1. Прямые методы решения СЛАУ. Метод Гаусса.

Рассмотрим СЛАУ $Ax = f$ $(\mathbf{1})$
где A - вещественная квадратная матрица порядка m, f - заданный и x - искомый векторы. Пусть $\det A \neq 0$ => для каждого вектора f *невырожденная* система (1) имеет единственное решение. (1) в развернутом виде:

$$\begin{gathered} a_{11}x_1 + \cdots + a_{1m}x_m = f_1 \\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots \\ a_{m1}x_1 + \cdots + a_{mm}x_m = f_m \end{gathered} \qquad (\mathbf{2})$$

Прямыми называются методы, которые позволяют точное решение невырожденной системы (2) за конечное число операций.
Формулы Крамера представляют компоненты x_j решения системы (2) в виде отношения двух определителей: $x_j = \Delta_j/\Delta$, где $\Delta_j = \det A_j$, $j = 1, \ldots, n$
Матрица A_j получается из матрацы A заменой ее j-го столбца столбцом правых частей системы (2):

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

Чтобы найти решение системы (2), 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. Определитель n-го 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, т.е. всего $Q_n = n \cdot n!$ арифметических операций Оценим это число. При $n \gg 1$ число $n!$ можно подсчитать с помощью асимптотической формулы Стирлинга:

$$n! = \sqrt{2\pi n}(n/e)^n \ \Rightarrow\ Q_n = \sqrt{2\pi}n^{3/2}(n/e)^n$$

При умеренном значении n = 20 эта формула дает астрономическое число: $Q_{20} = 5 \cdot 10^{19}$
Компьютеру, производительность которого составляет m операций/с, для вычисления определителя 20-го порядка понадобится $T_{20} = (5 \cdot 10^{19}/m)$ с. В частности, при $m = 10^{10}$ операций/с получим $T_{20} = 5 \cdot 10^9 = 170$ лет. Даже увеличение производительности компьютера на два, три порядка не спасает положения. Т.о., рассчитывать решение СЛАУ по формулам Крамера с вычислением определителей «в лоб» невозможно, т.е. практическая ценность этих формул невелика.
Метод Гаусса условно делят на два этапа. На 1-м этапе (прямой ход) система (1) приводится к треугольному виду. На 2-м этапе (обратный ход) осуществляется последовательное отыскание неизвестных $x_1, \ldots, x_n$ из этой треугольной системы.
Не ограничивая общности, будем считать, что коэффициент a_{11} , который называют ведущим элементом первого шага, отличен от 0 (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 1-го уравнения на a_{11} и введем в качестве новых коэффициентов c_{1i}, $i = 2, \ldots, n$ и правой части y_1 отношения

$$c_{1i} = \frac{a_{1i}}{a_{11}}, \quad i = 2,3, \ldots, n, \quad y_1 = \frac{f_1}{a_{11}} \qquad (\mathbf{3})$$

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

$$\begin{gathered} x_1 + c_{12}x_2 + \cdots + c_{1j}x_j + \cdots + c_{1n}x_n = y_1 \\ a_{22}^{(1)}x_2 + \cdots + a_{2j}^{(1)}x_j + \cdots + a_{2n}^{(1)}x_n = f_2^{(1)} \\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots \\ a_{n2}^{(1)}x_2 + \cdots + a_{nj}^{(1)}x_j + \cdots + a_{nn}^{(1)}x_n = f_n^{(1)} \end{gathered} \qquad (\mathbf{4})$$

$$\text{где} \quad a_{ij}^{(1)} = a_{ij} - a_{i1}\frac{a_{1j}}{a_{11}}, \quad f_i^{(1)} = f_i - a_{i1}\frac{f_1}{a_{11}}, \quad i,j = 2,3, \ldots, n \quad (\mathbf{5})$$

Естественно выделить из (4) «укороченную» систему, содержащую (n - 1) уравнение:

$$\begin{gathered} a_{22}^{(1)}x_2 + \cdots + a_{2j}^{(1)}x_j + \cdots + a_{2n}^{(1)}x_n = f_2^{(1)} \\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots \\ a_{n2}^{(1)}x_2 + \cdots + a_{nj}^{(1)}x_j + \cdots + a_{nn}^{(1)}x_n = f_n^{(1)} \end{gathered}$$

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

$$\begin{aligned} x_1 + c_{12}x_2 + \cdots + c_{1n}x_n &= y_1 \\ x_2 + \cdots + c_{2n}x_n &= y_2 \\ \ldots\ldots\ldots\ldots & \\ x_{n-1} + c_{n-1,n}x_n &= y_{n-1} \\ x_n &= n \end{aligned} \qquad (\mathbf{6})$$

или в матричной форме $\mathbf{Cx} = \mathbf{y}$

Здесь **C** является верхней треугольной матрицей с единицами на главной диагонали.

$$C = \begin{bmatrix} 1 & c_{12} & \ldots & c_{1,n-1} & c_{1n} \\ 0 & 1 & \ldots & c_{2,n-1} & c_{2n} \\ & & \ldots\ldots\ldots\ldots & & \\ 0 & 0 & \ldots & 1 & c_{n-1,n} \\ 0 & 0 & \ldots & 0 & 1 \end{bmatrix} \quad (\mathbf{7})$$

Построение системы (6) завершает прямой ход метода Гаусса. Обратный ход состоит в последовательном определении неизвестных из системы (6) в обратном порядке:

$$x_n = y_n, \qquad x_{n-1} = y_{n-1} - c_{n-1,n}x_n \quad \text{и т. д.}$$

Общие формулы обратного хода :

$$x_i = y_i - \sum_{j=i+1}^{n} c_{ij}\,x_j, \quad i = n-1, \ldots, 1, \; x_n = y_n \quad (\mathbf{8})$$

Подсчитаем число арифметических операций, которое требуется выполнить при решении СЛАУ по методу Гаусса. Первый шаг прямого хода, согласно формулам (3) и (5), требует n делений и n (n - 1) сложений и умножений Мы учитываем деления отдельно, поскольку для компьютера, как для человека, это более сложная операция. Переходя последовательно от n к (n - 1), потом к (n - 2) подсчитаем общее число арифметических операций на стадии прямого хода. Оно включает:

$Q_1 = n + (n-1) + \cdots + 1 = \dfrac{1}{2}n(n+1)$ делений

$Q_2 = n(n-1) + (n-1)(n-2) \ldots + 2 \cdot 1 = \frac{1}{3}n(n^2-1)$ сложений и умножений

Обратный ход не требует деления, а необходимое число сложений и умножений

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

Сравнивая, видим, что обратный ход существенно проще прямого. Сумма Q_2 и Q_3 дает общее число сложений и умножений, необходимое для решения СЛАУ по методу Гаусса:

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

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

Описанная выше процедура решения системы (2) 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 процессе приведения системы (2) к треугольному виду (6) у матрицы С (7) образовались большие по модулю элементы $|c_{ij}| > 1$ и даже $|c_{ij}| \gg 1$. Тогда при вычислении неизвестных но формулам (8) 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

$$|c_{ij}| \le 1 \quad (\mathbf{9})$$

то роль ошибок округления в процессе вычислений будет нивелироваться.

Опишем, как можно добиться выполнения условия (9). Приступая к первому шагу прямого хода метода Гаусса, рассмотрим элементы a_{1j} первой строки матрицы *A* и найдем среди них элемент, наибольший по модулю. Пусть он имеет номер j_1. Поменяем в системе (2) 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_{11}. Благодаря этому элементы c_{1j} первой строки матрицы *C,* которые рассчитываются по формулам (3), будут удовлетворять неравенству (9)

Процедуру выделения наибольшего по модулю элемента в очередной строке и превращения ею в ведущий элемент нужно затем повторять во время каждого шага прямого хода метода Гаусса. В этом

случае все элементы c_{ij} треугольной матрицы C будут удовлетворять неравенствам (9), обеспечивая устойчивость метода по отношению к ошибкам округления. Такой способ коррекции называется *выбором ведущего элемента по строке.*

Пример и вычисление определителя матрицы А см. учебник стр. 14.

При бесконтрольном применении метода Гаусса для решения больших систем возможностей для потери точности становится еще больше, в то время как выполнение процедуры выбора ведущих элементов по строкам снимает эту проблему.

Определение. *Назовем систему* (2) *системой* с *диагональным преобладанием по строке, если элементы матрицы А удовлетворяют неравенствам*

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \ , \quad 1 \le i \le n \quad (\mathbf{10})$$

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.*

Док-во. Рассмотрим соответствующую однородную систему

$$\sum_{j=1}^{n} a_{ij}x_j = 0 \ , \quad 1 \le i \le n \quad (\mathbf{11})$$

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

$$|\overline{x}_k| > 0, \ \ |\overline{x}_k| \ge |\overline{x}_j| \ , \ \ 1 \le i \le n \quad (\mathbf{12})$$

Запишем k-е уравнение системы (1) в виде

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

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

$$|a_{kk}||\overline{x}_k| \le \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}||\overline{x}_j| \le |\overline{x}_k| \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}| \quad (\mathbf{13})$$

Сокращая неравенство (13) на множитель $|\overline{x}_k|$, который, согласно (12), не равен 0, придем к противоречию с неравенством (10), выражающим диагональное преобладание. Полученное противоречие позволяет последовательно высказать три утверждения:

1. Однородная система (11) с диагональным преобладанием имеет только тривиальное решение
2. Определитель матрицы A с диагональным преобладанием не равен нулю.
3. Неоднородная система (2) с диагональным преобладанием всегда разрешима, и притом единственным образом. ■

2. Трехдиагональные системы линейных алгебраических уравнений. Метод прогонки.

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

$$A_i x_{i-1} + C_i x_i + B_i x_{i+1} = F_i \ , \quad i = 1, 2, \dots, n-1 \quad (\mathbf{1})$$

$$x_0 = q_0 \quad x_n = q_n \qquad (\mathbf{2})$$

где коэффициенты A_i, C_i , B_i и правые части F_i , $i = 1, 2, \dots, n-1$ известны вместе с числами q_0 и q_n Дополнительные соотношения (2) часто называют *краевыми условиями* для системы (1). Во многих случаях они могут иметь более сложный вид Например:

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

где p_0 , q_0 , p_n , q_n — заданные числа. Чтобы не усложнять изложение, ограничимся простейшей формой дополнительных условий (2). Т.к значения x_0 , x_n заданы, перепишем систему (1) в виде

$$\begin{gathered} C_1 x_1 + B_1 x_1 = F_1 - A_1 q_0 \\ A_2 x_1 + C_2 x_2 + B_2 x_3 = F_2 \\ \dots \qquad\qquad (\mathbf{3}) \\ A_{n-1} x_{n-2} + C_{n-1} x_{n-1} = F_{n-1} - B_{n-1} q_n \end{gathered}$$

Матрица этой системы имеет трехдиагональную структуру

$$\begin{bmatrix} C_1 & B_1 & 0 & 0 & \dots & 0 & 0 \\ A_2 & C_2 & B_2 & 0 & \dots & 0 & 0 \\ & & \dots\ \dots\ \dots\ \dots\ \dots\ \dots & & & & \\ 0 & 0 & 0 & 0 & \dots & A_{n-1} & C_{n-1} \end{bmatrix} \quad (\mathbf{4})$$

Это существенно упрощает решение системы (1) благодаря *методу прогонки.* Этот метод основан на предположении, что искомые неизвестные x_i, и x_{i+1} связаны рекуррентным соотношением

$$x_i = \alpha_{i+1}x_{i+1} + \beta_{i+1} \quad 0 \le i \le n-1 \quad (\mathbf{5})$$

Здесь величины α_{i+1} и β_{i+1}, получившие название *прогоночных коэффициентов,* подлежат определению исходя из условий задачи (1), (2). 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 (5) 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 (1). В результате получим

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

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

$$A_i\alpha_i\alpha_{i+1} - C_i\alpha_{i+1} + B_i = 0$$
$$A_i\alpha_i\beta_{i+1} + A_i\beta_i + C_i\beta_{i+1} - F_i = 0$$

Отсюда следуют рекуррентные соотношения для прогоночных коэффициентов.

$$\alpha_{i+1} = \frac{-B_i}{A_i\alpha_i + C_i} \qquad \beta_{i+1} = \frac{F_i - A_i\beta_i}{A_i\alpha_i + C_i} \quad i = 1, 2, \dots, n-1 \quad (\mathbf{6})$$

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

$$\alpha_1 = 0 \quad \beta_1 = q_0 \qquad (\mathbf{7})$$

Для $x_0 = p_0 x_1 + q_0$ и $x_0 = \alpha_1 x_1 + \beta_1$ имеем: $\alpha_1 = p_0 \quad \beta_1 = q_0 \quad (\mathbf{7'})$

Остальные значения коэффициентов прогонки $\alpha_2, \dots, \alpha_n$ и $\beta_2, \dots, \beta_n$ находим из (6).

Далее, согласно правому граничному условию, $x_n = q_n \quad (\mathbf{8})$

Для $x_n = p_n x_{n-1} + q_n$ и $x_{n-1} = \alpha_n x_n + \beta_n$ имеем:

$$x_n = \frac{p_n\beta_n + q_n}{1 - p_n\alpha_n}$$

Отсюда можно найти остальные неизвестные $x_{n-1}, \dots, x_1$ в процессе обратной прогонки с помощью рекуррентной формулы (5).

Число операций, которое требуется для решения системы общего вида $Ax = f$ методом Гаусса, растет при увеличении *n* пропорционально n^3 . Метод прогонки сводится к двум циклам: сначала по формулам (6) рассчитываются прогоночные коэффициенты, затем с их помощью по рекуррентным формулам (5) находится 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. Плюс особая простота его программной реализации на компьютере.

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

$$|C_i| > |A_i| + |B_i| \quad (\mathbf{9})$$

которые выражают свойство диагонального преобладания.

Теорема. *Пусть система уравнений*

$$A_i x_{i-1} + C_i x_i + B_i x_{i+1} = F_i \ , \quad i = 1, 2, \dots, n-1$$
$$x_0 = p_0 x_1 + q_0 \quad x_n = p_n x_{n-1} + q_n$$

такова, что $|C_i| > 0,\ |A_i| > 0, |B_i| > 0,\ |C_i| > |A_i| + |B_i|, i = 1, 2, \dots, n-1,\ |p_0| \le 1,\ |p_n| \le 1.$
Тогда метод прогонки корректен.

Док-во. Т.к. это система с диагональным преобладанием, то согласно теореме билета 1 решение таких систем всегда существует и является единственным.■

Лемма. *Если для системы с трехдиагоналъной матрицей выполняется условие диагонального преобладании* (9), *то прогоночные коэффициенты удовлетворяют неравенствам*

$$|\alpha_i| \le 1$$

Док-во по индукции. Согласно (7), (7') $\alpha_1 = 0$ или $\alpha_1 = p_0$. По условию $|p_0| \le 1$, т.е. при $i = 1$ утверждение леммы верно. Допустим теперь, что оно верно для α_i и рассмотрим α_{i+1} :

$$|\alpha_{i+1}| = \left|\frac{B_i}{C_i + A_i\alpha_i}\right| \le \frac{|B_i|}{|C_i| - |A_i|} \le 1 \quad \blacksquare$$

Неравенство для прогоночных коэффициентов α_i, делает прогонку устойчивой. Действительно, пусть в формуле (5) при $i = i_0 + 1$ вместо x_{i_0+1} вычислена величина $\tilde{x}_{i_0+1} = x_{i_0+1} + \delta_{i_0+1}$. Тогда на следующем шаге вычислений, т. е. при $i = i_0$ вместо $x_{i_0} = \alpha_{i_0+1}x_{i_0+1} + \beta_{i_0+1}$ получим величину $\tilde{x}_{i_0} = \alpha_{i_0+1}\left(x_{i_0+1} + \delta_{i_0+1}\right) + \beta_{i_0+1}$ и погрешность окажется равной

$$\delta_{i_0} = \tilde{x}_{i_0} - x_{i_0} = \alpha_{i_0+1}\delta_{i_0+1}$$

Отсюда получим, что $\left|\delta_{i_0}\right| \le \left|\alpha_{i_0+1}\right|\left|\delta_{i_0+1}\right| \le \left|\delta_{i_0+1}\right|$, т. е. погрешность не возрастает.

3. Обусловленность СЛАУ. Число обусловленности.

Задача поставлена *корректно,* если ее решение существует и единственно и если оно непрерывно зависит от входных данных, т.е. *устойчиво* относительно входных данных.

Рассмотрим вопросы корректности исходной задачи и численных алгоритмов ее решения на примере системы линейных алгебраических уравнений $Ax = f \quad (\mathbf{1})$

с квадратной матрицей *A* порядка *n*. Для каждого *n*-мерного вектора f решение *x* задачи (1) существует тогда и только тогда, когда $\det A \neq 0$. В этом случае можно определить матрицу A^{-1}, обратную матрице *A*, и записать решение в виде $x = A^{-1}f \quad (\mathbf{2})$

Чтобы убедиться в корректности задачи (1), необходимо еще установить непрерывную зависимость решения от входных данных. Входными данными являются правая часть f и элементы a_{ij} матрицы *A*. Соответственно различают *устойчивость по правой части* (когда возмущается только правая часть f, а матрица *A* остается неизменной) и *коэффициентную устойчивость* (когда возмущается только матрица *A,* а правая часть f остается неизменной).

Чтобы можно было говорить о непрерывной зависимости, необходимо ввести на множестве *n*-мерных векторов какую-либо метрику. Будем считать, что решение и правая часть задачи (1) принадлежат линейному пространству *H,* состоящему из *n*-мерных векторов. Введем в *H* норму $\|\cdot\|$; должны выполняться все аксиомы нормы:

$\|x\| > 0, \ \forall 0 \neq x \in H, \qquad \|0\| = 0$

$\|\alpha x\| = |\alpha|\|x\|, \qquad \forall x \in H, \forall \alpha$

$\|x + y\| \le \|x\| + \|y\|, \forall x, y \in H$

Нормой матрицы A, подчиненной данной норме вектора, называется число

$$\|A\| = \sup_{0 \neq x \in H} \frac{\|Ax\|}{\|x\|}$$

Из определений следует, что $\|Ax\| \le \|A\|\|x\| \ \forall x \in H$, $\|A + B\| \le \|A\| + \|B\|, \|AB\| \le \|A\|\|B\|$ для любых матриц *A, B*; $\|E\| = 1$, где *E* - единичная матрица. Наряду с основной системой уравнений (1) рассмотрим «возмущенную систему» $A\tilde{x} = \tilde{f} \quad (\mathbf{4})$

которая отличается от (1) правой частью. Предполагаем, что в матрицу *A* возмущений не вносится. Насколько сильно может измениться решение *x* в результате изменения правой части. Обозначим *абсолютные погрешности* решения и правой части : $\delta x = \tilde{x} - x, \ \delta f = \tilde{f} - f$.

Система (1) *устойчива по правой части,* если при любых $\tilde{f}, \ f$ справедлива оценка

$$\|\delta x\| \le M_1\|\delta f\| \quad (\mathbf{5})$$

где $M_1 > 0$ - постоянная, не зависящая от правых частей $\tilde{f}, \ f$. Оценка (5) выражает факт непрерывной зависимости решения от правой части, т. е. показывает, что $\|\delta x\| \to 0$ при $\|\delta f\| \to 0$. Наличие устойчивости очень важно при численном решении систем уравнений, поскольку почти никогда нельзя задать правую часть f точно, на самом деле вместо вектора f задается какой-то близкий ему вектор $\tilde{f}$. Погрешность $\delta f = \tilde{f} - f$ возникает, например, в результате погрешностей округления.

Если $\det A \neq 0$, то система (1) устойчива по правой части. Действительно, из (1) и (4) следует уравнение для погрешности $A(\delta x) = \delta f \ \Rightarrow \delta x = A^{-1}(\delta f) \ \Rightarrow \ \|\delta x\| \le \|A^{-1}\|\|\delta f\| \quad (\mathbf{6})$

т. е. выполняется неравенство (5) с константой $M_1 = \|A^{-1}\|$. Чем ближе к 0 $\det A$, тем больше постоянная M_1 => тем сильнее погрешность правой части может исказить искомое решение.

