Міністерство освіти України
Київський національний університет ім. Тараса Шевченко

КОНСПЕКТ ЛЕКЦІЙ

з курсу “Чисельні методи”

для студентів 3 курсу факультету кібернетики
Київського національного університету ім. Тараса Шевченко

КИЇВ 2006

ЗМІСТ

1. Аналіз похибок заокруглення [СГ, 16-25], [БЖК, 17-33]

1.1 Види похибок

Нехай необхідно розв'язати рівняння

$$Au = f\,. \tag{1}$$

За рахунок неточно заданих вхідних даних насправді ми маємо рівняння

$$\tilde{A}\tilde{u} = \tilde{f}\,. \tag{2}$$

Назвемо $\delta_1 = u - \tilde{u}$ - *неусувною похибкою.*

Застосування методу розв'язання (2) приводить до рівняння

$$\tilde{A}_h\tilde{u}_h = \tilde{f}_h\ , \tag{3}$$

де $h > 0$ - малий параметр. Назвемо $\delta_2 = \tilde{u} - \tilde{u}_h$ - *похибкою методу*.

Реалізація методу на ЕОМ приводить до рівняння

$$\tilde{A}_h^*\tilde{u}_h^* = \tilde{f}_h^*\,. \tag{4}$$

Назвемо $\delta_3 = \tilde{u}_h - \tilde{u}_h^*$ - *похибкою заокруглення*.

Тоді *повна похибка* $\delta = u - \tilde{u}_h^* = \delta_1 + \delta_2 + \delta_3$.

Означення. Кажуть, що задача (1) *коректна*, якщо

1) $\forall f \in F \quad \exists!\, u \in U$;

2) задача (1) *стійка* , тобто

$$\forall \varepsilon > 0\ \exists \delta > 0:\ \| A - \tilde{A} \| < \delta, \| f - \tilde{f} \| < \delta \Rightarrow \| u - \tilde{u} \| < \varepsilon\,.$$

Якщо задача (1) некоректна, то або розв'язок її не існує, або він неєдиний, або він нестійкий, тобто

$$\exists \varepsilon > 0\ \forall \delta > 0:\ \| A - \tilde{A} \| < \delta, \| f - \tilde{f} \| < \delta \Rightarrow \| u - \tilde{u} \| > \varepsilon\,.$$

Абсолютна похибка $\Delta x \le | x - x^* |$.

Відносна похибка $\delta x \le \dfrac{\Delta x}{| x |}$ або $\dfrac{\Delta x}{| x^* |}$.

Значущими цифрами називаються всі цифри, починаючи з першої ненульової зліва.

Вірна цифра – це значуща, якщо абсолютна похибка за рахунок відкидання всіх молодших розрядів не перевищує одиниці розряду цієї цифри. Тобто, якщо $x^* = \alpha_n \ldots \alpha_0, \alpha_{-1} \ldots \alpha_{-p} \ldots$, то α_{-p} - вірна, якщо $\Delta x \le 10^{-p}$ (інколи $\Delta x \le w \cdot 10^{-p}, \dfrac{1}{2} \le w < 1$, наприклад, $w = 0{,}55$).

1.2. Підрахунок похибок в ЕОМ

Підрахуємо відносну похибку заокруглення числа x на ЕОМ з плаваючою комою. В β-ічній системі числення число представляється у вигляді

$$x = \pm(\alpha_1\beta^{-1} + \alpha_2\beta^{-2} + \ldots + \alpha_t\beta^{-t} + \ldots)\beta^p, \tag{5}$$

де $0 \le \alpha_k < \beta, \quad \alpha_1 \ne 0, \quad k = 1,2,\ldots$

Якщо в ЕОМ t розрядів, то при відкиданні молодших розрядів ми оперуємо з наближеним значенням

$$x^* = \pm(\alpha_1\beta^{-1} + \alpha_2\beta^{-2} + \ldots + \alpha_t\beta^{-t})\beta^p$$

і відповідно похибка заокруглення $x - x^* = \pm\beta^p(\alpha_{t+1}\beta^{-t-1} + \ldots)$. Тоді її можна оцінити так

$$|x - x^*| \le \beta^{p-t-1}(\beta-1)(1+\beta^{-1}+\ldots) \le \beta^{p-t-1}(\beta-1)\frac{1}{1-\beta^{-1}} = \beta^{p-t}$$

Якщо в представлені (5) взяти $\alpha_1 = 1$, то $|x| \ge \beta^p\beta^{-1}$. Звідси остаточно

$$\delta x \le \frac{\beta^{p-t}}{\beta^{p-1}} = \beta^{-t+1}.$$

При більш точних способах заокруглення можна отримати оцінку $\delta x \le \frac{1}{2}\beta^{-t+1} = \varepsilon$. Число ε називається “машинним іпсилон”. Наприклад, для $\beta = 2, \quad t = 24, \quad \varepsilon = 2^{-24} \approx 10^{-7}$.

1.3. Підрахунок похибок обчислення значення функції

Нехай задана функція $y = f(x_1,\ldots,x_n) \in C^1(\Omega)$. Необхідно обчислити її значення при наближеному значенні аргументів $\vec{x}^* = (x_1^*,\ldots,x_n^*)$, де $|x_i - x_i^*| \le \Delta x_i$ та оцінити похибку обчислення значення функції $y^* = f(x_1^*,\ldots,x_n^*)$. Маємо

$$|y - y^*| = |f(\vec{x}) - f(\vec{x}^*)| = \left|\sum_{i=1}^n \frac{\partial f}{\partial x_i}(\vec{\xi})(x_i - x_i^*)\right| \le \sum_{i=1}^n B_i \cdot \Delta x_i, \text{ де } B_i = \max_{\vec{x}\in U}\left|\frac{\partial f}{\partial x_i}(\vec{x})\right|.$$

Тут $U = \{\vec{x} = (x_1,\ldots,x_n) : |x_i - x_i^*| \le \Delta x_i\} \in \Omega, \ i = \overline{1,n}$. Отже з точністю до величин першого порядку малості по $\Delta x = \max_i \Delta x_i$ $\Delta y = |y - y^*| \prec \sum_{i=1}^n b_i \cdot \Delta x_i$, де $b_i = \left|\frac{\partial f}{\partial x_i}(\vec{x}^*)\right|$ та "$\prec$" означає приблизно менше.

Розглянемо похибки арифметичних операцій.

1. Сума: $y = x_1 + x_2, \ x_1, x_2 > 0,$

$$\Delta y \le \Delta x_1 + \Delta x_2, \qquad \delta y \le \frac{\Delta x_1 + \Delta x_2}{x_1 + x_2} \le \max(\delta x_1, \delta x_2).$$

2. Різниця: $y = x_1 - x_2, \quad x_1 > x_2 > 0,$

why?

$$\Delta y \le \Delta x_1 + \Delta x_2, \qquad \delta y \le \frac{x_2\delta x_1 + x_1\delta x_2}{x_1 - x_2}.$$

При близьких x_1, x_2 зростає відносна похибка (за рахунок втрати вірних цифр).

3. Добуток: $y = x_1 \cdot x_2, \quad x_1, x_2 > 0,$

$$\Delta y \prec x_2\Delta x_1 + x_1\Delta x_2 \qquad \delta y \le \delta x_1 + \delta x_2 .$$

4. Частка: $y = \dfrac{x_1}{x_2},\ x_1, x_2 > 0,$

$$\Delta y \prec \frac{x_2\Delta x_1 + x_1\Delta x_2}{(x_2)^2} \qquad \delta y \le \delta x_1 + \delta x_2 .$$

При малих x_2 зростає абсолютна похибка (за рахунок зростання результату ділення).

Пряма задача аналізу похибок : обчислення Δy, δy по заданих $\Delta x_i,\ i = \overline{1,n}$.

Обернена задача: знаходження $\Delta x_i, i = \overline{1,n}$ по заданих Δy, δy. Якщо $n > 1$, маємо одну умову $\sum_{i=1}^{n} b_i \Delta x_i < \varepsilon$ для багатьох невідомих Δx_i. Вибирають їх із однієї з умов

$$\forall i : b_i \Delta x_i < \frac{\varepsilon}{n} \text{ або } \Delta x_i < \frac{\varepsilon}{\sum_{i=1}^{n} b_i}.$$

2. Методи розв'язання нелінійних рівнянь

Постановка задачі. Нехай маємо рівняння $f(x) = 0$, $\bar{x}$ - його розв'язок, тобто $f(\bar{x}) \equiv 0$.

Задача розв'язання цього рівняння розпадається на етапи:

1. Існування та кількість коренів.
2. Відділення коренів, тобто розбиття числової вісі на інтервали, де знаходиться один корінь.
3. Обчислення кореня із заданою точністю ε.

Для розв'язання перших двох задач використовуються методи математичного аналізу та алгебри, а також графічний метод. Далі розглядаються методи розв'язання третього етапу.

2.1. Метод ділення навпіл [СГ, 191], [В, 194-195]

Припустимо на $[a,b]$ знаходиться лише один корінь рівняння

$$f(x) = 0 \qquad (1)$$

для $f(x) \in C[a,b]$, який необхідно визначити. Нехай $f(a)f(b) < 0$. Припустимо, що $f(a) > 0, f(b) < 0$. Покладемо $x_1 = \dfrac{a+b}{2}$ і підрахуємо $f(x_1)$. Якщо $f(x_1) < 0$, тоді шуканий корінь $\bar{x}$ знаходиться на інтервалі (a, x_1). Якщо ж $f(x_1) > 0$, то $\bar{x} \in (x_1, b)$. Далі з двох інтервалів (a, x_1) і (x_1, b) вибираємо той, на границях якого функція $f(x)$ має різні знаки,

знаходимо точку x_2 – середину вибраного інтервалу, підраховуємо $f(x_2)$ і повторюємо вказаний процес.

В результаті отримаємо послідовність інтервалів, що містять шуканий корінь $\bar{x}$, причому довжина кожного послідуючого інтервалу вдвічі менше попереднього.

Цей процес продовжується до тих пір, поки довжина отриманого інтервалу (a_n, b_n) не стане меншою за $b_n - a_n < 2\varepsilon$. Тоді x_{n+1}, як середина інтервалу (a_n, b_n), пов'язане з $\bar{x}$ нерівністю

$$|x_{n+1} - \bar{x}| < \varepsilon . \qquad (2)$$

Ця умова для деякого n буде виконуватись за теоремою Больцано – Коші. Оскільки

$$|b_{k+1} - a_{k+1}| = \frac{1}{2}|b_k - a_k| ,$$

то

$$|x_{n+1} - \bar{x}| \le \frac{1}{2^{n+1}}(b-a) < \varepsilon . \qquad (3)$$

Звідси отримаємо нерівність для обчислення кількості ітерацій n для виконання умови (2):

$$n = n(\varepsilon) \ge \left[\log\left(\frac{b-a}{\varepsilon} \right) \right] + 1 .$$

Степінь збіжності – лінійна, тобто геометричної прогресії з знаменником $q = \frac{1}{2}$.

Переваги методу: простота, надійність. Недоліки методу: низька швидкість збіжності; метод не узагальнюється на системи.

2.2. Метод простої ітерації [СГ, 191-193], [В, 176-183]

Спочатку рівняння

$$f(x) = 0 \qquad (1)$$

замінюється еквівалентним

$$x = \varphi(x) . \qquad (2)$$

Ітераційний процес має вигляді

$$x_{n+1} = \varphi(x_n), \quad n = 0,1,... \qquad (3)$$

Початкове наближення x_0 задається.

Для збіжності велике значення має вибір функції $\varphi(x)$. Перший спосіб заміни рівняння полягає в відділенні змінної з якогось члена рівняння. Більш продуктивним є перехід від рівняння (1) до (2) з функцією $\varphi(x) = x + \tau(x) f(x)$, де $\tau(x)$ – знакостала функція на тому відрізку, де шукаємо корінь.

Кажуть, що ітераційний метод *збігається*, якщо $\lim\limits_{k\to\infty} x_k = \bar{x}$.

Далі $U_r = \{x : |x-a| \le r\}$ відрізок довжини $2r$ з серединою в точці a. З'ясуємо умови , при яких збігається метод простої ітерації.

Теорема 1

Якщо $\max\limits_{x\in[a,b]=U_r} |\varphi'(x)| \le q < 1$, то метод простої ітерації збігається і має місце оцінка

$$|x_n - \bar{x}| \le \frac{q^n}{1-q}|x_0 - x_1| \le \frac{q^n}{1-q}(b-a) \qquad (3)$$

◁ Нехай $x_{k+1}, x_k \in U_r$. Тоді

$$|x_k - x_{k-1}| = |\varphi(x_k) - \varphi(x_{k-1})| = |\varphi'(\xi_k)(x_k - x_{k-1})| \le$$

$$\xi_k = x_k + \theta_k(x_{k+1} - x_k),\ 0 < \theta_k < 1$$

$$\le |\varphi'(\xi_k)| \cdot |x_k - x_{k-1}| \le q\,|x_k - x_{k-1}| = \ldots = q^k |x_1 - x_0|$$

$$|x_{k+p} - x_k| = |x_{k+p} - x_{k+p-1} + \ldots + x_{k+1} - x_k| = |x_{k+p} - x_{k+p-1}| + \ldots + |x_{k+1} - x_k| \le$$

$$\le \left(q^{k+p-1} + q^{k+p-2} + \ldots + q^k\right)|x_1 - x_0| = \frac{q^k - q^{k+p-1}}{1-q}|x_1 - x_0| \xrightarrow[k\to\infty]{} 0 \qquad (4)$$

Бачимо що $\{x_k\}$ – фундаментальна послідовність. Значить вона збіжна. При $p \to \infty$ в (4) отримуємо (3). ▷

Визначимо кількість ітерацій для досягнення точності ε. З оцінки в теоремі 1 отримаємо

$$|x_n - \bar{x}| \le \frac{q^n}{1-q}(b-a) < \varepsilon \Rightarrow \quad n(\varepsilon) = n \ge \left[\frac{\ln\frac{\varepsilon(1-q)}{b-a}}{\ln q}\right] + 1.$$

Практично ітераційний процес зупиняємо при : $|x_n - x_{n-1}| < \varepsilon$. Але ця умова не завжди гарантує, що $|x_n - \bar{x}| < \varepsilon$.

Зауваження Умова збіжності методу може бути замінена на умову Ліпшиця

$$|\varphi(x) - \varphi(y)| \le q\,|x - y|,\ 0 < q < 1.$$

Переваги методу: простота; при $q < \frac{1}{2}$ – швидше збігається ніж метод ділення навпіл; метод узагальнюється на системи. Недоліки методу: 1) при $q > \frac{1}{2}$ – збігається повільніше ніж метод ділення навпіл, 2) виникають труднощі при зведенні $f(x) = 0$ до $x = \varphi(x)$.

2.3. Метод релаксації [СГ, 192-193]

Якщо в методі простої ітерації для рівняння $x = x + \tau f(x) \equiv \varphi(x)$ вибрати $\tau(x) = \tau = const$, то ітераційний процес приймає вигляд

$$x_{n+1} = x_n + \tau f(x_n), \qquad (1)$$

$k = 0,1,2,3\ldots$ x_0 – задано. Метод можна записати у вигляді $\frac{x_{k+1} - x_k}{\tau} = f(x_k)$, $k = 0,1,\ldots$. Оскільки $\varphi'(x) = 1 + \tau f'(x)$, то метод збігається при умові

$$|\varphi'(x)| = |1 + \tau f'(x)| \le q < 1.$$

Нехай $f'(x) < 0$, тоді (3) запишеться у вигляді: $-q \le 1 + \tau f'(x) \le q < 1$. Звідси

$$\tau |f'(x_k)| \le 1 + q < 2 \text{, і } 0 < \tau < \frac{2}{|f'(x)|}.$$

Поставимо задачу знаходження τ, для якого $q = q(\tau) \to \min$. Для того, щоб вибрати оптимальний параметр τ, розглянемо рівняння для похибки $z_k = x_k - \bar{x}$.

Підставивши $x_k = \bar{x} + z_k$ в (1), отримаємо

$$z_{k+1} = z_k + \tau f(\bar{x} + z_k).$$

В припущені $f(x) \in C^1[a,b]$ з теореми про середнє маємо

$$f(\bar{x} + z_k) = f(\bar{x}) + z_k f'(\bar{x} + \theta z_k) = z_k f'(\bar{x} + \theta z_k) = z_k f'(\xi_k)$$

$$z_{k+1} = z_k + \tau f'(\xi_k) \cdot z_k$$

$$|z_{k+1}| \le |1 + \tau f'(\xi_k)| \cdot |z_k| \le \max_U |1 + \tau f'(\xi_k)| |z_k|$$

$$|z_{k+1}| \le \max\{|1 - \tau M_1|, |1 - \tau m_1|\} |z_k|$$

$$m_1 = \min_{[a,b]} |f'(x)| \ , \ \ M_1 = \max_{[a,b]} |f'(x)|$$

Таким чином, задача вибору оптимального параметра зводиться до знаходження τ, для якого функція

$$q(\tau) = \max\{|1 - \tau M_1|, |1 - \tau m_1|\}$$

приймає мінімальне значення: $q(\tau) \to \min$.

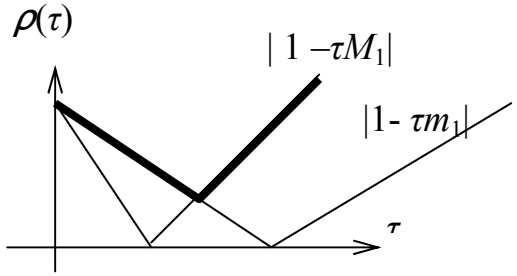


Рис. 1

З графіка видно, що точка мінімуму визначається умовою $|1 - \tau M_1| = |1 - \tau m_1|$. Тому

$$1 - \tau_0 m_1 = \tau_0 M_1 - 1 \Rightarrow \ \tau_0 = \frac{2}{M_1 + m_1} < \frac{2}{|f'(x)|}$$

При цьому значенні τ маємо

$$q(\tau_0) = \rho_0 = \frac{M_1 - m_1}{M_1 + m_1}.$$

Тоді для похибки вірна оцінка

$$|x_n - \bar{x}| \le \frac{(\rho_0)^n}{1 - \rho_0}(b - a) < \varepsilon$$

Кількість ітерацій

$$n = n(\varepsilon) \geq \left[\frac{\ln \varepsilon(1-\rho_0) \Big/ b-a}{\ln \rho_0} \right] + 1$$

Задача 1 Дати геометричну інтерпретацію методу простої ітерації для випадків:

$$0 < \varphi'(x) < 1;\ -1 < \varphi'(x) < 0;\ \varphi'(x) < -1;\ \varphi'(x) > 1.$$

Задача 2 Знайти оптимальне $\tau = \tau_0$ для методу релаксації при $f'(x) > 0$.

2.4. Метод Ньютона (метод дотичних) [СГ, 193-194], [БЖК, 323-324]

Припустимо, що рівняння $f(x) = 0$ має простий дійсний корінь $\bar{x}$, тобто $f(\bar{x}) = 0$, $f'(\bar{x}) \neq 0$. Нехай виконуються умови: $f(x) \in C^1[a,b]$, $f(a) \cdot f(b) < 0$. Тоді

$$0 = f(\bar{x}) = f(x_k + \bar{x} - x_k) = f(x_k) + f'(\xi_k)(\bar{x} - x_k),$$

де $\xi_k = x_k + \theta_k(\bar{x} - x_k)$, $0 < \theta_k < 1$, $\xi_k \approx x_k$. Тому наступне наближення виберемо з рівняння

$$f(x_k) + f'(x_k)(x_{k+1} - x_k) = 0.$$

Звідси маємо ітераційний процес

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},\ k = 0,1,2\ldots\ ;\ x_0\text{-задане}.$$

Метод Ньютона ще називають методом лінеаризації або методом дотичних.

Задача 3 Дати геометричну інтерпретацію методу Ньютона.

Метод Ньютона можна інтерпретувати як метод простої ітерації з

$$\varphi(x) = x - \frac{f(x)}{f'(x)},\ \text{тобто}\ \tau(x) = -\frac{1}{f'(x)}.$$

Тому $\varphi'(x) = 1 - \dfrac{[f'(x)f'(x) - f(x)f''(x)]}{(f'(x))^2} = \dfrac{f(x)f''(x)}{(f'(x))^2}$. Якщо $\bar{x}$ - корінь $f(x)$, то $\varphi'(\bar{x}) = 0$. Тому знайдеться окіл кореня, де

$$|\varphi'(x)| = \left| \frac{f(x)f''(x)}{(f'(x))^2} \right| < 1.$$

Це означає, що збіжність методу Ньютона залежить від вибору x_0.

Недолік методу Ньютона: необхідність обчислювати на кожній ітерації не тільки значення функції, а й похідної.

Модифікований метод Ньютона позбавлений цього недоліку і має вигляд:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)},\ k=0,1,2,\ldots$$

Цей метод має лише лінійну збіжність: $|x_{k+1} - \bar{x}| = O(|x_k - \bar{x}|)$.

Задача 4 Дати геометричну інтерпретацію модифікованого методу Ньютона.

В методі Ньютона, для якого $f'(x_k)$ замінюється на $\dfrac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}}$

дає метод січних:

$$x_{k+1}=x_k-\frac{x_k-x_{k-1}}{f(x_k)-f(x_{k-1})}f(x_k)\quad k=1,2\ldots;\ x_0,x_1\text{- задані}$$

Задача 5 Дати геометричну інтерпретацію методу січних .

2.5. Збіжність методу Ньютона [СГ, 199-203]

Теорема 1 Нехай $f(x)\in C^2[a,b]$; $\bar{x}$ простий дійсний корінь рівняння

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

і $f'(x)\neq 0$ при $x\in U_r=\{x:|x-\bar{x}|<r\}$. Якщо

$$q=\frac{M_2|x_0-\bar{x}|}{2m_1}<1, \tag{2}$$

де $m_1=\min\limits_{U_r}|f'(x)|, M_2=\max\limits_{U_r}|f''(x)|$, то для $x_0\in U_r$ метод Ньютона

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)} \tag{3}$$

збігається і має місце оцінка

$$|x_n-\bar{x}|\le q^{2^n-1}|x_0-\bar{x}|. \tag{4}$$

◁ З (3) маємо

$$x_{k+1}-\bar{x}=x_k-\frac{f(x_k)}{f'(x_k)}-\bar{x}=\frac{(x_k-\bar{x})f'(x_k)-f(x_k)}{f'(x_k)}=\frac{F(x_k)}{f'(x_k)}, \tag{5}$$

де $F(x)=(x-\bar{x})f'(x)-f(x)$, така, що

$$1)\ F(\bar{x})=0;\ 2)\ F'(x)=(x-\bar{x})f''(x).$$

Тоді

$$F(x_k)=F(\bar{x})+\int_{\bar{x}}^{x_k}F'(t)dt=\int_{\bar{x}}^{x_k}(t-\bar{x})f''(t)dt.$$

Так як $(t-\bar{x})$ не міняє знак на відрізку інтегрування, то скористаємося теоремою про середнє значення:

$$F(x_k)=f''(\xi_k)\int_{\bar{x}}^{x_k}(t-\bar{x})dt=\frac{(x_k-\bar{x})^2}{2}f''(\xi_k), \tag{6}$$

де $\xi_k=\bar{x}+\theta_k(x_k-\bar{x}),0<\theta_k<1$. З (5),(6) маємо

$$x_{k+1}-\bar{x}=\frac{(x_k-\bar{x})^2}{2f'(x_k)}f''(\xi_k) \tag{7}$$

Доведемо оцінку (3) за індукцією. Так як $x_0\in U_r$, то

$$|\xi_0-\bar{x}|=|\theta_0(x_0-\bar{x})|<|\theta_0||x_0-\bar{x}|<r\Rightarrow\xi_0\in U_r.$$

Тоді $|f''(\xi_0)|\le M_2$, тому

$$|x_1-\bar{x}|\le\frac{(x_0-\bar{x})^2M_2}{2m_1}=\frac{M_2|x_0-\bar{x}|}{2m_1}|x_0-\bar{x}|=q|x_0-\bar{x}|<r,x_1\in U_r,$$

Ми довели твердження (4) при $n=1$. Нехай воно справджується при $n=k$

$$|x_k-\bar{x}|\le q^{2^k-1}|x_0-\bar{x}|<r,|\xi_k-\bar{x}|=|\theta_k(x_k-\bar{x})|<r.$$

Тоді $x_k,\xi_k\in U_r$.

Доведемо (4) для $n=k+1$. З (7) маємо

$$|x_{k+1}-\bar{x}|\le\frac{|x_k-\bar{x}|^2M_2}{2m_1}\le\left(q^{2^k-1}\right)^2\frac{|x_0-\bar{x}|^2M_2}{2m_1}=$$

$$=q^{2^{k+1}-2}\frac{|x_0-\bar{x}|M_2}{2m_1}|x_0-\bar{x}|=q^{2^{k+1}-1}|x_0-\bar{x}|$$

Таким чином (4) справджується для $n=k+1$. Значить (4) виконується і для довільного n. Таким чином $x_n\xrightarrow[n\to\infty]{}x$.▷

З (4) маємо оцінку кількості ітерацій для досягнення точності ε

$$n\ge\left[\frac{\ln\frac{\varepsilon}{b-a}}{\log_2(1+\frac{}{\ln q})}\right]+1.$$

Кажуть, що ітераційний метод має *степінь збіжності m*, якщо

$$|x_{k+1}-\bar{x}|=O\left(|x_k-\bar{x}|^m\right).$$

Для методу Ньютона $|x_{k+1}-\bar{x}|=\frac{|x_k-\bar{x}|^2|f''(\xi_k)|}{2|f'(x_k)|}\Rightarrow|x_{k+1}-\bar{x}|=O\left(|x_k-\bar{x}|^2\right).$

Значить степінь збіжності методу Ньютона *m*=2. Для методу простої ітерації і ділення навпіл *m*=1.

Теорема 2 Нехай $f(x)\in C^2[a,b]$ та $\bar{x}$ простий корінь рівняння $f(x)=0$ $\left(f'(\bar{x})\ne0\right)$ $\forall x\in[a,b]$. Якщо $f'(x)f''(x)>0$ $\left(f'(x)f''(x)<0\right)$ то для методу Ньютона при $x_0=b$ послідовність наближень $\{x_k\}$ монотонно спадає (монотонно зростає при $x_0=a$).

Задача 6 Довести теорему 2 при а) $f'(x)f''(x)>0$, б) $f'(x)f''(x)<0$.

Задача 7 Знайти степінь збіжності методу січних [Калиткин Н.Н., Численные методы, с. 145-146]

Якщо $f(a)f''(a)>0$ та $f''(x)$ не міняє знак, то потрібно вибирати $x_0=a$; при цьому $\{x_k\}\uparrow\bar{x}$.

Якщо $f(b)f''(b)>0$, то $x_0=b$; маємо $\{x_k\}\downarrow\bar{x}$. Пояснення на рисунку 2.

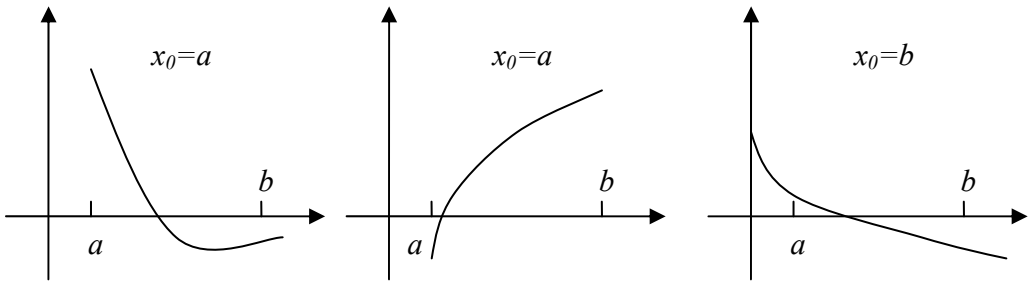


Рис. 2

Зауваження 1 Якщо $\bar{x}$ p-кратний корінь тобто

$$f^{(m)}(\bar{x}) = 0, m = 0,1,..., p-1; f^{(p)}(\bar{x}) \neq 0,$$

то в методі Ньютона необхідна наступна модифікація

$$x_{k+1} = x_k - p\frac{f(x_k)}{f'(x_k)} \text{ і } q = \frac{M_{p+1}\left|x_0 - \bar{x}\right|}{m_p p(p+1)} < 1.$$

Зауваження 2 Метод Ньютона можна застосовувати і для обчислення комплексного кореня $z_{k+1} = z_k - \frac{f(z_k)}{f'(z_k)}$. В теоремі про збіжність $q = \frac{\left|x_0 - \bar{x}\right| M_2}{2m_1}$, де $m_1 = \min_{U_r}\left|f'(z)\right|, M_2 = \max_{U_r}\left|f''(z)\right|$. Тут $\left|z\right|$ - модуль комплексного числа.

Переваги методу Ньютона**:** 1) висока швидкість збіжності; 2) узагальнюється на системи рівнянь; 3) узагальнюється на комплексні корені.

Недоліки методу Ньютона: 1) на кожній ітерації обчислюється не тільки $f(x_k)$, а і похідна $f'(x_k)$; 2) збіжність залежить від початкового наближення x_0, так як від нього залежить умова збіжності $q = \frac{M_2\left|x_0 - \bar{x}\right|}{2m_1} < 1$; 3) потрібно, щоб $f(x) \in C^2[a,b]$.

3. Методи розв'язання систем лінійних алгебраїчних рівнянь (СЛАР)

Методи розв'язування СЛАР поділяються на прямі та ітераційні. При умові точного виконання обчислень прямі методи за скінчену кількість операцій в результаті дають точний розв'язок. Використовуються вони для невеликих та середніх СЛАР n=10^2-10^4. Ітераційні методи використовуються для великих СЛАР n>10^5, як правило розріджених. В результаті отримуємо послідовність наближень, яка збігається до розв'язку.

3.1. Метод Гаусса [СГ, 49-67], [БЖК, 257-262]

Розглянемо задачу розв'язання СЛАР

$$A\vec{x} = \vec{b}, \quad (1)$$

причому $A=(a_{ij})_{i,j=1}^{n}, \det A\neq 0, \vec{x}=(x_i)_{i=1}^{n}, \vec{b}=(b_j)_{j=1}^{n}$. Метод Крамера з обчисленням визначників для такої системи має складність $Q=O(n!n)$.

Запишемо СЛАР у вигляді

$$\begin{cases} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=b_1\equiv a_{1n+1} \\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_x=b_2\equiv a_{2n+1} \\ \ldots \\ a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n=b_n\equiv a_{nn+1} \end{cases} \quad (1')$$

Якщо $a_{11}\neq 0$, то ділимо перше рівняння на нього і виключаємо x_1 з інших рівнянь:

$$\begin{cases} x_1+a_{12}^{(1)}x_2+\ldots+a_{1n}^{(1)}x_n=a_{1n+1}^{(1)} \\ \quad a_{22}^{(1)}x_2+\ldots+a_{2n}^{(1)}x_n=a_{2n+1}^{(1)} \\ \qquad\qquad\qquad\qquad \ldots \\ \quad a_{n2}^{(1)}x_2+\ldots+a_{nn}^{(1)}x_n=a_{nn+1}^{(1)} \end{cases}.$$

Процес повторюємо для $x_2,\ldots,x_n$. В результаті отримуємо систему з трикутною матрицею

$$\begin{cases} x_1+a_{12}^{(1)}x_2+\ldots+a_{1n}^{(1)}x_n=a_{1n+1}^{(1)} \\ \qquad x_2+\ldots+a_{2n}^{(2)}x_n=a_{2n+1}^{(2)} \\ \qquad\qquad\qquad \ldots \\ \qquad\qquad x_n=a_{nn+1}^{(n)} \end{cases}. \quad (2)$$

Це прямий хід методу Гаусса. Формули прямого ходу

$$\begin{cases} k=\overline{1,n-1}: \\ a_{kj}^{(k)}=\dfrac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}}, j=\overline{k+1,n+1}; \\ a_{ij}^{(k)}=a_{ij}^{(k-1)}-a_{ik}^{(k-1)}a_{kj}^{(k)}, \\ i=\overline{k+1,n}; j=\overline{k+1,n+1}. \end{cases}$$

Звідси

$$x_n=a_{nn+1}^{(n)}, x_i=a_{in+1}^{(i)}-\sum_{j=i+1}^{n}a_{ij}^{(i)}x_j, i=\overline{n-1,1}. \quad (3)$$

Це формули оберненого ходу.

Складність, тобто кількість операцій, яку необхідно виконати для реалізації методу, - $Q_{пр}=\frac{2}{3}n^3+O(n^2)$ для прямого ходу, $Q_{об}=n^2+O(n)$ для оберненого ходу.

Умова $a_{kk}^{(k-1)} \neq 0$ не суттєва, оскільки знайдеться m, для якого $\left|a_{mk}^{(k-1)}\right| = \max\limits_{i}\left|a_{ik}^{(k-1)}\right| \neq 0$ (оскільки $\det A \neq 0$).Тоді міняємо місцями рядки номерів k і m. Елемент $a_{kk}^{(k-1)} \neq 0$ називається ведучим.

Введемо матриці

$$M_k = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ 0 & & m_{kk} & & \\ & & \vdots & \ddots & \\ 0 & & m_{nk} & & 1 \end{pmatrix},$$

елементи якої обчислюється так: $m_{kk} = \dfrac{1}{a_{kk}^{(k-1)}}$, $m_{kk} = -\dfrac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$.

Нехай на k –му кроці $A_{k-1}\vec{x} = \vec{b}_{k-1}$. Множимо цю СЛАР зліва на M_k: $M_k A_{k-1}\vec{x} = M_k \vec{b}_{k-1}$. Позначимо $A_k = M_k A_{k-1}$; $A_0 = A$. Тоді прямий хід методу Гаусса можна записати у вигляді

$$M_n M_{n-1} \ldots M_1 A\vec{x} = M_n M_{n-1} \ldots M_1 \vec{b}.$$

Позначимо останню систему, яка співпадає з (2), так

$$U\vec{x} = \vec{c},\ U = (u_{ij})_{i,j=1}^{n}, \qquad (3)$$

причому

$$\begin{cases} u_{ii} = 1, \\ u_{ij} = 0, i > j. \end{cases}$$

Таким чином $U = M_n M_{n-1} \ldots M_1 A$. Введемо матриці

$$L_k = M_k^{-1} = \begin{pmatrix} 1 & & 0 & & 0 \\ & \ddots & & & \\ 0 & & a_{kk}^{(k-1)} & & 0 \\ & & \vdots & \ddots & \\ 0 & & a_{nk}^{(k-1)} & & 1 \end{pmatrix}.$$

Тоді

$$A = L_1 \ldots L_n U = LU;\ \ L = L_1 \ldots L_n$$

L- нижня трикутня матриця, U- верхня трикутня матриця. Таким чином метод Гаусса можна трактувати, як розклад матриці A в добуток двох трикутних матриць - (LU) – розклад

Введемо матрицю перестановок на k-му кроці (це матриця, отримана з одиничної матриці перестановкою k - того і m- того рядка). Тоді при множені на неї матриці A_{k-1} робимо ведучим елементом максимальний за модулем.

$$P_k = \begin{pmatrix} 1 & & \dots & & 0 \\ . & . & . & & \\ . & & 0 & 1 & \\ . & & 1 & 0 & \\ 0 & . & . & . & 1 \\ & & & & \end{pmatrix} \begin{matrix} \\ \\ \leftarrow k \\ \leftarrow m \\ \\ \\ \end{matrix} .$$

$$\begin{matrix} & & \uparrow & \uparrow \\ & & k & m \end{matrix}$$

За допомогою цих матриць перехід до трикутної системи (3) тепер має вигляд:

$$M_n M_{n-1} P_{n-1} \dots M_1 P_1 A\vec{x} = M_n M_{n-1} P_{n-1} \dots M_1 P_1 \vec{b} .$$

Твердження

Знайдеться така матриця P - перестановок, що $PA = LU$ - розклад матриці на нижню трикутну з ненульовими діагональними елементами і верхню трикутну матрицю з одиницями на діагоналі.

bred!!!

Висновки про переваги трикутного розкладу:

1. Розділення прямого і оберненого ходів дає змогу економно розв‘язувати декілька систем з одноковою матрицею та різними правими частинами.
2. Зберігання M, або L та U на місці A.
3. Обчислюючи l - кількість перестановок, можна встановити знак визначника.

i v chem + ???

3.2. Метод квадратних коренів [СГ, 69-73], [БЖК, 262-263]

Цей метод призначений для розв’язання систем рівнянь із симетричною матрицею

$$A\vec{x} = \vec{b} , A^T = A . \qquad (1)$$

Він оснований на розкладі матриці A в добуток:

$$A = S^T D S , \qquad (2)$$

S – верхня трикутна матриця, S^T – нижня трикутна матриця, D – діагональна матриця.

Виникає питання: як обчислити S, D по матриці A? Маємо

$$(DS)_{ij} = \begin{cases} d_{ii} s_{ij}, i \le j \\ 0, i > j \end{cases}$$

$$(S^T DS)_{ij} = \sum_{l=1}^{n} s_{il}^T d_{ll} s_{lj} = \sum_{l=1}^{i-1} s_{li}^T s_{lj} d_{ll} + s_{ii} s_{ij} d_{ii} + \underbrace{s_{li}^T \sum_{l=i+1}^{n} s^T{}_{li} s_{lj} d_{ll}}_{=0} = a_{ij}, i, j = \overline{1,n} \qquad (3)$$

Якщо $i = j$, то

$$\left| s_{ii}^2 \right| d_{ii} = a_{ii} - \sum_{l=1}^{i-1} \left| s_{li}^2 \right| d_{ll} \equiv p_i .$$

Тому

$$d_{ii} = sign\ (p_i) \quad , \quad s_{ii} = \sqrt{|p_i|}\ .$$

Якщо $i < j$, то

$$s_{ij} = \left(a_{ij} - \sum_{l=1}^{i-1} s_{li}^T d_{ll} s_{lj} \right) / (s_{ii} d_{ii}),\ i = \overline{1,n},\ j = \overline{i+1,n}\ .$$

Якщо A>0 (тобто головні мінори матриці A додатні), то всі $d_{ii} = +1$.

Знайдемо розв'язок рівняння (1). Враховуючи (2), маємо:

$$S^T D\vec{y} = \vec{b}\ , \tag{4}$$

$$S\vec{x} = \vec{y}\ . \tag{5}$$

Оскільки S – верхня трикутна матриця, а $S^T D$ - нижня трикутна матриця, то

$$y_i = \frac{b_i - \sum_{j=1}^{i-1} s_{ji} d_{jj} y_j}{s_{ii} d_{ii}}, \quad i = \overline{1,n} \tag{6}$$

nepravilnyj poryadok

$$x_n = \frac{y_n}{s_{nn}}, x_i = \frac{y_i - \sum_{j=1}^{i-1} s_{ij} x_j}{s_{ii}}, \quad i = \overline{n-1,1}. \tag{7}$$

Метод застосовується лише для симетричних матриць. Його складність - $Q = \frac{1}{3} n^3 + O(n^2)$. Переваги цього методу:

1. він витрачає в 2 рази менше пам'яті ніж метод Гаусса для зберігання $A^T = A$ (необхідний об'єм пам'яті $\frac{n(n+1)}{2} \sim \frac{n^2}{2}$);
2. метод однорідний, без перестановок;
3. якщо матриця A має багато нульових елементів, то і в матриця S також.

Остання властивість дає економію в пам'яті та кількості арифметичних операцій. Наприклад, якщо A має m ненульових стрічок по діагоналі, то $Q = O(m^2 n)$.

3.3. Обчислення визначника та оберненої матриці [СГ, 67-69]

Кількість операцій обчислення детермінанту за означенням - $Q_{\det} = n!$. В методі Гаусса - $PA = LU$. Тому

$$\det P \det A = \det L \det U \Rightarrow \det A = (-1)^l \det L \det U = (-1)^l \prod_{k=1}^{n} a_{kk}^{(k)}\ , \tag{1}$$

де l - кількість перестановок. Ясно, що за методом Гаусса

$$Q_{\det} = \frac{2}{3} n^3 + O(n^2)\ .$$

В методі квадратного кореня $A=S^TDS$. Тому

$$\det A = \det S^T \det D \det S = \prod_{k=1}^{n} d_{kk} \prod_{k=1}^{n} s_{kk}^2 . \qquad (2)$$

Тепер $Q_{\det} = \frac{1}{3} n^3 + O(n^2)$.

За означенням

$$AA^{-1} = E , \qquad (3)$$

де A^{-1} обернена до матриці A. Позначимо

$$A^{-1} = (\alpha_{ij})_{i,j=1}^{n} .$$

Тоді $\vec{\alpha}_j = (\alpha_{ij})_{i=1}^{n}$ - вектор-стовпчик оберненої матриці. З (3) маємо

$$A\vec{\alpha}_j = \vec{e}_j , j = \overline{1,n} \qquad (4)$$

$\vec{e}_j$ - стовпчики одиничної матриці: $\vec{e}_j = (\delta_{ij})_{i=1}^{n}, \delta_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$. Для знаходження

i chto dolshe?

A^{-1} необхідно розв'язати n систем . Для знаходження A^{-1} методом Гаусса необхідна кількість операцій $Q = 2n^3 + O(n^2)$.

3.4. Метод прогонки [СГ, 45-47]

Це економний метод для розв'язання СЛАР з три діагональною матрицею:

$$\begin{cases} -c_0 y_0 + b_0 y_1 = -f_0 , \\ a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i , \\ a_N y_{N-1} - c_N y_N = -f_N . \end{cases}$$

(1)

(2)

(3)

Матриця системи

$$A = \begin{pmatrix} -c_0 & b_1 & & 0 \\ a_0 & -c_1 & \ddots & \\ & \ddots & \ddots & b_N \\ 0 & & a_N & -c_N \end{pmatrix}$$

тридіагональна.

Розв'язок представимо у вигляді

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, i = \overline{0, N-1} \qquad (4)$$

Замінимо в (4) $i \to i-1$ і підставимо в (2), тоді

$$(a_i \alpha_i - c_i) y_i + b_i y_{i+1} = -f_i - a_i \beta_i$$

Звідси

$$y_i = \frac{b_i}{c_i - a_i\alpha_i} y_{i+1} + \frac{f_i + a_i\beta_i}{c_i - a_i\alpha_i}.$$

Тому з (5)

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i\alpha_i}, \beta_{i+1} = \frac{f_i + a_i\beta_i}{c_i - a_i\alpha_i}, i = \overline{1, N-1}.$$

Умова розв‘язності (1) $c_i - a_i\alpha_i \neq 0$.

Щоб знайти всі α_i, β_i, треба задати перші значення. З (1):

$$\alpha_1 = \frac{b_0}{c_0}, \beta_1 = \frac{f_0}{c_0} \tag{5}$$

Після знаходження всіх α_i, β_i обчислюємо y_N з системи

$$\begin{cases} a_N y_N - c_N y_N = -f_N \\ y_{N-1} = \alpha_N y_N + \beta_N \end{cases}.$$

Звідси

$$y_N = \frac{f_N + a_N\beta_N}{c_N - a_N\alpha_N}. \tag{6}$$

Алгоритм

1. Покладемо $\alpha_1 = \frac{b_0}{c_0}, \beta_1 = \frac{f_0}{c_0}$.
2. Позначимо $z_i = c_i - a_i\alpha_i$, обчислимо $\alpha_{i+1} = \frac{b_i}{z_i}, \beta_{i+1} = \frac{f_i + a_i\beta_i}{z_i}$, для $i = \overline{1, N-1}$
3. Знайдемо $y_N = \frac{f_N + a_N\beta_N}{c_N - a_N\alpha_N}$.
4. Обчислюємо $y_i = \alpha_{i+1}y_{i+1} + \beta_{i+1}, i = \overline{N-1, 0}$.

Складність алгоритму $Q = 8N - 2$.

Метод можна застосовувати, коли $c_i - a_i\alpha_i \neq 0 \; \forall i, |\alpha_i| \leq 1$. Якщо $|\alpha_i| \geq q > 1$, то $\Delta y_0 \geq q^N \Delta y_N$ (тут Δy_i абсолютна похибка обчислення y_i), а це приводить до експоненціального накопичення похибок заокруглення, тобто нестійкості алгоритму прогонки.

Теорема (про достатні умови стійкості метода прогонки).

Нехай

$$a_i, b_i \neq 0, \text{ та } |c_i| \geq |a_i| + |b_i|, \; \forall i, (a_0 = b_N = 0)$$

та хоча би одна нерівність строга. Тоді

$$|\alpha_i| \leq 1, \text{ та } z_i = c_i - a_i\alpha_i \neq 0, i = \overline{1, N}.$$

etu teoremu?

Задача 8 Довести теорему про стійкість методу прогонки.

3.5. Обумовленість систем лінійних алгебраїчних рівнянь [СГ, 74-81], [БЖК, 303-308]

Нехай задано СЛАР

$$A\vec{x} = \vec{b} \qquad (1)$$

Припустимо, що матриця і права частина системи задані неточно і фактично розв'язуємо систему

$$B\vec{y} = \vec{h} \qquad (2)$$

де $B = A + C, \vec{h} = \vec{b} + \vec{\eta}, \vec{y} = \vec{x} + \vec{z}$.

Малість детермінанту $|\det A| << 1$ не є необхідною умовою різкого збільшення похибки. Це ілюструє наступний приклад:

$$A = diag(\varepsilon), a_{ij} = \varepsilon\ \delta_{ij},$$

Тоді $\det A = \varepsilon^n << 1$, але $x_i = \frac{b_i}{\varepsilon}$. Тому $\Delta x_i = \frac{\Delta b_i}{\varepsilon}$.

Оцінимо похибку розв'язку. Підставивши значення $B, \vec{h}$ та $\vec{z} = \vec{y} - \vec{x}$, отримаємо:

$$(A + C)(\vec{x} + \vec{z}) = \vec{b} + \vec{\eta}.$$

Віднімемо від цієї рівності (1) $A\vec{z} + C\vec{x} + C\vec{z} = \vec{\eta}$. Тоді

$$A\vec{z} = \vec{\eta} - C\vec{x} - C\vec{z}, \vec{z} = A^{-1}(\vec{\eta} - C\vec{x} - C\vec{z}).$$

Введемо норми векторів: $\|\vec{z}\|$:

$$\|\vec{z}\|_1 = \sum_{i=1}^{n} |z_i|, \quad \|\vec{z}\|_2 = (\sum_{i=1}^{n} |z_i|^2)^{\frac{1}{2}}, \quad \|\vec{z}\|_\infty = \max_i |z_i|.$$

Норми матриці, що відповідають нормам вектора, тобто

$$\|A\|_m = \sup_{\|\vec{x}\|_m \neq 0} \frac{\|A\vec{x}\|_m}{\|\vec{x}\|_m}, m = 1,2,\infty,$$

такі: $\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|$, $\|A\|_2 = \max_i \sqrt{\lambda_i(A^T A)}$, $\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|$, де $\lambda_i(B)$ – власні значення матриці B.

