Численное решение СЛАУ

Прямые методы решения СЛАУ

Прямыми называются методы, которые позволяют получить точное решение невырожденной системы

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

за конечное число операций. Если система **невырожденная**, то ее решение всегда **существует и является единственным**

Формулы Крамера

Алгоритм

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

$$x_j = \frac{\Delta_j}{\Delta}, \quad \Delta_j = \det A_j, \quad j = \overline{1,n}. \tag{2}$$

Здесь матрица A_j получается из матрацы A заменой ее j-го столбца столбцом правых частей системы.

Число операций

Чтобы найти решение системы (1), нужно подсчитать $n+1$ определитель. Определитель n-ого порядка – это $n!$ слагаемых, каждое из которых является произведением чисел. Таким образом, для его вычисления нужно выполнить $(n-1)n!$ умножений и $n!$ сложений – всего $Q_n = n \cdot n!$ арифметических операций.

По формуле Стирлинга $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$, так что $Q_n \approx \sqrt{2\pi} \cdot n^{\frac{3}{2}}(\frac{n}{e})^n$

Метод Гаусса

Не ограничивая общности, будем считать, что коэффициент a_{11}, который называют ведущим элементом первого шага, отличен от нуля (в случае $a_{11} = 0$ поменяем местами уравнения с номерами 1 и i, при котором $a_{11} \neq 0$ поскольку система предполагается невырожденной, то такой номер i заведомо найдется)

Прямой и обратный ход

Прямой ход – это приведение матрицы системы $Ax = f$ к треугольному виду.

Нормировка первой строки. Поделим первую строку на a_{11}, т.е. выполним следующие преобразования:

$$a_{ij}^{(1)} = \frac{a_{1j}}{a_{11}}, \quad f_1^{(1)} = \frac{f_1}{a_{11}}, \quad j = \overline{1,n}. \tag{3}$$

Обнуление первого столбца. Вычтем из строки i первую строку, домноженную на a_{i1}, т.е.

$$a_{ij}^{(1)} = a_{ij} - a_{i1} \cdot a_{1j}^{(1)}, \quad f_i^{(1)} - a_{i1} \cdot f_1^{(1)}, \quad i = \overline{2,n}, \quad j = \overline{1,n}. \tag{4}$$

Повторяем эти два шага для полученной матрицы $A^{(1)}$, пока не получим треугольную матрицу, из которой можно *по цепочке выразить все неизвестные*.

Число операций

Первый шаг прямого хода требует n делений и $n(n-1)$ сложений и $n(n-1)$ умножений. Тогда на шагах от 1 до $n-1$ суммарно будет Q_1 делений и Q_2 сложений и умножений, где

$$\begin{gathered} Q_1 = n + (n-1) + \ldots + 1 = \frac{1}{2}n(n-1). \\ Q_2 = n(n-1) + (n-1)(n-2) + \ldots + 2 \cdot 1 = \frac{1}{3}n(n^2-1). \end{gathered} \tag{5}$$

На обратном ходе число сложений и умножений подсчитывается по формуле

$$Q_3 = 1 + 2 + \ldots + (n-1) = \frac{1}{2}n(n-1) \tag{6}$$

Сумма Q_2 и Q_3 дает общее число сложений и умножений, необходимое для решения СЛАУ по методу Гаусса:

$$Q = Q_2 + Q_3 = \frac{1}{3}n^3 + O(n^2). \tag{7}$$

Выбор ведущего элемента по строке

На обратном ходе неизвестные x_i вычисляются всегда с погрешностями.

Предположим, что в процессе приведения системы к треугольному виду у матрицы $A^{(1)}$ образовались большие по модулю элементы $|a_{ij}^{(1)} > 1|$.

Тогда на обратном ходе умножение чисел x_i на большие по модулю элементы $a_{ij}^{(1)}$ приводит к увеличению ошибок.

Поэтому перед нормировкой каждой строки нужно менять местами диагональный элемент и максимальный элемент. В итоге получится треугольная матрица, все элементы которой $|a_{ij}^{(n-1)} \leqslant 1|$. Благодаря этому ошибка не будет расти.

Системы с диагональным преобладанием

Определение

Говорят, что система является **системой с диагональным преобладанием по строке**, если элементы матрицы A удовлетворяют неравенствам:

$$|a_{ii}| > \sum_{j=1,\, j \neq i}^{n} |a_{ij}|, \; 1 \leq i \leq n \tag{8}$$

Теорема

> **Система с диагональным преобладанием всегда разрешима, причём единственным образом.**

Рассмотрим соответствующую однородную систему:

$$\sum_{j=1}^{n} a_{ij} x_j = 0, \; 1 \leq j \leq n. \tag{9}$$

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

$$|x_k| > |x_i|, \; i = \overline{1, n} \tag{10}$$

$$|x_k| \neq |x_j|,\ j = \overline{1, n} \tag{10}$$

Запишем k-ое уравнение системы в виде

$$a_{kk}x_k = -\sum_{j \neq k} a_{kj}x_j \tag{11}$$

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

$$|a_{kk}||x_k| \leqslant \sum_{j \neq k} |a_{kj}||x_j| \leqslant |x_k| \sum_{j \neq k} |a_{kj}| \tag{12}$$

Получили к противоречию с 8. Теорема доказана.

Метод прогонки.

Система с трехдиагональной матрицей

$$\begin{gathered} A_i x_{i-1} + B_i x_i + C_i x_{i+1} = F_i,\ i = \overline{1, n-1} \\ x_0 = q_0,\ x_n = q_n \end{gathered} \tag{13}$$

где коэффициенты A_i, B_i, C_i, правые части $F_i\ (i = 1, \ldots, n-1)$ известны вместе с числами q_0 и q_n. Дополнительные соотношения $x_0 = q_0,\ x_n = q_n$ называют **краевыми условиями**

Матрица этой системы имеет **трёхдиагональную структуру**:

$$\begin{bmatrix} B_1 & C_1 & 0 & 0 & \ldots & 0 & 0 \\ A_2 & B_2 & C_2 & 0 & \ldots & 0 & 0 \\ 0 & A_3 & B_3 & C_3 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & B_{n-2} & C_{n-2} \\ 0 & 0 & 0 & 0 & \ldots & A_{n-1} & B_{n-1} \end{bmatrix} \tag{14}$$

Правая часть: $f = (F_1, \ldots, F_{n-1})^T$. Неизвестные: $x = (x_1, \ldots, x_n)^T$.

Прогоночные коэффициенты.

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

$$x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1},\ 0 \leq i \leq n-1 \tag{15}$$

Здесь величины $\alpha_{i+1}, \beta_{i+1}$ называют **прогоночными коэффициентами**. Для реализации описанной программы выразим x_{i-1} через x_{i+1}:

$$x_{i-1} = \alpha_i x_i + \beta_i = \alpha_i \alpha_{i+1} x_{i+1} + \alpha_i \beta_{i+1} + \beta_i \tag{16}$$

Подставим x_{i-1} и x_i в исходные уравнения. В результате получим:

$$(A_i \alpha_i \alpha_{i+1} + B_i \alpha_{i+1} + C_i) x_{i+1} + A_i \alpha_i \beta_{i+1} + A_i \beta_i + B_i \beta_{i+1} - F_i = 0,\ i = \overline{1, n-1}. \tag{17}$$

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

$$\begin{gathered} A_i \alpha_i \alpha_{i+1} + B_i \alpha_{i+1} + C_i = 0 \\ A_i \alpha_i \beta_{i+1} + A_i \beta_i + B_i \beta_{i+1} - F_i = 0 \end{gathered} \tag{18}$$

Отсюда следуют рекуррентные соотношения для прогоночных коэффициентов:

$$\alpha_{i+1} = \frac{-C_i}{A_i \alpha_i + B_i}, \quad \beta_{i+1} = \frac{F_i - A_i \beta_i}{A_i \alpha_i + B_i}, \quad i = \overline{1, n-1}. \tag{19}$$

Прямой и обратный ходы

Нахождение прогоночных коэффициентов будем называть **прямым ходом** метода прогонки. Вслед за ним можно приступить к **обратному ходу** – нахождению неизвестных $x_{n-1}, \dots, x_1$.

Левое граничное условие $x_0 = q_0$ и соотношение $x_0 = \alpha_1 x_1 + \beta_1$ дают

$$\alpha_1 = 0, \quad \beta_1 = q_0, \tag{20}$$

после чего можно приступать к прямому ходу.

Правое граничное условие $x_n = q_n$ позволяет начать обратный ход с вычисления $x_{n-1} = \alpha_n x_n + \beta_n$.

Лемма о прогоночных коэффициентах

> **Если система с трёхдиагональной матрицей удовлетворяет условию диагонального преобладания, то**
>
> $$|\alpha_i| \leqslant 1, \ \forall i. \tag{21}$$

Воспользуемся ММИ:

- база: при $i = 1$ имеем $\alpha_i = 0 < 1$
- шаг: пусть верно для i, тогда для $i + 1$

$$|\alpha_{i+1}| = \left| \frac{C_i}{A_i \alpha_i + B_i} \right| \leqslant \frac{|C_i|}{|B_i| - |A_i|} \leqslant 1. \quad \blacksquare \tag{22}$$

Это неравенство обеспечивает следующее: если при округлении x_i была допущена ошибка, то в силу формулы

$$x_{i-1} = \alpha_i x_i + \beta_i \tag{23}$$

при вычислении следующих неизвестных эта ошибка не будет расти

Число обусловленности матрицы

Корректность по Адамару

> Математическая задача корректна, если выполняются три условия:
>
> 1. Решение задачи существует.
> 2. Решение задачи единственное.
> 3. Решение задачи непрерывно зависит от входных данных.

Рассмотрим решение СЛАУ с невырожденной матрицей.

$$Ax = f. \tag{24}$$

Запишем её решение как $x = A^{-1} f$. Пусть f подвергся возмущению δf и стал равным

$$\tilde{f} = f + \delta f. \tag{25}$$

Тогда решение $\tilde{x}$ возмущенной системы $A\tilde{x} = \tilde{f}$:

$$\tilde{x} = A^{-1}\tilde{f} = A^{-1} f + A^{-1}\delta f = x + \delta x \tag{26}$$

Отсюда получаем $\|\delta x\| \leq \|A^{-1}\| \|\delta f\|$. Это неравенство означает непрерывную зависимость δx от δf.

Итог: решение СЛАУ с невырожденной матрицей – это корректная математическая задача: для нее выполняются все три требования корректности Адамара.

Число обусловленности матрицы

По свойствам нормы матрицы:

$$\|f\| \leq \|A\| \, \|x\| \tag{27}$$

Перемножая его с неравенством $\|\delta x\| \leq \|A^{-1}\| \|\delta f\|$, получим:

$$\|f\| \|\delta x\| \leq \|A\| \|A^{-1}\| \|x\| \|\delta f\| \tag{28}$$

Пусть $f \neq 0$ и $x \neq 0$, тогда неравенство можно переписать в виде:

$$\frac{\|\delta x\|}{\|x\|} \leq M_A \frac{\|\delta f\|}{\|f\|} \tag{29}$$

где $M_A = \|A\| \|A^{-1}\|$ — **число обусловленности** матрицы A.

Оно позволяет оценить относительную погрешность решения через относительную погрешность возмущения правой части. Матрицы с большим M_A называют **плохо обусловленными**.

Оценка числа обусловленности

> **Для числа обусловленности матрицы A справедливо неравенство**
>
> $$M_A \geq \frac{|\lambda_{max}|}{|\lambda_{min}|} \tag{30}$$
>
> **Соотношение корректно, поскольку в силу невырожденности матрицы $\lambda_{min} \neq 0$.**

В самом деле пусть y – это собственный вектор матрицы A, отвечающий $\lambda_{\max}$: $Ay = \lambda_{max} y$. Тогда

$$|\lambda_{max}| \|y\| = \|Ay\| \leq \|A\| \cdot \|y\| \Leftrightarrow |\lambda_{max}| \leq \|A\| \tag{31}$$

Аналогичным образом для собственного вектора z, связанного с λ_{min} , имеем

$$Az = \lambda_{\min} z \Rightarrow z/\lambda_{\min} = A^{-1} z \Rightarrow \|z\| |\lambda_{min}|^{-1} \leq \|z\| \|A^{-1}\| \Leftrightarrow |\lambda_{min}|^{-1} \leq \|A^{-1}\| \tag{32}$$

Перемножая два последних неравенства, получим

$$M_A \geq \frac{|\lambda_{max}|}{|\lambda_{min}|}. \tag{33}$$

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

$$\|A\| = |\lambda_{\max}|, \;\; \|A^{-1}\| = \frac{1}{\lambda_{\min}}, \tag{34}$$

поэтому для таких матриц

$$M_A = \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{35}$$

Выводы о числе обусловленности

Из полученной оценки для M_A следуют два важных вывода:

1. $M_A \geq 1$
2. Число обусловленности тем больше, чем больше разброс характеристических чисел матрицы. Поэтому с увеличением размера матрицы, вообще говоря, её обусловленность имеет тенденцию к ухудшению.

Итерационные методы

Построение итерационных последовательностей

При применении **итерационных методов** ответ получается в процессе построения последовательных приближений x_k, сходящихся к решению системы в пространстве с евклидовой нормой $\|x\|$

Критерием предельного равенства в конечномерном евклидовом пространстве E_n является **покомпонентная сходимость**:

$$\lim_{k\to\infty} x_i^k = x_i,\ i \le i \le n. \tag{36}$$

Сходимость итерационной последовательности обеспечивает принципиальную возможность получить в процессе итераций ответ с **любой наперед заданной точностью.**

Итерационные алгоритмы

Если каждый x_k вычисляется через предудыщие m членов, то говорят об m**-шаговом итерационном алгоритме.** В простейшем случае имеем дело с одношаговым алгоритмом: $x_{k+1} = F(x_k)$.

Стандартная каноническая форма:

$$B_{k+1}\frac{x_{k+1}-x_k}{\tau_{k+1}} + Ax_k = f,\ \det B_{k+1} \neq 0,\ \tau_{k+1} > 0 \tag{37}$$

В такой записи процесс характеризуется последовательностью матриц B_{k+1} и числовых параметров τ_{k+1} , которые называют **итерационными параметрами**. Если они не зависят от k, то итерационный процесс называется **стационарным**.