При решении системы (1) на ЭВМ с плавающей запятой более естественными характеристиками являются *относительные погрешности*

$$\frac{\|\delta f\|}{\|f\|}, \ \frac{\|\delta x\|}{\|x\|}$$

Получим оценку, выражающую относительную погрешность решения через относительную погрешность правой части. Для этого используем неравенство $\|f\| \le \|A\|\|x\|$ (**7**)
которое следует из (1). Перемножив (6) и (7), придем к требуемой оценке

$$\frac{\|\delta x\|}{\|x\|} \le M_A \frac{\|\delta f\|}{\|f\|} \quad (\mathbf{8}) \quad \text{где } M_A = \|A^{-1}\|\,\|A\| \quad (\mathbf{9})$$

Число M_A называется *числом обусловленности матрицы A* и характеризует степень зависимости относительной погрешности решения от относительной погрешности правой части. Матрицы с большим числом обусловленности M_A называются *плохо обусловленными матрицами.* При численном решении систем с плохо обусловленными матрицами возможно сильное накопление погрешностей.
Задачу 1 см. учебник стр. 25.
Отметим следующие свойства числа обусловленности:
1°. $M_A \ge 1$.
2°. $M_A \ge |\lambda_{max}|/|\lambda_{min}|$, где λ_{max} и λ_{min} - соответственно наибольшее и наименьшее по модулю собственные числа матрицы *A*.(Число обусловленности тем больше, чем больше разброс характеристических чисел матрицы => с увеличением размера матрицы ее обусловленность имеет тенденцию к ухудшению). Соотношение корректно, т.к. в силу невырожденности матрицы $\lambda_{min} \ne 0$
3°. $M_{AB} \le M_A M_B$.
Док-во 2°. Число $\rho(A) = |\lambda_{max}|$ называется *спектральным радиусом матрицы A*. Покажем, что для любой нормы вектора подчиненная ей норма матрицы удовлетворяет неравенству

$$\rho(A) \le \|A\| \quad (\mathbf{10})$$

Рассмотрим собственный вектор *y* матрицы *A*, отвечающий наибольшему по модулю собственному значению. Справедливо равенство $Ax = \lambda_{max}x \quad => \quad \|Ax\| = |\lambda_{max}|$
С другой стороны, $\|Ay\| \le \|A\|\|y\| \quad => \quad |\lambda_{max}| \le \|A\|$ => получаем (10).
Поскольку $\lambda_{min}^{-1}A)$ является максимальным по модулю собственным значением матрицы A^{-1}, для него выполняется неравенство $|\lambda_{min}|^{-1} \le \|A^{-1}\|$
Отсюда и из (10) => свойство 2°. При этом правая часть неравенства 2° не зависит от выбора нормы.
Свойство 1° следует непосредственно из 2°, а свойство 3° - из аналогичного свойства матричных норм, $\|AB\| \le \|A\|\,\|B\|$.
Если матрица самосопряженная : $A = A^*$, то все ее собственные числа вещественны, причем

$$\|A\| = |\lambda_{max}(A)| \quad \text{и} \quad \|A^{-1}\| = |\lambda_{min}(A)|$$

поэтому для таких матриц

$$M_A = |\lambda_{max}(A)|/|\lambda_{min}(A)|$$

4. Одношаговые итерационные методы решения СЛАУ. Достаточные условия сходимости.

Процедура решения СЛАУ $A\mathbf{x} = \mathbf{f}$ (**1**)
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, в случае преобразования 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) $x_k = \{x_1^k, \dots, x_m^k\}$, сходящихся к решению системы (1) в евклидовом пространстве E_n с евклидовой нормой $\|x\|$

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

При записи вектора 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 x_k к решению системы *x* означает, что

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

Необходимым и достаточным условием предельного равенства в конечномерном евклидовом пространстве E_n является покомпонентная сходимость

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

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 $\mathbf{x}_{k+1} = F(\mathbf{x}_k, \mathbf{x}_{k-1}, \dots, \mathbf{x}_{k-m+1})$, то говорят о m-шаговом итерационном алгоритме. В частности, в простейшем случае очередной член последовательности $\mathbf{x}_{k+1}$ может выражаться только через предыдущий: $\mathbf{x}_{k+1} = F(\mathbf{x}_k)$. Такие итерационные алгоритмы называют *одношаговыми.*

Ограничимся линейными одношаговыми алгоритмами, которые обычно записывают в стандартной канонической форме

$$B_{k+1}\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\tau_{k+1}} + A\mathbf{x}_k = \mathbf{f} \quad \det B_{k+1} \neq 0 \quad \tau_{k+1} > 0 \quad (\mathbf{2})$$

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} параметры τ_{k+1} не меняются в процессе итераций, т е не зависят от индекса k, то итерационный процесс называется *стационарным.* Перепишем формулу (2) в виде

$$B_{k+1}\mathbf{x}_{k+1} = \mathbf{F}_{k+1} \quad (\mathbf{3})$$

$$\text{где} \quad \mathbf{F}_{k+1} = (B_{k+1} - \tau_{k+1}A)\mathbf{x}_k + \tau_{k+1}\mathbf{f} \quad (\mathbf{4})$$

Построение очередной итерации сводится к решению системы (3) с правой частью (4), зависящей от предыдущей итерации $\mathbf{x}_k$. Такую задачу приходится решать многократно, поэтому матрицы B_{k+1} следует выбирать достаточно простыми. Наиболее прост в реализации процесс с единичной матрицей: $B_{k+1} = E$. В этом случае формулы (3), (4) дают явное выражение

$$\mathbf{x}_{k+1} = (E - \tau_{k+1}A)\mathbf{x}_k + \tau_{k+1}\mathbf{f} \quad (\mathbf{5})$$

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

Итерационный процесс может быть использован для решения СЛАУ только при условии сходимости. Для исследования его сходимости введем 2 характеристики. Первая из них — **погрешность** решения:

$$\mathbf{z}_k = \mathbf{x}_k - \mathbf{x} \quad (\mathbf{6})$$

Смысл этого вектора ясен. Сходимость итерационного процесса означает, что

$$\lim_{k\to\infty} \mathbf{z}_k = 0 \quad \lim_{k\to\infty} z_i^k = 0 \quad 1 \le i \le n \quad (\mathbf{7})$$

Вторая характеристика - невязка уравнения: $\Psi_k = A\mathbf{x}_k - \mathbf{f}$ $(\mathbf{8})$

Она показывает, насколько хорошо или, наоборот, плохо член итерационной последовательности $\mathbf{x}_k$. удовлетворяет исходной системе. Установим связь между $\mathbf{z}_k$ и Ψ_k :

$$\Psi_k = A\mathbf{x}_k - \mathbf{f} = A(\mathbf{z}_k + \mathbf{x}) - \mathbf{f} = A\mathbf{z}_k \quad (\mathbf{9})$$

Можно также написать обратное соотношение: $\mathbf{z}_k = A^{-1}\Psi_k$ $(\mathbf{10})$

Из формул (9) и (10) вытекают оценки:

$$\|\Psi_k\| \le \|A\|\|\mathbf{z}_k\| \qquad \|\mathbf{z}_k\| \le \|A^{-1}\|\|\Psi_k\| \qquad (\mathbf{11})$$

Они показывают, что погрешность решения $\mathbf{z}_k$ стремится к 0 тогда, и только тогда, когда стремится к нулю невязка Ψ_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_{k+1} , в первую очередь такие, как самосопряженность и знакоопределенность. Напомним, что в вещественном евклидовом пространстве E_n для каждого линейного преобразования существует единственное сопряженное ему линейное преобразование, определяемое тождественным равенством скалярных произведений:

$$(A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A^*\mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in E_n \quad (12)$$

В частности,

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

Преобразование называется *самосопряженным,* если

$$(A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A\mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in E_n \quad (13)$$

Матрицы сопряженных преобразований в ортонормированном базисе связаны простым транспонированием: $a_{ij}^* = a_{ij} \quad \forall i, j = 1, \dots n$

Свойство самосопряженности преобразования равносильно в этом случае выполнению условия совпадения матриц A и A^*: $a_{ij} = a_{ji} = a_{ij}^* \quad \forall i, j = 1, \dots n$

Как известно, любая матрица представима в виде суммы эрмитовой ($A = A^*$) и косоэрмитовой ($A = -A^*$) матриц: $A = \overline{A} + \tilde{A}$ (14)

$$\text{где } \overline{A} = \frac{A + A^*}{2} = \overline{A}^* \qquad \tilde{A} = \frac{A - A^*}{2} = -\tilde{A}^* \qquad (15)$$

Нетрудно видеть, что

$$(A\mathbf{x}, \mathbf{x}) = (A^*\mathbf{x}, \mathbf{x}) = \left(\overline{A}\mathbf{x}, \mathbf{x}\right) \qquad \left(\tilde{A}\mathbf{x}, \mathbf{x}\right) = 0 \qquad (16)$$

В дальнейшем 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.

Матрица линейного преобразования A называется *положительно определенной,* если для любого, отличного от нуля $\mathbf{x} \in E_n$ справедливо неравенство $(A\mathbf{x}, \mathbf{x}) > 0$. В ортонормированном базисе :

$$\sum_{i,j=1}^{n} a_{ij} x_i a_j > 0 \quad \forall \mathbf{x} \in E_n \quad \mathbf{x} \neq 0 \qquad (17)$$

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

Необходимым и достаточным условием положительной определенности самосопряженной матрицы A является критерий Сильвестра (*Квадратичная форма* $\boldsymbol{A}(x, x)$ *положительно определена тогда и только тогда, когда угловые миноры* Δ_k, $k = \overline{1,n}$, *ее матрицы в произвольном базисе положительны*: $\Delta_k > 0$, $k = \overline{1,n}$), из которого, в частности, следует строгая положительность всех диагональных элементов: $a_{ij} > 0 \;\; 1 \leq i \leq n \qquad (18)$

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

Лемма 1. *Для того чтобы симметричная* $(A = A^*)$ *патрица была положительно определенной, необходимо и достаточно, чтобы все ее 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: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$. Поскольку по условию леммы $\lambda_i > 0$, то в ортонормированном базисе из собственных векторов преобразования с матрицей A для $\forall \mathbf{x} \neq 0$ имеем

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

Поэтому очевидно, что $A > 0$

Лемма 2. *Пусть* $A = A^* > 0$ и $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ *упорядоченный набор характеристических чисел этой матрицы, тогда*

$$\lambda_n \|\mathbf{x}\|^2 \leq (A\mathbf{x}, \mathbf{x}) \leq \lambda_1 \|\mathbf{x}\|^2 \qquad (19)$$

Доказательство предлагается провести самостоятельно. Наверное, так ???

$$(A\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n} \lambda_i {\xi_i}^2 \leq \lambda_1 \sum_{i=1}^{n} {\xi_i}^2 = \lambda_1 \|\mathbf{x}\|^2$$

$$(A\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n} \lambda_i {\xi_i}^2 \geq \lambda_n \sum_{i=1}^{n} {\xi_i}^2 = \lambda_n \|\mathbf{x}\|^2$$

Лемма 3. *Если* $A > 0$, *то всегда найдется постоянное число* $\delta > 0$, *такое, что*

$$(A\mathbf{x}, \mathbf{x}) \geq \delta \|\mathbf{x}\|^2 \qquad (20)$$

Док-во. Если $A = A^*$, то достаточно положить $\delta = \lambda_n$. В обшем случае напомним, что согласно (16) $(A\mathbf{x}, \mathbf{x}) = \left(\overline{A}\mathbf{x}, \mathbf{x}\right) > 0$, где $\overline{A} = \overline{A}^*$, поэтому согласно лемме 2

$$(A\mathbf{x}, \mathbf{x}) = \left(\overline{A}\mathbf{x}, \mathbf{x}\right) \geq \overline{\lambda}_n \|\mathbf{x}\|^2$$

где $\overline{\lambda}_n > 0$ — минимальное характеристическое число матрицы $\overline{A}$. Полагая, что $\delta = \overline{\lambda}_n$, приходим к требуемому неравенству (20). ■

Рассмотрим стационарный итерационный процесс (2), 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:*

$$A = A^*,\ A > 0 \qquad (21)$$

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

$$B - \frac{\tau}{2}A > 0 \quad \tau > 0 \qquad (22)$$

Тогда при любом выборе нулевого приближения $\mathbf{x}_0$ *итерационный процесс, который определяется рекуррентной формулой* (2)*, сходится к решению исходной системы* (1)*.*

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

$$(B\mathbf{x}, \mathbf{x}) > \frac{\tau}{2}(A\mathbf{x}, \mathbf{x}) \quad \forall \mathbf{x} \in E_n \quad \mathbf{x} \neq 0 \qquad (23)$$

т.е. оно в частности, предполагает, что матрица *B* является положительно определенной. Кроме того, неравенство (23) определяет интервал, в котором может изменяться параметр τ:

$$0 < \tau < \inf_{\mathbf{x}\neq 0} \frac{2(B\mathbf{x}, \mathbf{x})}{(A\mathbf{x}, \mathbf{x})} \qquad (24)$$

Док-во теоремы. Выразим из соотношения (6) $\mathbf{x}_k:\ \mathbf{x}_k = \mathbf{z}_k + \mathbf{x}$
и подставим в рекуррентную формулу для итерационной последовательности (2):

$$B\frac{\mathbf{z}_{k+1} - \mathbf{z}_k}{\tau} + Az_k = 0 \qquad (\mathbf{25})$$

Отличие итерационной формулы (25) от (2) заключается в том, что она является однородной.
Матрица *B* - положительно определенная. Следовательно, она невырожденная и имеет обратную B^{-1}.
С ее помощью рекуррентное соотношение (25) можно разрешить относительно $\mathbf{z}_{k+1}$:

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \tau B^{-1}Az_k = \mathbf{z}_k - \tau\omega_k \qquad (26)$$

$$\text{где}\ \ \omega_k = B^{-1}Az_k \ \ \text{так что}\ Az_k = B\omega_k \qquad (27)$$

Умножая обе части равенства (26) слева на матрицу *A*. лучаем еще одно рекуррентное соотношение

$$A\mathbf{z}_{k+1} = A\mathbf{z}_k - A\tau\omega_k \qquad (28)$$

Рассмотрим последовательность положительных функционалов:

$$J_k = (A\mathbf{z}_k, \mathbf{z}_k) \qquad (29)$$

Составим аналогичное выражение для J_{k+1} и преобразуем его с помощью рекуррентных (26) и (28):

$$J_{k+1} = (A\mathbf{z}_k - A\tau\omega_k, \mathbf{z}_k - \tau\omega_k) = (A\mathbf{z}_k, \mathbf{z}_k) - \tau(A\omega_k, \mathbf{z}_k) - \tau(A\mathbf{z}_k, \omega_k) + \tau^2(A\omega_k, \omega_k) \qquad (30)$$

Из самосопряженности матрицы *A* и формулы (27) следует, что

$$(A\omega_k, \mathbf{z}_k) = (A\mathbf{z}_k, \omega_k) = (B\omega_k, \omega_k)$$

В результате формула (30) принимает вид

$$J_{k+1} = J_k - 2\tau(B\omega_k, \omega_k) + \tau^2(A\omega_k, \omega_k) = J_k - 2\tau\left(\left(B - \frac{\tau}{2}A\right)\omega_k, \omega_k\right) \qquad (31)$$

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

$$J_k \geq J_{k+1} \geq \cdots \geq 0 \qquad (32)$$

поэтому она сходится Далее, согласно лемме 3,

$$\left(\left(B - \frac{\tau}{2}A\right)\omega_k, \omega_k\right) \geq \delta\|\omega_k\|^2$$

где $\delta > 0$ — строго положительная константа. В результате, согласно (30) и (32), будем иметь

$$J_k - J_{k+1} = \left(\left(B - \frac{\tau}{2}A\right)\omega_k, \omega_k\right) \geq \delta\|\omega_k\|^2$$

Из этого неравенства и сходимости последовательности функционалов J_k следует, что $\|\omega_k\| \to 0$ при $k \to \infty$. В свою очередь, $z_k = A^{-1}B\omega_k$, так что $\|z_k\| \leq \|A^{-1}\|\|B\|\|\omega_k\| \to 0$ ■

5. Метод простой итерации.

Решаем СЛАУ $A\mathbf{x} = \mathbf{f}$ $(\mathbf{1})$
Каноническая форма линейных одношаговых алгоритмов:

$$B_{k+1}\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\tau_{k+1}} + A\mathbf{x}_k = \mathbf{f} \quad \det B_{k+1} \neq 0 \ \ \tau_{k+1} > 0 \quad (2)$$

В качестве матрицы B выбирается единичная матрица: $B = E$, итерационный параметр τ предполагается независящим от номера итерации *k*.

Метод простой итерации — это явный стационарный метод, когда очередная итерация $\mathbf{x}_{k+1}$ вычисляется по рекуррентной формуле

$$\mathbf{x}_{k+1} = (E - \tau A)\mathbf{x}_k + \tau \mathbf{f} \quad (\mathbf{3})$$

Теорема. *Пусть A - самосопряженная положительно определенная матрица*: $A = A^*,\ A > 0$. *И пусть* $\tau \le \frac{2}{\|A\|}$. *Тогда метод простой итерации сходится*.

Док-во. Матрица *A* удовлетворяет условию теоремы Самарского(*Пусть A — самосопряженная положительно определенная матрица:* $A = A^*,\ A > 0$,$B - \frac{\tau}{2}A$ *— положительно определенная матрица;* τ *— положительное число. Тогда при любом выборе нулевого приближения* $\mathbf{x}_0$ *итерационный процесс, который определяется рекуррентной формулой* (2)*, сходится к решению исходной системы* (1)), Требование $B - \frac{\tau}{2}A > 0$ можно переписать в виде

$$(B\mathbf{x}, \mathbf{x}) > \frac{\tau}{2}(A\mathbf{x}, \mathbf{x}) \quad \forall \mathbf{x} \in E_n \quad \mathbf{x} \neq 0$$

т.е. оно в частности, предполагает, что матрица *B* является положительно определенной. Кроме того, неравенство (23) определяет интервал, в котором может изменяться параметр τ:

$$0 < \tau < \inf_{\mathbf{x}\neq 0} \frac{2(B\mathbf{x}, \mathbf{x})}{(A\mathbf{x}, \mathbf{x})}$$

Для метода простой итерации формула, определяющая границу интервала сходимости по итерационному параметру τ, принимает вид

$$\tau_0 = \inf_{\mathbf{x}\neq 0} \frac{2(\mathbf{x}, \mathbf{x})}{(A\mathbf{x}, \mathbf{x})} = \frac{2}{\sup\limits_{\mathbf{x}\neq 0} \frac{(A\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}}$$

Пусть $\mathbf{e}_1, \dots, \mathbf{e}_n$ - ортонормированный базис собственных векторов оператора, соответствующего матрице *A*. В силу положительной определенности все его собственные значения положительны. Будем считать их занумерованными в порядке убывания: $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n > 0$

Разложим вектор $\mathbf{x} \neq 0$ по базису собственных векторов

$$\mathbf{x} = \xi_1 e_1 + \cdots + \xi_n e_n$$

тогда

$$(\mathbf{x}, \mathbf{x}) = {\xi_1}^2 + \cdots + {\xi_n}^2 \qquad (A\mathbf{x}, \mathbf{x}) = \lambda_1 {\xi_1}^2 + \cdots + \lambda_n {\xi_n}^2$$

$$\sup_{\mathbf{x}\neq 0} \frac{(A\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})} = \sup_{\mathbf{x}\neq 0} \frac{\lambda_1 {\xi_1}^2 + \cdots + \lambda_n {\xi_n}^2}{{\xi_1}^2 + \cdots + {\xi_n}^2} = \lambda_1$$

В результате следует, что метод простой итерации сходится при любом τ, принадлежащем интервалу

$$0 < \tau < \tau_0 = \frac{2}{\lambda_1} \quad (4)$$

Если *:* $A = A^* > 0$, то $\|A\| = \lambda_1$ ■

Дальнейшее исследование метода простой итерации построим на конкретном анализе рекуррентной формулы (3) Введем матрицу оператора перехода

$$S = E - \tau A, \quad S = S^* \quad (5)$$

и перепишем формулу (3) в виде

$$\mathbf{x}_{k+1} = S\mathbf{x}_k + \tau \mathbf{f} \qquad (\mathbf{6})$$

При этом погрешность $\mathbf{z}_k = \mathbf{x}_k - \mathbf{x}$ будет удовлетворять аналогичному рекуррентному соотношению, только однородному: $\mathbf{z}_{k+1} = S\mathbf{z}_k$ (7)

Лемма. *Пусть оператор, который порождает матрица A, имеет собственный вектор* $\mathbf{e}_i$ *с собственным значением* λ_i *. Тогда оператор перехода, который порождается матрицей S , также имеет собственный вектор* $\mathbf{e}_i$*, но с собственным значением*

$$\mu_i(\tau) = 1 - \tau\lambda_i \quad (7)$$

Док-во проводится прямой проверкой: $S\mathbf{e}_i = (E - \tau A)\mathbf{e}_i = (1 - \tau\lambda_i) = \mu_i \mathbf{e}_i$ ■

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

$$\|S\| = \max_{1\le i\le n} |\mu_i(\tau)| \quad (8)$$

Теорема. *Для того чтобы метод простой итерации сходился* к *решению системы* (1) *при любом выборе начального приближения, необходимо и достаточно, чтобы все собственные значения оператора перехода S были по модулю меньше единицы:*

$$|\mu_i(\tau)| < 1, \ \ 1 \le i \le n \quad (9)$$

Достаточность. Условие (9) означает, что норма матрицы S, согласно (8), будет меньше единицы: $\|S\| < 1$. В результате получаем:

$$\|\mathbf{z}_{k+1}\| \le \|S\|\|\mathbf{z}_k\| \le \cdots \le \|S\|^{k+1}\|\mathbf{z}_0\| \to 0, \;\; k \to \infty \quad (10)$$

Необходимость. Допустим, что среди собственных значений μ_i нашлось хотя бы одно μ_j, которое не удовлетворяет условию (9), т.е. $|\mu_j| \ge 1$.

