

1. Аналіз похибок заокруглення

1.1. Види похибок

Нехай необхідно розв'язати рівняння

$$Au = f. \quad (1)$$

За рахунок неточно заданих вхідних даних насправді ми маємо рівняння

$$\tilde{A}\tilde{u} = \tilde{f}. \quad (2)$$

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

Застосування методу розв'язання (2) приводить до рівняння

$$\tilde{A}_h \tilde{u}_h = \tilde{f}_h, \quad (3)$$

де $h > 0$ — малий параметр.

Означення: Назвемо $\delta_2 = \tilde{u} - \tilde{u}_h$ похибкою методу.

Реалізація методу на ЕОМ приводить до рівняння

$$\tilde{A}_h^* \tilde{u}_h^* = \tilde{f}_h^*. \quad (4)$$

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

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

Означення: кажуть, що задача (1) коректна, якщо

- $\forall f \in F: \exists! u \in U$;
- Задача (1) стійка, тобто $\forall \varepsilon > 0: \exists \delta > 0$:

$$|A - \tilde{A}| < \delta, |f - \tilde{f}| < \delta \implies |u - \tilde{u}| < \varepsilon. \quad (5)$$

Якщо задача (1) некоректна, то або розв'язок її не існує, або він неєдиний, або він нестійкий, тобто $\exists \varepsilon > 0: \forall \delta > 0$:

$$|A - \tilde{A}| < \delta, |f - \tilde{f}| < \delta \implies |u - \tilde{u}| > \varepsilon. \quad (6)$$

Означення: Абсолютна похибка $\Delta x \leq |x - x^*|$.

Означення: Відносна похибка $\delta x \leq \Delta x / |x|$, або $\Delta x / |x^*|$.

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

Означення: Вірна цифра — це значуща, якщо абсолютна похибка за рахунок відкидання всіх молодших розрядів не перевищує одиниці розряду цієї цифри.

Тобто, якщо $x^* = \alpha_n \dots \alpha_0 . \alpha_{-1} \dots \alpha_{-p} \dots$, то α_{-p} вірна, якщо $\Delta x \leq 10^{-p}$ (інколи $\Delta x \leq w \cdot 10^{-p}$, де $1/2 \leq w < 1$ 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

$$x = \pm(\alpha_1\beta^{-1} + \alpha_2\beta^{-2} + \dots + \alpha_t\beta^{-t} + \dots) \cdot \beta^p, \quad (7)$$

де $0 \leq \alpha_k < \beta$, $\alpha_1 \neq 0$, $k = 1, 2, \dots$

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

$$x^* = \pm(\alpha_1\beta^{-1} + \alpha_2\beta^{-2} + \dots + \alpha_t\beta^{-t}) \cdot \beta^p \quad (8)$$

і відповідно похибка заокруглення

$$x - x^* = \pm\beta^p \cdot (\alpha_{t+1}\beta^{-t-1} + \dots). \quad (9)$$

Тоді її можна оцінити так

$$|x - x^*| \leq \beta^{p-t-1} \cdot (\beta - 1) \cdot (1 + \beta^{-1} + \dots) \leq \beta^{p-t-1} \cdot (\beta - 1) \cdot \frac{1}{1 - \beta^{-1}} = \beta^{p-t}. \quad (10)$$

Якщо в представленні (7) взяти $\alpha_1 = 1$, то $|x| \geq \beta^p \cdot \beta^{-1}$. Звідси остаточно

$$\delta x \leq \frac{\beta^{p-t}}{\beta^{p-1}} = \beta^{-t+1}. \quad (11)$$

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

1.3. Підрахунок похибок обчислення значення функції

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

$$|y - y^*| = |f(\vec{x}) - f(\vec{x}^*)| = \left| \sum_{i=1}^n \frac{\partial f}{\partial x_i}(\vec{\xi}) \cdot (x_i - x_i^*) \right| \leq \sum_{i=1}^n B_i \cdot \Delta x_i, \quad (12)$$

$$\text{де } B_u = \max_{\vec{x} \in U} \left| \frac{\partial f}{\partial x_i}(\vec{x}) \right|.$$