Построение очередной итерации равносильно решению СЛАУ, поэтому B нужно выбирать простым, например, единичная матрица, диагональная или треугольная.

$$B_{k+1}x_{k+1} = (B_{k+1} - \tau_{k+1}A)x_k + \tau_{k+1}f. \tag{38}$$

Сходимость

Итерационный метод можно применять только если последовательность действительно сходится к решению. Для исследования этого вопросы вводим две характеристики

- **Погрешность решения** $z_k = x_k - x$

$$\lim_{k\to\infty}\|z_k\| = 0 \Rightarrow \lim_{k\to\infty} z_i^k = 0, i = \overline{1,n}. \tag{39}$$

- **Невязка:** $\psi_k = Ax_k - f$. Она показывает, насколько хорошо или, наоборот, плохо член итерационной
  последовательности x_k удовлетворяет исходной системе.

Связь погрешности и невязки

Установим связь между z_k и ψ_k:

$$\psi_k = Ax_k - f = A(z_k + x) - f = Az_k \Rightarrow \|\psi_k\| \le \|A\|\cdot\|z_k\|,\ \|z_k\| \le \|A^{-1}\|\cdot\|\psi_k\| \tag{40}$$

Они показывают, что погрешность решения z_k эквивалентна невязке ψ_k. Это позволяет судить о сходимости к решению по невязке, а не по погрешности, т.е. судить о сходимости к решению, не зная самого решения.

Самосопряженные положительные операторы

Определения:

1. $A = A^*,$ если $(Ax, y) = (x, Ay)$ $(a_{ij} = a_{ji})$
2. $A > 0,$ если $(Ax, x) > 0,\ x \neq 0$

Свойства:

1. $A = A^* \Rightarrow$ все собственные значения *вещественны*
2. $A = A^* \Rightarrow \exists$ набор из собственных *ортонормированных* векторов
3. Критерием положительной определенности самосопряженной матрицы A является *критерий Сильвестра*.

Лемма о положительности собственных значений

> **Для того, чтобы эрмитова матрица была положительно определенной, необходимо и достаточно, чтобы все её характеристические числа были положительны:** $\lambda_i > 0$

Необходимость. Векторы e_i — ОНБ из собственных векторов,
$A > 0 \Rightarrow (Ae_i, e_i) = \lambda_i > 0,\ i = \overline{1, n}.$

Достаточность. $\lambda_i > 0,\ i = \overline{1, n};\ \forall x = \sum_{i=1}^{n} \xi_i e_i;\ (Ax, x) = \sum_{i=1}^{n} \xi_i^2 \lambda_i > 0$

Лемма об оценке (Ax, x)

> **Для матрицы эрмитовой матрицы** $A > 0$ **верно соотношение**
>
> $$\lambda_{min}\|x\|^2 \leq (Ax, x) \leq \lambda_{max}\|x\|^2 \tag{41}$$

Лемма 3

> **Для** $\forall A > 0$ **найдётся** $\delta > 0$ **такое, что**
>
> $$(Ax, x) > \delta\|x\|^2 \tag{42}$$

Теорема Самарского

> **Пусть** $A : A = A^*,\ A > 0, B - \frac{\tau}{2}A > 0,\ \tau > 0.$ **Тогда при любом выборе** x_0 **итерационный процесс**
>
> $$B\frac{x_{k+1} - x_k}{\tau} + Ax_k = f \tag{43}$$
>
> **сходится к решению системы.**

$$B - \frac{\tau}{2}A > 0 \Leftrightarrow (Bx, x) > \frac{\tau}{2}(Ax, x) > 0,\ \forall x \in E^n,\ x \neq 0 \tag{44}$$

То есть матрица B является положительной, а значит, $\exists B^{-1}$.

Выразим x_k как $x_k = z_k + x$ и подставим в формулу процесса. В результате получим:

$$B\frac{z_{k+1} - z_k}{\tau} + Az_k = 0 \Leftrightarrow Bz_{k+1} = (B - \tau A)z_k \Rightarrow z_{k+1} = z_k - \tau \cdot \underbrace{B^{-1}Az_k}_{\omega_k} \tag{45}$$

$$z_{k+1} = z_k - \tau\omega_k.$$

Рассмотрим последовательность положительных функционалов $J_k = (Az_k, z_k)$:

$$\begin{gathered} J_{k+1} = (Az_k - \tau A\omega_k, z_k - \tau\omega_k) = (Az_k, z_k) - \tau(A\omega_k, z_k) - \tau(Az_k, \omega_k) + \tau^2(A\omega_k, \omega_k) \Rightarrow \\ \{\text{поскольку } A = A^*, \text{ то } (A\omega_k, z_k) = (Az_k, \omega_k) = (B\omega_k, \omega_k)\} \\ J_{k+1} = J_k - 2\tau(B\omega_k, \omega_k) + \tau^2(A\omega_k, \omega_k) = J_k - 2\tau\left(\left[B - \frac{\tau}{2}A\right]\omega_k, \omega_k\right). \end{gathered} \tag{46}$$

Последовательность функционалов J_k с учетом $B - \frac{\tau}{2}A > 0$ образует монотонно невозрастающую последовательность, ограниченную снизу нулем, значит, она сходится. Тогда по лемме 3 $\exists \delta$:

$$\left(\left[B - \frac{\tau}{2}A\right]\omega_k, \omega_k\right) \geqslant \delta\|\omega_k\|^2 \geqslant 0 \Rightarrow \underbrace{J_k - J_{k+1}}_{\to 0} \geqslant 2\tau\delta\|\omega_k\|^2 > 0 \Rightarrow$$
$$\Rightarrow \|\omega_k\| \to 0 \tag{47}$$
$$\|z_k\| = \|A^{-1}B\omega_k\| \leqslant \|A^{-1}\|\|B\|\|\omega_k\| \Rightarrow$$
$$\|z_k\| \to 0$$

Сходимость доказана.

Замечание

Из неравенства 44 следует **интервал сходимости**:

$$0 < \tau < \tau_0 = \inf_{x \neq 0} \frac{2(Bx, x)}{(Ax, x)}. \tag{48}$$

Метод простой итерации.

Определение

В качестве матрицы B выбирается единичная матрица E. Это **явный стационарный метод**, когда очередная итерация x_{k+1} вычисляется по рекуррентной формуле

$$x_{k+1} = (E - \tau A)x_k + \tau f \tag{49}$$

Будем считать, что матрица A : $A = A^* > 0$, тогда **интервал сходимости** в силу свойств самосопряжённого оператора равен

$$0 < \tau < \tau_o = \inf_{x \neq \theta} \frac{2(x, x)}{(Ax, x)} = \frac{2}{\sup_{x \neq 0} \frac{(Ax, x)}{(x, x)}} = \frac{2}{\|A\|} = \frac{2}{\lambda_{max}}. \tag{50}$$

Введем матрицу оператора перехода $S = E - \tau A,\ S = S^*$ и перепишем формулу в виде

$$x_{k+1} = Sx_k + \tau f. \tag{51}$$

Вместе с погрешностью $z_k = x_k - x$:

$$z_{k+1} = Sz_k. \tag{52}$$

Лемма 1

> Если оператор A имеет собственный вектор e_i с собственным значением λ_i, то оператор $S = E - \tau A$ также имеет собственный вектор e_i, но с собственным значением
>
> $$\mu_i(\tau) = 1 - \tau\lambda_i \tag{53}$$

$$Se_i = (E - \tau A)e_i = (1 - \tau\lambda_i)e_i = \mu_i(\tau)e_i. \quad \blacksquare \tag{54}$$

Спектральная норма остаётся зависимой от τ :

$$\|S(\tau)\| = \max_{1 \leq i \leq n} |\mu_i(\tau)| \tag{55}$$

Лемма 2

> Для того, чтобы метод простой итерации сходился к решению системы при любом выборе начального приближения, необходимо и достаточно, чтобы все собственные значения оператора перехода $S = E - \tau A$ были по модулю меньше единицы:
>
> $$|\mu_i(\tau)| < 1,\ i = \overline{1, n}. \tag{56}$$

Достаточность. Условие $|\mu_i(\tau)| < 1$ означает, что норма матрицы S будет меньше единицы, тогда

$$\|z_{k+1}\| \leq \|S\| \cdot \|z_k\| \leq \ldots \leq \underbrace{\|S\|^k}_{\to 0} \cdot \|z_0\| \tag{57}$$

Необходимость. Допустим, что среди собственных значений μ_i нашлось хотя бы одно $\mu_j : |\mu_j| \geq 1$. Выберем $x_0 = x + e_j$, где x – решение системы. Тогда $z_0 = e_j$, следовательно

$$z_k = S^k e_j = \mu_j^k e_j,\ \|z_k\| = |\mu_j|^k \to \infty \tag{58}$$

То есть последовательность z_k не сходится к нулю. Получили противоречие.

Оптимальное значение итерационного параметра.

В ходе доказательства леммы 2 было установлено, что погрешность $\|z_k\|$ убывает по закону геометрической прогрессии, в которой знаменатель зависит от τ

$$q(\tau) = \|S\| = \max_{1 \leq i \leq n} |\mu_i(\tau)| \tag{59}$$

Значит, чем меньше $\|S\|$, тем метод быстрее сходится. Минимальное собственное значение $\mu_1(\tau) = 1 - \tau\lambda_1$ меняет знак в точке $\tau_0/2$ с плюса на минус.

Найдем на отрезке $[\tau_0/2, \tau_0]$ точку $\tau^* : \mu_n(\tau^*) = -\mu_1(\tau^*)$

$$1 - \tau\lambda_n = \tau\lambda_1 - 1 \Rightarrow \tau^* = \frac{2}{\lambda_1 + \lambda_n} < \tau_0 \tag{60}$$

В результате получаем:

$$\|S\| = \begin{cases} \mu_n(\tau), & 0 < \tau \leqslant \tau_* \\ -\mu_1(\tau), & \tau_* \leqslant \tau < \tau_0 \end{cases} \tag{61}$$

Причём

$$\min_{0<\tau<\tau_0} \|S\| = 1 - \tau^*\lambda_n = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} = \frac{M_A - 1}{M_A + 1} \tag{62}$$

Выводы и замечания

- Метод простой итерации $x_{k+1} = (E - \tau A)x_k + \tau f$ сходится быстрее всего при

$$\tau = \frac{2}{\lambda_1 + \lambda_n} \tag{63}$$

- В случае плохо обусловленной матрицы, метод простой итерации сходится медленно, т.к. при $M_A \gg 1$

$$\min_{0<\tau<\tau_0} \|S\| = \frac{M_A - 1}{M_A + 1} \approx 1. \tag{64}$$

- Обычно τ^* неизвестно заранее, поскольку найти λ_1, λ_n это отдельная трудоёмкая задача.

Метод Зейделя.

Вернемся к общей записи итерационного стационарного процесса в канонической форме

$$B_{k+1}\frac{x_{k+1}-x_k}{\tau_{k+1}}+Ax_k=f,\ \det B_{k+1}\neq 0,\ \tau_{k+1}>0 \tag{65}$$

Рассмотрим произвольную квадратную матрицу A, разложим её на сумму трех матриц:

$$A=D+T_H+T_B \tag{66}$$

где D – диагональная часть матрицы A, T_H – нижняя треугольная матрица, T_B – верхняя треугольная матрица. Положим $B=D+T_H,\ \tau=1$, тогда

$$\begin{gathered}(D+T_H)(x_{k+1}-x_k)+Ax_k=f\\(D+T_H)x_{k+1}+T_Bx_k=f\end{gathered} \tag{67}$$

Перейдем от матричной формы записи к построчной:

$$\begin{cases}a_{11}x_1^{k+1} & + & a_{12}x_2^k & + & a_{13}x_3^k & + & \dots & + & a_{1n}x_n^k & = & f_1\\ a_{21}x_1^{k+1} & + & a_{22}x_2^{k+1} & + & a_{23}x_3^k & + & \dots & + & a_{2n}x_n^k & = & f_2\\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots\\ a_{n1}x_1^{k+1} & + & a_{n2}x_2^{k+1} & + & a_{n3}x_3^{k+1} & + & \dots & + & a_{nn}x_n^{k+1} & = & f_n\end{cases} \tag{68}$$

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

$$x_i^{k+1}=\frac{1}{a_{ii}}\left[f_i-\sum_{j=1}^{i-1}a_{ij}x_j^{k+1}-\sum_{j=i+1}^{n}a_{ij}x_j^k\right],\ i=\overline{1,n}. \tag{69}$$

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

Метод верхней релаксации

Введем параметр ω :

$$(D+\omega T_H)\frac{(x_{k+1}-x_k)}{\omega}+Ax_k=f \tag{70}$$

В данном случае $B=(D+\omega T_H),\ \tau=\omega>0$. Соотношению можно придать вид

$$\begin{gathered}\left(\frac{1}{\omega}D+T_H\right)(x_{k+1}-x_k)+Ax_k=f\\ \left(\frac{1}{\omega}D+T_H\right)x_{k+1}+\left[\left(1-\frac{1}{\omega}\right)D+T_B\right]x_k=f\end{gathered} \tag{71}$$

Отсюда следуют формулы

$$x_i^{k+1}=x_i^k+\frac{\omega}{a_{ii}}\left[f_i-\sum_{j=1}^{i-1}a_{ij}x_j^{k+1}-\sum_{j=i}^{n}a_{ij}x_j^k\right],\ i=\overline{1,n}. \tag{72}$$

Пусть $A:A^*=A>0$. Тогда $T_H^*=T_B$. Отсюда следует

$$(T_Hx,x)=(T_B^*x,x)=(T_Bx,x) \tag{73}$$

тогда, составив матрицу $B-\frac{\tau}{2}A$

$$B-\frac{\tau}{2}A=(D+\omega T_H)-\frac{\omega}{2}(D+T_H+T_B)=\left(1-\frac{\omega}{2}\right)D+\frac{\omega}{2}(T_H-T_B) \tag{74}$$

получим достаточное условие сходимости в виде

$$\left(\left[B - \frac{\tau}{2}A\right]x, x\right) = \left(1 - \frac{\omega}{2}\right)(Dx, x) > 0 \tag{75}$$

Матрица A является, по предположению, положительно определенной. Следовательно, все ее диагональные элементы строго положительны: $a_{ii} > 0,\ i = \overline{1, n}$. Это означает положительную определенность матрицы $D : (Dx, x) > 0$.