Позначимо $\delta(\vec{x}) = \frac{\|\vec{z}\|}{\|\vec{x}\|}$, $\delta(\vec{b}) = \frac{\|\vec{\eta}\|}{\|\vec{b}\|}$, $\delta(A) = \frac{\|C\|}{\|A\|}$ - відносні похибки $\vec{x}, \vec{b}, A$, де $\|.\| = \|.\|_k$ одна з введених вище норм.

Для характеристики зв'язку між похибками правої частини та розв'язку вводять поняття обумовленості матриці системи.

Число обумовленості матриці A - $cond(A) = \|A\|A^{-1}\|$.

Теорема Якщо $\exists A^{-1}$ та $\|A^{-1}\|C\| < 1$, то

$$\delta(\vec{x}) \le \frac{condA}{1 - condA\,\delta(A)}(\delta(A) + \delta(\vec{b})), \qquad (3)$$

де $condA$ - число обумовленості.

$\triangleleft$ $A\vec{z} = \vec{\eta} - C\vec{x} - C\vec{z}$, $\vec{z} = A^{-1}\vec{\eta} - A^{-1}C\vec{x} - A^{-1}C\vec{z}$;

$$\|\vec{z}\| \le \left\|A^{-1}\vec{\eta}\right\| + \left\|A^{-1}C\vec{x}\right\| + \left\|A^{-1}C\vec{z}\right\| \le \left\|A^{-1}\right\|\|\vec{\eta}\| + \left\|A^{-1}\right\|\|C\|\|\vec{x}\| + \left\|A^{-1}\right\|\|C\|\|\vec{z}\|$$

$$\|\vec{z}\| \le \frac{\left\|A^{-1}\right\|\left(\vec{\eta} + \|C\|\|\vec{x}\|\right)}{1 - \left\|A^{-1}\right\|\|C\|}$$

Оцінка похибки

$$\delta(\vec{x}) \le \frac{\left\|A^{-1}\right\|}{1 - \left\|A^{-1}\right\|\|C\|}\left(\frac{\|\vec{\eta}\|}{\|\vec{x}\|} + \|C\|\right) = \frac{\left\|A^{-1}\right\|\|A\|}{1 - \left\|A^{-1}\right\|\|A\|\dfrac{\|C\|}{\|A\|}}\left(\frac{\|\vec{\eta}\|}{\|A\|\|\vec{x}\|} + \delta(A)\right) \le$$

$$\le \frac{condA}{1 - condA\,\delta(A)}\left(\frac{\|\vec{\eta}\|}{\|\vec{x}\|} + \delta(A)\right). \triangleright$$

Наслідок Якщо $C \equiv 0$, то $\delta(\vec{x}) \le condA\,\delta(\vec{b})$.

Властивості $condA$:

1^0. $condA \ge 1$;

2^0. $condA \ge \dfrac{\max\limits_i |\lambda_i(A)|}{\min\limits_i |\lambda_i(A)|}$;

3^0. $cond(AB) \le condA * condB$;

4^0. $A^T = A^{-1} \Rightarrow condA = 1$.

Друга властивість має місце оскільки довільна норма матриці не менше її найбільшого за модулем власного значення. Значить $\|A\| \ge \max|\lambda_A|$. Оскільки власні значення матриць A^{-1} та A взаємно обернені, то

$$\left\|A^{-1}\right\| \ge \max\frac{1}{|\lambda_A|} = \frac{1}{\min|\lambda_A|}.$$

Якщо $condA >> 1$, то система називається *погано обумовленою.*

Оцінка впливу похибок заокруглення при обчисленні розв‘язку СЛАР така (Дж. Уілкінсон): $\delta(A) = O(n\beta^{-t})$, $\delta(\vec{b}) = O(\beta^{-t})$, де β - розрядність ЕОМ, t - кількість розрядів, що відводиться під мантису числа. З оцінки (3) витікає: $\delta(\vec{x}) = condA \times O(n\beta^{-t})$. Висновок: найпростіший спосіб підвищити точність обчислення розв‘язку погано обумовленої СЛАР – збільшити розрядність ЕОМ при обчисленнях. Інші способи пов‘язані з розглядом цієї СЛАР як некоректної задачі із застосуванням відповідних методів її розв‘язання (див. [БЖК, с. 306]).

Приклад погано обумовленої системи – системи з матрицею Гільберта

$$H_n = \left(\frac{1}{i + j - 1}\right)_{i,j=1}^{n}, \text{ наприклад } condH_8 \approx 10^9.$$

4. Ітераційні методи для систем рівнянь

4.1. Ітераційні методи розв'язання СЛАР [СГ, 82-86], [БЖК, 269-270, 287-290]

Систему

$$A\vec{x} = \vec{b} \tag{1}$$

зводимо до вигляду

$$\vec{x} = B\vec{x} + \vec{f} \tag{2}$$

Будь яка система

$$\vec{x} = \vec{x} - C\left(A\vec{x} - \vec{b}\right) \tag{3}$$

має вигляд (2) і при $\det C \neq 0$ еквівалентна системі (1). Наприклад, для $C = \tau E$

$$\vec{x} = \vec{x} - \tau\left(A\vec{x} - \vec{b}\right) \tag{3'}$$

4.1.1. Метод простої ітерації (м.п.і.)

Цей метод застосовується до рівняння (2)

$$\vec{x}^{k+1} = B\vec{x}^{k} + \vec{f}\,, \tag{4}$$

$\vec{x}^{0}$ –початкове наближення задано. Ітераційний процес збігається, тобто $\left\|\vec{x}^{k} - \vec{x}\right\| \to 0,\ k \to \infty$, якщо

$$\|B\| \le q < 1 \tag{5}$$

При цьому має місце оцінка

$$\left\|\vec{x}^{n} - \vec{x}\right\| \le \frac{q^{n}}{1-q}\left\|\vec{x}^{1} - \vec{x}^{0}\right\| \tag{6}$$

4.1.2. Метод Якобі

Припустимо $a_{ii} \neq 0, \forall i$. Зведемо систему (1) до вигляду

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

Ітераційний процес запишемо у вигляді

$$x_i^{k+1} = -\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}x_j^{k} - \sum_{j=i+1}^{n}\frac{a_{ij}}{a_{ii}}x_j^{k} + \frac{b_i}{a_{ii}}\ \ k = 0,1....,i = \overline{1,n} \tag{7}$$

Ітераційний процес збігається до розв'язку, якщо виконується умова

$$\sum_{\substack{j=1\\ i\neq j}}^{n}\left|a_{ij}\right| \le \left|a_{ii}\right|, \forall i = \overline{1,n}$$

Це умова діагональної переваги матриці A. Якщо ж

$$\sum_{\substack{j=1\\ i\neq j}}^{n}\left|a_{ij}\right| \le q\cdot\left|a_{ii}\right|, \forall i = \overline{1,n}, 0 \le q < 1\,, \tag{8}$$

то має місце оцінка точності:

$$\left\|\vec{x}^{\,n}-\vec{x}\right\| \le \frac{q^n}{1-q}\left\|\vec{x}^{\,1}-\vec{x}^{\,0}\right\|.$$

4.1.3. Метод Зейделя

В компонентному вигляді ітераційний метод Зейделя записується так:

$$x_j^{k+1} = -\sum_{j=1}^{i-1}\frac{a_{ij}}{a_{ii}}x_j^{k+1} - \sum_{j=i+1}^{n}\frac{a_{ij}}{a_{ii}}x_j^{k} + \frac{b_i}{a_{ii}} \qquad (9)$$

На відміну від методу Якобі на *k*-м*у*-кроці попередні компоненти розв‘язку беруться з *k*+1- ої ітерації.

Достатня умова збіжності методу Зейделя - $A^T = A > 0$.

4.1.4. Матрична інтерпретація методів Якобі і Зейделя

Подамо матрицю A у вигляді

$$A = A_1 + D + A_2,$$

де A_1 - нижній трикутник матриці A, A_2 - верхній трикутник матриці A, D - її діагональ. Тоді систему (1) запишемо у вигляді

$$D\vec{x} = A_1\vec{x} + A_2\vec{x} + \vec{b}$$

або

$$\vec{x} = D^{-1}A_1\vec{x} + D^{-1}A_2\vec{x} + D^{-1}\vec{b}$$

Матричний запис методу Якобі:

$$\vec{x}^{\,k+1} = -D^{-1}(A_1 + A_2)\vec{x}^{\,k} + D^{-1}\vec{b}\,;$$

методу Зейделя:

$$\vec{x}^{\,k+1} = -D^{-1}A_1\vec{x}^{\,k+1} - D^{-1}A_2\vec{x}^{\,k} + D^{-1}\vec{b} \quad \text{або} \quad (D + A_1)\vec{x}^{\,k+1} = -A_2\vec{x}^{\,k} + \vec{b}\,.$$

Необхідна і достатня умова збіжності методу Якобі:

- всі корені рівняння $\det\left(D + \left(A_1 + A_2\right)\lambda\right) = 0$ по модулю більше 1.

Необхідна і достатня умова збіжності методу метода Зейделя:

- всі корені рівняння $\det\left(A_1 + D + A_2\lambda\right) = 0$ по модулю більше 1.

4.1.5. Однокрокові (двошарові) ітераційні методи

Канонічною формою однокрокового ітераційного методу розв’язку СЛАР є його запис у вигляді

$$B_k\frac{\vec{x}^{\,k+1} - \vec{x}^{\,k}}{\tau_{k+1}} + A\vec{x}^{\,k} = \vec{b} \qquad (10)$$

Тут $\{B_k\}$ - послідовність матриць (пере обумовлюючі матриці), що задають ітераційний метод на кожному кроці; $\{\tau_{k+1}\}$ - ітераційні параметри.

Якщо $B_k = E$, то ітераційний процес називається *явним*

$$x^{k+1} = x^k - \tau_{k+1}\left(Ax^k + b\right).$$

Якщо $B_k \neq E$, то ітераційний процес називається *неявним*

$$B_k\vec{x}^{\,k+1} = \vec{F}^{\,k}.$$

В цьому випадку на кожній ітерації необхідно розв‘язувати СЛАР.

Якщо $\tau_{k+1} \equiv \tau$, $B_k \equiv B$, то ітераційний процес називається *стаціонарним;* інакше – *нестаціонарним*.

Методам, що розглянуті вище відповідають:

методу простої ітерації (3′): $B_k = E, \tau_{k+1} = \tau$;

методу Якобі: $B_k = D, \tau_{k+1} = 1$;

методу Зейделя: $B_k = D + A_1, \tau_{k+1} = 1$.

4.1.6. Збіжності стаціонарних ітераційних процесів у випадку симетричних матриць

Розглянемо випадок симетричних матриць $A^T = A$ і стаціонарний ітераційний процес $B_k \equiv E, \tau_{k+1} \equiv \tau$

Нехай для A справедливі нерівності

$$\gamma_1 E \le A \le \gamma_2 E,\ \gamma_1, \gamma_2 > 0 \qquad (11)$$

Тоді при виборі $\tau = \tau_0 = \dfrac{2}{\gamma_1 + \gamma_2}$ ітераційний процес збігається. Найбільш точним значенням γ_1, γ_2 при яких виконуються обмеження (11) є $\gamma_1 = \min_i \lambda_i(A), \gamma_2 = \max_i \lambda_i(A)$. Тоді $q = q_0 = \dfrac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} = \dfrac{1-\xi}{1+\xi}, \xi = \dfrac{\gamma_1}{\gamma_2}$

(Зауважимо, що аналогічно обчислюється q і для методу релаксації розв‘язання нелінійних рівнянь, де $\gamma_1 = m_1 = \min|f'(x)|$, $\gamma_2 = M_1 = \max|f'(x)|$) і справедлива оцінка

$$\left\|\vec{x}^n - \vec{x}\right\| \le \frac{q^n}{1-q}\left\|\vec{x}^0 - \vec{x}\right\|.$$

Явний метод з багатьма параметрами $\{\tau_k\}$:

$$B \equiv E,\ \{\tau_k\}: \min_{\tau} q(\tau)\ \ n = n(\varepsilon) \to \min,$$

які обчислюються за допомогою нулів багаточлена Чебишова, називаються ітераційним методом з чебишевським набором параметрів.

4.1.7. Метод верхньої релаксації

Узагальненням методу Зейделя є метод верхньої релаксації: $(D + \omega A_1)\dfrac{\vec{x}^{k+1} + \vec{x}^k}{\omega} + A\vec{x}^k = \vec{b}$,

де D - діагональна матриця з елементами a_{ii} по діагоналі. $\omega > 0$ - заданий числовий параметр.

Тепер $B = D + \omega A_1, \tau = \omega$. Якщо $A^T = A > 0$, то метод верхньої релаксації збігається при умові $0 < \omega < 2$. Параметр підбирається експериментально з умови мінімальної кількості ітерацій.

4.1.8. Методи варіаційного типу

До цих методів відносяться: метод мінімальних нев'язок, метод мінімальних поправок, метод найшвидшого спуску, метод спряжених градієнтів. Вони дозволяють обчислювати наближення без використання апріорної інформації про γ_1, γ_2 в (11).

Нехай $B = E$. Для методу мінімальних нев'язок параметри τ_{k+1} обчислюються з умови

$$\left\|\vec{r}^{\,k+1}\right\|^2 = \left\|\vec{r}^{\,k}\right\|^2 - 2\tau_{k+1}\left(\vec{r}^{\,k}, A\vec{r}^{\,k}\right) + \tau_{k+1}^2 \left\|A\vec{r}^{\,k}\right\|^2 \to \min .$$

Тому

$$\tau_{k+1} = \frac{\left(A\vec{r}^{\,k}, \vec{r}^{\,k}\right)}{\left\|\vec{r}^{\,k}\right\|^2} \text{ де } \vec{r}^{\,k} = A\vec{x}^{\,k} - \vec{b} \text{ - нев'язка.}$$

Умова для завершення ітераційного процесу:

$$\left\|\vec{r}^{\,n}\right\| < \varepsilon .$$

Швидкість збіжності цього методу співпадає із швидкістю методу простої ітерації з одним оптимальним параметром $\tau_0 = \dfrac{2}{\gamma_1 + \gamma_2}$.

Аналогічно будуються методи з $B \neq E$. Матриця B називається переобумовлювачем і дозволяє підвищити швидкість збіжності ітераційного процесу. Його вибирають з умов а) легко розв'язувати СЛАР $B\vec{x}^{\,k} = \vec{F}_k$ (діагональний, трикутній, добуток трикутніх та інше); б) зменшення числа обумовленості матриці $B^{-1}A$ у порівнянні з A.

4.2. Методи розв'язання нелінійних систем [СГ, 209-211]

Розглянемо систему рівнянь:

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

Перепишемо її у векторному вигляді :

$$\vec{f}(\vec{x}) = 0 \qquad (1)$$

4.2.1. Метод простої ітерації

В цьому методі рівняння (1) зводиться до еквівалентного вигляду

$$\vec{x} = \vec{\Phi}(\vec{x}) \qquad (2)$$

Ітераційний процес представимо у вигляді:

$$\vec{x}^{\,k+1} = \vec{\Phi}(\vec{x}^{\,k}), \qquad (3)$$

початкове наближення $\vec{x}^{\,0}$ – задано.

Нехай оператор Φ визначений на множині H. За теоремою про стискуючі відображення ітераційний процес (3) сходиться, якщо виконується умова

$$\left\|\vec{\Phi}(\vec{x})-\vec{\Phi}(\vec{y})\right\|\le q\left\|\vec{x}-\vec{y}\right\|,\quad 0<q<1, \tag{4}$$

або

$$\left\|\Phi'(\vec{x})\right\|\le q<1 \tag{5}$$

$\vec{x}\in U_r$ $\Phi'(x)=\left(\dfrac{\partial\varphi_i}{\partial x_j}\right)_{i,j=1}^{n}$. Для похибки справедлива оцінка

$$\left\|\vec{x}^m-\vec{x}\right\|\le\frac{q^n}{1-q}\left\|\vec{x}^1-\vec{x}^0\right\|.$$

Частинним випадком методу простої ітерації є метод релаксації для рівняння (1)

$$\vec{x}^{k+1}=\vec{x}^k-\tau\vec{F}(\vec{x}^k)\ ,$$

де $\tau<\dfrac{2}{\left\|F'(\vec{x})\right\|}$.

4.2.2. Метод Ньютона

Розглянемо рівняння

$$\vec{F}(\vec{x})=0.$$

Представимо його у вигляді

$$\vec{F}(\vec{x}^k)+F'(\vec{\xi}^k)(\vec{x}-\vec{x}^k)=0, \tag{6}$$

де $\vec{\xi}^k=\vec{x}^k+\theta_k(\vec{x}^k-\vec{x}^k),\quad 0<\theta_k<1$ Тут $F(x)=\left(\dfrac{\partial f_i}{\partial x_j}\right)_{i,j=1}^{n}$ - матриця Якобі

для $\vec{F}(\vec{x})$:. Можемо наближено вважати $\vec{\xi}^k\approx\vec{x}^k$. Тоді з (6) матимемо:

$$\vec{F}(\vec{x}^k)+F'(\vec{x}^k)(\vec{x}^{k+1}-\vec{x}^k)=0 \tag{7}$$

Ітераційний процес представимо у вигляді:

$$\vec{x}^{k+1}=\vec{x}^k-\left[F'(\vec{x}^k)\right]^{-1}\vec{F}(\vec{x}^k). \tag{8}$$

Для реалізації методу Ньютона потрібно, щоб існувала обернена матриця $\left[F'(\vec{x}^k)\right]^{-1}$.

Можна не шукати обернену матрицю, а розв’язувати на кожній ітерації СЛАР

$$\begin{cases}A_k\vec{z}^k=\vec{F}(\vec{x}^k)\\ \vec{x}^{k+1}=\vec{x}^k-\vec{z}^k\end{cases},k=0,1,2,... \tag{9}$$

де $\vec{x}^0$—задано, а матриця $A_k=F'(\vec{x}^k)$.

Метод має квадратичну збіжність, якщо добре вибрано початкове наближення. Складність методу (при умові використання методу Гаусса розв‘язання СЛАР (9)) на кожній ітерації

$$Q_n=\frac{2}{3}n^3+O(n^2),$$

де n – розмірність системи (1).

4.2.3. Модифікований метод Ньютона

Ітераційний процес має вигляд :

$$\vec{x}^{k+1} = \vec{x}^k - \left[F'(\vec{x}^0)\right]^{-1} \vec{F}(\vec{x}^k).$$

Тепер обернена матриця обчислюється тільки на нульовій ітерації. На інших - обчислення нового наближення зводиться до множення матриці $A_0 = \left[F'\left(\vec{x}^0\right)\right]^{-1}$ на вектор $\vec{F}(\vec{x}^k)$ та додавання до $\vec{x}^k$.

Запишемо метод у вигляді системи лінійних рівнянь (аналог (9))

$$\begin{cases} A_0 \vec{z}^k = \vec{F}(\vec{x}^k) \\ \vec{x}^{k+1} = \vec{x}^k - \vec{z}^k \end{cases}. \qquad (10)$$

Оскільки матиця A_0 розкладається на трикутні (або обертається) один раз, то складність цього методу на одній ітерації (окрім нульової) $Q_n = O(n^2)$. Але цей метод має лінійну швидкість збіжності.

Можливе циклічне застосування модифікованого методу Ньютона, тобто коли обернену матрицю похідних шукаємо та обертаємо через певне число кроків ітераційного процесу.

Задача 9 Побудувати аналог методу січних для систем нелінійних рівнянь.

5. Алгебраїчна проблема власних значень.

Нехай задано матрицю $A:(n\times n)$. Тоді задача на власні значення ставиться так: знайти число λ та вектор $\vec{x} \neq 0$, що задовольняють рівнянню

$$A\vec{x} = \lambda\vec{x}. \qquad (1)$$

λ називається власним значенням A, а $\vec{x}$ - власним вектором. З (1)

$$\det(A - \lambda E) \equiv P_n(\lambda) \equiv (-1)^n \lambda^n + a_n \lambda^{n-1} + \ldots + a_0 = 0.$$

Тут $P_n(\lambda)$ – характеристичний багаточлен.

Для розв'язання багатьох задач механіки, фізики, хімії потрібне знаходження всіх власних значень λ_i, $i = \overline{1,n}$, а іноді й всіх власних векторів $\vec{x}_i$, що відповідають λ_i. Цю задачу називають повною проблемою власних значень.

В багатьох випадках потрібно знайти лише максимальне або мінімальне за модулем власне значення матриці. При дослідженні стійкості коливальних процесів іноді потрібно знайти два максимальних за модулем власних значення матриці.

Останні дві задачі (2)-(4) називають частковими проблемами власних значень.

5.1. Степеневий метод [БЖК, 309-314], [КБМ, 149-157]

1) *Знаходження* $\lambda_{\max}$: $|\lambda_1| \equiv \lambda_{\max} > |\lambda_2| \geq |\lambda_3| \geq \ldots.$

Нехай $\vec{x}^0$ - заданий вектор, будемо послідовно обчислювати вектори

$$\vec{x}^{k+1} = A\vec{x}^k, k = 0,1.... \quad (2)$$

Тоді $\vec{x}^k = A^k\vec{x}^0$. Нехай $\{\vec{e}_i\}_{i=1}^n$ - система власних векторів. Представимо $\vec{x}^0$ у вигляді:

$$\vec{x}^0 = \sum_{i=1}^n c_i\vec{e}_i .$$

Оскільки $A\vec{e}_i = \lambda_i\vec{e}_i$, то $\vec{x}^k = \sum_{i=1}^n c_i\lambda_i^k\vec{e}_i$. При великих k $\vec{x}^k \approx c_1\lambda_1^k\vec{e}_1$. Тому

$$\mu_1^{(k)} = \frac{x_m^{k+1}}{x_m^k} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right).$$

Значить $\mu_1^{(k)} \xrightarrow[k\to\infty]{} \lambda_1$.

Якщо матриця $A = A^T$ симетрична, то існує ортонормована система векторів $(\vec{e}_i, \vec{e}_j) = \delta_{ij}$. Тому

$$\mu_1^{(k)} = \frac{(\vec{x}^{k+1}, \vec{x}^k)}{(\vec{x}^k, \vec{x}^k)} = \frac{\left(\sum_i c_i\lambda_i^{k+1}\vec{e}_i, \sum_j c_j\lambda_j^k\vec{e}_j\right)}{\left(\sum_i c_i\lambda_i^k\vec{e}_i, \sum_j c_j\lambda_j^k\vec{e}_j\right)} = \frac{\sum_i c_i^2\lambda_i^{2k+1}}{\sum_i c_i^2\lambda_i^{2k}} =$$

$$= \frac{c_1^2\lambda_1^{2k+1} + c_2^2\lambda_2^{2k+1} + ...}{c_1^2\lambda_1^{2k} + c_2^2\lambda_2^{2k} + ...} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right) \xrightarrow[k\to\infty]{} \lambda_1$$

Це означає збіжність до максимального за модулем власного значення з квадратичною швидкістю.

Якщо $|\lambda_1| > 1$, то при проведенні ітерацій відбувається зріст компонент вектора $\vec{x}^k$, що приводить до "переповнення" (overflow). Якщо ж $|\lambda_1| < 1$, то це приводить до зменшення компонент (underflow). Позбутися негативу такого явища можна нормуючи вектори $\vec{x}^k$ на кожній ітерації.

Алгоритм степеневого методу знаходження максимального за модулем власного значення з точністю ε виглядає так:

1. $\vec{x}^0 \to \vec{e}_0 = \frac{\vec{x}^0}{\|\vec{x}^0\|}$;

2. $\vec{x}^{k+1} = A\vec{x}^k, \mu_1^{(k)} = (\vec{x}^{k+1}, \vec{e}^k)$, $\vec{e}^{k+1} = \frac{\vec{x}^{k+1}}{\|\vec{x}^{k+1}\|}$, $k = 0,1...$;

3. $\left|\mu_1^{(k+1)} - \mu_1^{(k)}\right| \geq \varepsilon$ $goto\, 2$;

4. $\lambda_1 \approx \mu_1^{(k+1)}$.

За цим алгоритмом для симетричної матриці $A^T = A$ швидкість прямування $\mu_1^{(k)}$ до $\lambda_{\max}$ - квадратична.

2. *Знаходження* λ_2: $|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \ldots$. Нехай λ_1, $\vec{e}_1$ відомі.

Задача 10 Довести, що якщо $|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \ldots$, то

$$\mu_2^{(k)} = \frac{x_m^{k+1} - \lambda_1 x_m^k}{x_m^k - \lambda_1 x_m^{k-1}} \xrightarrow[k\to\infty]{} \lambda_2, \text{ де } \vec{x}^{k+1} = A\vec{x}^k$$

x_m^k – m-та компонента $\vec{x}^k$.

Задача 11 Побудувати алгоритм обчислення λ_2, $\vec{e}_2$, використовуючи нормування векторів та скалярні добутки для обчислення $\mu_2^{(k)}$.

3. *Знаходження мінімального власного числа* $\lambda_{\min}(A) = \min_i |\lambda_i(A)|$.

Припустимо, що $\lambda_i(A) > 0$ та відоме $\lambda_{\max}$. Розглянемо матрицю $B = \lambda_{\max} E - A$. Маємо

$$\lambda_i(B) = \lambda_{\max} - \lambda_i(A), \ \forall i,$$

Тому $\mathbf{\max_i} \lambda_i(B) = \lambda_{\max} - \mathbf{\min_i} \lambda_i(A)$. Звідси $\lambda_{\min}(A) = \lambda_{\max}(A) - \lambda_{\max}(B)$.

Якщо $\exists i : \lambda_i(A) < 0$, то будуємо матрицю $\overline{A} = \sigma E + A, \sigma > 0$ $\overline{A} > 0$ і для неї попередній розгляд дає необхідний результат. Замість $\lambda_{\max}$ в матриці B можна використовувати $\|A\|$.

Ще один спосіб обчислення мінімального власного значення полягає в використання обернених ітерацій:

$$A\vec{x}^{k+1} = \vec{x}^k, k = 0,1\ldots \qquad (3)$$

Але цей метод вимагає більшої кількості арифметичних операцій: складність методу на основі формули (2) $Q = O(n^2)$, а на основі (3) - $Q = O(n^3)$, оскільки треба розв‘язувати СЛАР, але збігається метод (3) швидче.

5.2. Ітераційний метод обертання [КБМ, 157-161]

Це метод розв’язання повної проблеми власних значень для симетричних матриць $A^T = A$. Існує матриця U, що приводить матрицю A до діагонального виду:

$$A = U\Lambda U^T, \qquad (1)$$

де Λ - діагональна матриця, по діагоналі якої стоять власні значення λ_i; U - унітарна матриця, тобто: $U^{-1} = U^T$.

З (1) маємо

$$\Lambda = U^T A U \qquad (2)$$

Нехай $\exists \tilde{U}$ - матриця, така що $\tilde{\Lambda} = \tilde{U}^T A \tilde{U}$ i $\tilde{\Lambda} = (\tilde{\lambda}_{ij})_{i?j}^n$ $|\tilde{\lambda}_{ij}| < \delta << 1, i \neq j$.

Тоді діагональні елементи мало відрізняються від власних значень

$$|\tilde{\lambda}_{ii} - \lambda_i(A)| < \varepsilon = \varepsilon(\delta).$$

Введемо $t(A)=\sum\limits_{\substack{i,j=1\\ i\neq j}}^{n}(a_{ij})^2$. З малості величини $t(A)$ витікає, що діагональні елементи малі. По $A=A_0$ за допомогою матриць обертання U_k

$$U_k=\begin{pmatrix} 1 & 0 & 0 & 0\cdots 0 & 0 & 0 & 0 \\ 0 & \ddots & \vdots & \vdots\cdots\vdots & \vdots & \vdots & 0 \\ 0 & 0\cdots & \cos\varphi & 0\cdots 0 & -\sin\varphi & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0\cdots & \sin\varphi & 0\cdots 0 & \cos\varphi & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots\cdots\vdots & & \ddots & 0 \\ 0 & 0 & 0 & 0\cdots 0 & 0 & & 1 \end{pmatrix}\begin{matrix} \\ \\ \leftarrow i \\ \\ \leftarrow j' \\ \\ \\ \end{matrix}$$

$$\uparrow i \qquad\qquad \uparrow j$$

що повертають систему векторів на кут φ, побудуємо послідовність $\{A_k\}$ таку, що $A_k\to\Lambda$ при $k\to\infty$.

Задача 12 Показати , що матриця обертання U_k є унітарною: $U_k^{-1}=U_k^{T}$.

Послідовно будуємо:

$$A_{k+1}=U_k^T A_k U_k, \tag{3}$$

Процес (3) називається *монотонним*, якщо: $t(A_{k+1})<t(A_k)$.

Задача 13 Довести, що для матриці (3) виконується :

$$a_{ij}^{(k+1)}=a_{ij}^{(k)}\cos 2\varphi+\frac{1}{2}\left(a_{jj}^{(k)}-a_{ii}^{(k)}\right)\sin 2\varphi. \tag{4}$$

Показати, що $t(A_{k+1})=t(A_k)-2\left[a_{ij}^{(k)}\right]^2$, якщо вибирати φ з умови $a_{ij}^{(k+1)}=0$.

Звідси $\varphi=\varphi_k=\frac{1}{2}\arctan p^{(k)}$ $p^{(k)}=\frac{2a_{ij}^{(k)}}{a_{ii}^{(k)}-a_{jj}^{(k)}}$, де $\left|a_{ij}^{(k)}\right|=\max\limits_{\substack{m,l\\ m\neq l}}\left|a_{ml}^{(k)}\right|$. Тоді $t(A_k)\to 0, k\to\infty$. Чим більше *n* тим більше ітерацій необхідно для зведення *A* до Λ.

Якщо матриця несиметрична, то застосовують *QR*, *QL* методи.

6. Інтерполювання функцій

6.1 Постановка задачі інтерполювання [СГ, 151-155]

Нехай функція $f(x)\in C[a,b]$ задана своїми значеннями $y_i=f(x_i), x_i\in[a,b], i=\overline{0,n}$, причому при $x_i\neq x_j$ для $i\neq j$.

Функція $\Phi(x)$ називається *інтерполюючою* для $f(x)$ на сітці $\{x_i\}_{i=0}^{n}$, якщо $\Phi(x_i)=y_i, i=\overline{0,n}$.

Задача інтерполювання функції має не єдиний розв'язок. Виберемо систему лінійно незалежних функцій $\{\varphi_k(x)\}_{k=0}^{n}$, $\varphi_k(x) \in C[a,b]$ і побудуємо лінійну комбінацію

$$\Phi(x) = \Phi_n(x) = \sum_{k=0}^{n} c_k \varphi_k(x), \tag{1}$$

яка називається *узагальненим багаточленом*. Умови інтерполювання дають СЛАР

$$\sum_{k=0}^{n} c_k \varphi_k(x_i) = y_i, i = \overline{0,n}, \tag{2}$$

розв'язком якої є $\vec{c} = (c_0,...,c_n)$. Якщо

$$D(x_0,..,x_n) = \begin{vmatrix} \varphi_0(x_0) & ... & \varphi_n(x_0) \\ ... & ... & ... \\ \varphi_0(x_n) & ... & \varphi_n(x_n) \end{vmatrix} \neq 0,$$

то система (2) має єдиний розв'язок.

Система функцій $\{\varphi_k(x)\}$ $k = \overline{0,n}$ називається *системою Чебишова*, якщо, $D(x_0,..,x_n) \neq 0, \forall \{x_i\}_{i=0}^{n}$, $x_i \in [a,b], i = \overline{0,n}$, $x_i \neq x_j$ $\forall i \neq j$.

Приклади систем Чебишова.

1. $\varphi_k(x) = x^k$ - алгебраїчна система.

Визначник $D(x_0,..,x_n) \neq 0$ є визначником Вандермонда:

$$D(x_0,..,x_n) = \begin{vmatrix} 1 & x_0 & ... & x_0^{n-1} \\ ... & ... & ... & ... \\ 1 & x_n & ... & x_n^{n-1} \end{vmatrix} = \prod_{0 \le k < m \le n} (x_k - x_m) \neq 0.$$

2. $\varphi_k(x) = L_k(x)$ - ортогональні багаточлени Лежандра; $\varphi_k(x) = T_k(x)$ - ортогональні багаточлени Чебишова.

3. $\varphi_k(x)$: $1, \cos x, \sin x, ..., \cos nx, \sin nx$.

Тоді $\Phi_n(x) = T_n(x) = a_0 + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx)$ - тригонометричний багаточлен.

6.2. Інтерполяційна формула Лагранжа [СГ, 127-129],[БЖК,38-42]

Якщо $\varphi_k(x) = x^k$, то $\Phi_n(x) = P_n(x) = \sum_{k=0}^{n} c_k x^k$. Задача інтерполювання функції $f(x)$ алгебраїчним, багаточленом полягає в знаходженні коефіцієнтів c_k, $k = \overline{0,n}$ для яких виконується умова $f(x_i) = \varphi(x_i)$ $i = \overline{0,n}$. Представимо інтерполяційний багаточлен у вигляді

$$P_n(x) = L_n(x) = \sum_{k=0}^{n} f(x_k) \Phi_k^{(n)}(x) \tag{1}$$

Тут $L_n(x)$ - *інтерполяційна поліном*; $\Phi_k^{(n)}(x)$ - поліноми n-го степеня, які називають *множниками Лагранжа.* З умови $L_n(x_i)=f(x_i)$ випливає, що множник Лагранжа повинен задовольняти умови

$$\Phi_k^{(n)}(x_i)=\delta_{ik}. \tag{2}$$

Так як $\Phi_k^{(n)}(x)$- багаточлен степеня *n*, то він має вигляд

$$\Phi_k^{(n)}(x)=A_k(x-x_0)...(x-x_{k-1})(x-x_{k+1})...(x-x_n),$$

де A_k –число . Знайдемо його з умови $\Phi_k^{(n)}(x_k)=1$:

$$A_k=\frac{1}{(x_k-x_0)...(x_k-x_{k-1})(x_k-x_{k+1})...(x_k-x_n)}.$$

Таким чином багаточлен $\Phi_k^{(n)}(x)$ мають вигляд:

$$\Phi_k^{(n)}(x)=\frac{(x-x_0)...(x-x_{k-1})(x-x_{k+1})...(x-x_n)}{(x_k-x_0)...(x_k-x_{k-1})(x_k-x_{k+1})...(x_k-x_n)} \tag{3}$$

Позначивши $\omega_n(x)=\prod_{i=0}^{n}(x-x_i)$, маємо $\Phi_k^{(n)}(x)=\frac{\omega_n(x)}{(x-x_k)\omega_n'(x_k)}$. Остаточно формула Лагранжа має вигляд:

$$L_n(x)=\sum_{k=0}^{n}f(x_{\mathrm{k}})\frac{\omega_n(x)}{(x-x_k)\omega_n'(x_k)} \tag{4}$$

6.3. Залишковий член інтерполяційного полінома [СГ, 132-133], [БЖК,42]

В заданих точках (точки інтерполювання) значення функції та полінома співпадають, але в інших точках в загальному випадку не співпадають. Отже доцільно розглянути питання про похибку інтерполювання.

Замінюючи функцію $f(x)$ на $L_n(x)$ ми допускаємо похибку $r_n(x)=f(x)-L_n(x)$. Це *залишковий член* інтерполювання.

З означення випливає, що, $r_n(x_i)=0,\ x_i\in[a,b]$. Оцінимо похибку у кожній точці $x\in[a,b]$. Введемо допоміжну функцію:

$$g(t)=f(t)-L_n(t)-K\omega_n(t),\ t\in[a,b],\ g(x_i)=0,\ i=\overline{0,n}.$$

Знайдемо таке K, щоб $g(x)=0$, в деякій точці $x\in[a,b],\ x\neq x_i,\ i=\overline{0,n}$. Легко бачити, що

$$K=\frac{f(x)-L_n(x)}{\omega_n(x)}.$$

Припустимо що $f(x)\in C^{n+1}[a,b]$, тоді $g(t)\in C^{n+1}[a,b]$. Функція $g(t)=0$ в (n+2)-х точках, а саме $t=x,\ t=x_i,\ i=\overline{0,n}$. З теореми Ролля випливає, що існує (n+1)-а точка, де $g'(t_i)=0,\ i=\overline{0,n}$. Продовжуючи цей процес, отримаємо, що існує хоча б одна $\xi\in[a,b]$ така, що $g^{(n+1)}(\xi)=0$. Так як $g^{(n+1)}(t)=f^{(n+1)}(t)-0-K(n+1)!$, то $(\exists\xi)$, що

$$g^{(n+1)}(\xi)=f^{(n+1)}(\xi)-(n+1)!\frac{f(x)-L_n(x)}{\omega_n(x)}=0.$$

Звідси

$$r_n(x)=f(x)-L_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(x) \qquad (1)$$

Оскільки ξ невідомо, то використовують *оцінку залишкового члена:*

$$|r_n(x)|=|f(x)-L_n(x)|\le\frac{M_{n+1}}{(n+1)!}|\omega_n(x)|, \qquad (2)$$

де $M_{n+1}=\sup\limits_{x\in[a,b]}|f^{(n+1)}(x)|$.

6.4 Багаточлени Чебишова. Мінімізація залишкового члена інтерполяційного полінома [СГ 103-108],[БЖК 56-63]

Як вибрати вузли інтерполяції щоб похибка інтерполювання була мінімальною? Спочатку обґрунтуємо теоретичний апарат, завдяки якому будемо досліджувати це питання.

Багаточленом Чебишова (n-того степеня , 1–го роду) називається поліном, який задається такими рекурентними співвідношеннями:

$$T_{n+1}(x)-2xT_n(x)+T_{n-1}(x)=0 \qquad (1)$$

$$T_0(x)=1,\ T_1(x)=x. \qquad (2)$$

Знайдемо явний вигляд багаточлена Чебишова. Будемо шукати розв‘язок рівняння (1) у вигляді $T_n(x)=q^n$, де $q=q(x)$. Підставивши в (1), отримуємо характеристичне рівняння $q^2-2xq+1=0$. Тоді при $|x|\ge1\Rightarrow q_{1,2}=x\pm\sqrt{x^2-1}$, а при $|x|<1\Rightarrow q_{1,2}=\cos\varphi\pm i\sin\varphi$, $\varphi=\arccos x$.

Розглянемо обидва випадки детальніше:

а) при $|x|\le1$ $T_n(x)=A\cos(n\varphi)+B\sin(n\varphi)$. З (2) випливає A=1, B=0 і тому

$$T_n(x)=\cos(n\arccos x) \qquad (3)$$

б) при $|x|>1$

$$T_n(x)=\frac{1}{2}\left[\left(x+\sqrt{x^2-1}\right)^n+\left(x-\sqrt{x^2-1}\right)^n\right]. \qquad (4)$$

Знайдемо нулі та екстремуми багаточлена Чебишова:

$$T_n(x)=0,\ x\in[-1,1],\ \cos(n\arccos x)=0,\ \arccos x=\frac{2k+1}{2n}\pi\ ,\ k=\overline{0,n-1}.$$

Отже *нулі багаточлена Чебишова*:

$$x_k=\cos\frac{2k+1}{2n}\pi\in[-1,1],\ k=\overline{0,n-1}.$$

Локальні екстремуми багаточлена Чебишова на $x\in[-1,1]$:

$$x'_k=\cos\frac{k\pi}{2n},\ k=\overline{0,n}.$$

Коефіцієнт при старшому члені багаточлена дорівнює 2^{n-1}. Введемо *нормований багаточлен Чебишова* $\overline{T_n}(x) = 2^{1-n} T_n(x) = x^n + \ldots$ Тоді

$$\left\|\overline{T_n}(x)\right\|_{C[-1,1]} = \max_{x\in[-1,1]} \left|T_n(x)\right| = 2^{1-n}.$$

Відхиленням двох функцій $f(x)$ та $\Phi(x)$ називається величина

$$\Delta(f,\Phi) = \left\|f(x) - \Phi(x)\right\|_{C[a,b]}.$$

Теорема. (Чебишова) Серед усіх багаточленів *n*- того степеня з коефіцієнтом 1 при старшому степені $\overline{T_n}(x)$ найменше відхиляється від 0 на [-1,1], тобто $\left\|\overline{T}_n(x) - 0\right\|_{C[-1,1]} = \inf_{\overline{P}_n(x)} \left\|\overline{P}_n(x)\right\|_{C[-1,1]} = 2^{1-n}$.

◁ Будемо доводити від супротивного: припустимо, що існує багаточлен, такий, що

$$\left\|\overline{Q_n}(x)\right\| < 2^{1-n}.$$

Тоді $Q_{n-1}(x) = \overline{T_n}(x) - \overline{Q_n}(x)$ - поліном степеня не вище *n*-1 і не рівний тотожно нулю. Дослідимо його знаки:

$$\mathbf{sign}\left(Q_{n-1}\left(x_k'\right)\right) = \mathbf{sign}\left(\overline{T_n}\left(x_k'\right) - \overline{Q_n}\left(x_k'\right)\right) = \mathbf{sign}\left(\overline{T_n}\left(x_k'\right)\right) = \alpha(-1)^k, \ \alpha = \pm 1.$$

Значить $\exists z_k$, $k = \overline{0,n-1}$ таке, що $Q_{n-1}(z_k) = 0$. Це протиріччя, бо $Q_{n-1}(x)$ - поліном степеня $\le n$-1. ▷

Тепер узагальнимо наш багаточлен Чебишова на довільний проміжок. Нагадаємо $T_n(t) = \cos(n \arccos t)$, $-1 \le t \le 1$. Від змінної $t \in [-1,1]$ перейдемо до $x \in [a,b]$. Запровадимо заміну: $t = \dfrac{2}{b-a}x - \dfrac{b+a}{b-a}$, $x = \dfrac{b+a}{2} + \dfrac{b-a}{2}t$. Тоді $T_n^{[a,b]}(t) = \overline{T}_n\left(\dfrac{2}{b-a}x - \dfrac{b+a}{b-a}\right) = 2^{1-n}\cos\left(n\arccos\left(\dfrac{2}{b-a}x - \dfrac{b+a}{b-a}\right)\right)$. Побудований нами багаточлен Чебишова на $[a,b]$ не є нормованим. *Нормований багаточлен Чебишова* на $[a,b]$:

$$\overline{T_n}^{[a,b]}(x) = \frac{(b-a)^n}{2^{2n-1}}\cos\left(n\arccos\left(\frac{2x-(b+a)}{b-a}\right)\right).$$

Відповідно його нулі $x_k = \dfrac{a+b}{2} - \dfrac{b-a}{2}t_k$, $t_k = \cos\dfrac{(2k+1)\pi}{2n}$, $k = \overline{0,n-1}$, а точки екстремуму $x_k' = \dfrac{a+b}{2} - \dfrac{b-a}{2}t_k'$ $t_k' = \cos\dfrac{k\pi}{2n}$, $k = \overline{0,n}$.

Теорема Чебишова вірна і у випадку [*a*,*b*]. Тепер $\left\|\overline{T}_n^{[a,b]}(t)\right\|_{C[a,b]} = \dfrac{(b-a)^n}{2^{2n-1}}$.

Перейдемо до питання мінімізації залишкового члена. Нагадаємо, що

$$\left|r_n(x)\right| = \left|f(x) - L_n(x)\right| \le \frac{M_{n+1}}{(n+1)!}\left|\omega_n(x)\right|, \qquad (5)$$

де $M_{n+1} = \max_{x\in[a,b]}\left|f^{(n+1)}(x)\right|$, $\omega_n(x) = \prod_{i=0}^{n}(x - x_i) = x^{n+1} + \ldots$. Поставимо задачу

$$\inf_{\overline{P}_n(x)} \max_{x\in[a,b]} |\omega_n(x)|.$$

За теоремою Чебишова $\omega_n(x) = \overline{T}_{n+1}^{[a,b]}(x)$ поліном Чебишова. Якщо співпадають поліноми, то співпадають їх нулі. Отже: x_k – вузли інтерполяції співпадають з нулями багаточлена Чебишова $x_k = \frac{a+b}{2} - \frac{b-a}{2} t_k$, $t_k = \cos\frac{(2k+1)\pi}{2(n+1)}, k = \overline{0,n}$. В цьому випадку

$$|r_n(x)| = |f(x) - L_n(x)| \le \frac{M_{n+1}}{(n+1)!} \frac{(b-a)^{n+1}}{2^{2n+1}} \qquad (6)$$

Цю оцінку не можна покращити! Так для $f(x) = \overline{P}_{n+1}(x) = x^{n+1} + \ldots$ її $(n+1)$ похідна дорівнює $(n+1)!$, тому $M_{n+1} = (n+1)!$. Різниця $f(x) - L_n(x) = \overline{T}_{n+1}^{[a,b]}(x)$, отже $\max_{[a,b]} |f(x) - L_n(x)| = \frac{(b-a)^{n+1}}{2^{2n+1}}$.

6.5. Розділені різниці [БЖК 42-44], [СГ 129-130]

Розділені різниці є аналогом похідної для функцій, що задана таблицею.

Розділеною різницею 1- го порядку для функції $f(x)$ називатимемо

$$f(x_i; x_j) = \frac{f(x_i) - f(x_j)}{x_i - x_j}.$$

Розділеною різницею 2- го порядку для функції $f(x)$ називатимемо

$$f(x_i; x_j; x_k) = \frac{f(x_i; x_j) - f(x_j; x_k)}{x_i - x_k}.$$

Аналогічно визначаються розділені різниці довільного порядку.

Наведемо деякі властивості розділених різниць:

1^0. $f(x_0;\ldots;x_n) = \sum_{i=0}^{n} \frac{f(x_i)}{\prod_{i\ne j}(x_i - x_j)}$.

2^0. Розділена різниця – лінійний функціонал

$$(\alpha_1 f_1 + \alpha_2 f_2)(x_0; x_1) = \alpha_1 f_1(x_0; x_1) + \alpha_2 f_2(x_0; x_1)$$

3^0. Розділена різниця – симетричний функціонал

$$f(x_1,\ldots,x_i,\ldots,x_j,\ldots,x_n) = f(x_1,\ldots,x_j,\ldots,x_i,\ldots,x_n)$$

4^0. Для $f(x) \in C^n([a,b]), \exists \xi \in (a,b) : f(x_0, x_1,\ldots,x_n) = \frac{f^{(n)}(\xi)}{n!}$.