Выберем нулевой член итерационной последовательности в виде $\mathbf{x}_0 = \mathbf{x} + \mathbf{e}_j$, где $\mathbf{x}$ - решение системы (1). 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

$$\mathbf{z}_k = S^k\mathbf{e}_j = {\mu_j}^k\mathbf{e}_j \quad \|\mathbf{z}_k\| = |\mu_j|^k \ge 1$$

т.е. $\|\mathbf{z}_k\| \nrightarrow 0$. Необходимость выполнения неравенства (9) для всех собственных значении μ_j для сходимости метода простой итерации доказана. ■

Нужно установить диапазон изменения параметра τ , при котором все собственные значения удовлетворяют неравенству (9). Рассмотрим (см. рис. в учебнике) графики убывающих линейных функций $\mu_i(\tau)$. Все они выходят из одной точки $\tau = 0, \;\; \mu = 1$ и идут вниз из-за отрицательных коэффициентов при τ, причем быстрее всех убывает функция $\mu_1(\tau) = 1 - \tau\lambda_1$, т.к. λ_1 - максимальное собственное значение. Когда она принимает значение -1, условие (9) для нее перестает выполняться $\mu_1(\tau) = 1 - \tau\lambda_1 = -1$ при $\tau = \tau_0 = \frac{2}{\lambda_1}$ Границей интервала сходимости метода простой итерации является значение τ_0 :

$$0 < \tau < \tau_0 = \frac{2}{\lambda_1} \quad (11)$$

Т.о., мы установили, что принадлежность итерационного параметра τ этому интервалу является необходимым и достаточным условием сходимости метода простой итерации

Перейдем к исследованию скорости сходимости метода. Оценка погрешности (10) показывает, что она убывает по закону геометрической прогрессии со знаменателем

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

Рассмотрим (см. рис. в учебнике) графики модулей функций $|\mu_i(\tau)|$. При малых τ все собственные значения $\mu_i(\tau)$ 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 1.

Найдем на отрезке $[\tau_0/2, \tau_0]$ точку, в которой убывающая функция $\mu_n(\tau)$ сравнивается с возрастающей функцией $|\mu_1(\tau)| = -\mu_1(\tau)$. Она определяется уравнением

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

которое дает

$$\tau_* = \frac{2}{\lambda_1 + \lambda_n} < \tau_0 \quad (12)$$

В результате получаем

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

В заключение заметим, что формула (11), определяющая границу интервала сходимости τ_0 , и формула (12) для оптимального значения итерационного параметра τ_* представляют прежде всего теоретический интерес. Обычно при решении СЛАУ наибольшее и наименьшее характеристические шсла матрицы A неизвестны, поэтому подсчитать величины τ_0 и τ_* заранее невозможно. В результате итерационный параметр τ нередко приходится подбирать прямо в процессе вычислений методом проб и ошибок.

6. Метод Зейделя.

Решаем СЛАУ $A\mathbf{x} = \mathbf{f}$ (**1**)

Каноническая форма линейных одношаговых алгоритмов:

$$B_{k+1}\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\tau_{k+1}} + A\mathbf{x}_k = \mathbf{f} \quad \det B_{k+1} \ne 0 \;\; \tau_{k+1} > 0 \quad (2)$$

Рассмотрим произвольную квадратную матрицу и разложим ее на сумму трех матриц:

$$A = D + T_{\text{Н}} + T_{\text{В}}$$

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

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

$T_{\text{Н}}$ — нижняя треугольная матрица: $(T_{\text{Н}})_{ij} = \begin{cases} a_{ij}, & i > j \\ 0, & i \leq j \end{cases}$

$T_{\text{В}}$ — верхняя треугольная матрица: $(T_{\text{Н}})_{ij} = \begin{cases} a_{ij}, & i < j \\ 0, & i \geq j \end{cases}$

В классическом методе Зейделя, записанном в канонической форме, полагают $B = D + T_{\text{Н}}$, $\tau = 1$

В результате формула (2) принимает вид

$$(D + T_{\text{Н}})(\mathbf{x}_{k+1} - \mathbf{x}_k) + A\mathbf{x}_k = \mathbf{f}$$

или

$$(D + T_{\text{Н}})\mathbf{x}_{k+1} + T_{\text{В}}\mathbf{x}_k = \mathbf{f}$$

Перейдем от векторной формы записи рекуррентной формулы к индексной:

$$\sum_{j=1}^{i} a_{ij}x_j^{k+1} + \sum_{j=i+1}^{n} a_{ij}x_j^k = f_i\,, \qquad i = 1, \dots n$$

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

$$x_i^{k+1} = \frac{1}{a_{ii}}\left[f_i - \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}x_j^k\right] \qquad i = 1, \dots n$$

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

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.

Теорема. *Пусть $A = A^*$, $A > 0$. Тогда метод Зейделя сходится.*

Док-во. Метод Зейделя - частный случай метода верхней релаксации при $\omega = 1$:

$$(D + \omega T_{\text{Н}})\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\omega} + A\mathbf{x}_k = \mathbf{f}$$

Достаточное условие сходимости метода верхней релаксации : 0 < ω < 2 (док-во см. билет 7).

Следовательно, метод Зейделя сходится.

Теорема. *Пусть матрица такова, что*

$$\sum_{j \neq i} |a_{ij}| \leq q|a_{ii}| \;\;, \;\; i = 1, \dots, n \;\; |q| < 1$$

Тогда метод Зейделя сходится со скоростью геометрической прогрессии со знаменателем q :

$$\|\mathbf{z}_k\| = \|\mathbf{x}_k - \mathbf{x}\| \leq |q|^{k+1}\|\mathbf{x}_0 - \mathbf{x}\| = |q|^{k+1}\|\mathbf{z}_0\|$$

Т.е. метод Зейделя сходится для любой системы (1), в которой матрица A обладает свойством диагонального преобладания

Скорость сходимости см. на примерах из учебника.

7. Метод верхней релаксации.

Решаем СЛАУ $A\mathbf{x} = \mathbf{f}$ $(\mathbf{1})$

Каноническая форма линейных одношаговых алгоритмов:

$$B_{k+1}\frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\tau_{k+1}} + A\mathbf{x}_k = \mathbf{f} \quad \det B_{k+1} \neq 0 \quad \tau_{k+1} > 0 \quad (2)$$

Рассмотрим произвольную квадратную матрицу и разложим ее на сумму трех матриц:

$$A = D + T_{\text{Н}} + T_{\text{В}}$$

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

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

$T_{\text{Н}}$ — нижняя треугольная матрица: $(T_{\text{Н}})_{ij} = \begin{cases} a_{ij}, & i > j \\ 0, & i \leq j \end{cases}$

$T_{\text{В}}$ — верхняя треугольная матрица: $(T_{\text{Н}})_{ij} = \begin{cases} a_{ij}, & i < j \\ 0, & i \geq j \end{cases}$

Введем параметр ω и запишем рекуррентное соотношение (2) в виде

$$(D + \omega T_{\text{Н}}) \frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\omega} + A\mathbf{x}_k = \mathbf{f} \quad (\mathbf{3})$$

В данном случае $B = D + \omega T_{\text{Н}}$, $\tau = \omega$

Если $\omega < 1$, это метод нижней релаксации. При $\omega = 1$ получаем метод Зейделя. Если $\omega > 1$, это метод верхней релаксации.

Соотношению (3) можно придать вид

$$(\frac{1}{\omega} D + T_{\text{Н}})(\mathbf{x}_{k+1} - \mathbf{x}_k) + A\mathbf{x}_k = \mathbf{f} \quad (\mathbf{4})$$

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 (4) члены, содержащие $\mathbf{x}_{k+1}$ и $\mathbf{x}_k$:

$$\left(\frac{1}{\omega} D + T_{\text{Н}}\right)\mathbf{x}_{k+1} + \left[\left(1 - \frac{1}{\omega}\right) D + T_{\text{В}}\right]\mathbf{x}_k = \mathbf{f}$$

Если перейти от векторной записи к индексной, получим для компонент x_i^{k+1} очередной итерации :

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

Теорема. *Пусть A - самосопряженная положительно определенная матрица* : $A = A^*$, $A > 0$. *Тогда метод верхней релаксации*

$$(D + \omega T_{\text{Н}}) \frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\omega} + A\mathbf{x}_k = \mathbf{f}$$

сходится при условии 0 < ω < 2. *В частности, метод Зейделя* (ω = 1) *сходится.*

Док-во. Самосопряженность матрицы *A* означает, что $T_{\text{Н}}^* = T_{\text{В}}$, $T_{\text{В}}^* = T_{\text{Н}}$ Отсюда следует

$$(T_{\text{Н}}\mathbf{x}, \mathbf{x}) = (T_{\text{Н}}^*\mathbf{x}, \mathbf{x}) = (T_{\text{В}}\mathbf{x}, \mathbf{x}) \quad (\mathbf{5})$$

Составим для рассматриваемого случая матрицу $B - \frac{\tau}{2}A$ Согласно $B = D + \omega T_{\text{Н}}$, $\tau = \omega$

$$B - \frac{\tau}{2}A = (D + \omega T_{\text{Н}}) - \frac{\omega}{2}(D + T_{\text{Н}} + T_{\text{В}}) = \left(1 - \frac{\omega}{2}\right) D + \frac{\omega}{2}(T_{\text{Н}} - T_{\text{В}}) \quad (6)$$

Запишем условие ее положительной определенности.

$$\left((B - \frac{\tau}{2}A)\mathbf{x}, \mathbf{x}\right) = \left(1 - \frac{\omega}{2}\right)(D\mathbf{x}, \mathbf{x}) > 0 \quad (7)$$

Второе слагаемое в выражении (6) не дает вклада в квадратичную форму (7) в силу соотношения (5) Матрица *A* является, по предположению, положительно определенной. Следовательно, все ее диагональные элементы строго положительны $a_{ii} > 0$, $i = 1, .., n$. Это означает положительную определенность матрицы *D*: $(D\mathbf{x}, \mathbf{x})$. В результате знак выражения (7) определяется знаком первого множителя, поэтому достаточное условие для сходимости итерационной последовательности метода верхней релаксации принимает вид: 0 < ω < 2.

8. Интерполирование полиномами. Интерполяционные формулы Лагранжа и Ньютона.

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 $x_0, x_1, \dots, x_n$ этого отрезка, которые будем считать занумерованными в порядке возрастания:

$$a \leq x_0 < x_1 < \cdots < x_n \leq b \quad (1)$$

Требуется по известным значениям

$$y_i = f(x_i), \quad i = 0,1, \dots, n \quad (2)$$

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

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

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

где числовые коэффициенты c_i , $i=0,1,\dots,n$, подлежат определению согласно условиям

$$F(x_j)=f(x_j) \quad j=0,1,\dots,n \quad (4)$$

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

$$\sum_{i=0}^{n} c_i\varphi_i(x_j)=f(x_j) \quad j=0,1,\dots,n \quad (5)$$

Для того чтобы коэффициенты c_i , $i=0,1,\dots,n$ можно было определить, и притом единственным образом необходимо и достаточно, чтобы определитель полученной системных уравнений был отличен от нуля

$$\Delta=\begin{vmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \dots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \dots & \varphi_n(x_1) \\ & & \dots & \\ \varphi_0(x_n) & \varphi_1(x_n) & \dots & \varphi_n(x_n) \end{vmatrix} \neq 0 \quad (6)$$

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

Очевидно, что для однозначной разрешимости задачи интерполирования в классической постановке необходимо и достаточно, чтобы система функции $\varphi_i(x), i=0,1,\dots,n$, была чебышевской.

Необходимым условием принадлежности системы функций к чебышевской является их линейная независимость. Однако это условие не является достаточным (пример см. учебник).

Интерполирование полиномами.

При построении интерполирующей функции $F(x)$ в виде (3) функции $\varphi_i(x)$ выбираются в виде степенных функций:

$$\varphi_0(x)=1, \qquad \varphi_1(x)=x, \qquad \varphi_2(x)=x^2,\dots, \quad \varphi_n(x)=x^n$$

В этом случае интерполирующая функция представляет собой полином степени *n*:

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

с неизвестными коэффициентами c_i , $i=0,1,\dots,n$

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

$$\sum_{i=0}^{n} c_i x_j^i=f(x_j), \qquad j=0,1,\dots,n \ (8)$$

Определителем этой системы является определитель Ван-дер-Монда

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

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

Интерполяционный полином в форме Лагранжа

Представим искомый полином $P_n(x)$ в виде

$$P_n(x)=\sum_{i=0}^{n} f(x_i)Q_{n,i}(x) \quad (9)$$

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

$$Q_{n,i}(x)=\begin{cases}0, & x=x_j \\ 1, & x=x_i\end{cases} \quad \forall j\neq i \quad (10)$$

т.е. каждая из функций $Q_{n,i}(x)$, имеет не менее *n* нулей на $[a,b]$. Такие полиномы легко построить

Поскольку $P_n(x)$ - многочлен степени *n,* коэффициенты $Q_{n,i}(x)$ естественно искать также в виде многочленов степени *n,* а именно в виде

$$Q_{n,i}(x) = \lambda_i(x - x_0)(x - x_1) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n)$$

Из условия $Q_{n,i}(x_i) = 1$ находим

$$\lambda_i^{-1} = (x_i - x_0)(x_i - x_1) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)$$

Таким образом, коэффициенты $Q_{n,i}(x)$ интерполяционного многочлена (9) находятся по формулам

$$Q_{n,i}(x) = \prod_{\substack{j=0 \\ j \neq i}}^{j=n} \frac{(x - x_j)}{(x_i - x_j)} \quad (11)$$

или в развернутом виде:

$$Q_{n,0}(x) = \frac{(x - x_1) \dots (x - x_n)}{(x_0 - x_1) \dots (x_0 - x_n)}$$

$$Q_{n,i}(x) = \frac{(x - x_0) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n)}{(x_i - x_0) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)} \quad (12)$$

$$Q_{n,n}(x) = \frac{(x - x_0) \dots (x - x_{n-1})}{(x_n - x_1) \dots (x_n - x_{n-1})}$$

Иногда удобно записывать

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

Из (9) и (10) очевидно, что построенный полином $P_n(x)$ действительно является интерполяционным полиномом для функции $y = f(x)$ на сетке с узлами $x_0, x_1, \dots, x_n$. Его принято называть *интерполяционным полиномом* в форме Лагранжа

Интерполяционный полином в форме Ньютона

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

Перепишем интерполяционный полином Лагранжа в эквивалентной форме

$$P_n(x) = P_0(x) + \sum_{i=1}^{n} (P_i(x) - P_{i-1}(x)) \quad (13)$$

где $P_i(x)$ - полиномы Лагранжа степени $i \leq n$, соответствующие узлам интерполирования $x_0, x_1, \dots, x_i$. В частности, $P_0(x) = f(x_0)$ - полином 0-ой степени. Полином

$$Q_i(x) = P_i(x) - P_{i-1}(x) \quad (14)$$

имеет степень *i* и по построению обращается в нуль при $x = x_0, x = x_1, \dots, x = x_n$, поэтому его можно представить в виде

$$Q_i(x) = A_i(x - x_0) \dots (x - x_{i-1}) \quad (15)$$

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 (9) и (12) его можно записать:

$$A_i = \sum_{k=0}^{i} \frac{f(x_k)}{\omega_{k,i}} \quad (16)$$

где $\omega_{k,i} = (x_k - x_0) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_n)$ (17)

При этом $A_0 = f(x_0)$ (18)

Формулы (14) и (16) позволяют написать рекуррентное соотношение для полинома $P_n(x)$

$$P_n(x) = P_{n-1}(x) + A_n(x - x_0) \dots (x - x_{n-1})$$

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., получаем окончательную формулу для полинома $P_n(x)$:

$$P_n(x) = A_0 + A_1(x - x_0) + A_2(x - x_0)(x - x_1) + \dots + A_i(x - x_0) \dots (x - x_{i-1}) + \dots + A_n(x - x_0) \dots (x - x_{n-1})$$

Это представление удобно для вычислителя, поскольку увеличение *n* на 1 требует только добавления к старому многочлену одного дополнительного слагаемого. Такое представление интерполяционного полинома называют интерполяционным полиномом в форме Ньютона.

9. Погрешность интерполяционного полинома.

Рассматриваем на отрезке $[a,b]$ функцию $y=f(x)$, для которой известны ее значения в конечном числе точек $x_0, x_1, \dots, x_n$ этого отрезка, которые занумерованы в порядке возрастания:

$$a \le x_0 < x_1 < \cdots < x_n \le b \quad (1)$$

Строим интерполирующую функцию $F(x)$ в виде полинома степени *n*:

$$F(x) = P_n(x) = \sum_{i=0}^{n} c_i x^i \quad (2)$$

с неизвестными коэффициентами c_i , $i = 0,1,\dots,n$. Требуем, чтобы коэффициенты c_i удовлетворяли системе линейных уравнений

$$\sum_{i=0}^{n} c_i x_j^i = f(x_j), \qquad j = 0,1,\dots,n \quad (3)$$

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

$$R_n(x) = f(x) - P_n(x), \quad x \in [a,b] \quad (4)$$

По определению интерполяционного полинома

$$R_n(x_i) = 0, \qquad i = 0,1,\dots,n \quad (5)$$

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

В силу (5) $R_n(x)$ можно представить в виде

$$R_n(x) = \omega_{n+1}(x) r_n(x) \quad (6)$$

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

$$\omega_{n+1}(x) = (x-x_0)\dots(x-x_n)$$

Теорема. *Пусть функция* $y=f(x)$ *имеет*$(n+1)$ *непрерывную производную на* $[a,b]$, и $P_n(x)$ - *интерполяционный многочлен*, $P_n(x_i) = f(x_i), i = 0,1,\dots,n$. *Тогда для погрешности интерполяции справедлива оценка*

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x), \quad \xi \in [a,b] \quad (7)$$

Док-во. 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 $[a,b]$ 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 заведомо имеет $(n+1)$ нуля:

$$g(x_i) = 0, \qquad i = 0,1,\dots,n \quad , \quad g(x) = 0$$

Если $x \in [x_0, x_n]$, то все эти нули также лежат на $[x_0, x_n]$ Если $x < x_0$, то эти нули принадлежат отрезку $[x, x_n]$, а если $x > x_0$, то они находятся на отрезке $[x_0, x]$. Объединяя эти 3 случая, указанные нули функции $g(t)$ принадлежат отрезку $[\alpha,\beta]$, где $\alpha = \min(x_0, x) \ge a$, $\beta = \max(x, x_n) \le b$.

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 стрелке $[\alpha,\beta]$, и, наконец, $g^{(n+1)}(t)$ обращается хотя бы 1 раз в нуль в некоторой точке $t = \xi \in [\alpha,\beta]$, т.е.

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

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

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

и соответственно

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

Формула для $R_n(x)$ не позволяет вычислить погрешность, поскольку точное значение аргумента ξ неизвестно. Однако ее помощью погрешность можно оценить.

$$|R_n(x)| \le \frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)| \quad (8)$$

где

$$M_{n+1} = \max_{x\in[\alpha,\beta]}\left|f^{(n+1)}(x)\right| \le \max_{x\in[a,b]}\left|f^{(n+1)}(x)\right|$$

Обсудим роль полинома $\omega_{n+1}(x)$ в оценке (8) 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 (8) ухудшается из-за быстрого роста функции $|\omega_{n+1}(x)|$. Пример см. учебник.

Из сказанного можно сделать следующий вывод. Если $x \in [x_0, x_n]$, то множитель $|\omega_{n+1}(x)|$ не обесценивает оценку (8). Такой случай называют собственно интерполяцией $f(x)$. Противоположный случай, когда точка x лежит вне отрезка называют экстраполяцией функции $f(x)$. Отмеченная особенность поведения полинома $\omega_{n+1}(x)$ резко ухудшает оценку (86) 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.