Тут

$$U = \{ \vec{x} = (x_1, \dots, x_n) : |x_i - x_i^*| \leq \Delta x_i \} \subset \Omega, \quad (13)$$

для $i = \overline{1, n}$. Отже з точністю до величин першого порядку малості по

$$\Delta x = \max_i \Delta x_i, \quad (14)$$

$$\Delta y = |y - y^*| \prec \sum_{i=1}^n b_i \cdot \Delta x_i, \quad (15)$$

де $b_i = \left| \frac{\partial f}{\partial x_i}(\vec{x}^*) \right|$ та « \prec » означає приблизно менше.

Розглянемо похибки арифметичних операцій.

- Сума: $y = x_1 + x_2, x_1, x_2 > 0$:

$$\Delta y \leq \Delta x_1 + \Delta x_2, \quad (16)$$

$$\delta y \leq \frac{\Delta x_1 + \Delta x_2}{x_1 + x_2} \leq \max(\delta x_1, \delta x_2). \quad (17)$$

- Різниця: $y = x_1 - x_2, x_1 > x_2 > 0$:

$$\Delta y \leq \Delta x_1 + \Delta x_2, \quad (18)$$

$$\delta y \leq \frac{x_2 \delta x_1 + x_1 \delta x_2}{x_1 - x_2}. \quad (19)$$

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

- Добуток: $y = x_1 \cdot x_2, x_1, x_2 > 0$:

$$\Delta y \prec x_2 \Delta x_1 + x_1 \Delta x_2, \quad (20)$$

$$\delta y \leq \delta x_1 + \delta x_2. \quad (21)$$

- Частка $y = x_1/x_2, x_1, x_2 > 0$:

$$\Delta y \prec \frac{x_2 \Delta x_1 + x_1 \Delta x_2}{x_2^2}, \quad (22)$$

$$\delta y \leq \delta x_1 + \delta x_2. \quad (23)$$

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

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

Означення: *Обернена задача:* знаходження $\Delta x_i, i = \overline{1, n}$ по заданих $\Delta y, \delta y$. Якщо $n > 1$, маємо одну умову

$$\sum_{i=1}^n b_i \cdot \Delta x_i < \varepsilon \quad (24)$$

для багатьох невідомих Δx_i .

Вибирають їх із однієї з умов:

$$\forall i : b_i \cdot \Delta x_i < \frac{\varepsilon}{n} \quad (25)$$

або

$$\Delta x_i < \frac{\varepsilon}{\sum_{i=1}^n b_i}. \quad (26)$$

2. Методи розв'язання нелінійних рівнянь

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

Задача розв'язання цього рівняння розпадається на етапи:

- Існування та кількість коренів.
- Відділення коренів, тобто розбиття числової вісі на інтервали, де знаходиться один корінь.
- 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

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

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

для $f(x) \in C[a, b]$, який необхідно визначити. Нехай $f(a) \cdot f(b) < 0$.

Припустимо, що $f(a) > 0$, $f(b) < 0$. Покладемо $x_1 = \frac{a+b}{2}$ 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 послідовного 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

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

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

$$|b_{k+1} - a_{k+1}| = \frac{|b_k - a_k|}{2}, \quad (3)$$

то

$$|x_{n+1} - \bar{x}| \leq \frac{b - a}{2^{n+1}} < \varepsilon. \quad (4)$$

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

$$n = n(\varepsilon) \geq \left\lceil \log \left(\frac{b - a}{\varepsilon} \right) \right\rceil + 1. \quad (5)$$

Степінь збіжності — лінійна, тобто геометричної прогресії з знаменником $q = 1/2$.

- **Переваги методу:** простота, надійність.

- **Недоліки методу:** низька швидкість збіжності; метод не узагальнюється на системи.