Значит, условие выполнено, если

$$0 < \omega < 2 \tag{76}$$

Случай диагонального преобладания (без док-ва)

Пусть матрица такова, что

$$\sum_{j \neq i} |a_{ij}| < q|a_{ii}|,\ |q| \leqslant 1,\ i = \overline{1, n}. \tag{77}$$

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

$$\|z_k\| \leqslant q^k \|z_0\|. \tag{78}$$

Приближение функций

Задача интерполирования

`дано:` функция $f : \{x_0, \ldots, x_n\} \to \{y_0, \ldots, y_n\}$ заданная в $n + 1$ точках, $n \in \mathbb{N}$.

`найти:` функцию F такую, что $F(x_i) = y_i,\ \ i = \overline{0, n}$.

Другими словами, нужно **найти промежуточные значения**. Функцию S называют **интерполирующей**, а точки (x_i, y_i) – **узлами интерполяции**.

Выберем некоторую систему функций $\phi_0(x), \phi_1(x), \ldots, \phi_n(x)$, заданных на отрезке $[a, b]$, и будем строить $F(x)$ как их линейную комбинацию:

$$F(x) = \sum_{i=0}^{n} c_i \phi_i(x), \quad F(x_i) = f(x_i), \quad i = \overline{0, n}. \tag{79}$$

Эти равенства представляют собой систему линейных алгебраических уравнений относительно коэффициентов c_j:

$$\sum_{i=0}^{n} c_i \phi_i(x_j) = f(x_j),\ j = \overline{0, n} \tag{80}$$

Если эта СЛАУ имеет невырожденную матрицу, то систему функций ϕ_i называют **Чебышевской**.

Необходимым условием принадлежности системы функций $\phi_i(x)\ (i = 0, 1, \ldots, n)$ к Чебышевской является их линейная независимость. Однако это условие не является достаточным.

Интерполирование полиномами

Широкое распространение получило интерполирование с помощью **степенных функций:**

$$\phi_k(x) = x^k. \tag{81}$$

В этом случае интерполирующая функция является полиномом степени n:

$$F(x) = P_n(x) = \sum_{k=0}^{n} c_k x^k, \quad \sum_{k=0}^{n} c_k x_i^k = f(x_i),\ i = \overline{0, 1}. \tag{82}$$

Определителем этой системы является *определитель Вандермонда*:

$$\Delta = \begin{vmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{vmatrix} = \prod_{i>j}(x_i - x_j) \tag{83}$$

В нашем случае этот определитель отличен от нуля, поскольку все узлы интерполирования различны между собой.

> Интерполирование с помощью полиномов при сделанных предположениях всегда осуществимо и притом единственным образом.

Интерполяционный многочлен в форме Лагранжа.

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

$$P_n(x) = \sum_{i=1}^{n} y_i \varphi_i(x), \;\; \varphi_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}. \tag{84}$$

Из выражения и формул 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 многочлен в форме Ньютона.

Перепишем интерполяционный полином Лагранжа в иной, эквивалентной форме

$$P_n(x) = P_0(x) + \sum_{k=1}^{n} (P_k(x) - P_{k-1}(x)), \tag{85}$$

где $P_k(x)$ - полиномы Лагранжа, построенные для узлов $x_i, \; i = \overline{0,k}$. Полином $Q_k = P_k(x) - P_{k-1}(x)$ имеет степень k и по построению обращается в ноль в точках $x_i, \; i = \overline{0,k-1}, \;$ поэтому его можно представить в виде

$$Q_k(x) = A_k(x - x_0)\dots(x - x_1)(x - x_{k-1}), \tag{86}$$

В силу $Q_k = P_k(x) - P_{k-1}(x)$ вклад в A_k вносит только $P_k(x)$, причём коэффициент A_k совпадает с коэффициентом при старшей степени. Тогда поскольку в разложении $P_k(x) = \sum_{i=1}^{k} f(x_i)\varphi_i(x)$ у функций $\varphi_i(x)$ коэффициент при старшей степени равен $1/\omega_i(x_k)$, то коэффициент при x^k в $P_k(x)$ будет равен

$$A_i = \sum_{k=0}^{i} \frac{f(x_k)}{\omega_i(x_k)}, \;\; \omega_i(x) = (x - x_0)\dots(x - x_{k-1})(x - x_{k+1})\dots(x - x_i) \tag{87}$$

При этом $A_0 = f(x_0)$. Значит, в силу $P_k(x) = Q_k(x) + P_{k-1}(x)$ формула позволяет написать рекуррентное соотношение для полинома $P_k(x)$:

$$P_k(x) = P_{k-1}(x) + A_k\omega_{k-1}(x) \tag{88}$$

Выражая аналогичным образом по индукции $P_{k-1}(x)$, получим окончательную формулу для $P_n(x)$:

$$\begin{gathered} P_n(x) = A_0 + A_1(x - x_0) + A_2(x - x_0)(x - x_1) + \dots + A_n(x - x_0)\dots(x - x_{n-1}) \\ P_n(x) = A_0 + \sum_{k=1}^{n} A_k\omega_{k-1}(x) \end{gathered} \tag{89}$$

Числа $A_k = f(x_0; \dots; x_k)$ называют **разделённой разностью**.

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

Погрешность интерполяционного полинома

Теорема о погрешности интерполяции полиномом

Поставим вопрос о том, насколько хорошо интерполяционный полином $P_n(x)$ приближает функцию $f(x)$ на отрезке $[a,b]$, то есть **попытаемся оценить остаточный член** при $x \neq x_i,\ x \in [a,b]$.

$$R_n(x) = f(x) - P_n(x) \tag{90}$$

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

Поскольку $R_n(x) = 0$ при $x \in \{x_0, \ldots, x_n\}$ справедливо представление:

$$R_n(x) = \omega_{n+1}(x) r_n(x) \tag{91}$$

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

$$g(t) = f(t) - P_n(t) - \omega_{n+1}(t) r_n(x), \tag{92}$$

заданную на отрезке $[a,b]$ и содержащую переменную x в качестве параметра. Заметим, что $g(t = x_i) = 0$ и $g(t = x) = 0$, т. е. как функция аргумента t она имеет $(n+2)$ нуля:

$$g(x_i) = 0,\ g(x) = 0. \tag{93}$$

Указанные нули функции принадлежат отрезку $[\alpha, \beta]$, где

$$\alpha = \min(a,x) \geqslant a, \quad \beta = \max(b,x) \leqslant b \tag{94}$$

Согласно теореме Ролля можно утверждать, что

- $g'(t)$ имеет по крайней мере $(n+1)$ нуль на отрезке $[\alpha, \beta]$. Эти нули перемежаются с нулями самой функции $g(t)$
- $g''(t)$ имеет по крайней мере n нулей на отрезке $[\alpha, \beta]$
- . . .
- $g^{(n+1)}(t)$ имеет хотя бы один нуль в некоторой точке $\xi \in [\alpha, \beta]$, то есть

$$g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - P_n^{(n+1)}(\xi) - (n+1)! \cdot r_n(x) = 0 \tag{95}$$

Значит, поскольку $\deg g(t) = \deg P_n(t) = n$, то $g(t) \equiv 0$. Тогда

$$r_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!},\ \xi \in [\alpha, \beta] \Longrightarrow R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega_{n+1}(x). \tag{96}$$

Оценка погрешности

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

$$|R_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} |\omega_{n+1}(x)|, \quad M_{n+1} = \max_{[\alpha,\beta]} |f^{(n+1)}(x)| \leqslant \max_{[a,b]} |f^{(n+1)}(x)| \tag{97}$$

Сходимость интерполяционного процесса.

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

Сетку на отрезке $[a,b]$ обозначим для краткости Ω_n. Рассмотрим последовательность сеток с возрастающим числом узлов:

$$\Omega_0 = \{x_0^{(0)}\}, \quad \Omega_1 = \{x_0^{(1)}, x_1^{(1)}\}, \quad \ldots, \quad \Omega_n = \{x_0^{(n)}, x_1^{(n)}, \ldots, x_n^{(n)}\} \tag{98}$$

и отвечающую ей последовательность интерполяционных полиномов $P_n(x)$, построенных для фиксированной $f(x) \in C[a,b]$.

> Напомним, что интерполяционный процесс
>
> - сходится *поточечно* на $[a,b]$, если существует предел
>
> $$\forall x \in [a,b] \Rightarrow \lim_{n\to\infty} P_n(x) = f(x). \tag{99}$$
>
> - сходится *равномерно* на $[a,b]$, если существует предел
>
> $$\lim_{n\to\infty} \sup_{[a,b]} |f(x) - P_n(x)| = 0 \tag{100}$$

Сходимость или расходимость интерполяционного процесса зависит как от выбора последовательности сеток, так и от гладкости функции $f(x)$.

Если $f(x)$ — **целая аналитическая функция**, то при произвольном расположении узлов на отрезке $[a,b]$ интерполяционный многочлен *сходится равномерно*.

Если $f(x)$ разрывна или не определена в некоторых точках, как например $f(x) = |x|$, то значения $P_n(x)$ между узлами *неограниченно растёт.*

Если $f(x)$ гладкая, то можно добиться равномерной сходимости *путём поиска подходящей сетки.* Но построение такой сетки очень сложно.

Интерполяционный многочлен в форме Эрмита

Определение

Пусть в узлах x_i, среди которых нет совпадающих, заданы значения функции $f(x)$ и её производных $f^{(i)}(x_k)$ до порядка $i = N_k - 1$ включительно.

Числа N_k при этом называют **кратностью** узла x_k. В каждой точке x_k задано N_k величин:

$$f(x_k),\ f'(x_k),\ \ldots,\ f^{(N_k-1)}(x_k). \tag{101}$$

В общей сложности для всей совокупности узлов $x_0, x_1, \ldots, x_m$ известно $n+1 = N_0 + N_1 + \ldots + N_m$ величин, что дает возможность ставить вопрос о построении полинома H_n степени n такого, что

$$H_n^{(i)}(x_k) = f^{(i)}(x_k), \quad k = \overline{0,m}, \quad i = \overline{0, N_k - 1}. \tag{102}$$

Такой полином **называется интерполяционным полиномом Эрмита** для функции $f(x)$.

Теорема о существовании и единственности полинома Эрмита

> **Интерполяционный полином Эрмита существует и единствен.**

Представим его в стандартном виде

$$H_n(x) = a_0 + a_1 x + \ldots + a_n x^n. \tag{103}$$

Задача свелась к поиску коэффициентов a_i. Условия задачи есть СЛАУ относительно этих коэффициентов, причем число уравнений и число неизвестных равны $N_0 + N_1 + \ldots + N_m = n + 1$

Рассмотрим сначала однородную систему $\overline{H_n}^{(i)}(x_k) = 0$. Её смысл в том, что числа x_k являются корнями полинома $\overline{H}_n(x)$ кратности N_k.

Значит, полином $\overline{H}_n(x)$ имеет, не менее $N_0 + N_1 + \ldots + N_m = n + 1$ корней. Но $\deg \overline{H}_n = n$, следовательно $\overline{H}_n \equiv 0$, то есть $\bar{a}_0 = \bar{a}_1 = \ldots = \bar{a}_n = 0$, т.е. однородная система уравнений имеет только тривиальное решение, а значит, *матрица системы невырождена*. А значит, полином Эрмита существует и единствен. ■

Погрешность

Исследование $R_n(x) = f(x) - H_n(x)$ почти дословно повторяет проведенное ранее исследование для полинома с простыми узлами x_k, в которых заданы только $f(x_k)$. Достаточно представить $R_n(x)$ в виде

$$\begin{gathered} R_n(x) = \omega_{n+1}(x) r_n(x), \\ \omega_{n+1}(x) = (x - x_0)^{N_0} (x - x_1)^{N_1} \ldots (x - x_m)^{N_m}, \quad n + 1 = N_0 + \ldots + N_m \end{gathered} \tag{104}$$

и рассмотреть функцию

$$g(t) = f(t) - H_n(t) - \omega_{n+1}(t) r_n(x) \tag{105}$$

Эта функция имеет $n + 2$ нуля:

$$g(x_k) = 0, \quad g(x) = 0. \tag{106}$$

Тогда по теореме Ролля для функции $g(t)$ делаем тяп-тяп и получаем

$$\begin{gathered} R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega_{n+1}(x) \Rightarrow \\ |R_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} |\omega_{n+1}(x)|, \quad M_{n+1} = \max_{[a,b]} |f^{(n+1)}(x)|. \end{gathered} \tag{107}$$

Интерполирование сплайнами

Определение сплайна

Назовем **кубическим сплайном функции** $y = f(x), x \in [a, b]$ на сетке $a = x_0 < x_1 < x_2 < \ldots < x_n = b$ функцию $S(x)$ удовлетворяющую условиям:

1. На каждом отрезке $[x_{i-1}, x_i]$ функция $S(x)$ является полиномом третьей степени.
2. $S(x), S'(x), S''(x) \in C[a, b]$.
3. $S(x_i) = f(x_i) = f_i, \ i = \overline{0, n}$.
4. На концах сегмента $[a, b]$ функция $S''(x)$ удовлетворяет условиям $S''(a) = S''(b) = 0$.