Задача 14 Довести властивість розділених різниць

$$f(x_0;\ldots;x_n) = \sum_{i=0}^{n} \frac{f(x_i)}{\prod_{i\ne j}(x_i - x_j)}.$$

Таблиця розділених різниць має вигляд

x_i	f_i	p.p.1	p.p.2	.....	p.p.n
x_0	$f(x_0)$				
		$f(x_0;x_1)$			
x_1	$f(x_1)$		$f(x_0;x_1;x_2)$		
		$f(x_1;x_2)$			
x_2	$f(x_2)$				
.....	.....	.....	.....	.....	$f(x_0;\ldots;x_n)$
		$f(x_{n-1};x_n)$			
x_n	$f(x_n)$				

6.6. Інтерполяційна формула Ньютона [БЖК 44-47], [СГ 130-132]

Запишемо формулу Лагранжа інтерполяційного багаточлена

$$L_n(x) = \sum_{i=0}^{n} f(x_i) \frac{\omega_n(x)}{(x - x_i)\omega_n'(x_i)}, \qquad (1)$$

де $\omega_n(x) = \prod_{j=0}^{n} (x - x_j)$.

Позначимо $\Phi_j(x) = L_j(x) - L_{j-1}(x)$. Тоді, оскільки

$$L_n(x) = L_0(x) + (L_1(x) - L_0(x)) + \ldots + (L_n(x) - L_{n-1}(x)),$$
$$L_j(x_i) = L_{j-1}(x_i) = f(x_i) \quad \forall i \in \overline{0, j-1},$$

то

$$\Phi_j(x_i) = A_j(x - x_0)\ldots(x - x_{j-1}), \qquad (2)$$

де A_j визначається з умови $\Phi_j(x_j) = L_j(x_j) - L_{j-1}(x_j) = f(x_j) - L_{j-1}(x_j)$. Звідси

$$\Phi_j(x) = \frac{f(x_j) - L_{j-1}(x_j)}{(x_j - x_0)\ldots(x_j - x_{j-1})}(x - x_0)\ldots(x - x_j).$$

Тоді

$$A_j = \frac{f(x_j) - L_{j-1}(x_j)}{(x_j - x_0)\ldots(x_j - x_{j-1})} = \frac{f(x_j)}{(x_j - x_0)\ldots(x_j - x_{j-1})} -$$
$$- \sum_{i=0}^{j-1} \frac{f(x_i)}{(x_i - x_0)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_{j-1})(x_j - x_i)} =$$
$$= \frac{f(x_j)}{(x_j - x_0)\ldots(x_j - x_{j-1})} + \sum_{i=0}^{j-1} \frac{f(x_i)}{(x_i - x_0)\ldots(x_i - x_{j-1})} =$$
$$= \sum_{i=0}^{j} \frac{f(x_i)}{(x_i - x_0)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_j)} = f(x_0,\ldots,x_j).$$

Звідси маємо інтерполяційну формулу Ньютона вперед ($x_0 \to x_n$):

$$L_n(x) = f(x_0) + f(x_0, x_1)(x - x_0) + \ldots + f(x_0,\ldots,x_n)(x - x_0)\ldots(x - x_{n-1}). \qquad (3)$$

Аналогічно, інтерполяційна формула Ньютона назад ($x_n \to x_0$):

$$L_n(x) = f(x_n) + f(x_n, x_{n-1})(x - x_n) + ... + f(x_0, ..., x_n)(x - x_n)...(x - x_1). \quad (4)$$

Маємо рекурсію за степенем багаточлена

$$L_n(x) = L_{n-1}(x) + f(x_0; ...; x_n)(x - x_0)...(x - x_{n-1}).$$

Звідси $L_n(x) = f(x_0) + (x - x_0)\{f(x_0, x_1) + (x - x_1)\{... + (x - x_{n-1}) f(x_0, x_1, ..., x_n)\}\}$ і цю формулу розкриваємо починаючи з середини (це аналог формули Герона обчислення значення багаточлена).

Виведемо нову формулу для похибки інтерполювання. Для $x \neq x_i, i = \overline{0, n}$ розглянемо розділену різницю

$$f(x; x_0; ...; x_n) = \frac{f(x)}{(x - x_0)...(x - x_n)} + \sum_{k=0}^{n} \frac{f(x_k)}{\prod_{i \neq k} (x - x_i)}.$$

Звідси

$$f(x) = f(x_0)\frac{(x - x_1)...(x - x_n)}{(x_0 - x_1)...(x_0 - x_n)} + ... + f(x_n)\frac{(x - x_0)...(x - x_{n-1})}{(x_n - x_0)...(x_n - x_{n-1})} +$$

$$+ f(x; x_0; ...; x_n)(x - x_0)...(x - x_n) = L_n(x) + f(x; x_0; ...; x_n)\omega_n(x)$$

Тоді похибка має вигляд

$$r_n(x) = f(x) - L_n(x) = f(x; x_0; ...; x_n)\omega_n(x) \quad (5)$$

Це нова форма для залишкового члена.

Порівнюючи з формулою залишкового члена в пункті 6.3, маємо

$$f(x, x_0, ..., x_n) = \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

що доводить властивість 4^0 розділених різниць

Нехай маємо сітку рівновіддалених вузлів: $x_i = a + ih$,

$h = \frac{b-a}{n}, i = \overline{0, n}, x_0 = a, x_n = b$. Позначимо $\Delta f_0 = f_1 - f_0$,

$\Delta^2 f_0 = \Delta f_1 - \Delta f_0 = f_2 - 2f_1 + f_0$ … - скінчені різниці. Запишемо формули Ньютона у нових позначеннях:

$$L_n(x) = L_n(x_0 + th) = f_0 + t\Delta f_0 + ... + \frac{t(t-1)...(t-n+1)}{n!}\Delta^n f_0,\ t = \frac{x - x_0}{h}.$$

Це інтерполяційна. формула Ньютона вперед по рівновіддалених вузлах.

Задача 15 Побудувати інтерполяційну формулу Ньютона назад по рівновіддалених вузлах.

6.7. Інтерполювання з кратними вузлами [БЖК 47-50], [ЛМС 34]

Нехай $f(x)$ задана таблицею значень $f^{(j)}(x_i)$ $i = \overline{0, n}, j = \overline{0, k_i - 1}$, k_i - кратності відповідних вузлів. Побудуємо $H_m^{(i)}(x_j) = f^{(i)}(x_j)$ - інтерполяційний багаточлен Ерміта по кратних вузлах, причому $m = \sum_{i=1}^{n} k_i - 1$.

Якщо $k_i \equiv 1$, то $H_m(x) = L_n(x)$.

Для побудови $H_m(x)$ в загальному випадку для кожної точки x_i введемо k_i точок $x_{ij}^{\varepsilon} = x_i + j\varepsilon$, $i = \overline{0,n}; j = \overline{0,k_i - 1}$. З умови $\forall i$ $x_{ik_i-1}^{\varepsilon} = x_i + \varepsilon(k_i - 1) < x_{i+1}$ можна вибрати ε.

Нехай $f(x) \in C^{m+1}[a,b]$. Запишемо інтерполяційну формулу Ньютона

$$L_m{}^{\varepsilon}(x) = f(x_{00}^{\varepsilon}) + f(x_{00}^{\varepsilon}, x_{01}^{\varepsilon})(x - x_{00}^{\varepsilon}) + ... + f(x_{00},...,x_{nk_n-1})(x - x_{00})...(x - x_{nk_n-1})$$

При $\varepsilon \to 0$ маємо $x_{ij}^{\varepsilon} \to x_i$. Крім того $f(x_{io}^{\varepsilon};...;x_{ik_i-1}) = f(x_i;....;x_i) = \dfrac{f^{k_i}(x_i)}{k_i!}$.

Тому $L_m{}^{\varepsilon}(x) \to H_m(x)$ та

$$R_m(x) = f(x) - H_m(x) = \frac{f^{(m+1)}(\xi)}{(m+1)!}\Omega_m(x),$$

де $\Omega_m(x) = (x - x_0)^{k_0}...(x - x_n)^{k_n}$.

6.8. Збіжність процесу інтерполювання [ЛМС,42-46]; [СГ, 134-136]

Виникає питання, чи буде прямувати до нуля похибка інтерполювання $f(x) - L_n(x)$, якщо число вузлів *n* збільшувати?

Введемо норму $\|f(x) - L_n(x)\|_{C[a,b]} = \max\limits_{x\in[a,b]}|f(x) - L_n(x)|$. Тоді для

$f(x) \in C^{n+1}[a,b]$ справджується оцінка

$$\|f(x) - L_n(x)\|_{C[a,b]} \le \frac{M_{n+1}}{(n+1)!}\|\omega_n(x)\|_{C[a,b]}, \qquad (1)$$

де $M_{n+1} = \max\limits_{x\in[a,b]}\left|f^{(n+1)}(x)\right|$, $\omega_n(x) = \prod\limits_{i=0}^{n}(x - x_i)$. А яка оцінка буде для довільної неперервної функції?

Кажуть, що інтерполяційний процес для функції $f(x)$ *збігається* в точці $x \in [a,b]$, якщо

$$\lim_{n\to\infty} L_n(x) = f(x), \forall\{x_i\}_{i=1}^{n} : h = \max_{i=\overline{1,n}} h_i \to 0, (h_i = x_i - x_{i-1}).$$

Якщо $\|f(x) - L_n(x)\|_{C[a,b]} \xrightarrow[n\to\infty]{} 0$, то інтерполяційний процес *збігається рівномірно.*

Розглянемо приклади поведінки інтерполяційних багаточленів при $n \to \infty$ для деяких функцій.

Приклад 1 Послідовність інтерполяційних багаточленів (сітка рівномірна), побудованих для неперервної функції $f(x) = |x|, -1 \le x \le 1$ (функція неперервна, але негладка), не збігається на $x \in [-1,1]$, крім точок *x*= –1, 0, 1. На рис. 1 дано графіки самої функції (штрихова лінія) та інтерполяційного

багаточлена (суцільна лінія) на рівномірній сітці $x_i = -1 + ih,\ \ h = \frac{2}{n}, i = \overline{0,n}$ для n=10.

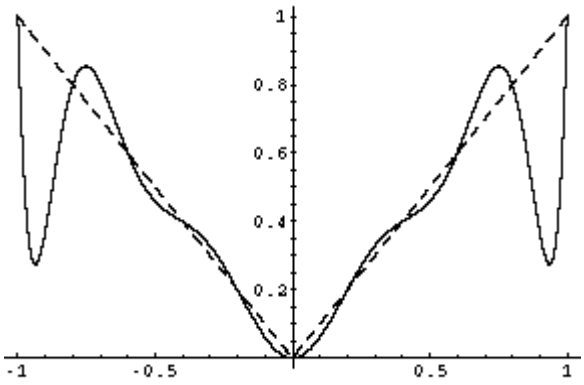


Рис. 3

Приклад 2 Функція Рунге $f(x) = \frac{1}{1+40x^2}$, $-1 \le x \le 1$ (функція аналітична!). Для рівномірної сітки

$$x_i = -1 + ih,\ \ h = \frac{2}{n}, i = \overline{0,n}$$

маємо графіки: суцільна лінія – інтерполяційного багаточлена; пунктирна – самої функції; n=10.

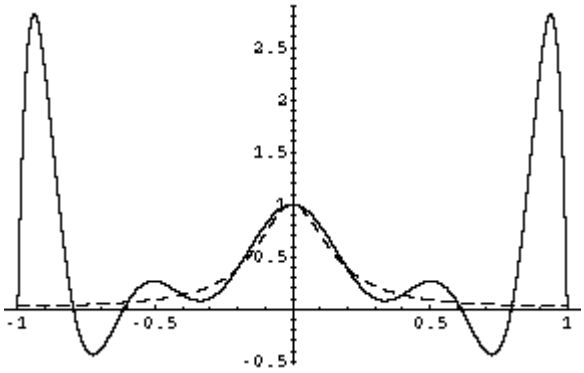


Рис. 4

Пояснити чому рівномірна сітка дає великі похибки інтерполювання біля кінців інтервалу інтерполювання допомагає рис. 3. На цьому рисунку суцільною лінією представлено графік функції $\omega_n(x) = \prod_{i=0}^{n}(x - x_i)$ (n=8) для рівномірної сітки. Як бачимо максимальні за модулем значення цієї функції припадають на кінці інтервалу.

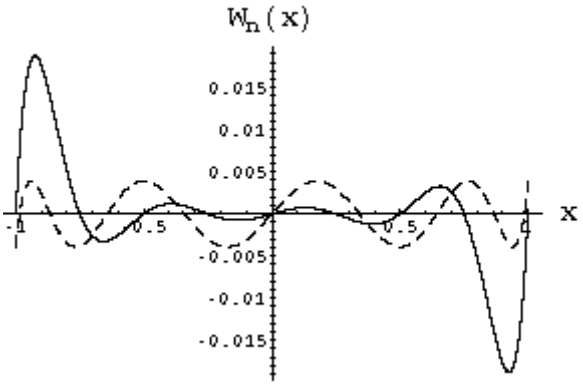


Рис. 5

Для порівняння на цьому ж рисунку (штрихова лінія) побудовано графік $\omega_n(x) = \prod_{i=0}^{n}(x - x_i)$, що відповідає чебишовським вузлам, які мінімізують похибку інтерполювання. Тепер відхилення $|\omega_n(x)|$ розподілено рівномірно по всьому проміжку інтерполювання.

Теорема 1 (Фабера) $\forall\{x_i\}_{i=0}^{n}$ існує $f(x) \in C[a,b]$, для якої інтерполяційний процес не збігається рівномірно.

Теорема 2 (Марцинкевича) $\forall f(x) \in C[a,b] \quad \exists\{x_i\}_{i=0}^{n}$ такі, що послідовність $\{L_n(x)\}$ збігається рівномірно до $f(x)$.

Розглянемо оператори $P_n : C[a,b] \to L_n$, тобто $P_n f(x) = L_n(x)$.

Теорема 3 Стала Лебега $\|P_n\| = \max\limits_{j,[a,b]} \sum\limits_{j=0}^{n} |\varphi_j^{(n)}(x)|$, де $\varphi_j^{(n)}(x) = \dfrac{\omega_n(x)}{(x - x_j)\omega_n'(x_i)}$

Теорема 4 Для $f(x) \in C[a,b]$

$$\|f(x) - L_n(x)\|_{C[a,b]} \le (1 + \|P_n\|)E_n(f),$$

де $E_n(f) = \inf\limits_{Q_n(x)} \|f(x) - Q_n(x\|_{C[a,b]}$ - відхилення багаточлена n-го степеня найкращого рівномірного наближення від $f(x)$.

Теорема 5 Нехай P_n^E - оператор інтерполяції на рівномірній сітці, P_n^T - оператор інтерполяції на чебишовській сітці. Тоді на $[-1;1]$ маємо наближені оцінки:

$$\|P_n^E\| \approx C_1 2^n, \|P_n^T\| \approx C_2 \ln(n).$$

Останні оцінки поясняють розбіжність процесу інтерполювання при великих n.

6.9. Кусково - лінійна інтерполяція

Інтерполяція багаточленом Лагранжа або Ньютона на відрізку $[a,b]$ з використанням великої кількості вузлів інтерполяції часто приводить до поганого наближення через розбіжність процесу інтерполювання. Для того щоб уникнути великої похибки, весь відрізок $[a,b]$ розбивають на частинні відрізки $[x_{i-1}, x_i]$ і на кожному з частинних відрізків замінюють функцію

$f(x)$ багаточленом невисокого степеню. В цьому і полягає кусково-поліноміальна інтерполяція.

Розглянемо найпростішу таку інтерполяцію – лінійну. Нехай задана $f(x)$ значеннями $f(x_i)$ $i=\overline{0,n}$. Побудуємо функцію $\Phi_1(x)$ – лінійну на $x\in[x_{i-1},x_i]$, що інтерполює ці значення:

$$\Phi_1(x)=L_1^i(x)=f(x_{i-1})\frac{x-x_{i-1}}{x_i-x_{i-1}}+f(x_i)\frac{x_i-x}{x_i-x_{i-1}},\ x\in[x_{i-1},x_i] \quad (2)$$

Представимо її у вигляді

$$\Phi_1(x)=\sum_{i=0}^{n}f(x_i)\varphi_i(x). \quad (3)$$

З умов інтерполювання маємо

$$\Phi_1(x_j)=\sum_{i=0}^{n}f(x_i)\varphi_i(x_j)=f(x_j).$$

Звідси

$$\varphi_i(x_j)=\delta_{ij}=\begin{cases}1, i=j\\ 0, i\neq j\end{cases}.$$

Значить

$$\varphi_i(x)=\begin{cases}\dfrac{x-x_{i-1}}{x_i-x_{i-1}}, x_{i-1}\le x\le x_i\\ \dfrac{x_{i+1}-x}{x_{i+1}-x_i}, x_i\le x\le x_{i+1}\\ 0 \qquad\quad , x\le x_{i-1},\ x\ge x_{i+1}\end{cases}$$

Теорема 1 Для $f(x)\in C^2[a,b]$ справедлива оцінка

$$\|f(x)-\Phi_1(x)\|_{C[a,b]}\le\frac{M_2}{8}|h|^2, \quad (4)$$

де $\Phi_1(x)$ - кусково-лінійна функція побудована по значеннях $f(x_i), i=\overline{0,n}$, $|h|=\max_i h_i, h_i=x_i-x_{i-1}$.

▷ Маємо для $x\in[x_{i-1},x_i]$

$$z(x)=f(x)-\Phi_1(x)=f(x)-L_1^i(x)=\frac{f''(\xi_i)}{2!}(x-x_{i-1})(x-x_i).$$

Звідси

$$|f(x)-\Phi_1(x)|\le\frac{M_2^i}{2}|(x-x_{i-1})(x-x_i)|\le\frac{M_2^i h_i^2}{8}, \quad (5)$$

де $M_2^i=\max_{x\in[x_{i-1},x_i]}|f''(x)|$. Остання оцінка отримана з нерівності

$$\max_{[x_{i-1},x_i]}|(x-x_{i-1})(x-x_i)|=\frac{h_i^2}{4}.$$

Тоді

$$\max_{i=\overline{1,n}} \max_{x\in[x_{i-1},x_i]} |z(x)| \le \frac{1}{8} h^2 M_2, \qquad (6)$$

де $M_2 = \max_{[a,b]} |f''(x)|$, $h = \max_i h_i$, що доводить (4)◁

Задача 16 Довести оцінку $|f'(x) - \Phi_1'(x)| \le |h| M_2$.

Отже маємо збіжність процесу інтерполювання за допомогою кусково-лінійної функції $\|f(x) - \Phi_1(x)\|_{C[a,b]} \xrightarrow[h\to 0 (n\to\infty)]{} 0 \quad \{\Phi_1(x)\} \Rightarrow f(x)$.

Розглянемо деякі простори:

1. $H_0 = L_2(a,b)$ - гільбертів простір, в якому скалярний добуток визначається так: $(u,v) = \int_a^b u(x)v(x)dx$, а норма $\|u\|_0 = \sqrt{(u,u)}$.

2. $H_k = W_2^k(a,b)$. Тепер скалярний добуток $(u,v)_k = \sum_{m=0}^{k} \int_a^b u^{(m)}(x) v^{(m)}(x) dx$, а норма $\|u\|_k = \sqrt{\|u\|_0^2 + \ldots + \|u^{(k)}\|^2}$.

Теорема 2 Нехай $f(x) \in H_2 = W_2^2(a,b)$. Тоді $\left\|f^{(k)} - \Phi_1^{(k)}\right\|_0 \le |h|^{2-k} \|f\|_2$, $k = 1,2$.

Зауважимо, що кусково-лінійна інтерполяція негладка, тому на практиці застосовують квадратичні, а найчастіше – кубічні поліноми на кожному проміжку.

6.10. Кусково-кубічна ермітова інтерполяція

Нехай деяка функція $f(x)$ задана в точках x_i своїми значеннями та значеннями похідної: $y_i = f(x_i)$, $y_i' = f'(x_i)$, $i = \overline{0,n}$. Позначимо через $\Phi_3(x)$ функцію, яка буде інтерполювати задану. Тоді

$$\Phi_3(x) = H_3^i(x), \quad x \in [x_{i-1}, x_i] \qquad (1)$$

Неважко написати явний вигляд цього багаточлена $H_3^i(x)$ на проміжку :

$$\begin{array}{c|cc} x_i & y_i & y_i' \\ x_i & y_i & \dfrac{y_{i-1,i} - y_i'}{h_i} \\ x_{i-1} & y_{i-1} & y_{i-1,i} \\ x_{i-1} & y_{i-1} & y_{i-1}' \end{array} \quad \frac{y_i' - 2y_{i-1,i} + y_{i-1}'}{h_i^2},$$

$$H_3^i(x) = y_i + y_i'(x - x_i) + \frac{y_{i-1,i} - y'}{h_i}(x - x_i)^2 + \frac{y_i' - 2y_{i-1,i} + y_{i-1}'}{h_i^2}(x - x_i)^2(x - x_{i-1}).$$

Можна представити кусково-кубічну функцію і в такому вигляді:

$$\Phi_3(x) = \sum_{i=0}^{n} \left[y_i \varphi_i^0(x) + y_i' \varphi_i^1(x) \right]. \qquad (2)$$

Умови інтерполювання: $\Phi_3(x_i)=y_i$, $\Phi_3'(x_i)=y_i'$, $i=\overline{0,n}$. Якщо ці умови підставити в (2), то отримаємо умови на базисні функції:

$$\varphi_i^0(x_j)=\delta_{ij},\ \left(\varphi_i^0\right)'(x_j)=0,\ i,j=\overline{0,n},$$

$$\varphi_i^1(x_j)=0,\ \left(\varphi_i^1\right)'(x_j)=\delta_{ij}.$$

Ці функції кусково-кубічні, тобто $\varphi_i^k(x)\in\pi_3$, $x\in[x_{i-1},x_{i+1}]$, $k=0,1$ (π_3 – множина багаточленів третього степеня), на всіх інших проміжках вони нульові. Нехай $h_i\equiv h$, і позначимо $s=\dfrac{x-x_i}{h}$, $x\in[x_{i-1},x_i]\ \Rightarrow\ s\in[-1,0]$.

1). Введемо $\overline{\varphi}_1^0(s)=\varphi_i^0(x)$, $x\in[x_i,x_{i+1}]$, $s\in[0,1]$. Побудуємо цю функцію. Вона задовольняє умовам :

$$\overline{\varphi}_1^0(0)=1,\ \overline{\varphi}_1^0(1)=0,\ \left(\overline{\varphi}_1^0\right)'(0)=\left(\overline{\varphi}_1^0\right)'(1)=0.$$

Її явний вигляд отримаємо за допомогою таблиці розділених різниць:

0	1				
		0			
0	1		−1		
		−1		1	2
1	0		1		
		0			
1	0				

$$H_3(s)=1+0\cdot s-1\cdot s^2+2s^2(s-1)=2s^3-3s^2+1\equiv\overline{\varphi}_1^0(s)$$

Аналогічно

2). $\overline{\varphi}_2^0(s)=-2s^3-3s^2+1$, $\varphi_i^0(x)=\overline{\varphi}_1^0(s)$, $x\in[x_{i-1},x_i]$, $s\in[-1,0]$;

3). $\overline{\varphi}_1^1(s)=s(s-1)^2$, $h\overline{\varphi}_1^0(s)=\varphi_i^0(x)$, $x\in[x_i,x_{i+1}]$, $s\in[0,1]$;

4). $\overline{\varphi}_2^1(s)=s(s+1)^2$, $h\overline{\varphi}_1^0(s)=\varphi_i^0(x)$, $x\in[x_{i-1},x_i]$, $s\in[-1,0]$.

А тепер будуємо явний вигляд функцій $\varphi_i^k(x)$ для довільного проміжку $x\in[x_{i-1},x_{i+1}]$:

$$\varphi_i^0(x)=\begin{cases}2s^3-3s^2+1,\ x_i\le x\le x_{i+1}\\ -2s^3-3s^2+1,\ x_{i-1}\le x\le x_i\\ 0,\ x_{i-1}>x\vee x>x_{i+1}\end{cases}\qquad \varphi_i^1(x)=\begin{cases}hs(s-1)^2,\ x_i\le x\le x_{i+1}\\ hs(s+1)^2,\ x_{i-1}\le x\le x_i\\ 0,\ x_{i-1}>x\vee x>x_{i+1}\end{cases}$$

де $s=\dfrac{x-x_i}{h}$ (якщо сітка нерівномірна, то в формулах замість h, буде h_i або h_{i+1} на відповідних інтервалах).

Оцінимо $\|f(x)-\Phi_3(x)\|_{C[a,b]}$. Розглянемо для $x\in[x_{i-1},x_i]$:

$$f(x)-\Phi_3(x)=f(x)-H_3^i(x)=\frac{f^{(4)}(\xi)}{4!}(x-x_{i-1})^2(x-x_i)^2.$$

Зразу потрібно зробити припущення, що $f(x)\in C^4[a,b]$. З тих же міркувань, що і для кусково-лінійної функції, максимум знаходиться в точці $\overline{x_i}=\dfrac{x_i+x_{i-1}}{2}$ і тому для модуля похибки маємо:

$$|f(x)-\Phi_3(x)| \le \frac{M_4^i}{24}\left(\frac{h^2}{4}\right)^2 = \frac{M_4^i h^4}{384},\ \|f(x)-\Phi_3(x)\|_{C[a,b]} \le \frac{M_4 h^4}{384}.$$

Звідси отримаємо теорему:

Теорема Якщо функція $f(x)\in C^4[a,b]$ задана в точках x_i своїми значеннями $y_i = f(x_i),\ y_i' = f'(x_i),\ i=\overline{0,n}$, то для кусково-кубічної ермітової інтерполяції $\Phi_3(x)=\sum_{i=0}^{n}\left(y_i\varphi_i^0(x)+y_i'\varphi_i^1(x)\right)$ має місце оцінка

$$\|f(x)-\Phi_3(x)\|_C \le \frac{M_4 h^4}{384}.$$

А для похідної

$$\|f'(x)-\Phi_3'(x)\|_{C[a,b]} \le M\cdot M_4 h^3,$$

де M – стала незалежна від h

Задача 17 Довести оцінку для $\|f'(x)-\Phi_3'(x)\|_{C[a,b]}$.

Порівняємо кусково-лінійну $\Phi_1(x)$ та кусково-кубічну інтерполяцію $\Phi_3(x)$: при згущенні сітки у 2 рази точність лінійної підвищується в 4 рази, а кубічної – у 16 разів, але треба задавати більше даних.

6.11. Кубічні інтерполяційні сплайни [СГ, 140-148], [БЖК, 194-202]

Сплайн (spline) в перекладі означає рейка, якою користувалися креслярі при проведені гладкої кривої, що з‘єднувала задані точки на площині.

Функція $s(x)$ називається *сплайном* степеня m і дефекту k, якщо

1) $s(x)\in\pi_m$ (множина поліномів степеня m) для $x\in[x_{i-1},x_i]$, $i=\overline{1,n}$;

2) $s(x)\in C^{m-k}[a,b]$.

Приклади: 1) $\Phi_1(x)$: $m=1$, $k=$; 2) $\Phi_3(x)$: $m=3$, $k=2$.

Зараз ми побудуємо сплайн, для якого $m=3$, $k=1$.

Функція $s_3(x)=s(x)$ називається *кубічним інтерполяційним природнім сплайном*, якщо

1) $s(x)\in\pi_3$ для $x\in[x_{i-1},x_i]$, $i=\overline{1,n}$, (кубічний); (1)

2) $s(x)\in C^2[a,b]$ (має дефект 1); (2)

3) $s(x_i)=f(x_i)$, $i=\overline{0,n}$, (інтерполює $f(x)$); (3)

4) $s''(a)=s''(b)=0$ (природній). (4)

Умови (4), так звані умови природності, необхідні, щоб разом було $4n$ умови для знаходження $4n$ коефіцієнтів сплайну. Замість них можуть бути такі умови:

$$s''(a)=A\ ,\ \ s''(b)=B \qquad (4а)$$

$$s'(a)=A\ ,\ \ s'(b)=B \qquad (4б)$$

$$s(a)=s(b),\ s'(a)=s'(b),\ s''(a)=s''(b) \qquad (4в)$$

Умови (4в) – це так звані умови періодичності.

Побудуємо природній сплайн. З (1) та (2) маємо

$$s''(x) = m_{i-1}\frac{x_i - x}{h_i} + m_i\frac{x - x_{i-1}}{h_i}\ , \tag{5}$$

де $m_i = s''(x_i)$ і вони є невідомими коефіцієнтами; $h_i = x_i - x_{i-1}$.

Двічі інтегруючи (5), маємо:

$$s(x) = m_{i-1}\frac{(x_i - x)^3}{6h_i} + m_i\frac{(x - x_{i-1})^3}{6h_i} + A_i\frac{x_i - x}{h_i} + B_i\frac{x - x_{i-1}}{h_i}.$$

A_i, B_i знаходимо з умов $s(x_{i-1}) = f_{i-1}$, $s(x_i) = f_i$. Остаточно

$$s(x) = m_{i-1}\frac{(x_i - x)^3}{6h_i} + m_i\frac{(x - x_{i-1})^3}{6h_i} + \\ + \left(f_{i-1} - \frac{m_{i-1}h_i^2}{6}\right)\frac{x_i - x}{h_i} + \left(f_i - \frac{m_i h_i^2}{6}\right)\frac{x - x_{i-1}}{h_i}, \quad x \in [x_{i-1}, x_i]. \tag{6}$$

З (4) маємо $m_0 = m_n = 0$.

Враховуючи, що $s'(x_i - 0) = s'(x_i + 0)$ отримаємо СЛАР для знаходження всіх $m_i = s''(x_i)$:

$$\begin{cases} \dfrac{h_i}{6}m_{i-1} + \dfrac{h_i + h_{i+1}}{3}m_i + \dfrac{h_{i+1}}{6}m_{i+1} = \dfrac{f_{i+1} - f_i}{h_i} - \dfrac{f_i - f_{i-1}}{h_i}, \quad i = \overline{1, n-1} \\ m_0 = m_n = 0. \end{cases} \tag{7}$$

Це тридіагональна СЛАР; її можна розв'язати методом прогонки за $Q = O(N)$ арифметичних операцій.

Задача 18 Написати СЛАР для кубічного інтерполяційного сплайну, якщо $s'(a) = A$, $s'(b) = B$ та розробити алгоритм її розв'язання (тобто написати формули методу прогонки).

Теорема Нехай $f(x) \in C^4[a,b]$, тоді має місце оцінка

$$\left\| f^{(k)}(x) - s^{(k)}(x) \right\|_{C[a,b]} \le M_4 |h|^{4-k},\ k = 0,1,2,$$

де $M_4 = \max\limits_{[a,b]} \left| f^{(4)}(x) \right|$, $|h| = \max\limits_i h_i$.

Введемо клас функцій $U = \left\{ u(x):\ \ u(x) \in W_2^2[a,b],\ u(x_i) = f_i,\ i = \overline{0,n} \right\}$ – це функції досить гладкі і приймають задані значення. Якщо ввести такий функціонал $\Phi(u) = \int_a^b (u''(x))^2\,dx$, то

$$\Phi(s) = \inf_{u \in U} \Phi(u),$$

де $s(x)$ – кубічний природній інтерполяційний сплайн.

Оскільки кривизна графіка кривої $u(x)$ пропорційна $u''(x)$, то це фактично означає, що сплайн має в середньоквадратичному розумінні найменшу кривизну серед всіх функцій $u(x) \in W_2^2[a,b]$, що інтерполюють значення $f(x_i)$.

Для того, щоб не розв'язувати СЛАР (7) інколи будують наближення до сплайну $\tilde{s}(x)$, яке отримується заміною $m_i = s''(x_i)$ на

$$f_{\bar{x}\hat{x},i} \equiv \frac{1}{\hbar_i}\left(\frac{f_{i+1}-f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i}\right) \approx f''(x_i) \approx s''(x_i),$$

де $\hbar_i = \frac{h_i + h_{i+1}}{2}$, причому $f''(x_i) - f_{\bar{x}\hat{x},i} = O\left(|h|^2\right)$. При цьому і $\tilde{s}(x) - s(x) = O\left(h^4\right)$. Відмітимо, що $\tilde{s}(x)$ не є сплайном дефекту 1.

Зауваження 1 Складемо матрицю A розмірності $(n\text{-}1)*(n\text{-}1)$:

$$A = \begin{pmatrix} \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & \dots & 0 \\ \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & \dots & 0 \\ 0 & \frac{h_3}{6} & \frac{h_3+h_4}{3} & \dots & 0 \\ . & . & . & . & . \\ 0 & \dots & 0 & \frac{h_{n-1}}{6} & \frac{h_{n-1}+h_n}{3} \end{pmatrix}$$

і матрицю H розмірності $(n+1)*(n-1)$:

$$H = \begin{pmatrix} \frac{1}{h_1} & -\left(\frac{1}{h_1}+\frac{1}{h_2}\right) & \frac{1}{h_2} & 0 & \dots \\ 0 & \frac{1}{h_2} & -\left(\frac{1}{h_2}+\frac{1}{h_3}\right) & \frac{1}{h_3} & \dots \\ 0 & 0 & \frac{1}{h_3} & -\left(\frac{1}{h_2}+\frac{1}{h_3}\right) & \dots \\ . & . & . & . & . \\ 0 & \dots & 0 & \frac{1}{h_{n-1}} & -\left(\frac{1}{h_{n-1}}+\frac{1}{h_n}\right) \end{pmatrix}$$

Тоді можна записати СЛАР (7) відносно моментів $\vec{m} = (m_1, m_2, \dots, m_{n-1})$ у вигляді:

$$A\vec{m} = H\vec{f}, \text{ де } \vec{f} = (f_0, f_1, \dots, f_n)^T.$$

Зауваження 2 Нагадаємо формулу для інтерполяційного багаточлена Лагранжа $L_n(x) = \sum_{i=0}^{n} f(x_i)\Phi_i^{(n)}(x)$, де $\Phi_i^{(n)}$ - множники Лагранжа. Це представлення інтерполяційного багаточлена Лагранжа по системі функцій $\left\{\Phi_i^{(n)}\right\}$. Для $\Phi_1(x) = \sum_{i=1}^{n} f(x_i)\varphi_i(x)$ маємо представлення по системі кусково-

лінійних функцій $\{\varphi_i(x)\}$. Для $\Phi_3(x) = \sum_{i=1}^{n} (f(x_i)\varphi_i^0(x) + f'(x_i)\varphi_i'(x))$ - представлення по системі $\{\varphi_i^0, \varphi_i'^1\}$.

Аналогічно, якщо представити кубічний сплайн у вигляді

$$s_3(x) = \sum_{i=0}^{n} c_i B_3^i(x),$$

то відповідна система для кубічного сплайну буде $\{B_3^i(x)\}_{i=1}^{n}$. Тут $B_3^i(x)$ - так званий кубічний B_3-сплайн (формула дається, а графік представлено на рис. 6):

$$B_3^i(z) = \frac{1}{6h}\begin{cases} \left(\frac{z - x_{i-2}}{h}\right)^3, & z \in [x_{i-2}, x_{i-1}]; \\ -3\left(\frac{z - x_{i-1}}{h}\right)^3 + 3\left(\frac{z - x_{i-1}}{h}\right)^2 + 3\left(\frac{z - x_{i-1}}{h}\right) + 1, & z \in [x_{i-1}, x_i]; \\ -3\left(\frac{x_{i+1} - z}{h}\right)^3 + 3\left(\frac{x_{i+1} - z}{h}\right)^2 + 3\left(\frac{x_{i+1} - z}{h}\right) + 1, & z \in [x_i, x_{i+1}]; \\ \left(\frac{x_{i+2} - z}{h}\right)^3, & z \in [x_{i+1}, x_{i+2}]; \\ 0, & z < x_{i-2}, z > x_{i+2}. \end{cases}$$

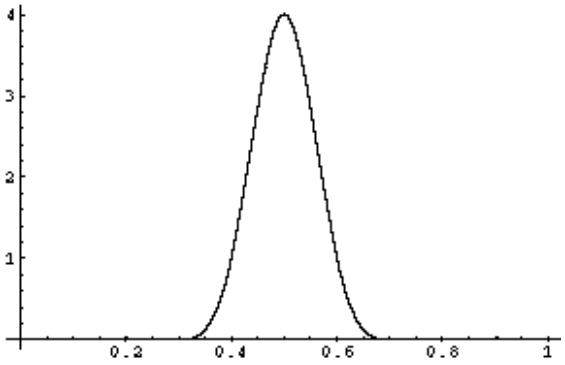


Рис. 6

Задача 19 Показати, що B_3^i - є кубічним сплайном дефекту 1.

Для знаходження коефіцієнтів c_i записується СЛАР з умов інтерполювання.

6.12. Параметричні сплайни

На практиці часто виникає задача побудови кривої по заданим точкам (x_i, y_i), $i = \overline{1,n}$. В цьому випадку використовують сплайни. Якщо відповідна функція $y = f(x)$ однозначна, то сплайн будується за алгоритмом, що розглянуто у пункті 6.12.

Окремо розглянемо випадок, коли точки (x_i, y_i), $i = \overline{1,n}$ в площині (x, y) розташовані у довільний спосіб:

В цьому випадку відповідна функція задається параметрично

$$x = x(t), y = y(t) \;\; t \in [A, B]. \qquad (1)$$

Для значень x_i, $i = \overline{1,n}$ побудуємо кубічний сплайн $s_x(t)$ такий, що $s_x(t_i) = x_i$, $i = \overline{1,n}$, а для y_i, $i = \overline{1,n}$ будуємо сплайн $s_y(t)$, для якого $s_y(t_i) = y_i$, $i = \overline{1,n}$. Тоді параметрична функція

$$\left(s_x(t), s_y(t)\right) t \in [A, B] \qquad (2)$$

називається параметричним сплайном для функції (1).

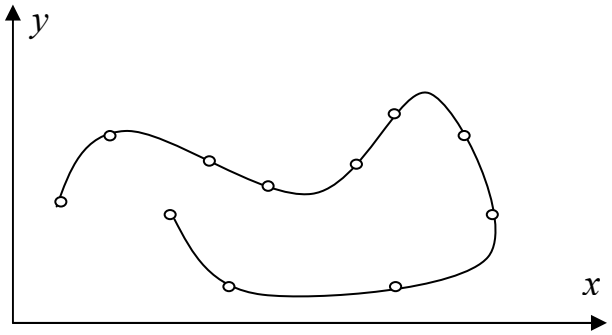


Рис. 7

Стає питання про вибір параметру t. Нехай $t_i = i, i = \overline{1,n}$, тобто для табличних даних (x_i, y_i), $i = \overline{1,n}$ параметром виступає номер точки в площині (x, y). Тоді для параметричного сплайну неперервний параметр t змінюється на інтервалі $t \in [1, n]$.

Побудова сплайнів $s_x(t)$ та $s_y(t)$ здійснюється за алгоритмом наведеним в пункті 6.12 по значенням $f_i = x_i$, $i = \overline{1,n}$ та $f_i = y_i$, $i = \overline{1,n}$.

Розглянемо тепер побудову замкненої гладкої кривої.

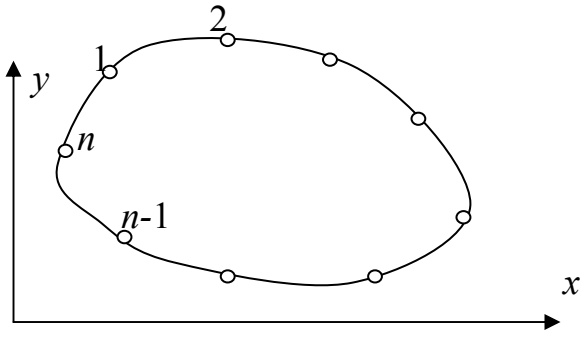


Рис. 8

Параметризуємо її як в попередньому випадку. Відмінність полягає в тому, що тепер функції $x = x(t)$ та $y = y(t)$ періодичні з періодом $T = n$, тобто

$$x(t) = x(t+n),\ y(t) = y(t+n)\ \ \forall t\,.$$

Наприклад, для значень в точках маємо:

$$x_1 = x_{n+1},\ y_1 = y_{n+1} \text{ і } x_0 = x_n,\ y_0 = y_n\,. \tag{3}$$

Побудуємо алгоритм реалізації періодичного параметричного кубічного сплайну. Як і для звичайного сплайну на інтервалі $t \in [t_i, t_{i+1}]\ \forall i$ маємо:

$$s(t) = m_{i-1}\frac{(t_i - t)^3}{6h_i} + m_i\frac{(t - t_{i-1})^3}{6h_i} + \left(f_{i-1} - \frac{m_{i-1}h_i^2}{6}\right)\frac{t_i - t}{h_i} + \left(f_i - \frac{m_i h_i^2}{6}\right)\frac{t - t_{i-1}}{h_i},$$

де $s(t)$ одна з функцій $s_x(t)$ або $s_y(t)$; $f_i = x_i,\ i = \overline{1,n}$ або $f_i = y_i,\ i = \overline{1,n}$; $h_i = t_{i+1} - t_i = 1,\ \forall i$. Для знаходження коефіцієнтів сплайну $m_i = s''(t_i)$ з умови неперервності першої похідної сплайна маємо СЛАР:

$$\begin{cases} \dfrac{h_i}{6}m_{i-1} + \dfrac{h_i + h_{i+1}}{3}m_i + \dfrac{h_{i+1}}{6}m_{i+1} = \dfrac{f_{i+1} - f_i}{h_i} - \dfrac{f_i - f_{i-1}}{h_i}, \quad i = \overline{1,n}; \\ m_0 = m_n,\ m_{n+1} = m_1. \end{cases} \tag{4}$$

Додаткові умови на коефіцієнти m_i випливають з періодичності сплайну та його других похідних.

Системі (4) відповідає матриця розмірності $(n \times n)$:

$$A = \begin{pmatrix} \dfrac{h_1 + h_2}{3} & \dfrac{h_2}{6} & 0 & \ldots & \left\langle \dfrac{h_n}{6} \right\rangle \\ \dfrac{h_2}{6} & \dfrac{h_2 + h_3}{3} & \dfrac{h_3}{6} & \ldots & 0 \\ 0 & \dfrac{h_3}{6} & \dfrac{h_3 + h_4}{3} & \ldots & 0 \\ . & . & . & . & . \\ \left\langle \dfrac{h_1}{6} \right\rangle & \ldots & 0 & \dfrac{h_n}{6} & \dfrac{h_n + h_1}{3} \end{pmatrix},$$

яка є майже тридіагональною: “заважають” два елементи матриці, що виділені кутовими дужками.

Для розв‘язання таких систем застосовують метод циклічної прогонки. Розглянемо алгоритм цього методу для більш загальної системи:

$$\begin{cases} a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i\,, \quad i = \overline{1,n}; \\ y_0 = y_n,\ y_{n+1} = y_1, a_1 = a_n, b_{n+1} = b_1\,. \end{cases} \tag{5}$$

Формули методу [ЛМС, 391-392]:

$1^0.\ \alpha_2 = \frac{b_1}{c_1};\ \beta_2 = \frac{f_1}{c_1};\ \gamma_2 = \frac{a_1}{c_1};$

$2^0.\ z_i = c_i - a_i\alpha_i; \alpha_{i+1} = \frac{b_i}{z_i}; \beta_{i+1} = \frac{f_i + a_i\beta_i}{z_i}; \gamma_{i+1} = \frac{a_i\gamma_i}{z_i},\ i = \overline{2,n};$

$3^0.\ p_{n-1} = \beta_n;\ q_{n-1} = \alpha_n + \gamma_n;$

$4^0.\ p_i = \alpha_{i+1}p_{i+1} + \beta_{i+1};\ q_i = \alpha_{i+1}q_{i+1} + \gamma_{i+1},\ i = \overline{n-2,1};$

$5^0.\ y_n = \frac{\beta_{n+1} + a_{n+1}p_1}{1 - \alpha_{n+1}q_1 - \gamma_{n+1}};$

$6^0.\ y_i = p_i + y_n q_i;\ i = \overline{1,n-1}..$

Метод стійкий ($|\alpha_i| < 1,\ 1 - \alpha_{n+1}\alpha_1 - \gamma_{n+1} \neq 0$), якщо $a_i, b_i > 0, |c_i| > b_i + a_i$. Для системи (4) ці умови виконані.

Метод економний, оскільки кількість арифметичних операцій, що витрачається на його реалізацію, $Q = O(n)$.

Розглянуті в цьому пункті параметричні сплайни мають хороші апроксимативні та екстремальні властивості, тому побудовані по ним криві добре відновлюють задані як при малій, так досить великій кількості точок інтерполювання.

6.13. Застосування інтерполювання [ЛМС 46-48], [БЖК 72-73]

1^0. *Складання таблиць*. Нехай $r_1^i(x)$ залишковий член лінійної інтерполяції по двох сусідніх точках x_{i-1}, x_i:

$$r_1^i(x) = f(x) - L_1^i(x) = \frac{f^n(\xi)}{2!}(x - x_{i-1})(x - x_i).$$

Тоді

$$\left|r_1^i(x)\right| \le \frac{M_2^i}{2} |(x - x_{i-1})(x - x_i)| \le \frac{M_2^i h^2}{8}, \forall x \in [x_{i-1}, x_i].$$

Таким чином $\left\|f(x) - L_1^i(x)\right\|_{C[a,b]} \le \frac{M_2 h^2}{8}$.

Ця оцінка може бути використана при складані таблиць функцій, які при відновлення проміжних значень лінійною інтерполяцією сусідніх значень забезпечують точність ε.

Для того, щоб похибка була меншою за ε потрібно вибрати $h \le \sqrt{\frac{8\varepsilon}{M_2}}$.

Аналогічно, для квадратичного інтерполювання маємо $\left|f(x) - L_2^i(x)\right| \le \frac{M_3 h^3}{9\sqrt{3}} < \varepsilon$. Звідси $h \le \sqrt[3]{\frac{9\sqrt{3}\varepsilon}{M_3}}$.

2^0. *Розв'язування рівнянь*. Нехай необхідно розв'язати рівняння

$$f(x)=\bar{y}. \qquad (1)$$

При $y=0$ маємо рівняння $f(x)=0$. Нехай $\bar{x}$ корінь рівняння (1).

а) *Обернене інтерполювання*. Якщо відома обернена функція $x=x(y)$, то $\bar{x}=x(\bar{y})$. Нехай функція $f(x)$ задана значеннями $y_i=f(x_i)$, $x_i\in[a,b]$. Побудуємо інтерполяційний багаточлен $L_n(y)$ по значеннях $\{y_i,\ x_i\}$, $i=\overline{0,n}$, де y_i вважаються значеннями аргументу, а x_i - значеннями оберненої функції. Тоді наближення до кореня є $x^*=L_n(\bar{y})$.

