расширим постановку задачи об интерполяции. Ранее полагалось, что в узлах интерполяции заданы только значения функции $f(x)$. Пусть теперь в узлах $x_k \in [a,b]$, $k = 0,1,\ldots m$, среди которых нет совпадающих, заданы значения функции $f(x_k)$, и её производных $f^{(i)}(x_k)$, $i = 1,2,\ldots,N_k - 1$ до $(N_k - 1)$-го порядка включительно. Числа N_k при этом называют кратностью узла x_k. В каждой точке x_k, таким образом, задано N_k величин:

$$f(x_k), f'(x_k), \ldots f^{(N_k-1)}(x_k).$$

В общей сложности на всей совокупности узлов $x_0, x_1, \ldots, x_m$ известно $N_0 + N_1 + \ldots + N_m$ величин, что дает возможность ставить вопрос о построении полинома $H_n(x)$ степени

$$n = N_0 + \ldots + N_m - 1, \quad (38)$$

удовлетворяющего требованиям:

$$H_n^{(i)}(x_k) = f^{(i)}(x_k),\ k = 0,1,\ldots,m,\ i = 0,1,\ldots,N_k - 1. \quad (39)$$

Такой полином называется интерполяционным полиномом Эрмита для функции $f(x)$. Рассмотренный ранее вариант построения интерполяционного полинома $P_n(x)$ по известным значениям функции $f(x)$ в узлах интерполяции является частным случаем построения полинома Эрмита при условии, что все узлы простые: $N_k = 1$, $k = 0,1,\ldots,m$.

Докажем, что интерполяционный полином Эрмита существует и является единственным. Представим его в стандартном виде

$$H_n(x) = a_0 + a_1 x + \ldots + a_n x^n.$$

Наше утверждение будет справедливо, если показать, что коэффициенты $a_0, a_1, \ldots, a_n$ определяются из условий (39) и притом единственным образом. Условия (39) представляют собой систему линейных алгебраических уравнений относительно этих коэффициентов, причем число уравнений и число неизвестных равны $N_0 + N_1 + \ldots + N_m = n+1$.

Рассмотрим соответствующую однородную систему

$$\overline{H}_n^{(i)}(x_k) = 0,\ k = 0,1,\ldots,m,\ i = 0,1,\ldots,N_k - 1. \quad (40)$$

Уравнения (40) просто указывают на то, что числа x_k являются корнями полинома $\overline{H}_n(x)$ кратности N_k. Мы видим, таким образом, что полином $\overline{H}_n(x)$ имеет, с учетом кратности, не менее $N_0 + N_1 + \ldots + N_m = n+1$ корней. Поскольку его степень равна n, то он должен тождественно равняться нулю. Это означает, что $\overline{a}_0 = \overline{a}_1 = \ldots = \overline{a}_n = 0$, т.е. однородная система уравнений (40) имеет только тривиальное решение. Отсюда следует, что неоднородная система (39) при любой правой части разрешима и при том единственным образом.

Исследование погрешности интерполирования полинома Эрмита $R_n(x) = f(x) - H_n(x)$ почти дословно повторяет проведенное ранее исследование для

полинома с простыми узлами x_k, в которых заданы только $f(x_k)$. Достаточно представить $R_n(x)$ в виде

$$R_n(x) = r_n(x)\omega_{n+1}(x), \quad (41)$$

где

$$\omega_{n+1}(x) = (x - x_0)^{N_0}(x - x_1)^{N_1}\ldots(x - x_m)^{N_m},\ n+1 = N_0 + \ldots + N_m \quad (42)$$

и рассмотреть функцию

$$g(t) = f(t) - H_n(t) - r_n(x)\omega_{n+1}(t).$$

Применяя теорему Ролля к функции $g(t)$ и ее производным с учетом кратности корней в узлах $t = x_k$ и условия $g(x) = 0$ придем к формуле

$$f(x) - H_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x), \quad (43)$$

которая по существу повторяет формулу (35). С ее помощью можно написать оценку типа (36):

$$|R_n(x)| \leq \frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)|, \quad (44)$$

где M_{n+1} - максимальное значение модуля функции $f^{(n+1)}(x)$ (37). Здесь полином $\omega_{n+1}(x)$ (42) является обобщением полинома (32) на случай кратных корней.

Построение полинома Эрмита в общем случае при произвольном числе узлов и их кратности приводит к довольно громоздким выражениям и редко используется. Поэтому мы ограничимся двумя примерами, встречающимися на практике.

Пример 1

Построить интерполяционный полином Эрмита для функции $f(x)$ по известным значениям в узлах $f(x_k) = f_k$, $k = 0,1,\ldots,m$ и значению $f'(x_j) = f'_j$ в одном из узлов $x = x_j$.

Степень полинома $H_n(x)$ в данном случае равна $m+1$.

Будем искать $H_{m+1}(x)$ в виде

$$H_{m+1}(x) = \sum_{\substack{k=0\\k\neq j}}^{m} f_k \frac{(x - x_0)\ldots[k]\ldots(x - x_m)}{(x_k - x_0)\ldots[k]\ldots(x_k - x_m)}\left(\frac{x - x_j}{x_k - x_j}\right) + \left[f_j + \alpha_j(x - x_j)\right]\frac{(x - x_0)\ldots[j]\ldots(x - x_m)}{(x_j - x_0)\ldots[j]\ldots(x_j - x_m)}$$

Здесь выражения, стоящие под знаком суммы, суть обычные составляющие полинома в форме Лагранжа в узлах x_k, $k \neq j$, «усиленные» дополнительными множителями $(x - x_j)/(x_k - x_j)$. Слагаемое, отвечающее кратному узлу $x = x_j$, выделено отдельно как особое. Постоянная α_j подлежит определению.

Из структуры $H_{m+1}(x)$ видно, что $H_{m+1}(x_i) = f_i$, $i = 0,1,\ldots,m$. Найдем производную $H'_{m+1}(x_j)$ в узле $x = x_j$. Слагаемые, стоящие под знаком суммы, содержат множители $(x - x_j)^2$ и потому их производные обращаются в нуль при $x = x_j$. Таким образом,

$$H'_{m+1}(x_j) = f_j\left(\frac{1}{x_j - x_0} + \ldots + \frac{1}{x_j - x_{j-1}} + \frac{1}{x_j - x_{j+1}} + \ldots + \frac{1}{x_j - x_m}\right) + \alpha_j. \quad (45)$$

Для соблюдения требования $H'_{m+1}(x_j) = f'_j$ следует положить

$$\alpha_j = f'_j - f_j A_j,$$

где для краткости обозначено

$$A_j = \frac{1}{x_j - x_0} + \ldots + \frac{1}{x_j - x_{j-1}} + \frac{1}{x_j - x_{j+1}} + \ldots + \frac{1}{x_j - x_m}. \quad (46)$$

Итак:

$$H_{m+1}(x) = \sum_{\substack{k=0\\k\neq j}}^{m} f_k \frac{(x - x_0)\ldots[k]\ldots(x - x_m)}{(x_k - x_0)\ldots[k]\ldots(x_k - x_m)}\left(\frac{x - x_j}{x_k - x_j}\right) + \left[f_j + \left(f'_j - f_j A_j\right)(x - x_j)\right]\frac{(x - x_0)\ldots[j]\ldots(x - x_m)}{(x_j - x_0)\ldots[j]\ldots(x_j - x_m)}. \quad (47)$$

Пример 2.

Построить интерполяционный полином Эрмита для функции $f(x)$ в случае, когда во всех узлах интерполяции x_k, $k = 0,1,\ldots,m$ заданы значения функции $f(x_k) = f_k$ и ее первой производной $f'(x_k) = f'_k$.

В данном случае $N_k = 2$, $k = 0,1,\ldots,m$, так что степень полинома $H_n(x)$ равна $2m+1$.

Запишем исходный полином в виде:

$$H_{2m+1}(x) = \sum_{k=0}^{m}\left[f_k + \alpha_k(x - x_k)\right]\frac{(x - x_0)^2\ldots[k]\ldots(x - x_m)^2}{(x_k - x_0)^2\ldots[k]\ldots(x_k - x_m)^2}. \quad (48)$$

Представление (48) удобно тем, что автоматически выполняются условия

$$H_{2m+1}(x_k) = f_k.$$

При вычислении производной полинома (48) в узле $x = x_k$ следует учесть, что все слагаемые суммы, кроме слагаемого, отвечающему самому узлу x_k, дают нулевой вклад в производную в этой точке, поэтому

$$H'_{2m+1}(x_k) = f_k\left(\frac{2}{x_k - x_0} + \ldots + \frac{2}{x_k - x_{k-1}} + \frac{2}{x_k - x_{k+1}} + \ldots + \frac{2}{x_k - x_k}\right) + \alpha_k = f'_k.$$

Отсюда

$$\alpha_k = f'_k - 2f_k A_k,$$

где, числа A_k определяются формулой (46). Таким образом, решением данной задачи является полиномом Эрмита

$$H_{2m+1}(x) = \sum_{k=0}^{m}\left[f_k + (f'_k - 2f_k A_k)(x - x_k)\right]\frac{(x - x_0)^2\ldots[k]\ldots(x - x_m)^2}{(x_k - x_0)^2\ldots[k]\ldots(x_k - x_m)^2}. \quad (49)$$

Задача 5

Построить полином Эрмита второй степени $H_2(x)$ для функции $\sin x$ по следующим данным:

$$\sin 0 = 0, \sin\frac{\pi}{2} = 1, \sin'\frac{\pi}{2} = 0.$$

Вычислить с помощью этого полинома приближенное значение синуса в точке $x = \pi/4$. Найти погрешность, сравнить ее с погрешностью, которую дает интерполяционный полином $P_2(x)$ (16) задачи 2 и с теоретической оценкой погрешности (44).

Здесь мы имеем задачу, которая в общем виде была разобрана в примере 1: согласно (49) узел $x_0 = 0$ является простым, а узел $x_1 = \pi/2$ - двукратным. В этом случае в формуле (47) сумма, соответствующая простым узлам, сводится к одному слагаемому, которое в силу нулевого значения синуса в точке $x_0 = 0$ обращается в ноль. Второй член в формуле (47) соответствует кратному корню $x_1 = \pi/2$. Подставляя сюда соответствующее значение синуса и его производной в этой точке, а также значение коэффициента $A_1 = 2/\pi$, будем иметь:

$$H_2(x) = \left[1 - \frac{2}{\pi}\left(x - \frac{\pi}{2}\right)\right]\frac{2x}{\pi} = \frac{4}{\pi^2}x(\pi - x). \quad (50)$$

Вычислим значение полинома $H_2(x)$ в точке $x = \pi/4$ и подсчитаем погрешность

$$H_2\left(\frac{\pi}{4}\right) = \frac{3}{4}, R_2\left(\frac{\pi}{4}\right) = \frac{1}{\sqrt{2}} - \frac{3}{4} = -0.04282. \quad (51)$$

Теоретическая формула для погрешности (43) принимает в данном случае вид:

$$R_2\left(\frac{\pi}{4}\right) = \frac{1}{6}(-\cos\xi)\omega_3\left(\frac{\pi}{4}\right) = -\cos\xi\frac{\pi^3}{384},\ 0 \leq \xi \leq \frac{\pi}{2}. \quad (52)$$

Она правильно определяет знак погрешности и позволяет написать для нее мажорантную оценку

$$\left|R_2\left(\frac{\pi}{4}\right)\right| \leq \frac{\pi^3}{384} < 0.081. \quad (53)$$

Данная оценка согласуется с величиной погрешности (51), подсчитанной «в лоб».

При подсчете приближенного значения $\sin x$ с помощью полинома Эрмита $H_2(x)$ (50) в точке $x = \pi/4$ мы получили погрешность (51), модуль которой в два с лишним раза превышает погрешность (17) полинома $P_2(x)$ (16). Чтобы понять причину такого расхождения, рассмотрим рис. 3, на котором приведены графики функций $\sin x$ (сплошная линия) и $H_2(x)$ (пунктир). Сравним его с рис. 1, на котором изображены графики функции $\sin x$ и полинома $P_2(x)$. Из-за нулевого значения производной $H'_2(x)$ в точке $x = \pi/2$ график полинома $H_2(x)$ качественно больше похож на график синуса, чем график полином $P_2(x)$. Однако равенство полинома

$P_2(x)$ синусу не только в граничных точках отрезка $\left[0, \frac{\pi}{2}\right]$, но и во внутренней точке $x = \pi/6$ приводит к тому, что полином $P_2(x)$ приближает синус на отрезке $\left[0, \frac{\pi}{2}\right]$ лучше чем полином $H_2(x)$. Это хорошо видно при сравнении рис. 1 и рис. 3. Подсчет погрешностей (17) и (51) в точке $x = \pi/4$ является дополнительным тому подтверждением.

§2. Интерполирование сплайнами.

Увеличение степени интерполяционного полинома может оказаться невыгодным из-за быстрого роста объема вычислений. К тому же далеко не всегда оно приводит к повышению точности. Во второй половине XX века с появлением компьютеров и развитием современной вычислительной математики при обработке больших таблиц получила развитие новая идея – строить приближение функций с помощью кусочно-полиномиальной интерполяции с использованием полиномов сравнительно невысоких степеней. Наиболее удобными оказались полиномы третьей степени. Такие конструкции получили название кубических сплайнов.

2.1. Определение кубического сплайна.

Пусть на отрезке $[a,b]$ задана функция $y = f(x)$. Рассмотрим сетку узлов

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b \quad (54)$$

и обозначим через h_i расстояние между смежными узлами

$$h_i = x_i - x_{i-1},\ i = 1,\ldots,n \quad (55)$$

Определение:

Назовем кубическим сплайном функции $y = f(x)$, $x \in [a,b]$ на сетке (54) функцию $S(x)$ удовлетворяющую условиям:

S1. *На каждом отрезке $[x_{i-1}, x_i]$ функция $S(x)$ является полиномом третьей степени.*

S2. *Функция $S(x)$, её первая $S'(x)$ и вторая $S''(x)$ производные непрерывны на сегменте $[a,b]$.*

S3. *$S(x_i) = f(x_i) = f_i$, $i = 0,\ldots,n$*

S4. *На концах сегмента $[a,b]$ функция $S''(x)$ удовлетворяет условиям $S''(a) = S''(b) = 0$.*

Замечание. На концах сегмента $[a,b]$ могут быть заданы в принципе и другие условия, например:

$$S''(a) = A,\ S''(b) = B.$$

Справедлива следующая теорема.

Теорема.

Существует единственный сплайн $S(x)$, удовлетворяющий требованиям (S1) – (S4).

Мы проведем конструктивное доказательство этой теоремы.

2.2. Формулировка системы уравнений для коэффициентов кубического сплайна.

Сведем задачу построения сплайна к отысканию коэффициентов упомянутых полиномов третьей степени на каждом из отрезков $[x_{i-1}, x_i]$. Для этого сопоставим отрезку $[x_{i-1}, x_i]$ полином $S_i(x)$, для удобства записанный в виде:

$$S_i(x) = a_i + b_i(x - x_i) + \frac{c_i}{2}(x - x_i)^2 + \frac{d_i}{6}(x - x_i)^3, \; x \in [x_{i-1}, x_i], \;\; i = 1, \ldots, n. \qquad (56)$$

При этом, очевидно:

$$S_i'(x) = b_i + c_i(x - x_i) + \frac{d_i}{2}(x - x_i^2), \qquad (57)$$

$$S_i''(x) = c_i + d_i(x - x_i), \qquad (58)$$

так, что

$$S_i(x_i) = a_i, \;\; S_i'(x_i) = b_i, \;\; S_i''(x_i) = c_i. \qquad (59)$$

Для выполнения требований (S3) в узлах интерполяции с номерами $i = 1, \ldots, n$ следует положить:

$$a_i = f(x_i) = f_i, \;\; i = 1, \ldots, n \qquad (60)$$

Требуя непрерывности сплайна в узлах x_i $(i = 1, \ldots, n-1)$ и выполнения условия (S3) при $i = 0$, получим:

$$S_i(x_{i-1}) = f_{i-1}, \;\; i = 1, \ldots, n \qquad (61)$$

или

$$f_i + b_i(x_{i-1} - x_i) + \frac{c_i}{2}(x_{i-1} - x_i)^2 + \frac{d_i}{6}(x_{i-1} - x_i)^3 = f_{i-1}, \;\; i = 1, \ldots n.$$

Это равенство можно переписать следующим образом:

$$b_i h_i - \frac{c_i}{2} h_i^2 + \frac{d_i}{6} h_i^3 = f_i - f_{i-1}, \;\; i = 1, \ldots, n. \qquad (62)$$

Условие (S2) непрерывности первой производной $S'(x)$ в узлах x_i $(i = 1, \ldots, n-1)$ принимает вид:

$$S_i'(x_{i-1}) = S_{i-1}'(x_{i-1}) = b_{i-1}, \;\; i = 2, \ldots, n \qquad (63)$$

и приводит к соотношениям

$$b_i - c_i h_i + \frac{d_i}{2} h_i^2 = b_{i-1}, \;\; i = 2, \ldots, n$$

или

$$c_i h_i - \frac{d_i}{2} h_i^2 = b_i - b_{i-1}, \;\; i = 2, \ldots, n. \qquad (64)$$

Аналогичным образом условия непрерывности второй производной $S''(x)$ в тех же узлах:

$$S_i''(x_{i-1}) = S_{i-1}''(x_{i-1}) = c_{i-1}, \;\; i = 2, \ldots, n \qquad (65)$$

означают, что

$$d_i h_i = c_i - c_{i-1}, \;\; i = 2, \ldots, n. \qquad (66)$$

Наконец, дополнительные граничные условия (S4) дают еще два уравнения

$$\begin{cases} S_1''(x_0) = S_1''(a) = c_1 - d_1 h_1 = 0 \\ S_n''(x_n) = S_n''(b) = c_n = 0 \end{cases}. \qquad (67)$$

В итоге мы получили замкнутую систему (62), (64), (66), (67), содержащую в сумме $3n$ линейных уравнений для отыскания $3n$ неизвестных: $b_i, c_i, d_i, \;\; i = 1, 2, \ldots, n$

2.3. Редукция системы.

Удобно формально ввести ещё одно неизвестное c_0, положив при этом $c_0 = 0$, и первое уравнение в (67) переписать в виде:

$$d_1 h_1 = c_1 - c_0,$$

то есть в форме аналогичной (66).

Теперь уравнения (66) и (67) естественно представить в единообразном виде

$$d_i h_i = c_i - c_{i-1}, \; i = 1, 2, \ldots, n \qquad (68)$$

$$c_0 = 0, \; c_n = 0. \qquad (69)$$

Обратим внимание на то, что из системы (68) можно выразить все коэффициенты d_i через разности $c_i - c_{i-1}$, а затем из системы (62) выразить через c_i и c_{i-1} коэффициенты b_i. Подставляя полученные выражения в (64), придем к системе линейных уравнений для c_i:

$$\frac{1}{3} c_{i-2} h_{i-1} + \frac{2}{3} c_{i-1}(h_{i-1} + h_i) + \frac{1}{3} c_i h_i = 2\left(\frac{f_i - f_{i-1}}{h_i} - \frac{f_{i-1} - f_{i-2}}{h_{i-1}} \right), \; i = 2, 3, \ldots, n. \qquad (70)$$

Сдвигая индекс i на единицу, получим симметричную форму записи уравнений (70):

$$h_i c_{i-1} + 2(h_i + h_{i+1}) c_i + h_{i+1} c_{i+1} = 6\left(\frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i} \right), \; i = 1, \ldots n-1. \qquad (71)$$

Кроме того, согласно (69)

$$c_0 = c_n = 0. \qquad (72)$$

Система (71) содержит $n-1$ уравнение с $(n-1)$-ой неизвестной: $c_1, c_2, \ldots, c_{n-1}$. Величины c_0 и c_n определены дополнительными соотношениями (72). Если сетка (54) равномерная, т. е. $h_i = h = const$, то уравнения (71) принимают особенно простой вид:

$$c_{i-1} + 4c_i + c_{i+1} = 6\frac{f_{i-1} - 2f_i + f_{i+1}}{h^2}. \qquad (73)$$

Для уравнений системы (71) выполнено условие диагонального преобладания. Отсюда следует существование и единственность решения задачи (71), (72). По найденным величинам c_i можно рассчитать остальные коэффициенты сплайна по формулам

$$d_i = \frac{c_i - c_{i-1}}{h_i}, \;\; i = 1, \ldots, n \qquad (74)$$

и

$$b_i = \frac{1}{2} h_i c_i - \frac{1}{6} h_i^2 d_i + \frac{f_i - f_{i-1}}{h}, \;\; i = 1, \ldots, n, \qquad (75)$$

завершив тем самым построение сплайна. Теорема доказана.

2.4. Замечание о решении системы.

Уравнения (71) имеют так называемую трехточечную структуру, общий вид таких систем

$$A_i y_{i-1} + C_i y_i + B_i y_{i+1} = F_i, \; i = 1, 2, \ldots, n-1, \qquad (76)$$

$$y_0 = 0, \; y_n = 0 \qquad (77)$$

соответствует системе линейных уравнений с трехдиагональной матрицей T для определения вектора неизвестных $y = (y_1, y_2, \ldots, y_{n-1})$:

$$Ty = F,$$

где

$$T = \begin{vmatrix} C_1 & B_1 & 0 & 0 & 0 & 0 \\ A_2 & C_2 & B_2 & 0 & 0 & 0 \\ 0 & A_3 & C_3 & B_3 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & A_{n-1} & C_{n-1} \end{vmatrix}, \quad F = \begin{vmatrix} F_1 \\ F_2 \\ F_3 \\ \ldots \\ \ldots \\ F_{n-1} \end{vmatrix}. \qquad (78)$$

При этом легко видеть, что в нашем случае

$$|C_i| > |A_i| + |B_i|, \;\; i = 1, \ldots, n-1, \qquad (79)$$

поскольку

$$C_i = 2(h_i + h_{i+1}), \;\; A_i = h_i, \;\; B_i = h_{i+1}. \qquad (80)$$

Как было показано в главе 1, решение подобных систем эффективно осуществляется методом прогонки.