Теорема о существовании и единственности кубического сплайна

> **Существует единственный сплайн $S(x)$, удовлетворяющий требованиям 1–4.**

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

$$S_i(x) = a_i + b_i(x - x_i) + \frac{c_i}{2}(x - x_i)^2 + \frac{d_i}{6}(x - x_i)^3, \ x \in [x_{i-1}, x_i]. \tag{108}$$

Тогда $S_i(x_i) = f_i = a_i,\ S_i'(x_i) = b_i,\ S_i''(x_i) = c_i.$

Пусть $h_i = x_i - x_{i-1},\ i = \overline{1,n}.$

Расшифруем условия

Непрерывность сплайна в узлах, $i = \overline{1,n}$:

$$\begin{gathered} S_i(x_{i-1}) = S_{i-1}(x_{i-1}) \Leftrightarrow \\ f_i + b_i(x_{i-1} - x_i) + \frac{c_i}{2}(x_{i-1} - x_i)^2 + \frac{d_i}{6}(x_{i-1} - x_i)^3 = a_{i-1} \Leftrightarrow \\ a_i - b_i h_i + \frac{c_i}{2}h_i^2 - \frac{d_i}{6}h_i^3 = a_{i-1} \Leftrightarrow \\ b_i h_i - \frac{c_i}{2}h_i^2 + \frac{d_i}{6}h_i^3 = a_i - a_{i-1}. \end{gathered} \tag{109}$$

Дифференцируемость сплайна в узлах, $i = \overline{2,n}$:

$$\begin{gathered} S_i'(x_{i-1}) = S_{i-1}'(x_{i-1}), \Leftrightarrow \\ b_i + c_i(x_{i-1} - x_i) + \frac{d^2}{2}(x_{i-1} - x_i)^2 = b_{i-1} \Leftrightarrow \\ b_i - c_i h_i + \frac{d_i}{2}h_i^2 = b_{i-1}, \Leftrightarrow \\ c_i h_i - \frac{d_i}{2}h_i^2 = b_i - b_{i-1}. \end{gathered} \tag{110}$$

Двойная дифференцируемость сплайна в узлах, $i = \overline{2,n}$:

$$\begin{gathered} S_i''(x_{i-1}) = S_{i-1}''(x_{i-1}) \Leftrightarrow \\ c_i + d(x_{i-1} - x_i) = c_{i-1} \Leftrightarrow \\ c_i - d_i h_i = c_{i-1} \Leftrightarrow \\ d_i h_i = c_i - c_{i-1}. \end{gathered} \tag{111}$$

Граничные условия:

$$\begin{cases} S_1''(x_0) = c_1 + d_1(x_0 - x_1) = 0 \\ S_n''(x_n) = c_n + d_n(x_n - x_n) = 0 \end{cases} \Leftrightarrow \begin{cases} c_1 - d_1 h_1 = 0 \\ c_n = 0 \end{cases} \tag{112}$$

Мы получили $n + (n-1) + (n-1) + 2 = 3n$ уравнений с неизвестными $c_i,\ b_i,\ d_i,\ i = \overline{1,n}.$

Если формально ввести неизвестную $c_0 = 0$, то первое граничное условие можно отнести под условие двойной дифференцируемости для $i = 1$.

Сведение к СЛАУ

Из 111 возьмём

$$d_i = (c_i - c_{i-1})/h_i,\ i = \overline{1,n} \tag{113}$$

И подставим в 109:

$$\begin{gathered} b_i = \frac{a_i - a_{i-1}}{h_i} + \frac{c_i}{2}h_i - \frac{d_i}{6}h_i^2 = \frac{a_i - a_{i-1}}{h_i} + \frac{1}{3}c_i h_i + \frac{1}{6}c_{i-1}h_i \Rightarrow \\ b_i - b_{i-1} = \frac{a_i - a_{i-1}}{h_i} - \frac{a_{i-1} - a_{i-2}}{h_{i-1}} + \frac{1}{3}c_i h_i + \frac{1}{6}c_{i-1}h_i - \frac{1}{3}c_{i-1}h_{i-1} - \frac{1}{6}c_{i-2}h_{i-1} \end{gathered} \tag{114}$$

Подставим $b_i - b_{i-1}$ в 110:

$$\frac{1}{3}c_{i-2}h_{i-1} + \frac{2}{3}c_{i-1}(h_{i-1} + h_i) + \frac{1}{3}c_i h_i = 2\left(\frac{a_i - a_{i-1}}{h_i} - \frac{a_{i-1} - a_{i-2}}{h_{i-1}}\right),\ i = \overline{2,n} \text{ или} \tag{115}$$

$$h_i c_{i-1} + 2(h_i + h_{i+1})c_i + h_{i+1}c_{i+1} = 6\left(\frac{ƒ_{i+1} - ƒ_i}{h_{i+1}} - \frac{ƒ_i - ƒ_{i-1}}{h_i}\right),\ i = \overline{1, n-1}.$$

Выводы

- Наш результат – это $n-1$ линейных уравнений относительно неизвестных $c_i,\ i = \overline{1, n-1}$
- Для матрицы этой СЛАУ выполнено условие **диагонального преобладания**, т.е. решение существует и единственно
- Матрица СЛАУ является трёхдиагональной, решить эту систему можно **методом прогонки**, найдя все c_i и подставив их в d_i и b_i. **Теорема доказана. Решение найдено**.

Условие и скорость сходимости сплайна к точному решению

Сплайн сходится равномерно. Если $f(x) \in C[a, b]$, то

$$\forall x \in [a, b] : \forall \varepsilon\ \exists \delta : \max_i h_i < \delta \Rightarrow |f(x) - S(x)| < \varepsilon \tag{116}$$

Сплайн сходится быстро :) Если $f(x) \in C^{(4)}[a, b]$ и $f''(a) = f''(b) = 0$, то $\forall x \in [a, b] \Rightarrow$

$$\begin{gathered}|f(x) - S(x)| \leqslant Mh^4, \quad |f'(x) - S'(x)| \leqslant Mh^3, \quad |f''(x) - S''(x)| \leqslant Mh^2, \\ M = \max_{[a,b]} |f^{(4)}(x)|.\end{gathered} \tag{117}$$

Метод наименьших квадратов

Постановка задачи

В методе наименьших квадратов **аппроксимирующая** функция $F(x)$ ищется в виде линейной комбинации:

$$F(x) = \sum_{k=0}^{m} a_k \phi_k(x) \tag{118}$$

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

$$\delta_i = y_i - F(x_i),\ i = \overline{0, n}. \tag{119}$$

Сумма квадратов этих величин называется **суммарной квадратичной погрешностью**

$$J = \sum_{i=0}^{n} \delta_i^2 \tag{120}$$

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

Мы можем менять значение $J(a_0,\ \ldots,\ a_m) = J(a)$. Вектор $\arg\min_a J$ называют **наилучшим приближением по методу наименьших квадратов**.

Построение наилучшего приближения по методу наименьших квадратов

Сведение к СЛАУ

Построение наилучшего приближения сводится к классической задаче математического анализа об экстремуме функции нескольких переменных.

Необходимым условием экстремума является равенство нулю в экстремальном точке всех первых частных производных рассматриваемой функции.

В данном случае это дает

$$\frac{\partial J}{\partial a_l} = -2\sum_{i=0}^{n}\left(y_i - \sum_{k=0}^{m} a_k\phi_k(x_i)\right)\phi_l(x_i) = 0,\ l = \overline{0,m}. \tag{121}$$

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

$$\sum_{k=0}^{m}\gamma_{lk}a_k = b_l,\ \text{где}\ \ \gamma_{lk} = \sum_{i=0}^{n}\phi_l(x_i)\phi_k(x_i),\quad b_l = \sum_{i=0}^{n}\phi_l(x_i)y_i,\quad l = \overline{0,m}. \tag{122}$$

Мы получили СЛАУ относительно $a_0, a_1, \ldots, a_m$. Число уравнений и число неизвестных в этой системе равно $m+1$.

Матрица Грама

Матрицу системы $\Gamma = (\gamma_{lk})$ называют **матрицей Грама для системы функций** $\phi_0(x), \phi_1(x), \ldots, \phi_m(x)$ **на сетке** $x_0, x_1, \ldots, x_n$. Заметим, что $\gamma_{lk} = \gamma_{kl}$.

Предположим, что функции $\phi_0(x), \phi_1(x), \ldots, \phi_m(x)$ выбраны такими, что

$$\Delta = \det\Gamma \neq 0 \tag{123}$$

В этом случае при $\forall b = (b_0,\ \ldots, b_m)^T$ решение $a = (a_0,\ \ldots,\ a_m)$ будет единственным.

Исследование решения

Рассмотрим произвольный вектор $\overline{a} = a + \delta,\ \|\delta\| \neq 0$. Сравним $J(a)$ и $J(\overline{a})$.

Квадрат погрешности и точке $x = x_i$ для функции $F(x)$ с коэффициентами $a_0, a_1, \ldots, a_m$ можно записать в виде

$$\begin{aligned}\delta_i^2 &= \left(y_i - \sum_{k=0}^{m}(a_k+\delta_k)\phi_k(x_i)\right)^2 = \left(y_i - \sum_{k=0}^{m}a_k\phi_k(x_i) - \sum_{k=0}^{m}\delta_k\phi_k(x_i)\right)^2 = \\ &= \left(y_i - \sum_{k=0}^{m}a_k\phi_k(x_i)\right)^2 - 2\left(y_i - \sum_{k=0}^{m}a_k\phi_k(x_i)\right)\sum_{l=0}^{m}\delta_l\phi_l(x_i) + \left(\sum_{k=0}^{m}\delta_k\phi_k(x_i)\right)^2\end{aligned} \tag{124}$$

Поскольку $J(\overline{a}) = \sum_{i=0}^{n}\delta_i^2$, то

- первое слагаемое сворачивается в $J(a)$
- второе слагаемое обращается в ноль в силу 121.

Тогда $J(\overline{a})$ строго больше $J(a)$. Это доказывает, что чтобы построить наилучшее приближение сеточной функции по методу наименьших квадратов, нужно взять в качестве коэффициентов разложения a_k решение системы линейных уравнений 122.

Численное интегрирование

Квадратурные формулы прямоугольников, трапеций и парабол

Постановка задачи численного интегрирования

Квадратурные формулы – это универсальные алгоритмы вычисления определенных интегралов. Они имеют следующий вид:

$$I = \int_a^b d(x)dx = \sum_{i=1}^{n} c_i f(x_i) + R_n \tag{125}$$

$i=1$

Здесь точки $x_i \in [a, b]$ называют **узлами**, коэффициенты c_i — **весовыми множителями**, величину R_n —погрешностью**.** Узлы и веса подбираются таким образом, чтобы выполнялось **условие сходимости**:

$$\lim_{n\to\infty} R_n = 0. \tag{126}$$

Таким образом, открывается возможность вычислить интеграл I с любой наперед заданной точностью по значениям функции $f(x)$, взятым в разных точках x_i отрезка $[a, b]$.

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

$$(b - a) = \sum_{i=1}^{n} c_i, \tag{127}$$

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

Квадратурная формула прямоугольников.

Определение

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

$$x_i = a + ih,\ i = \overline{0, n}. \tag{128}$$

Для дальнейшего нам также понадобятся средние точки этих отрезков

$$\xi_i = a + (i - 1/2)h, \quad \xi_i \in [x_{i-1}, x_i], \quad i = \overline{1, n}. \tag{129}$$

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:

$$P_n = \frac{b - a}{n} \sum_{i=0}^{n} f(\xi_i). \tag{130}$$

В квадратурной формуле **узлами** будут точки ξ_i, а **весовыми множителями** $h = (b - a)/n$. Такую квадратурную формулу называют **формулой прямоугольников**:

$$I = P_n + \alpha_n \tag{131}$$

Причина такого названия имеет простой геометрический смысл.

Сведение к интегральной сумме

Формула для величины P_n *изначально* строилась как интегральная сумма, поэтому метод прямоугольников сходится для любой интегрируемой функции. То есть любой интеграл можно вычислить с любой наперёд заданной точностью ε.

Квадратурная формула трапеций.

Определение

В качестве аппроксимирующей функции $g_n(x)$ берется кусочно-линейная функция,
$x \in [x_{i-1}, x_i],\ i = \overline{1, n}$

$$g_i(x) = f(x_{i-1})\frac{x - x_i}{x_{i-1} - x_i} + f(x_i)\frac{x - x_{i-1}}{x_i - x_{i-1}}$$

$$\text{подставим } h = x_i - x_{i-1}$$

$$g_i(x) = \frac{f(x_i) - f(x_{i-1})}{\ }x + +\frac{f(x_{i-1})}{\ }x_i - \frac{f(x_i)}{\ }x_{i-1}$$

$$h \qquad \underbrace{h \qquad h}_{q(x)}$$

$$\begin{gathered}\text{подставим в } q(x) \text{ выражение } x_i = x_{i-1} + h \\ q(x) = \frac{f(x_{i-1})}{h}(x_{i-1} + h) - \frac{f(x_i)}{h}x_{i-1} \\ q(x) = \frac{f(x_{i-1}) - f(x_i)}{h}x_i + f(x_{i-1}) \\ \text{тогда } g_i(x) \\ g_i(x) = f(x_{i-1}) + \frac{f(x_i) - f(x_{i-1})}{h}(x - x_{i-1}).\end{gathered} \tag{132}$$

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

Вычислим интеграл:

$$\begin{gathered}\int_{x_{i-1}}^{x_i} g_i(x)dx = \int_{x_{i-1}}^{x_i} \left[f(x_{i-1}) + \frac{f(x_i) - f(x_{i-1})}{h}(x - x_{i-1})\right]dx = \\ \text{пусть } z = x - x_{i-1} \\ = \int_0^{x_i - x_{i-1}} \left[f(x_{i-1}) + \frac{f(x_i) - f(x_{i-1})}{h}z\right]dz = \\ = f(x_{i-1})(x_i - x_{i-1}) + \frac{f(x_i) - f(x_{i-1})}{2h}(x_i - x_{i-1})^2 = \\ \text{вспомним, что } h = x_i - x_{i-1} \\ = \frac{f(x_{i-1}) + f(x_i)}{2}h\end{gathered} \tag{133}$$

Значит, интеграл от $g_n(x)$ по всему отрезку $[a, b]$ будет равен

$$\begin{gathered}T_n = \int_a^b g_n(x)dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} g_n(x)dx = \\ = \frac{b-a}{n}\left[\frac{1}{2}f(a) + f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + \frac{1}{2}f(b)\right]\end{gathered} \tag{134}$$

Тогда кадратурной формулой трапеций будет

$$I = \int_a^b f(x)dx = T_n + \beta_n \tag{135}$$

Сведение к интегральной сумме

При выводе формулы для величины T_n понятие интегральной суммы не использовалось. Однако видно, что величину T_nтоже можно интерпретировать как интегральную сумму. Чтобы убедиться в этом, рассмотрим разбиение отрезка $[a, b]$ на частичные отрезки точками ξ_i. Оно дает $n + 1$ отрезок. Два крайних $[a, \xi_1]$ и $[\xi_n, b]$ имеют длину $h/2$, а остальные — длину h. Выберем для образования интегральной суммы на крайних отрезках значения функции $f(x)$ в точках a и b, а на остальных отрезках $[\xi_i. \xi_{i+1}]$ — значения функции $f(x)$ в точках x_i $(i = \overline{1, n-1})$. Образованная таким образом интегральная сумма соответствует выражению для T_n.