Оцінимо похибку:

$$\left|\bar{x}-x^*\right|=\left|x(\bar{y})-L_n(\bar{y})\right|\le\frac{M_{n+1}}{(n+1)!}|\omega_n(\bar{y})|, \qquad (2)$$

де $M_{n+1}=\max\limits_{y_{\min}<y<y_{\max}}\left|\dfrac{d^{n+1}}{dy^{n+1}}x(y)\right|$, $\omega_n(y)=(y-y_0)...(y-y_n)$.

Недоліком методу є складність обчислення похідних старших порядків оберненої функції.

б) *Пряме інтерполювання*. Нехай знову відомі $y_i=f(x_i)$, $x_i\in[a,b]$. Тоді замість рівняння (1) розв‘язуємо рівняння

$$L_n(x^*)=y,$$

де $L_n(x)$ інтерполяційний багаточлен по значенням $\{x_i,\ y_i\}$, $i=\overline{0,n}$. Недоліками методу є необхідність розв’язування алгебраїчних рівнянь *n*-го степеня та необхідність вибирати шуканий розв’язок серед *n* коренів багаточлена степеня *n*. Але позитивним є те, що функція є все таки алгебраїчною (а саме багаточленом).

Оцінимо похибку такого способу обчислення кореня. Маємо

$$f(x^*)-L_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(x).$$

Далі $f(x^*)-y=f(x^*)-f(\bar{x})$, звідки $|f(x^*)-f(\bar{x})|\le\dfrac{M_{n+1}}{(n+1)!}|\omega_n(x)|$. Тут тепер $M_{n+1}=\max\limits_{x\in[a,b]}\left|f^{(n+1)}(x)\right|$. За теоремою Лагранжа $f(x^*)-f(\bar{x})=f'(\eta)(x^*-\bar{x})$. Припустимо, що $f'(x)\ne0$. Це означає, що на проміжку $[a,b]$ функція $f(x)$ монотонна. Звідси

$$\left|x^*-\bar{x}\right|\le\frac{|f(x^*)-f(\bar{x})|}{\min\limits_{x\in[a,b]}|f'(x)|}\le\frac{M_{n+1}}{\min\limits_{x\in[a,b]}|f'(x)|}\frac{|\omega_n(x)|}{(n+1)!}.$$

3^0. *Метод інтерполювання побудови характеристичного багаточлена*.

Одним з найпростіших методів побудови характеристичного багаточлена є наступний. Відомо, що багаточлен n-го степеня однозначно визначається своїми значеннями в $n+1$-й точці. Тому для побудови $P_n(\lambda)=\det(A-\lambda E)$ виберемо на проміжку де знаходяться власні значення

(наприклад, $\lambda \in \left[-\|A\|_k, \|A\|_k\right]$, де $k=1$ або $k=\infty$) деякі точки $\lambda_i, i=\overline{0,n}$. За допомогою методу Гаусса для несиметричних матриць або методу квадратних коренів для симетричних матриць обчислимо $P_n(\lambda_i)=\det(A-\lambda_i E)$ і по цих значення за формулою, наприклад, Ньютона побудуємо інтерполяційний багаточлен, який співпадатиме з характеристичним.

Далі розв‘язується рівняння $P_n(\lambda)=0$ одним з відомих методів для нелінійного рівняння. Характерно, що часто для цього використовують метод парабол (обернене інтерполювання по трьох точках, або заміна рівняння n-го степеня в околі кореня на квадратне рівняння за допомогою інтерполяційного багаточлена другого степеня).

Зауважимо, що знаходження власних значень за допомогою характеристичного багаточлена пов‘язана з проблемою нестійкості коренів характеристичного багаточлена відносно похибок обчислення коефіцієнтів цього багаточлена. Тому застосовують цей метод для невеликих розмірностях $n \le 10$ матриці A.

6.14. Тригонометрична інтерполяція [ЛМС 32-34][БЖК 166-170]

Інтерполяція відбувається за системою функцій

$$1, \sin x, \cos x, \sin 2x, \cos 2x, ..., \sin kx, \cos kx, ...$$

Тому

$$T_n(x)=\frac{a_0}{2}+\sum_{j=1}^{n}(a_k \cos jx + b_k \sin jx), \qquad (1)$$

що є відрізком тригонометричного ряду Фур‘є. Щоб знайти $T_n(x)$ потрібно визначити $2n+1$ коефіцієнт, а значить задати $(2n+1)$ значень періодичної з періодом 2π функції $y_i=f_i(x), i=\overline{0,2n}$.

Покажемо, що

$$T_n(x)=\sum_{i=0}^{2n} f(x_i)\Phi_i(x), \text{ де } \Phi_i(x)=\prod_{\substack{j=0\\ j\neq i}}^{2n}\frac{\sin\frac{x-x_j}{2}}{\sin\frac{x_i-x_j}{2}}, \qquad (2)$$

тобто $T_n(x_k)=f(x_k)$, та $\Phi_i(x_k)=\delta_{ik}$. Дійсно

$$\Phi_i(x_k)=\prod_{\substack{j=0\\ j\neq i}}^{2n}\frac{\sin\frac{x_k-x_j}{2}}{\sin\frac{x_i-x_j}{2}}=0,\ i\neq k;\quad \Phi_i(x_i)=\prod_{\substack{j=0\\ j\neq i}}^{2n}\frac{\sin\frac{x_i-x_j}{2}}{\sin\frac{x_i-x_j}{2}}=1.$$

Таким чином за допомогою формули (2) ми уникли необхідності підраховувати коефіцієнти Фур‘є a_k, b_k.

Нехай функція $f(x)$ є парною та неперервною на проміжку $[-\pi,\pi]$. Тоді по значенням в (n+1)-й точці, $y_i = f_i(x), i = \overline{0,n}$ $x_i \in [0,\pi]$ можна побудувати парний тригонометричний багаточлен

$$T_n(x) = \sum_{i=0}^{n} f(x_i) \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{\cos x - \cos x_i}{\cos x_i - \cos x_j}. \qquad (3)$$

Якщо ж функція є непарною на проміжку $[-\pi,\pi]$, то по значенням в n точках $y_i = f_i(x), i = \overline{1,n}$. $x_i \in [0,\pi]$ можна побудувати непарний інтерполяційний багаточлен

$$T_n(x) = \sum_{i=1}^{n} f(x_i) \frac{\sin x}{\sin x_i} \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{\cos x - \cos x_i}{\cos x_i - \cos x_j}. \qquad (4)$$

Задача 20 Показати, що тригонометричні багаточлени (3), (4) є інтерполюючими для функції $f(x)$. Яке значення функції інтерполює (4) при $x = 0$? Чому?

6.15. Двовимірна інтерполяція [БЖК, 226-228], [Калиткин, 47-51]

Побудова багаточлена для функції від двох змінних $z = f(x,y)$, що інтерполює значення $z_i = f(x_i, y_i)$ в точках $A_i(x_i, y_i)$, пов‘язана з такими труднощами.

1) Якщо в одновимірному випадку кількість вузлів та степінь багаточлена пов‘язані простою залежністю: n+1 точка x_i дозволяють побудувати багаточлен n -го степеня $L_n(x)$, то в двовимірному випадку багаточлен n -го степеня від двох змінних

$$P_n(x,y) = \sum_{0\le k+m\le n} a_{km} x^k y^m,$$

має $N = \dfrac{(n+1)(n+2)}{2}$ коефіцієнтів a_{km}. Тому необхідно задати значення в точках $A_i(x_i, y_i)$, $i = \overline{1,N}$.

2) Не всяке розташування вузлів допустиме. Якщо розглянути умови інтерполювання

$$P_n(x_i, y_i) = \sum_{0\le k+m\le n} a_{km} x_i^k y_i^m = z_i,$$

то для розв’язності цієї СЛАР необхідно виконання умови $\det B \neq 0$, де матриця B має коефіцієнти:

$$b_{ij} = (x_i^k y_i^m),\ 0 \le k+m \le n,\ i = \overline{1,N}$$

Ця умова, наприклад, для лінійної інтерполяції n=1 та N=3 вимагає, щоб вузли $A_i(x_i, y_i)$ не лежали на одній прямій. Якщо n=2, то N=6 і необхідно розглядати точки, які не лежать на деякій кривій другого порядку і т.д.

Розглянемо випадки, коли можна записати багаточлен для двовимірної інтерполяції в явному вигляді.

Нехай область, в якій інтерполюється функція є прямокутником $\overline{\Omega} = \{(x,y) : 0 \le x \le L_1, 0 \le y \le L_2\}$. Введемо сітку $x_i = ih_x, i = \overline{0,N}, h_x = \frac{L_1}{N}$, $y_j = jh_y, h_y = \frac{L_2}{M}$. Тоді інтерполяційний багаточлен має вигляд

$$P(x,y) = \sum_{i=0}^{N}\sum_{j=0}^{M} f(x_i,y_i) \prod_{\substack{p=0 \\ p\neq i}}^{N} \prod_{\substack{q=0 \\ q\neq j}}^{M} \frac{(x-x_i)}{(x_p-x_i)} \frac{(y-y_j)}{(y_q-y_j)} \qquad (1)$$

Розглянемо випадок , коли N=M=1. Тоді

$$P_{11}(x,y) = f(x_0,y_0)\frac{x-x_1}{x_0-x_1}\frac{y-y_1}{y_0-y_1} + f(x_0,y_1)\frac{x-x_1}{x_0-x_1}\frac{y-y_0}{y_1-y_0} +$$

$$+ f(x_1,y_0)\frac{x-x_0}{x_1-x_0}\frac{y-y_1}{y_0-y_1} + f(x_1,y_1)\frac{x-x_0}{x_1-x_0}\frac{y-y_0}{y_1-y_0} = \qquad (2)$$

$$= a_0 + a_1 x + a_2 y + a_{12} xy.$$

Це так звана білінійна інтерполяція, тобто лінійна по кожній окремій змінній.

Формула (1) являє собою приклад інтерполювання на всій області. В одновимірному випадку при великих степенях багаточлена отримують погане наближення через розбіжність процесу інтерполювання. Так ж картина має місце і в двовимірному випадку. Тому найчастіше застосовують кусково-поліноміальну апроксимацію.

Коротко наведемо деякі відомості про кусково-поліноміальне інтерполювання з теорії методу скінчених елементів розв‘язання крайових задач для диференціальних рівнянь в частинних похідних.

Нехай область $\Omega \subset R^2$ - багатокутник в площині. Представимо її у вигляді $\Omega = \bigcup_{i=1}^{n} K_i$, $K_i \in T_h$. T_h- називається “тріангуляцією” області Ω, а K_i задовольняють умовам:

1) K_i - багатокутник $\overset{0}{K}_i \neq \varnothing$;
2) $\overset{0}{K}_i \cap \overset{0}{K}_j = \varnothing, \forall K_i, K_j \in T_h$;
3) $F = K_i \cap K_j \neq \varnothing$, F- сторона K_i, K_j;
4) $diam K_i \le h$ - h-характеристика щільності розбиття.

Найчастіше K_i це трикутники або прямокутники.

Нехай $v \in X$ - функція, яку ми будемо інтерполювати. Позначимо X_h простір, що апроксимує X, а його елементи $v_h \in X_h$. Причому звуження цієї функції на область K_i, тобто $v_h\big|_{K_i}$ є поліномом.

Позначимо Π_k, $k \ge 0$- простір багаточленів степеня k по сукупності змінних x, y; його розмірність $\dim \Pi_k = \frac{(k+1)(k+2)}{2}$. Нехай Θ_k, $k \ge 0$-

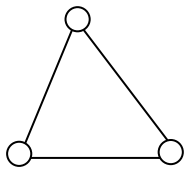
простір багаточленів степеня по кожній окремій змінний x, y; його розмірність $\dim \Theta_k = (k+1)^2$. Наприклад, $P_1(x,y) = a_0 + a_1 x + a_2 y \in \Pi_1$ - поліном степеня 1 по x, y, а $Q_1(x,y) = a_0 + a_1 x + a_2 y + a_{12} xy \in \Theta_1$ лінійна по кожній окремій змінній.

Позначимо через $X_h^k = \left\{ v_h \in C^0(\Omega) : v_h|_k \in \Pi_k, \forall k \in T_h \right\}$ - простір інтерполянтів при розбитті області на трикутники, а $Y_h^k = \left\{ v_h \in C^0(\Omega) : v_h|_k \in \Theta_k, \forall k \in T_h \right\}$ - при розбитті на прямокутники.

Приклад 1 Утворимо $X_h^1, k = 1$. Будуємо багаточлен 1-го степеня по двох змінних. Оскільки $\dim \Pi_1 = 3$, то для цього треба задати значення функції в трьох точках. Точки, які задано - $A_i, i = \overline{1,3}$ вибираємо вершинами трикутника, як на малюнку. Тоді поліном першого степеня $z = P_1(x,y)$ є розв'язком такого рівняння відносно z:

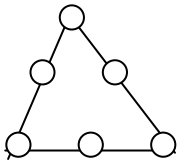
$$\begin{vmatrix} 1 & x & y & z \\ 1 & x_1 & y_1 & f_1 \\ 1 & x_2 & y_2 & f_2 \\ 1 & x_3 & y_3 & f_3 \end{vmatrix} = 0.$$

Тут $f_i = f(x_i, y_i), i = \overline{1,3}$.

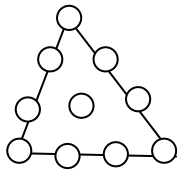


Задача 21 Знайти явний вигляд $z = P_1(x, y)$ - інтерполяційного багаточлена по значенням в точках $A_i = (x_i, y_i), i = 1,2,3$.

Приклад 2 Для $X_h^2, k = 2, \dim \Pi_2 = 6$. Треба задати 6 значень, щоб забезпечити однозначність наближення. Тому вибираємо точки інтерполювання так:



Приклад 3 $X_h^3, k = 3, \dim \Pi_2 = 10$. Потрібно задати 10 точок, як на наступному малюнку:

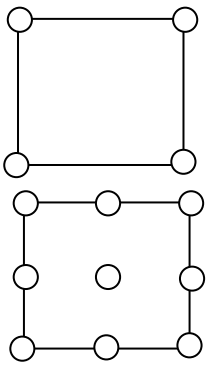


Точка в середині трикутника - є центром його мас.

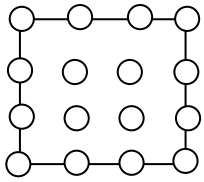
Приклад 4 $Y_h^1, k = 1, \dim \Theta_1 = 4$.

Формула для $Q_1(x,y)$ наведена в (2).

Приклад 5 $Y_h^2, k=2, \dim \Theta_2 = 9$



Приклад 6 $Y_h^3, k=3, \dim \Theta_3 = 16$



Нехай $X = W_2^m(\Omega) = H^m(\Omega)$ - це простір з нормою $\|v\|_m^2 = \sum_{k=0}^{m} |v|_k^2$:

$$|v|_k^2 = \int_\Omega (D^m v)^2 \, d\Omega = \int_\Omega [(D_{x^k}^k v)^2 + (D_{x^{k-1}y}^k v)^2 + ... + (D_{y^k}^k v)^2] d\Omega;$$

$$D_{x^k}^k v = \frac{\partial^k v}{\partial x^k}; \qquad D_{x^{k-1}y}^k v = \frac{\partial^k v}{\partial x^{k-1} \partial y}; ...$$

Якщо $\|v\|_m \le M < \infty$, то $v \in W_2^m(\Omega)$, класу функцій інтегрованих з квадратом до *m*- ї похідної.

Розглянемо розбиття на трикутники. Накладемо обмеження на них. Розбиття T_h називається *регулярним*, якщо $\exists \sigma \ge 1$ таке, що

$$\max_{k \in T_h} \frac{h_k}{\rho_k} \le \sigma, \quad h_k = diam K, \; S \subset K, \; \rho_k = mes S. \text{ (див. рис. 9)}$$

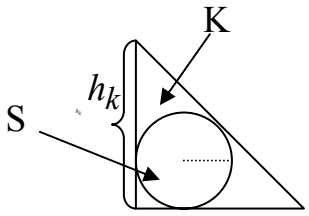


Рис. 9

Якщо $\frac{h_k}{\rho_k} >> 1$, то K вироджується в пряму і це погано.

Теорема Нехай $v \in W_2^{l+1}(\Omega), 1 \le l \le k$, T_h - регулярна тріангуляція, $v_h \in X_h = \{v_h : v_h|_k = P_k\}$. Тоді

$$|v - v_h|_m \le C h^{l+1-m} |v|_{l+1}, m = 0,1, k \ge 1.$$

Наприклад, для $k=1, m=0 \Rightarrow l=1: \|v - v_h\|_{L_2(\Omega)} = \|v - v_h\|_0 \le C h^2 |v|_2$. Якщо $k=3, l=3$, то $\|v - v_h\|_0 \le h^4 |v|_4$.

Ця теорема дозволяє стверджувати збіжність процесу інтерполювання. І чим більше степінь полінома на кожному елементі тим вища швидкість збіжності.

Узагальнимо результат теореми на область з гладкою границею (див. рис. 10). Для цього вибираємо точки на границі і будуємо вписаний багатогранник. Його тріангулюємо. Далі на кожному трикутнику будуємо інтерполянт. В результаті отримуємо $v_h \in X_h^k$.

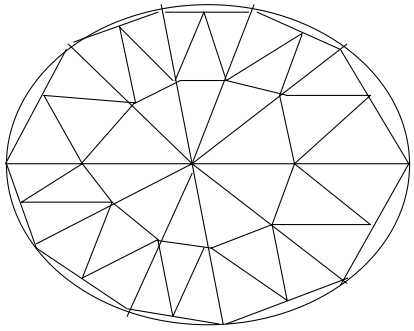


Рис. 10

Тоді для $k=1, l=1: \|v - v_h\|_0 \le ch^{\frac{3}{2}} |v|_2$.

7. Чисельне диференціювання

7.1. Побудова формул чисельного диференціювання [СГ, 188-190], [ЛМС, 48-51], [БЖК, 73-79]

Задача чисельного диференціювання виникає у випадку коли необхідно обчислити похідну функції, значення якої задані таблицею. Нехай задано

$$f_i = f(x_i), \quad i = \overline{0,n}, \quad x_i \in [a,b].$$

Проінтерполюємо ці значення. Тоді

$$f(x) = L_n(x) + r_n(x), \tag{1}$$

де залишковий член у формі Ньютона має вигляд:

$$r_n(x) = f(x; x_0, ..., x_n)\omega_n(x), \;\; \omega_n(x) = \prod_{i=0}^{n}(x - x_i).$$

Звідси

$$f^{(k)}(x) = L_n^{(k)}(x) + r_n^{(k)}(x). \tag{2}$$

За наближене значення похідної в точці x беремо $f^{(k)}(x) \approx L_n^{(k)}(x), x \in [a,b]$.

Оцінимо похибку наближення - $r_n^{(k)}(x)$. За формулою Лейбніца:

$$r_n^{(k)}(x) = \sum_{j=0}^{k} C_k^j f^{(j)}(x; x_0, ..., x_n)\omega_n^{(k-j)}(x). \tag{3}$$

З властивості розділених різниць маємо для $f(x) \in C^{n+k+1}[a,b]$:

$$f^{(j)}(x; x_0; ...; x_n) = j! f(\underbrace{x; ...x}_{j+1}; x_0; ...; x_n) = \frac{j!}{(n+j+1)!} f^{(n+j+1)}(\xi_j), \; \xi_j \in [a,b].$$

Остаточно вираз для похибки наближення похідної має вигляд:

$$r_n^{(k)}(x)=\sum_{j=0}^{k}\frac{k!}{(k-j)!(n+j+1)!}f^{(n+j+1)}(\xi_i)\omega_n^{(k-j)}(x), \qquad (4)$$

Оцінка похибки матиме вигляд:

$$\left|f^{(k)}(x)-L_n^{(k)}(x)\right|\le M\sum_{j=0}^{k}\frac{k!}{(k-j)!(n+j+1)!}\left|\omega_n^{(k-j)}(x)\right|,$$

де $M=\max\limits_{0\le j\le k}\max\limits_{x\in[a,b]}\left|f^{(n+j+1)}(x)\right|$.

Нагадаємо, що процес інтерполювання розбіжний. Крім того, якщо $k>n$, то $L_n^{(k)}(x)\equiv 0$. Тому не можна брати великими значення n та k. Як правило $k=1,2$, іноді $k=3,4$. Відповідно, $n=k$, або $n=k+1$, або $n=k+2$.

Подивимося як залежить порядок збіжності процесу чисельного диференціювання від кроку. Нехай $x_i=x_0+ih,\quad h>0-$ крок. Тоді за умови $x_n-x_0=O(h)$:

$$\omega_n(x)=(x-x_0)...(x-x_n)=O\left(h^{n+1}\right),\quad x\in[x_0,x_n].$$

Перша похідна від $\omega_n(x)$ має порядок на одиницю менше, тобто

$$\omega_n'(x)=O\left(h^n\right).$$

Далі

$$r_n^{(k)}(x)=O\left(h^{n+1-k}\right),\ \text{тому}\ f^{(k)}(x)-L_n^{(k)}(x)=O\left(h^{n+1-k}\right).$$

При умові $n\ge k$ останній вираз збігається до нуля, тобто

$$f^{(k)}(x)-L_n^{(k)}(x)\underset{h\to 0}{\to}0. \qquad (5)$$

Далі

$$r_n^{(k)}(x)=\underbrace{f(x;x_0;...;x_n)\omega_n^{(k)}(x)}_{O(h^{n+1-k})}+\underbrace{\sum_{j=1}^{k}C_k^j f^{(j)}(x;x_0;...;x_n)\omega_n^{(k-j)}(x)}_{O(h^{n+2-k})}.$$

Якщо

$$\omega_n^{(k)}(\bar{x})=0,\ \text{то}\ r_n^{(k)}(\bar{x})=O\left(h^{n+2-k}\right). \qquad (6)$$

Точки $x=\bar{x}$ називаються *точками підвищеної точності формул чисельного диференціювання.*

Приклад 1 Виведемо формули чисельного диференціювання для $k=1,n=1$. Виберемо точки x_0, $x_1=x_0+h$ і інтерполяційний багаточлен має вигляд:

$$L_1(x)=f_0+(x-x_0)\frac{f_1-f_0}{h}.$$

Для похідної отримаємо вираз:

$$f'(x)\approx L_1'(x)=\frac{f_1-f_0}{h},\ x\in[x_0,x_1].$$

Розписавши за формулою Тейлора, отримаємо вираз для похибки:

$$r_1'(x)=\frac{f^{(3)}(\xi_1)}{3!}(x-x_0)(x-x_1)+\frac{f^{(2)}(\xi_0)}{2!}(2x-x_1-x_0)=O(h).$$

Якщо $2\bar{x} - x_1 - x_0 = 0$, то $r_1'(\bar{x}) = O(h^2)$. Тобто $\bar{x} = \dfrac{x_1 + x_0}{2}$ – точка підвищеної точності. Більш точно (див. приклад 3)

$$|r_1'(\bar{x})| \le \frac{h^2}{24} M_3, \text{ де } M_3 = \max_{x\in[a,b]} \left| f^{(3)}(x) \right|.$$

Приклад 2 Аналогічно виведемо формули чисельного диференціювання для $k = 1, n = 2$. Виберемо точки $x_0, x_1 = x_0 + h, x_2 = x_0 + 2h$. Інтерполяційний поліном має вигляд:

$$L_2(x) = f_0 + (x - x_0)\frac{f_1 - f_0}{h} + (x - x_0)(x - x_1)\frac{f_2 - 2f_1 + f_0}{2h^2}.$$

Тоді замінимо $f'(x) \approx L_2'(x) = \dfrac{f_1 - f_0}{h} + (2x - x_0 - x_1)\dfrac{f_2 - 2f_1 + f_0}{2h^2}, \quad x \in [x_0, x_2]$.

Якщо сюди підставити $x = x_0$, то отримаємо $f'(x_0) \approx \dfrac{-f_2 + 4f_1 - 3f_0}{2h}$. Для точки $x = x_1$ маємо $f'(x_1) \approx \dfrac{f_2 - f_0}{2h} = f_{\underset{x,1}{0}}$. Для точки $x = x_2$ маємо $f'(x_2) \approx \dfrac{f_0 - 4f_1 + 3f_2}{2h}$, $x \in [x_0, x_2]$. Для похибки маємо оцінку $r_2'(x) = O(h^2)$.

Позначимо для $x \in [x_i, x_{i+1}]$ $\quad f_{x,i} = \dfrac{f_{i+1} - f_i}{h} \approx f'(x)$ (різницева похідна вперед);

для $x \in [x_{i-1}, x_i]$ - $f_{\bar{x},i} = \dfrac{f_i - f_{i-1}}{h} \approx f'(x)$ (різницева похідна назад);

для $x \in [x_{i-1}, x_{i+1}]$ - $f_{\underset{x,i}{0}} = \dfrac{f_{i+1} - f_{i-1}}{2h} \approx f'(x)$ (центральна різницева похідна).

Замість $f'(x_i)$ можна взяти будь-яке із значень: $f_{x,i}$, $f_{\bar{x},i}$ або $f_{\underset{x,i}{0}}$.

Задача 21 Знайти точки підвищеної точності формул чисельного диференціювання для $k = 1, n = 2$ і оцінити похибку в цих точках.

Приклад 3 При $n = 1, k = 1$ оцінимо точність формул чисельного диференціювання за формулою Тейлора.

а) Нехай $f(x) \in C^2[a,b]$. Тоді

$$f'(x_0) - \frac{f(x_0 + h) - f(x_0)}{h} = f'(x_0) - \frac{1}{h}\left[f_0 + hf_0' + \frac{h^2}{2} f''(\xi) - f_0 \right] = -\frac{h}{2} f''(\xi),$$

$$\left| f'(x_0) - \frac{f_1 - f_0}{h} \right| \le \frac{M_2 h}{2}, \; M_2 = \max_{[x_0, x_1]} |f''(\xi)|.$$

б) Нехай $f(x) \in C^3[a,b]$. Тоді, розписавши розклад по формулі Тейлора до третьої похідної, маємо оцінку:

$$f'(\bar{x})-\frac{f(x_0+h)-f(x_0)}{h}=f'(\bar{x})-\frac{1}{h}[f(\bar{x})+\frac{h}{2}f'(\bar{x})+\frac{h^2}{8}f''(\bar{x})+\frac{h^3}{48}f'''(\xi)-$$

$$-f(\bar{x})+\frac{h}{2}f'(\bar{x})-\frac{h^2}{8}f''(\bar{x})+\frac{h^3}{48}f'''(\eta)]=-\frac{h^2}{24}f'''(\varsigma).$$

$$\left|f'(\bar{x})-\frac{f_1-f_0}{h}\right|\le\frac{h^2M_3}{24},\bar{x}=\frac{x_1+x_0}{2}\ .$$

Задача 22 Показати, що якщо $f(x)\in C^3[a,b]$, то $\left|f'(x_1)-\frac{f_2-f_0}{2h}\right|\le\frac{M_3h^2}{6}$.

Приклад 4 При $n=2,k=2$ маємо:

$$L_2(x)=f_{i-1}+\frac{f_i-f_{i-1}}{h}(x-x_{i-1})+\frac{f_{i+1}-2f_i+f_{i-1}}{h^2}(x-x_{i-1})(x-x_i);$$

$$L_2''(x)=\frac{f_{i+1}-2f_i+f_{i-1}}{h^2}.$$

Для $f(x)\in C^4[a,b]$ оцінимо точність формул чисельного диференціювання за формулою Тейлора:

$$f''(x_1)-\frac{f_2-2f_1+f_0}{h^2}=f''(x_1)-\frac{f(x_1+h)-2f(x_1)+f(x_1-h)}{h^2}=$$

$$=f''(x_1)-\frac{1}{h^2}[f_1+hf_1''+\frac{h^2}{2}f_1''+\frac{h^3}{6}f_1'''+\frac{h^4}{24}f^{(4)}(\xi_1)-2f_1+$$

$$+f_1-hf_1''+\frac{h^2}{2}f_1''-\frac{h^3}{6}f_1'''+\frac{h^4}{24}f^{(4)}(\xi_2)]=\qquad .$$

$$=\frac{h^2}{12}f^{(4)}(\xi)\qquad \xi_1,\xi_2,\xi\in[x_0,x_2]$$

Отже, $\left|f_1''-\frac{f_2-2f_1+f_0}{h^2}\right|\le\frac{M_4h^2}{12}$.

Задача 23 Побудувати формулу чисельного диференціювання $k=2,n=2$ у випадку нерівновіддалених вузлів: x_0, $x_1=x_0+h_1$, $x_2=x_1+h_2$. Оцінити точність формули. Знайти точки підвищеної точності оцінити похибку.

Крім інтерполяційних формул для чисельного диференціювання можна застосовувати сплайни. Нехай $f_i=f(x_i)$. Побудуємо інтерполяційний сплайн першого степеня $s_1(x)$, для якого має місце оцінка $\left|f^{(k)}(x)-s_1^{(k)}(x)\right|=O(h^{2-k}),k=0,1$. Звідси при $k=1$ маємо $\left|f'(x)-s_1'(x)\right|=O(h)$.

Для кубічного інтерполяційного сплайну $s_3(x)$ маємо для першої та другої похідних

$$\left|f^{(k)}(x)-s_3^{(k)}(x)\right|=O(h^{4-k}),\ k=1,2.$$

7.2. Про обчислювальну похибку чисельного диференціювання [СГ, 186-188], [БЖК, 80-81]

Нехай значення функції обчисленні з деякою похибкою. Постає питання про вплив цих похибок на значення похідних обчислених за формулами чисельного диференціювання.

Перед цим зробимо зауваження про вплив збурення функції на значення звичайних похідних.

Нехай $f(x)\in C^1[a,b]$ і її збурення має вигляд:

$$\tilde{f}(x)=f(x)+\frac{1}{n}\sin(\omega x).$$

При $n\to\infty$ маємо $\left\|f(x)-\tilde{f}(x)\right\|_{C[a,b]}=\frac{1}{n}\to 0$ звідси $\tilde{f}(x)\underset{n\to\infty}{\Rightarrow} f(x)$. Таким чином це малі збурення. Маємо $\tilde{f}'(x)=f'(x)+\frac{\omega}{n}\cos(\omega x)$. Нехай $\omega=n^2$, тоді

$$\left\|\tilde{f}'-f'\right\|_{C[a,b]}=\frac{|\omega|}{n}=n\xrightarrow[n\to\infty]{}\infty.$$

Цей приклад ілюструє нестійкість оператора диференціювання. Є сподівання, що ця нестійкість має місце і для чисельного диференціювання.

Нехай $\tilde{f}_i=f_i+\delta_i,\quad f_i=f(x_i),\ i=\overline{0,n}\ |\delta_i|\le\delta$. Розглянемо вплив похибок δ_i на конкретних формулах чисельного диференціювання.

Приклад 1 Оцінімо вплив збурень на похибку обчислення першої похідної $n=1,k=1$.

$$f_i'-\frac{\tilde{f}_i-\tilde{f}_{i-1}}{h}=f_i'-\frac{f_i-f_{i-1}}{h}-\frac{\delta_i-\delta_{i-1}}{h},$$

$$\left|f_i'-\frac{\tilde{f}_i-\tilde{f}_{i-1}}{h}\right|\le\left|f_i'-\frac{f_i-f_{i-1}}{h}\right|+\left|\frac{\delta_i-\delta_{i-1}}{h}\right|\le\frac{M_2h}{2}+\frac{2\delta}{h}\xrightarrow[h\to 0]{}\infty,\qquad(1)$$

Таким чином, як і для аналітичного диференціювання, маємо некоректність: при малих збуреннях $|\delta_i|\le\delta$ можуть бути як завгодно великі похибки, якщо $\frac{\delta}{h}\to\infty$ при $h\to 0$.

Мінімізуємо вплив цих збурень. Позначимо

$$\varphi(h)=\frac{M_2h}{2}+\frac{2\delta}{h}.$$

Тоді мінімум цієї функції досягається для таких h:

$$\varphi'(h)=\frac{M_2}{2}-\frac{2\delta}{h^2}=0,\ h_0=2\sqrt{\frac{\delta}{M_2}}.$$

При такому значенні h оцінка похибки (1) така:

$$\varphi(h_0)=2\sqrt{M_2\delta}=O\left(\delta^{\frac{1}{2}}\right)\underset{\delta\to 0}{\to}0.$$

Приклад 2 Подивимося на вплив збурень на похибку обчислення першої похідної при використанні центральної різницевої похідної:

$$\left| f_i' - \frac{\tilde{f}_{i+1} - \tilde{f}_{i-1}}{2h} \right| \le \left| f_i' - \frac{f_{i+1} - f_{i-1}}{2h} \right| + \left| \frac{\delta_{i+1} - \delta_{i-1}}{2h} \right| \le \frac{M_3 h^2}{6} + \frac{\delta}{h} = \varphi(h).$$

З рівняння $\varphi'(h) = \frac{M_3 h}{3} - \frac{\delta}{h^2} = 0$ маємо: $h_0^3 = \frac{3\delta}{M_3}$, $h_0 = \sqrt[3]{\frac{3\delta}{M_3}}$. Отже,

$$\varphi(h_0) = \frac{M_3}{6}\sqrt[3]{\frac{9\delta^2}{M_3^2}} + \frac{\delta}{\sqrt[3]{\frac{3\delta}{M_3}}} = \frac{1}{2}\sqrt[3]{\frac{M_3\delta^2}{3}} + \sqrt[3]{\frac{\mathrm{M}_3\delta^2}{3}} = \frac{3}{2}\sqrt[3]{\frac{M_3\delta^2}{3}} = O\left(\delta^{\frac{2}{3}}\right).$$

Таким чином швидкість збіжності при $\delta \to 0$ похибки формули чисельного диференціювання центральною похідною вища ніж для формули з прикладу 1 (похідна вперед або назад).

Задача 24 Дослідити похибку чисельного диференціювання для $n = 2, k = 2$, вибрати оптимальний крок h_0, дати оцінку $\varphi(h_0)$.

8. Апроксимування функцій

8.1. Постановка задачі апроксимації [ЛМС, 8-13], [БЖК, 160-161].

Наближення функцій застосовують у випадках, якщо

- функція складна (трансцендентна або є розв‘язком складної задачі) і її замінюють функцією, яка легко обчислюється (найчастіше, поліномом);
- необхідно побудувати функцію неперервного аргументу для функції, яка задана своїми значеннями (таблична);
- таблична функція наближається табличною ж функцією (згладжування).

Інтерполювання не кращий спосіб наближення функцій через розбіжність цього процесу для поліномів. Тим більше доцільність застосування інтерполювання сумнівна, якщо функція таблична, а її значення неточні. Потрібно будувати апроксимуючу функцію з інших міркувань.

Найбільш загальний принцип: наблизити $f(x)$ функцією $\Phi(x)$ так, щоб досягалася деяка задана точність ε:

$$\|f(x) - \Phi(x)\| < \varepsilon.$$

Але розв’язок в такій постановці може не існувати або бути не єдиним.

Загальна постановка задачі наближення така. Нехай маємо елемент f лінійного нормованого простору R. Побудуємо підпростір M_n, в якому елементи є лінійною комбінацією

$$\Phi = \sum_{i=0}^{n} c_i \varphi_i \in M_n \subset R \qquad (1)$$

по елементах лінійно незалежної системи

$$\{\varphi_i\}_{i=0}^{\infty},\ \varphi_i \in R \qquad (2)$$

Відхилення $\Phi \in M_n$ від $f \in R$ є число

$$\Delta(f,\Phi) = \|f - \Phi\|.$$

Позначимо

$$\inf_{\Phi \in M_n} \|f - \Phi\| = \Delta(f).$$

Елемент Φ_0 такий, що

$$\Delta(f,\Phi_0) = \|f - \Phi_0\| = \inf_{\Phi \in M_n} \|f - \Phi\| = \Delta(f), \qquad (3)$$

називається *елементом найкращого наближення* (ЕНН).

Ясно, що умову точності треба перевіряти на цьому елементі. У випадку її невиконання треба збільшувати кількість елементів n в (1).

Теорема 1 Для будь-якого лінійного нормованого простору R існує елемент найкращого наближення $\Phi_0 \in M_n$.

◁ Введемо $F(\vec{c}) = F(c_0, c_1, ..., c_n) = \|f - \Phi\| = \left\| f - \sum_{i=0}^{n} c_i \phi_i \right\|$. Це неперервна функція аргументів $\vec{c} = (c_0, c_1, ..., c_n)$. Для елементів, які задовольняють умові

$$\|\Phi\| > 2\|f\| \quad f \in R_1 \quad \Phi \in M_n, \qquad (4)$$

маємо

$$F(\vec{c}) = \|f - \Phi\| \geq \|\Phi\| - \|f\| > 2\|f\| - \|f\| = \|f\| > \Delta(f).$$

Значить ЕНН $\Phi_0 \in \{\Phi : \|\Phi\| \leq 2\|f\|\} = \overline{U} \subset M_n$. За теоремою Кантора $\exists \Phi_0$, де $F(\vec{c})$ досягає мінімуму. Причому $\|f - \Phi_0\| \leq \|f - \Phi\|$. ▷

Елементів найкращого наближення в лінійному нормованому просторі може бути і декілька.

Простір R називається *строго нормованим,* якщо з умови

$$\|f + g\| = \|f\| + \|g\|, \|f\| \neq 0, \|g\| \neq 0$$

випливає, що $\exists \lambda \neq 0$ таке, що

$$g = \lambda f \qquad (5)$$

Теорема 2 Якщо простір R строго нормований, то елемент найкращого наближення Φ_0 єдиний.

◁ Доведення від супротивного. Нехай існують $\Phi_0^{(1)} \neq \Phi_0^{(2)}$ – два елементи найкращого наближення. Візьмемо $\alpha \in [0,1]$, тоді

$$\Delta(f) \leq \left\| f - \alpha \Phi_o^{(1)} - (1-\alpha)\Phi_0^{(2)} \right\| = \left\| \alpha\left(f - \Phi_0^{(1)}\right) + (1-\alpha)\left(f - \Phi_0^{(2)}\right) \right\| \leq$$

$$\leq \alpha \left\| f - \Phi_0^{(1)} \right\| + (1-\alpha)\left\| f - \Phi_0^{(2)} \right\| = \alpha\,\Delta(f) + (1-\alpha)\Delta(f) = \Delta(f)$$

Тобто всі "≤" можна замінити на "=".Отримаємо

$$\left\| \alpha\left(f - \Phi_0^{(1)}\right) + (1-\alpha)\left(f - \Phi_0^{(2)}\right) \right\| = \alpha \left\| f - \Phi_0^{(1)} \right\| + (1-\alpha)\left\| f - \Phi_0^{(2)} \right\|$$

За припущенням $\exists \lambda$ таке, що $\alpha\left(f - \Phi_0^{(1)}\right) = \lambda(1-\alpha)\left(f - \Phi_0^{(2)}\right)$. Виберемо $\alpha = \frac{1}{2}$. Тоді $\left(f - \Phi_0^{(1)}\right) = \lambda\left(f - \Phi_0^{(2)}\right)$. Оскільки $\left\| f - \Phi_0^{(1)} \right\| = \left\| f - \Phi_0^{(2)} \right\| = \Delta(f)$, то

остання рівність має місце тільки для $\lambda = 1$. Звідси

$f - \Phi_0^{(1)} = f - \Phi_0^{(2)} \Rightarrow \Phi_0^{(1)} = \Phi_0^{(2)}$.

Отже, ми отримали протиріччя з припущенням, що і доводить існування єдиного елемента найкращого наближення. ▹

Теорема 3 Гільбертів простір H – строго нормований.

◃ $\left(H:\ (u,v)\ \forall u,v \in H\ \ \|u\| = \sqrt{(u,u)}\right)$. Нехай

$$\|f+g\| = \|f\| + \|g\|, \qquad x, y \in H \tag{6}$$

$$\|f+g\|^2 = \|f\|^2 + 2\|f\|\|g\| + \|g\|^2.$$

З іншого боку

$$\|f+g\|^2 = (f+g, f+g) = \|f\|^2 + 2(f,g) + \|g\|^2$$

Звідси $\|f\|\|g\| = (f,g)$. Для довільного гільбертового простору $(f,g) \le \|f\|\|g\|$. Таким чином на елементах (6) нерівність Коші – Буняковського перетворюється в рівність. Розглянемо

$$\|f - \lambda g\|^2 = \|f\|^2 - 2\lambda(f,g) + \lambda^2\|g\|^2 = \|f\|^2 - \lambda\|f\|\|g\| + \lambda^2\|g\|^2 = \left(\|f\| - \lambda\|g\|\right)^2,$$

Тоді для $\lambda = \dfrac{\|f\|}{\|g\|}$ маємо $\|f - \lambda g\| = 0$. Звідси $\exists\lambda : f = \lambda g$, тобто Н – строго нормований. ▹

Наслідок $R = H \Rightarrow \exists!\, \Phi_0 \in M_n$.

Приклади строго нормованих просторів:

1) $L_2(a,b)$ з нормою $\|u\| = \sqrt{\int_a^b u^2 dx}$.

2) $L_p(a,b)$ з нормою $\|u\|_p = \left(\int_a^b u^p dx\right)^{\frac{1}{p}}$, $p > 1$.

Простір $C[a,b]$ не є строго нормованим, але в ньому існує єдиний елемент найкращого наближення (про цей факт в наступному пункті).

8.2. Найкраще рівномірне наближення [БЖК, 180-186], [ЛМС, 66-82]

Найкраще рівномірне наближення – це наближення в просторі $R = C[a,b]$, де $\|f\|_{C[a,b]} = \max\limits_{x\in[a,b]}|f(x)|$ – рівномірна метрика.

Теорема 1 (Хаара) Для того, щоб $\forall f \in C[a,b]$ існував єдиний елемент найкращого рівномірного наближення необхідно і достатньо, щоб система $\{\varphi_i\}_{i=0}^{\infty}$ була системою Чебишова.

Система $\{\varphi_i\}_{i=0}^{\infty}$ називається *системою Чебишова*, якщо елемент $\Phi_n(x)=\sum_{i=0}^{n} c_i\varphi_i(x)$ має не більше n нулів, причому $\sum_{i=1}^{n} c_i^2 \neq 0$. Наприклад, системою Чебишова є поліноміальна система $\left\{x^i\right\}_{i=0}^{\infty}$.

Позначимо $Q_n^0(x)$ – *багаточлен найкращого рівномірного наближення* (далі – БНРН.). Його відхилення від f: $\Delta(f)=\left\|Q_n^0(x)-f(x)\right\|_C=\inf_{Q_n(x)}\left\|Q_n(x)-f(x)\right\|$.

Теорема 2 (Чебишова) $Q_n^0(x)$– БНРН неперервної функції $f(x)$ тоді та тільки тоді, якщо на відрізку $[a,b]$ існує хоча б $(n+2)$-а точки $a\le x_0...\le x_m\le b,\ m\ge n+1$ такі, що

$$f(x_i)-Q_n^0(x_i)=\alpha(-1)^i\Delta(f), \tag{1}$$

де $i=\overline{0,m}$, $\alpha=\pm 1$.

Точки $\{x_i\}_{i=0}^{m}$, які задовольняють умовам теореми Чебишова, називаються *точками чебишовського альтернансу*.

Теорема 3 $Q_n^0(x)$ – БНРН для неперервної функції єдиний.

◁ Припустимо, існують два БНРН степеня n: $Q_n^{(1)}(x)\neq Q_n^{(2)}(x)$:

$$\Delta(f)=\left\|f-Q_n^{(1)}\right\|_C=\left\|f-Q_n^{(2)}\right\|_C.$$

Звідси випливає, що

$$\left\|f-\frac{Q_n^{(1)}+Q_n^{(2)}}{2}\right\|\le\left\|\frac{f-Q_n^{(1)}}{2}\right\|+\left\|\frac{f-Q_n^{(2)}}{2}\right\|=\Delta(f),$$

тобто багаточлен $\dfrac{Q_n^{(1)}(x)+Q_n^{(1)}(x)}{2}$ також є БНРН. Нехай $x_0,x_1,...,x_m$ – відповідні йому точки чебишовського альтернансу.

Це означає, що

$$\left|\frac{Q_n^{(1)}(x_i)+Q_n^{(2)}(x_i)}{2}-f(x_i)\right|=\Delta(f),$$

або

$$\left|\left[Q_n^{(1)}(x_i)-f(x_i)\right]+\left[Q_n^{(2)}(x_i)-f(x_i)\right]\right|=2\Delta(f) \tag{2}$$

Так як $\left|Q_n^{(k)}(x_i)-f(x_i)\right|\le\Delta(f)$, $k=1,2$, то (2) можливе лише у тому випадку, коли

$$Q_n^{(1)}(x_i)-f(x_i)=Q_n^{(2)}(x_i)-f(x_i),\ i=\overline{0,n+1}.$$

Звідки випливає, що $Q_n^{(1)}(x)=Q_n^{(2)}(x)$, а це суперечить початковому припущенню. ▷

8.3. Приклади побудови БНРН [БЖК, 186-181], [Волков, 81-91]

Скінченого алгоритму побудови БНРН для довільної функції не існує. Є ітераційний [ЛМС, 73-79]. Але в деяких випадках можна побудувати БНРН за теоремою Чебишова.

1^0. Потрібно наблизити багаточленом нульового степеня.

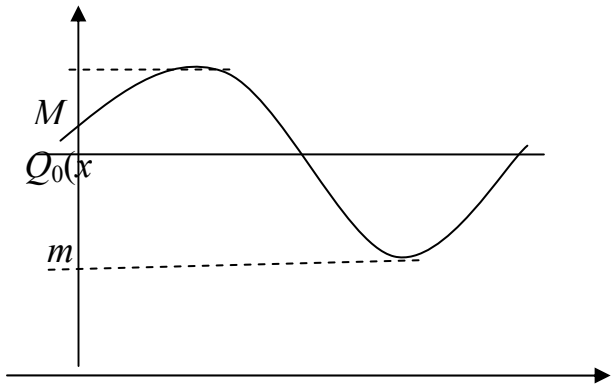


Рис. 11

Нехай $M = \max\limits_{[a,b]} f(x) = f(x_0)$, $m = \min\limits_{[a,b]} f(x) = f(x_1)$. тоді $Q_0(x)$ – БНРН має вигляд (див. рис. 11):

$$Q_0(x) = \frac{M+m}{2},$$

де $f(x_0) = \frac{M+m}{2} = \frac{M-m}{2}$, $f(x_1) = \frac{M+m}{2} = m - \frac{M+m}{2} = -\frac{M+m}{2}$,

$\Delta(f) = \frac{M+m}{2}$, а x_0, x_1 – точки чебишовського альтернансу.