Задача 6.

Рассмотреть функцию $y = f(x) = 3^x$ *на отрезке* $[-1,1]$ *с узлами интерполяции* $x_0 = -1, \; x_1 = 0, \; x_2 = 1$. *Построить кубический сплайн. Найти его значение при* $x = 1/2$, *т. е. вычислить приближенно* $\sqrt{3}$. *Подсчитать погрешность.*

В рассматриваемом случае мы имеем равномерную сетку с шагом $h = 1$. У нее одна внутренняя точка x_1 и две граничные - x_0 и x_2. Система (73) сводится к одному уравнению относительно коэффициента c_1, которое с учетом дополнительных соотношений (70), определяющих нулевые значения коэффициентов c_0 и c_2, принимает вид:

$$4c_1 = 6\left(\frac{1}{3} - 2 + 3 \right).$$

Таким образом, в нашей задаче:

$$c_0 = 0, \; c_1 = 2, \; c_2 = 0.$$

Остальные коэффициенты сплайна находятся по формулам (60), (74), (75):

$$a_1 = 1, \; a_2 = 3; \;\; d_1 = 2, \; d_2 = -2; \;\; b_1 = 4/3, \; b_2 = 7/3.$$

Теперь можно выписать кубические полиномы, определяющие сплайн:

$$S(x) = \begin{cases} S_1(x) = 1 + \frac{4}{3}x + x^2 + \frac{1}{3}x^3, & -1 \le x \le 0, \\ S_2(x) = 3 + \frac{7}{3}(x-1) - \frac{1}{3}(x-1)^3, & 0 \le x \le 1. \end{cases} \qquad (82)$$

Легко проверить, что построенная таким образом функция $S(x)$ непрерывна вместе с первой и второй производной во внутренней узловой точке $x = 0$.

В заключение вычислим значение сплайна в точке $x = 1/2$, т. е. подсчитаем приближенно $\sqrt{3}$:

$$\sqrt{3} \approx S_2\left(\frac{1}{2}\right) = \frac{15}{8}, \; \varepsilon = \sqrt{3} - \frac{15}{8} = -0{,}142949. \qquad (83)$$

Значительная погрешность обусловлена прежде всего большим шагом $h = 1$. Определенную роль играют также условия S4:

$$S''(-1) = S''(1) = 0. \qquad (84)$$

Вторая производная рассматриваемой функции $f(x) = 3^x$ в точках $x = \pm 1$ в ноль не обращается, т. е. условие (84) дает о ней искаженную информацию. Если учесть при построении сплайна истинные значения функции $f''(x)$ в точках ± 1, то точность аппроксимации улучшится.

2.5. Сходимость и точность интерполирования сплайнами.

При обсуждении эффективности численного метода в первую очередь обращают внимание на две характеристики:

1. Условие сходимости метода (сходимость).
   Речь идет о минимальных по возможности ограничениях, при которых приближенное решение задачи стремится к точному решению задачи.
   Сходимость означает, что данный метод в принципе позволяет найти решение задачи с любой степенью точности.
2. Скорость сходимости (точность).
   Это характеристика близости приближенного решения к точному (характеристика скорости убывания погрешности) при некоторых дополнительных ограничениях.

Посмотрим как решаются эти вопросы в теории сплайнов.

Итак, на сегменте $[a, b]$ задана функция $f(x)$ и построена сетка

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b; \;\; h_i = x_i - x_{i-1} > 0.$$

Введем в рассмотрение величину

$$h = \max_{1 \le i \le n} h_i. \qquad (85)$$

Приведем без доказательства две теоремы.

Теорема 1. *Пусть* $f(x)$ *непрерывна на сегменте* $[a, b]$, *тогда для любого* $\varepsilon > 0$ *можно указать* $\delta(\varepsilon) > 0$ *такое, что при любой сетке, удовлетворяющей условию* $h < \delta$ *справедливо неравенство*

$$|f(x) - S(x)| < \varepsilon \;\; \forall x \in [a, b], \qquad (86)$$

иными словами $S_h(x)$ *при* $h \to 0$ *равномерно сходится к непрерывной функции* $f(x)$.

Теорема 2. *Пусть* $f(x)$ *имеет на сегменте* $[a, b]$ *четыре непрерывных производных и дополнительно удовлетворяет условию* $f''(a) = f''(b) = 0$. *Тогда имеют место неравенства (оценки):*

$$|f(x) - S(x)| \le M_4 h^4 \;\; \forall x \in [a, b], \qquad (87)$$

$$|f'(x) - S'(x)| \le M_4 h^3 \;\; \forall x \in [a, b], \qquad (88)$$

$$|f''(x) - S''(x)| \le M_4 h^2 \;\; \forall x \in [a, b], \qquad (89)$$

$$M_4 = \max_{[a,b]} |f^{(4)}(x)|. \qquad (90)$$

§3. Метод наименьших квадратов.

Метод наименьших квадратов был предложен Гауссом и Лежандром в конце XVIII - начале XIX веков в связи с проблемой обработки экспериментальных данных. В этом случае задача построения функции непрерывного аргумента по дискретной информации (1), (2) характеризуется двумя особенностями:

1. Число точек x_i, в которых проводятся измерения, обычно бывает достаточно большим.
2. Значения функции y_i (2) в точках сетки x_i (1) определяются приближенно в связи с неизбежными ошибками измерения.

С учетом этих обстоятельств строить функцию $y(x)$ в виде суммы большого числа слагаемых (3) и добиваться ее точного равенства в точках сетки величинам y_i, как это делалось при интерполировании, становится нецелесообразным.

В методе наименьших квадратов апроксимирующая функция $y(x)$ ищется в виде суммы, аналогичной (3), но содержащей сравнительно небольшое число слагаемых

$$F(x) = \sum_{k=0}^{m} a_k \varphi_k(x), \quad m < n, \qquad (91)$$

в частности, возможен вариант $m \square n$.

Предположим, что мы каким-то образом выбрали коэффициенты a_k, тогда в каждой точке сетки x_i, можно подсчитать погрешность

$$\delta_i = y_i - F(x_i) = y_i - \sum_{k=0}^{m} a_k \varphi_k(x_i), \;\; i = 0, 1, 2, \ldots, n. \qquad (92)$$

Сумма квадратов этих величин называется суммарной квадратичной погрешностью

$$J = \sum_{i=0}^{n} \delta_i^2 = \sum_{i=0}^{n} (y_i - \sum_{k=0}^{m} a_k \varphi_k(x_i))^2. \qquad (93)$$

Она дает количественную оценку того, насколько близки значения функции $F(x)$ (91) в точках сетки к величинам y_i.

Меняя значения коэффициентов a_k, мы будем менять погрешность J, которая является их функцией. В результате естественно возникает задача:

Найти такой, набор коэффициентов a_k, *при которых суммарная квадратичная погрешность* J *оказывается минимальной.*

Функцию $F(x)$ (91) с набором коэффициентов, удовлетворяющих этому требованию, называют наилучшим приближением по методу наименьших квадратов.

Построение наилучшего приближения сводится к классической задаче математического анализа об экстремуме функции нескольких переменных. Метод решения этой задачи известен. Необходимым условием экстремума является равенство нулю в экстремальном точке всех первых частных производных рассматриваемой функции. В случае (93) это дает

$$\frac{\partial J}{\partial a_l} = -2\sum_{i=0}^{n} (y_i - \sum_{k=0}^{m} a_k \varphi_k(x_i)) \varphi_l(x_i) = 0 \quad l = 0, 1, \ldots, m. \qquad (94)$$

Оставим члены, содержащие a_k, слева и поменяем в них порядок суммирования по индексам i и k. Члены, содержащие y_i, перенесем направо. В результате уравнения (94) примут вид

$$\sum_{k=0}^{m} \gamma_{lk} a_k = b_l, \quad l = 0, 1, \ldots, m, \qquad (95)$$

где

$$\gamma_{lk} = \sum_{i=0}^{n} \varphi_l(x_i) \varphi_k(x_i), \qquad (96)$$

$$b_l = \sum_{i=0}^{n} \varphi_l(x_i) y_i. \qquad (97)$$

Мы получили систему линейных алгебраических уравнений (95), в которой роль неизвестных играют искомые коэффициенты разложения $a_0, a_1, \ldots, a_m$. Число уравнений и число неизвестных в этой системе совпадает и равно $m+1$. Матрица коэффициентов системы Γ состоит из элементов γ_{lk}, которые определяются формулой (96). Ее называют матрицей Грама для системы функций $\varphi_0(x), \varphi_1(x), \ldots, \varphi_m(x)$ на сетке (1). Отметим, что матрица Грама является симметричной: для ее элементов, согласно (96), справедливо равенство $\gamma_{lk} = \gamma_{kl}$. Числа b_l, стоящие в правой части уравнений (95), вычисляются по формуле (97) через значения y_i сеточной функции (2).

Предположим, что функции $\varphi_0(x), \varphi_1(x), \ldots, \varphi_m(x)$ выбраны такими, что определитель матрицы Грама, отличен от нуля:

$$\Delta = \det \Gamma \neq 0. \qquad (98)$$

В этом случае при любой правой части системы (95) имеет единственное решение

$$\overline{a}_0, \overline{a}_1, \ldots, \overline{a}_m. \qquad (99)$$

Рассмотрим наряду с набором коэффициентов (99), полученных в результате решения системы (95), любой другой набор коэффициентов $a_0, a_1, \ldots, a_m$. Представим числа a_k в виде

$$a_0 = \overline{a}_0 + \Delta a_0, a_1 = \overline{a}_1 + \Delta a_1, \ldots, a_m = \overline{a}_m + \Delta a_m,$$
$$(\Delta a_0)^2 + (\Delta a_1)^2 + \cdots + (\Delta a_m)^2 > 0 \qquad (100)$$

и сравним значения суммарной квадратичной погрешности J для функций $F(x)$ (91), построенных с помощью коэффициентов (99) и (100).

Квадрат погрешности и точке $x = x_i$ для функции $F(x)$ (91) с коэффициентами (100) можно записать в виде

$$\delta_i^2 = \left\{ y_i - \sum_{k=0}^{m} (\overline{a}_k + \Delta a_k) \varphi_k(x_i) \right\}^2 =$$

$$= \left\{ \left(y_i - \sum_{k=0}^{m} \overline{a}_k \varphi_k(x_i) \right) - \sum_{k=0}^{m} \Delta a_k \varphi_k(x_i) \right\}^2 = \left(y_i - \sum_{k=0}^{m} \overline{a}_k \varphi_k(x_i) \right)^2 - \qquad (101)$$

$$- 2\left(y_i - \sum_{k=0}^{m} \overline{a}_k \varphi_k(x_i) \right) \left(\sum_{l=0}^{m} \Delta a_l \varphi_l(x_i) \right) + \left(\sum_{k=0}^{m} \Delta a_k \varphi_k(x_i) \right)^2.$$

Здесь в среднем слагаемом мы заменили в одной из сумм индекс суммирования k на l, чтобы не использовать один и тот же индекс в двух разных суммах и иметь возможность перемножить их почленно.

Чтобы получить суммарную квадратичную погрешность, нужно просуммировать выражения (101) для δ_i^2 по индексу i Первые слагаемые не содержат Δa_k. Их сумма дает погрешность J, вычисленную для функции (91) с коэффициентами (99) $\overline{a}_k$.

Рассмотрим теперь сумму вторых слагаемых, которые зависят от Δa_l линейно:

$$-2\sum_{i=0}^{n} \left\{ \left(y_i - \sum_{k=0}^{m} \overline{a}_k \varphi_k(x_i) \right) \cdot \left(\sum_{l=0}^{m} \Delta a_l \varphi_l(x_i) \right) \right\} =$$

$$= -2\sum_{l=0}^{m} \Delta a_l \left\{ \sum_{i=0}^{n} y_i \varphi_l(x_i) - \sum_{k=0}^{m} \overline{a}_k \sum_{i=0}^{n} \varphi_k(x_i) \varphi_l(x_i) \right\} = \qquad (102)$$

$$= -2\sum_{l=0}^{m} \Delta a_l \{ b_l - \sum_{k=0}^{m} \overline{a}_k \gamma_{lk} \} = 0.$$

Здесь мы поменяли местами порядок суммирования и воспользовались тем, что коэффициенты $\overline{a}_k$, удовлетворяют системе уравнений (95).

С учетом (102) будем иметь

$$J(\overline{a}_0 + \Delta a_0, \overline{a}_1 + \Delta a_1, \ldots, \overline{a}_m + \Delta a_m) =$$

$$= J(\overline{a}_0, \overline{a}_1, \ldots, \overline{a}_m) + \sum_{i=0}^{n} \left(\sum_{k=0}^{m} \Delta a_k \varphi_k(x_i) \right)^2 > J(\overline{a}_0, \overline{a}_1, \ldots, \overline{a}_m). \qquad (103)$$

Формула (103) показывает, что функция $F(x)$ (91) с коэффициентами $\overline{a}_k$ (100), полученными в результате решения уравнений (95), действительно минимизирует суммарную квадратичную погрешность J. Если мы возьмем любой другой набор коэффициентов (100), отличный от (99), то согласно формуле (103) к погрешности $y(\overline{a}_0, \overline{a}_1, \ldots, \overline{a}_m)$ добавится положительное слагаемое и она увеличится.

Итак, чтобы построить наилучшее приближение (91) сеточной функции (1), (2) по методу наименьших квадратов, нужно взять в качестве коэффициентов разложения a_k решение системы линейных уравнений (95).

Задача 7

Сеточная функция задана таблицей 1

.

Таблица 1:

i	x_i	y_i
0	*0,0*	*0,95*
1	*0,5*	*1,54*
2	*1,0*	*2,04*
3	*1,5*	*2,46*
4	*2,0*	*2,95*

Построить линейную функцию

$$F(x) = a_0 + a_1 x, \qquad (104)$$

которая дает для нее наилучшее приближение по методу наименьших квадратов.

В рассматриваемом случае имеем:

$$n = 4, \;\; m = 1, \;\; \varphi_0(x) = 1, \;\; \varphi_1(x) = x.$$

Для определения коэффициентов a_0 и a_1 составим систему уравнений (95). Элементы $\gamma_{lk} (l = 0,1 \;\; k = 0,1)$ матрицы Грама вычисляются по формуле (96)

$$\gamma_{0,0} = \sum_{i=0}^{4} \varphi_0(x_i) \varphi_0(x_i) = 5,$$

$$\gamma_{0,1} = \gamma_{1,0} = \sum_{i=0}^{4} \varphi_0(x_i) \varphi_1(x_i) = 5,$$

$$\gamma_{1,1} = \sum_{i=0}^{4} \varphi_1(x_i) \varphi_1(x_i) = 7.5.$$

Числа b_0 и b_1, стоящие в правой части уравнений (95), находим по формуле (97)

$$b_0 = \sum_{i=0}^{4} \varphi_0(x_i) y_i = 9,94,$$

$$b_1 = \sum_{i=0}^{4} \varphi_1(x_i) y_i = 12,40.$$

В результате система (95) принимает в рассматриваемом случае вид

$$5a_0 + 5a_1 = 9.94$$
$$5a_0 + 7,5a_1 = 12.40. \qquad (105)$$

Определитель системы (105) $\Delta = 12,5 \neq 0$, так что система имеет единственное решение

$$\overline{a}_0 = 1,004, \quad \overline{a}_1 = 0,984.$$

В результате мы получаем следующую линейную аппроксимацию рассматриваемой табличной функции

$$F(x) = 1,004 + 0,984x. \qquad (106)$$

Теперь, когда функция (106) построена, можно подсчитать погрешность аппроксимации в точках сетки:

$$\delta_i = y_i - (1,004 + 0,984x_i), \; i = 0, 1, 2, 3, 4.$$

В результате получаем

$$\delta_0 = -0,054, \;\; \delta_1 = -0,044, \;\; \delta_2 = 0,052, \;\; \delta_3 = -0,020, \;\; \delta_4 = -0,022. \qquad (107)$$

Отметим, что наибольшая по модулю погрешность достигается в точке $x_0 = 0$: $|\delta_0| = 0.054 > |\delta_i|, \; i = 1, 2, 3, 4$.

В заключение сделаем важное замечание. Обычно бывает известна точность ε, с которой задаются значения функции y_i. Например, если речь идет об экспериментальных данных, то ошибка в определении y_i зависит от методики проведения измерений и точности приборов.

Предположим, что в разобранном примере числа y_i заданы с точностью $\varepsilon = 0,1$. В этом случае построенная линейная функция согласуется с доступной нам информацией о функции $y(x)$: погрешности (107) по модулю не превышают ε. В результате мы можем утверждать, что в пределах точности задания таблицы зависимость y от x можно принять линейной.

Это видно на рис.4. На нем показаны точки (x_i, y_i), соответствующие рассматриваемой таблице. Для каждой из них указан доверительный интервал $y_i - 0.1 \le y \le y_i + 0.1$, в пределах которого может реально находится значение функции $y(x_i)$ с учетом точности задания величины y_i. Прямая (106) везде проходит внутри доверительных интервалов, что подтверждает сделанный выше вывод.

Рассмотрим теперь противоположный случай: будем считать, что величины y_i заданы с более высокой точностью $\varepsilon = 0,01$. При такой точности построенная линейная аппроксимация (106) не согласуется с данными таблицы: погрешность аппроксимации (107) превышает по модулю ε. В этом случае нужно либо увеличить число членов в разложении функции $F(x)$, добавив к линейной функции квадратичный член $a_2 x^2$, либо заменить систему функций $\varphi_k(x)$, по которым ведется разложение, на какую-нибудь другую.

Глава 3. ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

§1. Формула Ньютона-Лейбница и численное интегрирование.

Из курса математического анализа Вы знакомы с вычислением определенных интегралов с помощью формулы Ньютона-Лейбница

$$I=\int_a^b f(x)dx=F(b)-F(a). \tag{1}$$

где $F(x)$- любая первообразная подынтегральной функции $f(x)$ на отрезке $[a,b]$. Формула Ньютона-Лейбница играет важную роль, устанавливая связь задачи определенного интегрирования с задачей отыскания первообразной (с задачей неопределенного интегрирования). Она позволяет вычислять интегралы от элементарных функций, первообразные которых тоже являются элементарными функциями. Например,

$$\int_1^2\frac{dx}{x}=\ln|x|_1^2=\ln 2. \tag{2}$$

Однако существует много простых функций, первообразные которых не выражаются через элементарные функции. В качестве примера можно привести такие функции как e^{-x^2} или $\sin x/x$. Для них описанный способ вычисления определенных интегралов неприменим. Формула Ньютона-Лейбница не позволяет также вычислять интегралы от функций, которые задаются графиком или таблицей. Иными словами, она не дает общего, универсального метода нахождения определенного интеграла от произвольной функции $f(x)$ по ее значениям на отрезке $[a,b]$, она не является алгоритмом решения рассматриваемой задачи.

Универсальные алгоритмы вычисления определенных интегралов дают формулы численного интегрирования или, как их обычно называют, квадратурные формулы (буквально формулы вычисления площадей). Квадратурные формулы имеют вид:

$$I=\int_a^b f(x)dx=\sum_{i=1}^n c_i f(x_i)+R_n. \tag{3}$$

Здесь точки $x_i\in[a,b]$ называют узлами, коэффициенты c_i -весовыми множителями или просто весами, величину R_n - остаточным членом или погрешностью. Узлы и веса подбираются таким образом, чтобы выполнялось предельное равенство:

$$\lim_{n\to\infty}R_n=0,\ \text{так что}\ \lim_{n\to\infty}\sum_{i=1}^n c_i f(x_i)=I. \tag{4}$$

Суть этого требования заключается в следующем. Если пренебречь в формуле (3) остаточным членом R_n, то получится приближенное равенство:

$$I=\int_a^b f(x)dx\approx\sum_{i=1}^n c_i f(x_i). \tag{5}$$

Условие (4), которое называют сходимостью, позволяет сделать погрешность в равенстве (5) меньше любого наперед заданного числа за счет выбора достаточно большого n.Таким образом, открывается возможность вычислить интеграл I с любой наперед заданной точностью по значениям функции $f(x)$, взятым в разных точках x_i отрезка $[a,b]$. Чем выше требование точности, тем больше слагаемых приходится удерживать в сумме. За точность приходится платить увеличением объема вычислений.

В заключение сделаем следующее замечание. Подставляя в формулу (3) функцию $f(x)=1$, получим:

$$(b-a)=\sum_{i=1}^n c_i+R_n.$$

Обычно весовые коэффициенты c_i подбираются таким образом, чтобы выполнялось равенство:

$$(b-a)=\sum_{i=1}^n c_i,$$

т. е., чтобы при интегрировании константы равенство (5) было не приближенным, а точным.

В следующих параграфах этой главы мы обсудим методы построения квадратурных формул и с разных сторон разберем проблему оценки их точности.

§2. Квадратурные формулы прямоугольников, трапеций, Симпсона.

2.1. Квадратурные формулы прямоугольников, трапеций, Симпсона и их особенности.