Из этого следует, что выполнена сходимость для любой интегрируемой функции.

Оценка погрешности

Матан 2 семестр “Приближённые методы вычислений”.

Квадратурные формулы парабол

Определение

Будем аппроксимировать функцию $f(x)$ функцией $g_j(x)$ на частичном сегменте $[x_{2j-2}, x_{2j}],\ j = \overline{1, n/2}$

$$g_j(x) = f(x_{2j-2})\frac{(x-x_{2j-1})(x-x_{2j})}{2h^2} + f(x_{2j-1})\frac{(x-x_{2j-2})(x-x_{2j})}{-(h^2)} + \\ + f(x_{2j})\frac{(x-x_{2j-2})(x-x_{2j-1})}{2h^2} \tag{136}$$

Проинтегрируем и получим

$$\int_{x_{2j-2}}^{x_{2j}} g_j(x)dx = \frac{h}{3}[f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})], \quad h = \frac{b-a}{n} \tag{137}$$

Тогда интеграл по всему отрезку равен

$$S_n = \int_a^b g_n(x)dx = \sum_{j=1}^{n/2}\int_{x_{2j-2}}^{x_{2j}} g_n(x)dx = \\ = \frac{b-a}{3n}\{f(a) + 4f(x_1) + 2f(x_2) + \ldots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(b)\} \tag{138}$$

Квадратурной формулой Симпсона называют

$$I = \int_a^b f(x)dx = S_n + \gamma_n \tag{139}$$

Сведение к интегральной сумме

Представление для S_n как и представление для T_n, также можно рассматривать как интегральную сумму. Для ее построения нужно разбить отрезок $[a, b]$ на $(n+1)$ частичный отрезок с помощью точек

$$\eta_{2j-1} = x_{2j-1} - \frac{2h}{3}, \quad \eta_{2j} = x_{2j} + \frac{2h}{3}, \quad j = \overline{1, \frac{n}{2}} \\ \eta_0 = a, \quad \eta_{n+1} = b. \tag{140}$$

В результате получаются отрезки $[\eta_{i-1}, \eta_i],\ 1 \le i \le n+1$ различной длины. Два крайних отрезка $[a, \eta_1]$ и $[\eta_n, b]$ имеют длину $h/3$. Отрезки, в центре которых лежат точки x_i с четными номерами, — длину $2h/3$, отрезки, в центре которых лежат точки x_i с нечетными номерами, — длину $4h/3$.

Из этого следует сходимость метода Симпсона для любой интегрируемой функции.

Связь с формулой парабол

Заканчивая обсуждение, установим связь между величинами T_n и S_n.

$$S_n = \frac{4}{3}T_n - \frac{1}{3}T_{n/2} \tag{141}$$

Здесь $T_{n/2}$ — сумма с вдвое меньшим числом слагаемых и, соответственно, с вдвое большим шагом. Благодаря этому при ее образовании в качестве узлов используются точки x_i только с четными номерами. Поскольку в формуле Симпсона n предполагается обязательно четным, то $n/2$ - целое число, так что выражение T_n определено. Соотношение проверяется «в лоб».

Квадратурная формула Гаусса

Постановка задачи

Предыдущие квадратурные формулы строились как интеграл от апроксимации, т.е. задача сводилась к выбору узлов и весовых коэффициентов. Естественно возникает задача поиска *наилучшей* квадратурной формулы с заданным числом узлов n.

В формулировке Гаусса: построить квадратурную формулу с числом узлов n, которая является точной для любого полинома степени $(2n-1)$ или ниже:

$$\int_{-1}^{1} f(x)dx = \sum_{i=1}^{n} c_i f(x_i) + \delta_n, \tag{142}$$

Для любого полинома степени $(2n-1)$ остаточный член в формуле должен быть равен нулю, т.е.

$$\int_{-1}^{1} x^m dx = \frac{1+(-1)^m}{m+1} = \sum_{i=1}^{n} c_i x_i^m, \quad m = \overline{0, 2n-1} \tag{143}$$

Получили систему из $2n$ уравнений относительно x_i и c_i, $i = \overline{1,n}$. Уравнение при $m=0$ даёт соотношение

$$\sum_{i=1}^{n} c_i = 2 \tag{144}$$

Свойства полиномов Лежандра

Они определяются формулами

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n \tag{145}$$

> 1°. Полином Лежандра $P_n(x)$ является полиномом n-ой степени, обладающим той же четностью, что и n:
>
> $$P_n(-x) = (-1)^n P_n(x) \tag{146}$$

Напрямую следует из формулы. ■

> 2°. Полиномы Лежандра $P_n(x)$ в точках $x = \pm 1$ принимают следующие значения:
>
> $$P_n(1) = 1, \quad P_n(-1) = (-1)^n \tag{147}$$

Представим выражение $(x^2-1)^n$ в виде произведения

$$(x^2-1)^n = (x+1)^n (x-1)^n \tag{148}$$

и выполним n-кратное дифференцирование. В результате по формуле Лейбница получим:

$$\begin{gathered} \frac{d^{n-k}}{dx^{n-k}}(x+1)^n = \frac{n!}{k!}(x+1)^k, \quad \frac{d^k}{dx^k}(x-1)^n = \frac{n!}{(n-k)!}(x-1)^{n-k} \Rightarrow \\ P_n(x) = \frac{1}{2^n n!}\sum_{k=0}^{n} C_n^k \cdot C_n^k \cdot n! \cdot (x+1)^{n-k} \cdot (x-1)^k \end{gathered} \tag{149}$$

При $x=1$ все члены этой суммы обращаются в ноль, кроме члена $k=0$, который равен 1, значит, $P_n(1)=1$. Равенство $P_n(-1) = (-1)^n$ следует из свойства 1°, согласно которому $P_n(-x) = (-1)^n P_n(x)$.

> 3°. Полином Лежандра $P_n(x)$ имеет на интервале $(-1,1)$ ровно n простых корней. В силу свойства 1° корни располагаются симметрично относительно $x=0$.

Функция $(x^2-1)^n$ имеет нули в точках $x=\pm 1$. Значит, её производная тоже. По теореме Ролля, её производная имеет по крайней мере 1 ноль на интервале $(-1,1)$. Значит, вторая производная имеет по крайнеё мере 2 нуля на интервале $(-1,1)$. И так далее.

> 4°. Любой полином $Q_m(x)$ степени $m<n$ ортогонален к полиному Лежандра $P_n(x)$ на сегменте $[-1,1]$:

$$\int_{-1}^{1} Q_m(x)P_n(x)dx=0. \tag{150}$$

Подставим в интеграл представление полинома Лежандра и проинтегрируем по частям. В результате получим:

$$J=\frac{1}{2^n n!}\int_{-1}^{1} Q_m(x)\frac{d^n}{dx^n}(x^2-1)^n dx=$$
$$=\frac{1}{2^n n!}\left\{Q_m(x)\underbrace{\frac{d^{n-1}}{dx^{n-1}}(x^2-1)^n\Big|_{-1}^{1}}_{=0}-\int_{-1}^{1}\frac{dQ_m(x)}{dx}\frac{d^{n-1}}{dx^{n-1}}(x^2-1)^n dx\right\}. \tag{151}$$

Выполняя процедуру интегрирования по частям $m+1\le n$ раз, получим:

$$J=(-1)^{m+1}\frac{1}{2^n n!}\int_{-1}^{1}\frac{d^{m+1}Q_m(x)}{dx^{m+1}}\frac{d^{n-m-1}}{dx^{n-m-1}}(x^2-1)^n dx=0. \tag{152}$$

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

> **Следствие 1. Полиномы Лежандра образуют систему полиномов, ортогональных на отрезке**
>
> $$\int_{-1}^{1} P_m(x)P_n(x)dx=0,\quad m\ne n. \tag{153}$$

Из линейной алгебры известно, что система полиномов, ортогональных на некотором множестве, определена однозначно с точностью до множителей. Поэтому следствию 1 можно сопоставить обратное.

> **Следствие 2. Любая система полиномов, ортогональных на отрезке $[-1,1]$, совпадает с точностью до множителя с системой полиномов Лежандра.**

Узлы и коэффициенты квадратуры Гаусса.

Теперь мы можем найти узлы $x_i,\ i=\overline{1,n}$. Составим полином n-ой степени

$$\omega_n(x)=(x-x_1)(x-x_2)\dots(x-x_n), \tag{154}$$

Проинтегрируем $Q_m(x)\omega_n(x)$ по отрезку $[-1,1]$ методом Гаусса, где $Q_m(x)$ – произвольный полином степени $m<n$. Поскольку $m+n\le 2n-1$, формула Гаусса должна быть точной. Таким образом, имеем

$$\int_{-1}^{1} Q_m(x)\omega_n(x)dx=\sum_{i=1}^{n} c_i Q_m(x_i)\omega_n(x_i)=0. \tag{155}$$

Мы видим, что $\omega_n(x)$ ортогонален к любому полиному степени $m<n$. Это означает, что он с точностью до множителя совпадает с n-ым полиномом Лежандра:

$$\omega_n(x)=A_n P_n(x). \tag{156}$$

> Вывод: *узлы квадратурной формулы Гаусса являются корнями полинома Лежандра*.

Для того, чтобы подсчитать весовые коэффициенты, введем используем базисные полиномы Лагранжа

$$\varphi_{n-1,k}(x) = \frac{(x-x_1)\dots(x-x_{k-1})(x-x_{k+1})\dots(x-x_n)}{(x_k-x_1)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_n)},\ \varphi_{n-1,k}(x_i) = \delta_{ik}. \quad (157)$$

Для $\varphi_{n-1,m}(x)$ квадратурная формула Гаусса должна быть точной, т.к. $\deg \varphi_{n-1,k} = n-1$, значит,

$$\int_{-1}^{1} \varphi_{n-1,m}(x)dx = \sum_{i=1}^{n} c_i \varphi_{n-1,m}(x_i) = c_m \quad (158)$$

Доказательство точности для полинома степени $\leqslant 2n-1$

Первый пункт

Сначала докажем, что такая формула является точной для любого полинома $Q_{n-1}(x)$ степени $\leqslant n-1$. Такой полином полностью совпадает со своей интерполяцией методом Лагранжа, потому что полиномы степени $n-1$, совпадающие в $n-1$ точках, тождественно равны:

$$Q_{n-1}(x) = \sum_{m=1}^{n} Q_{n-1}(x_m)\varphi_{n-1,m}(x). \quad (159)$$

Интегрируя равенство по отрезку $[-1,1]$, получим доказательство квадратурной формулы Гаусса:

$$\int_{-1}^{1} Q_{n-1}(x)dx = \sum_{i=1}^{n} Q_{n-1}(x_i) \int_{-1}^{1} \varphi_{n-1,i}(x)dx = \sum_{i=1}^{n} c_i Q_{n-1}(x_i) \quad (160)$$

Второй пункт

Теперь рассмотрим произвольный полином $Q_{2n-1}(x)$ степени $2n-1$. Разделим его с остатком на полином Лежандра $P_n(x)$:

$$Q_{2n-1}(x) = P_n(x)q_{n-1}(x) + r_{n-1}(x). \quad (161)$$

где $q_{n-1}(x)$ и $r_{n-1}(x)$ – степени $(n-1)$. В силу ортогональности полиномов Лежандра, получим:

$$\begin{aligned} \int_{-1}^{1} Q_{2n-1}(x)dx &= \int_{-1}^{1} [P_n(x)q_{n-1}(x) + r_{n-1}(x)]dx = \{\text{ортогональность}\} = \\ &= \int_{-1}^{1} r_{n-1}(x) = \{\text{Гаусс}\} = \sum_{i=1}^{n} c_i r_{n-1}(x_i) = \{\text{узлы это корни } P_n(x)\} = \\ &= \sum_{i=1}^{n} c_i \{P_n(x_i)q_{n-1}(x_i) + r_{n-1}(x_i)\} = \sum_{i=1}^{n} c_i Q_{2n-1}(x_i) \end{aligned} \quad (162)$$

Итак, построенная квадратурная формула действительно является точной для любого полинома степени $(2n-1)$, значит, *задача Гаусса решена*.

Численное решение дифференциальных уравнений

Разностная аппроксимация производных

Постановка задачи

Наиболее универсальными методами численного решения обыкновенных дифференциальных уравнений являются *разностные методы*. Они основаны на замене производных в дифференциальном уравнении разностными отношениями. В результате исходное дифференциальное уравнение сводится к системе алгебраических уравнений, которые называются *разностными*. Решение этой системы дает приближенное решение исходной задачи.

Сетка, сеточные функции

Будем рассматривать р/м сетку на $[a, b]$ из точек $x_i,\ i = \overline{0, n}$.

Функцию $y : x_i \mapsto y_i,\ i = \overline{0, n}$ называют **сеточной**. Сеточные функции, определённые на одной сетке, образуют $(n + 1)$-мерное линейное пространство.

Будем использовать следующую норму:

$$\|f\| = \max_i |y_i|. \tag{163}$$

Пусть дано дифференциальное уравнение $Lu(x) = f(x, u)$, например, $u' = f(x, u)$

Пусть u_h – проекция непрерывной функции $u(x)$ на сетку:

$$u_h = u(x_i) = u_i. \tag{164}$$

Оператор L можно **аппроксимировать** 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**.