2^0. Опукла функція $f(x) \in C[a,b]$ наближається багаточленом першого степеня

$$Q_1(x) = c_0 + c_1 x.$$

Оскільки $f(x)$ опукла, то різниця $f(x) - (c_0 + c_1 x)$ може мати лише одну внутрішню точку екстремуму. Тому точки a, b є точками чебишовського альтернансу. Нехай ξ третя – точка чебишовського альтернансу. Згідно з теоремою Чебишова, маємо систему:

$$\begin{cases} f(a) - c_0 - c_1 a = \alpha\Delta(f) \\ f(\xi) - c_0 - c_1 \xi = -\alpha\Delta(f) \\ f(b) - c_0 - c_1 b = \alpha\Delta(f) \end{cases}$$

Звідси $f(b) - f(a) = c_1(b-a)$ та $c_1 = \frac{f(b)-f(a)}{b-a}$.

Цю систему треба замкнути, використавши ще одне рівняння з умови: точка ξ є точкою екстремуму різниці $f(x) - (c_0 + c_1 x)$. Тому для диференційованої функції $f(x)$ для визначення ξ маємо рівняння (дотична і січна паралельні):

$$f'(\xi) = c_1 = \frac{f(b) - f(a)}{b - a}.$$

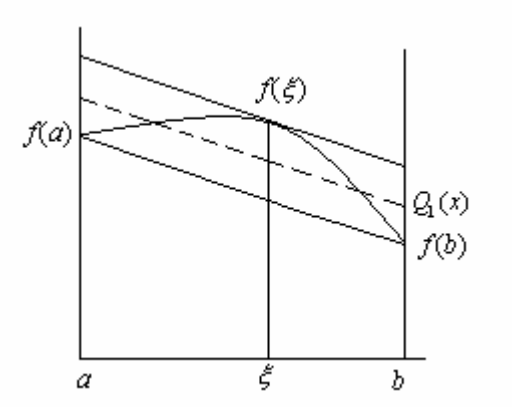


Рис. 12

Геометрично ця процедура виглядає наступним чином (див. рис. 12). Проводимо січну через точки $(a, f(a)), (b, f(b))$. Для неї тангенс кута дорівнює c_1. Проводимо паралельну їй дотичну до кривої $y = f(x)$, а потім пряму, рівновіддалену від січної та дотичної, яка і буде графіком $Q_1(x)$. При цьому $x_0 = a$, $x_1 = \xi$, $x_2 = b$.

3^0. Потрібно наблизити $f(x) = x^{n+1}$, $x \in [-1,1]$ багаточленом степеня n $Q_n^0(x)$. Введемо

$$\overline{P}_{n+1}(x) = x^{n+1} - Q_n(x) = x^{n+1} - a_1 x^n - \ldots$$

$$\Delta(f) = \inf_{Q_n(x)} \left\| x^{n+1} - Q_n^0(x) \right\|_C = \inf_{\overline{P}_{n+1}} \left\| \overline{P}_{n+1} - 0 \right\|_C = \left\| \overline{T}_{n+1}(x) \right\| \Rightarrow$$

$$x^{n+1} - Q_n^0(x) = \overline{T}_{n+1}(x) \Rightarrow Q_n^0(x) = x^{n+1} - \overline{T}_{n+1}(x).$$

Задача 25 Для прикладу 3 вказати точки чебишовського альтернансу $\{x_i\}, i = \overline{0, n+1}$.

4^0. Потрібно наблизити $f(x) = P_{n+1}(x) = a_0 + \ldots + a_{n+1} x^{n+1}$, $a_{n+1} \neq 0$, $x \in [a,b]$ БНРН степеня n. Запишімо його у вигляді:

$$Q_n^0(x) = P_{n+1}(x) - a_{n+1} \overline{T}_{n+1}^{[a,b]},$$

де $\overline{T}_{n+1}^{[a,b]}(x)$ - нормований багаточлен Чебишова на проміжку $x \in [a,b]$.

Дійсно це БНРН: вираз у правій частині є багаточленом степеня n, оскільки коефіцієнт при x^{n+1} дорівнює нулю, а його нулі $x_k = \frac{b+a}{2} + \frac{b-a}{2} t_k$, $t_k = \cos \frac{2k+1}{2(n+1)} \pi, k = \overline{0, n}$ є точками чебишевського альтернасу для $Q_n^0(x)$.

Задача 26 Показати, що для $f(x)$ парної (непарної) функції БНРН це багаточлен по парних (непарних) степенях x.

5^0. Телескопічний метод. Дуже часто БНРН точно знайти не вдається. В таких випадках шукається багаточлен, близький до нього. Бажано щоб цей

багаточлен був невисокого степеня (менше арифметичних операцій на його обчислення) Спочатку будують такий багаточлен $P_n(x)=\sum_{j=0}^{n} a_j x^j$, щоб відхилення від $f(x)$ була достатньо малою. (наприклад меншою за $\frac{\varepsilon}{2}$). Можна це зробити, наприклад, за формулою Тейлора. Потім наближають багаточлен $P_n(x)$ багаточленом найкращого рівномірного наближення $P_{n-1}(x)$ (за алгоритмом п. 4; для простоти $x\in[-1,1]$):

$$P_{n-1}(x)=P_n(x)-a_nT_n(x)2^{1-n}.$$

Оскільки $|T_n(x)|\le 1$ на відрізку $[-1,1]$, то

$$|P_{n-1}(x)-P_n(x)|\le|a_n|2^{1-n}.$$

Далі наближають багаточлен $P_{n-1}(x)$ багаточленом найкращого рівномірного наближення $P_{n-2}(x)$ і т. д. Пониження степеня продовжується до тих пір, поки сумарна похибка від таких послідовних апроксимацій залишається меншою за задане мале число ε.

8.4. Найкраще середньоквадратичне наближення [БЖК, 156-166], [ЛМС, 53-58]

Наблизимо функцію $f(x)\in H$ з гільбертового простору H функціями з скінченно-вимірного підпростору M_n простору H. Тут H гільбертів простір із скалярним добутком (u,v), норма і відстань для якого визначаються формулами:

$$\|u\|=\sqrt{(u,u)},\ \Delta(u,v)=\|u-v\|.$$

Побудуємо

$$u=\sum_{i=0}^{n} c_i\varphi_i\in M_n\subset H, \tag{1}$$

де $\{\varphi_i\}_{i=0}^{\infty}$ - лінійно незалежна система елементів з H.

Елемент найкращого середньоквадратичного наближення (в подальшому ЕНСКН) Φ_0 такий, що

$$\|f-\Phi_0\|=\sqrt{(f-\Phi_0,f-\Phi_0)}=\inf_{\Phi\in M_n}\|f-\Phi\|.$$

Теорема 1 Нехай $f\in H$, $\Phi_0\in M_n$ – елемент найкращого середньо-квадратичного наближення

$$\|f-\Phi_0\|=\inf_{\varphi\in M_n}\|f-\Phi\|.$$

Тоді

$$(f-\Phi_0,\Phi)=0,\ \forall\Phi\in M_n \tag{2}$$

◁ Нехай (2) не виконується, тобто $\exists\,\Phi_1$:

$$(f-\Phi_0,\Phi_1)=\alpha\ne 0,\ \Phi_1\in M_n,\ \|\Phi_1\|=1.$$

Побудуємо $\Phi_2 = \Phi_0 + \alpha\Phi_1$,

$\|f - \Phi_2\|^2 = (f - \Phi_2, f - \Phi_2) = \|f - \Phi_0\|^2 - \alpha^2 < \|f - \Phi_0\|^2$.

Отже, елемент Φ_2 кращий за елемент найкращого середньоквадратичного наближення Φ_0. А це суперечність. ▷

Наслідок

$f = \Phi_0 + v$, де $\Phi_0 \in M_n$, а $v \perp M_n$ (поправка v з ортогонального доповнення до M_n).

Знайти ЕНСКН

$$\Phi_0 = \sum_{i=0}^{n} c_i \varphi_i \tag{3}$$

означає знайти коефіцієнти c_i.

Для виконання (2) достатньо, щоб

$$(f - \Phi_0, \varphi_k) = 0,\ k = \overline{0,n}. \tag{4}$$

Підставимо (3) у формулу (4):

$$\left(f - \sum_{i=0}^{n} c_i \varphi_i, \varphi_k \right) = 0.$$

Таким чином маємо СЛАР для c_i:

$$\sum_{i=0}^{n} c_i (\varphi_i, \varphi_k) = (f, \varphi_k), k = \overline{0,n} \tag{5}$$

З теореми 1 витікає лише достатність умов (5) для знаходження коефіцієнтів c_i. Розглянемо задачу $\|f - \Phi_0\| = \inf\limits_{\varphi \in M_n} \|f - \Phi\|$, як задачу мінімізації функції багатьох змінних:

$$F(a_0, .., a_n) \equiv \|f - \Phi\|^2 = \left\| f - \sum_{i=0}^{n} a_i \varphi_i \right\|^2.$$

Умови мінімуму цієї функції приводять до (5).

Задача 27 Показати, що для коефіцієнтів c_i елемента найкращого середньо квадратичного наближення умови (5) є необхідними та достатніми.

Матриця СЛАР (5) складається з елементів $g_{ik} = (\varphi_i, \varphi_k)$, тобто це матриця Грамма: $G = (g_{ik})_{i,k=0}^{n}$. Оскільки це матриця Грамма лінійно незалежної системи, то $\det G \neq 0$, що ще раз доводить існування та єдиність ЕНСКН. Оскільки $G^T = G$, то для розв'язку цієї системи використовують метод квадратних коренів.

Якщо взяти $x \in [0,1]$ та $\varphi_i = x^i$, $i = \overline{0,n}$, $H = L_2(0,1)$, то

$$g_{ik} = \int_0^1 x^i x^k dx = \frac{1}{i + k + 1},\ i, k = \overline{0,n}.$$

Це матриця Гілберта, яка є погано обумовленою: $condG \cong 10^7$, $n = 6$. Праві частини

$$f_k = (f, \varphi_k) = \int_0^1 f(x_i) x^k dx,$$

як правило, обчислюються наближено, тому похибки обчислення c_i можуть бути великими.

Що робити? Якщо вибирати систему $\{\varphi_i\}_{i=0}^{\infty}$ ортонормованою, тобто

$$(\varphi_i, \varphi_k) = \delta_{ik} = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases},$$

то система (5) має явний розв‘язок:

$$\Phi_0 = \sum_{i=0}^{n} (f, \varphi_i)\varphi_i \,,. \tag{6}$$

Якщо $\{\varphi_i\}$ – повна ортонормована система, то довільну функцію можна представити у вигляді ряду Фур’є:

$$f = \sum_{i=0}^{\infty} (f_n, \varphi_i)\varphi_i \,, \tag{7}$$

і $f - \Phi_0 = \sum_{i=n+1}^{\infty} c_i \varphi_i = v$ - залишок (похибка). Таким чином ЕНСКН є відрізком ряду Фур’є. Далі

$$\|f - \Phi_0\|^2 = (f - \Phi_0, f - \Phi_0) = \|f\|^2 - 2(f, \Phi_0) + \|\Phi_0\|^2 =$$

$$= \|f\|^2 - 2\|\Phi_0\|^2 - \underbrace{2(v, \Phi_0)}_{=0 \ \text{за теоремою 1}} + \|\Phi_0\|^2 =$$

$$= \|f\|^2 - \|\Phi_0\|^2 = \sum_{i=0}^{\infty} c_i^2 - \sum_{i=0}^{n} c_i^2 = \underbrace{\sum_{i=n+1}^{\infty} c_i^2}_{\text{квадрат похибки}} \underset{n\to\infty}{\to} 0.$$

Останнє витікає з відповідної теореми математичного аналізу. Таким чином, якщо $\{\varphi_i\}_{i=0}^{\infty}$ - повна ортонормована система, то

$$\sum_{i=n+1}^{\infty} c_i \xrightarrow[n\to\infty]{} 0 \ \text{ та } \ \Phi_0 \xrightarrow[n\to\infty]{} f.$$

Значить вірна

Теорема 2 В гільбертовому просторі H послідовність ЕНСКН $\left\{\Phi_0^{(n)}\right\}$ по повній ортонормованій системі $\{\varphi_i\}_{i=0}^{\infty}$ збігається до f.

Зауваження 1. Відхилення можна обчислити за формулою:

$$\Delta^2(f) = \|f - \Phi_0\|^2 = \|f\|^2 - 2(f, \Phi_0) + \|\Phi_0\|^2 = \|f\|^2 - \|\Phi_0\|^2 = \|f\|^2 - \sum_{i=0}^{n} c_i^2 \,.$$

Якщо $\{\varphi_i\}_{i=0}^{\infty}$ – ортогональна система, але не нормована, тобто $(\varphi_i, \varphi_k) = \delta_{ik}\|\varphi_i\|^2$, то

$$c_i = \frac{(f,\varphi_i)}{\|\varphi_i\|^2}, \qquad \Phi_0 = \sum_{i=0}^{n} \frac{(f,\varphi_i)\varphi_i}{\|\varphi_i\|^2}, \qquad \|f-\Phi_0\|^2 = \|f\|^2 - \sum_{i=0}^{n} \frac{c_i^2}{\|\varphi_i\|^2}.$$

Для функції $f(x)$, щоб побудувати ЕНСКН покладемо $H = L_{2,\alpha}(a,b)$, в якому скалярний добуток виберемо наступним чином

$$(u,v) = \int_a^b u(x)v(x)d\alpha(x) \quad \text{(інтеграл Стільт‘єса)},$$

де $\alpha(x)$ - зростаюча функція. Можливі випадки:

1. $\alpha(x) \in C^1[a,b]$, тоді $\alpha'(x) = \rho(x) > 0$ та $(u,v) = \int_a^b \rho(x)u(x)v(x)dx$;

2. $\alpha(x)$ – функція стрибків, $\alpha(x) = \alpha(x_k - 0)$, де $x_{k-1} \le x \le x_k$, $k = \overline{1,N}$. Якщо ввести $\rho_k = \alpha(x_k + 0) - \alpha(x_k - 0)$, то $(u,v) = \sum_{i=1}^{n} \rho_i u(x_i)v(x_i)$.

Перший вибір $\alpha(x)$ використовується при апроксимації функцій неперервного аргументу, а другий – для табличних функцій.

8.5. Системи ортогональних функцій [БЖК, 19-102], [ЛМС, 388-382]

Як вибирати ортонормальну або ортогональну систему функцій $\{\varphi_i\}_{i=0}^{\infty}$? Розглянемо деякі з найбільш вживаних таких систем.

1^0. Якщо $H = L_2(-1,1)$; $\rho \equiv 1$ (ваговий множник), то $\varphi_i(x) = L_i(x)$ - система *багаточленів Лежандра*, які мають вигляд

$$L_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n.$$

Використовують також рекурентні формули

$$(n+1)L_{n+1}(x) = (2n+1)xL_n(x) - nL_{n-1}(x),$$

до яких додаємо умови

$$L_0(x) = 1, L_1(x) = x.$$

Це ортогональна система в тому сенсі, що

$$(L_i, L_k) = \int_{-1}^{1} L_i(x)L_k(x)dx = \delta_{ik}\|L_i(x)\|^2, \ \|L_i(x)\|^2 = \frac{2}{2i+1}$$

і тому $c_i = \dfrac{(f,L_i)}{\|L_i\|^2} = \dfrac{2i+1}{2}(f,L_i)$.

Зауваження Якщо потрібно побудувати наближення на довільному проміжку (a,b), то бажано перейти до проміжку $(-1,1)$, тобто по $f(x)$ на $[a,b]$ побудувати $\bar{f}(t)$ з $t \in [-1,1]$ заміною $x = At + B$, $t = \alpha x + \beta$ та для побудови багаточлена НСКН для $\bar{f}(t)$ використати багаточлени Лежандра $L_i(t)$.

Можна робити навпаки - систему багаточленів перевести з $[a,b]$ на $[-1,1]$, але це вимагає більше обчислень і процес побудови ЕНСКН складніше.

2^0. Якщо $H = L_{2,\rho}(-1,1)$, $\rho(x) = \dfrac{1}{\sqrt{1-x^2}}$, скалярний добуток $(u,v) = \int\limits_{-1}^{1} \dfrac{u(x)v(x)}{\sqrt{1-x^2}}dx$ (це невласні інтеграли другого роду), то $\varphi_i(x) = T_i(x)$, де $\{T_i(x)\}$ система ортогональних *багаточленів Чебишова 1-го роду*, які мають вигляд

$$T_n(x) = \cos(n\arccos(x)).$$

Рекурентна формула для цих багаточленів:

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),\ T_0 = 1\ \ T_1 = x,$$

$$\|T_n\|^2 = \begin{cases} \pi, & n = 0 \\ \dfrac{\pi}{2}, & n = 1,2... \end{cases}$$

3^0. H гільбертів простір з ваговим множником $\rho(x) = (1-x)^\alpha (1+x)^\beta$. Система $\varphi_i(x) = P_n^{(\alpha,\beta)}(x)$- *багаточленів Якобі*, $\alpha, \beta > -1$ (α, β – числові параметри) ортогональна в сенсі скалярного добутку $(u,v) = \int\limits_{-1}^{1} (1-x)^\alpha (1+x)^\beta u(x)v(x)dx$. Ця система є узагальненням випадків 1^0 та 2^0. Диференціальна формула для багаточленів:

$$P_n^{(\alpha,\beta)}(x) = \frac{(-1)^n}{2^n n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}\left[(1-x)^{n+\alpha}(1+x)^{n+\beta}\right];$$

Рекурентна формула:

$$2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)P_{n+1}^{(\alpha,\beta)}(x) =$$
$$= (2n+\alpha+\beta+1)[(2n+\alpha+\beta)(2n+\alpha+\beta+2)x + \alpha^2 - \beta^2]P_n^{(\alpha,\beta)}(x) -$$
$$- 2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)P_{n-1}^{(\alpha,\beta)}(x),\ \ n = 1,2,...$$

де

$$P_0^{(\alpha,\beta)} = 1,\quad P_{-1}^{(\alpha,\beta)} = 0,\quad \left\|P_n^{(\alpha,\beta)}\right\|^2 = \frac{2^{\alpha+\beta+1}\Gamma(\alpha+n+1)\Gamma(\beta+n+1)}{n!(\alpha+\beta+2n+1)\Gamma(\alpha+\beta+n+1)},$$

та

$$\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}dt,\quad \Gamma(z+1) = z\Gamma(z),\quad \Gamma(n+1) = n!$$

Коли $\alpha = \beta = 0$: $P_n^{(0,0)}(x) = L_n(x)$, а для $\alpha = \beta = -\dfrac{1}{2}$: $P_n^{(-\frac{1}{2},-\frac{1}{2})}(x) = T_n(x)$.

4^0. $H = L_{2,\rho}[0,\infty)$, $\rho(x) = x^\alpha e^{-x}$, $\alpha > -1$. Цьому ваговому множнику відповідає *система багаточленів Лагерра* $\varphi_i(x) = L_i^\alpha(x)$, які задаються диференціальною формулою:

$$L_n^\alpha(x) = (-1)^n x^{-\alpha}e^x\frac{d^n}{dx^n}\left[x^{\alpha+n}e^{-x}\right],$$

або в рекурентній формі

$$(n+1)L_{n+1}^{\alpha} = (2n+\alpha+1-x)L_n^{\alpha} - (n+\alpha)L_{n-1}^{\alpha},$$

де $L_0^{\alpha} = 1, \quad L_{-1}^{\alpha} = 0$ та з нормою $\left\|L_n^{\alpha}\right\|^2 = n!\Gamma(\alpha+n+1)$.

5^0. $H = L_{2,\alpha}(-\infty,\infty), \quad \rho(x) = e^{-x^2}$. Систему ортогональних функцій вибираємо як систему *багаточленів Ерміта* $\varphi_i(x) = H_i(x)$, які задаються диференціальною формулою:

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}(e^{-x^2})$$

або в рекурентній формі

$$H_{n+1} = 2xH_n - 2nH_{n-1},$$

де $H_0 = 1, \quad H_{-1} = 0$ та $\left\|H_n\right\|^2 = 2^n n!\sqrt{\pi}$.

6^0. $H = L_2(0,2\pi), \quad \rho(x) \equiv 1 \quad f(x) = f(x+2\pi).$ $f(x)$ – 2π- періодичні функції. За систему ортогональних функцій вибираємо тригонометричну систему

$$\varphi_0 = \frac{1}{\sqrt{2\pi}}, \qquad \varphi_{2k-1}(x) = \frac{1}{\sqrt{\pi}}\cos(kx), \qquad \varphi_{2k}(x) = \frac{1}{\sqrt{\pi}}\sin(kx),$$

де

$$\left\|\varphi_m\right\|^2 = 1.$$

Елемент найкращого середньоквадратичного наближення представляє собою *тригонометричний багаточлен*

$$\Phi_0(x) = T_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n}(a_k\cos(kx) + b_k\sin(kx)),$$

формули для обчислення цих коефіцієнтів наведені в наступному пункті.

7^0. Якщо потрібно апроксимувати табличну функцію, то

$$H = l_2, \quad x_i \equiv i, i = \overline{0,N}, \quad (u,v) = \frac{1}{N+1}\sum_{i=0}^{N} u_i v_i$$

і за систему ортогональних функцій вибираємо наступну систему багаточленів $\varphi_k(x) = p_k^{(N)}(x), \quad k = \overline{0,m} \quad (m \le N)$ – *систему багаточленів Чебишова дискретного аргументу*, які задається формулою

$$p_k^{(N)}(x) = \sum_{j=0}^{k} \frac{(-1)^j c_k^j c_{k+j}^j}{N^{(j)}} x^{(j)},$$

$x^{(j)} = x(x-1)...(x-j+1)$- факторіальний багаточлен; C_k^j - число сполук. Рекурентна формула

$$\frac{(m+1)(N-m)}{2(2m+1)}p_{m+1}^{(N)} = \left(\frac{N}{2} - x\right)p_m^{(N)} - \frac{m(N+m+1)}{2(2m+1)}p_{m-1}^{(N)}, \quad p_0^{(N)} = 1, \quad p_{-1}^{(N)} = 0.$$

Наприклад $p_1^{(N)} = 1 - \frac{2x}{N}$, $p_2^{(N)} = 1 - \frac{6x}{N} + \frac{6x^2}{N(N-1)}$.

У випадку, якщо задані вузли $t_i = t_0 + ih, i = \overline{0,N}$, то робимо заміну $x_i = \frac{t_i - t_0}{h} = i.$

8.6. Середньоквадратичне наближення періодичних функцій [ЛМС,60-61], [БЖК,166-182]

Нехай маємо періодичну функцію $f(x)$ неперервного аргументу, з періодом $T = 2\pi$, тобто $f(x + 2\pi) = f(x)$. В просторі $H_2 = L_2[0,2\pi]$ визначений скалярний добуток

$$(u,v) = \int_0^{2\pi} u(x)v(x)dx.$$

В якості системи лінійно-незалежних функцій $\{\varphi_i\}$ виберемо тригонометричну систему функцій:

$$\varphi_0(x) = 1;\ \varphi_{2k-1}(x) = \cos kx;\ \varphi_{2k}(x) = \sin kx,\quad k = \pm 1, \pm 2, \ldots;$$

яка є повною нормованою системою в $L_2[0,2\pi]$.

Будемо шукати $\Phi(x)$ у вигляді тригонометричного багаточлена

$$\Phi(x) \equiv T_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx). \qquad (1)$$

За теорією найкращого середньоквадратичного наближення коефіцієнти обчислюємо за формулами:

$$\begin{cases} a_0 = (f, \varphi_0) = \dfrac{1}{2\pi} \displaystyle\int_0^{2\pi} f(x)dx, \\ a_k = (f, \varphi_k^c) = \dfrac{1}{\pi} \displaystyle\int_0^{2\pi} f(x)\cos kxdx, \\ b_k = (f, \varphi_k^s) = \dfrac{1}{\pi} \displaystyle\int_0^{2\pi} f(x)\sin kxdx. \end{cases} \qquad (2)$$

Відхилення:

$$\Delta^2(f) = \|f\|^2 - \left(2\pi {a_0}^2 + \sum_{k=1}^{n} \pi({a_k}^2 + {b_k}^2) \right).$$

Тепер нехай функція $f(x)$ задана таблично:

$$f_i = f(x_i), i = \overline{1,N}.$$

Тригонометрична система $\varphi_0(x), \varphi_{2k-1}(x), \varphi_{2k}(x)$ ортогональна в $H = L_2(\omega)$ для $\omega = \left\{ x_i = i\frac{\pi}{N}, i = \overline{1,N} \right\}$ в сенсі скалярного добутку

$$(u,v)=\frac{1}{N}\sum_{i=1}^{N}u_i v_i\ ,\ u_i=u(x_i)\,.$$

Тоді

$$\begin{cases} a_0=\dfrac{1}{N}\sum\limits_{i=1}^{N}f_i, \\ a_k=\dfrac{2}{N}\sum\limits_{i=1}^{N}f_i\cos kx_i, \\ b_k=\dfrac{2}{N}\sum\limits_{i=1}^{N}f_i\sin kx_i. \end{cases} \tag{3}$$

Це формули Бесселя . В формулі (1) $\Phi(x)\equiv T_n(x)$ (тобто багаточлен той же), але коефіцієнти визначаємо за формулою (3).

Зауваження Як правило кількість даних значень $N>>2n+1$. Але якщо $N=2n+1,$ то $n=\dfrac{N-1}{2}$ і N -непарне. При цьому $T_{\frac{N-1}{2}}(x)-$ БНСКН і звідси

$$\Delta^2(f)=\left\|f(x)-T_{\frac{N-1}{2}}(x)\right\|^2=\frac{1}{N}\sum_{i=1}^{N}\left[f(x_i)-T_{\frac{N-1}{2}}(x_i)\right]^2\to\inf_{a_k,b_k}$$

Оскільки найменше значення відхилення $\Delta^2(f)=0$, то тригонометричний багаточлен найкращого середньоквадратичного наближення співпадає з інтерполяційним тригонометричним багаточленом і

$$T_{\frac{N-1}{2}}(x_i)=f(x_i)\,.$$

Для визначення коефіцієнтів a_i,b_i за формулою Бесселя (3) необхідна кількість операцій $Q=O(N^2)$. Існують алгоритми, які дозволяють обчислити за $Q=O(N\log N)$ операцій. Це так званий алгоритм швидкого перетворення Фур'є. Якщо в (3) існує група доданків , які рівні між собою, тобто число N можна представити як $N=p_1p_2,$ то можна так вибрати сітку, що $Q=O(N\max(p_1,p_2))$. Якщо ж $N=n^m$, то $Q=O(Nm)=O(N\log_2 N)$.

8.7. Метод найменших квадратів (МНК) [ЛМС 61-65], [В,88-93]

Нехай в результаті вимірювань функції $f(x)$ маємо таблицю значень:

$$y_i\approx f(x_i),\ i=\overline{1,N},\ x_i\in[a,b]\ . \tag{1}$$

За даними цієї таблиці треба побудувати аналітичну формулу $\Phi(x;a_0,a_2,\ldots,a_n)$ таку, що

$$\Phi(x_i;a_0,a_2,\ldots,a_n)\approx y_i\ ,\ i=\overline{1,N}\,. \tag{2}$$

Виконувати це інтерполюванням тобто задавати

$$\Phi(x_i;a_0,a_2,\ldots,a_n)=y_i\ ,\ i=\overline{1,N} \tag{3}$$

нераціонально, бо $N >> n$ і система перевизначена; її розв'язки як правило не існують. Вигляд функції $\Phi(x; a_0, a_2, \ldots, a_n)$ і число параметрів a_i у деяких випадках відомі. В інших випадках вони визначаються за графіком, побудованим за відомими значеннями $f(x_i)$ так, щоб залежність (2) була досить простою і добре відображала результати спостережень. Але такі міркування не дають змогу побудувати єдиний елемент та й ще найкращого наближення.

Тому визначають параметри $a_0, \ldots, a_n$ так, щоб у деякому розумінні всі рівняння системи (2) одночасно задовольнялись з найменшою похибкою, наприклад, щоб виконувалося

$$I(a_0, \ldots, a_n) = \sum_{i=1}^{N} [y_i - \Phi(x_i; a_0, \ldots, a_n)]^2 \to \min. \qquad (4)$$

Такий метод розв'язання системи (2) і називають методом найменших квадратів, оскільки мінімізується сума квадратів відхилення $\Phi(x; a_0, a_2, \ldots, a_n)$ від значень $f(x_i)$.

Для реалізації мінімуму необхідно та достатньо виконання умов:

$$\frac{\partial I}{\partial a_i} = 0,\ i = \overline{0, n}. \qquad (5)$$

Якщо $\Phi(x_i; a_0, \ldots, a_n)$ лінійно залежить від параметрів $a_0, \ldots, a_n$, тобто

$$\Phi(x_i; a_0, \ldots, a_n) = \sum_{k=0}^{n} a_k \varphi_k(x), \qquad (6)$$

то з (3) маємо СЛАР

$$\sum_{k=0}^{n} a_k \varphi_k(x_i) = y_i, \quad i = \overline{1, N}, \qquad (7)$$

яку називають *системою умовних рівнянь*. Позначивши

$$C = (\varphi_k(x_i))_{k=\overline{0,n}}^{i=\overline{1,N}},\ \vec{a} = (a_0, \ldots, a_n)^{\mathrm{T}},\ \vec{y} = (y_1, \ldots, y_N)^{\mathrm{T}},$$

маємо матричний запис СЛАР (7)

$$C\vec{a} = \vec{y} \qquad (8)$$

Помноживши систему умовних рівнянь (8) зліва на транспоновану до C матрицю C^T отримаємо систему *нормальних рівнянь*

$$C^T C\vec{a} = C^T \vec{y} \qquad (9)$$

$$G = A = C^{\mathrm{T}} C,\ \dim G = n + 1,\ G = [g_{ik}]_{i,k=0}^{n},$$

$$g_{ik} = \sum_{j=1}^{N} c_{ij}^{\mathrm{T}} c_{jk} = \sum_{j=1}^{N} c_{ji} c_{jk} = \sum_{j=1}^{N} \varphi_k(x_i)\varphi_j(x_i), \quad C^T \vec{y} = \left(\sum_{i=1}^{N} c_{ik} y_i \right)_{k=0}^{n},$$

з якої власно і обчислюють невідомі коефіцієнти.

Покажемо, що МНК є методом знаходження ЕНСКН, якщо визначити скалярний добуток

$$(u, v) = \sum_{i=1}^{N} u(x_i) v(x_i).$$

Поставимо задачу знаходження ЕНСКН:

$$\Delta(f,\Phi)=\|f-\Phi\|^2=(f-\Phi,f-\Phi)=\sum_{i=1}^{N}(y_i-\Phi(x_i,\vec{a}))^2\to\inf.$$

За теорією середньоквадратичного наближення для цього необхідно, щоб коефіцієнти $a_0,\ldots,a_n$ знаходилися з системи:

$$\sum_{j=0}^{n}a_k(\varphi_k,\varphi_j)=(\varphi_k,f)\ ,\ k=\overline{0,n},$$

а це співпадає з (9).

Якщо відома інформація про обчислювальну похибку для значень $f(x_i)$

$$|f(x_i)-y_i|<\varepsilon_i,$$

то вибирають такий скалярний добуток $(u,v)=\sum_{i=1}^{N}\rho_i u(x_i)v(x_i)$, де $\rho_i=\frac{1}{\varepsilon_i^2}$.

Нехай тепер $\Phi(x,a_0,...,a_n)$ - нелінійна функція параметрів $\vec{a}=(a_0,\ldots,a_n)$, наприклад:

$$\Phi=a_0e^{a_1x}+a_2e^{a_3x}+...,$$

або

$$\Phi=a_0\cos a_1x+a_2\sin a_3x+...\ .$$

Складемо функціонал:

$$S(a_0,...,a_n)=\sum_{i=1}^{N}\rho_i[y_i-\Phi(x,\vec{a})]^2\to\inf_{\vec{a}}\qquad(10)$$

Оскільки тепер $\Phi(x,a_0,...,a_n)$ нелінійна, то застосуємо метод лінеаризації.

Нехай відомі наближені значення $\vec{a}^0=(a_0^0,...,a_n^0)$. Розкладемо $\Phi(x,\vec{a})$ в околі $\vec{a}^0$. Тоді отримаємо лінійне наближення до $\Phi(x,\vec{a})$:

$$\Phi(x,\vec{a})\approx\Phi(x,\vec{a}^0)+\sum_{k=0}^{n}\frac{\partial\Phi}{\partial a_k}(x,\vec{a}^0)(a_k-a_k^0).$$

Якщо ввести позначення

$$\vec{z}=\vec{a}-\vec{a}^0\quad y_i^*=y_i-\Phi(x,\vec{a}^0)\ ,\ c_{ik}=\Phi'_{a_k}(x_i,\vec{a}^0),$$

то отримаємо систему умовних рівнянь відносно поправок до $\vec{a}^0$:

$$C\vec{z}=\vec{y}^*\qquad(11)$$

Замінимо її на систему нормальних рівнянь

$$C^TC\vec{z}=C^T\vec{y}^*\qquad(12)$$

Знайшовши $\vec{z}$, обчислюємо наступне наближення: $\vec{a}^1=\vec{a}^0+\vec{z}$. Цей процес можна продовжувати: на кожній ітерації знаходимо $\vec{z}^m, m=0,1,...$ і уточнюємо наближення до $\vec{a}$: $\vec{a}^m=\vec{a}^{m-1}+\vec{z}^{m-1}$.

Умова припинення ітерацій

$$\left\|\vec{z}^{\,m}\right\| = \left(\sum_{k=0}^{n}\left(z_k^m\right)^2\right)^{\frac{1}{2}} < \varepsilon .$$

Важливим є вибір початкового наближення $\vec{a}^{\,0}$. З системи умовних рівнянь (нелінійної) виберемо деякі $n+1$. Розв‘язок цієї системи і дасть початкове наближення.

Для деяких простих нелінійних залежностей від невеликої кількості параметрів задачу можна ліанеризувати аналітично. Наприклад, розглянемо наближення даних алометричним законом

$$y_i \approx f(x_i)\ ,\ \Phi(x, A, \alpha) = Ax^{\alpha} .$$

Система умовних рівнянь має вигляд:

$$\Phi(x_i) = Ax_i^{\alpha} = y_i\ \ , i = \overline{1, N} .$$

Прологарифмуємо її:

$$\psi(x_i) = \ln\Phi(x_i) = \ln A + \alpha \ln x_i = \ln y_i , i = \overline{1, N}$$

Введемо $a = \ln A$. Тепер функція $\psi(x, a, \alpha)$ лінійна. Система умовних рівнянь відносно параметрів a та α має вигляд.

$$C\vec{z} = \vec{b}\ ,\ \vec{z} = (a, \alpha),\ \vec{b} = (\ln y_i)_{i=1}^{\mathrm{N}},$$

$$C = \begin{pmatrix} 1 & \ln x_1 \\ \ldots\ldots\ldots\ldots & \\ 1 & \ln x_N \end{pmatrix}.$$

Запишемо систему нормальних рівнянь для методу найменших квадратів

$$C^T C\vec{z} = C^T\vec{b}\ , \tag{13}$$

$$G = C^T C = \begin{pmatrix} N & \sum_{i=1}^{N}\ln x_i \\ \sum_{i=1}^{N}\ln x_i & \sum_{i=1}^{N}(\ln x_i)^2 \end{pmatrix},\ C^T\bar{b} = \begin{pmatrix} \sum_{i=1}^{N}\ln y_i \\ \sum_{i=1}^{N}\ln x_i \ln y_i \end{pmatrix}.$$

Розв’язавши систему (13), знаходимо і α , та $A = \exp(a)$.

8.8. Згладжуючі сплайни [Марчук Г. И. Методы вычислительной математики, с. 184–181]

Якщо значення в точках x_i неточно $\tilde{f}_i = f_i + \varepsilon_i$, то застосовують згладжування. Для цього треба побудувати нову таблицю із згладженими значеннями $\bar{f}_i$.

Наведемо деякі прості формули згладжування:

$m = 1$: $\quad \bar{f}_i = \frac{1}{3}\left[\tilde{f}_{i-1} + \tilde{f}_i + \tilde{f}_{i+1}\right]\ ,\ N = 3 .$

$$\bar{f}_i = \frac{1}{5}\left[\tilde{f}_{i-2} + \ldots + \tilde{f}_{i+2}\right]\ , N = 5 .$$

$$\bar{f}_i = \frac{1}{N}\left[\tilde{f}_{i-\frac{N}{2}} + \ldots + \tilde{f}_{i+\frac{N}{2}}\right], N \text{ - парне.}$$

$m=3$: $$\bar{f}_i = \frac{1}{3\cdot 5}\left[-3\tilde{f}_{i-2} + 12\tilde{f}_{i-1} + 17\tilde{f}_i + 12\tilde{f}_{i+1} - 3\tilde{f}_{i+2}\right], N = 5.$$

Їх отримуємо в такий спосіб: до $\tilde{f}_i$ застосовуємо апроксимацію, будуємо багаточлен НСКН

$$Q_m(x) = \sum_{k=0}^{m} c_k p_k^N(x),$$

де p_k^N –система багаточленів Чебишова дискретного аргументу. Беремо значення

$$\bar{f}_i = Q_m(x_i),$$

які приводять до наведених вище формул.

Але ці формули не дають гарантію, що в результаті ми отримаємо функцію, яка задовольняє умові:

$$\left|\bar{f}_i - f_i\right| < \varepsilon_i$$

Згладжуючі сплайни дають можливість побудувати наближення з заданою точністю. Нагадаємо деякі відомості про сплайни. Явний вигляд кубічного сплайна:

$$\begin{aligned} s(x) &= m_{i-1}\frac{(x_i - x)^3}{6h_i} + m_i\frac{(x - x_{i-1})^3}{6h_i} + \left(f_{i-1} - \frac{m_{i-1}h_i^{\,2}}{6}\right)\frac{x_i - x}{h_i} \\ &+ \left(f_i - \frac{m_i h_i^{\,2}}{6}\right)\frac{x - x_{i-1}}{h_i}, \quad x \in [x_{i-1}, x_i], \quad h_i = x_i - x_{i-1}. \end{aligned} \tag{1}$$

Тут $s(x_i) = f_i, i = \overline{0,n}$, а $m_i = s''(x_i)$ задовольняють систему:

$$\begin{cases} \frac{h_i}{6}m_{i-1} + \frac{h_i + h_{i+1}}{3}m_i + \frac{h_{i+1}}{6}m_{i+1} = \frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i}, & i = \overline{1, n-1}, \\ m_0 = m_n = 0. \end{cases} \tag{2}$$

В матричній формі ця система має вигляд:

$$A\vec{m} = H\vec{f}. \tag{3}$$

Тут

$$\vec{m} = (m_1, \ldots, m_{n-1})^{\mathrm{T}}, \ \vec{f} = (f_0, \ldots, f_n)^{\mathrm{T}},$$

$$A = \underbrace{\begin{pmatrix} \frac{h_1 + h_2}{3} & \frac{h_2}{6} & 0 \\ \frac{h_2}{6} & \frac{h_2 + h_3}{3} & \ldots \\ 0 & \ldots & \ldots \end{pmatrix}}_{n-1} n-1, \quad H = \underbrace{\begin{pmatrix} \frac{1}{h_1} & -\left(\frac{1}{h_1} + \frac{1}{h_2}\right) & \frac{1}{h_2} & 0 \ldots. \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & \ldots \end{pmatrix}}_{n+1} n-1.$$

Кубічний інтерполяційний сплайн мінімізує функціонал:

$$\Phi(u)=\int_a^b (u'')^2 dx:$$

$$\Phi(s)=\inf_{u\in U}\Phi(u)\ U=\left\{u(x): u(x_i)=f_i, i=\overline{0,n},\ u(x)\in W_2^2(a,b)\right\} \qquad (5)$$

Введемо функціонал

$$\Phi_1(u)=\Phi(u)+\sum_{i=0}^{n}\rho_i\left[\tilde{f}_i-u(x_i)\right]^2 .$$

Згладжуючим сплайном назвемо функцію g, яка є розв‘язком задачі:

$$\Phi_1(g)=\inf_{u\in W_2^2(a.b)}\Phi_1(u), \qquad (5)$$

Перший доданок в $\Phi_1(u)$ дає мінімум „згину”, другий – середньоквадратичне наближення до значень $\tilde{f}_i$. Покажемо, що g є сплайном.

Нехай існує функція $g(x)$. Побудуємо кубічний сплайн такий, що $s(x_i)=g(x_i)$. З того, що $g(x)$ є розв’язком задачі (5), маємо

$$\Phi_1(s)\ge\Phi_1(g)\Rightarrow\int_a^b (s'')^2\,dx+\sum\rho_i\left[\tilde{f}(x_i)-s(x_i)\right]^2\ge\int_a^b (g'')^2\,dx+\sum_i\rho_i\left[\tilde{f}_i-g(x_i)\right]^2$$

Звідси

$$\Phi(s)\ge\Phi(g).$$

Так як кубічний інтерполяційний сплайн $s(x)$ мінімізує функціонал (4), то $\Phi(s)\le\Phi(g)$. Тому $\Phi(s)=\Phi(g)$. Звідки $s=g$.

Позначимо

$$\mu_i=g(x_i),\ i=\overline{0,n} \qquad (6)$$

Якщо б ми знали μ_i, то для побудови g достатньо було б розв‘язати систему

$$A\vec{m}=H\vec{\mu} \qquad (7)$$

Підставимо (1) та (6) в $\Phi_1(g)$:

$$\Phi_1(g)=\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left(m_{i-1}\frac{x_i-x}{h_i}+m_i\frac{x-x_{i-1}}{h}\right)^2 dx+\sum_{i=0}^{n}\rho_i\left(\tilde{f}_i-\mu_i\right)^2=\inf\Phi_1(u). \qquad (8)$$

Після перетворень маємо

$$\Phi_1(g)=\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left(m_{i-1}\frac{x_i-x}{h_i}+m_i\frac{x-x_{i-1}}{h_i}\right)^2 dx+\sum_{i=0}^{n}\rho_i\left(\tilde{f}_i-\mu_i\right)^2=$$

$$=\sum_{i=1}^{n}m_i\left(m_{i-1}\frac{h_i}{6}+\frac{h_i+h_{i+1}}{3}m_i+\frac{h_{i+1}}{6}m_{i+1}\right)+\sum_{i=0}^{n}\rho_i\left(\tilde{f}_i-\mu_i\right)^2=$$

$$=\left(A\vec{m},\vec{m}\right)+\sum_{i=0}^{n}\rho_i\left(\tilde{f}_i-\mu_i\right)^2$$

Задача 28 Показати, що для кубічного згладжуючого сплайну g має місце формула:

$$\Phi_1(g)=\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}\left(m_{i-1}\frac{x_i-x}{h_i}+m_i\frac{x-x_{i-1}}{h_i}\right)^2dx+\sum_{i=0}^{n}\rho_i\left(\widetilde{f}_i-\mu_i\right)^2=$$

$$=\sum_{i=1}^{n}m_i\left(m_{i-1}\frac{h_i}{6}+\frac{h_i+h_{i+1}}{3}m_i+\frac{h_{i+1}}{6}m_{i+1}\right)+\sum_{i=0}^{n}\rho_i\left(\widetilde{f}_i-\mu_i\right)^2=$$

$$=(A\vec{m},\vec{m})+\sum_{i=0}^{n}\rho_i\left(\widetilde{f}_i-\mu_i\right)^2.$$

Так як $\Phi_1(g)$ представляє собою квадратичну функція відносно $\vec{m}=(m_0,...,m_N)$, то необхідною і достатньою умовою мінімуму є

$$\frac{\partial\Phi_1}{\partial\mu_j}=0,\ j=\overline{0,n}.$$

Знаходимо:

$$\frac{\partial\Phi_1}{\partial\mu_j}=\frac{\partial}{\partial\mu_j}(A\vec{m},\vec{m})+2\rho_j\left(\mu_j-\widetilde{f}_j\right)=$$

$$=2\left[\frac{\partial}{\partial\mu_j}(A\vec{m}),\vec{m}\right]+2\rho_j\left(\mu_j-\widetilde{f}_j\right)=2\left(\frac{\partial}{\partial\mu_j}(H\vec{\mu}),\vec{m}\right)+$$

$$`+2\rho_j\left(\mu_j-\widetilde{f}_j\right)=2\left(\frac{\partial\vec{m}}{\partial\mu_j},H^{\mathrm{T}}\vec{m}\right)+2\rho_j\left(\mu_j-\widetilde{f}_j\right)=$$

$$=2\left(H^{\mathrm{T}}\vec{m}\right)_j+2\rho_j\left(\mu_j-\widetilde{f}_j\right)=0$$

Отже, з умови мінімізації функціоналу

$$\Phi(u)=\int_a^b(u'')^2dx+\sum_{i=0}^{n}\rho_i(f_i-u_i)^2$$

ми отримали таку систему рівнянь :

$$2\left(H^T\vec{m}\right)_i+2\rho_i(\mu_i-f_i)=0 \qquad (9)$$

де, як і раніше μ_i – це невідомі значення згладжуючого сплайну :

$$\mu_i=s(x_i),\ m_i=s''(x_i).$$

Можна записати (9) матричному вигляді, якщо ввести матрицю $R=diag\,\rho_i$:

$$H^T\vec{m}+R\vec{\mu}=R\vec{f}\,. \qquad (10)$$

Тут $\vec{f}$ – вектор заданих значень функції.

Таким чином маємо для $\vec{m}$ та $\vec{\mu}$ дві системи (7) і (10). Виключаючи $\vec{\mu}$ отримаємо таку систему лінійних рівнянь

$$\left(A+HR^{-1}H^T\right)\vec{m}=H\vec{f} \qquad (11)$$

Розв’язавши її, можемо обчислити

$$\vec{\mu}=\vec{f}-R^{-1}H^T\vec{m} \qquad (12)$$

і підставити знайдені значення μ_i та m_i в формулу для сплайну

$$g(x)=m_{i-1}\frac{(x_i-x)^3}{6h_i}+m_i\frac{(x-x_{i-1})^3}{6h_i}+\left(\mu_{i-1}-\frac{m_{i-1}h_i^2}{6}\right)\frac{x_i-x}{h_i}+\left(\mu_i-\frac{m_ih_i^2}{6}\right)\frac{x-x_{i-1}}{h_i}. \quad (13)$$

Тепер звернемо увагу на матрицю системи (10)

$$A'=A+HR^{-1}H^T.$$

Оскільки матриці H, H^T – трьохдіагональні, то матриця $HR^{-1}H^T$ буде п'ятидіагональною, а тому п'ятидіагональною буде й A'. Розв'язують зазвичай системи з такими матрицями наступним чином :

1. або методом квадратних коренів; для матриць із такою структурою цей метод має складність $Q=O(nm)=O(2n)=O(n)$, так як в нашому випадку півширина діагональної смуги $m=2$.