С квадратурными формулами прямоугольников, трапеций, Симпсона Вы уже встречались в курсе математического анализа, поэтому их вывод будет изложен конспективно.

Возьмем произвольное целое число n и разобьем отрезок $[a,b]$, по которому ведется интегрирование, на n равных отрезков длиной $h=(b-a)/n$ точками

$$x_i=a+ih,\quad 0\le i\le n. \tag{6}$$

Для дальнейшего нам также понадобятся средние точки этих отрезков

$$\xi_i'=a+(i-1/2)h,\quad \xi_i'\in[x_{i-1},x_i],\quad 1\le i\le n. \tag{7}$$

Идея вывода формулы прямоугольников очень проста. Построим с помощью проведенного разбиения интегральную сумму, в которой значения функции $f(x)$ для каждого отрезка $[x_{i-1},x_i]$ вычисляются в его средней точке ξ_i' (7):

$$P_n=\frac{b-a}{n}\sum_{i=1}^n f(\xi_i'). \tag{8}$$

Принимая во внимание то, что интегральная сумма дает приближенное значение интеграла, можно написать:

$$I=P_n+\alpha_n. \tag{9}$$

В квадратурной формуле (9) узлами являются точки ξ_i' (7), все весовые множители одинаковы и равны $h=(b-a)/n$. Для остаточного члена введено специальное обозначение α_n.

Формулу (9) называют формулой прямоугольников. Причина такого названия имеет простой геометрический смысл. Величина P_n (8) представляет собой сумму площадей прямоугольников с одинаковыми основаниями $h=(b-a)/n$ и высотами $f(\xi_i')$. Она аппроксимирует с точностью до α_n площадь криволинейной трапеции, соответствующей исходному интегралу (см. рис. 1).

Идея вывода квадратурных формул трапеций и Симпсона иная. Она заключается в том, чтобы сопоставить подынтегральной функции $f(x)$ близкую ей функцию $g_n(x)$, которую можно проинтегрировать, и приближенно заменить искомый интеграл I интегралом от этой функции.

Рассмотрим, как данная идея реализуется при выводе формулы трапеций. В этом случае в качестве аппроксимирующей функции $g_n(x)$ берется кусочно – линейная функция. На каждом из частичных сегментов $[x_{i-1},x_i]$ она задается формулой

$$g_n(x)=f(x_{i-1})+\frac{f(x_i)-f(x_{i-1})}{h}(x-x_{i-1}),\qquad x\in[x_{i-1},x_i],\quad 1\le i\le n. \tag{10}$$

В граничных точках отрезка $x=x_{i-1}$ и $x=x_i$ функция $g_n(x)$ принимает те же значения, что и функция $f(x)$:

$$g_n(x_{i-1})=f(x_{i-1}),\quad g_n(x_i)=f(x_i), \tag{11}$$

т. е. она осуществляет кусочно – линейную интерполяцию функции $f(x)$ на отрезке $[a,b]$ (см. рис. 2).

Вычислим интеграл:

$$\int_{x_{i-1}}^{x_i}g_n(x)dx=\int_{x_{i-1}}^{x_i}\left\{f(x_{i-1})+\frac{f(x_i)-f(x_{i-1})}{h}(x-x_{i-1})\right\}dx=\frac{h}{2}\left(f(x_{i-1})+f(x_i)\right). \tag{12}$$

Этот результат имеет простой геометрический смысл: фигура ограниченная снизу отрезком $[x_{i-1},x_i]$ оси x, сверху отрезком прямой (10), с боков вертикальными прямыми $x=x_{i-1}$ и $x=x_i$, представляет собой трапецию, площадь которой дается формулой (12).

Интеграл от функции $g_n(x)$ по всему отрезку $[a,b]$ является суммой интегралов (12)

$$\begin{aligned}T_n=\int_a^b g_n(x)dx=\sum_{i=1}^n\int_{x_{i-1}}^{x_i}g_n(x)dx=\\=\frac{b-a}{n}\left\{\frac{1}{2}f(a)+f(x_1)+f(x_2)+\ldots+f(x_{n-1})+\frac{1}{2}f(b)\right\}.\end{aligned} \tag{13}$$

Он дает приближенное значение интеграла I:

$$I=\int_a^b f(x)dx=T_n+\beta_n, \tag{14}$$

В квадратурной формуле (14) узлами являются точки x_i (6). Все весовые коэффициенты, кроме двух, одинаковы и равны $h=(b-a)/n$, а весовые коэффициенты при $i=0$ и $i=n$ имеют значения в два раза меньше. Для остаточного члена введено специальное обозначение β_n. Формулу (14) называют квадратурной формулой трапеций. С точностью до β_n она выражает площадь криволинейной трапеции, соответствующую интегралу I, через сумму площадей трапеций (12) (см. рис. 2).

Формула (8) для величины P_n изначально строилась как интегральная сумма. При выводе формулы (13) для величины T_n понятие интегральной суммы не использовалась. Однако теперь, когда формула уже получена, видно, что величину T_n тоже можно интерпретировать как интегральную сумму. Чтобы убедиться в этом, рассмотрим разбиение отрезка $[a,b]$ на частичные отрезки точками ξ_i' (7). Оно дает $n+1$ отрезок. Два крайних $[a,\xi_1']$ и $[\xi_n',b]$ имеют длину $h/2$, а остальные - длину h. Выберем для образования интегральной суммы на крайних отрезках значения функции $f(x)$ в точках a и b, а на остальных отрезках $[\xi_i',\xi_{i+1}']$ - значения функции $f(x)$ в их средних точках x_i $(1\le i\le n-1)$. Образованная таким образом интегральная сумма соответствует выражению (13) для T_n.

Вывод квадратурной формулы Симпсона развивает описанный подход дальше. Теперь для аппроксимации функции $f(x)$ используется не кусочно – линейное, а кусочно – квадратичное интерполирование.

Будем считать n четным и сгруппируем отрезки $[x_{i-1},x_i]$ парами: первая пара $[a,x_1]$, $[x_1,x_2]$, вторая пара $[x_2,x_3]$, $[x_3,x_4]$ и т. д. Для каждого двойного отрезка $[x_{2j-2},x_{2j}]$ построим интерполяционный полином второй степени в форме Лагранжа, принимающий в узлах x_{2j-2}, x_{2j-1}, x_{2j} значения функции $f(x)$. В результате получим аппроксимирующую функцию $g_n(x)$ на отрезке $[a,b]$ в виде кусочно – квадратичной функции:

$$\begin{aligned}g_n(x)=f(x_{2j-2})\frac{(x-x_{2j-1})(x-x_{2j})}{2h^2}+f(x_{2j-1})\frac{(x-x_{2j-2})(x-x_{2j})}{(-h^2)}+\\+f(x_{2j})\frac{(x-x_{2j-2})(x-x_{2j-1})}{2h^2},\\x\in[x_{2j-2},x_{2j}],\quad 1\le j\le n/2.\end{aligned} \tag{15}$$

Проинтегрировав полином второй степени (15) по отрезку $[x_{2j-2},x_{2j}]$, получим

$$\int_{x_{2j-2}}^{x_{2j}}g_n(x)dx=\frac{h}{3}\left\{f(x_{2j-2})+4f(x_{2j-1})+f(x_{2j})\right\},\ h=\frac{b-a}{n}. \tag{16}$$

Интеграл от функции $g_n(x)$ по всему отрезку $[a,b]$ равен сумме интегралов (16)

$$\begin{aligned}S_n=\int_a^b g_n(x)dx=\sum_{j=1}^{n/2}\int_{x_{2j-2}}^{x_{2j}}g_n(x)dx=\\=\frac{b-a}{3n}\left\{f(a)+4f(x_1)+2f(x_2)+\ldots+2f(x_{n-2})+4f(x_{n-1})+f(b)\right\}.\end{aligned} \tag{17}$$

(Напомним, что число n должно быть обязательно четным.) Величина S_n (17) дает приближенное значение интеграла I:

$$I=\int_a^b f(x)dx=S_n+\gamma_n. \tag{18}$$

Узлами квадратурной формулы (17), как и формулы трапеций (14), являются точки x_i (6). Весовые коэффициенты в узлах с четными и нечетными номерами имеют разные значения. Для остаточного члена введено обозначение γ_n. Формула (18) называется квадратурной формулой Симпсона.

Представление (17) для S_n как и представление (13) для T_n, также можно рассматривать как интегральную сумму. Для ее построения нужно разбить отрезок $[a,b]$ на $(n+1)$ частичный отрезок с помощью n внутренних точек

$$\eta_{2j-1}=x_{2j-1}-2h/3,\ \eta_j=x_{2j-1}+2h/3,\ 1\le j\le n/2 \tag{19}$$

и двух граничных точек

$$\eta_0=a\ \text{и}\ \eta_{n+1}=b. \tag{20}$$

В результате получаются отрезки $[\eta_{i-1},\eta_i]$, $1\le i\le n+1$ различной длины. Два крайних отрезка $[a,\eta_1]$ и $[\eta_n,b]$ имеют длину $h/3$. Отрезки, в центре которых лежат точки x_i с четными номерами, - длину $2h/3$, отрезки, в центре которых лежат точки x_i с нечетными номерами, – длину $4h/3$.

Для построения интегральной суммы, соответствующей данному разбиению, возьмем для крайних отрезков значения функции $f(x)$ в точках a и b, для остальных отрезков – значение функции $f(x)$ в их средних точках x_i. В результате получим интегральную сумму в виде выражения (17). Разные длины частичных отрезков приводит к своеобразному чередованию коэффициентов в виде двоек, четверок и единиц в крайних точках.

Заканчивая обсуждение формул (13) для T_n и (17) для S_n, установим полезную для дальнейшего связь между этими величинами

$$S_n=\frac{4}{3}T_n-\frac{1}{3}T_{\frac{n}{2}}. \tag{21}$$

Здесь $T_{\frac{n}{2}}$ - сумма (13) с вдвое меньшим числом слагаемых и, соответственно, с вдвое большим шагом. Благодаря этому при ее образовании в качестве узлов используются точки x_i (6) только с четными номерами. Поскольку в формуле Симпсона n

предполагается обязательно четным, то $n/2$ - целое число, так что выражение $T_{\frac{n}{2}}$ определено.

Соотношение (21) проверяется «в лоб». Из (13) следует, что:

$$\begin{aligned}\frac{4}{3}T_n=\frac{b-a}{3n}\left\{2f(a)+4f(x_1)+4f(x_2)+\ldots+4f(x_{n-2})+4f(x_{n-1})+2f(b)\right\},\\\frac{1}{3}T_{\frac{n}{2}}=\frac{b-a}{3n}\left\{f(a)+2f(x_2)+\ldots+2f(x_{n-2})+2f(b)\right\}.\end{aligned}$$

Вычитая теперь вторую строку из первой, получим равенство (21).

2.2. Сходимость и точность квадратурных формул прямоугольников, трапеций и Симпсона.

После того, как мы установили, что величины P_n , T_n, S_n являются интегральными суммами, проблема сходимости рассмотренных методов численного интегрирования решается элементарно. Их сходимость имеет место для любой интегрируемой функции:

$$\lim_{n\to\infty}\alpha_n=0,\ \lim_{n\to\infty}P_n=I, \tag{22}$$

$$\lim_{n\to\infty}\beta_n=0,\ \lim_{n\to\infty}T_n=I. \tag{23}$$

$$\lim_{n\to\infty}\gamma_n=0\ \lim_{n\to\infty}S_n=I. \tag{24}$$

Этот вывод является прямым следствием определения интегрируемости.

Предельные соотношения (22) – (24) доказывают принципиальную возможность вычисления интеграла от произвольной интегрируемой функции каждым из трех методов с любой точностью ε за счет выбора достаточно большого n и, соответственно, малого шага $h=(b-a)/n$.

После общего вывода о сходимости методов перейдем к обсуждению основного вопроса, связанного с организацией реального вычислительного процесса: каким нужно взять n, чтобы добиться при вычислении интеграла нужной точности. Ответ на него требует анализа остаточных членов. При этом на функцию $f(x)$ приходится накладывать дополнительные ограничения, выходящие за рамки предположения об интегрируемости.

Начнем с обсуждения остаточных членов в квадратурных формулах прямоугольников и трапеций. Предположим, что функция $f(x)$ дважды непрерывно дифференцируема на отрезке $[a,b]$. В курсе математического анализа при этом предположении устанавливаются формулы

$$\int_{x_{i-1}}^{x_i}f(x)dx=f(\xi_i)h+\frac{h^3}{24}f''(\eta_i'), \tag{25}$$

$$\int_{x_{i-1}}^{x_i}f(x)dx=\frac{f(x_{i-1})+f(x_i)}{2}h-\frac{h^3}{12}f''(\eta_i''), \tag{26}$$

где η_i' и η_i'' - некоторые точки отрезка $[x_{i-1},x_i]$. Существование таких точек гарантировано, но их точное положение неизвестно. (См В. А. Ильин, Э. Г. Позняк «Основы математического анализа». М. 1965. С. 389-397.)

Суммируя равенства (25) и (26) по i, получим формулы (9) и (14) со следующими выражениями для остаточных членов

$$\alpha_n=\frac{h^3}{24}\sum_{i=1}^n f''(\eta_i'), \tag{27}$$

$$\beta_n=-\frac{h^3}{12}\sum_{i=1}^n f''(\eta_i''). \tag{28}$$

Рассмотрим суммы

$$h\sum_{i=1}^n f''(\eta_i')\ \text{и}\ h\sum_{i=1}^n f''(\eta_i''). \tag{29}$$

Функция $f''(x)$ по предположению непрерывна и, следовательно, интегрируема на отрезке $[a,b]$. С учетом этого замечания выражения (29) можно рассматривать как интегральные суммы для интеграла $\int_a^b f''(x)dx$. Отсюда следует вывод:

$$\lim_{n\to\infty}h\sum_{i=1}^n f''(\eta_i')=\int_a^b f''(x)dx=f'(b)-f'(a), \tag{30}$$

$$\lim_{n\to\infty}h\sum_{i=1}^n f''(\eta_i'')=\int_a^b f''(x)dx=f'(b)-f'(a). \tag{31}$$

Предельные равенства (30) и (31) позволяют записать остаточные члены квадратурных формул прямоугольников и трапеций в виде

$$\alpha_n=\frac{1}{n^2}(A+\mu_n), \tag{32}$$

$$\beta_n=\frac{1}{n^2}(B+\nu_n), \tag{33}$$

где

$$A=\frac{(b-a)^2}{24}\left\{f'(b)-f'(a)\right\}, \tag{34}$$

$$\mu_n=\frac{(b-a)^2}{24}\left\{h\sum_{i=1}^n f''(\eta_i')-\int_a^b f''(x)dx\right\}\to 0,\ \text{при}\ n\to\infty, \tag{35}$$

$$B=-\frac{(b-a)^2}{12}\left\{f'(b)-f'(a)\right\}, \tag{36}$$

$$\nu_n=-\frac{(b-a)^2}{12}\left\{h\sum_{i=1}^n f''(\eta_i'')-\int_a^b f''(x)dx\right\}\to 0,\ \text{при}\ n\to\infty. \tag{37}$$

Формулы (32) и (33) выделяют в остаточных членах главные слагаемые A/n^2 и B/n^2, которые при возрастании n стремятся к нулю как n^{-2}. Важно подчеркнуть, что

коэффициенты A (34) и B (36) от n не зависят. Дополнительные слагаемые μ_n/n^2 и ν_n/n^2 являются бесконечно малыми более высокого порядка. Если ими пренебречь по сравнению с главными слагаемыми, то получатся простые асимптотические представления остаточных членов

$$\alpha_n\approx An^{-2}\ \text{и}\ \beta_n\approx Bn^{-2}. \tag{38}$$

Их относительная точность возрастает при увеличении n.

Теперь получим другие представления остаточных членов. Из курса математического анализа известно следующее утверждение.

Лемма.

Пусть функция $\varphi(x)$ непрерывна на отрезке $[a,b]$ и пусть $x_1,x_2,\ldots,x_n$ -некоторые точки этого отрезка. Тогда на отрезке $[a,b]$ найдется такая точка η, что

$$\frac{1}{n}\sum_{i=1}^n\varphi(x_i)=\varphi(\eta),\ |\beta_n|\le\frac{(b-a)^3M_2}{12n^2}. \tag{39}$$

Иными словами, среднее арифметическое значений непрерывной функции в нескольких точках отрезка $[a,b]$, равно ее значению в одной из точек этого отрезка.

Применяя это утверждение к суммам (27) и (28), получим другое представление остаточных членов α_n и β_n:

$$\alpha_n=\frac{(b-a)^3}{24n^2}f''(\eta^*),\ a\le\eta^*\le b, \tag{40}$$

$$\beta_n=-\frac{(b-a)^3}{12n^2}f''(\eta^{**}),\ a\le\eta^{**}\le b. \tag{41}$$

Формулы (40) и (41) не позволяют вычислить остаточные члены: существование точек η^* и η^{**} на отрезке $[a,b]$ гарантировано, но их положение неизвестно. Однако эти формулы можно использовать для оценки остаточных членов. Пусть известно число M_2, которое является мажорантой для второй производной функции $f(x)$:

$$|f''(x)|\le M_2,\ a\le x\le b, \tag{42}$$

тогда равенства (40) и (41) можно заменить неравенствами:

$$|\alpha_n|\le\frac{(b-a)^3M_2}{24n^2}, \tag{43}$$

$$|\beta_n|\le\frac{(b-a)^3M_2}{12n^2}. \tag{44}$$

При заданной точности ε они позволяют определить число узлов n, которое нужно использовать при вычислении интеграла по рассматриваемым квадратурным формулам.

В случае, когда вторая производная функции $f(x)$ является знакоопределенной на отрезке $[a,b]$, формулы (40) и (41) позволяют определить знаки остаточных членов. При этом существенно то, что они оказываются противоположными. Пусть, например,

$f''(x)\ge 0$, в этом случае $\alpha_n\ge 0$, $\beta_n\le 0$ так что для интеграла получается двухсторонняя оценка

$$P_n\le I\le T_n. \tag{45}$$

При отрицательной второй производной $f''(x)$ сохраняется двухсторонняя оценка, но знаки неравенств (45) меняются на противоположные. Такие оценки очень удобны, поскольку позволяют легко контролировать точность вычислений: в случае (45) P_n и T_n дают значение интеграла с недостатком и избытком с ошибкой, не превышающей $\varepsilon_n=T_n-P_n$, в противоположном случае P_n и T_n меняются ролями.

Заканчивая обсуждение методов прямоугольников и трапеций, сделаем следующее замечание. Формулы (32), (33), оценки (43), (44) показывают, что в случае дважды непрерывно дифференцируемой подынтегральной функции остаточные члены α_n и β_n убывают как n^{-2}. Однако, если отказаться от этого требования гладкости, то данные результаты теряют силу. В этом случае для интегрируемых функций можно гарантировать стремление остаточных членов к нулю, но нельзя утверждать, что оно происходит со скоростью n^{-2}.

Можно поставить прямо противоположный вопрос. Нельзя ли, повышая требование гладкости подынтегральной функции, увеличить скорость сходимости методов? Ответ на него отрицательный. Предположение о существовании у функции $f(x)$ четырех или шести производных не может изменить формул (32) и (33), так что скорость убывания остаточных членов при возрастании n останется прежней - n^{-2}. Поэтому методы прямоугольников и трапеций называют методами второго порядка точности, добавляя при этом – для дважды непрерывно дифференцируемых функций.

Задача 1.

Вычислить по формулам прямоугольников и трапеций при $n=2$ интеграл

$$I=\int_0^{\pi/2}\sin x dx=1. \tag{46}$$

В данном случае

$$P_2=\frac{\pi}{4}\left(\sin\frac{\pi}{8}+\sin\frac{3\pi}{8}\right)=1.026172, \tag{47}$$

$$T_2=\frac{\pi}{4}\left(\frac{1}{2}\sin 0+\sin\frac{\pi}{4}+\frac{1}{2}\sin\frac{\pi}{2}\right)=0.948059. \tag{48}$$

Зная точный ответ (46), найдем погрешности

$$\alpha_2=-0.026172\ \text{и}\ \beta_2=0.051941. \tag{49}$$

Вторая производная функции $\sin x$ на отрезке $[0,\pi/2]$ отрицательна, ее модуль не превышает единицы: $M_2=1$. Мы видим, что знаки погрешности α_2 и β_2 (49) согласуются с формулами (40) и (41). Они противоположны, так что для интеграла I справедлива двусторонняя оценка, аналогичная (45), но другого знака:

$$T_2\le I\le P_2. \tag{50}$$

Величина погрешностей (49) удовлетворяет неравенствам (43) и (44):

$$|\alpha_2| \le \frac{1}{96}\left(\frac{\pi}{2}\right)^3 < 0{,}041,\ |\beta_2| \le \frac{1}{48}\left(\frac{\pi}{2}\right)^3 < 0{,}081. \qquad (51)$$

Перейдем к обсуждению остаточного члена γ_n в методе Симпсона, которое проведем при предположении о четырехкратной непрерывной дифференцируемости подынтегральной функции $f(x)$. Напомним, что в методе Симпсона число точек n выбирается четным, так что $n/2$ является целым числом.

Рассмотрим отрезок двойной длины $2h$, расположенный между точками разбиения (6) с четными номерами $\left[x_{2j-2}, x_{2j}\right]$, $1 \le j \le n/2$. В курсе математического анализа выводится формула:

$$\int_{x_{2j-2}}^{x_{2j}} f(x)dx = \frac{h}{3}\left\{f\left(x_{2j-2}\right) + 4f\left(x_{2j-1}\right) + f\left(x_{2j}\right)\right\} - \frac{h^5}{90} f^{(4)}\left(\eta_j\right), \qquad (52)$$

где $\eta_j \in \left[x_{2j-2}, x_{2j}\right]$. Существование такой точки гарантировано, но ее точное положение на отрезке неизвестно.