Вместе аппроксимация оператора L и функции f называется **разностной схемой**. Пример:

$$\frac{y_{i+1} - y_i}{h} = f(x_i, y_i), \quad i = \overline{1, n - 1}. \tag{165}$$

Погрешностью аппроксимации производной назовем величину

$$\psi_1 = (Lu)_h - L_h u_h. \tag{166}$$

Погрешностью аппроксимации правой части f сеточной функции ϕ_h назовем величину

$$\psi_2 = f_h - \varphi_h. \tag{167}$$

где f_h — проекция на сетку функции $f(x, u)$, например, $f(x_i, u_i)$.

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

$$\psi = \psi_1 - \psi_2 = (Lu)_h - L_h u_h - (f_h - \varphi_h) = \{Lu = f\} = \varphi_h - L_h u_h. \tag{168}$$

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.

Говорят, что погрешность аппроксимации разностной схемы **имеет порядок** m на сетке, если

$$\|\psi\| = O(h^m) \tag{169}$$

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

Аппроксимация первой производной

Односторонние производные

В качестве аппроксимации левой, правой и общей производной будем использовать следующие операторы:

$$\begin{aligned} L_h^+[y_i] &= \frac{y_{i+1} - y_i}{h}, \quad i = \overline{0, n-1} \\ L_h^-[y_i] &= \frac{y_i - y_{i-1}}{h}, \quad i = \overline{1, n} \\ L_h[y_i] &= \frac{y_{i+1} - y_{i-1}}{2h}, \quad i = \overline{1, n-1} \end{aligned} \tag{170}$$

Отношение $L_h^+[y_i]$ называют правой разностной производной, отношение $L_h^-[y_i]$ – левой разностной производной и отношение $L_h^{(0)}[y_i]$ – центральной разностной производной.

Введём погрешности аппроксимации производных:

$$\begin{aligned} \psi_i^+ &= L_h^+[y_i] - y'(x_i), \quad i = \overline{0, n-1} \\ \psi_i^- &= L_h^-[y_i] - y'(x_i), \quad i = \overline{1, n} \\ \psi_i^{(0)} &= L_h^{(0)}[y_i] - y'(x_i), \quad i = \overline{1, n-1} \end{aligned} \tag{171}$$

Предположим, что функция $y(x)$ дважды непрерывно дифференцируема на $[a, b]$. Используя формулу Тейлора в форме Лагранжа, получим

$$\begin{gathered} y_{i+1} = y(x_i + h) = y_i + y'(x_i)h + \frac{1}{2}y''(x_i + \theta_i h)h^2 \Rightarrow \\ \Rightarrow \psi_i^+ = y''(x_i + \theta_i h)\frac{h}{2}, \quad \psi_i^- = -y''(x_i - \theta_i h)\frac{h}{2} \end{gathered} \tag{172}$$

Производная $y''(x)$, по предположению, непрерывна на отрезке $[a, b]$, и, следовательно, ограничена, тогда

$$|\psi^+| \leqslant \frac{1}{2}M_2 h, \quad |\psi^-| \leqslant \frac{1}{2}M_2 h, \quad M_2 = \sup_{[a,b]} |y''(x)| \tag{173}$$

То есть введённая аппроксимация имеет *первый* порядок точности относительно h.

Центральная производная

Для оценки $\psi_i^{(0)}$ предположим, что функция $y(x)$ три раза непрерывно дифференцируема на отрезке $[a, b]$. Используя формулу Тейлора в форме Лагранжа, получим

$$\begin{gathered} y_{i+1} = y(x_i + h) = y_i + y'(x_i)h + \frac{1}{2}y''(x_i)h^2 + \frac{1}{6}y'''(x_i + \theta_{1,i}h)h^3 \\ y_{i-1} = y(x_i - h) = y_i - y'(x_i)h + \frac{1}{2}y''(x_i)h^2 - \frac{1}{6}y'''(x_i - \theta_{2,i}h)h^3 \\ \Rightarrow \psi_i^{(0)} = \frac{h^2}{12}\left[y'''(x_i + \theta_{1,i}h) + y'''(x_i - \theta_{2,i}h)\right] \Rightarrow \\ |\psi^{(0)}| \leqslant \frac{1}{6}M_3 h^2, \quad M_3 = \sup_{[a,b]} |y'''(x)| \end{gathered} \tag{174}$$

То есть введённая аппроксимация имеет *второй* порядок точности относительно h.

Аппроксимация второй производной

Составим разностное отношение первых разностных производных:

$$L_h[y] = \frac{\dfrac{y_{i+1} - y_i}{h} - \dfrac{y_i - y_{i-1}}{h}}{h} = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} \tag{175}$$

Введём погрешности аппроксимации производной:

$$\psi_i = L_h[y_i] - y''(x_i), \quad i = \overline{1, n-1} \tag{176}$$

Для оценки ψ_i предположим, что функция $y(x)$ четыре раза непрерывно дифференцируема на $[a, b]$. Используя формулу Тейлора в форме Лагранжа, получим

$$\begin{gathered} y_{i+1} = y(x_i + h) = y_i + y'(x_i)h + \frac{1}{2}y''(x_i)h^2 + \frac{1}{6}y'''(x_i)h^3 + \frac{1}{24}y^{(4)}(x_i + \theta_{1,i}h)h^4 \\ y_{i-1} = y(x_i + h) = y_i - y'(x_i)h + \frac{1}{2}y''(x_i)h^2 - \frac{1}{6}y'''(x_i)h^3 + \frac{1}{24}y^{(4)}(x_i - \theta_{2,i}h)h^4 \\ \Rightarrow \psi_i = \frac{h^2}{24}\left[y^{(4)}(x_i + \theta_{1,i}h) + y^{(4)}(x_i - \theta_{2,i}h)\right] \Rightarrow \\ |\psi| \leq \frac{1}{12}M_4 h^2, \quad M_4 = \sup_{[a,b]} |y^{(4)}(x)| \end{gathered} \tag{177}$$

То есть введённая аппроксимация имеет *второй* порядок точности относительно h.

Численное решение задачи Коши

При численном интегрировании диф. уравнений производные меняют на разностные отношения. Результат – некая сеточная функция y_i. Естественный вопрос: как сильно она отличается от прямого решения y?

1. По узлам y_i можно построить интерполирующую функцию u и вычислить следующую норму:

$$\|z\| = \max_{[a,b]} |u(x) - y(x)| \tag{178}$$

2. Функцию y можно спроецировать в сеточную функцию u_i и посчитать следующую норму:

$$\|z\| = \max_i |u_i - y_i| \tag{179}$$

Первый способ точнее, а второй проще. Обычно выбирают второй.

Метод Эйлера

Рассмотрим следующую задачу Коши:

$$u' = f(x, u), \quad u(x_0) = u_0. \tag{180}$$

Если функция $f(x, u)$ непрерывна и удовлетворяет условию Липшица по аргументу u в некоторой окрестности начальной точки (x_0, u_0), то можно указать такой отрезок $[a, b],\ a < x_0 < b$, на котором решение задачи $u(x)$ существует и является единственным.

Пусть нам нужно построить решение задачи на отрезке $[x_0, x_0 + l]$. Образуем сетку и сопоставим разностную задачу:

$$\begin{gathered} x_i = x_0 + ih, \quad h = \frac{l}{n}, \qquad i = \overline{0, n} \\ \frac{y_{i+1} - y_i}{h} = f(x_i, y_i), \quad y_0 = u_0, \quad i = \overline{0, n-1}. \end{gathered} \tag{181}$$

Уравнение является разностным уравнением первого порядка, которое принято называть **схемой Эйлера**. В рекуррентном виде:

$$y_{i+1} = y_i + hf(x_i, y_i), \quad i = \overline{0, n-1}. \tag{182}$$

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

Доказательство сходимости

Рассмотрим решение задачи в точках сетки, образуя из функции непрерывного аргумента сеточную функцию $\{u_i = u(x_i)\}$, и сравним ее с рассчитанной сеточной функцией $\{y_i\}$. Для этого **образуем две сеточные функции** z, ψ

$$\begin{aligned} z_i &= y_i - u_i, & i &= \overline{0, n} \\ \psi_i &= \frac{u_{i+1} - u_i}{h} - f(x_i, u_i), & i &= \overline{0, n-1} \end{aligned} \tag{183}$$

Функцию z называют **погрешностью решения**. Функцию ψ называют **погрешностью аппроксимации уравнения.**

Установим связь между z и ψ. Подставим $y_i = u_i + z_i$ в разностное уравнение:

$$\begin{aligned} \frac{z_{i+1} - z_i}{h} + \frac{u_{i+1} - u_i}{h} &= f(x_i, u_i + z_i) \\ \frac{z_{i+1} - z_i}{h} &= f(x_i, u_i + z_i) - f(x_i, u_i) - \left[\frac{u_{i+1} - u_i}{h} - f(x_i, u_i)\right] \\ \frac{z_{i+1} - z_i}{h} &= f(x_i, u_i + z_i) - f(x_i, u_i) - \psi_i \end{aligned} \tag{184}$$

По теореме Лагранжа:

$$f(x_i, u_i + z_i) - f(x_i, u_i) = \frac{d}{du} f(x_i, u_i + \theta_i z_i) z_i. \tag{185}$$

Так что

$$\begin{aligned} \frac{z_{i+1} - z_i}{h} &= \frac{d}{du} f(x_i, u_i + \theta_i z_i) z_i - \psi_i \\ z_{i+1} &= z_i \left[1 + \frac{d}{du} f(x_i, u_i + \theta_i z_i) \cdot h\right] - \psi_i h, \quad z_0 = 0, \quad i = \overline{0, n-1} \end{aligned} \tag{186}$$

Необходимым условием существования и единственности решения задачи Коши является ограниченность f'_u в интересующей нас области изменения ее аргументов:

$$\left|\frac{d}{du} f(x, u)\right| \leqslant C \tag{187}$$

Это позволяет написать оценку

$$\left|1 + \frac{df}{du}(x_i, u_i + \theta_i z_i) \cdot h\right| \leqslant 1 + Ch < e^{Ch} = q \Rightarrow |z_{i+1}| \leqslant q|z_i| + \|\psi\| h \tag{188}$$

Тогда

$$\begin{aligned} z_0 &= 0 \\ |z_1| &\leqslant \|\psi\| h \\ |z_1| &\leqslant (1+q)\|\psi\| h \\ |z_2| &\leqslant (1 + q + q^2)\|\psi\| h \\ &\dots \\ |z_n| &\leqslant (1 + q + q^2 + \ldots + q^{n-1})\|\psi\| h \end{aligned} \tag{189}$$

Воспользуемся следующими неравенствами:

$$1+q+q^2+\ldots+q^{n-1}<nq^n=ne^{Chn}\Rightarrow|z_i|\le ne^{Chn}\|\psi\|h,\quad i=\overline{0,n}\Leftrightarrow$$
$$\Leftrightarrow\|z\|\leqslant le^{Cl}\|\psi\| \tag{190}$$

Отсюда вывод: *чем лучше разностное уравнение аппроксимирует дифференциальное, тем меньше погрешность решения.*

Оценка скорости сходимости

Оценим $\|\psi\|$. Предположим, что функция $f(x,u)$ имеет в рассматриваемой области изменения аргументов непрерывные и ограниченные первые частные производные f'_x и f'_y. Это обеспечивает существование у решения задачи непрерывной и ограниченной второй производной

$$u''(x)=(f(x,u))'=f'_x+f'_uu'=f'_x+f'_uf(x,u). \tag{191}$$

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

$$u_{i+1}=u(x_i+h)=u_i+u'(x_i)h+\frac{1}{2}u''(x_i+\theta_ih)h^2. \tag{192}$$

Подставляя разложение в формулу 183 и пользуясь оценкой 190, получим

$$\psi_i=\frac{1}{2}u''(x_i+\theta_ih)h\Rightarrow$$
$$\|\psi\|\leqslant\frac{1}{2}M_2h,\quad\|z\|\leqslant\frac{M_2}{2}le^{Cl}h \tag{193}$$

Неравенства показывают, что при $h\to0$ *аппроксимация уравнения и погрешность решения стремятся к идеальным* со скоростью h.

В связи с этим метод Эйлера называют методом первого порядка точности относительно h.

Повышение точности разностной схемы

Предположим, что решение $u(x)$ имеет производные достаточно высокого порядка, тогда

$$u_{i+1}=u(x_i+h)=u_i+u'(x_i)h+\frac{1}{2}u''(x_i)h^2+\frac{1}{6}u'''(x_i)h^3+\ldots \tag{194}$$

Если оборвать разложение на члене порядка h и подставить $u'=f(x,y)$, то получим схему Эйлера. Оборвем разложение на члене порядка h^2 и воспользуемся формулой

$$u''(x)=\frac{df}{dx}(x,u)+\frac{df}{du}(x,u)f(x,u). \tag{195}$$

В результате получим новое рекуррентное соотношение

$$y_{i+1}=y_i+f(x_i,y_i)h+\frac{1}{2}\left[f'_x(x_i,y_i)+f'_y(x_i,y_i)f(x_i,y_i)\right]h^2\Leftrightarrow$$
$$\frac{y_{i+1}-y_i}{h}=f(x_i,y_i)+\frac{1}{2}[f'_x(x_i,y_i)+f'_y(x_i,y_i)f(x_i,y_i)]h \tag{196}$$

Уравнение, дополненное начальным условием $y_0=u_0$, дает явную разностную схему численного решения рассматриваемой задачи Коши. По рекуррентной формуле можно последовательно рассчитать все значения сеточной функции $y_i,\ 0\le i\le n$ и получить таким образом приближенное решение задачи.

Исследование показывает, что такая усложненная схема имеет второй порядок точности относительно как для аппроксимации уравнения, так и для погрешности решения. Причём, если оборвать разложение на членах большей степени, точность будет ещё больше. Но возрастает и сложность: для численного решения такой задачи приходится считать производные $f(x,y)$.