2. або методом п'ятидіагональної прогонки [Самарский А.А., Николаев С.Н., Методы решения сеточных уравнений], що також має складність $O(n)$.

Зауваженн ρ_i вибирають так $\rho_i=\frac{1}{\varepsilon_i^2}$.

9. Чисельне інтегрування

9.1. Постановка задачі чисельного інтегрування [ЛМС, 13-14], [СГ, 161-162]

Нехай потрібно знайти

$$I(f)=\int_a^b\rho(x)f(x)dx \quad (1)$$

де f – задана функція, $\rho(x)>0$ – деякий ваговий множник. Ця задача часто вимагає чисельного вирішення, оскільки

- значна кількість інтегралів типу (1) не можуть бути обчислені аналітично;
- інформація про функцію f може бути задана у вигляді таблиці.

Нагадаємо, що за означенням

$$I(f)=\lim_{\Delta\to0}\sum_{i=1}^n\rho(\xi_i)f(\xi_i)\Delta x_i,$$

де $\Delta x_i=x_i-x_{i-1}$, а $\{x_i\}_{i=0}^n$ – розбиття проміжку $[a,b]$, $x_i\in[a,b]$, $\xi_i\in[x_{i-1},x_i]$. Тому візьмемо як наближення таку суму, яка називається *квадратурною формулою*:

$$I_n(f)=\sum_{k=0}^n c_kf(x_k), \quad (2)$$

де x_k – вузли квадратурної формули (2), а c_k – її вагові множники. Задача полягає в тим, щоб вибрати $\{x_k,c_k\}_{k=0}^n$, так щоб похибка була найменша:

$$R_n(f)=I(f)-I_n(f)\to\min$$

Квадратурну формулу (2) називають *квадратурною формулою замкненого типу*, якщо $x_0=a$ та $x_n=b$, і *відкритого типу* , якщо $x_0>a$ та $x_n<b$.

Кажуть, що квадратурна формула (2) має *m-ий степінь алгебраїчної точності*, якщо

$$R_n(f)=0,\ \forall f\in\pi_m, \tag{3}$$

π_m – множина поліномів m-го степеня, і $\exists P_{m+1}(x)\in\pi_{m+1}$, такий що $R_n(P_{m+1})\neq 0$.

Цю умову можна замінити умовою

$$R_n\left(x^\alpha\right)=0,\ \alpha=\overline{0,m},\ R_n\left(x^{m+1}\right)\neq 0 \tag{4}$$

(вона більш зручна для перевірки).

Розглянемо деякі підходи до побудови квадратурних формул.

1) *Інтерполяційний*. Він приводить до квадратурних формул інтерполяційного типу. В інтегралі (1) покладають $f(x)\approx L_n(x)$ по деяких вузлах $\{x_k\}_{k=0}^n$ (вузли фіксовані). Тоді:

$$I_n(f)=I(L_n(x))=\int_a^b\rho(x)\sum_{k=0}^n\frac{f(x_k)\omega_n(x)}{(x-x_k)\omega_n'(x_k)}dx=\sum_{k=0}^n f(x_k)\int_a^b\rho(x)\frac{\omega_n(x)}{(x-x_k)\omega_n'(x_k)}dx$$

Отже вузлами цієї квадратурної формули є вузли інтерполяційного багаточлена, а вагові множники

$$c_k=\int_a^b\rho(x)\frac{\omega_n(x)}{(x-x_k)\omega_n'(x_k)}dx.$$

2) *Найвищого алгебраїчного степеня точності.* Вибираємо одночасно x_k і c_k з умови $R_n\left(x^\alpha\right)=0,\ \alpha=\overline{0,m}$, щоб m було максимальним. Отримуємо систему нелінійних алгебраїчних рівнянь, розв'язавши яку отримуємо квадратурні формули найвищого алгебраїчного степеня точності.

3) *Складені квадратурні формули.* Проміжок $[a,b]$ розбиваємо на підпроміжки (наприклад однокової довжини), а потім на кожному проміжку використовуємо, з невеликим степенем, формули з пункту 1 або 2. Наприклад, для формул інтерполяційного типу:

$$I(f)=\sum_{i=1}^N\int_{x_{i-1}}^{x_i}\rho(x)f(x)dx_i\approx\sum_{i=1}^N\sum_{k=0}^n c_k^i f\left(x_k^i\right)=I_h(f).$$

Кажуть, що квадратурна формула складеного типу I_h має *порядок (степінь) точності p по кроку h*, якщо $R_h(f)=I(f)-I_h(f)=O\left(h^p\right)$.

4). *Квадратурні формули оптимальні на класі функцій.* Вибираємо $\{x_k,c_k\}$ так, щоб досягався $\inf\limits_{\{x_k,c_k\}}\sup\limits_{f\in F}R_n(f)$. Це ми можемо робити, коли знаємо з яким класом функцій маємо справу.

Зауваження 1 (про квадратурні формули інтерполяційного типу)

При підвищенні степеня інтерполяції погіршується якість наближення функції внаслідок розбіжності процесу інтерполяції: $\|f-L_n\|_c\underset{n\to\infty}{\not\to}0$. Але

$R_n(f) \xrightarrow[n\to\infty]{} 0$, наприклад, для $f \in C[a,b]$. І все ж таки розбіжність процесу інтерполювання дає взнаки: при $n \to \infty$ $\max\limits_k |c_k| \to \infty$ і це приводить до поганих наслідків чисельного інтегрування. Дійсно, розглянемо випадок, коли функція задана неточними значеннями:

$$\tilde{f}(x_k) = f(x_k) + \delta_k, \ |\delta_k| < \delta .$$

Тоді

$$\delta I_n(f) = I_n(\tilde{f}) - I_n(f) = \sum_{k=0}^{n} c_k \delta_k .$$

Якщо всі $c_k > 0$, то

$$|\delta I| = \sum_{k=0}^{n} c_k |\delta_k| \le \delta \sum_{k=0}^{n} c_k = \delta(b-a).$$

При $\rho \equiv 1$, якщо підставити $f \equiv 1$, то отримаємо $b - a = \int\limits_a^b dx = \sum\limits_{k=0}^{n} c_k$. При $\rho \neq 1$ $\sum\limits_{k=0}^{n} c_k = \int\limits_a^b \rho(x)dx$, бо хоча б нульовий степінь точності будь-яка квадратурна формула повинна мати.

Нагадаємо, що при $n \to \infty$ $\max\limits_k |c_k| \to \infty$, а оскільки $\sum c_k > 0$, то $\exists c_k > 0$ і $\exists c_k < 0$, тому з ростом n зростає $|c_k|$, а відповідно і вплив похибки на результат. Тому не можна використовувати великі степені і використовують складені квадратурні формули.

Зауваження 2 Ясно, що квадратурні формули інтерполяційного типу мають алгебраїчний степінь точності принаймні $m = n$, бо ми заміняємо $f \to L_n$, а якщо $f \in \pi_n$, то $f \equiv L_n$. Але виявляється, що для парних n та симетричному розташуванні вузлів інтегрування, $m = n+1$, тобто алгебраїчний степінь точності на одиницю вищий степеня інтерполяції.

9.2. Квадратурні формули прямокутників [СГ, 162-163]

Припустимо, що $\rho \equiv 1$. Тоді можна побудувати такі квадратурна формули інтерполяційного типу при $n = 0$:

а) лівих прямокутників: $x_0 = a$: $I_0^{лів} = (b-a)f(a)$;

б) правих прямокутників: $I_0^{прав} = (b-a)f(b)$ $x_0 = b$;

в) середніх прямокутників:

$$I_0 = (b-a)f(x_0),\ x_0 = \frac{a+b}{2}. \qquad (1)$$

Знайдемо тепер алгебраїчну степінь точності цих квадратурних формул. Для лівих прямокутників:

$I_0^{лів}(1) = b - a = I(1)$,

$$I_0^{лів}(x) = a(b-a) \neq I(x) = \int_a^b x dx = \frac{x^2}{2}\Big|_a^b = \frac{b^2 - a^2}{2},$$

отже степінь точності $m = 0$. Така ж вона буде і для $I_0^{прав}$. А для середніх прямокутників

$$I_0(x) = (b-a)\frac{a+b}{2} = I(x),\ I_0\left(x^2\right) \neq I\left(x^2\right),$$

тому $m = 1$. Отож нею і будемо користуватися.

Оцінимо для неї похибку. Взагалі для формули інтерполяційного типу:

$$R_n(f) = I(f) - I_n(f) = I(f) - I(L_n) = I(f - L_n) = I(r_n) = \int_a^b r_n(x) dx,$$

де $r_n(x)$– залишковий член інтерполяції. Далі

$$|R_n(f)| \leq (b-a)\max_x |r_n(x)| \leq (b-a)\frac{M_{n+1}}{n+1}\max_x |\omega_n(x)|.$$

Для I_0:

$$|R_0(f)| = \left|\int_a^b r_0(x) dx\right| \leq \int_a^b |r_0(x)| dx \leq \int_a^b \frac{M_1}{1!}|x - x_0| dx = M_1 \int_a^b |x - x_0| dx \leq M_1 \frac{b^2 - a^2}{4},$$

Але це погана оцінка, вона не використовує той факт, що квадратурна формула має степінь точності на одиницю вищу. Отримаємо кращу оцінку. Маємо при $f \in C^2[a,b]$:

$$f(x) = f(x_0) + (x - x_0) f'(x_0) + \frac{(x - x_0)^2}{2} f''(\xi),\ x_0 \equiv \frac{a+b}{2}$$

де $\xi \in [a,b]$. Тоді

$$R_0(f) = \int_a^b f(x) dx - \int_a^b L_0(x) dx = \int_a^b [f(x) - f(x_0)] dx =$$

$$= \int_a^b \left[(x - x_0) f'(x_0) + \frac{(x - x_0)^2}{2} f''(\xi) \right] dx = \int_a^b \frac{(x - x_0)^2}{2} f''(\xi) dx =$$

$$= f''(\eta) \int_a^b \frac{(x - x_0)^2}{2} dx = \frac{f''(\eta)}{24}(b-a)^3.$$

Таким чином

$$|R_0(f)| \leq \frac{M_2}{24}(b-a)^3 \qquad (2)$$

Але тут у нас немає впливу на точність (величину похибки). Тому використовують формулу складеного типу. Якщо сітка рівномірна, то складена квадратурна формула прямокутників має вигляд

$$I(f)=\sum_{i=1}^{N}\int_{x_{i-1}}^{x_i} f(x)dx \approx \sum_{i=1}^{N} hf(\bar{x}_i)=I_h^{np}(f), \qquad (3)$$

де $\bar{x}_i = x_{i-\frac{1}{2}} = x_i - \frac{h}{2}$.

Оцінимо похибку цієї квадратурної формули:

$$R_h(f)=I(f)-I_h(f)=\sum_{i=1}^{N}\int_{x_{i-1}}^{x_i}[f(x)-f(\bar{x}_i)]dx=\sum_{i=1}^{N} f''(\eta_i)\frac{h^3}{24},$$

$$|R_h(f)| \le \frac{M_2}{24}\sum_{i=1}^{N} h^3 = \frac{M_2 h^2}{24}\sum_{i=1}^{N} h = \frac{M_2 h^2 (b-a)}{24} \qquad (4)$$

Тобто ця формула має степінь точності $p=2$ по кроку h. (Не слід плутати з алгебраїчним степенем точності $m=1$ для цієї формули).

Якщо $f(x)\in C^4[a,b]$, то

$$f(x)-f(\bar{x}_i)=f(\bar{x}_i)+(x-\bar{x}_i)f'(\bar{x}_i)+\frac{(x-\bar{x}_i)^2}{2}f''(\bar{x}_i)+$$

$$+\frac{(x-\bar{x}_i)^3}{6}f'''(\bar{x}_i)+\frac{(x-\bar{x}_i)^4}{24}f^{(4)}(\xi_i)-f(\bar{x}_i).$$

При непарних степенях інтеграли пропадуть і тому:

$$R_h(f)=\sum_{i=1}^{N}\int_{x_{i-1}}^{x_i}\frac{(x-\bar{x}_i)^2}{2}f''(\bar{x}_i)dx+\sum_{i=1}^{N}\int_{x_{i-1}}^{x_i}\frac{(x-\bar{x}_i)^4}{24}f^{(4)}(\xi_i)dx=\frac{h^2}{24}\sum_{i=1}^{N} hf''(\bar{x}_i)+\sum_{i=1}^{N}\frac{h^5 f^{(4)}(\eta_i)}{1920}.$$

Оскільки $\sum_{i=1}^{N} hf''(\bar{x}_i)$ - це квадратурна формула середніх прямокутників для $f''(x)$ з похибкою $O(h^2)$, то

$$R_h(f)=\frac{h^2}{24}\int_a^b f''(x)dx+O(h^4)=O(h^4).$$

$$R_h^{np}(f)=\overset{0}{R}_h(f)+\alpha(h), \text{ де } \overset{0}{R}_h(f)=\frac{h^2}{24}\int_a^b f''(x)dx,\ \alpha(h)=O(h^4). \qquad (5)$$

Вона використовується для побудови програм, що автоматично вибирають крок інтегрування.

9.3. Формула трапеції [СГ, 164-165]

Нехай $x_0=a$, $x_1=b$, $L_1(x)=f(x)$. Тоді отримаємо формулу:

$$I_1(f)=\frac{b-a}{2}[f(a)+f(b)] \qquad (1)$$

Формула має алгебраїчний степінь точності $m=1$, оскільки $I(x^2)\neq I_1(x^2)$. Це формула замкненого типу. Залишковий член:

$$R_1(f)=\int_a^b \frac{f''(\xi)(x-a)(x-b)}{2}dx=-\frac{(b-a)^3}{12}f''(\xi) \qquad (2)$$

Оцінка залишкового члена:

$$|R_1(f)|\le M_2\frac{(b-a)^3}{12}. \qquad (3)$$

З геометричної точки зору замінюється площа криволінійної трапецій площею звичайної трапеції.

Складена квадратурна формула трапецій:

$$I_h(f)=\sum_{i=1}^{N}\frac{h}{2}\left[f(x_{i-1})+f(x_i)\right]=\frac{h}{2}f(a)+\sum_{i=1}^{N-1}hf(x_i)+\frac{h}{2}f(b), \qquad (4)$$

де $x_i=a+ih,\ h=\dfrac{b-a}{N},\ i=\overline{0,N}$ та

$$|R_h(f)|\le M_2\frac{(b-a)}{12}h^2,\ f\in C^2[a,b]\ ,. \qquad (5)$$

Якщо $f\in C^4[a,b]$, то

$$R_h(f)=\overset{0}{R}_h(f)+\alpha(h),\ \text{де}\ \overset{0}{R}_h(f)=-\frac{h^2}{12}\int_a^b f''(x)dx,\ \alpha(h)=O\left(h^4\right). \qquad (6)$$

Задача 29 Використовуючи явний вигляд головних членів похибки складених квадратурних формул прямокутників та трапецій, побудувати лінійною комбінацією цих двох формул квадратурну формулу четвертого степеня точності за кроком h.

9.4 Квадратурна формула Сімпсона [СГ, 165-167]

Нехай $x_0=a,\ x_1=\dfrac{a+b}{2},\ x_2=b$. Замість f використовуємо $L_2(x)$. Тоді отримаємо квадратурну формулу:

$$I_2(f)=\frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right]. \qquad (1)$$

Це *квадратурна формула Сімпсона.*

Задача 30 Довести, що алгебраїчна степінь точності квадратурної формули Сімпсона $m=3$.

Задача 31 Довести, що для $f\in C^4[a,b]$ залишковий член квадратурної формули Сімпсона має місце представлення:

$$R_2(f)=\frac{1}{24}\int_a^b(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)f^{(4)}(\xi)dx=\frac{f^{(4)}(\xi)}{2880}(b-a)^5, \qquad (2)$$

та вірна оцінка:

$$|R_2(f)| \le \frac{M_4}{2880}(b-a)^5. \qquad (3)$$

Складена квадратурна формула Сімпсона має вигляд:

$$I_h(f) = \sum_{i=1}^{N} \frac{h}{6}\left[f(x_{i-1}) + 4f\left(x_{i-\frac{1}{2}} \right) + f(x_i) \right] =$$

$$= \frac{h}{6}\left[f(x_0) + 4f\left(x_{\frac{1}{2}} \right) + 2f(x_1) + \ldots + 2f(x_{N-1}) + 4f\left(x_{N-\frac{1}{2}} \right) + f(x_N) \right] \qquad (4)$$

Якщо $f \in C^4[a,b]$, то має місце оцінка:

$$|R_h(f)| \le \frac{M_4}{2880}(b-a)h^4,\ p=4. \qquad (5)$$

Якщо $f \in C^6[a,b]$, то

$$R_h(f) = \overset{0}{R}_h(f) + \alpha(h), \text{ де } \overset{0}{R}_h(f) = \frac{h^4}{2880}\int_a^b f^{(4)}(x)\,\mathrm{d}x\ , \alpha(h) = O(h^6). \qquad (6)$$

Задача 32 Побудувати інтерполяційну квадратурну формулу для $n=3$, $x_0 = a$, $x_1 = \frac{3a+b}{4}, x_2 = \frac{a+3b}{4}, x_3 = b$. Який алгебраїчний степінь точності вона має?

9.5. Принцип Рунге [СГ, 169-171]

Нехай задана деяка величина I (сіткова функція, інтеграл, неперервна функція). Нехай $I_h \approx I$ та $I_h \to I$ при $h \to 0$. Нехай похибка послідовності I_h представляється у вигляді

$$R_h = I - I_h = \overset{0}{R}_h + \alpha(h) \qquad (1)$$

де $\overset{0}{R}_h = C \cdot h^m$ - головний член похибки, C не залежить від h, $\alpha(h) = o(h^m)$. Обчислимо $I_{h/2}$. З (1) слідує, що

$$I = I_h + Ch^m + \alpha(h),\ I = I_{h/2} + C\frac{h^m}{2^m} + \alpha(h).$$

Звідси

$$I_{h/2} - I_h = \frac{Ch^m}{2^m}\left(2^m - 1\right) + \alpha(h).$$

З (1)

$$\overset{0}{R}_{h/2} = \frac{Ch^m}{2^m} = \frac{I_{h/2} - I_h}{2^m - 1} \qquad (2)$$

та $\overset{0}{R}_h = \frac{2^m}{2^m - 1}\left(I_{h/2} - I_h \right)$. Формула (2) носить назву *апостеріорної оцінки* похибки обчислення I за допомогою наближення $I_{h/2}$. (*Апріорні оцінки* це оцінки отримані до обчислення величини I_h, *апостеріорні оцінки* – під час її обчислення).

З формули (2) витікає такий алгоритм обчислення інтегралу із заданою точністю ε:

- обчислюємо I_h, $I_{h/2}$, $\overset{0}{R}_{h/2}$; перевірюємо чи $\left|\overset{0}{R}_{h/2}\right| < \varepsilon$. Якщо так, то $I \approx I^0_{h/2}$.

  Якщо ж ні, то
- обчислюємо $I_{h/2}$, $I_{h/4}$, $\overset{0}{R}_{h/4}$; перевіряємо $\left|\overset{0}{R}_{h/4}\right| < \varepsilon$ і т.д.
- Процес продовжуємо поки не буде виконана умова $\left|\overset{0}{R}_{h2^{-k}}\right| < \varepsilon \quad k = 1,2,\ldots$.

Зауваження Ми даємо оцінку не похибки, а її головного члена з точністю $\alpha(h)$, тому такий метод може давати збої, якщо не виконана умова $|\alpha(h)| << \left|R^0_h\right|$.

За допомогою головного члена похибки можна отримати краще значення для I:

$$\widetilde{I}_{h/2} = I^{(1)}_{h/2} = I_{h/2} + R^0_{h/2} = \frac{2^m}{2^m - 1} I_{h/2} - \frac{1}{2^m - 1} I_h . \tag{3}$$

Це *екстраполяційна формула Річардсона*: $I_h - \widetilde{I}_{h/2} = \alpha(h)$.

Для квадратурної формули трапецій $p = 2$ і

$$I - I_h = Ch^2 + O\left(h^4\right),\ \overset{0}{R}_{h/2} = \frac{I_{h/2} - I_h}{3}.$$

Маємо

$$R_h = -\frac{h^2}{12}\int_a^b f''(x)\,\mathrm{d}x + O\left(h^4\right) = O\left(h^2\right).$$

Отже, якщо застосовувати екстраполяційну формулу Річардсона, то

$$\widetilde{I}_{h/2} = \frac{4}{3} I_{h/2} - \frac{1}{3} I_h \tag{4}$$

і $I_h - \widetilde{I}_{h/2} = O\left(h^4\right)$.

Задача 33 Написати явний вигляд квадратурної формули, яка отримується екстраполяцією Річардсона з квадратурної формули трапецій.

Якщо невідомо m, то можна використати принцип Рунге для його знаходження. Для цього використаємо I_h, $I_{h/2}$, $I_{h/4}$:

$$I_{h/2} - I_h = \frac{Ch^m}{2^m}\left(2^m - 1\right) + \alpha(h),\ I_{h/4} - I_{h/2} = \frac{Ch^m}{2^{2m}}\left(2^m - 1\right) + \alpha(h).$$

З точністю $\alpha(h)$ маємо

$$2^m = \frac{I_{h/2} - I_h}{I_{h/4} - I_{h/2}}.$$

Звідки

$$m = \log_2 \frac{I_{h/2} - I_h}{I_{h/4} - I_{h/2}}.$$

Оцінка $\left|R^0_{h/4}\right| < \varepsilon$ - найбільш точна, тому $I \approx I_{h/4}$.

Покажемо чому формулу прямокутників рідко використовують з принципом Рунге._Нехай точки, в яких обчислюється значення функції позначаються: в $I_h \leftrightarrow \circ$, в - $I_{h/2} \leftrightarrow *$.

Для формули трапецій використовуються такі точки:

$$\circ - \ -\circ - \ -\circ - \ -\circ - \ -\circ - \ -\circ - \ -\circ - \ -\circ \ I_h$$

$$\circ - * - \circ - * - \circ - * - \circ - * - \circ - * - \circ - * - \circ - * - \circ - * - \circ \ I_{h/2}.$$

Для формули прямокутників:

$$- \circ - \ -\circ - \ -\circ - \ -\circ - \ -\circ - \ - \circ - \ -\circ - \ -\circ - \ I_h$$

$$-*-*-*-*-* \ -*-* \ -*-*-*-*-* \ -*-* \ -*-*-*- \ I_{h/2}.$$

Як бачимо для формули трапецій необхідно підраховувати нові значення в N точках, а для формули прямокутників в $2N$.

Для економного використання обчислених значень функції на сітці з кроком h для формули трапецій запишемо

$$I_h^{mp} = \frac{h}{2}\left(f(a) + 2\sum_{i=1}^{N-1} f(x_i) + f(b)\right),$$

$$I_{h/2}^{mp} = \frac{h}{4}\left(f(a) + 2\sum_{i=1}^{N-1} f(x_i) + 2\sum_{i=1}^{N-1} f\left(x_{i-1/2}\right) + f(b)\right) = \frac{1}{2} I_h^{mp} + \frac{h}{2}\sum_{i=1}^{N-1} f\left(x_{i-1/2}\right). \quad (5)$$

Отже, на одному кроці принципу Рунге кількість обчислень $Q^{mp} = O(N)$, а для $Q^{np} = O(2N)$.

Цей принцип застосовується і для формули Сімпсона $m = 4$. Головна частина залишкового члена для цієї формули:

$$\overset{0}{R}_{h/2} = \frac{I_{h/2} - I_h}{15}.$$

$$\tilde{I}_{h/2} = \frac{16}{15} I_{h/2} - \frac{1}{15} I_h,\ I_h - \tilde{I}_{h/2} = O\left(h^6\right)$$

Розглянемо використання так званих *адаптивних квадратурних формул*, в яких змінний крок вибирається за принципом Рунге. Для цього запишемо формулу трапецій із змінним кроком:

$$I_h^{mp}(f)=\sum_{i=1}^{N}\frac{h_i}{2}\left[f(x_{i-1})+f(x_i)\right],\ \text{де}\ h_i=x_i-x_{i-1}.$$

Оцінимо похибку на кожному інтервалі:

$$R_{h_i}=I_i-I_{h_i}=\int_{x_{i-1}}^{x_i}f(x)\,\mathrm{d}\,x-\frac{h_i}{2}\left[f(x_{i-1})+f(x_i)\right]=-\frac{h_i^3}{6}f''\left(x_{i-1/2}\right)+O\left(h_i^{\,5}\right).$$

Таким чином $p=3$ і головний член похибки:

$$R^0_{h_i/2}=\frac{\left(I_{h_i/2}-I_{h_i}\right)}{7}.$$

Умова припинення ділення навпіл проміжку $[x_{i-1},x_i]$:

$$\left|R^0_{h_i/2}\right|\le\frac{\varepsilon\cdot h_i}{b-a}.$$

Це забезпечує точність на всьому інтервалі

$$\left|R_{h/2}\right|=\left|\sum_{i=1}^{N}R_{h_i/2}\right|\le\sum_{i=1}^{N}\frac{\varepsilon h_i}{b-a}=\varepsilon\frac{b-a}{b-a}=\varepsilon.$$

Ще одне застосування принципу Рунге – високоточне обчислення інтегралу від достатньо гладкої функції за допомогою таблиці Ромберга. Для побудови цієї таблиці обчислимо за допомоги складеної квадратурної формули трапецій із сталим кроком h послідовність значень $I_h=I_h^{(0)},I_{h/2}=I_{h/2}^{(0)},I_{h/4}=I_{h/4}^{(0)},I_{h/8}=I_{h/8}^{(0)},\ldots$, які мають похибку $O(h^2)$. За допомогою екстраполяції Ричардсона (3) з коефіцієнтами лінійної комбінації $\left(\frac{4}{3},-\frac{1}{3}\right)$ уточнимо ці значення (див. також формулу (4)). Отримаємо $I_{h/2}^{(1)},I_{h/4}^{(1)},I_{h/8}^{(1)},\ldots$. Вони мають похибку $O(h^4)$. Знову використовуємо екстраполяцію Ричардсона з коефіцієнтами лінійної комбінації $\left(\frac{16}{15},-\frac{1}{15}\right)$. Отримаємо $I_{h/4}^{(2)},I_{h/8}^{(2)},\ldots$, які мають точність $O(h^6)$ і т.д.. Отримані значення можна розмістити в такій таблиці Ромберга:

$$\begin{array}{cccc}
I_h^{(0)} & & & \\
I_{h/2}^{(0)} & I_{h/2}^{(1)} & & \\
I_{h/4}^{(0)} & I_{h/4}^{(1)} & I_{h/4}^{(2)} & \\
I_{h/8}^{(0)} & I_{h/8}^{(1)} & I_{h/8}^{(2)} & I_{h/8}^{(3)}
\end{array}$$

Всі значення крім останнього $I_{h/8}^{(3)}$ можна оцінити за принципом Рунге (див. формулу (2)). Використання формули (5) для обчислення $I_{h2^{-k}}^{(0)}$ та лінійні

комбінації (2) дають простий та економічний алгоритм обчислення I. Початкове значення h можна брати рівним $b-a$, або $\frac{b-a}{n}$, де n ціле.

9.6. Квадратурні формули найвищого алгебраїчного степеня точності [ЛМС, 89-95], [СГ, 180-183], [БЖК, 113-115, 102-108]

Розглянемо інтеграл

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

де

$$\rho(x)>0 \quad x\in[a,b] \quad \left|\int_a^b \rho(x)x^i dx\right|<\infty .$$

Розглянемо задачу побудови квадратурної формули

$$I_n(f)=\sum_{k=1}^{n} c_k f(x_k), \tag{2}$$

яка при заданому n була б точною для алгебраїчного багаточлена можливо більшого степеня. Такі квадратурні формули існують, вони називаються *квадратурні формули найвищого алгебраїчного степеня точності* або формули Гаусса (або Гаусса – Кристофеля).

В (2) невідомими є c_k, x_k, $k=\overline{1,n}$. Їх обирають з умови, що (2) точна для будь-якого багаточлена степеня p, а це еквівалентно умові, щоб формула була точною для функції $f(x)=x^\alpha, \alpha=0,1,...,p$. Звідси отримуємо умови:

$$I_n(x^\alpha)=\int_a^b \rho(x)x^\alpha dx=\sum_{k=1}^{n} c_k x_k^\alpha, \ a=\overline{0,p}. \tag{3}$$

Ми хочемо отримати формули для $m\to\max$. Щоб кількість рівнянь була рівною кількості невідомих нам потрібно, щоб $p+1=2n$.

Задача 34 Побудувати квадратурну формулу найвищого степеня точності (розв'язати систему рівнянь (3)) для $a=-1, b=1, \rho(x)=1$.

Теорема Гаусса Квадратурна формула (2) буде точною для будь-якого багаточлена степеня $p=2n-1$, тобто $\forall f(x)\in\pi_{2n-1}$ тоді і тільки тоді, коли виконуються умови:

1) поліном $\omega(x)=(x-x_1)(x-x_2)...(x-x_n)$ ортогональний з вагою $\rho(x)$ до будь-якого багаточлена степеня менше n Q_{n-1}:

$$\int_a^b \omega(x)Q_{n-1}(x)\rho(x)dx=0; \tag{4}$$

2) формула (2) є квадратурною формулою інтерполяційного типу , тобто коефіцієнти обчислюються за формулою

$$c_k=\int_a^b \rho(x)\frac{\omega(x)}{(x-x_k)\omega'(x_k)}dx \tag{5}$$

◁ Необхідність. Нехай формула (2) точна для багаточлена степеня $p = 2n-1$, тобто $I(f) = I_n(f)\ \forall f(x) \in \pi_{2n-1}$. Тоді

$$I(f) = \int_a^b \rho(x)\omega(x)Q_{n-1}(x)dx = \sum_{k=1}^{n} c_k \omega(x_k)Q_{n-1}(x_k) = 0.$$

Тобто виконується (4). Тепер покладемо

$$f(x) = \frac{\omega(x)}{(x-x_j)\omega'(x_j)} \in \pi_{n-1} \subset \pi_{2n-1}.$$

Отримаємо

$$\int_a^b \rho(x)f(x)dx = \int_a^b \rho(x)\frac{\omega(x)}{(x-x_j)\omega'(x_j)}dx = \sum_{k=1}^{n} c_k \frac{\omega(x_k)}{(x_k-x_j)\omega'(x_j)} = \sum_{k=1}^{n} c_k \delta_{kj} = c_j,$$

тобто виконується і умова (5).

Достатність. Нехай виконується (4) і (5). Подамо $\forall f(x) \in \pi_{2n-1}$ у вигляді

$$f(x) = \omega(x)Q_{n-1}(x) + R_{n-1}(x).$$

Розглянемо

$$I(f) = \int_a^b \rho(x)f(x)dx = \int_a^b \rho(x)(\omega(x)Q_{n-1}(x) + R_{n-1}(x))dx =$$

$$= \sum_{k=1}^{n} c_k \omega(x_k)Q_{n-1}(x_k) + \sum_{k=1}^{n} c_k R_{n-1}(x_k)$$

Так як $R_{n-1}(x_k) = f(x_k) - \omega(x_k)Q_{n-1}(x_k) = f(x_k)$, то

$$I(f) = \sum_{k=1}^{n} c_k f(x_k) = I_n(f).$$

Тобто формула (2) є точною для будь-якого багаточлена степеня $2n-1$. ▷

Отже, з точністю до сталого множника багаточлени $\omega(x)$ співпадають з багаточленами *n*-того степеня ортогональної системи багаточленів. Ця система ортогональна на проміжку $[a,b]$ з вагою $\rho(x)$.

Вивчимо деякі властивості квадратурних формул Гаусса.

1) Покажемо, що c_k, x_k визначаються однозначно.

Представимо багаточлен $\omega(x)$ у вигляді $\omega(x) = a_0 + a_1 x + ... + a_{n-1}x^{n-1} + x^n$. Умови ортогональності (4) приймуть вигляд

$$\int_a^b \rho(x)\omega(x)x^\alpha dx = \int_a^b \rho(x)(a_0 + a_1 x + ... + a_{n-1}x^{n-1} + x^n)x^\alpha dx = 0,\ \alpha = \overline{0, n-1}$$

або

$$\int_a^b \rho(x)(a_0 + a_1 x + ... + a_{n-1}x^{n-1})x^\alpha dx = -\int_a^b \rho(x)x^n x^\alpha dx$$

Покажемо, що відповідна однорідна система рівнянь

$$\int_a^b \rho(x)(a_0 + a_1 x + ... + a_{n-1}x^{n-1})x^\alpha dx = 0,\ \alpha = \overline{0, n-1} \qquad (6)$$

має єдиний розв'язок $a_0 = a_1 = ... = a_{n-1} = 0$.

Помножимо систему (6) на a_α і просумуємо по всіх $\alpha = \overline{0, n-1}$:

$$\sum_{\alpha=0}^{n-1} a_\alpha \int_a^b \rho(x)\left(a_0 + a_1 x + ... + a_{n-1}x^{n-1}\right)x^\alpha dx =$$

$$= \int_a^b \rho(x)\sum_{\alpha=0}^{n-1}\sum_{j=0}^{n-1} a_j a_\alpha x^\alpha x^j dx = \int_a^b \rho(x)\left(\sum_{j=0}^{n-1} a_j x^j\right)^2 dx = 0.$$

Звідси і з умови $\rho(x) > 0$ випливає, що $a_0 = a_1 = ... = a_{n-1} = 0$. Тому і відповідна неоднорідна система має єдиний розв'язок. Отже існує єдиний багаточлен $\omega(x)$ степеня n, який ортогональний з вагою $\rho(x)$ до будь-якого багаточлена степеня $n-1$.

2) Покажемо, що найвищий степінь точності формули Гаусса $p = 2n-1$. З теореми випливає, що $p \geq 2n-1$. Покажемо, що існує багаточлен степеня $2n$, для якого формула не є точною. Для цього введемо функцію $f(x) = \omega^2(x) \in \pi_{2n}$. Маємо

$$I(f) = \int_a^b \rho(x)\omega^2(x)dx > 0,$$

але

$$I_n(f) = \sum_{k=1}^n c_k \omega^2(x_k) = 0.$$

Отже, $I(f) \neq I_n(f)$. Звідси $p \leq 2n-1$, тобто $p = 2n-1$.

3) Коефіцієнти формул Гаусса додатні, тобто $c_k > 0$. Розглянемо багаточлени

$$\varphi_j = \left[\frac{\omega(x)}{(x-x_j)\omega'(x_j)}\right]^2,$$

які мають степінь $2n-2$ і властивості:

$$\varphi_i(x_k) = \delta_{ik},\ I(\varphi_j) = \int_a^b \rho(x)\varphi_j(x)dx > 0.$$

Так як для цих багаточленів справедлива теорема Гаусса, то

$$I(\varphi_j) = I_n(\varphi_j) = \sum_{k=1}^n c_k \varphi_j(x_k) = \sum_{k=1}^n c_k \left[\frac{\omega(x_k)}{(x-x_j)\omega'(x_j)}\right]^2 = \sum_{k=1}^n c_k \delta_{jk}^2 = c_j.$$

Звідси випливає, що $c_j > 0,\ j = \overline{1,n}$.

4) Теорема Нехай вагова функція $\rho(x) > 0 \quad x \in [a,b] \quad f(x) \in C^{2n}[a,b]$. Тоді існує точка $\xi \in [a,b]$ така, що

$$R_n(f) = \frac{f^{(2n)}(\xi)}{(2n)!}\int_a^b \rho(x)\omega^2(x)dx. \qquad \textbf{(7)}$$

◁ Розглянемо інтерполяційний багаточлен Ерміта з двократними вузлами

$$H_{2n-1}(x):\ f(x_i) = H_{2n-1}(x_i)\ ,\ f'(x_i) = H'_{2n-1}(x_i),\ \ i = \overline{1,m}.$$

Для нього

$$r_{2n-1}(x)=f(x)-H_{2n-1}(x)=\frac{f^{(2n)}(\xi)}{(2n)!}\omega^2(x).$$

Звідси

$$R_{2n-1}(x)=\int_a^b \rho(x)r_{2n-1}(x)dx=\frac{f^{(2n)}(\xi)}{(2n)!}\int_a^b \rho(x)\omega^2(x)dx. \quad \triangleright$$

9.7. Частинні випадки квадратурної формули Гаусса [ЛМС, 99]

1) Розглянемо відрізок $[-1,1]$ і ваговий множник $\rho(x)=1$, тобто виведемо формули Гаусса для обчислення інтегралу

$$\int_{-1}^{1} f(x)dx.$$

Щоб знайти вузли квадратурна формули розглянемо багаточлени Лежандра

$$(n+1)L_{n+1}(x)=(2n+1)xL_n(x)-nL_{n-1}(x),\ \ L_0=1,\ L_1(x)=x.$$

Багаточлени Лежандра задовольняють умовам теореми 1 (пункт 1), тому $\omega(x)=L_n(x)$ і вузлами квадратурної формули є корені цього багаточлена. Вагові множники цієї формули обчислюються за формулою

$$c_k=\int_{-1}^{1}\frac{\omega(x)}{(x-x_k)\omega'(x_k)}dx,$$

а залишковий член

$$R_n(f)=\frac{2^{2n+1}}{(2n+1)!(2n)!}\frac{(n!)^2}{(2n)!}f^{(2n)}(\xi)\ .$$

Приклад. Побудуємо квадратурну формулу

$$\int_0^1 f(x)dx\approx\sum_{k=1}^{n}c_k f(x_k).$$

При $n=2$ потрібно знайти c_0,c_1,x_0,x_1. Заміною $t=2x-1$ переведемо $x\in[0,1]$ на проміжок $t\in[-1,1]$. Запишемо $L_2(t)=\frac{3}{2}t^2-\frac{1}{2}$. Звідси

$$L_2(x)=\frac{3(2x-1)-1}{2}=\frac{12x^2-12x+2}{2}=6x^2-6x+1=0.$$

Звідси $x_1=\frac{3-\sqrt{3}}{6}, x_2=\frac{3+\sqrt{3}}{6}$. За формулою (5) знайдемо

$$c_1=\int_0^1\frac{x-x_2}{x_1-x_2}dx=\frac{1}{2},\ \ c_2=\int_0^1\frac{x-x_1}{x_2-x_1}dx=\frac{1}{2}.$$

Тобто квадратурна формула має вигляд

$$\int_0^1 f(x)dx = \frac{1}{2}\left(f\left(\frac{3-\sqrt{3}}{6}\right) + f\left(\frac{3+\sqrt{3}}{6}\right)\right).$$

2) Розглянемо відрізок $[-1,1]$ і вага $\rho(x) = \left(1-x^2\right)^{-\frac{1}{2}}$, тобто виведемо формули Гаусса для обчислення інтегралу

$$I(f) = \int_{-1}^1 \frac{f(x)dx}{\sqrt{1-x^2}}.$$

Багаточлени Чебишова задовольняють умовам теореми 1 (п.1), тому

$$\omega(x) = \overline{T}_n(x) = \frac{1}{2^{n-1}}\cos(n\arccos x).$$

Вузлами квадратурної формули є корені цього багаточлена Чебишова, тобто корені рівняння $\cos(n\arccos x) = 0$. Звідси

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

Відповідні коефіцієнти обчислюються за формулами

$$c_k = \int_{-1}^1 \frac{T_n(x)dx}{\sqrt{1-x^2}\ T_n'(x_k)(x-x_k)} = \frac{\pi}{n},\ k = \overline{1,n}.$$

Отже, квадратурні формули найвищого степеня точності (ці формули називають формулами Ерміта) мають вигляд

$$I_n(f) = \frac{\pi}{n}\sum_{k=1}^n f(x_k)\ , \qquad (1)$$

де x_k – корені багаточлена Чебишова.

Залишковий член має вигляд

$$R_n(f) = \frac{\pi}{2^{2n-1}(2n)!}f^{(2n)}(\xi).$$

3) Розглянемо проміжок $(-\infty,\infty)$ і вагу $\rho(x) = e^{-x^2}$, тобто виведемо формули Гаусса для обчислення інтегралу

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx.$$

За теорією

$$\omega(x) = H_n(x) = (-1)^n e^{x^2}\frac{d^n}{dx^2}e^{-x^2},$$

де $H_n(x)$ багаточлени Ерміта. Багаточлени Ерміта обчислюються також за рекурентними формулами

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x),\quad H_{-1} = 0,\ H_0 = 1\ .$$

Коефіцієнти квадратурної формули обчислюються за формулами

$$c_k = \frac{2^{n+1}n!\sqrt{\pi}}{[H_n'(x_k)]^2}.$$

Залишковий член

$$R_n(f)=\frac{n!\sqrt{\pi}}{2^n(2n)!}f^{(2n)}(\xi).$$

Наприклад, при $n=1$ $H_1(x)=2x$. Корінь $x_0=0$,

$$c_0=\int_{-\infty}^{\infty}e^{-x^2}dx=\sqrt{\pi}.$$

Квадратурна формула має вигляд:

$$I_1(x)=\sqrt{\pi}f(0).$$

4) Розглянемо відрізок $[0,\infty]$ і ваговий множник $\rho(x)=x^{\alpha}e^{-x}$, тобто виведемо формули Гаусса для обчислення інтегралу

$$\int_0^{\infty}x^{\alpha}e^{-x}f(x)dx.$$

За теорією

$$\omega(x)=L_n^{\alpha}(x)=(-1)^n x^{-\alpha}e^{x}\frac{d^n}{dx^n}\left(x^{\alpha+n}e^{-x}\right),$$

де $L_n^{\alpha}(x)$ багаточлени Лагера. Коефіцієнти обчислюються за формулами

$$c_k=\frac{P(n+1)P(n+\alpha+1)}{x_k\left[L_n^{\alpha}(x_k)\right]^2}.$$

Залишковий член при $\alpha=0$ рівний

$$R_n(f)=\frac{(n!)^2}{(2n)!}f^{(2n)}(\xi).$$

7) Інтегрування швидко осцилюючих функцій.

Розглянемо інтеграл

$$I(f)=\int_a^b f(x)e^{j\omega x}dx,\ j^2=-1.$$

При великих ω застосування складених квадратурних формул інтерполяційного типу приводить до великої похибки і при малих кроках h. Розглянемо $e^{j\omega x}$ як ваговий коефіцієнт, тобто $\rho(x)=e^{j\omega x}$. Замінимо $[a,b]$ на $[-1,1]$: $x_i=\frac{a+b}{2}+\frac{b-a}{2}d_i$, $d_i\in[-1,1]$, $i=\overline{1,n}$ (вузли можуть бути не рівновіддалені, якщо рівновіддалені, то $d_i=-1+i\frac{2}{n}$, $i=\overline{1,n}$).

Замінимо $f(x)$ на інтерполяційний багаточлен Лагранжа $L_{n-1}(x)$ з вузлами x_i і отримаємо формулу

$$I_n(f)=\int_a^b L_{n-1}(x)e^{j\omega x}dx, \qquad (2)$$

яка буде точною для всіх багаточленів не вище $n-1$ степеня . Тобто, якщо в (2) підставити багаточлен Лагранжа, то можна обчислити інтеграл і отримати квадратурну формулу

$$S_n^{\omega}(f)=\frac{b-a}{2}\exp\left\{j\omega\frac{a+b}{2}\right\}\sum_{i=1}^{n}D_i\left(\omega\frac{b-a}{2}\right)f(x_i),\ D_i(p)=\int_{-1}^{1}\left(\prod_{\substack{k=1\\k\neq i}}^{n}\frac{\xi-d_k}{d_i-d_k}\right)\exp(jp\xi)d\xi$$

.

При $n=3, d_1=-1, d_2=0, d_3=1$ – це формула Філона. Можна брати і більше точок, наприклад $n=5, d_1=-1, d_2=-\frac{1}{2}, d_3=0, d_4=\frac{1}{2}, d_5=1$. Ці формули не потрібно застосовувати, коли немає швидко осцилюючого множника.

9.8. Обчислення невласних інтегралів [БЖК, 146-153]

Розглянемо обчислення інтегралів з такими особливостями :

а) інтеграли другого роду, тобто

$$I=\int_a^b F(x)dx,\ F(x)\underset{x\to a\vee x\to b}{\to}\infty;$$

б) інтеграли першого роду

$$I=\int_a^{\infty}F(x)dx.$$

Розглянемо спочатку інтеграли другого роду, тобто

$$I=\int_a^b F(x)dx \qquad F(x)\underset{x\to a\vee x\to b}{\to}\infty.$$

1) *Мультиплікативний спосіб*. Представимо підінтегральну функцію у вигляді $F(x)=\rho(x)f(x)$, причому $\rho(x)$– особлива, а $f(x)$– гладка. Далі для обчислення інтегралу

$$I=\tilde{I}(f)=\int_a^b\rho(x)f(x)dx$$

використовуємо відповідні квадратурні формули Гаусса.

Приклад 1 Потрібно обчислити інтеграл

$$I=\int_{-1}^{1}\frac{dx}{\sqrt{1-x^4}}.$$

Точки $x=\pm1$ є особливими.

Представимо підінтегральну функцію у вигляді:

$$F(x)=\underbrace{\frac{1}{\sqrt{1-x^2}}}_{\rho(x)}\underbrace{\frac{1}{\sqrt{1+x^2}}}_{f(x)},$$

отримаємо інтеграл вигляду

$$I=\int_0^1\frac{1}{\sqrt{1+x^2}}\frac{dx}{\sqrt{1-x^2}},$$

де $\rho(x)=\frac{1}{\sqrt{1-x^2}}$.

Далі використовуємо квадратурну формулу Ерміта (1) з попереднього пункту і обчислюємо інтеграл.

Приклад 2 Обчислити інтеграл

$$I = \int_0^{\pi} \ln(\sin x)dx.$$

Особливі точки $x = 0, x = \pi$.

Зведемо цю особливість до степеневої:

$$\rho(x) = \frac{1}{\sqrt{x(\pi - x)}},$$

тоді

$$f(x) = \sqrt{x(\pi - x)}\ln(\sin x) \xrightarrow[x\to 0,\pi]{} 0.$$

Для знаходження інтегралу з таким $\rho(x)$ застосовуємо квадратурні формули Чебишова. Неприємності виникають, оскільки $f'(x) \xrightarrow[x\to 0,\pi]{} \infty$ (хоча квадратурні формули даватимуть наближене значення). Тому застосовують другий спосіб розв'язання проблеми:

2) Адитивний спосіб. Представимо підінтегральну функцію у вигляді $F(x) = f(x) + \psi(x)$,

причому $\psi(x)$–особлива, $f(x)$– гладка. Розбиваємо інтеграл на два: $I = I_1 + I_2$.