Суммируя равенства (52) по j, получим квадратурную формулу (18) со следующим выражением для остаточного члена:

$$\gamma_n = -\frac{h^5}{90}\sum_{j=1}^{n/2} f^{(4)}\left(\eta_j\right). \qquad (53)$$

Из формулы (53), аналогичной формулам (27), (28), можно вывести различные представления остаточного члена и изучить его свойства.

Рассмотрим сумму

$$2h\sum_{j=1}^{n/2} f^{(4)}\left(\eta_j\right). \qquad (54)$$

Функция $f^{(4)}(x)$ предполагается непрерывной и, следовательно, интегрируемой на отрезке $[a,b]$. С учетом этого сумму (54) можно рассматривать как интегральную сумму для интеграла $\int_a^b f^{(4)}(x)dx$. Отсюда следует вывод

$$\lim_{h\to\infty} 2h\sum_{j=1}^{n/2} f^{(4)}\left(\eta_j\right) = \int_a^b f^{(4)}(x)dx = f'''(b) - f'''(a). \qquad (55)$$

Предельное равенство (55) позволяет записать остаточный член квадратурной формулы Симпсона (53) в виде

$$\gamma_n = \frac{1}{n^4}\left(C + \sigma_n\right), \qquad (56)$$

$$C = -\frac{(b-a)^4}{180}\left\{f'''(b) - f'''(a)\right\}, \qquad (57)$$

$$\sigma_n = -\frac{(b-a)^4}{180}\left\{2h\sum_{j=1}^{n/2} f^{(4)}\left(\eta_j\right) - \int_a^b f^{(4)}(x)dx\right\} \to 0, \text{ при } n \to \infty. \qquad (58)$$

Эта формула, как и формулы (32), (33) для методов прямоугольников и трапеций, выделяет в остаточном члене γ_n главное слагаемое C/n^4, которое стремится к нулю как n^{-4}. Коэффициент C (57) не зависит от n. Дополнительное слагаемое σ_n/n^4 является бесконечно малой более высокого порядка. Если им пренебречь, то получится асимптотическое представление остаточного члена

$$\gamma_n \approx Cn^{-4}. \qquad (59)$$

Его относительная точность возрастает с увеличением n.

Другое представление остаточного члена γ_n можно вывести с помощью формулы (39). Она позволяет записать формулу (53) в виде

$$\gamma_n = -\frac{(b-a)^5}{180n^4} f^{(4)}(\eta), \qquad (60)$$

где η - какая-то точка отрезка $[a,b]$. Вычислить погрешность по формуле (60) нельзя, поскольку положение точки η неизвестно, но можно ее оценить. Пусть $\left|f^{(4)}(x)\right| \le M_4$, тогда

$$|\gamma_n| \le \frac{(b-a)^5 M_4}{180n^4}. \qquad (61)$$

Данная оценка позволяет определить, с каким n нужно проводить вычисления, чтобы погрешность не превышала заданной точности ε. Кроме того, если четвертая производная функции $f(x)$ является знакоопределенной, то формула (60) дает знак погрешности, что также может оказаться полезным при организации вычислений.

Метод Симпсона является методом более высокого порядка точности – четвертого. В этом его преимущество перед методами прямоугольников и трапеций. Правда, приведенные выше оценки остаточного члена, требуют большей гладкости подынтегральной функции – она должна быть четыре раза непрерывно дифференцируема.

Задача 2.

Вычислить интеграл (46) по формуле Симпсона при $n = 2$.

В данном случае

$$S_2 = \frac{\pi}{12}\left(\sin 0 + 4\sin\frac{\pi}{4} + \sin\frac{\pi}{2}\right) = 1.002280, \qquad (62)$$

$$\gamma_2 = -0.002280. \qquad (63)$$

Четвертая производная функции $\sin x$ на отрезке $\left[0, \frac{\pi}{2}\right]$ положительна и не превосходит единицы, так что знак погрешности согласуется с формулой (60), а ее величина – с оценкой (61):

$$|\gamma_2| \le \frac{1}{180\cdot 16}\left(\frac{\pi}{2}\right)^5 < 0.0034.$$

2.3. Апостериорные оценки погрешности при численном интегрировании.

В латинском языке существуют два термина – антонима: *априори* (***a priori***) и *апостериори* (***a posteriori***). Первый означает изначально, независимо от опыта, второй – на основании опыта. Оба они часто используются в вычислительной математике, подразделяя информацию на ту, которая известна до начала вычислений, и ту, которая получается в процессе вычислений.

Оценки погрешности квадратурных формул прямоугольников (43), трапеций (44), Симпсона (61) называют априорными. Они справедливы изначально и предсказывают точность вычисления интеграла независимо от того, будем мы фактически проводить вычисления или нет. Эти результаты позволяют понять структуру остаточных членов, определить скорость их убывания при возрастании n.

Однако недаром говорят, что недостатки являются продолжением достоинств. Постановка задачи численного интегрирования предполагает, что известен алгоритм вычисления подынтегральной функции $f(x)$ при любом значении аргумента x на отрезке $[a,b]$ и все. В оценки же (43), (44), (61) входят константы M_2 и M_4, мажерирующие вторую и четвертую производные функции $f(x)$ в асимптотические формулы (38) и (59) – значения первой и третьей производных в граничных точках отрезка $[a,b]$. Такая информация выходит за рамки первоначальной постановки задач. Чтобы ее получить и использовать в процессе вычислений, нужно провести дополнительное исследование функции $f(x)$. В случае, когда функция $f(x)$ задана сравнительно простой формулой, такое исследование возможно, хотя требует определенных усилий и времени. В случае же, когда она задается графиком, таблицей, определяется как сложная неявная функция и т. д., на этом пути возникают большие или даже непреодолимые трудности. В связи с этим перейдем к обсуждению методов оценки погрешности численного интегрирования, которые не требуют предварительного анализа производных подынтегральной функции. Они используют сопоставление результатов вычислений с разным числом точек n и называются апостериорными (буквально, основанными на опыте, что в данном случае означает основанными на результатах вычислений).

Начнем обсуждение идеи апостериорных оценок погрешности с методов второго порядка – прямоугольников и трапеций. Предположим, что мы провели расчеты по методу прямоугольников с числом точек $n/2$ (n - четное число), а потом с числом точек n и в результате получили два числа - $P_{n/2}$ и P_n. Согласно формулам (9) и (32) это позволяет написать соотношения

$$\begin{aligned} I &= P_{n/2} + \frac{4}{n^2}\left(A + \mu_{n/2}\right), \\ I &= P_n + \frac{1}{n^2}\left(A + \mu_n\right). \end{aligned} \qquad (64)$$

Вычитая теперь второе равенство из первого, получим

$$\left(P_{n/2} - P_n\right) + \frac{3}{n^2}A + \frac{1}{n^2}\left(4\mu_{n/2} - \mu_n\right) = 0$$

или

$$\alpha_n = \frac{1}{n^2}\left(A + \mu_n\right) = \frac{1}{3}\left(P_n - P_{n/2}\right) + \frac{4}{3n^2}\left(\mu_n - \mu_{n/2}\right). \qquad (65)$$

Первый член в правой части этого представления остаточного члена нам известен из результатов вычислений. Он является главным. Второй член неизвестен, но он, по сравнению с первым, представляет собой бесконечно малую более высокого порядка. Если им пренебречь, то для погрешности получится простая асимптотическая формула:

$$\alpha_n \approx \frac{1}{3}\left(P_n - P_{n/2}\right). \qquad (66)$$

Ее относительная точность возрастает при увеличении n.

Аналогичные формулы имеют место для погрешности метода трапеций

$$\beta_n = \frac{1}{3}\left(T_n - T_{n/2}\right) + \frac{4}{3n^2}\left(\nu_n - \nu_{n/2}\right) \approx \frac{1}{3}\left(T_n - T_{n/2}\right). \qquad (67)$$

Для метода Симпсона, который является методом четвертого порядка, формулы немного изменяются. Теперь соотношения, аналогичные (64), будут иметь вид:

$$\begin{aligned} I &= S_{n/2} + \frac{16}{n^4}\left(C + \sigma_{n/2}\right), \\ I &= S_n + \frac{1}{n^4}\left(C + \sigma_n\right). \end{aligned} \qquad (68)$$

(Здесь число n предполагается кратным четырем, так что $n/2$ четное число.) Проводя в (68) вычитание второй строки из первой, получим

$$\gamma_n = \frac{1}{n^4}\left(C + \sigma_n\right) = \frac{1}{15}\left(S_n - S_{n/2}\right) + \frac{16}{15n^4}\left(\sigma_n - \sigma_{n/2}\right). \qquad (69)$$

Здесь опять первый член в правой части равенства известен из вычислений. Он является главным. Второй член неизвестен, но он представляет собой бесконечно малую более высокого порядка по сравнению с первым. Если им пренебречь, то получим асимптотическую формулу для приближенного вычисления погрешности по результатам двух вычислений

$$\gamma_n \approx \frac{1}{15}\left(S_n - S_{n/2}\right). \qquad (70)$$

Ее относительная точность возрастает с увеличением n.

Обычно апостериорные оценки погрешности с помощью асимптотических формул (66), (67), (70) включают в компьютерные программы численного интегрирования. Они служат критерием для завершения вычислений после того, как нужная точность достигнута.

В заключение отметим следующее. Можно подставить полученные выражения для остаточных членов (65), (67), (69) в исходные квадратурные формулы (9), (14) и (18). В результате они примут вид:

$$I = \frac{4}{3}P_n - \frac{1}{3}P_{n/2} + \tilde{\alpha}_n, \qquad (71)$$

$$I = \frac{4}{3}T_n - \frac{1}{3}T_{n/2} + \tilde{\beta}_n, \qquad (72)$$

$$I = \frac{16}{15}S_n - \frac{1}{15}S_{n/2} + \tilde{\gamma}_n, \qquad (73)$$

где $\tilde{\alpha}_n$, $\tilde{\beta}_n$, $\tilde{\gamma}_n$ - остаточные члены этих модифицированных формул

$$\tilde{\alpha}_n = \frac{4}{3n^2}\left(\mu_n - \mu_{n/2}\right) = o\left(n^{-2}\right), \qquad (74)$$

$$\tilde{\beta}_n = \frac{4}{3n^2}\left(\nu_n - \nu_{n/2}\right) = o\left(n^{-2}\right), \qquad (75)$$

$$\tilde{\gamma}_n = \frac{16}{15n^2}\left(\sigma_n - \sigma_{n/2}\right) = o\left(n^{-4}\right). \qquad (76)$$

Формулы (71), (72), (73), написанные по результатам двух расчетов с числом точек $n/2$ и n, являются асимптотически более точными, чем исходные. В исходных формулах погрешности убывают, соответственно, как n^{-2}, n^{-2}, n^{-4}, в модифицированных формулах погрешности, согласно (74), (75), (76) являются бесконечно малыми более высокого порядка. Однако для исходных формул известны оценки погрешностей (43), (44). (61). Для модифицированных формул в нашем распоряжении оценок нет. Если мы хотим ими пользоваться, то нужно провести соответствующее исследование. Исключение составляет формула (72). Согласно формуле (21) ее можно переписать в виде

$$I = S_n + \tilde{\beta}_n, \qquad (77)$$

т. е. модифицированная формула трапеций оказалась просто формулой Симпсона с уже известным остаточным членом $\gamma_n = \tilde{\beta}_n$.

Задача 3.

Вычислить по формуле Симпсона интеграл (46) с $n = 4$. Используя результаты задачи 2, найти приближенную апостериорную погрешность (70).

В данном случае

$$S_4 = \frac{\pi}{24}\left(\sin 0 + 4\sin\frac{\pi}{8} + 2\sin\frac{\pi}{4} + 4\sin\frac{3\pi}{8} + \sin\frac{\pi}{2}\right) = 1.000135, \qquad (78)$$

$$\gamma_4 = -0.000135.$$

Апостериорная оценка погрешности по результатам двух расчетов дает

$$\gamma_4 \approx \frac{1}{15}\left(S_4 - S_2\right) = -0.000143.$$

Несмотря на маленькое число точек, она хорошо согласуется с фактической погрешностью (78), сосчитанной «в лоб» по известному значению интеграла (46).

Задача 4.

Используя результаты решения задач 2 и 3, посчитать интеграл (46) по модифицированной формуле Симпсона (73).

В данном случае

$$I \approx \frac{16}{15}S_4 - \frac{1}{15}S_2 = 0.999992, \qquad (79)$$

$$\tilde{\gamma}_4 = 0.000008.$$

Модифицированная формула Симпсона (73) без дополнительных вычислений позволила на порядок улучшить результат, полученный по обычной формуле Симпсона. Отметим, что погрешности при расчетах по формулам (78) и (79) имеют противоположные знаки.

§3. Квадратурные формулы Гаусса.

3.1. Задача построения оптимальных квадратурных формул.

Точность квадратурной формулы определяется выбором узлов и весовых коэффициентов. Например, формулы трапеций и Симпсона имеют одинаковые узлы, но различные веса и, как следствие, их точность оказывается разной. В связи с этим естественно возникает задача поиска наилучшей квадратурной формулы с заданным числом узлов n. Обсудим постановку и решение такой задачи в формулировке Гаусса: построить квадратурную формулу с числом узлов n, которая является точной для любого полинома степени $(2n-1)$ или ниже. Такая постановка задачи вполне оправдана: квадратурная формула, точная для полиномов, будет хорошо работать для гладких функций.

Переходя к решению задачи, поставленной Гауссом, будем считать, что интеграл предварительно приведен к стандартной форме, когда областью интегрирования является отрезок $[-1,1]$. С учетом этого замечания запишем искомую квадратурную формулу в виде:

$$\int_{-1}^{1} f(x)dx = \sum_{i=1}^{n} c_i f\left(x_i\right) + \delta_n, \qquad (80)$$

где x_i узлы, $x_i \in [-1,1]$, c_i весовые коэффициенты, δ_n остаточный член. Для любого полинома степени $(2n-1)$ остаточный член в формуле (80) должен быть равен нулю. На протяжении этого параграфа каждый раз, когда мы будем говорить о произвольных полиномах какой-нибудь степени, всегда будем включать в их число полиномы более низких степеней, не оговаривая это особо.

Полагая последовательно $f(x) = 1, x, x^2, \ldots, x^{2n-1}$ и принимая во внимание, что для этих функций, согласно требованию Гаусса, остаточный член должен равняться нулю, получим:

$$\int_{-1}^{1} x^m dx = \frac{1}{(m+1)}\left\{1 + (-1)^m\right\} = \sum_{i=1}^{n} c_i x_i^m, \ 0 \le m \le 2n-1. \qquad (81)$$

Соотношения (81) представляют собой систему $2n$ нелинейных уравнений с $2n$ неизвестными, в качестве которых выступают узлы x_i и веса c_i $(1 \le i \le n)$.

Уравнение (81), соответствующее индексу $m = 0$, дает

$$\sum_{i=1}^{n} c_i = 2. \qquad (82)$$

Таким образом, сумма весовых коэффициентов в квадратурной формуле Гаусса при любом n равна двум.

Задача 5.

Составить и решить систему уравнений (81) для квадратурной формулы Гаусса с одним узлом.

В этом случае в задаче подлежат определению два параметра: узел x_1 и весовой коэффициент c_1. Система уравнений для их определения получается из (81) при $m = 0$ и $m = 1$:

$$\begin{cases} c_1 = 2 \\ c_1 x_1 = 0 \end{cases}.$$

Ее решение имеет вид: $x_1 = 0$, $c_1 = 2$, так что искомая квадратурная формула запишется следующим образом:

$$\int_{-1}^{1} f(x)dx = 2f(0) + \delta_1. \qquad (83)$$

Выбор в качестве единственного узла средней точки отрезка $[-1,1]$ выглядит по соображениям симметрии вполне естественно. Требование, чтобы сумма весовых коэффициентов равнялась двум (82), определяет в данном случае единственный весовой коэффициент c_1. Квадратурная формула (83) является точной для любой линейной функции $Q_1 = a_0 + a_1 x$.

3.2. Полиномы Лежандра

Мы решили систему уравнений (81) при $n = 1$. Однако решить ее «в лоб» в общем случае при произвольном n сложно. Поэтому мы будем вынуждены воспользоваться обходным путем. Для этой цели нам понадобятся полиномы Лежандра, с которыми Вы уже встречались в курсе линейной алгебры. Они определяются формулами

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\left(x^2 - 1\right)^n. \qquad (84)$$

Выпишем, используя эту формулу, несколько первых полиномов Лежандра

$$P_0(x) = 1,\ P_1(x) = x,\ P_2(x) = \frac{3}{2}x^2 - \frac{1}{2},\ P_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x. \qquad (85)$$

Полиномы Лежандра обладают следующими свойствами:

1. Полином Лежандра $P_n(x)$ номера n является полиномом n -ой степени, обладающим той же четностью, что и n:

$$P_n(-x) = (-1)^n P_n(x). \qquad (86)$$

2. Полиномы Лежандра $P_n(x)$ в точках $x = \pm 1$ принимают следующие значения:

$$P_n(1) = 1,\ P_n(-1) = (-1)^n.$$

3. Полином Лежандра $P_n(x)$ имеет на интервале $(-1,1)$ n простых корней. В силу свойства 1 корни располагаются симметрично относительно точки $x = 0$.

4. Любой полином $Q_m(x)$ степени $m < n$ ортогонален к полиному Лежандра $P_n(x)$ на сегменте $[-1,1]$:

$$\int_{-1}^{1} Q_m(x)P_n(x)dx = 0. \qquad (87)$$

Докажем перечисленные свойства.

1. Свойство 1 напрямую следует из формулы (84).

2. Представим выражение $\left(x^2 - 1\right)^n$ в виде произведения

$$\left(x^2 - 1\right)^n = (x+1)^n (x-1)^n$$

и выполним n - кратное дифференцирование. В результате получим:

$$P_n(x) = \frac{1}{2^n n!}\sum_{k=0}^{n}\left(C_n^k\right)^2 n!(x+1)^{n-k}(x-1)^k.$$

Все члены этой суммы, кроме нулевого, содержат множители $(x-1)^k$: $1 \le k \le n$ и при $x = 1$ обращаются в ноль, а нулевой член дает нужное равенство: $P_n(1) = 1$. Второе равенство следует из (86): $P_n(-1) = (-1)^n$.

3. Функция $\left(x^2 - 1\right)^n$ обращается на концах отрезка $[-1,1]$ в ноль. Согласно теореме Ролля ее первая производная должна иметь по крайней мере один ноль на интервале $(-1,1)$. Кроме того, производная обращается в ноль в граничных точках $x = \pm 1$. Применяя таким же образом теорему Ролля ко второй производной $\left\{\left(x^2 - 1\right)^n\right\}'$, убеждаемся в том, что она имеет два нуля на интервале $(-1,1)$ и обращается в ноль в граничных точках $x = \pm 1$.

Будем продолжать этот процесс, пока не дойдем до n -ой производной выражения $\left(x^2 - 1\right)^n$. Эта производная определяет полином Лежандра с точностью до множителя. Она должна иметь n корней на интервале $(-1,1)$. Поскольку число корней равно степени полинома, все они должны быть простыми. Корни, как мы уже отмечали выше, располагаются на интервале $(-1,1)$ симметрично относительно его средней точки $x = 0$.

4. Подставим в интеграл (87) представление полинома Лежандра (84) и проинтегрируем по частям. В результате получим:

$$J = \frac{1}{2^n n!}\int_{-1}^{1} Q_m(x)\frac{d^n}{dx^n}\left(x^2 - 1\right)^n dx =$$

$$= \frac{1}{2^n n!}\left\{Q_m(x)\frac{d^{n-1}}{dx^{n-1}}\left(x^2 - 1\right)^n\Big|_{-1}^{1} - \int_{-1}^{1}\frac{dQ_m(x)}{dx}\frac{d^{n-1}}{dx^{n-1}}\left(x^2 - 1\right)^n dx\right\}.$$

Подстановки на концах отрезка $[-1,1]$ обращаются в ноль, поскольку степень n у выражения $\left(x^2 - 1\right)^n$ больше $(n-1)$-го порядка производной.

Выполняя процедуру интегрирования по частям $m+1 \le n$ раз, получим:

$$J = (-1)^{m+1}\frac{1}{2^n n!}\int_{-1}^{1}\frac{d^{m+1}Q_m(x)}{dx^{m+1}}\frac{d^{n-m-1}}{dx^{n-m-1}}\left(x^2 - 1\right)^n dx = 0.$$

Здесь под знаком интеграла в качестве множителя стоит $(m+1)$-ая производная от полинома m-ой степени $Q_m(x)$, тождественно равная нулю. Ортогональность доказана.

Сделаем важное замечание. Соотношение ортогональности (87) справедливо, в частности, в случае, когда в качестве полинома $Q_m(x)$ взят полином Лежандра $P_m(x)$:

$$\int_{-1}^{1} P_m(x)P_n(x)dx = 0, \text{ при } m < n.$$

Фактически в этом условии ортогональности не важно, какой именно из двух индексов m или n больше, а какой меньше. Важно лишь, что они не равны. Таким образом, из свойства 4 вытекает следствие.

Следствие 1.