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

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

$$\Omega_0 = \left\{x_0^{(0)}\right\}, \qquad \Omega_1 = \left\{x_0^{(1)}, x_1^{(1)}\right\}, \quad \dots, \quad \Omega_n = \left\{x_0^{(n)}, x_1^{(n)}, \dots, x_n^{(n)}\right\}, \dots$$

и отвечающую ей последовательность интерполяционных полиномов $P_n(x)$, построенных для фиксированной 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 зависят 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, притом равномерной на $[a, b]$. Однако построить такие сетки 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.

10. Интерполирование с кратными узлами. Полиномы Эрмита.

Рассматриваем на отрезке $[a, b]$ функцию $y = f(x)$ и конечное число точек $x_0, x_1, \dots, x_m$ этого отрезка, которые занумерованы в порядке возрастания:

$$a \le x_0 < x_1 < \dots < x_m \le b$$

Пусть в узлах интерполяции $x_k \in [a, b], k = 0,1, \dots, m$ заданы значения функции $f(x_k)$ и ее производных $f^{(i)}(x_k),\ i = 1,2, \dots, N_k - 1$ до $(N_k - 1)$ порядка включительно. Числа N_k при этом называют *кратностью узла* k. В каждой точке x_k задано N_k величин:

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

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

$$n = N_0 + \dots + N_m - 1$$

удовлетворяющего требованиям

$$H_n^{(i)}(x_k) = f^{(i)}(x_k), \qquad k = 0,1, \dots, m, \qquad i = 0,1, \dots, N_k - 1 \quad (\mathbf{1})$$

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

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

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

Наше утверждение будет справедливо, если показать, что коэффициенты $a_0, a_1, \dots, a_n$ определяются из условий (1), и притом единственным образом. Условия представляют собой систему линейных алгебраических уравнений относительно этих коэффициентов, причем число уравнений равно числу неизвестных, а именно$N_0 + \cdots + N_m = n + 1$ Рассмотрим соответствующую однородную систему

$$\overline{H}_n^{(i)}(x_k) = 0, \qquad k = 0,1, \dots, m, \qquad i = 0,1, \dots, N_k - 1 \quad (\mathbf{2})$$

Уравнения (2) 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 $N_0 + \cdots + N_m = n + 1$ корней. Поскольку его степень равна *n*, то он должен тождественно равняться нулю. Это означает, что $\overline{a}_0 = 0, \overline{a}_1 = 0, \dots, \overline{a}_n = 0$, те однородная система уравнений (2) имеет только тривиальное решение Отсюда следует, что неоднородная система (1) при любой правой части разрешима, и притом единственным образом.

Исследуем погрешность интерполирования полинома Эрмита

$$R_n(x) = f(x) - H_n(x), \quad x \in [a, b] \quad (3)$$

Представим $R_n(x)$ в виде

$$R_n(x) = \omega_{n+1}(x) r_n(x) \quad (6)$$

где $\omega_{n+1}(x)$ - полином:

$$\omega_{n+1}(x) = (x - x_0)^{N_0} \dots (x - x_m)^{N_m}, \quad N_0 + \cdots + N_m = n + 1$$

и рассмотрим функцию

$$g(t) = f(t) - H_n(t) - \omega_{n+1}(t) r_n(x) \quad (4)$$

заданную на $[a, b]$ и содержащую переменную *x* в качестве параметра.

Узлы x_k являются корнями кратности N_k функции $g(t),\ k = 1, \dots, m$. Кроме того, точка $x \in [a, b]$ является корнем $g(t)$. Таким образом, функция $g(t)$ имеет с учетом кратности $N_0 + N_1 + \cdots + N_m + 1 = n + 2$ корня на отрезке $[a, b]$. По теореме Ролля производная $g'(s)$ имеет по крайней мере 1 нуль между двумя соседними корнями функции $g(t)$. Следовательно, $g'(s)$ имеет не менее $m + 1$ корня на $[a, b]$ в точках, не совпадающих ни с одной из точек $x_0, x_1, \dots, x_m, x$. Кроме того, $g'(t)$ имеет в точке x_k корень кратности $N_k - 1,\ k = 0,1, \dots, m$. Таким образом, $g'(s)$ имеет с учетом кратности не менее

$$(N_0 - 1) + \cdots + (N_m - 1) + (m + 1) = N_0 + N_1 + \cdots + N_m = n + 1$$

корней на $[a, b]$. Аналогично $g''(s)$ имеет не менее *n* корней и т. д. Производная $g^{(n+1)}(s)$ по крайней мере 1 раз обращается в 0 на $[a, b]$, т. е. существует точка $\xi \in [a, b]$, в которой $g^{(n+1)}(\xi) = 0$. Из (4) имеем

$$g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - H_n^{(n+1)}(\xi) - \omega^{(n+1)} r_n(x) = 0$$

Так как $\omega(t)$ - многочлен степени $n + 1$ со старшим коэффициентом 1, имеем $\omega^{(n+1)}(s) = (n + 1)!$ Учитывая, что $(n + 1)$-я производная полинома степени *n* тождественно равна нулю, получаем, что

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

Формула для $R_n(x)$ не позволяет вычислить погрешность, поскольку точное значение аргумента ξ неизвестно. Однако ее помощью погрешность можно оценить.

$$|R_n(x)| \le \frac{M_{n+1}}{(n+1)!} |\omega_{n+1}(x)|$$

где

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

Построение полинома Эрмита в общем случае при произвольном числе узлов и их кратности приводит к довольно громоздким выражениям и редко используется.

11. Интерполирование сплайнами.

Интерполирование многочленом Лагранжа или Ньютона на всем отрезке с использованием большого числа узлов интерполяции часто приводит к плохому приближению, что объясняется сильным накоплением погрешностей в процессе вычислений. Кроме того, из-за расходимости процесса интерполяции увеличение числа узлов не обязано приводить к повышению точности. Для того чтобы избежать больших погрешностей, весь отрезок разбивают на частичные отрезки и на каждом из

частичных отрезков приближенно заменяют функцию $f(x)$ многочленом невысокой степени (так называемая *кусочно-полиномиальная интерполяция*).
Одним из способов интерполирования на всем отрезке является интерполирование с помощью сплайн-функций. Сплайн-функцией или сплайном называют кусочно-полиномиальную функцию, определенную на отрезке $[a,b]$ и имеющую на этом отрезке некоторое число непрерывных производных. Преимуществом сплайнов перед обычной интерполяцией является, во-первых, их сходимость и, во-вторых, устойчивость процесса вычислений.
Рассмотрим частный, но распространенный в вычислительной практике случай, когда сплайн определяется с помощью многочленов 3-й степени (кубический сплайн).
Пусть на $[a,b]$ задана непрерывная функция $f(x)$. Введем сетку $a = x_0 < x_1 < \cdots < x_n \le b$ и обозначим $f_i = f(x_i), i = 0,1,\dots,n.$
Кубическим сплайном, соответствующим данной функции $f(x)$ и данным узлам $\{x_i\}_{i=0}^{n}$, называется функция $S(x)$, удовлетворяющая следующим условиям:
1. На каждом сегменте $[x_{i-1},x_i],\ i = 1,2,..,n$ функция $S(x)$ является многочленом 3-ей степени;
2. Функция $S(x)$, а также ее 1-ая и 2-я производные непрерывны на $[a,b]$;
3. $S(x_i) = f(x_i), i = 0,1,\dots,n.$
4. На концах сегмента $[a,b]$ функция $S''(x)$ удовлетворяет условиям: $S''(a) = S''(b) = 0.$
Замечание. На концах сегмента могут быть заданы и другие условия, например: $S''^{(a)} = A, S''(b) = B.$
Справедлива следующая теорема.
Теорема. *Существует единственный сплайн $S(x)$, удовлетворяющий условиям* 1—4.
Мы проведем конструктивное доказательство этой теоремы
На каждом из отрезков $[x_{i-1},x_i],\ i = 1,2,..,n$, будем искать функцию $S(x) = S_i(x)$ в виде многочлена третьей степени

$$S_i(x) = a_i + b_i(x - x_i) + \frac{c_i}{2}(x - x_i)^2 + \frac{d_i}{6}(x - x_i)^3 \quad (\mathbf{1})$$
$$x_{i-1} \le x \le x_i \quad i = 1,2,..,n$$

где a_i, b_i, c_i, d_i — коэффициенты, подлежащие определению. Поясним смысл введенных коэффициентов. Имеем

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

поэтому $a_i = S_i(x_i), \quad b_i = S_i'(x_i), \quad c_i = S_i''(x_i), \quad d_i = S_i'''(x_i)$
Из условий $S(x_i) = f(x_i), i = 1,\dots,n$ получаем, что $a_i = f(x_i) = f_i,\ i = 1,\dots,n$
Далее, требование непрерывности функции $S(x)$ в узлах приводит к условиям

$$S_i(x_{i-1}) = f_{i-1},\ i = 1,\dots,n$$

Отсюда, учитывая выражения для функций $S_i(x)$ получаем при $i = 1,\dots,n$ уравнения

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

Обозначая $h_i = x_i - x_{i-1}$, перепишем эти уравнения в виде

$$b_i h_i - \frac{c_i}{2}h_i^2 + \frac{d_i}{6}h_i^3 = f_i - f_{i-1}\ , \quad i = 1,2,\dots,n \quad (\mathbf{2})$$

Условия непрерывности первой производной

$$S_i'(x_{i-1}) = S_{i-1}'(x_{i-1}) = b_{i-1}\ , \quad i = 2,\dots,N$$

приводят к уравнениям $c_i h_i + \frac{d_i}{2}h_i^2 = b_i - b_{i-1}\ , \quad i = 2,\dots,N \quad (\mathbf{3})$
Из условия непрерывности второй производной

$$S_i''(x_{i-1}) = S_{i-1}''(x_{i-1}) = c_{i-1}\ , \quad i = 2,\dots,N$$

получаем уравнения

$$d_i h_i = c_i - c_{i-1}\ , \quad i = 2,\dots,N \quad (\mathbf{4})$$

Наконец, дополнительные граничные условия 4 дают еое два уравнения:

$$\begin{cases} S_1''(x_0) = S_1''(a) = c_1 - d_1 h_1 = 0 \\ S_n''(x_n) = S_n''(a) = c_n = 0 \end{cases} \quad (5)$$

Объединяя (2) - (5), получим систему $3N$ уравнений относительно $3N$ неизвестных $b_i, c_i, d_i\ , i = 1,2,\dots,N$. Удобно формально ввести еще одно неизвестное c_0, положив при этом, что оно равно 0, и первое уравнение в (5) переписать в виде $d_1 h_1 = c_1 - c_0$, т.е. в форме, аналогичной (4)
Таким образом, приходим к замкнутой системе уравнений для определения коэффициентов кубического сплайна:

$$d_i h_i = c_i - c_{i-1}\ , \quad i = 1,2,\dots,N \quad c_0 = c_n = 0 \quad (\mathbf{6})$$

$$c_i h_i + \frac{d_i}{2} h_i^2 = b_i - b_{i-1}\ , \quad i = 2,3,\dots,n \quad (\mathbf{7})$$

$$b_i h_i - \frac{c_i}{2} h_i^2 + \frac{d_i}{6} h_i^3 = f_i - f_{i-1}\ , \quad i = 1,2,\dots,n \quad (\mathbf{8})$$

Убедимся в том, что эта система имеет единственное решение. Исключим из (6) - (8) переменные $b_i, d_i\ , i = 1,2,\dots,n$, и получим систему, содержащую только $c_i, i = 1,2,\dots,n$. Для этого рассмотрим два соседних уравнения (8):

$$b_i = \frac{c_i}{2} h_i - \frac{d_i}{6} h_i^2 + \frac{f_i - f_{i-1}}{h_i}$$

$$b_{i-1} = \frac{c_{i-1}}{2} h_{i-1} - \frac{d_{i-1}}{6} h_{i-1}^2 + \frac{f_{i-1} - f_{i-2}}{h_{i-1}}$$

и вычтем второе уравнение из первого. Тогда получим

$$b_i - b_{i-1} = \frac{1}{2}(h_i c_i - h_{i-1} c_{i-1}) - \frac{1}{6}(h_i^2 d_i - h_{i-1}^2 d_{i-1}) + \frac{f_i - f_{i-1}}{h_i} - \frac{f_{i-1} - f_{i-2}}{h_{i-1}}$$

Подставляя найденное выражение для $b_i - b_{i-1}$ в правую часть уравнения (6), получим

$$h_i c_i - h_{i-1} c_{i-1} - \frac{d_{i-1}}{3} h_{i-1}^2 - \frac{2d_i}{3} h_i^2 = 2\left(\frac{f_i - f_{i-1}}{h_i} - \frac{f_{i-1} - f_{i-2}}{h_{i-1}}\right) \quad (\mathbf{9})$$

Далее, из уравнения (6) получаем

$$h_i^2 d_i = h_i(c_i - c_{i-1}), \quad h_{i-1}^2 d_{i-1} = h_{i-1}(c_{i-1} - c_{i-2})$$

и, подставляя эти выражения в (9), приходим к уравнению

$$h_{i-1} c_{i-2} + 2(h_{i-1} + h_i) c_{i-1} + h_i c_i = 6\left(\frac{f_i - f_{i-1}}{h_i} - \frac{f_{i-1} - f_{i-2}}{h_{i-1}}\right)$$

Окончательно для определения коэффициентов c_i получаем систему уравнений

$$h_i c_{i-1} + 2(h_i + h_{i+1}) c_i + h_{i+1} c_{i+1} = 6\left(\frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i}\right) \quad (\mathbf{10})$$

$$i = 1,2,\dots,n-1, \ \ c_0 = c_n = 0$$

В силу диагонального преобладания ($h_i + h_{i+1} < 2(h_i + h_{i+1})$) система (10) имеет единственное решение. Так как матрица системы трехдиагональная, решение легко найти методом прогонки, которая в данном случае устойчива. По найденным коэффициентам c_i коэффициенты b_i, d_i определяются с помощью явных формул

$$d_i = \frac{c_i - c_{i-1}}{h_i}, \quad b_i = \frac{h_i}{2} c_i - \frac{h_i^2}{6} d_i + \frac{f_i - f_{i-1}}{h_i} \quad i = 1,2,\dots,n$$

Таким образом, доказано, что существует единственный кубический сплайн, определяемый условиями 1) - 4)

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. 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 $a \leq= x_0 < x_1 < \cdots < x_n \leq b,\ h_i = x_i - x_{i-1} > 0$

Введем в рассмотрение величину

$$h = \max_{1 \leq i \leq n} h_i$$

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*

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

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

Теорема 2. *Пусть* $f(x)$ *имеет на сегменте* $[a,b]$ 4 *непрерывных производных и дополнительно удовлетворяет условию* $f''(a) = f''(b) = 0$. *Тогда имеют место оценки*:

$$|f(x) - s(x)| < M_4 h^4 \ , \quad \forall x \in [a,b]$$
$$|f'(x) - S'(x)| < M_4 h^3 \ , \quad \forall x \in [a,b]$$
$$|f''(x) - S''(x)| < M_4 h^2 \ , \quad \forall x \in [a,b]$$
$$M_4 = \max_{[a,b]} \left|f^{(4)}(x)\right|$$

12. Квадратурные формулы прямоугольников и трапеций.

Рассматриваем способы приближенного вычисления определенных интегралов

$$I = \int_a^b f(x)dx \quad (\mathbf{1})$$

Универсальные алгоритмы вычисления определенных интегралов дают формулы численного интегрирования - *квадратурные формулы* вида:

$$I = \int_a^b f(x)dx = \sum_{i=0}^{n} c_i f(x_i) + R_n \quad (2)$$

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

$$\lim_{n\to\infty} R_n \to 0, \text{так что} \quad \lim_{n\to\infty} \sum_{i=0}^{n} c_i f(x_i) \to I \quad (3)$$

Суть этого требования заключается в следующем. Если в формуле (2) пренебречь остаточным членом R_n , то получится приближенное равенство

$$I = \int_a^b f(x)dx = \sum_{i=0}^{n} c_i f(x_i) \quad (4)$$

Условие (3), которое называют *сходимостью,* позволяет сделать погрешность в равенстве (4) 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 $x_i \in [a,b]$. Чем выше требование точности, тем больше слагаемых следует удерживать в сумме.

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

$$x_i = a + ih, \ \ i = 0,1,\dots,n \quad (5)$$

Для дальнейшего нам также понадобятся средние точки этих отрезков:

$$\xi_i = a + (i - 1/2)h, \ \ \xi_i \in [x_{i-1}, x_i], \qquad i = 0,1,\dots,n \quad (6)$$

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:

$$P_n = \frac{(b-a)}{n} \sum_{i=0}^{n} f(\xi_i) \quad (7)$$

Т.к. интегральная сумма дает приближенное значение интеграла, можно написать

$$I = P_n + \alpha_n \quad (8)$$

В квадратурное формуле (8) узлами являются точки ξ_i , все весовые множители одинаковы и равны $h = (b-a)/n$, α_n - остаточный член.

Формулу (8) называют **формулой прямоугольников**. Геометрический смысл см. учебник.

При выводе **формулы трапеций** в качестве аппроксимирующей функции $g_n(x)$ берется кусочно-линейная функция. На каждом частичном сегменте $[x_{i-1}, x_i]$ она задается формулой

$$g_n(x) = f(x_{i-1}) + \frac{f(x_i) - f(x_{i-1})}{h}(x - x_{i-1}) \, , \qquad x \in [x_{i-1}, x_i], \qquad i = 1,\dots,n \quad (9)$$

В граничных точках отрезка $x = x_{i-1}$ и $x = x_i$ функция $g_n(x)$ принимает те же значения, что и функция $f(x)$: $g_n(x_{i-1}) = f(x_{i-1}), \ g_n(x_i) = f(x_i)$ (10)

т.е. она осуществляет кусочно-линейную интерполяцию функции $f(x)$ на отрезке $[a,b]$. Вычислим интеграл:

$$\int_{x_{i-1}}^{x_i} g_n(x)dx = \int_{x_{i-1}}^{x_i} \left\{ f(x_{i-1}) + \frac{f(x_i) - f(x_{i-1})}{h}(x - x_{i-1}) \right\} dx = \frac{h}{2}\left(f(x_{i-1}) + f(x_i)\right) \quad (11)$$

Геометрический смысл см. учебник.

Интеграл от функции $g_n(x)$ по всему отрезку $[a,b]$ является суммой интегралов (11):

$$T_n = \int_a^b g_n(x)dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} g_n(x)dx = \frac{(b-a)}{n}\left\{ \frac{1}{2}f(a) + f(x_1) + f(x_2) + \cdots + f(x_{n-1}) + \frac{1}{2}f(b) \right\} \quad (12)$$

Он дает приближенное значение интеграла:

$$I = \int_a^b f(x)dx = T_n + \beta_n \quad (13)$$

В квадратурной формуле (13) узлами являются точки x_i. Все весовые коэффициенты, кроме двух, одинаковы и равны $h = (b-a)/n$,а весовые коэффициенты при $i = 0$ и $i = n$ имеют вдвое меньшие значения. Для остаточного члена введено специальное обозначение β_n Формулу (13) называют *квадратурной формулой трапеций.*

Анализ остаточных членов. Пусть функция $f(x)$ дважды непрерывно дифференцируема на отрезке $[a,b]$. В курсе математического анализа при этом предположении устанавливаются формулы (вывод см. прошлогодние билеты стр. 20-22)

$$\int_{x_{i-1}}^{x_i} f(x)dx = f(\xi_i)h + \frac{h^3}{24}f''(\eta_i^*) \quad (14)$$

$$\int_{x_{i-1}}^{x_i} f(x)dx = \frac{f(x_{i-1}) + f(x_i)}{2}h - \frac{h^3}{12}f''(\eta_i^{**}) \quad (15)$$

где η_i^* и $\eta_i^{**} \in [x_{i-1}, x_i]$. Существование таких точек гарантировано, но их точное положение неизвестно. Суммируя равенства (14) и (15) по i, получаем формулы (8) и (13) со следующими выражениями для остаточных членов

$$\alpha_n = \frac{h^3}{24}\sum_{i=1}^{n} f''(\eta_i^*) \quad (16)$$

$$\beta_n = -\frac{h^3}{12}\sum_{i=1}^{n} f''(\eta_i^{**}) \quad (17)$$

Лемма. Пусть функция $f(x)$ непрерывна на отрезке $[a,b]$ и пусть $x_1, \ldots, x_n$ - некоторые точки этого отрезка. Тогда на отрезке x_i найдется такая точка η, что

$$\frac{1}{n}\sum_{i=0}^{n} f(x_i) = f(\eta)$$

Применяя лемму к суммам (16) и (17), получаем:

$$\alpha_n = \frac{(b-a)^3}{24n^2}f''(\eta^*) \quad (18)$$

$$\beta_n = -\frac{(b-a)^3}{12n^2}f''(\eta^{**}) \quad (19)$$

Формулы (18) и (19) не позволяют вычислить остаточные члены: существование точек η^* и η^{**} на $[a,b]$ гарантировано, но их положение на отрезке неизвестно. Но можно оценить остаточные члены. Если обозначить

$$M_2 = \max_{x\in[a,b]} |f''(x)|$$

тогда равенства (18) и (19) можно заменить:

$$|\alpha_n| \le \frac{(b-a)^3}{24n^2}M_2 \quad (20)$$

$$|\beta_n| = -\frac{(b-a)^3}{12n^2} M_2 \quad (21)$$

При заданной точности ε они позволяют определить число узлов n, которое нужно использовать при вычислении интеграла по рассматриваемым квадратурным формулам.