$I_1 = \int_a^b f(x)dx$ – обчислюють чисельно (наприклад, формули Сімпсона чи трапецій),

$I_2 = \int_a^b \psi(x)dx$ – пробують обчислити аналітично (можливо апроксимувати функцію $\psi(x)$, наприклад, рядом).

Приклад 3 Обчислити інтеграл з прикладу 2:

$$I = \int_0^{\pi} \ln(\sin x)dx.$$

$$I = 2\int_o^{\pi/2} \ln(\sin x)dx.$$

Представимо

$$\ln(\sin x) = \ln\frac{\sin x}{x} + \ln x.$$

Отримаємо два інтеграли:

$I_1 = \int_a^b \ln\frac{\sin x}{x}dx$ обчислюємо чисельно,

а $I_2 = \int_0^{\pi/2} \ln x dx = (x\ln x - x)\Big|_0^{\pi/2} = \frac{\pi}{2}\left(\ln\frac{\pi}{2} - 1\right)$.

Розглянемо тепер інтеграли першого роду

$$I = \int_a^{\infty} F(x)dx$$

Нехай $a>0$.Зробимо заміну $t = \frac{x-a}{x}$, $x = \frac{a}{1-t}$. Тоді

$$I = a\int_0^1 F\left(\frac{a}{1-t}\right)\frac{dt}{(1-t)^2},$$

а це інтеграл другого роду.

Якщо $a=0$, то робимо заміну $t = e^{-x}$, $x = -\ln t$, тоді

$$I = \int_0^1 F(-\ln t)\frac{dt}{t}.$$

Знову отримуємо інтеграл другого роду.

Якщо $a<0$ (не можна зробити заміну $t = \frac{x-a}{x}$, тому що виникає особливість в точці $x = 0$), розбиваємо інтеграл на два:

$$I = \int_a^0 F(x)dx + \int_o^{\infty} F(x)dx$$

і обчислюємо за допомогою попередніх пунктів.

Мультиплікативний спосіб обчислення інтегралів першого роду ґрунтується на представлені підінтегральної функції у вигляді

$$F(x) = \rho(x)f(x),$$

де, наприклад,

$$\rho(x) = x^{\alpha}e^{-x},\ x \in [0,\infty).$$

Такий ваговий коефіцієнт відповідає багаточленам Лагера. При $x \in (-\infty,\infty)$, $\rho(x) = e^{-x^2}$ приходимо до багаточленів Ерміта.

Ще один спосіб ґрунтується на обрізанні верхньої границі. Для цього інтеграл запишемо у вигляді

$$I = \int_a^{\infty} F(x)dx = \int_a^b F(x)dx + \int_b^{\infty} F(x)dx.$$

Верхня границя b обчислюють з умови

$$\left|\int_b^{\infty} F(x)dx\right| < \frac{\varepsilon}{2},$$

де ε – задана точність. Для обчислення $\int_a^b F(x)dx$ використовують квадратурні формули складеного типу.

9.9. Обчислення кратних інтегралів [Волков, 125-129], [Калиткин, 108-113, 121-123]

Розглянемо інтеграл

$$I = \int_a^b \int_c^d f(x,y)dxdy \tag{1}$$

Цей інтеграл зводиться до повторного, якщо ввести

$$F(x) = \int_c^d f(x,y)dy \tag{2}$$

Тоді

$$I = \int_a^b F(x)dx\,. \tag{3}$$

До однократних інтегралів можна застосувати квадратурну формулу середніх прямокутників. Тоді

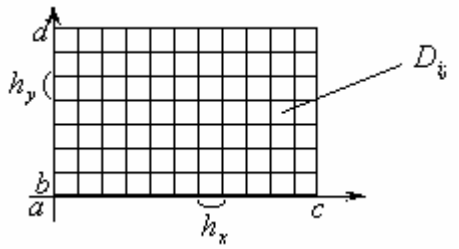
$$I \approx I_0 = (b-a)F(\bar{x}) = (b-a)\int_c^d f(\bar{x},y)dy \approx (b-a)(d-c)f(\bar{x},\bar{y}), \tag{4}$$

$$\bar{x} = \frac{a+b}{2},\ \bar{y} = \frac{c+d}{2}$$

Кубатурна формула (це формула наближеного обчислення інтегралу (1)), якщо використовується формула трапеції має вигляд:

$$I \approx I_1 = \frac{(b-a)(d-c)}{4}[f(a,c) + f(b,c) + f(a,d) + f(b,a)].$$

Точність залежить від поведінки функції та від довжини інтервалів $[a,b], [c,d]$. Аналог формул складеного типу для (1) будується таким чином: розбиваємо D на комірки (див. малюнок):



$$D = \bigcup_{i,j} D_{ij} \quad D_{ij} = \{x_{i-1} \le x \le x_i, y_{j-1} \le y \le y_j\},\ x_i = a + ih_x,\ i = \overline{0, N_x},$$

$$y_j = c + jh_y,\ j = \overline{0, N_y},\ h_x = \frac{b-a}{N_x},\ h_y = \frac{d-c}{N_y}$$

Тоді для кожного інтегралу по комірці застосовуємо кубатурну формулу прямокутників (4):

$$I = \sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\iint_{D_{ij}} f(x,y)dxdy \approx I_{0,h} = \sum_{i=1}^{N_x}\sum_{j=1}^{N_y} h_x h_y f(\bar{x}_i, \bar{y}_i)$$

$$\bar{x}_i = x_i - \frac{h_x}{2} \qquad \bar{y}_j = y_j - \frac{h_y}{2}.$$

Якщо $f(x,y) \subset C^2(\bar{D})$, то $I - I_{0,h} = O(h_x^2 + h_y^2)$. Степінь точності по крокам сітки – 2. Складність методу пропорційна кількості комірок

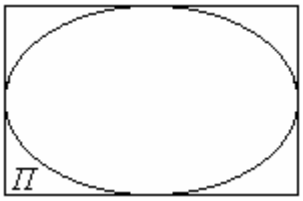
$$Q = O(N_x, N_y) = O(N^2),$$

$N \approx N_x \approx N_y$. В 3-вимірному просторі $f = f(x, y, z)$ складність - $Q = O(N^3)$.

Якщо D не прямокутник, то замість f введемо

$$\bar{f}(x, y) = \begin{cases} f(x, y), & x \in D, \\ 0, & x \in \Pi \setminus D. \end{cases}$$

де Π – найменший охоплюючий D прямокутник $D \subset \Pi$



Тоді $I = \iint_{\Pi} \bar{f}(x, y) dxdy$, що розглядався вище.

Недолік такого підходу: $\bar{f}(x, y)$ може бути розривною функцією і через це низька точність обчислення інтегралу.

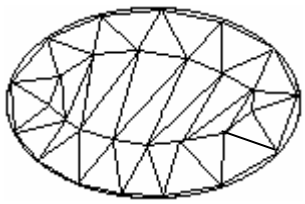
Наступний підхід базується на відповідній заміні змінних

$$x = x(\xi, \eta) \quad y = y(\xi, \eta) \quad D \to \Pi,$$

$$I = \iint_{\Pi} f(x(\xi, \eta), y(\xi, \eta)) J(\xi, \eta) d\xi d\eta,$$

де Π - прямокутник в площині (ξ, η); $J(\xi, \eta)$ – Якобіан переходу. Якщо границя області D гладка, то якобіан буде мати особливості, що також знижує швидкість збіжності.

Ще один підхід в обчисленні інтегралу по довільній області D базується на триангулювання області. Якщо область довільного вигляду, то її можна розбити на трикутники таким чином $D = \bigcup_{k=1}^{N} D_k$:



Тоді $I = \sum_{k=1}^{N} I_k$, $I_k = \iint_{D_k} f(x, y) dxdy$. Застосуємо кубатурні формули до кожного $I_k \approx I_k^h$. Для цього замінимо функцію поліномом першого степеня

$$f(x, y) \cong L_1(x, y) = A + Bx + Cy.$$

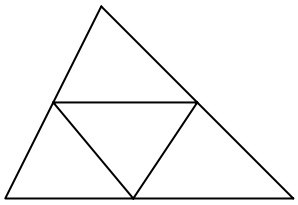
Задача 35 Побудувати явний вигляд кубатурної формули, яка дозволяє наближено обчислити I_k по трикутнику D_k, якщо замінити $f(x, y) \approx L_1(x, y)$ на інтерполяційний багаточлен 1-го степеня.

Точність такого підходу

$$I - I^h = I - \sum_{k=1}^{N} I_k^h = O(h^2),$$

де $h = \max_k diam D_k$.

Складність обчислення інтеграла: $Q = O(h^{-2})$ для $D \subset R_2$; $Q = O(h^{-3})$ для



$D \subset R_3$.

Можна згустити сітку, поділивши один трикутник D_k на чотири (див. малюнок):

І, нарешті, розглянемо ідею методу статистичних випробувань (метод Монте-Карло) для обчислення інтегралу $I = \iint\limits_D f(x,y)dxdy$. Замінимо наближено

$$I = \iint\limits_{\Pi} \bar{f}(x,y)dxdy \approx \frac{1}{N}\sum_{i=1}^{N} \bar{f}(\xi_i,\eta_i)\cdot mes\Pi,$$

де Π – найменший охоплюючий D прямокутник $D \subset \Pi$; $\bar{f}(x,y)$ - продовження функції (4); ξ_i, η_i - незалежні реалізації рівномірно розподілених на $[a,b]$ та $[c,d]$ випадкових величин ξ та η. Складність цього методу $Q = O(N)$. Оцінка точності $I - I_N = O(N^{-\frac{1}{2}})$ носить ймовірнистний характер. Позитивна сторона методу – незалежна від розмірності складність; негативна – низька точність.

10. Чисельні методи розв'язання задачі Коші для звичайних диференційних рівнянь

Постановка задачі : нехай потрібно знайти розв'язок диференційного рівняння з початковими умовами

$$\frac{du}{dt} = f(t,u),\ t > t_0,\ u(t_0) = u_0, \tag{1}$$

$u = (u_1,...,u_m)^T$, $f = (f_1,...,f_m)^T$, $u_k = u_k(t)$, $f_k = f_k(t,u_1,...,u_m)$.

Якщо довільна функція f_k неперервна по кожній своїй змінній та по u_j вони Ліпшиць-неперервні, тобто

$$\left|f_k(t,...,u_j,...) - f_k(t,...,v_j,...)\right| \le L_j\left|u_j - v_j\right|,\ \forall j,k,$$

то розв'язок задачі (1) існує і єдиний.

Нехай задано рівняння *m*-ого порядку та початкові умови:

$$\begin{cases} v^{(m)}(t) = F(t,v,v',...,v^{(m-1)}), \\ t = t_0 : v = v_1, v' = v_2,..., v^{(m-1)} = v_m. \end{cases} \tag{2}$$

Введемо компоненти вектора $\vec{u}$: $u_k(t) = v^{(k-1)}(t)$. Тоді задача (2) записується у вигляді системи (1):

$$\begin{cases} \dfrac{du_1}{dt} = u_2, & u_1(t_0) = v_1, \\ \ldots \\ \dfrac{du_{m-1}}{dt} = u_m, & u_{m-1}(t_0) = v_{m-1}, \\ \dfrac{du_m}{dt} = F(t, u_1, \ldots, u_m), & u_m(t_0) = v_m. \end{cases}$$

Тому далі, в основному, розглядаються методи розв‘язання задачі (1).

10.1. Наближені аналітичні методи [БЖК 358-360; ЛМС 254-255]

1) *Метод послідовних наближень (метод Пікара)*

Потрібно розв’язати диференційне рівняння з відповідними початковими умовами:

$$\frac{du}{dt} = f(t,u), \quad u(t_0) = u_0. \tag{1}$$

Проінтегруємо (1)

$$u(t) = u(t_0) + \int_{t_0}^{t} f(\xi, u(\xi))d\xi \tag{2}$$

Задаємо $u^{(0)}(t)$ і запишемо ітераційний процес:

$$u^{(k+1)}(t) = u_0 + \int_{t_0}^{t} f(\xi, u^{(k)}(\xi))d\xi, \; k = 0,1,\ldots. \tag{3}$$

Існує $T = T(u_0, L)$ ($L = \max\limits_j L_j$ - стала Ліпшиця) така, що

$$u^{(n)}(t) \underset{n\to\infty}{\to} u(t), \; t \in [t_0, T].$$

Тому $u(t) \approx u^{(n)}(t)$.

Недолік методу: необхідно проведення аналітичного інтегрування

2) *Метод рядів Тейлора*

Нехай розв’язок задачі (1) можна представити у вигляді ряду

$$u(t) = \sum_{k=0}^{\infty} \frac{u^{(k)}(t_0)}{k!}(t - t_0)^k.$$

Будемо шукати наближення у вигляді скінченої суми

$$u(t) \approx u^N(t) = \sum_{k=0}^{N} \frac{u^{(k)}(t_0)}{k!}(t - t_0)^k, \; t \in [t_0, t_1]. \tag{4}$$

Для визначення $u^{(k)}(t_0)$ диференціюємо рівняння (1) по t:

$$u^{(0)}(t_0) = u_0, \; u^{(1)}(t_0) = f(t_0, u_0),$$

$$u^{(2)}(t_0) = f_{t,0} + f_{u,0}f_0, \; u^{(3)}(t_0) = f_{tt,0} + 2f_{tu,0}f_0 + f_{uu,0}{f_0}^2, \ldots.$$

Якщо $\tau = t_1 - t_0$ малий параметр, то

$$\left|u(t)-u^N(t)\right| = O(\tau^{N+1}).$$

Недоліки методу:

– зростання кількості доданків при обчисленні $u^{(n)}(t_0)$;

– необхідно аналітичного диференціювання.

10.2. Методи типу Ейлера [СГ, 214- 218]

Розглянемо задачу Коші:

$$\frac{du}{dt} = f(t,u),\ u(t_0)=u_0. \qquad (1)$$

Використаємо перше наближення за допомогою рядів Тейлора на проміжку $t_0 \le t \le t_1$:

$$u(t) \approx u^1(t) = u(t_0) + (t-t_0) f(t_0,u_0).$$

Обчислимо наближене значення в точці t_1:

$$u(t_1) = u_1 \approx u^1(t_1) = u_0 + \tau f(t_0,u_0),$$

де $\tau = t_1 - t_0$ - деякий крок. Якщо позначити $y_1 = u(t_1)$, $y_0 = u(t_0)$, то маємо формулу

$$y_1 = y_0 + \tau f(t_0, y_0).$$

Застосовуючи такий підхід для $t_n \le t \le t_{n+1}$, отримаємо рекурентну формулу

$$y_{n+1} = y_n + \tau f(t_n, y_n), n = 0,1,\dots, y_0 = u_0, \qquad (2)$$

Це формула методу Ейлера. Крок інтегрування може змінюватися:

$$\tau = \tau_n = t_{n+1} - t_n.$$

Геометрична інтерпретація методу Ейлера представлена на рис.1. Його друга назва – метод ламаних.

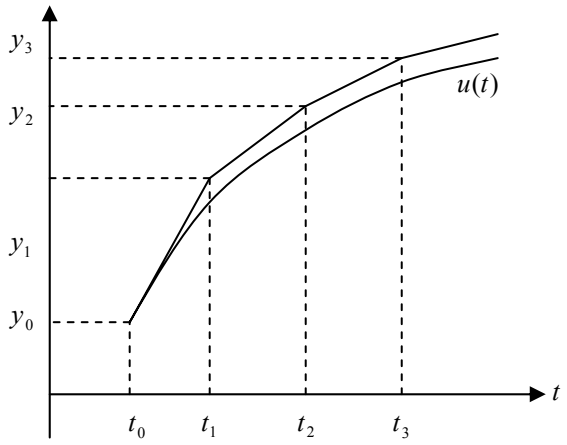


Рис. 1

Цей метод відноситься до однокрокових, тобто розв'язок на наступному кроці обчислюється тільки по одному значенню на попередньому кроці. Загальна формула однокрокових методів

$$y_{n+1} = y_n + \tilde{\Phi}(t_n, y_n, f) \qquad (3)$$

Для методу Ейлера $\tilde{\Phi} = \tau f(t_n, y_n)$.

Величина $R(\tau)=y(t_{n+1})-u(t_{n+1})$, де y_{n+1} обчислюється за формулою (3), причому $y_n \equiv u(t_n)$ називається *похибка методу* (3) *на одному кроці*.

Загальна похибка на $(n+1)$-му кроці складається з похибок на всіх попередніх кроках.

Величина $z_{n+1}=y_{n+1}-u(t_{n+1})$, де y_{n+1} і всі попередні y_k, $k=1,2,...$ також наближені, називається *похибкою методу* (3).

Якщо виразити $y_{n+1}=u(t_{n+1})+z_{n+1}$ і підставити в (3), то

$$u(t_{n+1})+z_{n+1}=u(t_n)+z_n+\tau\widetilde{\Phi}(t_n,u_n+z_n).$$

Тоді можна записати рівняння для z_n:

$$z_{n+1}=z_n+\tau\left[-\frac{u_{n+1}-u_n}{\tau}+\widetilde{\Phi}(t_n,u_n)\right]+\tau\left[\widetilde{\Phi}(t_n,u_n+z_n)-\widetilde{\Phi}(t_n,u_n)\right].$$

Величина $\psi_n=-\frac{u_{n+1}-u_n}{\tau}+\widetilde{\Phi}(t_n,u_n)$ називається *похибкою апроксимації методу* (3).

Для методу Ейлера

$$\psi_n=-\frac{u_{n+1}-u_n}{\tau}+f(t_n,u_n)=-\frac{1}{\tau}\left[u_n+\tau\frac{du}{dt}(t_n)+O(\tau^2)-u_n\right]+f(t_n,u_n)=$$

$$=-\frac{du}{dt}(t_n)+f(t_n,u_n)+O(\tau)=O(\tau)\xrightarrow[\tau\to 0]{}0.$$

Похибка апроксимації це нев'язка, коли замість y в різницеве рівняння (3) підставляємо точний розв'язок задачі Коші u.

Метод (3) має *похибку на одному кроці степеня* m, якщо $|R_n(\tau)|=O(\tau^{m+1})$.

Кажуть, що чисельний метод має *похибку апроксимації степеня* p, якщо $|\psi_n|=O(\tau^p)$.

Для методу Ейлера похибка на одному кроці

$$R_n(\tau)=y_{n+1}-u(t_{n+1})=u(t_n)+\tau f(t_n,u(t_n))-u(t_n+\tau)=u(t_n)+\tau f(t_n,u(t_n))-$$
$$-u(t_n)-\tau u'(t_n)+O(\tau^2)=O(\tau^2).$$

Отже $m=1$. Маємо зв'язок похибок: $\psi(\tau)=\frac{1}{\tau}R(\tau)$ і тому $p=1$.

Кажуть, що метод (3) має *степінь точності* m, якщо

$$\forall n\quad z_n=y_n-u(t_n)=O(\tau^m).$$

<u>Теорема</u> Нехай $f(t,u)\in C^1(\overline{D}_T)$, та $|f_u(t,u)|\le L$, де $\overline{D}_T=\{t_0<t\le T,\ |u(t)|\le K\}$. Тоді метод Ейлера (3) має степінь точності $m=1$ і для нього має місце оцінка

$$|z_n|\le M\max_j|\psi_j|=O(\tau),$$

де $M=M(T,L)$.

▹ Для z маємо задачу

$$z_{n+1}=z_n+\tau\psi_n+\tau\left[f(t_n,u_n+z_n)-f(t_n,u_n)\right].$$

Оцінимо z_{n+1}:

$$|z_{n+1}| \le |z_n| + +\tau|\psi_n| + \tau L|z_n| \le (1+\tau L)|z_n| + \tau|\psi_n| \le$$

$$\dots \le (1+\tau L)^n |z_0| + \sum_{j=0}^{n} \tau(1+\tau L)^{n-j}|\psi_j| \le$$

$$\le (1+\tau L)^n \max_j|\psi_j| \cdot \sum_{j=0}^{n}\tau \le Te^{L\tau}\max_j|\psi_j| \le M\max_j|\psi_j| = O(\tau),$$

так як $(1+\tau L)^n = (1+\tau L)^{\frac{1}{\tau}n\tau} = (1+\tau L)^{\frac{1}{\tau}t_n} \le (1+\tau L)^{\frac{1}{\tau}T} \le e^{LT}$. Позначимо $Te^{LT} = M$ і отримуємо бажану оцінку. ◁

Таким чином порядок точності методу Ейлера $m = 1$.

Метод Ейлера можна вивести із таких міркувань. Інтегруємо (1) по t: $t_n < t < t_{n+1}$

$$u(t_{n+1}) = u(t_n) + \int_{t_n}^{t_{n+1}} f(t,u(t))dt. \qquad (4)$$

Застосуємо формулу лівих прямокутників для інтегралу

$$\int_{t_n}^{t_{n+1}} f(t,u(t))dt \approx \tau f(t_n, u(t_n))$$

і підставимо в (4). Отримаємо формулу для явного методу Ейлера

$$y_{n+1} = y_n + \tau f(t_n, y_n).$$

Застосуємо формулу правих прямокутників для інтегрування

$$\int_{t_n}^{t_{n+1}} f(t,u(t))dt \approx \tau f(t_{n+1}, u(t_{n+1}))$$

і отримаємо з (4) формулу для неявного методу Ейлера

$$y_{n+1} = y_n + \tau f(t_{n+1}, y_{n+1}). \qquad (5)$$

Ці формули 1-го степеня точності по кроку τ. Заміняючи інтеграл за квадратурною формулою трапеції

$$\int_{t_n}^{t_{n+1}} f(t,u(t))dt \approx \frac{\tau}{2}[f(t_n,u(t_n)) + f(t_{n+1},u(t_{n+1}))],$$

ми отримаємо формулу *методу трапеції* інтегрування задачі Коші

$$y_{n+1} = y_n + \frac{\tau}{2}[f(t_n,u(t_n)) + f(t_{n+1},u(t_{n+1}))]. \qquad (6)$$

Це неявний метод.

Обчислимо похибку апроксимації цього методу

$$\psi_n = -\frac{u_{n+1}-u_n}{\tau} + \frac{1}{2}[f(t_n,u_n) + f(t_{n+1},u_{n+1})] =$$

$$= -\frac{1}{\tau}[u_{n+\frac{1}{2}} + \frac{\tau}{2}\dot{u}_{n+\frac{1}{2}} + \frac{1}{2}\left(\frac{\tau}{2}\right)^2 \ddot{u}_{n+\frac{1}{2}} + O(\tau^3) -$$

$$- u_{n+\frac{1}{2}} + \frac{\tau}{2}\dot{u}_{n+\frac{1}{2}} - \frac{1}{2}\left(\frac{\tau}{2}\right)^2 \ddot{u}_{n+\frac{1}{2}} + O(\tau^3)] +$$

$$+\frac{1}{2}\left[f_{n+\frac{1}{2}}+\frac{\tau}{2}\dot{f}_{n+\frac{1}{2}}+O\left(\tau^2\right)+f_{n+\frac{1}{2}}-\frac{\tau}{2}\dot{f}_{n+\frac{1}{2}}+O\left(\tau^2\right)\right]=$$
$$=-\dot{u}_{n+\frac{1}{2}}+O\left(\tau^2\right)+f_{n+\frac{1}{2}}=O\left(\tau^2\right).$$

Таким чином метод трапецій має другий порядок апроксимації.

Задача 36 Показати, що похибка на одному кроці методу трапецій є величина порядку $O\left(\tau^3\right)$.

Таким чином ми отримали більш точний метод. Недолік методу трапецій – неявність (треба розв'язувати нелінійне рівняння). Для розв'язання проблеми неявності застосуємо метод предиктор-корректор (предиктор - попередник, коректор - уточнювач).

Обчислимо попередньо за явним методом Ейлера

$$\bar{y}_{n+1}=y_n+\tau f\left(t_n,y_n\right). \tag{7}$$

Уточнимо за методом трапецій

$$y_{n+1}=y_n+\frac{\tau}{2}\left[f\left(t_n,y_n\right)+f\left(t_{n+1},\bar{y}_{n+1}\right)\right]. \tag{8}$$

Формули (7) та (8) утворюють *метод Ейлера-Коші*.

Оцінимо похибку на одному кроці:

$$R(\tau)=y_1-u(t_0+\tau)=u_0+\frac{\tau}{2}\left[f(t_0,y_0)+f(t_0+\tau,u_0+\tau f(t_0,y_0)\right]-u(t_0+\tau),$$

Звідси $R(0)=0$. Далі

$$R'(\tau)=\frac{1}{2}\left[f(t_0,y_0)+f\left(t_0+\tau,u_0+\tau f_0\right)\right]-\frac{du}{dt}\left(t_0+\tau\right)+\frac{\tau}{2}f_t\left(t_0+\tau,u_0+\tau f\left(t_0,u_0\right)\right)+$$
$$+\frac{\tau}{2}f_u\left(t_0+\tau,u_0+\tau f\left(t_0,u_0\right)\right)\cdot f(t_0,u_0),\ R'(0)=\frac{1}{2}(f_0+f_0)-\frac{du}{dt}(t_0)=0.$$

Наступна похідна

$$R''(\tau)=f_t\left(t_0+\tau,u_0+\tau f_0\right)+f_u\left(t_0+\tau,u_0+\tau f_0\right)\cdot f_0+\tau(...)-u''\left(t_0\right),$$
$$R''(0)=f_{t,0}+f_{u,0}\cdot f_0-\frac{d^2u}{dt^2}\left(t_0\right)=0,$$

так як $\frac{d^2u}{dt^2}=\frac{d}{dt}f\left(t,u\right)=f_t(...)+f_u(...)f_0$.

І, нарешті, $R'''(0)=\frac{3}{2}\left(f_{tt,0}+2f_{tu,0}+f_{uu}\cdot f_0^2\right)-u'''(t_0)\neq 0$. Таким чином похибка на одному кроці має порядок $p=2$.

Ще один метод цього типу:

$$y_{n+\frac{1}{2}}=y_n+\frac{\tau}{2}f\left(t_n,y_n\right), \tag{9}$$

$$y_{n+1}=y_n+\tau f\left(t_{n+\frac{1}{2}},y_{n+\frac{1}{2}}\right). \tag{10}$$

Це *модифікований метод Ейлера*. На етапі коректор (11) використовуємо формулу середніх прямокутників, а предиктор – це метод Ейлера на інтервалі $(t_n, t_{n+\frac{1}{2}})$.

Задача 37 Показати, що для модифікованого методу Ейлера $R(\tau) = O(\tau^3)$, тобто $m = 2$.

10.3. Методи типу Рунге-Кутта [СГ, 218-221], [ЛМС, 256-261]

Методи типу Ейлера мають низьку точність ($m = 1,2$). Рунге запропонував метод третього, а Кутта розвинув його ідею та отримав методи четвертого порядку точності.

Розглянемо задачу Коші

$$\frac{du}{dt} = f(t,u),\ u(t_0) = u_0. \tag{1}$$

Явний m- *етапний (стадійний) метод Рунге-Кутта* полягає в наступному. Нехай розв'язок $y_n = y(t_n)$ вже відомий. Задаються числові коефіцієнти α_i, β_{ij}, $i = \overline{2,m}$, $j = \overline{1,i-1}$ та $p_i, i = \overline{1,m}$ і послідовно обчислюються за формулами:

$$k_1 = \tau f(t_n, y_n),$$
$$k_2 = \tau f(t_n + \alpha_2\tau, y_n + \beta_{21}\tau k_1),$$
$$k_3 = \tau f(t_n + \alpha_3\tau, y_n + \beta_{31}\tau k_1 + \beta_{32}\tau k_2),\ldots,$$
$$k_m = \tau f(t_n + \alpha_m\tau,\ y_m + \beta_{m1}\tau k_1 + \beta_{m2}\tau k_2 + \ldots + \beta_{m,m-1}\tau k_{m-1}).$$

Потім з формули

$$y_{n+1} = y_n + \sum_{i=1}^{m} p_i k_i \tag{2}$$

знаходимо нове значення $y_{n+1} = y(t_{n+1}) \approx u(t_{n+1})$.

Коефіцієнти α_i, β_{ij}, p_i вибираємо з міркувань точності. Наприклад, для того, щоб рівняння (2) апроксимувало рівняння (1), необхідно, щоб $\sum_{i=1}^{m} p_i = 1$.

Інформація про m- стадійний метод записується в *таблиці Батчера*:

$$\begin{array}{c|ccc} \alpha_1 = 0 & & & \\ \alpha_2 & \beta_{21} & & \\ \vdots & \vdots & & \\ \alpha_m & \beta_{m1} & \cdots & \beta_{m\,m-1} \\ \hline & p_1 & \ldots. & p_{m-1} \quad p_m \end{array}$$

Похибка на одному кроці

$$R(\tau) = y_1 - u(t_0 + \tau) = u_0 + \sum_{i=1}^{m} p_i k_i(\tau) - u(t_0 + \tau) =$$

$$= R(0) + R'(0) + ... + \frac{\tau^p}{p!} R^{(p)}(0) + O(\tau^{p+1})$$

Для того, щоб $R(\tau) = O(\tau^{p+1})$, тобто метод мав p-й степінь точності необхідно і достатньо, щоб $R(0) = R'(0) = ... = R^{(p)}(0) = 0$. Загального розв'язку цієї нелінійної системи немає, тому розглянемо частинні випадки.

1. Методи першого порядку $m = 1$. Невідоме p_1.

$$k_1(\tau) = \tau f(t_0, y_0),\ R(\tau) = u_0 + p_1 k_1(\tau) - u(t_0 + \tau),\ R(0) = u_0 - u(t_0) = 0,$$

$$R'(\tau) = p_1 k'(\tau) - \dot{u}(t_0 + \tau) = p_1 f(t_0, y_0) - \dot{u}(t_0 + \tau),$$

$R'(0) = (p_1 - 1) f_0 = 0 \Rightarrow p_1 = 1$.

Ясно, що $R''(0) \neq 0$. Тому $R(\tau) = O(\tau^2)$. Отримуємо *явний метод Ейлера*

$$y_{n+1} = y_n + \tau f(t_n, y_n).$$

2. Методи другого порядку $m = 2$. Тут отримуємо сімейство методів. Невідомі $p_1, p_2, \alpha_2, \beta_{21}$. Вони вибираються з умови: $R(\tau) = O(\tau^3)$, $p = 2$. Маємо

$$p_1 + p_2 = 1,\ 2\alpha_2 p_2 = 1,\ 2\beta_{21} p_2 = 1$$

і один параметр залишається вільним. Далі один параметр p_1 фіксуємо і отримуємо конкретний метод.

а) $p_1 = 0,\ p_2 = 1,\ \alpha_2 = \beta_{21} = \frac{1}{2}$ – *модифікований метод Ейлера*

$$y_{n+1} = y_n + \tau f\left(t_{n+\frac{1}{2}}, y_n + \frac{\tau}{2} f_n\right).$$

б) $p_1 = \frac{1}{2},\ p_2 = \frac{1}{2},\ \alpha_2 = \beta_{21} = 1$ – *метод Ейлера-Коші*

$$y_{n+1} = y_n + \frac{\tau}{2}[f(t_n, y_n) + f(t_{n+1}, y_n + \tau f_n)].$$

в) $p_1 = \frac{1}{3}, p_2 = \frac{2}{3}, \alpha_2 = \beta_{21} = \frac{3}{4}$ - ще один метод другого порядку точності:

$$y_{n+1} = y_n + \frac{\tau}{3}[f(t_n, y_n) + 2f(\left(t_{n+\frac{3}{4}}, y_n + \frac{3}{4} f_n\right)].$$

3. Методи третього порядку $m = 3$. $R(\tau) = O(\tau^4)$. Тому $p = 3$.

Запишемо результат вибору параметрів у вигляді таблиці Батчера (один з частинних випадків):

0			
$\frac{1}{3}$	$\frac{1}{3}$		
$\frac{2}{3}$	0	$\frac{2}{3}$	
	$\frac{1}{4}$	0	$\frac{3}{4}$

Задача 38 Довести. що метод, якому відповідає ця таблиця Батчера, має третій степінь точності.

4. Методи четвертого порядку $m=4$, $R(\tau)=O(\tau^5)$, $p=4$. Найбільш поширені методи:

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & \frac{1}{2} & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \qquad \begin{array}{c|cccc} 0 & & & & \\ \frac{1}{4} & \frac{1}{4} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 1 & -2 & 2 & \\ \hline & \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6} \end{array}$$

Розглянемо як пов'язаний порядок методу p і кількість стадій m:

кількість стадій m	1	2	3	4	5	6
точність методу p	1	2	3	4	4	5

Ясно чому методи Рунге-Кутта при $m>5$ рідко застосовуються: збільшення кількості стадій (а значить збільшення об'єму обчислювальної роботи) не дає виграшу в точності.

Теорема Нехай m- стадійний метод Рунге-Кутта має p-й степінь точності на кроці, а $f(t,u)$ задовольняє умову Ліпшиця. Тоді метод має p-й степінь точності і для похибки методів Рунге-Кутта має місце така оцінка:

$$|z_n| = |y_n - u(t_n)| \le T \cdot e^{\alpha_0 T} \cdot \max_j |\psi_j| = O(\tau^p),$$

де $\psi_j = \frac{1}{\tau} R_j(\tau)$- похибка апроксимації метода, $\alpha_0 = pLm(1+BL\tau_0)^{m-1}$, L – стала Ліпшиця, $B = \max_{i,j} |f_{ij}|, T = \max_n |t_n|$.

10.4. Методи з контролем точності на кроці [БЖК, 365 - 367].

Часто в ході розрахунків необхідно змінювати крок інтегрування, контролюючи величину похибки методу на кроці. При практичній оцінці цієї величини можна. наприклад, поступати так.

Перший підхід використовує принцип Рунге. Головний член похибки на одному кроці (t_n, t_{n+1}) інтегрування має вигляд

$$\frac{\varphi^{(p+1)}(0)\tau^{p+1}}{(p+1)!}.$$

В результаті двох кроків інтегрування однокроковим методом, наприклад, методом Рунге – Кутта p-го степеня точності, буде отримано наближення $y^{\tau}(t_n+2\tau)$ до значення $u(t_n+2\tau)$ таке, що

$$y^{\tau}(t_n+2\tau) - u(t_n+2\tau) \approx 2\frac{\varphi^{(p+1)}(0)\tau^{p+1}}{(p+1)!} = 2\overset{0}{R}(\tau).$$

Якщо тепер застосувати метод Рунге – Кутта p-го степеня з одним кроком довжини 2τ на інтервалі (t_n, t_{n+2}), то отримаємо наближене значення $y^{2\tau}(t_n + 2\tau)$, для якого

$$y^{2\tau}(t_n + 2\tau) - u(t_n + 2\tau) \approx \frac{\varphi^{(p+1)}(0)(2\tau)^{p+1}}{(p+1)!} = \overset{0}{R}(2\tau) = 2^{p+1}\overset{0}{R}(\tau).$$

З цих співвідношень випливає представлення головного члена похибки на кроці

$$y^{\tau}(t_n + 2\tau) - u(t_n + 2\tau) \approx 2\overset{\circ}{R}(\tau) = \frac{y_{n+1}^{2\tau} - y_{n+1}^{\tau}}{2^p - 1}.$$

При необхідності можна уточнити отримане наближене значення, додавши до нього величину головного члена похибки, тобто покласти

$$u(t_n + 2\tau) \approx y^{\tau} + \frac{y^{\tau} - y^{2\tau}}{2^p - 1}.$$

Позначимо $g(\tau) = \frac{y_{n+1}^{2\tau} - y_{n+1}^{\tau}}{2^p - 1}$. Якщо $|g(\tau)| \le \varepsilon$, де ε – деяка задана мала величина (похибка на одному кроці), то τ – успішний крок і $y_{n+1}^{\tau} \approx u(t_n + 2\tau)$. Якщо $|g(\tau)| > \varepsilon$, то зменшуємо крок $\tau := \frac{\tau}{2}$. Треба ще передбачити умову збільшення кроку. Задаємо деяке $\delta << \varepsilon$ і якщо $|g(\tau)| \le \delta$, то $y_{n+1}^{\tau} \approx u(t_n + 2\tau)$ і далі беремо $\tau := 2\tau$. Параметр δ вибирають, наприклад, так: $\delta = \alpha \cdot \varepsilon \cdot 2^{-p}$, $0 < \alpha < 1$, де p – порядок точності методу.

Інший підхід вибору кроку інтегрування заключається в використанні методів різного степеня точності. Отже, якщо в нас є два методи степеня точності на кроці p та $p+1$:

$$y_{n+1}^{(1)} - u(t_n) = O(\tau^{p+1}),$$
$$y_{n+1}^{(2)} - u(t_n) = O(\tau^{p+2}),$$

то головний член похибки першого методу

$$g(\tau) = y_{n+1}^{(1)} - y_{n+1}^{(2)} = O(\tau^{p+1}).$$

Далі з головним членом похибки $g(\tau)$ методу степені p оперуємо так як і в першому підході.

Бажано мати можливість здійснювати крок інтегрування і оцінювати похибку при меншій кількості обчислення значень правих частин. Виграш досягається, якщо використовують методи, які називаються *вкладеними*. Таблиця Батчера для них має вигляд:

$$\begin{array}{c|ccccc} \alpha_2 & \beta_{21} & & & & \\ \vdots & \vdots & & & & \\ \alpha_m & \beta_{m1} & \cdots & \beta_{m\,m-1} & & \\ \hline & p_1 & & & p_{m-1} & p_m \\ & \overline{p}_1 & & & \overline{p}_{m-1} & \overline{p}_m \end{array}$$

Метод з параметрами $p_1,...,p_{m-1},p_m$ має порядок точності p, а з параметрами $\overline{p}_1,...,\overline{p}_{m-1},\overline{p}_m$ - $p+1$. Коефіцієнти α_i, β_{ij} у обох методів однакові.

Найпростіший приклад вкладених методів для $m=2$ має таблицю Батчера:

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & 1 & 0 \\ & 0 & 1 \end{array}$$

Перший метод, якому відповідають коефіцієнти $p_1=1,\ p_2=0$, це метод Ейлера, $p=1$. Другий - $p_1=0,\ p_2=1$ - метод Ейлера – Коші, $p=2$.

Іншим прикладом може служити сукупність формул шестістадійного $m=6$ методу Рунге – Кутта – Фельберга:

$$k_1=\tau f(t_n,y_n),\ k_2=\tau f\left(t_n+\frac{\tau}{2},\ y_n+\frac{k_1}{2}\right),\ k_3=\tau f\left(t_n+\frac{\tau}{2},\ y_n+\frac{1}{4}(k_1+k_2)\right),$$

$$k_4=\tau f(t_n+\tau,y_n-k_2+2k_3),\ k_5=\tau f\left(t_n+\frac{2\tau}{3},y_n+\frac{1}{27}(7k_1+10k_2+k_4)\right),$$

$$k_6=\tau f\left(t_n+\frac{\tau}{5},y_n+\frac{1}{625}(28k_1-125k_2+546k_3+54k_4-378k_5)\right),$$

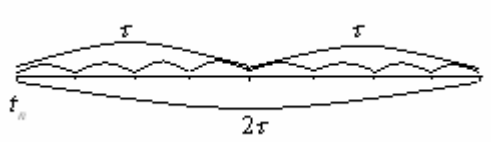
$$\Delta y_n=\frac{1}{6}(k_1+4k_3+k_4),$$

з головним членом похибки

$$y_n(t_n+\tau)-u(t_n+\tau)=g(\tau)+O(\tau^5),$$

$$g(\tau)=-\frac{1}{366}(42k_1+224k_3+21k_4-162k_5-125k_6).$$

Методи Рунге – Кутта – Фельберга мають четвертий та п’ятий степінь точності. Порівняємо кількість обчислень правих частин в методах Рунге-Кутта та Рунге – Кутта – Фельберга ($p=4$ для обох). Згідно схеми кроків по змінній t та по стадіях $i=\overline{1,4}$ для методу Рунге – Кутта необхідно обчислити для оцінки похибки 11 значень функції, а для методу Рунге – Кутта – Фельберга 6 значень функції.



Основний недолік методів Рунге-Кутта: щоб отримати досить високий степінь точності потрібно багато раз обчислювати значення функції.

Достойністю методів Рунге – Кутта є можливість зміни кроку інтегрування $\tau = \tau_n$ і за рахунок цього задовольняти умову точності на кроці.

10.5. Багатокрокові методи розв'язання задачі Коші. Методи Адамса [СГ, 230-231], [БЖК, 372-375]

Недолік методів Рунге-Кутта: велика кількість обчислень значень функцій на одному кроці (особливо, чутливо це для систем). Висока точність в цих методах досягається за рахунок обчислень для m - стадій коефіцієнтів $k_i(\tau) = \tau f(\xi_i, \eta_i)$ в проміжних точках між t_n та $t_n + \tau$. А чи не можна для цього використати попередні значення $f(t_n, y_n), f(t_{n-1}, y_{n-1}),..., f(t_{n-m}, y_{n-m})$?

Для розв'язання задачі Коші

$$\frac{du}{dt} = f(t,u),\ t > 0,\ u(0) = u_0 \qquad (1)$$

введемо сітку

$$\omega_\tau = \{t_n = n\tau,\ n = 0,1,...\}$$

з постійним кроком $\tau > 0$. Позначимо через $y_n = y(t_n),\ f_n = f(t_n, y_n)$ функції, визначені на сітці ω_τ.

Лінійним m - кроковим різницевим методом називається рівняння:

$$\frac{a_0 y_n + a_1 y_{n-1} + ... + a_m y_{n-m}}{\tau} = b_0 f_n + b_1 f_{n-1} + ... + b_m f_{n-m},\ n = m, m+1,... \qquad (2)$$

де a_k, b_k – числові коефіцієнти, які не залежать від n, причому $a_0 \neq 0$.

Рівняння (2) слід розглядати як рекурентне співвідношення, яке виражає нове значення $y_n = y(t_n)$ через знайдені раніше значення $y_{n-1}, y_{n-2},..., y_{n-m}$.

Розрахунок починається з $n = m$, тобто з рівняння

$$\frac{a_0 y_m + a_1 y_{m-1} + ... + a_m y_0}{\tau} = b_0 f_m + b_1 f_{m-1} + ... + b_m f_0.$$

Як бачимо з рівняння, для початку розрахунків необхідно задати m початкових значень. Значення y_0 визначається початковою умовою задачі (1), а саме покладають $y_0 = u_0$. Величини $y_1,..., y_{m-1}$ можна обчислити, наприклад, за методом Рунге-Кутта, або за методом рядів Тейлора. В подальшому будемо вважати, що початкові значення $y_0, y_1,..., y_{m-1}$ задані.

З рівняння (2) видно, що на відміну від методу Рунге-Кутта багатокрокові різницеві методи допускають обчислення правих частин тільки в точках основної сітки ω_τ.

Метод (2) називається *явним*, якщо $b_0 = 0$, і відповідно, шукане значення y_n виражається явно через $y_{n-1},...,y_{n-m}$. Якщо $b_0 \neq 0$, то метод (2) називається *неявним*.

В неявному методі для пошуку y_n потрібно розв'язувати нелінійне рівняння:

$$\frac{a_0}{\tau} y_n - b_0 f(t_n, y_n) = F[y_{n-1}, y_{n-2},..., y_{n-m}],$$

де

$$F[y_{n-1}, y_{n-2},..., y_{n-m}] = \sum_{k=1}^{m} \left(b_k f_{n-k} - \frac{a_0}{\tau} y_{n-k} \right).$$

Як правило це рівняння розв'язують методом Ньютона, обираючи, наприклад, початкове наближення $y_n^{(0)}$ рівним y_{n-1}.

Коефіцієнти рівняння (2) визначені з точністю до множника. Для уникнення цієї неоднозначності, будемо вважати, що виконується умова:

$$\sum_{k=0}^{m} b_k = 1, \tag{3}$$

яка означає, що права частина різницевого рівняння (2) апроксимує праву частину диференціального рівняння (1).

В практичному використанні найбільш поширені методи Адамса, які являють собою частинний випадок багатокрокових методів (2), коли похідна $u'(t)$ апроксимується тільки по двох точках, t_n і t_{n-1}, тобто

$$a_0 = -a_1 = 1,\ a_k = 0,\ k = \overline{2,m}.$$

Таким чином, *методи Адамса* мають вигляд:

$$\frac{y_n - y_{n-1}}{\tau} = \sum_{k=0}^{m} b_k f_{n-k}.$$

У випадку, коли $b_0 = 0$, методи Адамса називають *явними*, А при $b_0 \neq 0$, методи Адамса називають *неявними*.

Розглянемо детальніше процедуру побудови методів Адамса. Інтегруємо рівняння (1) по $t \in [t_{n-1}, t_n]$:

$$\frac{u(t_{n-1}) - u(t_n)}{\tau} = \frac{1}{\tau} \int_{t_{n-1}}^{t_n} f(t, u(t)) dt. \tag{4}$$

Замінимо $f(t, u(t))$ на інтерполяційний поліном. Виберемо вузлами інтерполювання точки $t_n, t_{n-1},...t_{n-m}$. Використаємо багаточлен степеня m за формулою Ньютона по рівновіддалених вузлах, тобто

$$f(t, u(t)) \approx L_m(t) =$$

$$= f_n + s\Delta f_{n-1} + ... + \frac{s(s+1)\cdot...\cdot(s+m-1)}{m!} \Delta^m f_{n-m},$$

де $s = \frac{t - t_n}{\tau}$, крок τ сталий. В результаті підстановки в (4) $f(t, u(t)) \approx L_m(t)$ отримуємо метод

$$\frac{y_n - y_{n-1}}{\tau} = f_n + \beta_1 \Delta f_{n-1} + \ldots + \beta_m \Delta^m f_{n-m},$$

де

$$\beta_k = \int_{-1}^{0} \frac{s(s+1)\cdot\ldots\cdot(s+k-1)}{k!} ds, k = \overline{1,m} \quad (4)$$

$$\Delta f_{n-1} = f_n - f_{n-1},\ \Delta^2 f_{n-2} = f_n - 2f_{n-1} + f_{n-2}.$$

Похибка методу на одному кроці

$$R(\tau) = \int_{t_{n-1}}^{t_n} r_m(t)dt = O(\tau^{m+2}),$$

де

$$r_m(t) = f(t) - L_m(t),$$

звідки степінь точності на одному кроці $p = m+1$.