Полиномы Лежандра образуют систему полиномов, ортогональных на отрезке $[-1,1]$

$$\int_{-1}^{1} P_m(x)P_n(x)dx = 0, \text{ при } m \ne n. \qquad (89)$$

Из линейной алгебры известно, что система полиномов, ортогональных на некотором множестве, определена однозначно с точностью до множителей. Поэтому следствию 1 можно сопоставить обратное утверждение.

Следствие 2

Любая система полиномов, ортогональных на отрезке $[-1,1]$, совпадает с точностью до множителя с системой полиномов Лежандра.

3.3. Узлы и весовые коэффициенты квадратурных формул Гаусса.

Изучив свойства полиномов Лежандра, перейдем к решению основной задачи – определению узлов и весовых коэффициентов квадратурных формул Гаусса. Составим полином n-ой степени

$$\omega_n(x) = (x - x_1)(x - x_2)\cdots(x - x_n), \qquad (90)$$

где x_i - искомые узлы. Возьмем произвольный полином $Q_m(x)$ степени $m < n$, помножим его на полином $\omega_n(x)$ и проинтегрируем произведение по отрезку $[-1,1]$ с помощью квадратурной формулы (80). Поскольку это произведение представляет

собой полином степени $m+n\le 2n-1$, формула Гаусса должна быть для него точной. В результате согласно (90) получим:

$$\int_{-1}^{1} Q_m(x)\omega_n(x)dx=\sum_{i=1}^{n} c_i Q_m(x_i)\omega_n(x_i)=0. \qquad (91)$$

Мы видим, что полином $\omega_n(x)$ ортогонален к любому полиному степени $m<n$ в том числе и к полиномам Лежандра индекса $m<n$. Это означает, что он с точностью до множителя совпадает с n-ым полиномом Лежандра: $\omega_n(x)=A_n P_n(x)$. Отсюда следует вывод: узлы квадратурной формулы Гаусса являются корнями полинома Лежандра $P_n(x)$. Напомним, что корни полиномов Лежандра располагаются на интервале $(-1,1)$ симметрично относительно его средней точки $x=0$.

Для того, чтобы подсчитать весовые коэффициенты c_i, введем специальные полиномы

$$Q_{n-1,m}(x)=\frac{(x-x_1)\cdots(x-x_{m-1})(x-x_{m+1})\cdots(x-x_n)}{(x_m-x_1)\cdots(x_m-x_{m-1})(x_m-x_{m+1})\cdots(x_m-x_n)} \qquad (92)$$

Каждый из них является полиномом степени $(n-1)$. В числителе у него стоит полином $\omega_n(x)$ с опущенным множителем $(x-x_m)$, в знаменателе - значение числителя в точке $x=x_m$. В результате такой структуры полином $Q_{n-1,m}(x)$ в точках x_i удовлетворяет соотношениям:

$$Q_{n-1,m}(x_i)=\begin{cases}0, & i\ne m\\ 1, & i=m\end{cases}. \qquad (93)$$

Для полинома $Q_{n-1,m}(x)$ квадратурная формула Гаусса должна быть точной. С учетом (93) это дает

$$\int_{-1}^{1} Q_{n-1,m}(x)dx=\sum_{i=1}^{n} c_i Q_{n-1,m}(x_i)=c_m. \qquad (94)$$

В результате получаем следующее интегральное выражение для весовых коэффициентов квадратурной формулы Гаусса:

$$c_m=\int_{-1}^{1} Q_{n-1,m}(x)dx=\int_{-1}^{1}\frac{(x-x_1)\cdots(x-x_{m-1})(x-x_{m+1})\cdots(x-x_n)}{(x_m-x_1)\cdots(x_m-x_{m-1})(x_m-x_{m+1})\cdots(x_m-x_n)}dx. \qquad (95)$$

3.4. Исследование квадратурной формулы.

Нам осталось решить последний вопрос – доказать, что квадратурная формула, у которой в качестве узлов x_i берутся корни полинома Лежандра, а весовые коэффициенты c_i вычисляются по формулам (95), действительно решают задачу Гаусса, являясь точной для любого полинома степени $(2n-1)$.

Проведем доказательство в два этапа. Сначала докажем, что такая формула является точной для любого полинома $Q_{n-1}(x)$ степени $(n-1)$. Такой полином можно представить в виде суммы специальных полиномов (92)

$$Q_{n-1}(x)=\sum_{m=1}^{n} Q_{n-1}(x_m)Q_{n-1,m}(x). \qquad (96)$$

Справедливость данного разложения вытекает из следующих соображений. Здесь левая и правая части равенства совпадают в n точках $x_i, 1\le i\le n$. Но, если два полинома $(n-1)$-ой степени совпадают в n точках, то они тождественно равны.

Интегрируя равенство (96) по отрезку $[-1,1]$, получим

$$\int_{-1}^{1} Q_{n-1}(x)dx=\sum_{m=1}^{n} Q_{n-1}(x_m)\int_{-1}^{1} Q_{n-1,m}(x)dx=\sum_{m=1}^{n} c_m Q_{n-1}(x_m). \qquad (97)$$

Итак, для полиномов $(n-1)$-ой степени утверждение доказано.

Теперь рассмотрим произвольный полином $Q_{2n-1}(x)$ степени $(2n-1)$. Разделим его с остатком на полином Лежандра $P_n(x)$ и представим в виде:

$$Q_{2n-1}(x)=P_n(x)q_{n-1}(x)+r_{n-1}(x), \qquad (98)$$

где $q_{n-1}(x)$ и $r_{n-1}(x)$ полиномы степени $(n-1)$. Проинтегрировав равенство (98) по отрезку $[-1,1]$, будем иметь:

$$\begin{aligned}\int_{-1}^{1} Q_{2n-1}(x)dx&=\int_{-1}^{1}\{P_n(x)q_{n-1}(x)+r_{n-1}(x)\}dx=\int_{-1}^{1} r_{n-1}(x)dx=\\ &=\sum_{i=1}^{n} c_i r_{n-1}(x_i)=\sum_{i=1}^{n} c_i\{P_n(x_i)q_{n-1}(x_i)+r_{n-1}(x_i)\}=\sum_{i=1}^{n} c_i Q_{2n-1}(x_i).\end{aligned} \qquad (99)$$

Поясним выполненные преобразования. Интеграл $\int_{-1}^{1} P_n(x)q_{n-1}(x)dx$ опущен, поскольку полином Лежандра $P_n(x)$ ортогонален к любому полиному $(n-1)$-ой степени. Оставшийся интеграл от полинома $r_n(x)$ вычислен с помощью квадратурной формулы (97). Выше уже доказано, что для полиномов степени $(n-1)$ она является точной. Последний переход заключается в том, что в сумму $\sum_{i=1}^{n} c_i r_{n-1}(x_i)$ добавлены слагаемые $P_n(x_i)q_{n-1}(x_i)$. Они не меняют значения суммы, поскольку все равны нулю: ведь узлами квадратурной формулы являются корни полинома Лежандра $P_n(x)$.

Итак, построенная квадратурная формула действительно является точной для любого полинома степени $(2n-1)$, т. е. задача Гаусса решена. На оценке погрешности квадратурных формул Гаусса мы останавливаться не будем, однако задачи, к разбору которых переходим, показывают, что эти формулы обеспечивают для гладких функций очень высокую точность.

Задача 5.
Построить квадратурную формулу Гаусса с двумя и тремя узлами.

Выведем сначала квадратурную формулу с двумя узлами. Узлы определяются как корни второго полинома Лежандра, выражение для которого мы выписывали выше (85). В данном случае имеем:

$$x_1=-1/\sqrt{3},\ x_2=1/\sqrt{3}. \qquad (100)$$

Узлы расположены симметрично относительно точки $x=0$.

Весовые коэффициенты рассчитываются по формуле (95):

$$\begin{aligned}c_1&=\int_{-1}^{1}\frac{x-x_2}{x_1-x_2}dx=\int_{-1}^{1}\frac{x-1/\sqrt{3}}{-2/\sqrt{3}}dx=1,\\ c_2&=\int_{-1}^{1}\frac{x-x_1}{x_2-x_1}dx=\int_{-1}^{1}\frac{x+1/\sqrt{3}}{2/\sqrt{3}}dx=1.\end{aligned} \qquad (101)$$

Они равны между собой, а их сумма, в соответствии с общим соотношением (82), равна двум. В результате искомая квадратурная формула принимает вид:

$$\int_{-1}^{1} f(x)dx=f\left(-1/\sqrt{3}\right)+f\left(1/\sqrt{3}\right)+\delta_2. \qquad (102)$$

Она является точной для любого полинома третьей степени.

Перейдем теперь к выводу квадратурной формулы Гаусса с тремя узлами. Согласно формуле (85) для третьего полинома Лежандра ее узлами являются числа:

$$x_1=-\sqrt{3/5},\ x_2=0,\ x_3=\sqrt{3/5}. \qquad (103)$$

Остается подсчитать весовые коэффициенты:

$$\begin{aligned}c_1&=\int_{-1}^{1}\frac{x\left(x-\sqrt{3/5}\right)}{\left(-\sqrt{3/5}\right)\left(-2\sqrt{3/5}\right)}dx=\frac{5}{9},\\ c_2&=\int_{-1}^{1}\frac{\left(x+\sqrt{3/5}\right)\left(x-\sqrt{3/5}\right)}{-3/5}dx=\frac{8}{9},\\ c_3&=\int_{-1}^{1}\frac{\left(x+\sqrt{3/5}\right)x}{\left(\sqrt{3/5}\right)\left(2\sqrt{3/5}\right)}dx=\frac{5}{9}.\end{aligned} \qquad (104)$$

В результате квадратурная формула Гаусса с тремя узлами запишется в виде:

$$\int_{-1}^{1} f(x)dx=\frac{5}{9}f\left(-\sqrt{3/5}\right)+\frac{8}{9}f(0)+\frac{5}{9}f\left(\sqrt{3/5}\right)+\delta_3. \qquad (105)$$

Она является точной для любого полинома пятой степени.

Задача 6.
Вычислить по формулам Симпсона и Гаусса при $n=2$ интеграл:

$$\int_{-1}^{1} e^x dx=2sh1=2.350402.$$

Сравнить результаты численного интегрирования с точным значением интеграла и между собой.

Формулы Симпсона и Гаусса дают в данном случае следующие результаты:

$$\begin{aligned}S_2&=\frac{1}{3}\left(e^{-1}+4+e\right)=\frac{4}{3}+\frac{2}{3}ch1=2.362054,\\ \gamma_2&=-0.011651,\\ G_2&=\left(e^{-1/\sqrt{3}}+e^{1/\sqrt{3}}\right)=2ch\frac{1}{\sqrt{3}}=2.342696,\\ \delta_2&=0.007706.\end{aligned}$$

Мы видим, что даже с двумя узлами формула Гаусса дает хороший ответ. Его точность выше точности ответа, полученного по формуле Симпсона.

В заключение сделаем следующее замечание. Несмотря на высокую точность квадратурных формул Гаусса, при компьютерных расчетах ими пользуются сравнительно редко. Дело в том, что для применения метода Гаусса нужно либо ввести в компьютер до начала расчетов корни полинома Лежандра и весовые коэффициенты, либо составить специальную подпрограмму для их вычисления. В результате потери человеческого и машинного времени на подготовку программы к основному расчету, связанному с вычислением интеграла, могут не окупиться точностью метода Гаусса. Вычисление интеграла по более простой схеме метода Симпсона имеет с этой точки зрения преимущество.

§4. Построение первообразной с помощью численного интегрирования.

Формулы Ньютона-Лейбница (1) позволяет выразить значение определенного интеграла от функции $f(x)$ через ее первообразную $F(x)$. В математическом анализе устанавливается и прямо противоположная возможность: первообразная функции $f(x)$, непрерывной на отрезке $[a,b]$, может быть записана в виде определенного интеграла с переменным верхним пределом:

$$F(x)=\int_{x_0}^{x} f(t)dt. \qquad (106)$$

Здесь x_0, x - две точки отрезка $[a,b]$, причем нижний предел интегрирования x_0 предполагается фиксированным, верхний x - переменным. В случае непрерывной функции $f(x)$ функция $F(x)$, определенная с помощью интеграла (106), является дифференцируемой и ее производная равна $f(x)$:

$$F'(x)=\frac{d}{dx}\left(\int_{x_0}^{x} f(t)dt\right)=f(x). \qquad (107)$$

Формула (106) в сочетании с какой-нибудь формулой численного интегрирования, например, Симпсона, представляет собой универсальный алгоритм построения первообразной. Приведем два примера, иллюстрирующие этот алгоритм.

Функция $f(x)=\sin x/x$ непрерывна и, следовательно, имеет первообразные. Они не могут быть выражены через элементарные функции, но представление в виде интеграла с переменным верхним пределом для них справедливо. Одну из

первообразных мы получим, выбирая нижний предел интегрирования $x_0=0$. Ее называют интегральным синусом и обозначают

$$Si(x)=\int_{0}^{x}\frac{\sin t}{t}dt.$$

Интегральный синус определен на всей числовой прямой, является нечетной функцией x, имеет конечные предельные значения на бесконечности

$$\lim_{x\to\pm\infty} Si(x)=\pm\frac{\pi}{2}.$$

Согласно (107)

$$Si'(x)=\sin x/x.$$

По знаку производной легко определить области возрастания и убывания функции, разделенные точками экстремума $x_k=k\pi\ \ (k=\pm1,\pm2,\ldots)$. Методы численного интегрирования позволяют вычислить значения $Si(x)$ при любом x. График интегрального синуса при $x\ge 0$ приведен на рис. 3.

В качестве второго примера рассмотрим функцию ошибок $erf(x)$, играющую важную роль в теории вероятности. Ее обозначение образовано с помощью первых букв английского названия функции ошибок – error function. Подобно интегральному синусу, функция ошибок вводится в виде интеграла с переменным пределом от функции e^{-x^2}, которая не имеет первообразных в классе элементарных функций:

$$erf(x)=\frac{2}{\sqrt{\pi}}\int_{0}^{x} e^{-t^2}dt.$$

Функция ошибок определена на всей числовой прямой, является нечетной функцией x, имеет конечные предельные значения на бесконечности:

$$\lim_{x\to\pm\infty} erf(x)=\pm1.$$

Согласно (107)

$$erf'(x)=\frac{2}{\sqrt{\pi}}e^{-x^2}.$$

Производная всюду положительная, следовательно, функция ошибок монотонно возрастает. Ее график приведен на рис. 4.

Существует ряд других специальных функций, которые вводятся как интегралы с переменным верхним пределом. Не будем останавливаться на их описании, отметим лишь, что разобранные примеры показывают, насколько условно деление функций на элементарные и неэлементарные. По существу, чтобы работать с какой-нибудь функцией, нужно знать ее свойства и иметь алгоритм вычисления при любом значении аргумента. С этой точки зрения применение интегрального синуса или функции ошибок ничем не отличается от применения привычных нам элементарных функций.

Глава 4. ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ ОБЫКНОВЕНЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Наиболее универсальными методами численного решения обыкновенных дифференциальных уравнений являются разностные методы. Они основаны на замене производных в дифференциальном уравнении разностными отношениями. В результате исходное дифференциальное уравнение сводится к системе алгебраических уравнений, которые называются разностными. Решение этой системы дает приближенное решение исходной задачи.

§1. Разностная аппроксимация производных.

1.1. Сеточные функции.

Пусть на отрезке $[a,b]$ задан набор точек

$$a=x_0<x_1<x_2<\cdots<x_{n-1}<x_n=b. \qquad (1)$$

Будем называть его сеткой. Чтобы не усложнять изложения, условимся считать сетку равномерной:

$$x_{i+1}-x_i=h=(b-a)/n,\ 0\le i\le n-1. \qquad (2)$$

Пусть каждой точке сетки x_i сопоставлено по определенному закону число y_i. Совокупность этих чисел $\mathbf{y}=\{y_0,y_1,\ldots,y_n\}=\{y_i\}\ \ (0\le i\le n)$ назовем сеточной функцией. Сеточные функции, определенные на сетке (1), образуют $(n+1)$-мерное линейное пространство.

Чтобы иметь возможность сравнивать сеточные функции между собой, говорить об их близости, нужно ввести в этом пространстве норму. В этой главе мы будем пользоваться нормой C, которая определяется следующим образом:

$$\|\mathbf{y}\|_C=\max_{0\le i\le n}|y_i|. \qquad (3)$$

Это определение законно, поскольку удовлетворяет трем аксиомам нормы:

1. Норма неотрицательна

$$\|\mathbf{y}\|_C\ge 0,$$

причем равенство нулю имеет место только для нулевого элемента.

2. Модуль числового множителя можно вынести за знак нормы

$$\|\alpha\mathbf{y}\|_C=|\alpha|\|\mathbf{y}\|_C.$$

3. Неравенство треугольника

$$\|\mathbf{y}+\mathbf{z}\|_C\le\|\mathbf{y}\|_C+\|\mathbf{z}\|_C. \qquad (4)$$

Справедливость последнего утверждения вытекает из свойства максимума:

$$\max_{0\le i\le n}|y_i+z_i|\le\max_{0\le i\le n}|y_i|+\max_{0\le i\le n}|z_i|.$$

1.2. Разностные аппроксимации первой производной.

Для сеточных функций нельзя ввести обычное понятие производной, включающее операцию предельного перехода при $\Delta x\to 0$. Вместо производной здесь вводятся разностные отношения:

$$L_h^+[y_i]=\frac{y_{i+1}-y_i}{h},\ 0\le i\le n-1; \qquad (5)$$

$$L_h^-[y_i]=\frac{y_i-y_{i-1}}{h},\ 1\le i\le n; \qquad (6)$$

$$L_h^{(0)}[y_i]=\frac{y_{i+1}-y_{i-1}}{2h},\ 1\le i\le n-1. \qquad (7)$$

Отношение (5) называют правой разностной производной, отношение (6) – левой разностной производной и отношение (7) – центральной разностной производной.

Чтобы установить связь разностных отношений (5) – (7) с обычной производной, предположим, что на отрезке $[a,b]$ определена дифференцируемая функция $y(x)$, значения которой в точках сетки (1) равны значениям рассматриваемой сеточной функции: $y_i=y(x_i)$. Вычислим первую производную функции $y(x)$ в точках x_i и сопоставим с разностными отношениями (5) – (7):

$$\psi_i^+=L_h^+[y_i]-y'(x_i),0\le i\le n-1; \qquad (8)$$

$$\psi_i^-=L_h^-[y_i]-y'(x_i),1\le i\le n; \qquad (9)$$

$$\psi_i^{(0)}=L_h^{(0)}[y_i]-y'(x_i),1\le i\le n-1. \qquad (10)$$

Эти величины представляют собой погрешности аппроксимации производной с помощью разностных отношений (5) – (7) в точке x_i.

Предположим, что функция $y(x)$ дважды непрерывно дифференцируема на отрезке $[a,b]$ и запишем для нее формулу Тейлора с остаточным членом в форме Лагранжа

$$y_{i+1}=y(x_i+h)=y_i+y'(x_i)h+\frac{1}{2}y''(x_i+\theta_i h)h^2, \qquad (11)$$

где θ_i какое-то неизвестное нам число между нулем и единицей. Подставляя разложение (11) в формулу (8), получим

$$\psi_i^+=\frac{1}{2}y''(x_i+\theta_i h)h. \qquad (12)$$

Аналогичное представление можно получить для величины ψ_i^- (9)

$$\psi_i^-=-\frac{1}{2}y''(x_i-\theta_i h)h. \qquad (13)$$

Формулы (12) и (13) не позволяют вычислить соответствующие погрешности, но дают возможность их оценить. Функция $y''(x)$, по предположению, непрерывна на отрезке $[a,b]$, и, следовательно, ограничена:

$$|y''(x)|\le M_2,\ a\le x\le b. \qquad (14)$$

В результате получаем

$$|\psi_i^+|\le\frac{1}{2}M_2h,\ |\psi_i^-|\le\frac{1}{2}M_2h. \qquad (15)$$

Оценки (15) являются равномерными, поскольку не зависят от индекса i. Таким образом, левое и правое разностное отношение аппроксимируют производную $y'(x)$ с первым порядком точности относительно h.

Для оценки $\psi_i^{(0)}$ (10) предположим, что функция $y(x)$ три раза непрерывно дифференцируема на отрезке $[a,b]$ и продолжим разложение (11) еще на один член

$$\begin{aligned}y_{i+1}&=y_i+y'(x_i)h+\tfrac{1}{2}y''(x_i)h^2+\frac{1}{6}y'''(x_i+\theta_{1,i}h)h^3,\\ y_{i-1}&=y_i-y'(x_i)h+\tfrac{1}{2}y''(x_i)h^2-\frac{1}{6}y'''(x_i-\theta_{2,i}h)h^3.\end{aligned} \qquad (16)$$

Подставляя разложения (16) в формулу (10), будем иметь

$$\psi_i^{(0)}=\frac{1}{6}\{y'''(x_i+\theta_{1,i}h)+y'''(x_i-\theta_{2,i}h)\}h^2. \qquad (17)$$

По предположению функция $y'''(x)$ непрерывна и, следовательно, ограничена на отрезке $[a,b]$:

$$|y'''(x)|\le M_3,\ a\le x\le b. \qquad (18)$$

В результате из равенства (17) получим оценку

$$|\psi_i^{(0)}|\le\frac{1}{3}M_3h^2. \qquad (19)$$

Оценка (19), как и раньше (15), не зависит от индекса i, она является равномерной. Таким образом, центральная разностная производная дает более хороший результат: она аппроксимирует производную $y'(x_i)$ со вторым порядком точности относительно h для функций, трижды непрерывно дифференцируемых на отрезке $[a,b]$.