Если вторая производная функции $f(x)$ является знакоопределенной на отрезке $[a, b]$, формулы (18) и (19) позволяют определить знаки остаточных членов, при этом они оказываются противоположными При $f''(x) \geq 0 \ : \alpha_n \geq 0, \beta_n \leq 0$. Таким образом для интеграла получается оценка: $P_n \leq I \leq T_n$. И наоборот. Такие оценки очень удобны, поскольку позволяют легко контролировать точность вычислений.

Оценки (20), (214) показывают, что в случае дважды непрерывно дифференцируемой подынтегральной функции остаточные члены 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} .

Методы прямоугольников и трапеций называют методами второго порядка точности.

13. Квадратурные формулы Симпсона.

Рассматриваем способы приближенного вычисления определенных интегралов

$$I = \int_a^b f(x)dx \quad (\mathbf{1})$$

Универсальные алгоритмы вычисления определенных интегралов дают формулы численного интегрирования - *квадратурные формулы* вида:

$$I = \int_a^b f(x)dx = \sum_{i=0}^{n} c_i f(x_i) + R_n \quad (2)$$

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

$$\lim_{n\to\infty} R_n \to 0, \text{так что} \quad \lim_{n\to\infty} \sum_{i=0}^{n} c_i f(x_i) \to I \quad (3)$$

Если в формуле (2) пренебречь остаточным членом R_n , то получится приближенное равенство

$$I = \int_a^b f(x)dx = \sum_{i=0}^{n} c_i f(x_i) \quad (4)$$

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

$$x_i = a + ih, \ \ i = 0,1, \dots, n \quad (5)$$

Для аппроксимации функции $f(x)$ используем кусочно-квадратичное интерполирование.

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

$$g_n(x) = f(x_{2j-2}) \frac{(x - x_{2j-1})(x - x_{2j})}{2h^2} + f(x_{2j-1}) \frac{(x - x_{2j-2})(x - x_{2j})}{(-h^2)} +$$

$$+ f(x_{2j}) \frac{(x - x_{2j-2})(x - x_{2j-1})}{2h^2}, \qquad x \in [x_{2j-2}, x_{2j}], \qquad j = 1, \dots, n/2 \quad (6)$$

Проинтегрировав полином второй степени (6) по отрезку $[x_{2j-2}, x_{2j}]$, получим

$$\int_{x_{2j-2}}^{x_{2j}} g_n(x)dx = \frac{h}{3}\{f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})\}, \qquad h = \frac{(b-a)}{n} \quad (7)$$

Интеграл от $g_n(x)$ по всему отрезку $[a, b]$ равен сумме интегралов

$$S_n = \int_a^b g_n(x)dx = \sum_{i=1}^{\frac{n}{2}} \int_{x_{2j-2}}^{x_{2j}} g_n(x)dx =$$

$$\frac{(b-a)}{3n}\{f(a) + 4f(x_1) + 2f(x_2) + \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(b)\} \quad (8)$$

Величина S_n дает приближенное значение интеграла *I*:

$$I = \int_a^b f(x)dx = S_n + \gamma_n \quad (9)$$

Узлами квадратурной формулы (8) являются точки д x_i . Весовые коэффициенты в узлах с четными и нечетными номерами имеют разные значения. Для остаточного члена введено обозначение γ_n . Формула (9) называется *квадратурной формулой Симпсона*.

Пусть функция $f(x)$ четырежды непрерывно дифференцируема на отрезке $[a, b]$.Рассмотрим отрезок двойной длины $2h$ расположенный между точками разбиения (5) с четными номерами $[x_{2j-2}, x_{2j}], j = 1, \dots, n/2$. В курсе математического анализа выводится формула (вывод см. прошлогодние билеты)

$$\int_{x_{i-1}}^{x_i} f(x)dx = \frac{h}{3}\{f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})\} + \frac{h^5}{90}f^{(4)}(\eta_j) \quad (10)$$

где $\eta_j \in [x_{2j-2}, x_{2j}]$. Существование такой точки гарантировано, но ее точное положение неизвестно. Суммируя равенства (10) по *j* получаем квадратурную формулу (9) со следующим выражением для остаточного члена

$$\gamma_n = -\frac{h^5}{90}\sum_{J=1}^{n/2} f^{(4)}(\eta_j) \quad (11)$$

Из формулы (11) можно вывести различные представлении остаточного члена и изучить его свойства. Рассмотрим сумму

$$2h\sum_{J=1}^{n/2} f^{(4)}(\eta_j) \quad (12)$$

Функция $f^{(4)}(x)$ предполагается непрерывной и следовательно, интегрируемой на отрезке $[a, b]$. С учетом этого (12) можно рассматривать как интегральную сумму для интеграла

$$\int_a^b f^{(4)}(x)dx$$

Отсюда следует вывод

$$\lim_{n\to\infty} 2h\sum_{J=1}^{n/2} f^{(4)}(\eta_j) = \int_a^b f^{(4)}(x)dx = f^{(3)}(b) - f^{(3)}(a) \quad (13)$$

Предельное равенство (13) позволяет записать остаточный член формулы Симпсона (11) в виде

$$\gamma_n = \frac{1}{n^4}(C + \sigma_n) \quad (14)$$

$$C = -\frac{(b-a)^4}{180}f^{(3)}(b) - f^{(3)}(a) \quad (15)$$

$$\sigma_n = -\frac{(b-a)^4}{180}\left\{2h\sum_{J=1}^{n/2} f^{(4)}(\eta_j) - \int_a^b f^{(4)}(x)dx\right\} \to 0, \ n \to \infty \quad (16)$$

Эта формула выделяет в остаточном члене γ_n главное слагаемое C/n^4 , которое стремится к 0 как n^{-4}. Коэффициент С не зависит от *n*. Дополнительное слагаемое σ_n/n^4 является бесконечно малой более высокого порядка. Если им пренебречь, получится асимптотическое представление остаточного члена

$$\gamma_n \approx Cn^{-4}$$

Его относительная точность возрастает с увеличением *n*.

Лемма. Пусть функция $f(x)$ непрерывна на отрезке$[a,b]$ и пусть $x_1, \dots, x_n$ - некоторые точки этого отрезка. Тогда на отрезке x_i найдется такая точка η, что

$$\frac{1}{n}\sum_{i=0}^{n} f(x_i) = f(\eta)$$

Применяя лемму к сумму (11), получим другое представление остаточного члена:

$$\gamma_n = -\frac{(b-a)^5}{180n^4} f^{(4)}(\eta) \quad (17)$$

где η - какая-то точка отрезка $[a,b]$. Вычислить погрешность по формуле (17) нельзя, поскольку положение точки η неизвестно, но можно ее оценить. Пусть

$$M_4 = \sup_{x\in[a,b]} |f^{(4)}(x)|$$

тогда

$$|\gamma_n| \le \frac{(b-a)^5 M_4}{180n^4}$$

Данная оценка позволяет определить, с каким n нужно проводить вычислении, чтобы погрешность не превышала заданной точности ε. Кроме того, если четвертая производная функции $f(x)$ - знакоопределенная, то формула (17) даст знак погрешности, что также может оказаться полезным при организации вычислении.

Метод Симпсона является методом более высокого порядка точности - четвертого.

14. Квадратурные формулы Гаусса.

Рассматриваем способы приближенного вычисления определенных интегралов

$$I = \int_a^b f(x)dx \quad (\mathbf{1})$$

Универсальные алгоритмы вычисления определенных интегралов дают формулы численного интегрирования - *квадратурные формулы* вида:

$$I = \int_a^b f(x)dx = \sum_{i=0}^{n} c_i f(x_i) + R_n \quad (2)$$

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

$$\lim_{n\to\infty} R_n \to 0, \text{так что} \quad \lim_{n\to\infty} \sum_{i=0}^{n} c_i f(x_i) \to I \quad (3)$$

Если в формуле (2) пренебречь остаточным членом R_n , то получится приближенное равенство

$$I = \int_a^b f(x)dx = \sum_{i=0}^{n} c_i f(x_i) \quad (4)$$

Постановка 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 $[-1,1]$. Запишем искомую квадратурную формулу в виде

$$\int_{-1}^{1} f(x)dx = \sum_{i=0}^{n} c_i f(x_i) + \delta_n \quad (5)$$

гае x_i - узлы, $x_i \in [-1,1]$; c_i - весовые коэффициенты, δ_n -остаточный член. Для любого полинома степени $(2n-1)$ остаточный член и формуле (5) должен быть равен нулю. Далее под произвольными полиномами будем иметь в виду также полиномы более низких степеней.
Полагая последовательно $f(x) = 1, x, x^2, x^{2n-1}$ и принимая во внимание, что для этих функций, согласно требованию Гаусса, остаточный член должен равняться 0, получаем

$$\int_{-1}^{1} x^m dx = \frac{1}{(m+1)}\{1 + (-1)^m\} = \sum_{i=0}^{n} c_i f(x_i^m), \quad 0 \le m \le 2n-1 \quad (6)$$

Соотношения (6) представляют собой систему $2n$ нелинейных уравнений с $2n$ неизвестными (узлы x_i и веса c_i $(1 \le i \le n)$. Уравнение (6), соответствующее индексу $m = 0$, дает

$$\sum_{i=0}^{n} c_i = 2 \quad (7)$$

Т. о. сумма весовых коэффициентов в квадратурной формуле Гаусса при любом *n* равна 2.
Полиномы Лежандра определяются формулой

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n \quad (8)$$

Выпишем первые полиномы:

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{3}{2}x^2 - \frac{1}{2}, \ , \quad P_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x \quad (9)$$

Свойства полиномов Лежандра:
1. Полином Лежандра $P_n(x)$ номера n является полиномом *n*-й степени, обладающим той же четностью что и n:

$$P_n(-x) = (-1)^n P_n(x) \quad (10)$$

Док-во: напрямую следует из (8).
2. Полиномы Лежандра $(P_n x)$ в точках $x = \pm 1$ принимают следующие значения:

$$P_n(1) = 1, \quad P_n(-1) = (-1)^n$$

Док-во: Представим в виде произведения выражение

$$(x^2-1)^n = (x+1)^n (x-1)^n$$

и выполним n-кратное дифференцирование. В результате получим

$$P_n(x) = \frac{1}{2^n n!}\sum_{k=0}^{n} (C_n^k)^2 n!\,(x+1)^{n-k}(x-1)^k$$

Все члены этой суммы, кроме нулевого, содержат множители $(x-1)^k, 1 \le k \le n$, и при $x = 1$ обращаются в 0, а нулевой член дает нужное равенство $P_n(1) = 1$. Второе равенство следует из (10)
3. Полином Лежандра $(-1)^n$ имеет на интервале $[-1,1]$ n простых корней. В силу свойства 1 корни располагаются симметрично относительно точки $x = 0$.
Док-во: Функция $(x^2-1)^n$ обращается на концах отрезка $[-1,1]$ в нуль. Согласно теореме Ролля ее первая производная должна иметь по крайней мере один нуль на интервале $(-1,1)$ Кроме того, производная обращается в нуль в граничных точках $x = \pm 1$. Применяя таким же образом теорему Ролля ко второй производной $\{(x^2-1)^n\}''$, убеждаемся что она имеет 2 нуля на интервале $(-1,1)$ и обращается и нуль в граничных точка $x = \pm 1$.
Будем продолжать этот процесс пока не дойдем до n-производной выражения $(x^2-1)^n$. Эта производная определяет полином Лежандра с точностью до множителя. Она должна иметь n корней на интервале $(-1,1)$. Т.к. число корней равно степени полинома, все они должны быть простыми. И располагаются они симметрично относительно точки $x = 0$.
4. Любой полином $Q_m(x)$ степени $m < n$ ортогонален к полиному Лежандра $P_n(x)$ на $[-1,1]$:

$$\int_{-1}^{1} Q_m(x) P_n(x) dx = 0 \quad (11)$$

Док-во: Подставим в интеграл (11) представление полинома Лежандра (8) и проинтегрируем по частям. В результате получим

$$J = \frac{1}{2^n n!}\int_{-1}^{1} Q_m(x)\frac{d^n}{dx^n}(x^2-1)^n dx =$$

$$= \frac{1}{2^n n!}\left\{ Q_m(x)\frac{d^{n-1}}{dx^{n-1}}(x^2-1)^n\Big|_{-1}^{1} - \int_{-1}^{1}\frac{dQ_m(x)}{dx}\frac{d^{n-1}}{dx^{n-1}}(x^2-1)^n dx \right\}$$

Подстановки на концах отрезка $[-1,1]$ обращаются в ноль, т.к. степень n у выражения $(x^2-1)^n$ больше (n - 1)-го порядка производной.

Выполняя процедуру интегрирования по частям $m+1 \le n$ раз получаем

$$J = (-1)^{m+1}\frac{1}{2^n n!}\int_{-1}^{1}\frac{d^{m+1}Q_m(x)}{dx^{m+1}}\frac{d^{n-m-1}}{dx^{n-m-1}}(x^2-1)^n dx = 0$$

Здесь под знаком интеграла в качестве множителя стоит $(m+1)$-я производная от полинома m-й степени $Q_m(x)$, тождественно равная нулю. Ортогональность доказана.

Замечание. Соотношение ортогональности (11) справедливо, в частности в случае, когда в качестве полинома $Q_m(x)$ при $m \le n$ взят полином Лежандра $P_m(x)$

$$\int_{-1}^{1} P_m(x)P_n(x)dx = 0, \quad m \ne n$$

Фактически в этом условии ортогональности не важно, какой из двух индексов больше. Важно, что они не равны. Т.о., из св-ва 4 вытекает следствие.

Следствие 1. *Полиномы Лежандра образуют систему полиномов, ортогональных на отрезке* $[-1,1]$.

Из линейной алгебры известно, что система полиномов, ортогональных на некотором множестве, определена однозначно с точностью до множителей. Поэтому справедливо обратное утверждение.

Следствие 2. *Пусть $Q_n(x)$ — полином n-й степени ортогональный на отрезке* $[-1,1]$ *к любому полиному степени m < n , тогда с точностью до множителя он является полиномом Лежандра* $Q_n(x) = A_n P_n(x)$.

Перейдем к решению основной задачи — определению узлов и весовых коэффициентов квадратурных формул Гаусса. Составим полином n -й степени

$$\omega_n(x) = (x-x_1)(x-x_2)\dots(x-x_n) \quad (12)$$

где x_i — искомые узлы. Возьмем произвольный полипом $Q_m(x)$ степени $m < n$, помножим его на полином $\omega_n(x)$ и проинтегрируем произведение по отрезку $[-1,1]$ с помощью квадратурной формулы (5). Поскольку это произведение представляет собой полином степени $n+m \le 2n-1$, формула Гаусса должна быть для него точной. В результате, согласно (12), получим

$$\int_{-1}^{1} Q_m(x)\omega_n(x)dx = \sum_{i=0}^{n} c_i Q_m(x)\omega_n(x_i) = 0 \quad (13)$$

Как видим полином $\omega_n(x)$ ортогонален к любому полиному степени $m < n$. Согласно следствию 2 он с точностью до множителя совпадает с n-м полиномом Лежандра $\omega_n(x) = A_n P_n(x)$. Отсюда следует вывод: узлы квадратурной формулы Гаусса являются корнями полинома Лежандра $P_n(x)$, которые располагаются на интервале $(-1,1)$ симметрично относительно точки $x = 0$.

Чтобы посчитать весовые коэффициенты c_i введем для $m = 1,..,n$ специальные полиномы типа

$$Q_{n-1,m}(x) = \frac{(x-x_1)\dots(x-x_{m-1})(x-x_{m+1})\dots(x-x_n)}{(x_m-x_1)\dots(x_m-x_{m-1})(x_m-x_{m+1})\dots(x_m-x_n)} \quad (14)$$

Каждый из них является полиномом степени $(n-1)$ В числителе у него стоит полином $\omega_n(x)$ с опушенным 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

$$Q_{n-1,m}(x) = \begin{cases} 0, & i \ne m \\ 1, & i = m \end{cases} \quad (15)$$

Для полинома $Q_{n-1,m}(x)$ квадратурная формула Гаусса должна быть точной. С учетом (15) это дает

$$\int_{-1}^{1} Q_{n-1,m}(x)dx = \sum_{i=0}^{n} c_i Q_{n-1,m}(x_i) = c_m \quad (16)$$

В результате получаем следующее интегральное выражение для весовых коэффициентов квадратурной формулы Гаусса

$$c_m = \int_{-1}^{1} Q_{n-1,m}(x)dx = \int_{-1}^{1}\frac{(x-x_1)\dots(x-x_{m-1})(x-x_{m+1})\dots(x-x_n)}{(x_m-x_1)\dots(x_m-x_{m-1})(x_m-x_{m+1})\dots(x_m-x_n)}dx \quad (17)$$

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 (17), действительно решает задачу Гаусса, являясь точной для любого полинома степени $(2n-1)$.
Проведем доказательство в два этапа. Сначала докажем, что данная формула является точной для любого полинома $Q_{n-1}(x)$ степени $(n-1)$. Такой полином можно представить в виде суммы специальных полиномов (14):

$$Q_{n-1}(x) = \sum_{m=1}^{n} Q_{n-1}(x_m) Q_{n-1,m}(x) \qquad (18)$$

Справедливость такого разложения вытекает из следующих соображений. Здесь левая и правая части равенства совпадают в n точках x_i , $i = 1, \dots, n$. Но если два полинома $(n-1)$-й степени совпадают в n точках, то они тождественно равны. Интегрируя равенство (18) по отрезку $[-1,1]$, получаем

$$\int_{-1}^{1} Q_{n-1}(x)dx = \sum_{m=1}^{n} Q_{n-1}(x_m) \int_{-1}^{1} Q_{n-1,m}(x)dx = \sum_{m=1}^{n} c_m Q_{n-1}(x_m) \qquad (19)$$

Итак, для полиномов $(n-1)$ -й степени утверждение доказано. Теперь рассмотрим произвольный полином $Q_{2n-1}(x)$ степени $(2n-1)$. Разделим его с остатком на полином Лежандра $P_n(x)$:

$$Q_{2n-1}(x) = P_n(x) q_{n-1}(x) + r_{n-1}(x)$$

где $q_{n-1}(x)$ и $r_{n-1}(x)$ — полиномы степени $(n-1)$. Проинтегрировав это равенство по отрезку $[-1,1]$, будем иметь

$$\int_{-1}^{1} Q_{2n-1}(x)dx = \int_{-1}^{1} \{P_n(x) q_{n-1}(x) + r_{n-1}(x)\}dx = \int_{-1}^{1} r_{n-1}(x)dx = \sum_{i=0}^{n} c_i r_{n-1}(x_i) =$$

$$= \sum_{i=0}^{n} c_i \{P_n(x_i) q_{n-1}(x_i) + r_{n-1}(x_i)\} = \sum_{i=0}^{n} c_i Q_{2n-1}(x_i)$$

Поясним выполненные преобразования. Интеграл

$$\int_{-1}^{1} P_n(x) q_{n-1}(x)dx$$

опущен, поскольку полином Лежандра $P_n(x)$ ортогонален любому полиному$(n-1)$ -й степени. Оставшийся интеграл от полинома $r_n(x)$ вычислен с помощью квадратурной формулы (19), т.к. уже доказано, что для полиномов степени $(n-1)$ она является точной. Последний переход заключается в том, что в сумму $\sum_{i=0}^{n} c_i r_{n-1}(x_i)$ 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.
Пример. Построить квадратурную формулу Гаусса с двумя узлами (с тремя см. учебник).
Узлы определяются как корни второго полинома Лежандра:

$$P_2(x) = \frac{3}{2}x^2 - \frac{1}{2} \quad => \quad x_1 = -\frac{1}{\sqrt{3}}, \quad x_2 = \frac{1}{\sqrt{3}}$$

Весовые коэффициенты рассчитываются по формулам (17):

$$c_1 = \int_{-1}^{1} \frac{(x-x_2)}{(x_1-x_2)}dx = \int_{-1}^{1} \frac{x-1/\sqrt{3}}{-2/\sqrt{3}}dx = 1$$

$$c_2 = \int_{-1}^{1} \frac{(x-x_1)}{(x_2-x_1)}dx = \int_{-1}^{1} \frac{x+1/\sqrt{3}}{2/\sqrt{3}}dx = 1$$

Они равны между собой и их сумма =2 => выполнено условие (7). Искомая квадратурная формула имеет вид

$$\int_{-1}^{1} f(x)dx = f\left(-1/\sqrt{3}\right) + f\left(1/\sqrt{3}\right) + \delta_2$$

Она точна для любого полинома 3-й степени.