Формула (4) називається *формулою Адамса-Мултона* (неявний метод Адамса, інтерполяційний метод Адамса).

Виберемо вузли інтерполювання точки $t_{n-1},\ldots t_{n-m}$. Отримаємо багаточлен степеня $m-1$:

$$f(t,u(t)) \approx L_{m-1}(t) = f_{n-1} + v\Delta f_{n-2} + \ldots + \frac{v(v+1)\cdot\ldots\cdot(v+m-2)}{(m-1)!}\Delta^{m-1} f_{n-m},$$

де $v = \frac{t - t_{n-1}}{\tau}$. Підставляючи в (4), маємо метод

$$\frac{y_n - y_{n-1}}{\tau} = f_{n-1} + \gamma_1 \Delta f_{n-1} + \ldots + \gamma_{m-1}\Delta^{m-1} f_{n-m}\ , \quad (5)$$

де

$$\gamma_k = \int_0^1 \frac{v(v+1)\cdot\ldots\cdot(v+k-1)}{k!} dv,\ k = \overline{1,m-1}.$$

Формула (5) називається *формулою Адамса-Башфорта* (явний метод Адамса, екстраполяційний метод Адамса). Похибка цього методу на одному кроці $R(\tau) = O(\tau^{m+1})$. Степінь точності на одному кроці $p = m$.

Задача 39 Побудувати явний та неявний двокрокові методи Адамса. Який степінь точності вони мають?

10.6. Метод невизначених коефіцієнтів побудови багатокрокових методів для розв'язку задачі Коші [БЖК, 375-379], [СГ, 230-231]

Розглянемо задачу Коші

$$\frac{du}{dt} = f(t,u),\ t > 0,\ u(t_0) = u_0. \quad (1)$$

і багатокроковий метод

$$\sum_{k=0}^{m} \frac{a_k}{\tau} y_{n-k} = \sum_{k=0}^{m} b_k f_{n-k} \quad (2)$$

Підберемо коефіцієнти a_k, b_k так, щоб досягти найвищої точності методу (2).

Введемо похибку апроксимації $\psi(\tau)$ для формули (2). *Похибкою апроксимації* на розв'язку або нев'язкою різницевого методу (2) називається функція

$$\psi(\tau)=-\sum_{k=0}^{m}\frac{a_k}{\tau}u_{n-k}+\sum_{k=0}^{m}b_k f(t_{n-k},u_{n-k}), \qquad (3)$$

де $u_{n-k}=u(t_n-k\tau)$. Її отримують підстановкою точного розв'язку $u(t)$ задачі (1) в різницеве рівняння (2). Розглянемо питання про порядок похибки апроксимації при $\tau\to 0$ в залежності від вибору коефіцієнтів a_k, b_k, $k=\overline{0,m}$.

Розкладемо функції $u_{n-k}=u(t_n-k\tau)$ в точці $t=t_n$ по формулі Тейлора:

$$u(t_n-\tau k)=\sum_{j=0}^{p}u^{(j)}(t_n)\cdot\frac{(-k\tau)^j}{j!}+O(\tau^{p+1}). \qquad (4)$$

Тут $u^{(j)}$ – j-та похідна. Далі

$$f(t_n-k\tau,u(t_n-k\tau))=\frac{du}{dt}(t_n-k\tau)=\sum_{j=0}^{p-1}u^{(j+1)}(t_n)\frac{(-k\tau)^j}{j!}+O(\tau^p),\ k=\overline{1,m}. \qquad (5)$$

Підставляючи ці вирази в формулу (3), отримаємо:

$$\psi(\tau)=-\sum_{k=0}^{m}\frac{a_k}{\tau}\left(\sum_{j=0}^{p}\frac{(-k\tau)^j u^{(j)}(t_n)}{j!}\right)+\sum_{k=0}^{m}b_k\left(\sum_{j=0}^{p-1}\frac{(-k\tau)^j u^{(j+1)}(t_n)}{j!}\right)+O(\tau^p)=$$

$$=-\sum_{j=0}^{p}\left(\frac{a_k}{\tau}\cdot\sum_{k=0}^{m}\frac{(-k\tau)^j u^{(j)}(t_n)}{j!}\right)+\sum_{j=1}^{p}\left(\sum_{k=0}^{m}b_k\cdot\frac{(-k\tau)^{j-1}u^j(t_n)}{(j-1)!}\right)+O(\tau^p).$$

Після перетворень, приходимо до розкладу:

$$\psi(\tau)=-\left(\sum_{k=0}^{m}\frac{a_k}{\tau}\right)u(t_n)+\sum_{j=1}^{p}\left(\sum_{k=0}^{m}(-k\tau)^{j-1}\left(a_k\cdot\frac{k}{j}+b_k\right)\right)\frac{u^{(j)}(t_n)}{(j-1)!}+O(\tau^p).$$

Звідки видно, що похибка апроксимації має порядок p, якщо виконуються умови:

$$E_0=\sum_{k=0}^{m}a_k=0,\ E_l=\sum_{k=0}^{m}\frac{k^{l-1}}{l!}(ka_k+lb_k)=0,\ l=\overline{1,p} \qquad (6)$$

Разом з умовою нормування $\sum_{k=0}^{m}b_k=1$, рівняння (6) утворюють систему з $p+2$ лінійних алгебраїчних рівнянь відносно $2(m+1)$ невідомих $a_0,a_1,...,a_m,b_0,b_1,...,b_m$.

Умови нормування запишуться у вигляді:

$$\lim_{\tau\to0}\sum_{k=0}^{m}\frac{a_k}{\tau}u_{n-k}=\frac{du}{dt}(t_n),\ \lim_{\tau\to0}\sum_{k=0}^{m}b_k f(t_{n-k},u_{n-k})=f(t_n,u_n).$$

З урахуванням (4), (5) маємо

$$\sum_{k=0}^{m}b_k=1,\ \sum_{k=0}^{m}ka_k=1.$$

З рівняння (6) при $l=0$ отримуємо

$$\sum_{k=0}^{m} k a_k - \sum_{k=0}^{m} b_k = 0.$$

Тому умови нормування додатково до (6) дають тільки рівняння

$$\sum_{k=0}^{m} b_k = 1.$$

Для того, щоб система (5) не була перевизначеною, необхідно вимагати, щоб $p \le 2m$. Ця вимога означає, що порядок апроксимації лінійних m-крокових різницевих методів не може перевищувати $2m$ (неявні методи). Найвищий порядок апроксимації явних методів – $2m-1$.

Для методів Адамса умови p-го порядку апроксимації мають вигляд:

$$\sum_{k=1}^{m} k^{l-1} b_k = 1,\ l = 2,3,...,p,\ b_0 = 1 - \sum_{k=1}^{m} b_k .$$

Звідки бачимо, що найвищий порядок апроксимації неявного m- крокового методу Адамса дорівнює $m+1$, а найвищий порядок апроксимації явного методу Адамса ($b_0 = 0$) дорівнює m.

Задача 40 Побудувати явний та неявний двокрокові методи найвищого степеня точності.

10.7. Питання реалізації багатокрокових методів

1). Перша проблема, яка виникає при застосування багатокрокових методів це вибір додаткових початкових умов. Треба знайти $(m-1)$ додаткове початкове значення $y_1,...,y_{m-1}$. Шляхи вирішення проблеми такі.

а) Можна використовувати *методи Рунге-Кутта* для пошуку цих початкових значень. Але треба, щоб ці методи мали або точність апроксимації p, або точність похибки на кроці p. Недолік такого способу: сама процедура обчислення за методами Рунге-Кутта займає великий об'єм, та й багатокрокові методи об'ємні.

б) Можна використовувати *метод рядів Тейлора.* Знову ж таки потрібно узгоджувати точність p. Функцію розкладають в ряд:

$$\overset{(p-1)}{u}(t) = \sum_{k=0}^{p-1} \frac{u^{(k)}(t_0)}{k!}(t-t_0)^k ,$$

де індекс $p-1$ над u вказує кількість членів ряду. Похибка

$$u - \overset{(p-1)}{u} = O(\tau^p).$$

2). Друга проблема: реалізація неявних методів

$$y_n = \tau \frac{b_0}{a_0} f(t_n, y_n) + \Phi(y_{n-1},...,y_{n-m}) = \varphi(y_n).$$

а) Можна використовувати для знаходження розв'язку нелінійного рівняння метод простої ітерації:

$$y_n^{(k+1)} = \varphi\left(y_n^{(k)}\right),$$

де індекс k означає k-ту ітерацію. Умова збіжності методу простої ітерації $\varphi'(u) \le q < 1$. Тоді маємо:

$$|\varphi'(u)| = \left|\tau \frac{b_0}{a_0} f_u'(t,u)\right| \le q < 1 \Rightarrow \tau < \frac{a_0}{b_0 \cdot L}, \text{ де } |f_u'(t,u)| \le L,$$

L – максимум похідної чи стала Ліпшиця. Тобто для збіжності методу необхідно, щоб крок τ був досить малим.

б) Можна використовувати метод Ньютона. Як відомо, умови збіжності методу Ньютона залежать від вдалого вибору початкових умов. Хороше початкове наближення таке: або $y_n^{(0)} = y_{n-1}$, або $y_n^{(0)}$ обчислюється за явною m-кроковою формулою.

В результаті умови збіжності методу Ньютона менш жорсткі, ніж у методу простої ітерації.

в) можна використовувати *метод предиктор-коректор*. Запишемо формули неявного методу (C – коректор, P – предиктор):

C: $y_n = \tau \frac{b_0}{a_0} f(t_n, y_n) + \Phi(y_{n-1}, ..., y_{n-m})$ – неявний m-кроковий метод, $p = 2m$.

P: $y_n = \overline{\Phi}(y_{n-1}, ..., f_{n-1}, ...)$ – явний m-кроковий метод, $p = 2m - 1$.

Далі виконується така процедура:

P: $\overline{y_n} = \overline{\Phi}(y_{n-1}, ..., f_{n-1}, ...)$;

E: $\overline{f_n} = f(t_n, y_n)$;

C: $y_n = \tau \frac{b_0}{a_0} f(t_n, \overline{y_n}) + \Phi(y_{n-1}, ...)$;

E: $f_n = f(t_n, y_n)$,

де E – підрахування правої частини рівняння (Equation). Схема $PECE$ – це метод *предиктор-коректор*.

Іноді, щоб підвищити точність, використовують схему $PE(CE)^j$. Ця схема аналогічна методу простої ітерації, де j – це максимальна кількість ітерацій.

10.8. Стійкість методів розв'язання задачі Коші [ЛМС,273,281-282], [БЖК,379-382], [СГ,235-238]

Важливими є питання стійкості чисельних методів для задач Коші, тобто неперервна залежність розв'язків від збурень початкових даних, похибок заокруглення, тощо.

Чисельний метод розв'язання задачі Коші

$$\frac{du}{dt} = f(t,u),\ u(t_0) = u_0 \qquad (1)$$

називається *стійкий за початковими даними*, якщо для довільної задачі Коші (тобто $\forall f(t,u)$)

$$\forall\, \varepsilon > 0\ \exists\, \sigma = \sigma(\varepsilon) > 0 \ \text{ при } \ \left|u - \bar{u}_0\right| < \sigma, \text{ маємо } \left|y_n - \overline{y_n}\right| = \varepsilon, \quad \forall n .$$

Чисельний метод називається *нестійким,* якщо $\exists\, \varepsilon > 0, \forall\, \sigma = \sigma(\varepsilon) > 0, \exists\, f(t,u)$, що при $\left|u - \bar{u}_0\right| < \sigma$, маємо $\left|y_n - \overline{y_n}\right| > \varepsilon, \ \ n \geq N(\varepsilon).$

Перше означення складніше перевіряти ніж друге. Перевіркою умови нестійкості можна вибраковувати погані методи на таких класах правих частин:

$$\text{а) } f \equiv 0\,;\ \text{б) } f = \lambda u, \lambda < 0\,;\ \text{в) } f = \lambda u, \operatorname{Re}\lambda < 0\,.$$

Розглянемо кожен випадок.

а) $f \equiv 0$. Якщо застосувати метод Рунге-Кутта, то $y_{n+1} = y_n = \ldots = y_0$, а це означає, що малі збурення y_0 приводять до таких самих збурень в y_{n+1}: $\delta = \varepsilon$.

Розглянемо тепер багатокроковий метод, що приводить до різницевого рівняння n-го порядку

$$\sum_{k=0}^{m} \frac{a_k}{\tau} y_{n-k} = 0 \tag{2}$$

Часткові розв'язки шукаємо у вигляді $y_n = q^n$, $y_{n-k} = q^{n-k}$. Підставимо їх і отримаємо характеристичне рівняння

$$\sum_{k=0}^{m} a_k q^{m-k} = 0 \tag{3}$$

з коренями $q_1, \ldots, q_m$.

Розв'язок рівняння (2) знаходимо у вигляді:

якщо q – дійсні різні, то $y^n = \sum_{k=0}^{m} c_k q_k^n$;

якщо q – дійсні кратні, то $y^n = \ldots + c_j q_j^n + \widetilde{c}_j n q_j^n + \ldots$;

якщо q – комплексні, то $y^n = .. + c_j \rho_j^n \cos(n \operatorname{Im} q_j) + \widetilde{c}_j \rho_j^n \sin(n \operatorname{Im} q_j) + ..,$

де $\rho_j = \operatorname{mod} q_j$, якщо $\exists\, \rho_j > 1 \Rightarrow n \to \infty, y_n \to \infty$, c_k знаходять з початкових умов.

При $\left|q_k\right| > 1$ багатокроковий метод – нестійкий.

Кажуть, що багатокроковий метод розв'язання задач задовольняє *кореневій умові*, якщо всі корені характеристичного рівняння (3) лежать в одиничному крузі, а на колі немає кратних коренів, тобто $\left|q_j\right| \leq 1$, для $q_j = q_m$, $\left|q_j\right| = \left|q_m\right| < 1$.

Щоб перевірити, як працює коренева умова розглянемо задачу. Побудуємо явний двокроковий метод найвищого степеня точності. $m = 2$, $p = 3$:

$$\frac{y_n + 4y_{n-1} - 5y_{n-2}}{6\tau} = \frac{2}{3} f_{n-1} + \frac{1}{3} f_{n-2}.$$

Його характеристичне рівняння

$$q^n + 4q^{n-1} - 5q^{n-2} = 0,$$

$$q^2 + 4q - 5 = 0,\ q_1 = 1,\ q_2 = -5.$$

Явний двокроковий метод найвищого степеня – нестійкий метод оскільки $q_2 = -5$.

Теорема 1 Нехай багатокроковий метод задовольняє кореневій умові та права частина задовольняє умову Ліпшиця, $|f(t,u) - f(t,v)| \le L|u - v|$, тоді має місце оцінка точності

$$|z_n| = |y_n - u(t_n)\|z_n| \le M\left(\max_{k=0,m-1}|z_k| + \max_{k=0,n-1}|\psi_k|\right), \tag{4}$$

де ψ_k – похибка апроксимації.

б) $f = \lambda u$, $\lambda < 0$. В цьому випадку розв'язок задачі Коші (1) $u(t) = u_0 e^{\lambda(t-t_0)} \underset{t\to\infty}{\to} 0$ і він є асимптотично стійким відносно 0. Слід очікувати, що чисельний метод має ту ж властивість. Тобто

$$|y_n| < |y_{n-1}| < \ldots < |y_0| = |u_0| \tag{5}$$

Для явного методу Ейлера

$$y_{n+1} = y_n + \tau f(F_n, y_n) = y_n + \tau\lambda y_n = (1 + \tau\lambda)y_n = qy_n,\ q = 1 + \tau\lambda\ .$$

Тому (5) має місце тільки якщо $|q| < 1$ або $-1 < q < 1$. Права нерівність виконується для довільного τ, а ліва тільки для $\tau < \dfrac{2}{|\lambda|}$.

Метод розв'язання задачі Коші називається *умовно стійким*, якщо він стійкий при $\tau \le \tau_0$. Якщо ж він стійкий для $\forall\tau > 0$, то такий метод називається *абсолютно стійким.*

Таким чином явний метод Ейлера – умовно стійкий. Для неявного методу Ейлера:

$$y_{n+1} = y_n + \tau f(t_{n+1}, y_{n+1}) = y_n + \tau\lambda y_{n+1}.$$

Звідси запишемо $y_{n+1} = \dfrac{1}{1-\tau\lambda} y_n$, $0 < q = \dfrac{1}{1-\tau\lambda} < 1, \lambda < 0$. Таким чином неявний метод Ейлера абсолютно стійкий на тестовому рівнянні.

в) $f = \lambda u$, $\operatorname{Re}\lambda < 0$. Розглянути систему

$$\frac{d\vec{u}}{dt} = A\vec{u},$$

де A – матриця простої структури, тобто $\exists H:\ \Lambda = HAH^{-1} = diag(\lambda_i)$, $\lambda_i = \lambda_i(A)$. Попередня система зводиться до такої:

$$\frac{d\vec{v}}{dt} = \Lambda\vec{v}\text{, або } \frac{dv_i}{dt} = \lambda_i v_i,\ i = \overline{1,n},$$

де $v_i = v_i^0 e^{\lambda_i t} = v_i^0 e^{\operatorname{Re}\lambda_i t}\left(\cos\varphi(t) + j\sin\varphi(t)\right)$, φ – кут для якого $tg\varphi = \dfrac{\operatorname{Im}\lambda_i}{\operatorname{Re}\lambda_i}$.

Дослідимо явний та неявний метод Ейлера на такому тестовому прикладі.

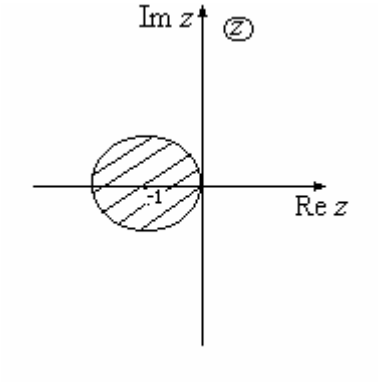


Рис. 13

Явний метод Ейлера: $y_{n+1} = qy_n$, $q = 1 + \tau\lambda = 1 + z$, $z = \tau\lambda = \mathrm{Re}\, z + i\,\mathrm{Im}\, z$. Умова $|q| \le 1$ виконується для внутрішності кола радіуса 1 з центром в точці (-1, 0) (рис. 13):

$$|q_i| = |1 + \tau\lambda_i| < 1, \forall i.$$

Звідси

$$\tau \le \frac{2\,\mathrm{Re}\,\lambda_i}{|\lambda_i|^2} = \tau_0,$$

Таким чином явний метод Ейлера умовно стійкий і для тестового рівняння з $\mathrm{Re}\,\lambda < 0$.

Для неявного методу Ейлера $q = \dfrac{1}{1 - z}$.

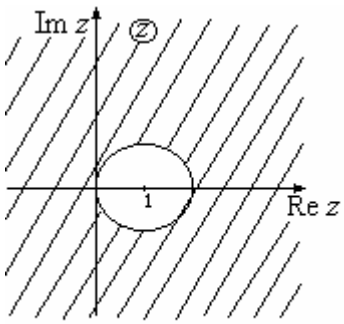


Рис.14

Умова $|q| \le 1$ виконується для зовнішності кола радіуса $R = 1$, з центром в точці $z_0 = (1,0)$ (рис. 14), в той час, як точки $z_i = \tau\lambda_i$ лежать в лівій півплощині z Значить умова стійкості задовольняється для довільних τ. Таким чином неявний метод Ейлера є абсолютно стійкий. Недолік неявного методу Ейлера - низька точність: $p = 1$.

Кажуть, що метод розв'язання задачі Коші є *A-стійким*, якщо область його стійкості для тестового рівняння $f = \lambda u$, $\mathrm{Re}\,\lambda < 0$ включає всю півплощину $\mathrm{Re}\, z < 0$, $z = \tau\lambda$.

Задача 41 Показати , що метод трапеції розв'язання задач Коші є A-стійким.

Теорема 2 (Далквіст). Явний лінійний багатокроковий метод не може бути A-стійким. Порядок неявного A- стійкого методу не може бути вищим за 2. Найбільш точний серед цих методів є метод трапецій.

Задача Коші (1) називається *жорсткою* для $t \in [t_0, T]$, якщо $\operatorname{Re}\lambda_i < 0$ та

$$S(t) = \frac{\max\limits_i |\operatorname{Re}\lambda_i(t)|}{\min\limits_i |\operatorname{Re}\lambda_i(t)|} >> 1,$$

де $\lambda_i = \lambda_i(t)$ власні значення матриці Якобі правої частини рівняння

$$A_n = \left(\frac{\partial f_i}{\partial u_j}(t_n) \right)_{i,j=1}^{m}.$$

Характерна особливість розв'язків жорсткої задачі Коші є наявність компонент $e^{\lambda_i t}$, які повільно змінюються (коли $|\operatorname{Re}\lambda_i| \approx \min\limits_i |\operatorname{Re}\lambda_i|$) та швидко спадають (коли $|\operatorname{Re}\lambda_i| \approx \max\limits_i |\operatorname{Re}\lambda_i|$). Стійкість явних методів (Рунге-Кутта, явного методу Ейлера) визначається компонентами, які швидко змінюються. В той час як в точному розв'язку вони швидко прямують до нуля і дають малий вклад. Сам розв'язок змінюється повільно разом з повільними компонентами. Але умова стійкості явних методів на всьому проміжку $[t_0, T]$ повинна бути орієнтована на крок $\tau < \dfrac{2}{\max\limits_i |\lambda_i|}$. Це дуже жорстка умова. Тому такі системи називаються жорсткими.

Застосування A- стійких методів знімає проблему стійкості: крок τ можна міняти тільки з умови виконання малої похибки на одному кроці. Але точність A- стійких методів не більше 2. Що робити? Послаблюють означення A-стійкості.

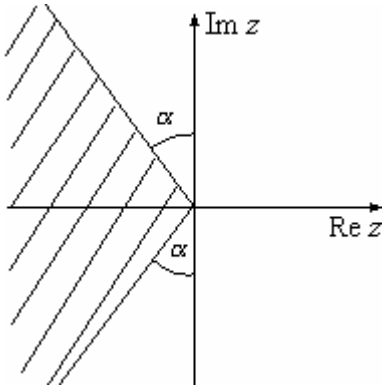


Рис. 15

Метод розв'язання задачі Коші називають $A(\alpha)$- *стійким* $0 \le \alpha \le \dfrac{\pi}{2}$, якщо область стійкості методу для тестового рівняння з $f = \lambda u$, $\operatorname{Re}\lambda < 0$, включає частину півплощини $\operatorname{Re} z$ $(z = \tau\lambda)$, що лежить в середині між променями, що утворюють з віссю $\operatorname{Im} z$ кути α (рис. 15). Зауважимо, що A - стійкість = $A(0)$ - стійкість !

Найбільш поширеними для розв'язування жорстких систем є методи Куртіса-Хершфільта:

$$\sum_{k=0}^{m}\frac{a_k}{\tau}y_{n-k}=f(t_n,y_n).$$

Області їх стійкості для різних m приведена на рис. 16:

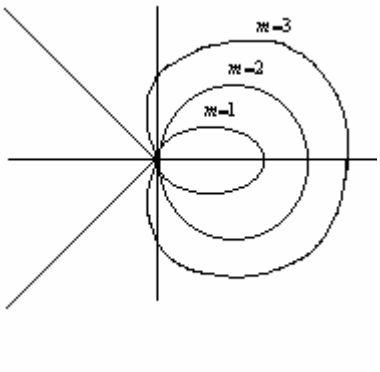


Рис. 16

Це неявні m- крокові методи з $b_0=1,\ b_k=0\quad k=\overline{1,m}$ (інші назви: методи диференціювання назад, чисто неявні методи).

Задача 42 Методом невизначених коефіцієнтів побудувати метод диференціювання назад для $m=3$.

Для жорстких задач треба використовувати реалізацію неявних методів за допомогою ітераційного методу Ньютона з вибором початкового наближення за явною схемою.

Для жорстких задач також застосовують неявні методи Рунге-Кутта. Але для їх реалізації треба розв'язувати систему з $m\times N$ нелінійних рівнянь, m – кількість стадій, N – розмірність системи. Таблиця Батчера цих методів має вигляд

$$\left(\begin{array}{c|ccc}\alpha_1 & \beta_{11} & \cdots & \beta_{1m}\\ \vdots & \vdots & \ddots & \vdots\\ \alpha_m & \beta_{m1} & \cdots & \beta_{mm}\\ \hline & p_1 & \cdots & p_m\end{array}\right).$$

11. Методи розв'язання крайових задач для звичайних диференційних рівнянь

Почнемо з постановки крайових задач.

1). Нелінійна двоточкова крайова задача

$$\frac{d\vec{U}}{dx}=\vec{F}\left(x,\vec{U}\right)\ a<x<b \tag{1}$$

$$\vec{\varphi}\left(\vec{U}(a),\vec{U}(b)\right)=\vec{d} \tag{2}$$

Тут

$\vec{U}=(u_1,...,u_m)^T$, $u_k=u_k(x)$, $\vec{F}(f_1,...,f_m)^T$, $f_k=f_k\left(x,\vec{U}\right)$, $\vec{\varphi}=(\varphi_1,...,\varphi_m)^T$, $\varphi_k=\varphi_k\left(\vec{U}(a),\vec{U}(b)\right)$, $\vec{d}=(d_1,...,d_m)$, d_k - числа.

2). Лінійна двоточкова крайова задача

$$\frac{d\vec{U}}{dx} = A(x)\vec{U}(x) + \vec{F}(x) \qquad (3)$$

$$B_1\vec{U}(a) + B_2\vec{U}(b) = \vec{d} \qquad (4)$$

$A(x) = \left(a_{ij}(x)\right)_{i,j=1}^{m}$, $\vec{F} = \left(f_1,...,f_m\right)$, $f_k = f_k(x)$, B_1, B_2 - числові матриці $m \times m$; $\vec{d}$ - вектор.

Крайові умови (2) і (4) називаються *нероздільними*. Часто зустрічаються *розділені крайові умови* (наприклад, для лінійної задачі):

$$C_1\vec{U}(a) = \vec{d}_1,\ C_2\vec{U}(b) = \vec{d}_2, \qquad (4')$$

де $C_1 - (m-k)\times m$ - матриця, $C_2 - k \times m$ - матриця; $rangC_1 = m-k$, $rangC_2 = k$; $\vec{d}_1 - (m-k)$ - вектор, $\vec{d}_2 - k$ - вектор.

До (3),(4/) зводиться крайова задача для рівнянь вищих порядків. Нехай задана крайова задача:

$$\begin{cases} u^{(m)}(x) = p_1(x)u^{(m-1)}(x) + ... + p_m(x)u(x) + f(x) \\ \alpha_{i1}u^{(m-1)}(a) + ... + \alpha_{im}u(a) = \mu_i, i = \overline{1, m-k} \\ \beta_{i1}u^{(m-1)}(b) + ... + \beta_{im}(b) = \nu_i, i = \overline{1, k} \end{cases}.$$

Вона зводиться до задачі (3), (4) з

$$A(x) = \begin{pmatrix} 0 & 1 & 0 & .. & 0 \\ 0 & 0 & 1 & .. & 0 \\ .. & .. & .. & .. & .. \\ 0 & 0 & 0 & .. & 1 \\ p_m & p_{m-1} & p_{m-2} & .. & p_1 \end{pmatrix},$$

$$C_1 = \left(\alpha_{ij}\right)_{i=\overline{1,m-k}}^{j=\overline{i,m}},\ C_2 = \left(\beta_{ij}\right)_{i=\overline{1,k}}^{j=\overline{1,m}},$$

$$\vec{d}_1 = (\mu_1,...,\mu_{m-k})^T,\ \vec{d}_2 = (\nu_1,...,\nu_k)^T.$$

Вважаємо, що всі задачі мають єдині розв'язки. Розглянемо методи розв'язання цих задач.

11.1. Метод стрільби [ЛМС, 288 – 290],[БЖК, 434 – 435]

Розглянемо лінійну крайову задачу з нерозділеними крайовими умовами:

$$\frac{d\vec{U}}{dx} = A(x)\vec{U}(x) + \vec{F}(x) \qquad (1)$$

$$B_1\vec{U}(a) + B_2\vec{U}(b) = \vec{d} \qquad (2)$$

Метод стрільби зводить крайову задачу до послідовності задач Коші. Для цього розв'яжемо (m+1) задачу Коші:

$$\frac{d\vec{Y}_0}{dx} = A(x)\vec{Y}_0 + \vec{F}(x),\ \vec{Y}_0(a) = 0 \qquad (3)$$

$$\frac{d\vec{Y}_i}{dx} = A(x)\vec{Y}_i(x), \quad \vec{Y}_i(a) = \vec{\delta}_i, \quad \vec{\delta}_i = (\delta_{ij})_{j=1}^{m}, \; i = \overline{1,m} \qquad (4)$$

Матриця $Y(x) = \left(\vec{Y}_i(x)\right)_{i=\overline{1,m}}$ називається *фундаментальною матрицею* однорідної системи (1). Розв'язок (1) шукаємо у вигляді:

$$\vec{Y}(x) = \vec{Y}_0(x) + \sum_{i=1}^{m} c_i \vec{Y}_i(x) \qquad (5)$$

Він задовольняє (1) $\forall c_i$. Самі c_i знаходяться з (2):

$$B_1\left(\vec{Y}_0(a) + \sum_i c_i \vec{Y}_i(a)\right) + B_2\left(\vec{Y}_0(b) + \sum_i c_i \vec{Y}_i(b)\right) = \vec{d} \text{ або}$$

$$[B_1 + B_2 Y(b)]\vec{c} = \vec{d} - B_2 \vec{Y}_0(b) \qquad (6)$$

Розв'язуючи СЛАР (6) знаходимо c_i. За єдиністю $\vec{Y}(x) = \vec{U}(x)$.

Алгоритм А1.

1) Розв'язуємо задачу Коші (3), знаходимо $\vec{Y}_0(b)$.

2) Розв'язуємо m задач Коші (4), знаходимо $Y(b)$.

3) Розв'язуємо СЛАР (6), що дає нам $c_i, i = \overline{1,m}$.

4) $\vec{Y}(x) = U(x)$ знаходимо з (5).

Складність цього алгоритму співпадає зі складністю розв'язання $(m+1)$ задач Коші.

Якщо крайові умови розділені

$$C_1\vec{Y}(a) = \vec{d}_1, \; C_2\vec{Y}(b) = \vec{d}_2,$$

то можна зменшити кількість задач Коші, які необхідно розв'язати. Для цього побудуємо вектор $\vec{V}_0$ такий, що

$$C_1\vec{V}^0 = \vec{d}_1 \qquad (7)$$

Це завжди можна зробити, оскільки кількість рівнянь менше кількості невідомих. Далі будуємо $\vec{V}_i, i = \overline{1,k}$ такі, що

$$C_1\vec{V}^i = 0, \; i = \overline{1,k} \qquad (8)$$

Знову це можна здійснити бо $rang C_1 = m - k$.

Розв'яжемо задачі Коші:

$$\frac{d\vec{Y}^0}{dx} = A\vec{Y}^0 + \vec{F}, \; \vec{Y}^0(a) = \vec{V}^0 \qquad (9)$$

$$\frac{d\vec{Y}^i}{dx} = A\vec{Y}^i, \; \vec{Y}^i(a) = \vec{V}^i, \; i = \overline{1,k}. \qquad (10)$$

Сталі c_i знаходимо з умови виконання другої крайової умови.

Алгоритм А2.

1). Розв'язуємо СЛАР (8) – (9).

2). Розв'язуємо задачу Коші (9).

3). Розв'язуємо k задач Коші (10).

4). Розв‘язуємо СЛАР:

$$B_2\vec{Y}(b)\equiv C_2\left[\vec{Y}^0(b)+\sum_{i=1}^{k}c_i\vec{Y}(b)\right]=\vec{d}_2. \qquad (11)$$

5). Розв’язок

$$\vec{Y}(x)=\vec{Y}^0(x)+\sum_{i=1}^{k}c_i\vec{Y}^i(x). \qquad (12)$$

Оскільки для А1 та А2 розв’язок задачі Коші шукається чисельно, то маємо фактично значення

$$\vec{Y}^i(x_n),\ n=\overline{0,N},\ x_n\in[a,b].$$

Їх треба запам‘ятовувати для (12). Для запобігання цього у випадку розділеної крайової задачі запишемо алгоритм А3.

Алгоритм А3.

1). Розв’язуємо (8) – (9).

2). Розв’язуємо задачу Коші (9).

3). Розв’язуємо задачі Коші (10) і не запам’ятовуємо $\overline{Y}^i(x_n)$, а знаходимо тільки $\overline{Y}^i(x_N)=\overline{Y}^i(b)$.

4). Розв’язуємо СЛАР (11).

5). Розв’язуємо ще одну задачу Коші:

$$\frac{d\vec{Y}}{dx}=A\vec{Y}+F,\ \vec{Y}(a)=\vec{V}^0+\sum_{i=1}^{k}c_i\vec{V}^i\ .$$

Тоді за формулою (12) $\vec{Y}(x)=\vec{U}(x)$.

Зрозуміло, що “стріляти”, тобто починати розв‘язувати задачу Коші, треба з того боку, де задано більше крайових умов.

Суттєвий недолік алгоритмів!. Серед власних значень $A(x)$, як правило, є такі, що $\mathrm{Re}\,\lambda_i(x)>0$. Тоді лінійно незалежні розв’язки задачі Коші наростають експоненціально. Це призводить до наростання похибок заокруглень та погано обумовленої матриці системи (6) або (11) (розв‘язки $\overline{Y}^i(x)$ стають майже лінійно залежні).

Тому $[a,b]$ розбивають на проміжки $[x_{p-1},x_p],\ p=\overline{1,M}$, і розв’язують задачу Коші на підпроміжках, а в кінці $x=x_p$ ортогоналізують отримані розв’язки. Зрозуміло, що для $x=b$ отримують не $\vec{Y}^i(b)$, а деякі $\vec{W}^i(b)$, які залежать від $\vec{Y}^i(b)$ та відповідних перетворень ортогоналізації. З їх допомогою по $\vec{W}^i(b)$ обчислюють $\vec{Y}^i(b)$ та “прогоняють” ці умови для всіх значень $\vec{Y}(a)=\vec{Y}^0(a)+\sum c_i\vec{Y}^i(a)$. Така ідея метода ортогональної прогонки Годунова, що широко застосовується на практиці.

11.2. Метод пристрілки

Це метод для розв‘язання крайової задачі для нелінійних рівнянь аналогічний методу стрільби.

Розглянемо крайову задачу з розділеними крайовими умовами:

$$\frac{d\vec{U}}{dx} = \vec{F}\left(x, \vec{U}\right), a < x < b, \tag{1}$$

$$u_i(a) = c_i, i = \overline{k+1, m}, \tag{2}$$

$$\varphi_i\left(\vec{U}(b)\right) = d_i, i = \overline{1, k}. \tag{3}$$

При $x = a$ невідомі k початкових умов $u_i(a), i = \overline{1, k}$. Будемо їх шукати.

Розв’яжемо задачу Коші:

$$\begin{cases} \dfrac{d\vec{Y}}{dx} = \vec{F}\left(x, \vec{Y}\right), a < x < b; \\ \vec{Y}(a) = \vec{C} = (c_i)_{i=1}^{m}. \end{cases},$$

де $c_i, i = \overline{1, k}$ невідомі. Їх шукаємо з крайової умови (3):

$$f_i(c_1, ..., c_k) \equiv \varphi_i\left(\vec{\varphi}(b; c_1, ..., c_k)\right) - d_i = 0, \; i = \overline{1, k}.$$

Це система нелінійних рівнянь. Задаємо початкові значення $c_i^{(0)}, i = \overline{1, k}$. За якимось ітераційним методом знаходимо її розв‘язок. Найзручніше використовувати метод січних.

Метод пристрілки найбільш прозоро виглядає для $k = 1$. В цьому випадку нам необхідно знайти тільки c_1. Використаємо метод ділення навпіл. Знайдемо $c_1^{(0)}$ таке, що

$$\varphi_1\left(\vec{y}\left(b; c_1^{(0)}\right)\right) - d_1 > 0,$$

та $c_1^{(1)}$ таке, що

$$\varphi_1\left(\vec{y}\left(b; c_1^{(1)}\right)\right) - d_1 < 0.$$

Тоді вибираємо

$$c_1^{(2)} = \frac{1}{2}\left(c_1^{(0)} + c_1^{(1)}\right).$$

З трьох $c_1^{(0)}, c_1^{(1)}, c_1^{(2)}$ вибираємо таке, що $\varphi_1\left(\vec{y}(b; c_1)\right) - d_1$ міняє знак. Процес продовжуємо до виконання умови

$$\left|\varphi_1\left(\vec{y}\left(b, c_1^{(k)}\right)\right) - d_1\right| < \varepsilon,$$

де $\varepsilon > 0$ - задана точність.

11.3. Метод лінеаризації [ЛМС, 293 – 294], [БЖК, 441 – 442]

Розглянемо задачу:

$$\frac{d\vec{U}}{dx} = \vec{F}\left(x, \vec{U}\right) \qquad a < x < b, \tag{1}$$

$$\vec{\varphi}\left(\vec{U}(a), \vec{U}(b)\right) = \vec{d}. \tag{2}$$

Метод лінеаризації для задачі (1) це аналог методу Ньютона для систем нелінійних рівнянь. Нехай $\vec{Y}^0(x)$ – деяке наближення. Побудуємо його уточнення $\vec{Z}^0(x)$ до точного розв'язку $\vec{U}(x)$:

$$\vec{U}(x) = \vec{Y}^0(x) + \vec{Z}^0(x).$$

З (1) маємо $\frac{d\vec{Z}^0}{dx} = \Phi_F\left(x, \vec{V}\right)\vec{Z}^0(x) + \vec{F}\left(x, \vec{Y}^0\right) - \frac{d\vec{Y}^0}{dx}$. Замінюючи середнє значення $\vec{V}(x)$ на $\vec{Y}^0(x)$ отримаємо лінійне рівняння:

$$\frac{d\vec{Z}^0}{dx} = \Phi_F\left(x, \vec{Y}^0\right)\vec{Z}^0 + \vec{F}\left(x, \vec{Y}^0\right) - \frac{d\vec{Y}^0}{dx} \qquad (3)$$

Аналогічно:

$$\Phi_a\left(\vec{Y}^0(a), \vec{Y}^0(b)\right)\vec{Z}^0(a) + \Phi_b\left(\vec{Y}^0(a), \vec{Y}^0(b)\right)\vec{Z}^0(b) = \vec{d} - \vec{\varphi}\left(\vec{Y}^0(a), \vec{Y}^0(b)\right) \qquad (4)$$

Тут $\Phi_F = \left(\frac{\partial F_i}{\partial u_i}\right)_{i,j=\overline{1,m}}$ – матриця Якобі правої частини $\vec{F}\left(x, \vec{U}\right)$;

$\Phi_a = \left(\frac{\partial \varphi_i}{\partial u_j(a)}\right)_{i,j=\overline{1,m}}$, $\Phi_b = \left(\frac{\partial \varphi_i}{\partial u_j(b)}\right)_{i,j=\overline{1,m}}$ - матриці Якобі для $\vec{\varphi}\left(\vec{U}(a), \vec{U}(b)\right)$

по крайовим умовам в точках $x = a$ та $x = b$ відповідно. Задача (3)-(4) лінійна і розв'язується методом стрільби (з ортогоналізацією). Розв'язавши цю задачу, маємо наступне наближення $\vec{Y}^1(x) = \vec{Y}^0(x) + \vec{Z}^0(x)$. Цей процес продовжуємо до виконання умови точності $\left\|\vec{Z}^k(x)\right\| < \varepsilon$.

Недоліки методу:

1) Наявність похідної $\frac{d\vec{Y}^0}{dx}$ в правій частині. Оскільки розв'язок задач Коші чисельний, то для її обчислення треба застосовувати формули чисельного диференціювання. Це може привести до великих похибок за рахунок нестійкості задачі чисельного диференціювання.

2) Збіжність залежить від вибору $\vec{Y}^0$.

11.4. Метод продовження за параметром [ЛМС, 293 – 294]

Суттєвим недоліком методу ліанерізації є необхідність задавати хороше початкове наближення та чисельне диференціювання попереднього наближення. Розглянемо метод, який позбавлений цих недоліків.

Розглянемо задачу знаходження вектора $\vec{U}(x) = (u_i)_{i=\overline{1,n}}$, що задовольняє умовам:

$$\frac{d\vec{U}}{dx} = \vec{F}\left(x, \vec{U}\right) \qquad a < x < b, \qquad (1)$$

$$\vec{\varphi}\left(\vec{U}(a), \vec{U}(b)\right) = \vec{d}. \qquad (2)$$

Нехай розв'язок цієї задачі існує та єдиний.

Розв'яжемо задачу Коші

$$\frac{d\vec{Y}}{dx} = \vec{F}\left(x, \vec{Y}\right), \quad \vec{Y}(a) = \vec{Y}^0 \tag{3}$$

Вибір $\vec{Y}^0$ здійснимо так, щоб було задовольнялося як можна більша кількість з крайових умов (2). Наприклад, якщо $\varphi_i\left(\vec{U}(a), \vec{U}(b)\right) \equiv u_i(a)$, то вибираємо

$$y_i^0 = d_i.$$

Обчислимо $\vec{d}^{\,0} = \vec{\varphi}\left(\vec{Y}(a), \vec{Y}(b)\right)$. Якщо $\vec{d}^{\,0} \equiv \vec{d}$, то $\vec{Y} \equiv \vec{U}$. Але, як правило, $\vec{d}^{\,0} \neq \vec{d}$ і тому необхідно уточнювати початкове наближення. Розглянемо параметричну крайову задачу

$$\frac{d\vec{V}}{dx} = \vec{F}\left(x, \vec{V}\right), \quad a < x < b, \tag{4}$$

$$\vec{\varphi}\left(\vec{V}(a), \vec{V}(b)\right) = \lambda\vec{d} + (1-\lambda)\vec{d}^{\,0}, \tag{5}$$

яка залежить від параметра λ: $\vec{V} = \vec{V}(x, \lambda)$, Ясно, що $\vec{V}(x,0) = \vec{Y}(x)$, а $\vec{V}(x,1) = \vec{U}$.

Спробуємо продовжити розв‘язок задачі (4), (5) від відомого $\vec{Y}(x)$ до шуканого $\vec{U}(x)$. Для цього продиференціюємо (4), (5) по λ:

$$\frac{d}{dx}\frac{\partial V_i}{\partial \lambda} = \sum_{j=1}^{n} \frac{\partial F_i}{\partial u_j}\frac{\partial V_j}{\partial \lambda},$$

$$\sum_{j=1}^{n} \frac{\partial \varphi_i}{\partial u_j(a)}\frac{\partial V_j(a)}{\partial \lambda} + \sum_{j=1}^{n} \frac{\partial \varphi_i}{\partial u_j(b)}\frac{\partial V_j(b)}{\partial \lambda} = d_i - d_i^0.$$

Позначимо $\vec{Z} = \dfrac{\partial \vec{V}}{\partial \lambda}$. Тоді останню систему можна записати у вигляді:

$$\frac{d\vec{Z}}{dx} = \Phi_F\left(x, \vec{V}\right)\vec{Z}, \quad a < x < b, \tag{4}$$

$$\Phi_a\left(\vec{V}(a), \vec{V}(b)\right)\vec{Z}(a) + \Phi_b\left(\vec{V}(a), \vec{V}(b)\right)\vec{Z}(b) = \vec{d} - \vec{d}^{\,0}, \tag{5}$$

$$\frac{\partial \vec{V}}{\partial \lambda} = \vec{Z}, \quad \vec{V}(x,0) = \vec{Y}^0. \tag{6}$$

де $\Phi_F = \left(\dfrac{\partial F_i}{\partial u_j}\right)_{i,j=\overline{1,n}}$ - матриця Якобі правої частини рівняння (1) $\vec{F}(x.\vec{U})$;

$\Phi_a = \left(\dfrac{\partial \varphi_i}{\partial u_j(a)}\right)_{i,j=\overline{1,n}}$ - матриця Якобі лівої частини $\vec{\varphi}\left(\vec{U}(a), \vec{U}(b)\right)$ крайової умови (2) по першому аргументу $\vec{U}(a)$; $\Phi_b = \left(\dfrac{\partial \varphi_i}{\partial u_j(b)}\right)_{i,j=\overline{1,n}}$ - матриця Якобі лівої частини $\vec{\varphi}\left(\vec{U}(a), \vec{U}(b)\right)$ крайової умови (2) по другому аргументу $\vec{U}(b)$.

Задача (4)-(6) не простіше ніж вихідна задача (1)-(2), а ще й складніша за неї. Спростимо її, застосувавши до задачі Коші (6) чисельний метод, наприклад, метод Ейлера:

$$\vec{V}^{k+1}(x) = \vec{V}^{k}(x) + \Delta\lambda \vec{Z}^{k}(x),\ \vec{V}^{0}(x) = \vec{Y}(x).$$

Тут $\vec{V}^{k}(x) = \vec{V}(x, \lambda_k)$, $\Delta\lambda = \lambda_{k+1} - \lambda_k$, $\lambda_0 = 0$, $\lambda_K = 1$, $k = \overline{1, K}$.

Знайдене наближення $\vec{V}^{k+1}(x)$ використовується для знаходження наступного наближення $\vec{Z}^{k+1}(x)$ лінійної крайової задачі (4), (5).

Повністю алгоритм розв‘язання крайової задачі (1), (2) цим методом такий:

1. Розв‘язуємо задачу Коші (3). Задаємо початкові значення $\vec{V}^{0}(x) = \vec{Y}(x)$.
2. Для $k = \overline{1, K}$ розв‘язуємо лінійні крайові задачі:

$$\frac{d\vec{Z}^{k}}{dx} = \Phi_F\left(x, \vec{V}^{k}\right)\vec{Z}^{k},\quad a < x < b,$$

$$\Phi_a\left(\vec{V}^{k}(a), \vec{V}^{k}(b)\right)\vec{Z}^{k}(a) + \Phi_b\left(\vec{V}^{k}(a), \vec{V}^{k}(b)\right)\vec{Z}^{k}(b) = \vec{d} - \vec{d}^{\,0};$$

3. Продовжуємо розв‘язок по параметру λ:

$$\vec{V}^{k+1}(x) = \vec{V}^{k}(x) + \Delta\lambda \vec{Z}^{k}(x).$$

4. Шуканий розв‘язок $\vec{U}(x) \approx \vec{V}^{K}(x)$.

Лінійні крайові задачі пункту 2 розв‘язуються, наприклад, методом стрільби. Для розв‘язання задачі Коші (6) можна застосовувати більш точні методи ніж метод Ейлера.