Задача 1.
Рассмотреть функцию $y=1/(1-x)$ на сетке

$$x_0=-0.1,\ x_1=0,x_2=0.1. \qquad (20)$$

Вычислить в точке $x_1=0$ правую, левую и центральную разностные производные, найти погрешности аппроксимации производной $y'(0)=1$, сравнить их с априорными оценками по формулам (15) и (19).

В данном случае

$$\begin{aligned}L_h^+[y_1]&=\frac{1/0.9-1}{0.1}=1.111111,\\ \psi_1^+&=0.111111;\\ L_h^-[y_1]&=\frac{1-1/1.1}{0.1}=0.909090,\\ \psi_1^-&=-0.090909;\\ L_h^{(0)}[y_1]&=\frac{1/0.9-1/1.1}{0.2}=1.010101,\end{aligned}$$

$$\psi_1^{(0)}=0.010101.$$

Перейдем к априорной оценке погрешности. Вторая и третья производные рассматриваемой функции $y(x)$ имеют вид

$$y''(x)=\frac{2}{(1-x)^3},\ y'''(x)=\frac{6}{(1-x)^4}.$$

Для них на отрезке $[-0.1,0.1]$ справедливы оценки

$$|y''(x)|\le\frac{2}{(0.9)^3}<2.8,\ |y'''(x)|=\frac{6}{(0.9)^4}<9.3.$$

Так что неравенства (15) и (19) запишутся следующим образом

$$|\psi_i^+|\le 0.14,\ |\psi_i^-|\le 0.14,\ |\psi_i^{(0)}|\le 0.031.$$

Они выполняются.

1.3. Разностная аппроксимация второй производной.

Для разностной аппроксимации второй производной составим разностное отношение первых разностных производных

$$L_h[y]=\frac{\frac{y_{i+1}-y_i}{h}-\frac{y_i-y_{i-1}}{h}}{h}=\frac{y_{i-1}-2y_i+y_{i+1}}{h^2} \qquad (21)$$

Чтобы установить связь выражения (21) со второй производной, предположим, что на отрезке $[a,b]$ определена дважды непрерывно дифференцируемая функция $y(x)$, значения которой в точках сетки (1) дают значения сеточной функции y_i. Вычислим ее вторую производную в точках сетки x_i и составим разность

$$\psi_i=L_h[y_i]-y''(x_i),1\le i\le n-1. \qquad (22)$$

Она представляет собой погрешность аппроксимации второй производной с помощью разностного отношения второго порядка (21).

Оценим величину погрешности при предположении, что функция $y(x)$ четыре раза непрерывно дифференцируема на отрезке $[a,b]$. Это предположение позволяет написать разложения Тейлора

$$\begin{aligned}y_{i+1}&=y(x_i+h)=y_i+y'(x_i)h+\frac{1}{2}y''(x_i)h^2+\frac{1}{6}y'''(x_i)h^3+\frac{1}{24}y^{(4)}(x_i+\theta_{1,i}h)h^4,\\ y_{i-1}&=y(x_i-h)=y_i-y'(x_i)h+\frac{1}{2}y''(x_i)h^2-\frac{1}{6}y'''(x_i)h^3+\frac{1}{24}y^{(4)}(x_i-\theta_{2,i}h)h^4.\end{aligned} \qquad (23)$$

Подставляя их в формулы (21), (22), получим

$$\psi_i=\frac{1}{24}\{y^{(4)}(x_i+\theta_{1,i}h)+y^{(4)}(x_i-\theta_{2,i}h)\}h^2. \qquad (24)$$

Мы не можем вычислить погрешность по этой формуле, поскольку значения аргументов у функции $y^{(4)}(x)$ нам неизвестны, но можем ее оценить. Функция $y^{(4)}(x)$ непрерывна и, следовательно, ограничена на отрезке

$$|y^{(4)}(x)|\le M_4,\ a\le x\le b. \qquad (25)$$

В результате из формулы (24) получаем

$$|\psi_i| \le \frac{1}{12} M_4 h^2. \qquad (26)$$

Таким образом, разностное отношении (21) аппроксимирует вторую производную со вторым порядком точности относительно h для функций, имеющих четыре непрерывные производные на отрезке $[a,b]$. Совершенно аналогично можно строить разностные аналоги производных более высокого порядка.

Задача 2.

Для функции $y = 1/(1-x)$ вычислить на сетке (20) вторую разностную производную в точке $x_1 = 0$. Найти погрешность аппроксимации второй производной $y''(0) = 2$ и сравнить результат с априорной оценкой (26).

В данном случае

$$L_h[y_1] = \frac{1/1.1 - 2 + 1/0.9}{0.01} = 2.020202,$$

$$\psi_1 = 0.020202.$$

Четвертая производная рассматриваемой функции $y(x)$ и мажоранта для нее на отрезке $[-0.1, 0.1]$ имеют вид

$$y^{(4)}(x) = \frac{24}{(1-x)^5}, |y^{(4)}(x)| \le \frac{24}{(0.9)^5} < 41.$$

Так что неравенство (26) запишется следующим образом

$$|\psi_1| \le 0.034.$$

Оно выполняется.

При численном интегрировании дифференциальных уравнений производные в них приближенно заменяются соответствующими разностными отношениями. В результате задача сводится к системе разностных уравнений, которые решаются на компьютере. В качестве ответа получается сеточная функция $\{y_i\}$ $(0 \le i \le n)$. После этого встает вопрос, в какой степени и с какой точностью ее можно рассматривать в качестве приближенного решения исходной задачи. Нужно иметь в виду, что прямое сравнение решения дифференциального уравнения $u(x)$ и рассчитанной сеточной функции невозможно: они принадлежат разным пространствам и их, прежде всего, нужно свести в одно пространство. Это можно сделать двояко.

Во-первых, по сеточной функции с помощью методов интерполирования можно построить функцию непрерывного аргумента $y(x)$ и оценить разность $z(x) = y(x) - u(x)$, например, в норме C

$$\|z\|_C = \max_{a \le x \le b} |y(x) - u(x)|.$$

Во-вторых, наоборот, решению дифференциального уравнения можно сопоставить сеточную функцию $\{u_i = u(x_i)\}$ и сравнить между собой две сеточные функции $\{y_i\}$ и $\{u_i\}$, составив их разность $z_i = y_i - u_i$. При этом погрешность приближенного решения задачи будет характеризовать норма разности

$$\|\mathbf{z}\|_C = \max_{0 \le i \le n} |y_i - u_i|.$$

Наиболее последовательным является первый путь, но обычно выбирают более простой - второй.

В следующих параграфах мы рассмотрим численное решение с помощью метода конечных разностей задачи Коши и краевой задачи для линейного дифференциального уравнения второго порядка.

§2. Численное решение задачи Коши.

Рассмотрим задачу Коши для дифференциального уравнения первого порядка

$$u' = f(x,u), \qquad (27)$$

$$u(x_0) = u_0. \qquad (28)$$

Если функция $f(x,u)$ непрерывна и удовлетворяет условию Липшица по аргументу u в некоторой окрестности начальной точки (x_0, u_0), то можно указать такой отрезок $[a,b]$, $a < x_0 < b$, на котором решение задачи (27), (28) $u(x)$ существует и является единственным. В этом параграфе мы обсудим численные методы ее решения.

2.1. Метод Эйлера.

Пусть нам нужно построить решение задачи (27), (28) на отрезке $[x_0, x_0 + l]$ длины l. Возьмем некоторое целое число n, введем шаг $h = l/n$ и образуем на отрезке сетку

$$x_i = x_0 + ih,\ 0 \le i \le n. \qquad (29)$$

Сопоставим задаче (27), (28) на отрезке разностную задачу

$$\frac{y_{i+1} - y_i}{h} = f(x_i, y_i),\ 0 \le i \le n-1; \qquad (30)$$

$$y_0 = u_0. \qquad (31)$$

Здесь мы заменили производную $u'(x)$ в уравнении (27) правой разностной производной и сохранили неизменным начальное условие (28).

Уравнение (30) является разностным уравнением первого порядка, которое принято называть схемой Эйлера. Его можно переписать в виде рекуррентного соотношения

$$y_{i+1} = y_i + hf(x_i, y_i),\ 0 \le i \le n-1. \qquad (32)$$

Это позволяет последовательно рассчитать все значения сеточной функции $\{y_i\}$, решив тем самым задачу (30), (31). Такую разностную схему называют явной.

Перейдем теперь к обсуждению главного вопроса: с какой точностью рассчитанная сеточная функция $\{y_i\}$ дает решение исходной задачи Коши $u(x)$. Для ответа на него рассмотрим решение задачи (27), (28) в точках сетки (29), образуя из функции непрерывного аргумента сеточную функцию $\{u_i = u(x_i)\}$, и сравним ее с

рассчитанной сеточной функцией $\{y_i\}$. Для этого образуем две сеточные функции $\mathbf{z}$, $\boldsymbol{\psi}$:

$$z_i = y_i - u_i,\ 0 \le i \le n; \qquad (33)$$

$$\psi_i = \frac{u_{i+1} - u_i}{h} - f(x_i, u_i),\ 0 \le i \le n-1. \qquad (34)$$

Смысл первой функции (33) очевиден. Она характеризует разницу между рассчитанными числами y_i и решением $u(x)$ задачи (27), (28) в точках сетки x_i. В соответствии с этим сеточную функцию $\mathbf{z}$ называют погрешностью решения.

Функция $\boldsymbol{\psi}$ (34) получается в результате подстановки решения дифференциального уравнения (27) в разностное уравнение (30). Если бы эти уравнения совпадали, то мы получили бы нуль. Но они различаются и нуля мы не получим. Сеточную функцию $\boldsymbol{\psi}$, характеризующую степень близости дифференциального и разностного уравнений, называют погрешностью аппроксимации уравнения на решении.

Установим связь между сеточными функциями $\mathbf{z}$ и $\boldsymbol{\psi}$. С этой целью выразим из формулы (33) y_i:

$$y_i = u_i + z_i \qquad (35)$$

и подставим в разностное уравнение (30). В результате получим

$$\frac{z_{i+1} - z_i}{h} + \frac{u_{i+1} - u_i}{h} = f(x_i, u_i + z_i)$$

или

$$\frac{z_{i+1} - z_i}{h} = \left\{ f(x_i, u_i + z_i) - f(x_i, u_i) \right\} - \left\{ \frac{u_{i+1} - u_i}{h} - f(x_i, u_i) \right\}. \qquad (36)$$

Здесь в обе фигурные скобки мы добавили величину $f(x_i, u_i)$. Добавленные члены входят в соотношение (36) с противоположными знаками и благодаря этому не нарушают равенство. После таких преобразований во вторых фигурных скобках получается величина ψ_i.

В первых фигурных скобках стоит разность значений функции f при одинаковом первом аргументе x_i и разных значениях второго аргумента. Эту разность с помощью формулы Лагранжа можно представить в виде

$$f(x_i, u_i + z_i) - f(x_i, u_i) = \frac{\partial f}{\partial u}(x_i, u_i + \theta_i z_i) z_i$$

и записать формулу (36) в виде рекуррентного соотношения

$$z_{i+1} = \left\{ 1 + h \frac{\partial f}{\partial u}(x_i, u_i + \theta_i z_i) \right\} z_i - \psi_i h,\ 0 \le i \le n-1. \qquad (37)$$

Согласно (28) и (31) его следует дополнить нулевым начальным условием

$$z_0 = 0. \qquad (38)$$

В отличие от формул (30), (31) формулы (37), (38) не могут быть использованы для вычисления величин z_i. В них входят неизвестные величины: ψ_i, u_i, θ_i. Однако из этой системы рекуррентных равенств можно получить рекуррентные неравенства.

Введем для оценки сеточной функции $\boldsymbol{\psi}$ ее норму

$$\|\boldsymbol{\psi}\|_C = \max_{0 \le i \le n-1} |\psi_i|, \text{ при этом } |\psi_i| \le \|\boldsymbol{\psi}\|_C. \qquad (39)$$

Предположим далее, что функция $\frac{\partial f}{\partial u}(x,u)$ в интересующей нас области изменения ее аргументов ограничена

$$\left| \frac{\partial f}{\partial u}(x,u) \right| \le C. \qquad (40)$$

Это позволяет написать оценку

$$\left| 1 + h \frac{\partial f}{\partial u}(x_i, u_i + \theta_i z_i) \right| \le 1 + Ch < e^{Ch} = q,\ q > 1. \qquad (41)$$

С учетом (39) и (41) из формулы (37) следуют рекуррентные неравенства

$$|z_{i+1}| \le q|z_i| + \|\boldsymbol{\psi}\|_C h, \qquad (42)$$

которые порождают цепочку оценок

$$\begin{aligned} &z_0 = 0, \\ &|z_1| \le \|\boldsymbol{\psi}\|_C h, \\ &|z_2| \le (1+q)\|\boldsymbol{\psi}\|_C h, \\ &|z_3| \le (1+q+q^2)\|\boldsymbol{\psi}\|_C h, \\ &\vdots \\ &|z_n| \le (1+q+q^2+\cdots+q^{n-1})\|\boldsymbol{\psi}\|_C h. \end{aligned} \qquad (43)$$

Согласно (41) $q > 1$, так что

$$1 + q + q^2 + \cdots + q^{n-1} < nq^n = ne^{Cnh}.$$

Это позволяет заменить индивидуальные оценки (43) универсальной оценкой

$$|z_i| \le nhe^{Cnh} \|\boldsymbol{\psi}\|_C,\ 0 \le i \le n. \qquad (44)$$

Неравенства (44) справедливы при любом i, в частности, при том, при котором $|z_i|$ достигает своего наибольшего значения и определяет тем самым норму сеточной функции $\|\mathbf{z}\|_C$. В результате оценка погрешности решения принимает вид

$$\|\mathbf{z}\|_C \le le^{Cl} \|\boldsymbol{\psi}\|_C, \qquad (45)$$

где l - длина отрезка, на котором рассматривается решение исходной задачи (27), (28).

Мы получили важный результат: оценку погрешности решения через оценку погрешности аппроксимации уравнения с коэффициентом, который не зависит от шага h. Чем лучше разностное уравнение аппроксимирует дифференциальное, тем меньше погрешность решения.

Чтобы завершить исследование метода Эйлера, оценим норму погрешности аппроксимации уравнения $\|\boldsymbol{\psi}\|_C$. Предположим, что функция $f(x,u)$ имеет в рассматриваемой области изменения аргументов непрерывные и ограниченные первые

частные производные $\frac{\partial f}{\partial x}$ и $\frac{\partial f}{\partial u}$. Это обеспечивает существование у решения задачи (27), (28) непрерывной и ограниченной второй производной

$$u''(x) = \frac{\partial f}{\partial x}(x,u) + \frac{\partial f}{\partial u}(x,u) f(x,u). \qquad (46)$$

Запишем для функции $u(x)$ формулу Тейлора с остаточным членом в форме Лагранжа

$$u_{i+1} = u(x_i + h) = u_i + u'(x_i)h + \frac{1}{2}u''(x_i + \theta_i h)h^2. \qquad (47)$$

Подставляя разложение (47) в формулу (34) для погрешности аппроксимации уравнения, получим

$$\psi_i = \frac{1}{2}u''(x_i + \theta_i h)h. \qquad (48)$$

Согласно формуле (46) функция $u''(x)$ непрерывна и ограничена

$$|u''(x)| \le M_2,\ x_0 \le x \le x_0 + l. \qquad (49)$$

Это позволяет написать оценки

$$\|\boldsymbol{\psi}\|_C \le \frac{M_2}{2}h,\ \|\mathbf{z}\|_C \le \frac{M_2 l}{2}e^{Cl}h. \qquad (50)$$

Неравенства (50) показывают, что при $h \to 0$ погрешность аппроксимации уравнения и связанная с ней неравенством (45) погрешность решения стремятся к нулю со скоростью h. В связи с этим метод Эйлера называют методом первого порядка точности относительно h.

Задача 3.

Рассмотреть задачу Коши

$$u' = \frac{1}{2}u + x, \qquad (51)$$

$$u(0) = 0. \qquad (52)$$

Построить ее численное решение на отрезке $[0,2]$ по схеме Эйлера с шагами $h_1 = 0.25$, $h_2 = 0.05$, $h_3 = 0.01$. Сравнить результаты расчетов между собой и с аналитическим решением задачи

$$u(x) = -2(x+2) + 4e^{\frac{x}{2}}. \qquad (53)$$

Результаты расчетов приведены в таблице 1.

Таблица 1

x_i	$h_1 = 0.25$	$h_2 = 0.05$	$h_3 = 0.01$	$u(x_i)$
0,00	0,000000	0,000000	0,000000	0,000000
0,25	0,000000	0,025633	0,031182	0,032594
0,50	0,062500	0,120338	0,132903	0,136102
0,75	0,195313	0,293193	0,314530	0,319966
1,00	0,407227	0,554466	0,586674	0,594885
1,25	0,708130	0,915776	0,961355	0,972984
1,50	1,109146	1,390270	1,452190	1,468000
1,75	1,622789	1,992821	2,074604	2,095501
2,00	2,263138	2,740255	2,846068	2,873127

Здесь в первом столбце выписаны значения независимой переменной x с шагом $h_1 = 0.25$, в трех следующих столбцах - решения разностной задачи с шагами h_1, h_2, h_3. При этом результаты расчетов с шагами h_2 и h_3 в промежуточных точках x_i, которые не вошли в первый столбец, опущены. В последнем пятом столбце приведены для сравнения значения функции $u(x)$ (53), дающей аналитическое решение задачи. Из таблицы видно, как по мере уменьшения шага повышается точность. В то же время следует отметить, что даже при маленьком шаге $h_3 = 0.01$ метод не может обеспечить решению хорошую точность: ошибка в последней точке $x = 2$ составляет $z = -0.027059$.

Результаты проведенных расчетов представлены также на рис. 1. На нем приведены три кривые, соответствующие численному решению задачи по схеме Эйлера с шагами h_1, h_2, h_3. При выбранном масштабе кривая III практически совпадает с графиком аналитического решения задачи (53) (пунктирная линия). Рисунок наглядно показывает повышение точности приближенного решения по мере уменьшения шага h.

Мы подробно разобрали метод Эйлера, поскольку на примере простой разностной схемы (30) он позволяет поставить и обсудить все основные вопросы численного решения задачи Коши методом конечных разностей. Однако следует отметить, что полученные в этом разделе результаты представляют прежде всего теоретический интерес. Для решения реальных задач разностную схему Эйлера обычно не применяют из-за ее низкой точности: погрешность с уменьшением h убывает как $O(h)$. В следующих разделах мы обсудим пути построения разностных схем более высокого порядка точности.

2.2. Повышение точности разностного метода.

Оценка погрешности решения через погрешность аппроксимации уравнения в методе Эйлера (45) приводит к вполне естественному выводу: чтобы повысить точность метода, нужно улучшить аппроксимацию дифференциального уравнения разностным. Рассмотрим возможные пути реализации этой идеи.

Предположим, что решение дифференциального уравнения $u(x)$ имеет производные достаточно высокого порядка и напишем для него разложение по формуле Тейлора

$$u_{i+1} = u(x_i + h) = u_i + u'(x_i)h + \frac{1}{2}u''(x_i)h^2 + \frac{1}{6}u'''(x_i)h^3 + \cdots. \qquad (54)$$

Если его оборвать на члене порядка h и положить в соответствии с дифференциальным уравнением (27) $u'(x_i) = f(x_i, u_i)$, то мы придем к схеме Эйлера.

Сделаем следующий шаг. Оборвем разложение (54) на члене порядка h^2 и воспользуемся для вычисления производной $u''(x_i)$ формулой (46). В результате получим новое рекуррентное соотношение, более сложное чем (32),

$$y_{i+1} = y_i + f(x_i, y_i)h + \frac{1}{2}\left\{ \frac{\partial f}{\partial x}(x_i, y_i) + \frac{\partial f}{\partial y}(x_i, y_i) f(x_i, y_i) \right\} h^2, \qquad (55)$$

которое можно также записать в виде разностного уравнения

$$\frac{y_{i+1} - y_i}{h} = f(x_i, y_i) + \frac{1}{2}\left\{ \frac{\partial f}{\partial x}(x_i, y_i) + \frac{\partial f}{\partial y}(x_i, y_i) f(x_i, y_i) \right\} h. \qquad (56)$$

Здесь, как и в предыдущем разделе, мы обозначили искомую функцию в разностном уравнении (56) буквой y, а не u, чтобы подчеркнуть, что (27) и (56) – это два разных уравнения.

Уравнение (55), дополненное начальным условием (31), дает явную разностную схему численного решения рассматриваемой задачи Коши. По рекуррентной формуле можно последовательно рассчитать все значения сеточной функции y_i, $0 \le i \le n$ и получить таким образом приближенное решение задачи (27), (28). Исследование показывает, что такая усложненная схема имеет второй порядок точности относительно h как для аппроксимации уравнения, так и для погрешности решения. Существенно то, что основная идея данного подхода допускает дальнейшее развитие. Если оборвать разложение (54) на члене порядка h^3, h^4 и т. д., то получатся разностные схемы третьего, четвертого и более высоких порядков точности.