15. Сеточные функции. Разностная аппроксимация 1-й и 2-й производной.

Пусть на отрезке сегменте $[a,b]$ задан набор точек и построена сетка

$$a \leq= x_0 < x_1 < \cdots < x_n \leq b \quad (1)$$

Условимся считать сетку равномерной:

$$x_i - x_{i-1} = h = (b-a)/n, \quad 0 \leq i \leq n \qquad (2)$$

Пусть каждой точке сетки x_i сопоставлено по определенному закону число y_i . Совокупность этих чисел $\boldsymbol{y} = (y_0, \dots y_n) = \{y_i\},\ 0 \leq i \leq n$ назовем *сеточной функцией.* Сеточные функции, определенные на сетке (1), образуют $(n+1)$-мерное линейное пространство.

Чтобы иметь возможность сравнивать сеточные функции между собой, нужно ввести в этом пространстве норму, например, такую:

$$\|\boldsymbol{y}\|_c = \max_{0 \leq i \leq n} |y_i| \quad (3)$$

Это определение законно, т.к. удовлетворяет аксиомам нормы:

1. Норма неотрицательна $\|\boldsymbol{y}\|_c \geq 0$, причем равенство 0 имеет место только для нулевого элемента.
2. Модуль числового множителя можно вынести за знак: $\|\alpha \boldsymbol{y}\|_c = |\alpha| \|\boldsymbol{y}\|_c$
3. Неравенство треугольника : $\|\boldsymbol{y} + \boldsymbol{z}\|_c \leq \|\boldsymbol{y}\|_c + \|\boldsymbol{z}\|_c$

Вытекает из свойства максимума

$$\max_{0 \leq i \leq n} |y_i + z_i| \leq \max_{0 \leq i \leq n} |y_i| + \max_{0 \leq i \leq n} |z_i|$$

Пусть дано дифференциальное уравнение

$$Lu(x) = f(x,u), \qquad \text{например,} \qquad u' = f(x,u)$$

Заменим оператор Lu в узле сетки x_i линейной комбинацией значений сеточной функции y_i на некотором множестве узлов сетки, называемом шаблоном. Такая замена Lu на $L_h u_h$ называется аппроксимацией на сетке дифференциального оператора L разностным оператором L_h . Замена непрерывной функции $f(x,u)$ в узлах сетки на сеточную функцию $\varphi(x_h, y_h)$ называется аппроксимацией правой части.

Т. о. дифференциальное уравнение можно аппроксимировать (заменить) на сетке **разностной схемой**

$$L_h u_h = \varphi(x_h, y_h), \qquad \text{например } \frac{y_{i+1} - y_i}{h} = f(x_i, y_i), \quad 1 \leq i \leq n-1$$

Изучение разностных аппроксимаций проводится сначала локально, т.е. в любом фиксированном узле сетки.

Пусть u_h - проекция непрерывной функции $u(x)$ на сетку (например, $u_h = u(x_i) = u_i$) .

Опр. Погрешностью аппроксимации дифференциального оператора Lu разностным оператором L_h назовем величину $\psi_1 = (Lu)_h - L_h u_h$, где $(Lu)_h$ - проекция на сетку результата действия дифференциального оператора L на функцию u (например, $u'(x_i)$)

Опр. Говорят, что погрешность аппроксимации дифференциального оператора имеет в узле x_i порядок k , если $\psi_1(x_i) = O(h^k) \to 0$ при $h \to 0$.

Опр. Погрешностью аппроксимации правой части f сеточной функцией φ_h назовем величину $\psi_1 = f_h - \varphi_h$, где f_h - проекция на сетку функции $f(x,u)$ (например, $f(x_i, u_i)$).

Опр. Погрешность аппроксимации правой части имеет в узле x_i порядок m , если $\psi_2 = O(h^m) \to 0$ при $h \to 0$

Опр. Погрешностью аппроксимации разностной схемы на решении в узле x_i (локальной погрешностью) назовем величину ψ , равную

$$\psi = \psi_1 - \psi_2 = (Lu)_h - L_h u_h - (f_h - \varphi_h) = \varphi_h - L_h u_h\ , \qquad \text{т. к. } Lu = f$$

$$\left(\text{например,} \qquad \psi = u'(x_i) - \frac{u_{i+1} - u_i}{h} - \big(f(x_i, u_i) - f(x_i, u_i)\big), \quad \text{здесь} \quad \psi_2 = 0\right)$$

Опр. Значения локальной погрешности аппроксимации в каждом узле x_i образуют сеточную функцию погрешности аппроксимации ψ_i .

Обычно требуется оценка погрешности аппроксимации на сетке, т.е. оценка функции ψ_i в некоторой сеточной норме.

Опр. Говорят, что погрешность аппроксимации разностной схемы имеет m-ый порядок на сетке, если $\|\psi\| = O(h^m)$.

Опр. Решение разностной схемы сходится к решению диф. уравнения с порядком k на сетке, если погрешность решения $\|z_h\| = \|u_h - y_h\| = O(h^k) \to 0$ при $h \to 0$.

Для сеточных функций нельзя ввести обычное понятие производной, включающее операцию предельного перехода при $\Delta x \to 0$. Вместо производной вводятся разностные отношения:

правая разностная производная: $L_h^+[y_i] = \frac{y_{i+1} - y_i}{h}, \quad 0 \le i \le n-1 \quad (4)$

левая разностная производная: $L_h^-[y_i] = \frac{y_i - y_{i-1}}{h}, \quad 1 \le i \le n \quad (5)$

центральная разностная производная: $L_h^{(0)}[y_i] = \frac{y_{i+1} - y_{i-1}}{2h}, \quad 1 \le i \le n-1 \quad (6)$

Чтобы установить связь разностных отношений (4)-(6) с обычной производной, предположим, что на отрезке $[a, b]$ определена дифференцируемая функции $y(x)$,я 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 н сопоставим с разностными отношениями (4)-(6):

$$\psi_i^+ = L_h^+[y_i] - y'(x_i), \ 0 \le i \le n-1 \quad (7)$$

$$\psi_i^- = L_h^-[y_i] - y'(x_i), \ 1 \le i \le n \quad (8)$$

$$\psi_i^{(0)} = L_h^{(0)}[y_i] - y'(x_i), \ 1 \le i \le n-1 \quad (9)$$

Эти величины представляют собой погрешности аппроксимации производной с помощью разностных отношения (4)-(6) 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:

$$y_{i+1} = y(x_i + h) = y_i + y'(x_i)h + \frac{1}{2}y''(x_i + \theta_i h)h^2 \quad (10)$$

где θ_i - какое-то неизвестное нам число между 0 и 1. Подставляя разложение (10) в формулу (7), получаем

$$\psi_i^+ = \frac{1}{2}y''(x_i + \theta_i h)h \quad (11)$$

Аналогично получаем

$$\psi_i^- = -\frac{1}{2}y''(x_i - \theta_i h)h \quad (12)$$

Формулы (11) и (12) не позволяют вычислить погрешности, но дают возможность их оценить. Функция $y''(x)$ по предположению непрерывна на отрезке $[a, b]$ и, следовательно, ограничена:

$$|y''(x)| \le M_2, \quad x \in [a, b]$$

В результате получаем

$$|\psi_i^+| \le \frac{1}{2}M_2 h, \quad |\psi_i^-| \le \frac{1}{2}M_2 h \quad (13)$$

Оценки (13) являются равномерными, поскольку не зависят от индекса i . Таким образом, левое и правое разностные отношения аппроксимируют производную $y'(x)$ с первым порядком точности относительно h

Для оценки $\psi_i^{(0)}$ предположим, что функция $y(x)$, три раза непрерывно дифференцируема на отрезке $[a, b]$ и продолжим разложение (10) еще на один член.

$$y_{i+1} = y_i + y'(x_i)h + \frac{1}{2}y''(x_i)h^2 + \frac{1}{6}y'''(x_i + \theta_{1,i}h)h^3 \quad (14)$$

$$y_{i-1} = y_i - y'(x_i)h + \frac{1}{2}y''(x_i)h^2 - \frac{1}{6}y'''(x_i - \theta_{2,i}h)h^3$$

Подставляя разложения (14) в формулу (9), будем иметь

$$\psi_i^{(0)} = \frac{1}{12}\{y'''(x_i + \theta_{1,i}h) + y'''(x_i - \theta_{2,i}h)\}h^2 \quad (15)$$

По предположению функция $y'''(x)$ непрерывна и, следовательно, ограничена на отрезке $[a, b]$:

$$|y'''(x)| \le M_3, \quad x \in [a, b]$$

В результате из равенства (15) получим оценку

$$\left|\psi_i^{(0)}\right| \le \frac{1}{6}M_3 h^2 \quad (16)$$

Оценка (16), не зависит от индекса i, она является равномерной. Таким образом, центральная разностная производная дает более хороший результат аппроксимирует производную $y'(x)$ со вторым порядком точности относительно h для функций, трижды непрерывно дифференцируемых на отрезке $[a, b]$.

Для разностной аппроксимации второй производной составим разностное отношение первых разностных производных:

$$L_h[y_i] = \frac{\frac{y_{i+1} - y_i}{h} - \frac{y_i - y_{i-1}}{h}}{h} = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} \quad (17)$$

Чтобы установить связь выражения (17) 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

$$\psi_i = L_h[y_i] - y''(x_i), \quad 1 \le i \le n-1 \quad (18)$$

Она представляет собой погрешность аппроксимации второй производной с помощью разностного отношения второго порядка (17).

Оценим величину погрешности при предположении, что функция $y(x)$ четыре раза непрерывно дифференцируема на отрезке $[a, b]$. Это предположение позволяет написать разложения Тейлора:

$$y_{i+1} = y(x_i + h) = y_i + y'(x_i)h + \frac{1}{2}y''(x_i)h^2 + \frac{1}{6}y'''(x_i)h^3 + \frac{1}{6}y^{(4)}\big(x_i + \theta_{1,i}h\big)h^4$$

$$y_{i-1} = y(x_i - h) = y_i - y'(x_i)h + \frac{1}{2}y''(x_i)h^2 - \frac{1}{6}y'''(x_i)h^3 + \frac{1}{6}y^{(4)}\big(x_i - \theta_{1,i}h\big)h^4$$

Подставляя их в формулы (17), (18), получаем

$$\psi_i = \frac{1}{24}\{y^{(4)}\big(x_i + \theta_{1,i}h\big) + y^{(4)}\big(x_i - \theta_{1,i}h\big)\}h^2 \quad (19)$$

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 $[a, b]$

$$\left|y^{(4)}(x)\right| \le M_4\,, \quad x \in [a, b]$$

в результате из формулы (19) получаем

$$|\psi_i| \le \frac{1}{12}M_4h^2$$

Таким образом, разностное отношение (17) аппроксимирует вторую производную со вторым порядком точности относительно h для функций, имеющих четыре непрерывные производные на отрезке $[a, b]$ Совершенно так же можно строить разностные аналоги производных более высокого порядка.

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 двумя способами.

1. 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 С:

$$\|\boldsymbol{z}\|_c = \max_{a \le x \le b}|y(x) - u(x)|$$

Во-вторых, наоборот, решению дифференциального уравнения можно сопоставить сеточную функцию $u_i = u(x_i)$ и сршмип 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

$$\|\boldsymbol{z}\|_c = \max_{0 \le i \le n}|y_i - u_i|$$

Наиболее последовательным является первой способ, но обычно выбирают более простой - второй.

16. Метод Эйлера.

Рассмотрим задачу Коши для дифференциального уравнения первого порядка.

$$u' = f(x, u) \quad (1)$$

$$u(x_0) = u_0 \quad (2)$$

Если функция $f(x, u)$ непрерывна и удовлетворяет условию Липшица по аргументу u в некоторой окрестности начальной точки (x_0, u_0),

$$|f(x, u_1) - f(x, u_2)| \le L|u_1 - u_2|, \quad \forall (x, u_1), (x, u_2) \in \text{окрестности } (x_0, u_0)$$

где L - положительная постоянная, то можно указать такой отрезок $[a, b], a < x_0 < b$, на котором решение задачи (1), (2) $u(x)$ существует и единственно.

Пусть нам нужно построить решение задачи (1), (2) на отрезке $[x_0, x_0 + l]$ длины l. Возьмем некоторое целое число n, введем шаг $h = l/n$ и образуем на отрезке сетку

$$x_i = x_0 + ih, \ \ 1 \le i \le n \quad (3)$$

Сопоставим задаче (1), (2) на отрезке разностную дачу

$$\frac{y_{i+1} - y_i}{h} = f(x_i, y_i), \quad 1 \le i \le n-1 \quad (4)$$

$$y_0 = u_0 \quad (5)$$

Здесь производная $u'(x)$ в уравнении (1) заменена правой разностной производной и сохранено неизменным начальное условие (2).

Уравнение (4) является разностным уравнением первого порядка, которое принято называть *схемой Эйлера*. Его можно переписать в виде рекуррентного соотношения:

$$y_{i+1} = y_i + hf(x_i, y_i), \quad 1 \le i \le n-1 \quad (6)$$

Это позволяет последовательно рассчитать все значения сеточной функции $\{y_i\}$, решив тем самым задачу (4), (5). Такую разностную схему называют *явной*.

Перейдем теперь к обсуждению главного вопроса: с какой точностью рассчитанная сеточная функция $\{y_i\}$ дает решение исходной задачи Коши $u(x)$? Рассмотрим решение задачи (1), (2) в точках сетки (3), образовав 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 $\boldsymbol{z}$ и ψ:

$$z_i = y_i - u_i, \ \ 1 \le i \le n \quad (7)$$

$$\psi_i = \frac{u_{i+1} - u_i}{h} - f(x_i, u_i), \qquad 1 \le i \le n-1 \quad (8)$$

Первая функция (7) характеризует разницу между рассчитанными числами y_i и решением $u(x)$ задачи (1), (2) в точках сетки x_i. В соответствии с этим сеточную функцию $\boldsymbol{z}$ называют *погрешностью решения*.

Функции ψ (8) получается в результате подстановки решения диф. уравнения (1) в разностное уравнение (4). Если бы эти уравнения совпадали, то мы получили бы нуль. Но они различаются, и нуля мы не получим. Сеточную функцию ψ характеризующую степень близости дифференциального и разностного уравнений, называют *погрешностью аппроксимации* схемы на решении.

Установим связь между сеточными функциями $\boldsymbol{z}$ и ψ. Из формулы (7) найдем

$$y_i = u_i + z_i \quad (9)$$

и подставим в разностное уравнение (4):

$$\frac{z_{i+1} - z_i}{h} + \frac{u_{i+1} - u_i}{h} = f(x_i, u_i + z_i)$$

или $$\frac{z_{i+1} - z_i}{h} = \{f(x_i, u_i + z_i) - f(x_i, u_i)\} - \left\{\frac{u_{i+1} - u_i}{h} - f(x_i, u_i)\right\} \quad (10)$$

Здесь в обе {} вставлена величина $f(x_i, u_i)$ с противоположными знаками, т.е. равенство не нарушается. Вторые {} =ψ_i. В первых {} стоит разность значений функции f при одинаковом 1-м аргументе x_i и разных значениях 2-го аргумента. Эта разность представляется по формуле Лагранжа:

$$f(x_i, u_i + z_i) - f(x_i, u_i) = \frac{\partial f}{\partial u}(x_i, u_i + \theta_i z_i) z_i \qquad 0 \le \theta_i \le 1$$

Формула (10) записывается в виде рекуррентного соотношения:

$$z_{i+1} = \left\{1 + h\frac{\partial f}{\partial u}(x_i, u_i + \theta_i z_i)\right\} z_i - \psi_i h \ , \quad 1 \le i \le n-1 \quad (11)$$

Согласно (2) и (5) его следует дополнить нулевым начальным условием

$$z_0 = 0 \quad (12)$$

В отличие от (4), (5) формулы (11), (12) не могут быть использованы для вычисления величин $\boldsymbol{z}_i$. В них входят неизвестные величины ψ_i, u_i, θ_i. Но из этой системы рекуррентных равенств можно получить рекуррентные неравенства.

Введем для оценки сеточной функции ψ ее норму:

$$\|\psi\|_c = \max_{0 \le i \le n-1} |\psi_i| \qquad \text{при этом } \ |\psi_i| \le \|\psi\|_c \quad (13)$$

Пусть функция $\frac{\partial f}{\partial u}(x, u)$ в интересующей нас области изменения ее аргументов ограничена:

$$\left|\frac{\partial f}{\partial u}(x, u)\right| \le C$$

Тогда получаем оценку:

$$\left|1 + h\frac{\partial f}{\partial u}(x_i, u_i + \theta_i z_i)\right| \le 1 + Ch < e^{Ch} = q \ , \quad q > 1 \quad (14)$$

С учетом (13) и (14) из формулы (11) следуют рекуррентные неравенства

$$|z_{i+1}| \le q|z_i| + \|\psi\|_c h$$

которые порождают цепочку оценок

$$z_0 = 0$$
$$|z_1| \le \|\psi\|_c h$$
$$|z_2| \le (1+q)\|\psi\|_c h$$
$$|z_3| \le (1+q+q^2)\|\psi\|_c h \quad (15)$$
$$\dots$$
$$|z_n| \le (1+q+q^2+\dots+q^{n-1})\|\psi\|_c h$$

Согласно (14) $q > 1$, поэтому

$$1+q+q^2+\dots+q^{n-1} < nq^n = ne^{Chn}$$

Это позволяет заменить индивидуальные оценки (15) универсальной оценкой

$$|z_i| \le nhe^{Chn}\|\psi\|_c \quad 1 \le i \le n \quad (16)$$

Неравенства (16) справедливы при любом i в частности при том, при котором $|z_i|$ достигает своего наибольшего значения и определяет тем самым норму сеточной функции $\|\boldsymbol{z}\|_c$. В результате оценка погрешности решения принимает вид

$$\|\boldsymbol{z}\|_c \le le^{Cl}\|\psi\|_c \quad (17)$$

где *l* — длина отрезка, на котором рассматривается решение исходной задачи (1), (2)

Мы получили важный результат, оценку погрешности решения через оценку погрешности аппроксимации схемы 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 $\|\psi\|_c$. Пусть функция $f(x,u)$ имеет в рассматриваемой области изменения аргументов непрерывные и ограниченные первые частные производные $\partial f/\partial x$ и $\partial f/\partial u$. Это обеспечивает существование у решения задачи (1), (2) непрерывной и ограниченной второй производной

$$u''(x) = \frac{\partial f}{\partial x}(x,u) + \frac{\partial f}{\partial u}(x,u)f(x,u) \quad (18)$$

Запишем для функции $u(x)$ формулу Тейлора с остаточным членом в форме Лагранжа:

$$u_{i+1} = u(x_i + h) = u_i + u'(x_i)h + \frac{1}{2}u''(x_i + \theta_i h)h^2 \quad 0 \le \theta_i \le 1 \quad (19)$$

Подставляя разложение (19) в формулу (8) для погрешности аппроксимации уравнения, получаем

$$\psi_i = \frac{1}{2}u''(x_i + \theta_i h)h$$

Согласно формуле (18) функция $u''(x)$ непрерывна и ограничена:

$$|u''(x)| \le M_2 \quad x \in [x_0, x_0 + l]$$

Это позволяет написать оценки

$$\|\psi\|_c \le \frac{M_2}{2}h \qquad \|\boldsymbol{z}\|_c \le \frac{M_2 l}{2}e^{Cl}h$$

Эти неравенства показывают, что при $h \to 0$ аппроксимации уравнения и связанная с ней неравенством (17) погрешность решения стремятся к нулю со скоростью h. В связи с этим метод Эйлера называют *методом 1-го порядка точности* относительно h.

Мы подробно разобрали метод Эйлера, поскольку примере простой разностной схемы (4) 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 $O(h)$.