Поэтому в разностных схемах высокго порядка стараются заменить вычисление производных.

Схема Рунге-Кутта второго порядка

Вывод

Главная идея метода Рунге-Кутта состоит в том, чтобы приближенно заменить производные в уравнении 196 на сумму значений функции f с точностью до членов порядка h^2.

С этой целью положим:

$$\begin{aligned} f(x_i, y_i) + \frac{1}{2}[f'_x(x_i, u_i) + f'_y(x_i, y_i) f(x_i, y_i)]h = \\ = \beta f(x_i, y_i) + \alpha f(x_i + \gamma h, y_i + \delta h) + O(h^2) \end{aligned} \tag{197}$$

где $\alpha, \beta, \gamma, \delta$ — четыре свободных параметра, которые нужно подобрать.

Разложим функцию $f(x_i + \gamma h, y_i + \delta h)$ по степеням h:

$$f(x_i + \gamma h, y_i + \delta h) = f(x_i, y_i) + [\gamma f'_x(x_i, y_i) + \delta f'_y(x_i, y_i)]h + O(h^2). \tag{198}$$

Подставим разложение в 197 и приравняем слева и справа члены соинаковыми степенями h. В результате получим параметрическое решение:

$$\begin{aligned} \alpha + \beta = 1, \quad \alpha\gamma = \frac{1}{2}, \quad \alpha\delta = \frac{1}{2} f(x_i, y_i) \Rightarrow \\ \beta = 1 - \alpha, \quad \gamma = \frac{1}{2\alpha}, \quad \delta = \frac{1}{2\alpha} f(x_i, y_i). \end{aligned} \tag{199}$$

Заменяя левую часть уравнения и отбрасывая члены порядка $O(h^2)$, получим однопараметрическое семейство разностных схем Рунге-Кутта:

$$\begin{aligned} y_{i+1} = y_i + h\left[(1-\alpha) f(x_i, y_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, y_i + \frac{h}{2\alpha} f(x_i, y_i)\right)\right] \\ \frac{y_{i+1} - y_i}{h} = (1-\alpha) f(x_i, y_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, y_i + \frac{h}{2\alpha} f(x_i, y_i)\right) \end{aligned} \tag{200}$$

Наиболее удобны схемы при $\alpha = 1/2$ и $\alpha = 1$. Первая схема называется **“предиктор–корректор”**, а вторая является “двойной” схемой Эйлера. Подробнее об этом в учебнике на стр. 162.

Погрешность аппроксимации

Рассмотрим погрешность аппроксимации уравнения:

$$\psi_i = \frac{u_{i+1} - u_i}{h} - \left[(1-\alpha) f(x_i, y_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, y_i + \frac{h}{2\alpha} f(x_i, y_i)\right)\right] \tag{201}$$

Подставим $y_i = z_i + u_i$ в разностное уравнение 200. В результате получим

$$\begin{aligned} \frac{z_{i+1} - z_i}{h} + \frac{u_{i+1} - u_i}{h} = \\ = (1-\alpha) f(x_i, u_i + z_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, u_i + z_i + \frac{h}{2\alpha} f(x_i, u_i + z_i)\right) \end{aligned} \tag{202}$$

Эту формулу можно переписать в виде

$$\begin{aligned} \frac{z_{i+1} - z_i}{h} = \left\{\left[(1-\alpha) f(x_i, u_i + z_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, u_i + z_i + \frac{h}{2\alpha} f(x_i, u_i + z_i)\right)\right] - \right. \\ \left. - \left[(1-\alpha) f(x_i, u_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, u_i + \frac{h}{2\alpha} f(x_i, u_i)\right)\right]\right\} - \\ \left\{ \frac{u_{i+1} - u_i}{h} - \left[(1-\alpha) f(x_i, u_i) + \alpha f\left(x_i + \frac{h}{2\alpha}, u_i + \frac{h}{2\alpha} f(x_i, u_i)\right)\right]\right\} \end{aligned} \tag{203}$$

$$-\left\{\frac{\quad}{h} - \underbrace{\left[(1-\alpha)f(x_i,u_i)+\alpha f\left(x_i+\frac{}{2\alpha},u_i+\frac{}{2\alpha}f(x_i,u_i)\right)\right]}_{F(u_i)}\right\}.$$

Вторая фигурная скобка – это ψ_i. А первая – это $F(u_i+z_i)-F(u_i)$. Тогда по теореме Лагранжа:

$$F(u_i+z_i)-F(u_i)=F'(u_i+\theta_i z_i)z_i. \tag{204}$$

где

$$F'(v)=(1-\alpha)f'_v(x_i,v)+\alpha f'_v\left(x_i+\frac{h}{2\alpha},v+\frac{h}{2\alpha}f(x_i,v)\right)\left(1+\frac{h}{2\alpha}f'_v(x_i,v)\right) \tag{205}$$

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

$$z_{i+1}=z_i\left[1+hF'(u_i+\theta_i z_i)\right]-\psi_i h,\quad i=\overline{0,n-1} \tag{206}$$

которую нужно дополнить нулевым начальным условием $z_0=0$

Необходимым условием решения задачи Коши являтся ограниченность производной:

$$\left|\frac{df}{du}(x,u)\right|\le C. \tag{207}$$

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

$$\begin{gathered}|1+hF'(u_i+\theta_i z_i)|\le 1+Ch+\frac{1}{2}C^2h^2<e^{Ch}=q\Rightarrow\\ \Rightarrow |z_{i+1}|\le q|z_i|+\|\psi\|_c h\end{gathered} \tag{208}$$

Отсюда получается цепочка оценок, в точности такая же, как 189. В результате оценка погрешности решения принимает вид.

$$\|z\|\le le^{Cl}\|\psi\| \tag{209}$$

где l – длина отрезка, на котором рассматривается решение исходной задачи. Вывод: *чем лучше разностное уравнение аппроксимирует дифференциальное, тем меньше погрешность решения.*

Скорость сходимости

Теперь нужно оценить норму погрешности аппроксимации уравнения ψ_i.

Предположим, что $f(x,u)$ имеет в интересующей нас области изменения своих аргументов непрерывные вторые производные. Это позволяет написать разложение Тейлора для u_{i+1}

$$u_{i+1}=u(x_i+h)=u_i+u'(x_i)h+\frac{1}{2}u''(x_i)h^2+\frac{1}{6}u'''(\widehat{x}_i)h^3. \tag{210}$$

И для $f(x,y)$:

$$\begin{gathered}f\left(x_i+\frac{h}{2\alpha},u_i+\frac{h}{2\alpha}f(x_i,u_i)\right)=f(x_i,u_i)+\frac{h}{2\alpha}\left[f'_x(x_i,u_i)+f'_u(x_i,u_i)f(x_i,u_i)\right]+\\ +\frac{h^2}{8\alpha^2}\left[f''_{xx}(\tilde{x}_i,\tilde{u}_i)+2f''_{xu}(\tilde{x}_i,\tilde{u}_i)f(\tilde{x}_i,\tilde{u}_i)+f''_{uu}(\tilde{x}_i,\tilde{u}_i)f^2(\tilde{x}_i,\tilde{u}_i)\right]\end{gathered} \tag{211}$$

где

$$\widehat{x}_i=x_i+\hat{\theta}_i h,\quad \tilde{x}_i=x_i+\tilde{\theta}\frac{h}{2\alpha},\quad \tilde{u}_i=u_i+\tilde{\theta}\frac{h}{2\alpha}f(x_i,u_i). \tag{212}$$

Подставим разложения в ψ. Члены нулевого и первого порядков сокращаются и остаются только члены второго порядка. В результате получаем

$$\left(\frac{1}{\quad}-\frac{1}{\quad}\right)$$

$$\psi_i = h^2 \left\{ \frac{1}{6}u'''(\bar{x}_i) - \frac{1}{8\alpha}[f''_{xx}(\tilde{x}_i,\tilde{u}_i) + 2f''_{xu}(\tilde{x}_i,\tilde{u}_i)f(\tilde{x}_i,\tilde{u}_i) + f''_{uu}(\tilde{x}_i,\tilde{u}_i)f^2(\tilde{x}_i,\tilde{u}_i)] \right\}. \quad (213)$$

Все производные непрерывны по предположению, а значит,

$$\|\psi\|_c \leqslant Mh^2 \Rightarrow \|z\| \leqslant le^{Cl} \cdot Mh^2. \quad (214)$$

Таким образом, скорость сходимости $O(h^2)$. Это означает, что разностное уравнение, полученное по схеме Рунге-Кутта, имеет второй порядок точности относительно h.

Схема Рунге-Кутта четвертого порядка

Как сказано в учебнике, “второй порядок точности лучше, чем первый, однако практика показывает, что этой точности тоже не хватает.” Наиболее часто при проведении реальных расчетов используется схема Рунге-Кутта четвертого порядка точности:

$$\frac{y_{i+1} - y_i}{n} = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \quad (215)$$

где

$$\begin{aligned} k_1 &= f(x_i, y_i), \\ k_2 &= f\left(x_i + \frac{h}{2},\ x_i + \frac{h}{2}k_1\right) \\ k_3 &= f\left(x_i + \frac{h}{2},\ x_i + \frac{h}{2}k_2\right) \\ k_4 &= f(x_i + h,\ x_i + hk_3). \end{aligned} \quad (216)$$

Схема Адамса

Вывод

Рассмотрим задачу Коши для дифференциального уравнения первого порядка

$$u' = f(x,u), \quad u(x_0) = u_0. \quad (217)$$

Если функция $f(x,u)$ непрерывна и удовлетворяет условию Липшица по аргументу u в некоторой окрестности начальной точки (x_0,u_0), то можно указать такой отрезок $[a,b],\ a < x_0 < b$, на котором решение задачи $u(x)$ существует и является единственным.есть

Пусть $u(x)$ — решение дифференциального уравнения. Для производной этой функции имеет место равенство

$$u'(x) = f(x,u(x)) = F(x) \quad (218)$$

Интегрируя его между двумя точками сетки, получим соотношение

$$u_{i+1} = u_i + \int_{x_i}^{x_{i+1}} F(x)dx. \quad (219)$$

Пусть в процессе численного решения задачи мы довели расчет до точки x_i. В результате проведенных расчетов нам оказались известными величины y_j и $f(x_j,y_j)$.

Возьмем некоторое фиксированное целое число $m \leqslant i$ и построим интерполяционный многочлен m -ой степени, принимающий в точках $x_j,\ j = \overline{i-m,i}$ значения $f(x_j,y_j)$

$$\begin{aligned} P_m(x_j) &= f(x_j,y_j), \quad j = \overline{i-m,i}, \\ P_m(x) &= \sum_{j=i-m}^{i} f(x_j,y_j)Q_{j,m}(x), \\ Q_{j,m}(x) &= \frac{(x-x_{i-m})\dots(x-x_{j-1})(x-x_{j+1})\dots(x-x_i)}{} \end{aligned} \quad (220)$$

$$\varphi_{j,m}(x) = \frac{}{(x_j - x_{i-m})\dots(x_j - x_{j-1})(x_j - x_{j+1})\dots(x_j - x_i)}.$$

Главная идея метода Адамса заключается в том, чтобы для расчета y_{i+1} использовать формулу типа $u_{i+1} = u_i + \int_{x_i}^{x_{i+1}} F(x)dx$, приближенно заменяя в ней функцию $F(x)$ на интерполяционный многочлен $P_m(x)$, составленный по результатам предыдущих вычислений. Это приводит к рекуррентной формуле

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} P_m(x)dx = y_i + \sum_{j=i-m}^{i} a_j f(x_j, y_j), \quad \text{где} \quad a_j = \int_{x_i}^{x_{i+1}} Q_{j,m}(x)dx \quad (221)$$

Погрешность аппроксимации для схемы с $m = 1$

Перейдем к исследованию варианта $m = 1$. В этом случае для аппроксимации функции $F(x)$ используется полином первой степени, построенный по значениям функции f в двух точках (x_{i-1}, y_{i-1}) и (x_i, y_i)

$$\begin{gathered} P_1(x) = f(x_{i-1}, y_{i-1})\varphi_{i-1,1} - f(x_i, y_i)\varphi_{i,1} \Rightarrow \\ c_{i-1} = \int_{x_i}^{x_{i+1}} \varphi_{i-1,1}(x)dx = \int_{x_i}^{x_{i+1}} \frac{x - x_i}{x_{i-1} - x_i}dx = -\frac{1}{h}\int_0^{x_{i+1}-x_i} zdz = \\ = -\frac{1}{2h}(x_{i+1} - x_i)^2 = -\frac{1}{2}h, \\ c_i = \int_{x_i}^{i-1} \varphi_{i,1}(x)dx = \int_{x_i}^{x_{i+1}} \frac{x - x_{i-1}}{x_i - x_{i-1}}dx = \frac{1}{h}\int_{x_i - x - i - 1}^{x_{i+1}-x_{i-1}} zdx = \\ = \frac{1}{2h}(4h^2 - h^2) = \frac{3}{2}h \Rightarrow \\ y_{i+1} = y_i + \left[\frac{3}{2}f(x_i, y_i) - \frac{1}{2}f(x_{i-1}, y_{i-1})\right]h. \end{gathered} \quad (222)$$

Отметим следующую особенность рекуррентной формулы. 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 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.