Однако у данного подхода есть существенный недостаток. При расчетах по схеме Эйлера требуется вычислять только значения функции $f(x_i, y_i)$. В схеме же (55) на каждом шаге приходится вычислять не только функцию f, но и ее первые производные $\frac{\partial f}{\partial x}(x_i, y_i)$ и $\frac{\partial f}{\partial y}(x_i, y_i)$. Если мы, оставив в разложении (54) члены до h^4 включительно, построим схему четвертого порядка точности, то на каждом шаге придется вычислять десять величин: функцию $f(x_i, y_i)$, две ее первых производных, три вторых производных и четыре третьих производных. Это существенно усложнит разработку программы и нарушит важный принцип вычислительной математики – использовать в расчетах только те величины, которые задаются условиями задачи. Формулировка задачи Коши предполагает, что известен алгоритм вычисления функции $f(x,u)$ по значениям ее аргументов. Если этот алгоритм сводится к расчету по простой формуле, то вычисление производных не составляет труда. Однако

возможны и такие варианты представления алгоритма, при которых вычисление производных функции $f(x,u)$ либо очень сложно, либо практически невозможно. Поэтому при разработке разностных схем высокого порядка точности стремятся заменить вычисление производных функции $f(x,u)$ вычислением самой функции в нескольких точках. В следующих разделах мы рассмотрим, как это удается сделать.

2.3. Метод Рунге-Кутта.

Рассмотрим правую часть разностного уравнения (56), содержащую первые производные от функции $f(x,u)$. Главная идея метода Рунге-Кутта состоит в том, чтобы приближенно заменить ее на сумму значений функции f в двух разных точках с точностью до членов порядка h^2. С этой целью положим:

$$\begin{aligned} &f(x_i, y_i) + \frac{1}{2}\left\{ \frac{\partial f}{\partial x}(x_i, y_i) + \frac{\partial f}{\partial y}(x_i, y_i) f(x_i, y_i) \right\} h = \\ &= \beta f(x_i, y_i) + \alpha f(x_i + \gamma h, y_i + \delta h) + O(h^2), \end{aligned} \qquad (57)$$

где α, β, γ, δ - четыре свободных параметра, которые нужно подобрать так, чтобы правая часть равнялась левой с нужной степенью точности.

Разложим функцию $f(x_i + \gamma h, y_i + \delta h)$ по степеням h:

$$f(x_i + \gamma h, y_i + \delta h) = f(x_i, y_i) + \left\{ \gamma \frac{\partial f}{\partial x}(x_i, y_i) + \delta \frac{\partial f}{\partial y}(x_i, y_i) \right\} h + O(h^2), \qquad (58)$$

подставим разложение (58) в формулу (57) и приравняем слева и справа члены, не содержащие h и содержащие h в первой степени. В результате получим для четырех параметров три уравнения

$$\alpha + \beta = 1,\ \alpha\gamma = \frac{1}{2},\ \alpha\delta = \frac{1}{2} f(x_i, y_i). \qquad (59)$$

Они позволяют выразить параметры β, γ, δ через α:

$$\beta = 1 - \alpha,\ \gamma = \frac{1}{2\alpha},\ \delta = \frac{1}{2\alpha} f(x_i, y_i). \qquad (60)$$

Заменяя с помощью (57) левую часть уравнения (56) и отбрасывая члены порядка $O(h^2)$, получим однопараметрическое семейство разностных схем Рунге-Кутта:

$$\frac{y_{i+1} - y_i}{h} = (1-\alpha) f(x_i, y_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, y_i + \frac{h}{2\alpha} f(x_i, y_i) \right). \qquad (61)$$

Уравнение (61), как и (30), можно записать в виде удобного для расчетов рекуррентного соотношения

$$y_{i+1} = y_i + \left[(1-\alpha) f(x_i, y_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, y_i + \frac{h}{2\alpha} f(x_i, y_i) \right) \right] h. \qquad (62)$$

Наиболее удобные разностные схемы этого семейства соответствуют двум значениям параметра α: $\alpha = \frac{1}{2}$ и $\alpha = 1$. При $\alpha = \frac{1}{2}$ рекуррентная формула (62) принимает вид

$$y_{i+1} = y_i + \frac{h}{2}\left\{ f(x_i, y_i) + f(x_i + h, y_i + hf(x_i, y_i)) \right\}. \qquad (63)$$

Она определяет следующую процедуру расчета y_{i+1}. Сначала делается шаг h по схеме Эйлера и вычисляется величина

$$\tilde{y}_{i+1} = y_i + f(x_i, y_i)h. \qquad (64)$$

Затем находится значение функции f в точке $(x_{i+1}, \tilde{y}_{i+1})$, составляется полусумма

$$\frac{f(x_i, y_i) + f(x_{i+1}, \tilde{y}_{i+1})}{2}$$

и проводится окончательный расчет величины

$$y_{i+1} = y_i + \frac{f(x_i, y_i) + f(x_{i+1}, \tilde{y}_{i+1})}{2} h. \qquad (65)$$

Такая схема вычислений называется «предиктор-корректор» или буквально «предсказание-исправление». Вычисление $\tilde{y}_{i+1}$ по схеме Эйлера (64) – это грубое предсказание результата. Вторичный расчет (65), сделанный на основании первого, является уточнением результата, его коррекцией.

При $\alpha = 1$ рекуррентная формула (62) имеет вид

$$y_{i+1} = y_i + f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2} f(x_i, y_i) \right). \qquad (66)$$

Здесь схема расчета заключается в следующем. Сначала делается половинный шаг $\frac{h}{2}$ по схеме Эйлера вычисляется величина

$$y_{i+\frac{1}{2}} = y_i + \frac{h}{2} f(x_i, y_i). \qquad (67)$$

Затем находится значение функции f в точке $\left(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}} \right)$. Оно определяет по формуле (66) очередное значение y_{i+1}.

Следует заметить, что процедура расчета приближенного решения задачи Коши (27), (28) по схеме (61) по сравнению со схемой Эйлера усложняется: теперь на каждом шаге функцию $f(x,u)$ приходится считать не один, а два раза. Однако такое усложнение оказывается оправданным благодаря более высокой точности метода. К исследованию проблемы точности мы теперь и переходим.

Введем, как и в предыдущем разделе, две сеточные функции: погрешность решения $\mathbf{z}$ (33) и погрешность аппроксимации уравнения $\boldsymbol{\psi}$. В рассматриваемом случае она определяется формулой

$$\psi_i = \frac{u_{i+1} - u_i}{h} - \left[(1-\alpha) f(x_i, u_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, u_i + \frac{h}{2\alpha} f(x_i, u_i) \right) \right]. \qquad (68)$$

Выразим y_i по формуле (35) через u_i и z_i и подставим в разностное уравнение (56). В результате получим

$$\frac{z_{i+1}-z_i}{h}+\frac{u_{i+1}-u_i}{h}=(1-\alpha)f\left(x_i,u_i+z_i\right)+\alpha f\left(x_i+\frac{h}{2\alpha},u_i+z_i+\frac{h}{2\alpha}f\left(x_i,u_i+z_i\right)\right). \quad (69)$$

Формулу (69) можно переписать в виде

$$\frac{z_{i+1}-z_i}{h}=\left\{\left[(1-\alpha)f\left(x_i,u_i+z_i\right)+\alpha f\left(x_i+\frac{h}{2\alpha},u_i+z_i+\frac{h}{2\alpha}f\left(x_i,u_i+z_i\right)\right)\right]-\right.$$
$$-\left[(1-\alpha)f\left(x_i,u_i\right)+\alpha f\left(x_i+\frac{h}{2\alpha},u_i+\frac{h}{2\alpha}f\left(x_i,u_i\right)\right)\right]\Bigg\}- \quad (70)$$
$$-\left\{\frac{u_{i+1}-u_i}{h}-\left[(1-\alpha)f\left(x_i,u_i\right)+\alpha f\left(x_i+\frac{h}{2\alpha},u_i+\frac{h}{2\alpha}f\left(x_i,u_i\right)\right)\right]\right\}.$$

Здесь мы перенесли член $(u_{i+1}-u_i)/h$ слева направо и в каждое из двух выражений, собранных в фигурных скобках, добавили одно и то же слагаемое. Поскольку между фигурными скобками стоит знак минус, значение правой части формулы (70) в целом при этом не меняется. Однако благодаря таким преобразованиям мы собрали во вторых фигурных скобках члены, которые дают погрешность аппроксимации дифференциального уравнения $\boldsymbol{\psi}$ (68).

Перейдем к дальнейшему исследованию соотношения (70). Рассмотрим функцию

$$F(v)=(1-\alpha)f\left(x_i,v\right)+\alpha f\left(x_i+\frac{h}{2\alpha},v+\frac{h}{2\alpha}f\left(x_i,v\right)\right). \quad (71)$$

Выражение, стоящее в первых фигурных скобках формулы (70), можно записать как разность значений этой функции при $v=u_i+z_i$ и $v=u_i$ и преобразовать эту разность с помощью формулы конечных приращений Лагранжа

$$F\left(u_i+z_i\right)-F\left(u_i\right)=F'\left(u_i+\theta_i z_i\right)z_i,\ 0<\theta_i<1, \quad (72)$$

где

$$F'(v)=(1-\alpha)\frac{\partial f}{\partial v}\left(x_i,v\right)+\alpha\frac{\partial f}{\partial v}\left(x_i+\frac{h}{2\alpha},v+\frac{h}{2\alpha}f\left(x_i,v\right)\right)\left(1+\frac{h}{2\alpha}\frac{\partial f}{\partial v}\left(x_i,v\right)\right). \quad (73)$$

Подставим полученные выражения для отдельных слагаемых в формулу (70). В результате она примет вид рекуррентной формулы

$$z_{i+1}=\left\{1+hF'\left(u_i+\theta_i z_i\right)\right\}z_i-\psi_i h,\ 0\le i\le n-1, \quad (74)$$

которую нужно дополнить нулевым начальным условием (38). Использовать эту формулу для последовательного вычисления значений сеточной функции $\mathbf{z}$ нельзя: в ее правую часть входят неизвестные аргументы: ψ_i, u_i, θ_i. Однако эту систему рекуррентных равенств можно заменить системой рекуррентных неравенств для последующей оценки z_i.

Предположим, как и при исследовании метода Эйлера, что частная производная $\frac{\partial f}{\partial u}(x,u)$ в интересующей нас области изменения ее аргументов ограничена (40). Тогда с учетом формулы (73) для производной $F'(v)$ получим

$$\left|1+hF'\left(u_i+\theta_i z\right)\right|\le 1+Ch+\frac{1}{2}C^2h^2<e^{Ch}=q,\ q>1. \quad (75)$$

С учетом этого рекуррентные равенства (74) можно заменить рекуррентными неравенствами

$$\left|z_{i+1}\right|\le q\left|z_i\right|+\left\|\boldsymbol{\psi}\right\|h, \quad (76)$$

которые полностью совпадают с неравенствами (42) предыдущего раздела. Мы уже знаем, что из них следует оценка нормы погрешности решения через норму погрешности аппроксимации уравнения

$$\left\|\mathbf{z}\right\|\le le^{Cl}\left\|\boldsymbol{\psi}\right\|. \quad (77)$$

Теперь нужно оценить норму погрешности аппроксимации уравнения (68). Предположим, что функция $f(x,u)$ имеет в интересующей нас области изменения своих аргументов непрерывные вторые производные и, следовательно, решение дифференциального уравнения $u(x)$ трижды непрерывно дифференцируемо. Это позволяет написать следующие разложения Тейлора

$$u_{i+1}=u\left(x_i+h\right)=u_i+u'\left(x_i\right)h+\frac{1}{2}u''\left(x_i\right)h^2+\frac{1}{6}u'''\left(\bar{x}_i\right)h^3, \quad (78)$$

$$f\left(x_i+\frac{h}{2\alpha},u_i+\frac{h}{2\alpha}f\left(x_i,u_i\right)\right)=f\left(x_i,u_i\right)+\frac{h}{2\alpha}\left\{\frac{\partial f}{\partial x}\left(x_i,u_i\right)+\frac{\partial f}{\partial u}\left(x_i,u_i\right)f\left(x_i,u_i\right)\right\}+$$
$$+\frac{h^2}{8\alpha^2}\left\{\frac{\partial^2 f}{\partial\bar{x}^2}\left(\bar{x}_i,\bar{u}_i\right)+2\frac{\partial^2 f}{\partial\bar{x}\partial\bar{u}}\left(\bar{x}_i,\bar{u}_i\right)f\left(\bar{x}_i,\bar{u}_i\right)+\frac{\partial^2 f}{\partial\bar{u}^2}\left(\bar{x}_i,\bar{u}_i\right)f^2\left(\bar{x}_i,\bar{u}_i\right)\right\}, \quad (79)$$

где

$$\bar{x}_i=x_i+\bar{\theta}_i h,\ \bar{x}_i=x_i+\bar{\theta}_i\frac{h}{2\alpha},\ \bar{u}_i=u_i+\bar{\theta}_i\frac{h}{2\alpha}f\left(x_i,u_i\right),$$
$$0<\bar{\theta}_i<1,\ 0<\bar{\theta}_i<1.$$

Здесь последние слагаемые в обоих разложениях представляют собой остаточные члены в форме Лагранжа, которые берутся в неизвестных нам промежуточных точках.

Подставим разложения (78), (79) в формулу (68) для погрешности аппроксимации дифференциального уравнения (27) и примем во внимание соотношения, вытекающие из этого уравнения

$$u'\left(x_i\right)=f\left(x_i,u_i\right),$$
$$u''\left(x_i\right)=\frac{\partial f}{\partial x}\left(x_i,u_i\right)+\frac{\partial f}{\partial u}\left(x_i,u_i\right)f\left(x_i,u_i\right). \quad (80)$$

Благодаря (80) члены нулевого и первого порядков относительно h сокращаются и остаются только члены второго порядка, обязанные своим происхождением остаточным членам в разложениях (78), (79). В результате получается следующее представление для погрешности аппроксимации уравнения

$$\psi_i=h^2\left\{\frac{1}{6}u'''\left(\bar{x}\right)-\frac{1}{8\alpha}\left[\frac{\partial^2 f}{\partial x^2}\left(\bar{x}_i,\bar{u}_i\right)+\right.\right.$$
$$\left.\left.+2\frac{\partial^2 f}{\partial x\partial u}\left(\bar{x}_i,\bar{u}_i\right)f\left(\bar{x}_i,\bar{u}_i\right)+\frac{\partial^2 f}{\partial u^2}\left(\bar{x}_i,\bar{u}_i\right)f^2\left(\bar{x}_i,\bar{u}_i\right)\right]\right\}. \quad (81)$$

Функции, входящие в правую часть этого соотношения, по предположению непрерывны и ограничены в интересующей нас области изменения своих аргументов. Это позволяет заменить равенство (81) неравенством

$$\left|\psi_i\right|\le\left\|\boldsymbol{\psi}\right\|\le Mh^2, \quad (82)$$

где M - константа, мажорирующая выражение в фигурных скобках формулы (81). Подставляя оценку (82) в неравенство (77), получим

$$\left\|\mathbf{z}\right\|\le Mle^{Cl}h^2. \quad (83)$$

Таким образом, при $h\to 0$ погрешность аппроксимации уравнения и, как следствие, погрешность решения стремятся к нулю со скоростью h^2. Это означает, что разностное уравнение (61), полученное по схеме Рунге-Кутта, имеет второй порядок точности относительно h.

Второй порядок точности лучше, чем первый, однако практика показывает, что этой точности также недостаточно. Наиболее часто при проведении реальных расчетов используется схема Рунге-Кутта четвертого порядка точности следующего вида

$$\frac{y_{i+1}-y_i}{h}=\frac{1}{6}\left(k_1+2k_2+2k_3+k_4\right), \quad (84)$$

где

$$k_1=f\left(x_i,y_i\right),\quad k_2=f\left(x_i+\frac{h}{2},y_i+\frac{h}{2}k_1\right),$$
$$k_3=f\left(x_i+\frac{h}{2},y_i+\frac{h}{2}k_2\right),\quad k_4=f\left(x_i+h,y_i+hk_3\right). \quad (85)$$

Если в схеме второго порядка точности на каждом шаге функцию $f(x,y)$ приходилось вычислять два раза, то здесь – четыре раза. Однако это усложнение схемы расчета окупается высокой точностью. На более подробном обсуждении схемы (84), (85) останавливаться не будем и ограничимся конкретным примером.

Задача 4.

Построить решение задачи Коши (51), (52) на отрезке $[0,2]$ *с шагом* $h=0.25$ *по схеме Рунге-Кутта второго порядка типа «предиктор-корректор» (65) и по схеме Рунге-Кутта четвертого порядка (84), (85). Сравнить результаты расчетов между собой и с аналитическим решением задачи (53).*

Результаты расчетов по этой задаче приведены в таблице 2. Здесь в первом столбце даны значения переменной x_i, во втором и третьем столбцах – результаты расчетов по схемам Рунге-Кутта второго и четвертого порядков, в последнем шестом столбце – значения аналитического решения (53) в узлах сетки. В четвертом и пятом

столбцах приведены результаты расчетов по методу Адамса. Они будут обсуждаться в следующем разделе.

Таблица 2.

x_i	Р.К. - II	Р.К. - IV	Ад. - II	Ад. - IV	$u(x_i)$
0,00	0,000000	0,000000	0,000000	0,000000	0,000000
0,25	0,031250	0,032593	0,031250	0,032593	0,032594
0,50	0,133057	0,136099	0,130859	0,136099	0,136102
0,75	0,314791	0,319962	0,309692	0,319962	0,319966
1,00	0,587068	0,594879	0,578331	0,594826	0,594885
1,25	0,961913	0,972975	0,948662	0,972847	0,972984
1,50	1,452948	1,467988	1,434141	1,467772	1,468000
1,75	2,075605	2,095486	2,050001	2,095159	2,095501
2,00	2,847365	2,873107	2,813492	2,872644	2,873127

Сравнение результатов второго столбца таблицы 1, рассчитанных по методу Эйлера с шагом $h=0.25$, с результатами второго и третьего столбца таблицы 2 показывает как уменьшается погрешность при фиксированном шаге h по мере перехода к более точным методам. Так метод Рунге-Кутта четвертого порядка, несмотря на достаточно крупный шаг, дает погрешность решения $\left\|\mathbf{z}\right\|=0.00002$. Это на много лучше, чем при расчете по схеме Эйлера с шагом $h=0.01$ (см. четвертый столбец в таблице 1). В то же время при расчете по схеме Эйлера было сделано двести шагов с однократным вычислением функции $f(x,y)$ на каждом шаге, а при расчете по схеме Рунге-Кутта – восемь шагов с четырехкратным вычислением функции $f(x,y)$ на каждом шаге. Таким образом, более сложный, но и более совершенный метод позволяет при меньшем объеме вычислений получить более точный результат.

В заключение сделаем следующее замечание. Априорные оценки погрешности по схеме Эйлера (50) или Рунге-Кутта (83) представляют теоретический интерес. Они определяют скорость, с которой погрешность стремится к нулю при $h\to 0$. Однако на практике оценки подобного типа неэффективны, поскольку содержат производные искомого решения $u(x)$. Обычно точность численного решения задачи устанавливают с помощью апостериорных оценок, основанных на сравнении результатов расчетов с шагом h и $h/2$. Процедура их вывода и применения была описана в предыдущей главе в связи с задачей численного интегрирования.

2.4. Метод Адамса.

Адамс – английский астроном и математик XIX века, который много занимался небесной механикой. При изучении траекторий планет ему постоянно приходилось численно интегрировать уравнения их движения. Желая минимизировать объем вычислений, Адамс разработал один из наиболее экономичных методов численного решения дифференциальных уравнений, к обсуждению которого мы теперь переходим.

Пусть $u(x)$ - решение дифференциального уравнения (27). Для производной этой функции имеет место равенство

$$u'\left(x\right)=f\left(x,u\left(x\right)\right)=F\left(x\right). \quad (86)$$

Интегрируя его между двумя точками сетки, получим соотношение

$$u_{i+1}=u_i+\int_{x_i}^{x_{i+1}}F\left(x\right)dx. \quad (87)$$

Мы не можем использовать это соотношение непосредственно для перехода в процессе решения задачи от i-ой точки сетки к $(i+1)$-ой, поскольку функция $F(x)$ нам не известна. Чтобы сделать следующий шаг, нужно приближенно заменить эту функцию на такую функцию, которую можно вычислить. Опишем, как эта проблема решается в методе Адамса.

Пусть в процессе численного решения задачи мы довели расчет до точки x_i. В результате проведенных расчетов нам оказались известными величины y_j и $f\left(x_j,y_j\right)$, $0\le j\le i$. Возьмем некоторое фиксированное целое число $m\le i$ и построим интерполяционный многочлен m-ой степени, принимающий в точках x_j, $i-m\le j\le i$ значения $f\left(x_j,u_j\right)$

$$P_m\left(x_j\right)=f\left(x_j,u_j\right),\ i-m\le j\le i. \quad (88)$$

Его можно записать по формуле Лагранжа

$$P_m\left(x\right)=\sum_{j=i-m}^{i}f\left(x_j,y_j\right)Q_{m,j}\left(x\right), \quad (89)$$

где $Q_{m,j}(x)$ специальные многочлены вида

$$Q_{m,j}\left(x\right)=\frac{\left(x-x_{i-m}\right)\cdots\left(x-x_{j-1}\right)\left(x-x_{j+1}\right)\cdots\left(x-x_i\right)}{\left(x_j-x_{i-m}\right)\cdots\left(x_j-x_{j-1}\right)\left(x_j-x_{j+1}\right)\cdots\left(x_j-x_i\right)}, \quad (90)$$

которые мы уже рассматривали в третьей главе.