17. Метод Рунге-Кутта.

Рассмотрим задачу Коши для дифференциального уравнения первого порядка.

$$u' = f(x,u) \quad (1)$$
$$u(x_0) = u_0 \quad (2)$$

Если функция $f(x,u)$ непрерывна и удовлетворяет условию Липшица по аргументу *u* в некоторой окрестности начальной точки (x_0, u_0),

$$|f(x,u_1) - f(x,u_2)| \le L|u_1 - u_2|, \quad \forall (x,u_1),(x,u_2) \in \text{окрестности } (x_0,u_0)$$

где *L* - положительная постоянная, то можно указать такой отрезок $[a,b], a < x_0 < b$, на котором

решение задачи (1), (2) $u(x)$ существует и единственно.
Пусть нам нужно построить решение задачи (1), (2) на отрезке $[x_0, x_0 + l]$ длины l. Возьмем некоторое целое число n, введем шаг $h = l/n$ и образуем на отрезке сетку

$$x_i = x_0 + ih, \quad 1 \le i \le n \quad (3)$$

Сопоставим задаче (1), (2) на отрезке разностную задачу по схеме Эйлера

$$\frac{y_{i+1} - y_i}{h} = f(x_i, y_i), \quad 1 \le i \le n-1 \quad (4)$$

$$y_0 = u_0 \quad (5)$$

Здесь производная $u'(x)$ в уравнении (1) заменена правой разностной производной и сохранено неизменным начальное условие (2). Или в виде рекуррентного соотношения:

$$y_{i+1} = y_i + hf(x_i, y_i), \quad 1 \le i \le n-1 \quad (6)$$

Пусть решение диф. уравнения $u(x)$ имеет производные достаточно высокого порядка, напишем разложение по формуле Тейлора:

$$u_{i+1} = u(x_i + h) = u_i + u'(x_i)h + \frac{1}{2}u''(x_i)h^2 + \frac{1}{6}u'''(x_i)h^3 + \cdots \quad (7)$$

Оборвем разложение (7) на члене порядка h^2 . Пусть функция $f(x, u)$ имеет в рассматриваемой области изменения аргументов непрерывные и ограниченные первые частные производные $\partial f/\partial x$ и $\partial f/\partial u$. Это обеспечивает существование у решения задачи (1), (2) непрерывной и ограниченной второй производной

$$u''(x) = \frac{\partial f}{\partial x}(x, u) + \frac{\partial f}{\partial u}(x, u)f(x, u) \quad (8)$$

Используя (8) для вычисления производной $u''(x_i)$, получим новое рекуррентное соотношение:

$$y_{i+1} = y_i + f(x_i, y_i)h + \frac{1}{2}\left\{\frac{\partial f}{\partial x}(x_i, y_i) + \frac{\partial f}{\partial y}(x_i, y_i)f(x_i, y_i)\right\}h^2 \quad (9)$$

или в виде разностного уравнения

$$\frac{y_{i+1} - y_i}{h} = f(x_i, y_i) + \frac{1}{2}\left\{\frac{\partial f}{\partial x}(x_i, y_i) + \frac{\partial f}{\partial y}(x_i, y_i)f(x_i, y_i)\right\}h \quad (10)$$

Главная идея метода Рунге-Кутта: приближенно заменить правую часть формулы (10) на сумму значений функции f в двух разных точках с точностью до членов порядка h^2. Положим

$$f(x_i, y_i) + \frac{1}{2}\left\{\frac{\partial f}{\partial x}(x_i, y_i) + \frac{\partial f}{\partial u}(x_i, y_i)f(x_i, y_i)\right\}h = \beta f(x_i, y_i) + \alpha f(x_i + \gamma h, y_i + \delta h) + O(h^2) \quad (11)$$

где $\alpha, \beta, \gamma, \delta$ — 4 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:

$$f(x_i + \gamma h, y_i + \delta h) = f(x_i, y_i) + \left\{\gamma\frac{\partial f}{\partial x}(x_i, y_i) + \delta\frac{\partial f}{\partial y}(x_i, y_i)\right\}h + O(h^2) \quad (12)$$

подставим разложение (12) в формулу (11) и приравняем слева и справа члены, не содержащие h и содержащие h в первой степени. В результате получим для четырех параметров три уравнения:

$$\alpha + \beta = 1, \quad \alpha\gamma = \frac{1}{2}, \quad \alpha\delta = \frac{1}{2}f(x_i, y_i)$$

Они позволяют выразить параметры β, γ, δ через α:

$$\beta = 1 - \alpha, \quad \gamma = \frac{1}{2\alpha}, \quad \delta = \frac{1}{2\alpha}f(x_i, y_i) \quad (13)$$

Заменяя с помощью (11) левую часть уравнения (10) и отбрасывая члены порядка $O(h^2)$, получаем однопараметрическое семейство разностных схем Рунге-Кутта

$$\frac{y_{i+1} - y_i}{h} = (1 - \alpha)f(x_i, y_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, y_i + \frac{h}{2\alpha}f(x_i, y_i)\right) \quad (14)$$

Уравнение (14) можно записать в виде удобного для расчетов рекуррентного соотношения

$$y_{i+1} = y_i + \left[(1 - \alpha)f(x_i, y_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, y_i + \frac{h}{2\alpha}f(x_i, y_i)\right)\right]h \quad (15)$$

Наиболее удобные разностные схемы этого семейства соответствуют двум значениям параметра $\alpha = 1/2$ и $\alpha = 1$.
При $\alpha = 1/2$ рекуррентная формула (15) принимает

$$y_{i+1} = y_i + \frac{h}{2}\{f(x_i, y_i) + f(x_i + h, y_i + hf(x_i, y_i))\} \quad (16)$$

Она определяет следующую процедуру расчета y_{i+1}. Сначала делается шаг h по схеме Эйлера и вычисляется величина

$$\tilde{y}_{i+1} = y_i + f(x_i, y_i)h \quad (17)$$

Затем находится значение функции f точке $f(x_{i+1}, \tilde{y}_{i+1})$, составляется полусумма

$$\frac{f(x_i, y_i) + f(x_{i+1}, \tilde{y}_{i+1})}{2}$$

и проводится окончательный расчет величины

$$y_{i+1} = y_i + \frac{f(x_i, y_i) + f(x_{i+1}, \tilde{y}_{i+1})}{2}h \quad (18)$$

Такая схема вычислений называется «предиктор -корректор» или «предсказание - исправление». Вычисление $\tilde{y}_{i+1}$ по схеме Эйлера (17) - это грубое предсказание результата. Вторичный расчет (18), сделанный на основании первого, является уточнением результата, его коррекцией.

При $\alpha = 1$ рекуррентная формула (5) имеет вид

$$y_{i+1} = y_i + f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2}f(x_i, y_i)\right) \quad (19)$$

Схема расчета заключается в следующем. Сначала делается половинный шаг $h/2$: по схеме Эйлера вычисляется величина

$$y_{i+\frac{1}{2}} = y_i + \frac{h}{2}f(x_i, y_i) \quad (20)$$

Затем находится значение функции f в точке $\left(x_{i+\frac{1}{2}}, y_{i+\frac{1}{2}}\right)$. Оно определяет по формуле (19) очередное значение y_{i+1}.

Процедура расчета приближенного решения задачи Коши (1), (2) по схеме (14) по сравнению со схемой Эйлера усложняется: теперь на каждом шаге функцию $f(x, u)$ приходится считать 2 раза. Однако такое усложнение оказывается оправданным благодаря более высокой точности метода.

Введем две сеточные функции: погрешность решения $\boldsymbol{z}$ и погрешность аппроксимации схемы ψ:

$$z_i = y_i - u_i, \quad 1 \le i \le n \quad (21)$$

$$\psi_i = \frac{u_{i+1} - u_i}{h} - \left[(1-\alpha)f(x_i, u_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, u_i + \frac{h}{2\alpha}f(x_i, u_i)\right)\right], \qquad 1 \le i \le n-1 \quad (22)$$

Выразим $y_i = u_i + z_i$ и подставим в разностное уравнение (10). В результате получим

$$\frac{z_{i+1} - z_i}{h} + \frac{u_{i+1} - u_i}{h} = (1-\alpha)f(x_i, u_i + z_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, u_i + z_i + \frac{h}{2\alpha}f(x_i, u_i + z_i)\right) \quad (23)$$

Формулу (23) можно переписать в виде

$$\frac{z_{i+1} - z_i}{h} = \left\{\left[(1-\alpha)f(x_i, u_i + z_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, u_i + z_i + \frac{h}{2\alpha}f(x_i, u_i + z_i)\right)\right] - \left[(1-\alpha)f(x_i, u_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, u_i + \frac{h}{2\alpha}f(x_i, u_i)\right)\right]\right\} - \left\{\frac{u_{i+1} - u_i}{h} - \left[(1-\alpha)f(x_i, u_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, u_i + \frac{h}{2\alpha}f(x_i, u_i)\right)\right]\right\} \quad (24)$$

Здесь в каждую {} добавлено одно и то же слагаемое. Учитывая знаки, в целом значение выражения не меняется. Но во вторых {} собраны члены, которые дают погрешность аппроксимации.

Перейдем к дальнейшему исследованию соотношения (24). Рассмотрим функцию

$$F(v) = (1-\alpha)f(x_i, v) + \alpha f\left(x_i + \frac{h}{2\alpha}, v + \frac{h}{2\alpha}f(x_i, v)\right) \quad (25)$$

Выражение, стоящее в первых {} формулы (24), 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:

$$F(u_i + z_i) - F(u_i) = F'(u_i + \theta_i z_i)z_i \quad 0 < \theta_i < 1 \quad (26)$$

где

$$F'(v) = (1-\alpha)\frac{\partial f}{\partial v}(x_i, v) + \alpha\frac{\partial f}{\partial v}\left(x_i + \frac{h}{2\alpha}, v + \frac{h}{2\alpha}f(x_i, v)\right)\left(1 + \frac{h}{2\alpha}\frac{\partial f}{\partial v}(x_i, v)\right) \quad (27)$$

Подставим полученные выражения для отдельных слагаемых в формулу (24). В результате она примет вид рекуррентной формулы

$$z_{i+1} = \{1 + hF'(u_i + \theta_i z_i)\}z_i - \psi_i h \quad 0 \le i \le n-1 \quad (28)$$

которую нужно дополнить нулевым начальным условием $z_0 = 0$. Использовать эту формулу для вычисления значений сеточной функции $\boldsymbol{z}$ нельзя, т.к. неизвестны величины ψ_i, u_i, θ_i. Но эту систему равенств можно заменить системой рекуррентных неравенств для оценки z_i. Пусть функция $\frac{\partial f}{\partial u}(x, u)$ в интересующей нас области изменения ее аргументов ограничена:

$$\left|\frac{\partial f}{\partial u}(x, u)\right| \le C$$

Тогда с учетом формулы (27) для производной $F'(v)$ получим

$$|1 + hF'(u_i + \theta_i z_i)| \le 1 + Ch + \frac{1}{2}C^2h^2 < e^{Ch} = q, \quad q > 1 \quad (29)$$

=> равенства (28) можно заменить рекуррентными неравенствами

$$|z_{i+1}| \le q|z_i| + \|\psi\|_c h$$

которые порождают цепочку оценок

$$\begin{gathered} z_0 = 0 \\ |z_1| \le \|\psi\|_c h \\ |z_2| \le (1+q)\|\psi\|_c h \\ |z_3| \le (1+q+q^2)\|\psi\|_c h \quad (30) \\ \dots \\ |z_n| \le (1+q+q^2+\dots+q^{n-1})\|\psi\|_c h \end{gathered}$$

Согласно (29) $q > 1$, поэтому

$$1 + q + q^2 + \dots + q^{n-1} < nq^n = ne^{Chn}$$

Это позволяет заменить индивидуальные оценки (30) универсальной оценкой

$$|z_i| \le nhe^{Chn}\|\psi\|_c \quad 1 \le i \le n \quad (31)$$

Неравенства (31) справедливы при любом i в частности при том, при котором $|z_i|$ достигает своего наибольшего значения и определяет тем самым норму сеточной функции $\|\boldsymbol{z}\|_c$. В результате оценка погрешности решения принимает вид

$$\|\boldsymbol{z}\|_c \le le^{Cl}\|\psi\|_c \quad (32)$$

где *l* — длина отрезка, на котором рассматривается решение исходной задачи (1), (2).

Теперь нужно оценить норму погрешности аппроксимации уравнения (22) Пусть функция $f(x, u)$ имеет в рассматриваемой области изменения аргументов непрерывные вторые производные и, следовательно, решение диф. уравнения $u(x)$ трижды непрерывно дифференцируемо. Это позволяет написать следующие разложения Тейлора:

$$u_{i+1} = u(x_i + h) = u_i + u'(x_i)h + \frac{1}{2}u''(x_i)h^2 + \frac{1}{6}u'''(\bar{x}_i)h^3 \quad (33)$$

$$f\left(x_i + \frac{h}{2\alpha}, u_i + \frac{h}{2\alpha}f(x_i, u_i)\right) = f(x_i, u_i) + \frac{h}{2\alpha}\left\{\frac{\partial f}{\partial x}(x_i, u_i) + \frac{\partial f}{\partial u}(x_i, u_i)f(x_i, u_i)\right\} +$$
$$+ \frac{h^2}{8\alpha^2}\left\{\frac{\partial^2 f}{\partial x^2}(\tilde{x}_i, \tilde{u}_i) + \frac{\partial^2 f}{\partial x \partial u}(\tilde{x}_i, \tilde{u}_i)f(\tilde{x}_i, \tilde{u}_i) + \frac{\partial^2 f}{\partial u^2}(\tilde{x}_i, \tilde{u}_i)f^2(\tilde{x}_i, \tilde{u}_i)\right\} \quad (34)$$

где $\bar{x}_i = x_i + \overline{\theta}_i h$, $\tilde{x}_i = x_i + \tilde{\theta}_i\frac{h}{2\alpha}$, $\tilde{u}_i = u + \tilde{\theta}_i\frac{h}{2\alpha}f(x_i, u_i)$, $0 < \overline{\theta}_i < 1$, $0 < \tilde{\theta}_i < 1$,

Здесь последние слагаемые в обоих разложениях являются остаточными членами в форме Лагранжа, которые берутся в неизвестных нам промежуточных точках.

Подставим разложения (33), (34) в формулу (22) для погрешности аппроксимации диф. уранения (1) и примем во внимание соотношения, вытекающие ю этого уравнения

$$u'(x_i) = f(x_i, u_i)$$
$$u''(x_i) = \frac{\partial f}{\partial x}(x_i, u_i) + \frac{\partial f}{\partial u}(x_i, u_i)f(x_i, u_i)$$

Благодаря этому члены нулевого и первого порядков относительно *h* сокращаются и остаются только члены второго порядка, обязанные своим происхождением остаточным членам в разложениях (33), (34). В результате получается следующее представление для погрешности аппроксимации:

$$\psi_i = h^2\left\{\frac{1}{6}u'''(\bar{x}_i) - \frac{1}{8\alpha}\left[\frac{\partial^2 f}{\partial x^2}(\tilde{x}_i,\tilde{u}_i) + \frac{\partial^2 f}{\partial x \partial u}(\tilde{x}_i,\tilde{u}_i)f(\tilde{x}_i,\tilde{u}_i) + \frac{\partial^2 f}{\partial u^2}(\tilde{x}_i,\tilde{u}_i)f^2(\tilde{x}_i,\tilde{u}_i)\right]\right\} \quad (35)$$

Функции, входящие в правую часть этого предположению непрерывны и ограничены в интересующей нас области изменения своих аргументов => можно заменить равенство (35) неравенством

$$|\psi_i| \le \|\psi\|_c \le Mh^2 \quad (36)$$

гае *M* - константа, мажорирующая выражение в {} формулы (35). Подставляя оценку (36) в неравенство (32) получаем

$$\|\boldsymbol{z}\|_c \le Mle^{Cl}h^2$$

Т.о., при $h \to 0$ погрешность аппроксимации и, как следствие, погрешность решения стремятся к нулю со скоростью h^2 Это означает, что разностное уравнение по схеме Рунге-Кутта имеет 2-й порядок точности относительно h. 2-й порядок точности лучше, чем 1-й, но практика показывает, что этой точности также недостаточно. Наиболее часто при проведении реальных расчетов используется схема Рунге-Кутта 4-го порядка точности следующего вида:

$$\frac{y_{i+1} - y_i}{h} = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$k_1 = f(x_i, y_i), \quad k_2 = f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2}k_1\right), \quad k_3 = f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2}k_2\right), \quad k_4 = f(x_i + h, y_i + hk_3)$$

Если в схеме 2-го порядка точности на каждом шаге функцию $f(x,y)$ приходилось вычислять 2 раза, то здесь - 4 раза. Но усложнение схемы расчета окупается высокой точностью.

18. Метод Адамса.

Рассмотрим задачу Коши для дифференциального уравнения первого порядка.

$$u' = f(x,u) \quad (1)$$

$$u(x_0) = u_0 \quad (2)$$

Если функция $f(x,u)$ непрерывна и удовлетворяет условию Липшица по аргументу *u* в некоторой окрестности начальной точки (x_0, u_0),

$$|f(x,u_1) - f(x,u_2)| \le L|u_1 - u_2|, \quad \forall (x,u_1),(x,u_2) \in \text{окрестности } (x_0,u_0)$$

где *L* - положительная постоянная, то можно указать такой отрезок $[a,b], a < x_0 < b$, на котором решение задачи (1), (2) $u(x)$ существует и единственно.

Пусть нам нужно построить решение задачи (1), (2) на отрезке $[x_0, x_0 + l]$ длины *l*. Возьмем некоторое целое число *n*, введем шаг $h = l/n$ и образуем на отрезке сетку

$$x_i = x_0 + ih, \quad 1 \le i \le n \quad (3)$$

Пусть $u(x)$ - решение диф. уравнения (1). Для производной этой функции выполняется равенство

$$u' = f(x,u(x)) = F(x) \quad (4)$$

Интегрируя его между двумя точками сетки, получаем соотношение

$$u_{i+1} = u_i + \int_{x_i}^{x_{i+1}} F(x)dx \quad (5)$$

Мы не можем использовать это соотношение непосредственно для перехода в процессе решения задачи от *i*-й точки сетки к $(i+1)$-й, поскольку функция $F(x)$ нам неизвестна. Чтобы сделать следующий шаг, нужно приближенно заменить эту функцию на такую, которую можно вычислить. Опишем, как это проблема решается в методе Адамса. Пусть в процессе численного решения задачи расчет доведен до точки x_i . В результате расчетов стали известны величины y_j и $f(x_j, y_j)$, $0 \le j \le i$ Возьмем некоторое фиксированное целое число $m \le i$ и построим интерполяционный многочлен *m*-й степени $P_m(x)$, принимающий в точках x_j , $(i-m) \le j \le i$ значения $f(x_j, y_j)$:

$$P_m(x_j) = f(x_j, y_j), \quad (i-m) \le j \le i \quad (6)$$

Его можно записать по формуле Лагранжа:

$$P_m(x) = \sum_{j=i-m}^{i} f(x_j, y_j)Q_{m,j}(x) \quad (7)$$

где $Q_{m,j}(x)$ — специальные многочлены вила

$$Q_{m,j}(x) = \frac{(x - x_{i-m}) \dots (x - x_{j-1})(x - x_{j+1}) \dots (x - x_i)}{(x_j - x_{i-m}) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_i)} \quad (8)$$

Главная идея метода Адамса заключается в том, чтобы для расчета y_{j+1} использовать формулу типа (5), приближенно заменяя в ней функцию $F(x)$ на интерполяционный многочлен $P_m(x)$, составленный, согласно (7), по результатам предыдущих вычислений. Это приводит к рекуррентной формуле

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} P_m(x)dx = y_i + \sum_{j=i-m}^{i} a_j f(x_j, y_j) \qquad (9)$$

$$\text{где} \quad a_j = \int_{x_i}^{x_{i+1}} Q_{m,j}(x)dx \qquad (10)$$

Рассмотрим более подробно данную схему численного решения задачи Копи в простейших случаях $m = 0$ и $m = 1$. При $m = 0$ для аппроксимации функции $F(x)$ используется полином 0-й степени, т.е. постоянная $F(x) = P_0 = f(x_j, y_j)$. Т.е. формула (9) переходит в рекуррентную формулу метода Эйлера

$$\frac{y_{i+1} - y_i}{h} = f(x_i, y_i)$$

обеспечивающую 1-й порядок точности.