Подсчитаем погрешность аппроксимации:

$$\begin{gathered} \psi_i = \frac{u_{i+1} - u_i}{h} - \left[\frac{3}{2}f(x_i, y_i) - \frac{1}{2}f(x_{i-1}, y_{i-1})\right] = \\ = \frac{u_{i+1} - u_i}{h} - \left[\frac{3}{2}u'(x_i) - \frac{1}{2}u'(x_{i-1})\right] \end{gathered} \quad (223)$$

Предположим, что функция $f(x, u)$ имеет в интересующей нас области изменения аргументов непрерывные вторые производные. В этом случае решение задачи $u(x)$ трижды непрерывно дифференцируемо. Запишем разложения Тейлора для $u'(x_i)$ и $u'(x_{i-1})$:

$$\begin{gathered} u_{i+1} = u_i + u'(x_i)h + \frac{1}{2}u''(x_i)h^2 + \frac{1}{6}u'''(x_i + \tilde{\theta}_i h)h^3 \\ u'(x_{i-1}) = u'(x_i - h) = u'(x_i) - u''(x_i)h + \frac{1}{2}u'''(x_i + \tilde{\tilde{\theta}}_i h)h^2 \end{gathered} \quad (224)$$

Подставляя их в формулу, получим

$$\psi_i = \left[\frac{1}{6}u'''(x_i + \tilde{\theta}_i h) + \frac{1}{4}u'''(x_i + \tilde{\tilde{\theta}}_i h)\right]h^2 \Rightarrow \|\psi\| \leqslant \frac{5}{12}M_3 h^2, \; M_3 = \sup_{[a,b]}|u''(x)| \quad (225)$$

Мы видим, что разностное уравнение метода Адамса, соответствующее случаю $m = 1$, аппроксимирует дифференциальное уравнение со вторым порядком точности относительно h. Как и в случае метода Рунге-Кутта, это обеспечивает второй порядок точности для погрешности решения $\|z\|_c$ при предположении, что значение y_1, которое рассчитывается нестандартно, вычислено со вторым порядком точности.

Схема с $m = 3$

Если написать интерполяционный полином $P_3(x)$ то формула 221 примет вид

$$y_{i+1} = y_i + h\left[\frac{55}{24}f(x_i, y_i) - \frac{59}{24}f(x_{i-1,y_{i-1}}) + \frac{37}{24}f(x_{i-2}, y_{i-2}) - \frac{9}{24}f(x_{i-3}, y_{i-3})\right]. \quad (226)$$

В другой форме:

$$\begin{aligned} y_{i+1} &= i_i + hf_i + \frac{h^2}{2}\Delta^1 f_i + \frac{5h^3}{12}\Delta^2 f_i + \frac{3h^4}{8}\Delta^3 f_i, \\ f_i &= f(x_i, y_i) \\ \Delta^1 f_i &= \frac{1}{h}[f(x_i, y_i) - f(x_{i-1}, y_{i-1})]/h, \\ \Delta^2 f_i &= \frac{1}{h^2}[f(x_i, y_i) - 2f(x_{i-1}, y_{i-1}) + f(x_{i-2}, y_{i-2})], \\ \Delta^3 f_i &= \frac{1}{h^3}[f(x_i, y_i) - 3f(x_{i-1}, y_{i-1}) + 3f(x_{i-2}, y_{i-2}) - f(x_{i-3}, y_{i-3})], \end{aligned} \quad (227)$$

Сеточные функции $\Delta^k f_i,\ k = \overline{1,3}$ аппроксимируют производные функции $F(x) = f(x, u(x))$.

Если $f(x, u) \in C^4$, то в такой схеме $\psi = O(h^4)$.

Численное решение краевой задачи для ЛДУ 2-го порядка

Краевая задача

Рассмотрим следующую задачу Коши на отрезке $[a, b]$:

$$u'' - q(x)u = -f(x), \quad u(a) = u_1, \quad u(b) = u_2 \quad (228)$$

В ней начальные условия заданы в граничных точках отрезка, поэтому задачу называют **краевой**.

Пусть $f(x), q(x) \in C[a, b]$, причем $q(x) \geqslant q_0 > 0$.

При сделанных предположениях, как известно из курса дифференциальных уравнений, решение задачи существует и является единственным.

Разностная схема

Построим сетку для некоторого n :

$$x_i = a + ih, \quad i = \overline{0,n}, \quad h = (b - a)/n. \quad (229)$$

Заменим дифференциальное уравнение разностной задачей:

$$\begin{gathered} \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} - q_i y_i = -f_i, \quad i = \overline{1, n-1}, \\ y_0 = u_0, \quad y_n = u_n. \end{gathered} \quad (230)$$

Разностные уравнения можно переписать в виде

$$y_{i-1} - (2 + q_i h^2)y_i + y_{i+1} = -f_i h^2, \quad i = \overline{1, n-1} \quad (231)$$

Мы получили линейную систему из $(n-1)$-го уравнения с $(n-1)$-им неизвестным. Значения y_0 и y_n неизвестными не являются: они задаются граничными условиями.

В отличие от предыдущих схем, такая разностная схема называется **неявной**.

Из записи разностных уравнений видно, что *мы получили систему уравнений с трехдиагональной матрицей с диагональным преобладанием*: диагональный элемент $(2 + q_i h^2)$ больше суммы двух других элементов той же строки, равной 2.

Диагональное преобладание гарантирует существование и единственность решения системы, которое может быть построено методом прогонки.

Погрешность аппроксимации разностной задачи

Погрешность решения и аппроксимации будут следующими:

$$\begin{aligned} z_i &= y_i - u_i, & i &= \overline{0, n} \\ \psi_i &= \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} - q_i u_i + f_i, & i &= \overline{1, n-1} \end{aligned} \tag{232}$$

Подставим $y_i = z_i + u_i$ в разностное уравнение:

$$\frac{z_{i-1} - 2z_i + z_{i+1}}{h^2} - q_i z_i = -\underbrace{\left[\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} - q_i u_i + f_i\right]}_{\psi_i} \tag{233}$$

Граничные условия $z_0 = z_n = 0$. Тогда если $j : \|z\| = |z_j|$, то j не равен ни 0, ни n. Рассмотрим уравнение для этого значения индекса и запишем его в виде:

$$\begin{gathered} (2 + q_j h^2) z_j = z_{j-1} + z_{j+1} + \psi_j h^2 \Rightarrow \\ (2 + q_j h^2)|z_j| = (2 + q_j h^2)\|z\| \leqslant |z_{j-1}| + |z_{j+1}| + |\psi_j| h^2 \leq 2\|z\|_c + \|\psi\|_c h^2 \Rightarrow \\ \|z\| \leqslant \frac{1}{q_0}\|\psi\|, \quad q_0 = \inf_{[a,b]} |q(x)| \end{gathered} \tag{234}$$

Таким образом, нам удалось оценить погрешность решения $\|z\|$ через погрешность аппроксимации уравнения $\|\psi\|$.

Скорость сходимости

Для оценки погрешности аппроксимации уравнения предположим, что функции $f(x)$ и $q(x) \in C^2[a, b]$. Это позволяет написать разложения:

$$\begin{aligned} u_{i-1} = u(x_i - h) &= u_i - u'(x_i)h + \frac{1}{2}u''(x_i)h^2 - \frac{1}{6}u'''(x_i)h^3 + \frac{1}{24}u^{(4)}(x_i - \tilde{\theta}_i h)h^4. \\ u_{i+1} = u(x_i + h) &= u_i + u'(x_i)h + \frac{1}{2}u''(x_i)h^2 + \frac{1}{6}u'''(x_i)h^3 + \frac{1}{24}u^{(4)}(x_i + \tilde{\tilde{\theta}}_i h)h^4 \end{aligned} \tag{235}$$

Подставляя их в формулу, получим следующее выражение для ψ_i:

$$\begin{gathered} \psi_i = [u''(x_i) - q_i u_i + f_i] + \frac{h^2}{24}\left[u^{(4)}(x_i - \tilde{\theta}_i h) + u^{(4)}(x_i + \tilde{\tilde{\theta}}_i h)\right] = \\ = \{u'' - q(x)u = -f(x)\} = \frac{h^2}{24}\left[u^{(4)}(x_i - \tilde{\theta}_i h) + u^{(4)}(x_i + \tilde{\tilde{\theta}}_i h)\right] \Rightarrow \\ \Rightarrow \|\psi\| \leqslant \frac{M_4}{12}h^2, \quad \|z\| \leqslant \frac{M_4}{12 q_0}h^2. \end{gathered} \tag{236}$$

Мы видим, что разностная схема обеспечивает второй порядок аппроксимации уравнения и второй порядок точности для погрешности решения.

Разностная краевая задача на собственные значения

Постановка задачи

Краевая задача на собственные значения для дифференциального уравнения второго порядка, собственные числа, собственные функции. Разностная задача на собственные значения, собственные значения и собственные функции

Рассмотрим задачу Штурма–Луивилля:

$$u''(x) + \lambda u(x) = 0, \quad a < x < b, \quad u(a) = u(b) = 0. \tag{237}$$

Она имеет решение только при $\lambda > 0$:

$$\lambda_k = \left(\frac{\pi k}{b-a}\right)^2,\ u_k = \sin\frac{\pi k(x-a)}{b-a},\ k \in \mathbb{N} \tag{238}$$

Введём равномерную сетку:

$$\omega_h = \Big\{x : x_i = a + ih, \quad i = \overline{0, N}, \quad h = (b-a)/N\Big\}. \tag{239}$$

Тогда разностная схема будет такой:

$$\frac{y_{j-1} - 2y_j + y_{j+1}}{h^2} + \lambda^{(h)} y_j = 0, \quad j = \overline{1, N-1}. \tag{240}$$

Система уравнений представляет собой задачу на собственные значения $Ay = \lambda^{(h)} y$ для симметричной трёхдиагональной матрицы A порядка $N-1$ с диагональным преобладанием:

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{bmatrix} \tag{241}$$

Поэтому существует ровно $N-1$ вещественных собственных значений $\lambda_k^{(h)}$ матрицы A.

Собственные значения и функции

Построим в явном виде собственные значения и собственные функции задачи. Перепишем разностное уравнение в виде

$$y_{i-1} - (2-\mu)y_i + y_{i+1} = 0, \quad \mu = h^2\lambda^{(h)}. \tag{242}$$

Отвечающее ему характеристическое уравнение: $q^2 - (2-\mu)q + 1 = 0$. Следовательно, общее решение уравнения имеет вид

$$\begin{gathered} y_j = c_1 q_1^j + c_2 q_2^j,\ \forall C_1, C_2, \\ q_{1,2} = 1 - \frac{\mu}{2} \pm \sqrt{\left(1 - \frac{\mu}{2}\right)^2 - 1}. \end{gathered} \tag{243}$$

Из граничных условий $y_0 = y_N = 0$ получаем

$$\begin{cases} c_1 + c_2 = 0, \\ c_1 q_1^N + c_2 q_2^N = 0. \end{cases} \tag{244}$$

Эта система имеет нетривиальное решение только при $q_1^N = q_2^N$. Из характеристического уравнения по теореме Виета имеем $q_1 q_2 = 1 \Rightarrow$

$$q_1^{2N} = 1 \Rightarrow q_1 = \exp\frac{\imath\pi k}{N} = \cos\frac{\pi k}{N} + i\sin\frac{\pi k}{N}, \quad k = \overline{1, N-1}. \tag{245}$$

Следовательно, зная q_1, q_2, получим

$$\begin{gathered}\cos\frac{\pi k}{N} = 1 - \frac{\mu}{2} \Rightarrow \mu = 2\left(1 - \cos\frac{\pi k}{N}\right) = 4\sin^2\frac{\pi k}{2N} \Rightarrow \\ \lambda^{(h)} = \frac{4}{h^2}\sin^2\frac{\pi k}{2N}, \quad k = \overline{1, N-1}\end{gathered} \tag{246}$$

Собственные функции y_j вычисляются согласно 243, где $c_2 = -c_1$. Так как $q_1 q_2 = 1$, то

$$\begin{gathered}y_j = c_1(q_1^j - q_2^j) = c_1(q_1^j - q_1^{-j}) = c_1(e^{ij\varphi} - e^{-ij\varphi}),\ \varphi = \pi k/N \Rightarrow \{\text{пусть } c_1 = -i/2\} \Rightarrow \\ y_j = -\frac{i}{2}[\cos j\varphi + i\sin j\varphi - \cos j\varphi + i\sin j\varphi] = \sin\frac{\pi k j}{N}, \quad j = \overline{1, N-1}.\end{gathered} \tag{247}$$

Собственные функции определены с точностью до произвольного постоянного (не зависящего от j) множителя.

Свойства собственных значений и собственных функций

1°. Перечислим свойства собственных значений и собственных функций разностной задачи. Прежде всего из 246 следует цепочка строгих неравенств

$$0 < \lambda_1^{(h)} < \lambda_2^{(h)} < \ldots < \lambda_{N-1}^{(h)} < \frac{4}{h^2}. \tag{248}$$

2°. Оценку $\lambda_1^{(h)}$ можно уточнить:

$$\alpha = \frac{\pi h}{2(b-a)}, \quad \lambda_1 = \frac{\pi^2}{(b-a)^2} \Rightarrow \lambda_1^{(h)} = \lambda_1\left(\frac{\sin\alpha}{\alpha}\right)^2 \tag{249}$$

Не ограничивая общности, можно предположить, что $h \leqslant (b-a)/3 \Rightarrow \alpha \leqslant \pi/6$, и поскольку функция $\sin\alpha/\alpha$ монотонно убывает при $\alpha \in [0, \pi/6]$, получим

$$\left(\frac{\sin\alpha}{\alpha}\right)^2 \geqslant \left(\frac{1/2}{\pi/6}\right)^2 = \frac{9}{\pi^2} \Rightarrow 0 < \frac{9}{\pi^2} \leqslant \lambda_1^{(h)}. \tag{250}$$

3°. Собственные функции задачи, отвечающие различным собственным значениям, ортогональны в смысле скалярного произведения.

4°. Спектр дифференциальной задачи не ограничен, а спектр разностной ограничен числом $4/h^2$:

$$\lim_{k\to\infty}\lambda_k = +\infty, \quad \lambda_k^{(h)} < \frac{4}{h^2} \tag{251}$$

5°. Собственные значения $\lambda_k^{(h)}$ сходятся слева к λ_k

$$\lim_{h\to 0}\frac{4}{h^2}\sin^2\frac{\pi k h}{2(b-a)} = \lim_{h\to 0}\left[\frac{\pi k}{b-a}\right]^2\left[\frac{\sin\frac{\pi k h}{2(b-a)}}{\frac{\pi k h}{2(b-a)}}\right]^2 = \left[\frac{\pi k}{b-a}\right]^2 = \lambda_k \tag{252}$$

Погрешность $\lambda_k - \lambda_k^{(h)}$ сильно возрастаёт с ростом k.