Главная идея метода Адамса заключается в том, чтобы для расчета y_{i+1} использовать формулу типа (87), приближенно заменяя в ней функцию $F(x)$ на интерполяционный многочлен $P_m(x)$, составленный согласно (89) по результатам предыдущих вычислений. Это приводит к рекуррентной формуле

$$y_{i+1}=y_i+\int_{x_i}^{x_{i+1}}P_m\left(x\right)dx=y_i+\sum_{j=i-m}^{i}a_j f\left(x_j,y_j\right), \quad (91)$$

где

$$a_j=\int_{x_i}^{x_{i+1}}Q_{m,j}\left(x\right)dx. \quad (92)$$

Рассмотрим более подробно данную схему численного решения задачи Коши в простейших случаях $m=0$ и $m=1$, когда технические трудности не закрывают прозрачную идею метода. При $m=0$ для аппроксимации функции $F(x)$ используется полином нулевой степени, т. е. постоянная

$$F\left(x\right)\approx P_0=f\left(x_i,y_i\right).$$

В этом случае формула (91) переходит в рекуррентную формулу метода Эйлера

$$y_{i+1}=y_i+hf\left(x_i,y_i\right),$$

обеспечивающую первый порядок точности. Такой результат сам по себе тривиален. Мы привели его только для того, чтобы показать, что для метода Адамса, как и для метода Рунге-Кутта, исходной точкой является схема Эйлера.

Перейдем к исследованию варианта $m=1$. В этом случае для аппроксимации функции $F(x)$ используется полином первой степени, построенный по значениям функции f в двух точках (x_{i-1},y_{i-1}) и (x_i,y_i):

$$P_1\left(x\right)=f\left(x_i,y_i\right)\frac{x-x_{i-1}}{h}-f\left(x_{i-1},y_{i-1}\right)\frac{x-x_i}{h}.$$

Подставляя его в формулу (91) и проводя интегрирование, получим

$$y_{i+1}=y_i+\left\{\frac{3}{2}f\left(x_i,y_i\right)-\frac{1}{2}f\left(x_{i-1},y_{i-1}\right)\right\}h. \quad (93)$$

Отметим следующую особенность рекуррентной формулы (93). Для расчета очередного значения сеточной функции y_{i+1} нужно знать ее значения в двух предыдущих точках y_i и y_{i-1}. Таким образом, формула (93) начинает работать только со второй точки. Вычислить по ней y_1 нельзя. Это значение решения разностной задачи приходится вычислять каким-нибудь другим методом, например, методом Рунге-Кутта.

Рекуррентную формулу (93) можно записать в виде разностного уравнения

$$\frac{y_{i+1}-y_i}{h}=\frac{3}{2}f\left(x_i,y_i\right)-\frac{1}{2}f\left(x_{i-1},y_{i-1}\right). \quad (94)$$

Подсчитаем для него погрешность аппроксимации дифференциального уравнения

$$\psi_i=\frac{u_{i+1}-u_i}{h}-\left\{\frac{3}{2}f\left(x_i,u_i\right)-\frac{1}{2}f\left(x_{i-1},u_{i-1}\right)\right\}=\frac{u_{i+1}-u_i}{h}-\left\{\frac{3}{2}u'\left(x_i\right)-\frac{1}{2}u'\left(x_{i-1}\right)\right\}. \quad (95)$$

Предположим, что функция $f(x,u)$ имеет в интересующей нас области изменения аргументов непрерывные вторые производные, так что решение задачи $u(x)$ трижды непрерывно дифференцируемо. Запишем разложения Тейлора

$$u_{i+1}=u_i+u'\left(x_i\right)h+\frac{1}{2}u''\left(x_i\right)h^2+\frac{1}{6}u'''\left(x_i+\bar{\theta}_i h\right)h^3,$$
$$u'\left(x_{i-1}\right)=u'\left(x_i\right)-u''\left(x_i\right)h+\frac{1}{2}u'''\left(x_i-\bar{\theta}_i h\right)h^2. \quad (96)$$

Подставляя их в формулу (95), получим

$$\psi_i=\left\{\frac{1}{6}u'''\left(x_i+\bar{\theta}_i h\right)+\frac{1}{4}u'''\left(x_i-\bar{\theta}_i h\right)\right\}h^2. \quad (97)$$

Отсюда можно написать оценку

$$\left|\psi_i\right|\le\left\|\boldsymbol{\psi}\right\|\le\frac{5}{12}M_3h^2, \quad (98)$$

где M_3 - постоянная, мажорирующая третью производную функции $u(x)$:

$$\left|u'''\left(x\right)\right|\le M_3,\ x_0\le x\le x_0+l. \quad (99)$$

Мы видим, что разностное уравнение метода Адамса, соответствующее случаю $m=1$, аппроксимирует дифференциальное уравнение (27) со вторым порядком точности относительно h. Как и в случае метода Рунге-Кутта, это обеспечивает второй порядок точности для погрешности решения $\left\|\mathbf{z}\right\|$, при предположении, что значение y_1, которое рассчитывается нестандартно, вычислено со вторым порядком точности.

Процесс построения более точных схем можно продолжить за счет увеличения m. При $m=2$ получается схема третьего порядка точности, при $m=3$ - четвертого и т.д. Схема четвертого порядка, как и в методе Рунге-Кутта, является наиболее употребительной, поэтому мы коротко остановимся на ее выводе и обсуждении.

Если написать интерполяционный полином третьей степени $P_3(x)$ (89) на сетке из четырех точек x_i, x_{i-1}, x_{i-2}, x_{i-3} и провести интегрирование (92), то рекуррентная формула (91) примет вид:

$$y_{i+1}=y_i+h\left\{\frac{55}{24}f\left(x_i,y_i\right)-\frac{59}{24}f\left(x_{i-1},y_{i-1}\right)+\frac{37}{24}f\left(x_{i-2},y_{i-2}\right)-\frac{9}{24}f\left(x_{i-3},y_{i-3}\right)\right\}. \quad (100)$$

Приведем еще одно форму записи этой формулы через так называемые конечные разности

$$y_{i+1}=y_i+hf_i+\frac{1}{2}h^2\Delta^1 f_i+\frac{5}{12}h^3\Delta^2 f_i+\frac{3}{8}h^4\Delta^3 f_i, \quad (101)$$

где

$$f_i=f\left(x_i,y_i\right),$$
$$\Delta^1 f_i=\frac{1}{h}\left\{f\left(x_i,y_i\right)-f\left(x_{i-1},y_{i-1}\right)\right\},$$
$$\Delta^2 f_i=\frac{1}{h^2}\left\{f\left(x_i,y_i\right)-2f\left(x_{i-1},y_{i-1}\right)+f\left(x_{i-2},y_{i-2}\right)\right\}, \quad (102)$$
$$\Delta^3 f_i=\frac{1}{h^3}\left\{f\left(x_i,y_i\right)-3f\left(x_{i-1},y_{i-1}\right)+3f\left(x_{i-2},y_{i-2}\right)-f\left(x_{i-3},y_{i-3}\right)\right\}.$$

Первая, вторая и третья разности (102) приближенно соответствуют первой, второй и третьей производной функции $F(x)=f\left(x,u(x)\right)$. Эквивалентность формул (100) и (101) легко проверить непосредственно. Формула (101) иногда более удобна для организации вычислительного процесса и контроля точности.

Особенность метода Адамса проявляется в формуле (100) еще сильнее, чем в формуле (93). Здесь для расчета очередного значения y_{i+1} нужно знать значения y в четырех предыдущих точках - y_i, y_{i-1}, y_{i-2}, y_{i-3}. Таким образом, формула (100) начинает работать только с четвертой точки. Вычислить по ней y_1, y_2, y_3 нельзя. Эти значения решения разностной задачи приходится рассчитывать другим методом, например, методом Рунге-Кутта.

Перейдем к обсуждению точности схемы (100). Если функция $f(x,u)$ имеет непрерывные четвертые производные по своим аргументам в интересующей нас области их изменения, так что решение задачи $u(x)$ пять раз непрерывно

дифференцируемо, то разностное уравнение (100) аппроксимирует дифференциальное уравнение (27) с четвертым порядком точности относительно h. Доказательство этого утверждения проводится также, как и для схемы второго порядка (93), только теперь в разложениях типа (96) нужно удерживать больше членов. Четвертый порядок точности при аппроксимации уравнения обеспечивает четвертый порядок точности для погрешности решения $\left\|\mathbf{z}\right\|$, при предположении, что начальные значения для метода Адамса y_1, y_2, y_3 вычислены с такой же точностью. Они рассчитываются независимо и при этом важно, чтобы начальный этап вычислительного процесса не внес такую погрешность, которая исказит все последующие результаты.

Задача 5.

Построить решение задачи Коши (51), (52) на отрезке $[0,2]$ *с шагом* $h=0.25$ *по схеме Адамса второго (93) и четвертого (100) порядка. Сравнить результаты расчетов между собой, с результатами расчетов по схеме Рунге-Кутта и с аналитическим решением задачи.*

Результаты расчетов приведены в четвертом и пятом столбцах таблицы 2. В соответствии с заданием, нужно сравнивать четвертый столбец со вторым и шестым, а пятый – с третьим и шестым. Напомним, что в шестом столбце приведено аналитическое решение (53) рассматриваемой задачи, так что сравнение с ним позволяет судить о точности приближенного решения по схеме Рунге-Кутта и схеме Адамса.

Расчет по схеме Адамса второго порядка точности начинается с y_2, четвертого - с y_4. Значение y_1 в четвертом столбце, y_1, y_2, y_3 в пятом столбце рассчитывались по схеме Рунге-Кутта соответствующего порядка, поэтому в таблице они оказываются одинаковыми с соответствующими данными второго и третьего столбцов. Сравнение результатов проведенных расчетов двумя методами с аналитическим решением задачи показывает, что их точность примерно одинакова.

Сравним схемы четвертого порядка точности в методе Рунге-Кутта (84) и Адамса (100) с точки зрения организации вычислительного процесса. Чтобы сделать один шаг по методу Рунге-Кутта, необходимо вычислить функцию $f(x,y)$ четыре раза (85), а в методе Адамса только один раз. В трех предшествующих точках функция $f(x,y)$ была уже вычислена на предыдущих шагах и вычислять ее снова нет необходимости. В этом заключается главное достоинство метода Адамса, которое особенно высоко ценилось в докомпьютерную эру.

Главный недостаток метода Адамса мы уже отмечали: при его применении первые шаги приходится делать с помощью другого метода, например, с помощью метода Рунге-Кутта и только после этого можно перейти на расчет по схеме Адамса. Таким образом, программа решения задачи Коши по методу Адамса должна включать в себя как элемент программу метода Рунге-Кутта для расчета начальной стадии вычислительного процесса.

С этой особенностью метода Адамса связана еще одна проблема. При численном интегрировании дифференциального уравнения часто приходится менять шаг h. В методе Рунге-Кутта это не составляет труда, поскольку каждый шаг делается

независимо от предыдущего. В методе Адамса ситуация иная. Здесь нужно либо изначально программировать весьма сложные формулы расчета с переменным шагом, либо после каждой смены шага заново проводить расчет первых трех точек по методу Рунге-Кутта. Только после этого можно переходить на стандартный счет по методу Адамса. Эти недостатки приводят к тому, что сегодня при компьютерных расчетах предпочтение часто отдается более удобному методу Рунге-Кутта.

§3. Численное решение краевой задачи для линейного дифференциального уравнения второго порядка.

Рассмотрим следующую задачу для линейного дифференциального уравнения второго порядка:

$$u''-q\left(x\right)u=-f\left(x\right),\quad a<x<b, \quad (103)$$
$$u\left(a\right)=u_1,\quad u\left(b\right)=u_2. \quad (104)$$

Здесь два дополнительных условия заданы в граничных точках отрезка $[a,b]$, поэтому задачу (103), (104) называют краевой.

Пусть функции $f(x)$ и $q(x)$ непрерывны на отрезке $[a,b]$, причем

$$q\left(x\right)\ge q_0>0. \quad (105)$$

При сделанных предположениях, как известно из курса дифференциальных уравнений, решение задачи (103), (104) существует и является единственным.

Перейдем к обсуждению вопросов, связанных с его расчетом с помощью численного метода. Возьмем некоторое целое число n, введем шаг $h=(b-a)/n$ и построим сетку

$$x_i=a+ih,\quad 0\le i\le n. \quad (106)$$

Заменим дифференциальное уравнение (103) его разностным аналогом. В результате получим следующую задачу:

$$\frac{y_{i-1}-2y_i+y_{i+1}}{h^2}-q_i y_i=-f_i,\quad 1\le i\le n-1, \quad (107)$$
$$y_0=u_0,\quad y_n=u_2. \quad (108)$$

Здесь $q_i=q\left(x_i\right)$, $f_i=f\left(x_i\right)$, граничные условия (108) для сеточной функции $\{y_i\}$ взяты такими же, что и в дифференциальной задаче.

Разностные уравнения (107) можно переписать в виде

$$y_{i-1}-\left(2+q_i h^2\right)y_i+y_{i+1}=-f_i h^2,\quad 1\le i\le n-1. \quad (109)$$

Мы получили линейную систему из $(n-1)$-го уравнения с $(n-1)$-им неизвестным y_i, $1\le i\le n-1$. Значения y_0 и y_n неизвестными не являются: они задаются граничными условиями (108).

Между разностными схемами для задачи Коши и для краевой задачи есть существенное различие. В первом случае для определения сеточной функции $\{y_i\}$ мы имели рекуррентные соотношения, которые позволяли последовательно рассчитать все ее значения. Такие разностные схемы называются явными. В краевой задаче (107),

(108) сеточная функция $\{y_i\}$ определяется из решения системы линейных алгебраических уравнений. Такая разностная схема называется неявной.

Из записи разностных уравнений в форме (109) видно, что мы получили систему уравнений с трехдиагональной матрицей с диагональным преобладанием: диагональный элемент $(2+q_ih^2)$ больше суммы двух других элементов той же строки, равной 2. Системы такого типа мы уже встречали в третьей главе в связи с задачей интерполяции кубическим сплайном. Диагональное преобладание гарантирует существование и единственность решения системы, которое может быть построено методом прогонки.

Перейдем к обсуждению основного вопроса: с какой точностью сеточная функция $\{y_i\}$, полученная в результате решения задачи (107), (108), приближает решение краевой задачи (103), (104). Пусть $u(x)$ решение исходной краевой задачи. Обозначим через $u_i = u(x_i)$ его значения в узлах сетки и введем две сеточные функции: погрешность решения и погрешность аппроксимации уравнения

$$z_i = y_i - u_i, \quad 0 \le i \le n, \tag{110}$$

$$\psi_i = \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} - q_iu_i + f_i, \quad 1 \le i \le n-1. \tag{111}$$

Выразим из соотношения (110) y_i через u_i и z_i и подставим в разностное уравнение (107). Оставим члены, содержащие z_i, слева, а остальные члены перенесем направо. В результате получим

$$\frac{z_{i-1} - 2z_i + z_{i+1}}{h^2} - q_iz_i = -\left\{\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} - q_iu_i + f_i\right\} = -\psi_i, \; 1 \le i \le n-1. \tag{112}$$

Граничные условия в дифференциальной и разностной задачах совпадают, так что значения сеточной функции z_i в граничных точках будут нулевыми

$$z_0 = z_n = 0. \tag{113}$$

Мы не можем рассчитать погрешность $\{z_i\}$, решая задачу (112), (113), поскольку в правые части уравнений входят неизвестные величины u_i и ψ_i. Однако задача (112), (113) позволяет оценить погрешность.

Пусть максимальное по модулю число z_j соответствует индексу $i = j$:

$$\|\mathbf{z}\|_C = |z_j| \ge |z_i|, \; 0 \le i \le n. \tag{114}$$

В граничных точках z_i обращается в ноль (113), так что индекс j не равен ни нулю, ни n. Рассмотрим уравнение (112) для этого значения индекса и запишем его в виде:

$$(2+q_jh^2)z_j = z_{j-1} + z_{j+1} + \psi_jh^2. \tag{115}$$

Возьмем модуль от обеих частей равенства и оценим правую часть сверху

$$(2+q_jh^2)|z_j| = (2+q_jh^2)\|\mathbf{z}\|_C \le |z_{j-1}| + |z_{j+1}| + |\psi_j|h^2 \le 2\|\mathbf{z}\|_C + 2\|\boldsymbol{\psi}\|_C h^2$$

или

$$\|\mathbf{z}\|_C \le \frac{1}{q_0}\|\boldsymbol{\psi}\|_C. \tag{116}$$

Здесь мы сократили одинаковые члены слева и справа, разделили обе части неравенства на множитель q_jh^2 и заменили q_j в знаменателе на минимально возможное значение функции $q(x)$ на отрезке $[a,b]$, равное q_0 (105). Таким образом нам удалось оценить погрешность решения $\|\mathbf{z}\|_C$ через погрешность аппроксимации уравнения $\|\boldsymbol{\psi}\|_C$.

Для оценки погрешности аппроксимации уравнения предположим, что функции $f(x)$ и $q(x)$ дважды непрерывно дифференцируемы на отрезке $[a,b]$. В этом случае уравнение (103) допускает двухкратное дифференцирование, что обеспечивает существование у решения краевой задачи (103), (104) четырех непрерывных производных и позволяет написать разложения

$$\begin{aligned} u_{i-1} &= u_i - u'(x_i)h + \frac{1}{2}u''(x_i)h^2 - \frac{1}{6}u'''(x_i)h^3 + \frac{1}{24}u^{(4)}(x_i - \bar{\theta}_ih)h^4, \\ u_{i+1} &= u_i + u'(x_i)h + \frac{1}{2}u''(x_i)h^2 + \frac{1}{6}u'''(x_i)h^3 + \frac{1}{24}u^{(4)}(x_i + \tilde{\theta}_ih)h^4. \end{aligned} \tag{117}$$

Подставляя их в формулу (111), получим следующее выражение для ψ_i:

$$\psi_i = \{u''(x_i) - q_iu_i + f_i\} + \frac{h^2}{24}\left(u^{(4)}(x_i - \bar{\theta}_ih) + u^{(4)}(x_i + \tilde{\theta}_ih)\right). \tag{118}$$

Выражение в первых фигурных скобках равно нулю в силу дифференциального уравнения (103). В результате в правой части формулы (118) остается только вторая группа членов, обязанная своим происхождением остаточным членам в разложениях (117). Оценим ее следующим образом. Функция $u^{(4)}(x)$ непрерывна и, следовательно, ограничена на отрезке $[a,b]$. Пусть

$$|u^{(4)}(x)| \le M_4, \quad a \le x \le b, \tag{119}$$

тогда из формул (116) и (118) получаем

$$\|\boldsymbol{\psi}\|_C \le \frac{M_4}{12}h^2, \; \|\mathbf{z}\|_C \le \frac{M_4}{12q_0}h^2. \tag{120}$$

Мы видим, что разностная схема (107) обеспечивает второй порядок аппроксимации уравнения и, как следствие неравенства (116), второй порядок точности для погрешности решения.

Задача 6.

Рассмотреть на отрезке $[-1,1]$ *краевую задачу*

$$u'' - u = -1, \tag{121}$$

$$u(-1) = u(1) = 0. \tag{122}$$

Выписать и решить соответствующую разностную задачу с шагом $h = 0.5$. *Сравнить решение разностной задачи с аналитическим решением*

$$u(x) = 1 - \frac{ch x}{ch 1}. \tag{123}$$

Система трех уравнений относительно y_1, y_2, y_3 с учетом нулевых граничных условий имеет вид

$$\begin{cases} -2.25y_1 + y_2 = -0.25 \\ y_1 - 2.25y_2 + y_3 = -0.25 \\ y_2 - 2.25y_3 = -0.25 \end{cases} \tag{124}$$

Решение системы (124), как и решение исходной дифференциальной задачи, симметрично относительно средней точки, так что $u_1 = u_3$. С учетом этой особенности система (124) сводится к системе двух уравнений с двумя неизвестными:

$$-2.25y_1 + y_2 = -0.25$$

$$2y_1 - 2.25y_2 = -0.25,$$

решение которой имеет вид

$$y_1 = y_3 = \frac{0.8125}{3.0625} = 0.265306, \; y_2 = \frac{1.0625}{3.0625} = 0.346939.$$

В таблице 3 приведены значения x_i, соответствующие узлам сетки, решение разностной задачи y_i, аналитическое решение (123), вычисленное в узлах сетки $u_i = u(x_i)$, погрешность решения z_i (110) и погрешность аппроксимации уравнения ψ_i (111). Согласно двум последним столбцам

$$\|\mathbf{z}\|_C = 0.005007, \; \|\boldsymbol{\psi}\|_C = 0.015352 \tag{125}$$

Таблица 3

x_i	y_i	$u(x_i)$	z_i	ψ_i
-1,0	0,000000	0,000000	0,000000	
-0,5	0,265306	0,269237	-0,003931	-0,015352
0,0	0,346939	0,351946	-0,005007	-0,013614
0,5	0,265306	0,269237	-0,003931	-0,015352
1,0	0,000000	0,000000	0,000000	

Погрешность аппроксимации уравнения $\boldsymbol{\psi}$ определена только для внутренних точек сетки, поэтому первая и последняя строчки последнего столбца остались незаполненными.

Теперь обратимся к теоретической оценке погрешности решения и погрешности аппроксимации уравнения (120). В данном случае

$$u^{(4)}(x) = -\frac{ch x}{ch 1}, \text{ так что } |u^{(4)}(x)| \le M_4 = 1.$$

В результате оценки (120) с учетом того, что $q = 1$, дают

$$\|\mathbf{z}\|_C \le \frac{0.25}{12} = 0.020833, \; \|\boldsymbol{\psi}\|_C \le \frac{0.25}{12} = 0.020833.$$

Это согласуется с фактическими значениями погрешности (125), подсчитанными непосредственно по известному решению краевой задачи (123).