Исследуем варианта $m = 1$. В этом случае для аппроксимации функции $F(x)$ используется полином 1-й степени, построенный по значениям функции f в двух точках (x_{i-1}, y_{i-1}) и (x_i, y_i):

$$P_1(x) = f(x_i, y_i)\frac{x - x_{i-1}}{h} - f(x_{i-1}, y_{i-1})\frac{x - x_i}{h}$$

Подставляя его в формулу (9) и проводя интегрирование, получаем

$$y_{i+1} = y_i + \left\{\frac{3}{2}f(x_i, y_i) - \frac{1}{2}f(x_{i-1}, y_{i-1})\right\}h \qquad (11)$$

Отметим следующую особенность рекуррентной формулы (1). 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 (11) 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 (11) можно записать в виде разностного уравнения

$$\frac{y_{i+1} - y_i}{h} = \frac{3}{2}f(x_i, y_i) - \frac{1}{2}f(x_{i-1}, y_{i-1}) \qquad (12)$$

Подсчитаем для него погрешность аппроксимации схемы:

$$\psi_i = \frac{u_{i+1} - u_i}{h} - \left\{\frac{3}{2}f(x_i, y_i) - \frac{1}{2}f(x_{i-1}, y_{i-1})\right\} = \frac{u_{i+1} - u_i}{h} - \left\{\frac{3}{2}u'(x_i) - \frac{1}{2}u'(x_{i-1})\right\} \qquad (13)$$

Пусть функция $f(x, u)$ имеет в интересующей нас области изменения аргументов непрерывные вторые производные, тогда решение задачи $u(x)$ трижды непрерывно дифференцируемо. Запишем разложения Тейлора

$$u_{i+1} = u_i + u'(x_i)h + \frac{1}{2}u''(x_i)h^2 + \frac{1}{6}u'''(x_i + \overline{\theta}_i h)h^3, \quad 0 \le \overline{\theta}_i \le 1$$

$$u'(x_{i-1}) = u'(x_i) - u''(x_i)h + \frac{1}{2}u'''\left(x_i - \overline{\overline{\theta}}_i h\right)h^2, \quad 0 \le \overline{\overline{\theta}}_i \le 1$$

Подставляя их в формулу (13), получаем

$$\psi_i = \{\frac{1}{6}u'''(x_i + \overline{\theta}_i h) + \frac{1}{4}u'''\left(x_i - \overline{\overline{\theta}}_i h\right)\}h^2$$

Отсюда можно дать оценку

$$|\psi_i| \le \|\psi\|_c \le \frac{5}{12}M_3 h^2$$

где M_3 - постоянная, мажорирующая 3-ю производную функции $u(x)$:

$$|u'''(x)| \le M_3 \quad x \in [x_0, x_0 + l]$$

Мы видим, что разностное уравнение метода Адамса, соответствующее случаю $m = 1$, аппроксимирует диф. уравнение (1) со 2-м порядком точности относительно h . Как и в случае метода Рунге-Кутта, это обеспечивает 2-й порядок точности для погрешности решения $\|\boldsymbol{z}\|_c$ 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 3-го порядка точности, при $m = 3$ - 4-го и т.д. Схема 4-го порядка является наиболее употребительной.

Если написать интерполяционный полином 3-ей степени $P_3(x)$ на сетке из четырех точек x_i , x_{i-1}, x_{i-2} , x_{i-3} и провести интегрирование (10), то рекуррентная формула (9) примет вид:

$$y_{i+1} = y_i + h\left\{\frac{55}{24}f(x_i, y_i) - \frac{59}{24}f(x_{i-1}, y_{i-1}) + \frac{37}{24}f(x_{i-2}, y_{i-2}) - \frac{9}{24}f(x_{i-3}, y_{i-3})\right\} \quad (14)$$

Приведем еще одну форму записи этой формулы через так называемые конечные разности:

$$y_{i+1} = y_i + hf_i + \frac{1}{2}h^2\Delta^1 f_i + \frac{5}{12}h^3\Delta^2 f_i + \frac{3}{8}h^4\Delta^3 f_i \quad (15)$$

где

$$f_i = f(x_i, y_i) \qquad \Delta^1 f_i = \frac{1}{h}\{f(x_i, y_i) - f(x_{i-1}, y_{i-1})\}$$

$$\Delta^2 f_i = \frac{1}{h^2}\{f(x_i, y_i) - 2f(x_{i-1}, y_{i-1}) + f(x_{i-2}, y_{i-2})\} \quad (16)$$

$$\Delta^3 f_i = \frac{1}{h^3}\{f(x_i, y_i) - 3f(x_{i-1}, y_{i-1}) + 3f(x_{i-2}, y_{i-2}) - f(x_{i-3}, y_{i-3})\}$$

1-я, 2-я и 3-я разности (16) приближенно соответствуют 1-й, 2-й и 3-й производной функции $F(x) = f(x, u(x))$. Эквивалентность формул (14) и (15) легко проверить непосредственно. Формула (15) иногда более удобна для организации вычислительного процесса и контроля точности.

Особенность метода Адамса проявляется в формуле (15) еще сильнее, чем в формуле (11). 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 (14) начинает работать только с четвертой точки => y_1, y_2 , y_3 приходится рассчитывать другим методом, например Рунге-Кутта.

Перейдем к обсуждению точности схемы (14). Если функция $f(x, u)$ имеет непрерывные четвертые производные по своим аргументам в интересующей нас области их изменения, т.е. решение задачи $u(x)$ 5 раз непрерывно дифференцируемо, то разностное уравнение (14) аппроксимирует диф. уравнение (1) с четвертым порядком точности относительно h. Доказательство этого утверждения проводится так же, как и для схемы второго порядка (11), только теперь в разложениях нужно удерживать больше членов. Четвертый порядок точности при аппроксимации уравнения обеспечивает четвертый порядок точности для погрешности решения $\|\boldsymbol{z}\|_c$ 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.

Сравним схемы 4-го порядка точности в методах Рунге-Кутта и Адамса с точки зрения организации вычислительного процесса. Чтобы сделать один шаг по методу Рунге-Кутта, необходимо вычислить функцию $f(x, u)$ 4 раза, а в методе Адамса только 1 раз, т.к. в трех предшествующих точках функция $f(x, u)$ была уже вычислена на предыдущих шагах. В этом заключается главное достоинство метода Адамса, которое особенно высоко ценилось в докомпьютерное время.

Главный недостаток метода Адамса в том, что при его применении первые шаги приходится делать с помощью другого метода, например Рунге-Кутта, и только тогда можно перейти на расчет по схеме Адамса. Т. о., 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 диф. уравнения часто приходится менять шаг *h*. В методе Рунге-Кутта это не составляет труда, поскольку каждый шаг делается независимо от предыдущего. В методе Адамса 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.

19. Разностная аппроксимация краевой задачи для линейного диффер. уравнения второго порядка.

Рассмотрим следующую задачу для линейного дифференциального уравнения второго порядка:

$$u'' - q(x)u = -f(x), \quad a < x < b \quad (\mathbf{1})$$

$$u(a) = u_1 \quad u(b) = u_2 \quad (\mathbf{2})$$

Здесь два дополнительных условия заданы в граничных точках отрезка $[a, b]$, поэтому задачу (1), (2) называют *краевой.*

Пусть функции $f(x)$ и $q(x)$ непрерывны на отрезке $[a,b]$, причем $q(x) \geq q_0 > 0$ $(\mathbf{3})$
При сделанных предположениях, как известно из курса дифференциальных уравнений, решение задачи (1), (2) существует и является единственным. Возьмем некоторое целое число *n*, введем шаг $h = (b-a)/n$ и построим сетку $x_i = a + ih, \quad 0 \leq i \leq n$ $(\mathbf{4})$
Заменим диф. уравнение (1) его разностным аналогом. В результате получим следующую задачу:

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} - q_i y_i = -f_i \quad 1 \leq i \leq n-1 \quad (\mathbf{5})$$

$$y_0 = u_1 \quad y_n = u_2 \qquad (\mathbf{6})$$

Здесь $q_i = q(x_i)$, $f_i = f(x_i)$, граничные условия (6) для сеточной функции $\{y_i\}$ взяты такими же, что и в дифференциальной задаче.
Разностные уравнения (5) можно переписать в виде

$$y_{i-1} - (2 + q_i h^2) y_i + y_{i+1} = -f_i h^2 \qquad 1 \leq i \leq n-1 \quad (\mathbf{7})$$

Получили линейную систему из $(n-1)$-го уравнения с $(n-1)$-м неизвестным y_i, $1 \leq i \leq n-1$
Значения y_0, y_n неизвестными не являются, они задаются граничными условиями (6).
В краевой задаче (5), (6) сеточная функция $\{y_i\}$ определяется из решения системы линейных алгебраических уравнений. Такая разностная схема называется *неявной*.
Из записи разностных уравнений в форме (7) видно, что мы получили систему уравнений, содержащих трехдиагональную матрицу с диагональным преобладанием: диагональный элемент $(2 + q_i h^2)$ больше суммы двух других элементов той же строки, равной 2. Системы такого рода уже встречались в билете 11 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 (5), (6), приближает решение краевой задачи (1), (2). 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 (подставляем точное решение u_i вместо приближенного y_i):

$$z_i = y_i - u_i \quad 0 \leq i \leq n \quad (\mathbf{8})$$

$$\psi_i = \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} - q_i u_i + f_i \quad 1 \leq i \leq n-1 \quad (\mathbf{9})$$

Выразим y_i в соотношении (8) $y_i = z_i + u_i$ и подставим в разностное уравнение (5). Оставим члены, содержащие z_i, слева, а остальные члены перенесем направо. В результате получим

$$\frac{z_{i-1} - 2z_i + z_{i+1}}{h^2} - q_i z_i = -\left\{\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} - q_i u_i + f_i\right\} = -\psi_i, \quad 1 \leq i \leq n-1 \quad (\mathbf{10})$$

Граничные условия в дифференциальной и разностной задачах совпадают, поэтому значения сеточной функции z_i в граничных точках будут нулевыми: $z_0 = z_n = 0$ $(\mathbf{11})$
Погрешность $\{z_i\}$ нельзя рассчитать, решая задачу (10), (11), поскольку в правые части уравнений входят неизвестные величины u_i и ψ_i Однако задача (10), (11) позволяет оценить эту погрешность.
Пусть максимальное по модулю число z_i соответствует индексу $i = j$:

$$\|z\|_C = |z_j| \geq |z_i|, \qquad 0 \leq i \leq n, \quad (\mathbf{12})$$

$$\text{где } \|z\|_C = \max_{0 \leq i \leq n} |z_i|$$

В граничных точках z_i обращается в 0, поэтому индекс *j* не равен ни 0, ни *n*. Рассмотрим уравнение (10) для этого значения индекса и запишем его в виде

$$(2 + q_j h^2) z_j = z_{j-1} + z_{j+1} + \psi_j h^2 \quad (\mathbf{13})$$

Возьмем модуль от обеих частей равенства и оценим правую часть сверху:

$$(2 + q_j h^2)|z_j| = (2 + q_j h^2)\|z\|_C \leq |z_{j-1}| + |z_{j+1}| + |\psi_j| h^2 \leq 2\|z\|_C + \|\psi\|_C h^2$$

$$\text{или } \|z\|_C \leq \frac{1}{q_0}\|\psi\|_C \quad (\mathbf{14})$$

Здесь мы сократили одинаковые члены слева и справа, разделили обе части неравенства на множитель $q_j h^2$ и заменили q_j в знаменателе на минимально возможное значение функции $q(x)$на отрезке $[a,b]$ равное q_0 (см. (3)) Таким образом, нам удалось оценить погрешность решения $\|z\|_C$ через погрешность аппроксимации уравнения $\|\psi\|_C$.
Для оценки погрешности аппроксимации уравнения предположим, что функции $f(x)$ и $q(x)$ дважды непрерывно дифференцируемы на отрезке $[a,b]$. В этом случае уравнение (1) допускает двухкратное дифференцирование, что обеспечивает существование у решения краевой задачи (1), (2) четырех непрерывных производных и позволяет написать разложения:

$$u_{i-1} = u_i - u'(x_i)h + \frac{1}{2}u''(x_i)h^2 - \frac{1}{6}u'''(x_i)h^3 + \frac{1}{24}u^{(4)}(x_i - \tilde{\theta}_i h)h^4\,,\ \ 0 \le \tilde{\theta}_i \le 1$$

$$u_{i+1} = u_i + u'(x_i)h + \frac{1}{2}u''(x_i)h^2 + \frac{1}{6}u'''(x_i)h^3 + \frac{1}{24}u^{(4)}(x_i + \overline{\tilde{\theta}}_i h)h^4\,,\ \ 0 \le \overline{\tilde{\theta}}_i \le 1$$

Подставляя их в формулу (9), получим следующее выражение:

$$\psi_i = \left\{u''^{(x_i)} - q_i u_i + f_i\right\} + \frac{h^2}{24}\left\{u^{(4)}\left(x_i - \tilde{\theta}_i h\right) + u^{(4)}\left(x_i + \overline{\tilde{\theta}}_i h\right)\right\} \quad (\mathbf{15})$$

Выражение в первых фигурных скобках =0 в силу дифференциального уравнения (1). В результате в правой части формулы (15) остается только вторая группа членов, обязанная своим происхождением остаточным членам. Оценим ее следующим образом. Функция $u^{(4)}(x)$ *непрерывна* и следовательно, ограничена на отрезке $[a, b]$. Пусть $\left|u^{(4)}(x)\right| \le M_4\,, a \le x \le b$, тогда из формул (14), (15) получаем

$$\|\psi\|_C \le \frac{M_4}{12}h^2\,,\ \ \|z\|_C \le \frac{M_4}{12q_0}h^2$$

Мы видим, что разностная схема (5) обеспечивает второй порядок аппроксимации уравнения и, как следствие неравенства (14) второй порядок точности для погрешности решения.

20. Разностная задача на собственные значения.

Задача на собственные значения

$$u''(x) + \lambda u(x) = 0, \qquad a < x < b, \qquad u(a) = u(b) = 0 \quad (1)$$

имеет решение $\lambda_k = \left(\frac{\pi k}{b-a}\right)^2$, $u_k = \sin\frac{\pi k(x-a)}{b-a}$, $k = 1,2, \ldots$

Рассмотрим на равномерной сетке $\omega_h = \{x_i = a + ih, i = 0,1, \ldots, N,\ h = (b-a)/N\}$ $(\mathbf{2})$ разностный аналог задачи (1)

$$\frac{y_{j-1} - 2y_j + y_{j+1}}{h^2} + \lambda^{(h)}y_j = 0 \quad j = 1,2, \ldots, N-1 \quad (\mathbf{3})$$

$$y_0 = y_N = 0,\ \ hN = b - a, \quad y_j = y\left(x_j\right),\ \ x_j = a + jh$$

Система уравнений (3) представляет собой задачу на собственные значения $Ay = \lambda^{(h)}y$ для симметричной матрицы

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ & & & \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots & & & \\ 0 & 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{bmatrix}$$

порядка $N-1$. Поэтому существует ровно $N-1$ вещественных собственных значений $\lambda_k^{(h)}, k = 1,2, \ldots N-1$, матрицы *A*. Построим в явном виде собственные значения и собственные функции задачи (3). Перепишем разностное уравнение (3) в виде

$$y_{j-1} - (2-\mu)y_j + y_{j+1} = 0,\ \ \mu = h^2\lambda^{(h)} \quad (\mathbf{4})$$

и рассмотрим отвечающее (4) характеристическое уравнение $q^2 - (2-\mu)q + 1 = 0$ $(\mathbf{5})$

Общее решение уравнения (4) имеет вид $y_j = c_1 q_1^j + c_2 q_2^j$ $(\mathbf{6})$

где c_1, c_2 - произвольные постоянные и q_1, q_2 - корни уравнения (5).

Из граничных условий $y_0 = y_N = 0$ получаем $c_1 + c_2 = 0,\ c_1 q_1^N + c_2 q_2^N = 0$

Эта однородная система уравнений имеет нетривиальное решение при условии $q_1^N = q_2^N$.

Учитывая, что $q_1 q_2 = 1$ приходим к условию $q_1^{2N} = 1$ $(\mathbf{7})$

Отсюда, представляя q_1 в тригонометрической форме: $q_1 = \rho e^{i\varphi}$, получим $\rho = 1$ и

$$\varphi = \frac{\pi k}{N},\ \ k = 1,2, \ldots, N-1 \quad (\mathbf{8})$$

С другой стороны, из уравнения (5) имеем

$$q_1 = 1 - \frac{\mu}{2} + \sqrt{\left(1 - \frac{\mu}{2}\right)^2 - 1}$$

следовательно, $\cos\varphi = 1 - 0{,}5\mu$

и из (8) получим

$$\mu = 2(1-\cos\varphi) = 4\sin^2\frac{\varphi}{2} = 4\sin^2\frac{\pi k}{2N}$$

Таким образом, собственные числа задачи (3) имеют вид

$$\lambda^{(h)} = \frac{4}{h^2}\sin^2\frac{\pi k}{2N}, \quad k = 1,2,\dots N-1, \qquad hN = b-a \quad (\mathbf{9})$$

Собственные функции y_j вычисляются согласно (6), где $c_2 = -c_1$. Так как $q_1 q_2 = 1$, то

$$y_j = c_1\left(q_1^j - q_2^j\right) = c_1\left(q_1^j - q_1^{-j}\right) = c_1(e^{ij\varphi} - e^{-ij\varphi})$$

где φ определено согласно (8). Полагая $c_1 = -0{,}5i$, получим

$$y_j^{(k)} = \sin\frac{\pi kj}{N}\ , \quad k,j = 1,2,\dots N-1 \quad (\mathbf{10})$$

Собственные функции (10) определены с точностью до произвольного постоянного (не зависящего от j) множителя. Интересно сопоставить решения дифференциальной (1) и разностной (3) задач на собственные значения. Значения собственных функций (10) разностной задачи совпадают в точках сетки со значениями собственных функций дифференциальной задачи. Спектр дифференциальной задачи не ограничен, т. е. $\lambda_k \to \infty$ при $k \to \infty$, в то время как спектр разностной задачи ограничен сверху при каждом фиксированном шаге h числом $4h^{-2}$. Для каждого фиксированного номера $k \le k_0$, где k_0 не зависит от h, собственные значения $\lambda_k^{(h)}$ разностной задачи сходятся при $h \to 0$ к соответствующему собственному значению h_k дифференциальной задачи, т. е.

$$\lim_{h\to 0}\frac{4}{h^2}\sin^2\frac{\pi kh}{2(b-a)} = \left(\frac{\pi k}{b-a}\right)^2 = \lambda_k$$

При этом собственные значения разностной задачи (3) всегда меньше соответствующих собственных значений дифференциальной задачи (1). Погрешность $\lambda_k - \lambda_k^{(h)}$ минимальна для малых номеров k и сильно возрастает с ростом k.

Свойства собственных значений и собственных функций. Перечислим свойства собственных значений и собственных функций разностной задачи (3). Прежде всего из (9) видно, что

$$0 < \lambda_1^{(h)} < \lambda_2^{(h)} < \dots < \lambda_k^{(h)} < \lambda_{k+1}^{(h)} < \dots \lambda_{N-1}^{(h)} < \frac{4}{h^2}$$

Последнее неравенство неулучшаемо, так как

$$\lambda_{N-1}^{(h)} = \frac{4}{h^2}\cos^2\frac{\pi h}{2(b-a)} \quad \text{и} \quad \cos^2\frac{\pi h}{2(b-a)} \to 1 \quad \text{при } h \to 0$$

Оценку снизу для наименьшего собственного значения λ_1 можно уточнить. Обозначая $\alpha = \pi h/(2(b-a))$, получим

$$\lambda_1^{(h)} = \lambda_1\left(\frac{\sin\alpha}{\alpha}\right)^2, \quad \text{где } \lambda_1 = \left(\frac{\pi}{b-a}\right)^2$$

- наименьшее собственное значение дифференциальной задачи. Не ограничивая общности, можно предположить, что $h \le (b-a)/3$. Тогда $\alpha \le \pi/6$, и поскольку функция $\sin\alpha/\alpha$ монотонно убывает при $\alpha \in [0, \frac{\pi}{6}]$, получим

$$\left(\frac{\sin\alpha}{\alpha}\right)^2 \ge \left(\frac{1}{2}\frac{6}{\pi}\right)^2 = \frac{9}{\pi^2}\ , \quad \text{т.е. } \lambda_1^{(h)} \ge \frac{9}{(b-a)^2} \quad (\mathbf{11})$$

Т.о., наименьшее собств. значение задачи (3) отделено от 0 константой $\delta_1 = \frac{9}{(b-a)^2}$, не зависящей от h.

Док-во, что собственные функции (10) задачи (3), отвечающие различным собственным значениям, ортогональны в смысле скалярного произведения, см. учебник стр. 42-43.