2.2. Метод простої ітерації

Спочатку рівняння

$$f(x) = 0 \quad (6)$$

замінюється еквівалентним

$$x = \varphi(x). \quad (7)$$

Ітераційний процес має вигляд

$$x_{n+1} = \varphi(x_n), \quad n = 0, 1, \dots \quad (8)$$

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 (6) до (7) з функцією $\varphi(x) = x + \tau(x) \cdot f(x)$, де $\tau(x)$ — знакостала функція на тому відрізку, де шукаємо корінь.

Означення: Кажуть, що ітераційний метод збігається, якщо $\lim_{k \rightarrow \infty} x_k = \bar{x}$.

Далі $U_r = \{x : |x - a| \leq r\}$ відрізок довжини $2r$ з серединою в точці a .

З'ясуємо умови, при яких збігається метод простої ітерації.

Теорема 1: Якщо

$$\max_{x \in [a, b] = U_r} |\varphi'(x)| \leq q < 1 \quad (9)$$

то метод простої ітерації збігається і має місце оцінка

$$|x_n - \bar{x}| \leq \frac{q^n}{1 - q} \cdot |x_0 - x_1| \leq \frac{q^n}{1 - q} \cdot (b - a). \quad (10)$$

Доведення: Нехай $x_{k+1}, x_k \in U_r$. Тоді

$$\begin{aligned} |x_k - x_{k-1}| &= |\varphi(x_k) - \varphi(x_{k-1})| = |\varphi'(\xi_k) \cdot (x_k - x_{k-1})| \leq |\varphi'(\xi_k)| \cdot |x_k - x_{k-1}| \leq \\ &\leq q \cdot |x_k - x_{k-1}| = \dots = q^k \cdot |x_1 - x_0|, \end{aligned} \quad (11)$$

де $\xi_k = x_k + \theta_k \cdot (x_{k+1} - x_k)$, а у свою чергу $0 < \theta_k < 1$. Далі

$$\begin{aligned} |x_{k+p} - x_k| &= |x_{k+p} - x_{k+p-1} + \dots + x_{k+1} - x_k| = |x_{k+p} - x_{k+p-1}| + \dots + |x_{k+1} - x_k| \leq \\ &\leq (q^{k+p-1} + q^{k+p-2} + \dots + q^k) \cdot |x_1 - x_0| = \frac{q^k - q^{k+p-1}}{1 - q} \cdot |x_1 - x_0| \xrightarrow[k \rightarrow \infty]{} 0. \end{aligned} \quad (12)$$

Бачимо що $\{x_k\}$ — фундаментальна послідовність. Значить вона збіжна. При $p \rightarrow \infty$ в (12) отримуємо (10). \square

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

$$|x_n - \bar{x}| \leq \frac{q^n}{1 - q} \cdot (b - a) < \varepsilon, \quad (13)$$

звідки безпосередньо маємо

$$n(\varepsilon) = n \geq \left\lceil \frac{\ln\left(\frac{\varepsilon(1-q)}{b-a}\right)}{\ln q} \right\rceil + 1. \quad (14)$$

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

$$|\varphi(x) - \varphi(y)| \leq q \cdot |x - y|, \quad 0 < q < 1. \quad (15)$$

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

2.3. Метод релаксації

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

$$x_{n+1} = x_n + \tau \cdot f(x_n), \quad (16)$$

де $k = 0, 1, 2, 3 \dots$, а x_0 — задано. Метод можна записати у вигляді

$$\frac{x_{k+1} - x_k}{\tau} = f(x_k), \quad k = 0, 1, \dots \quad (17)$$

Оскільки $\varphi'(x) = 1 + \tau \cdot f'(x)$, то метод збігається при умові

$$|\varphi'(x)| = |1 + \tau \cdot f'(x)| \leq q < 1. \quad (18)$$

Нехай $f'(x) < 0$, тоді (8) запишеться у вигляді: $-q \leq 1 + \tau \cdot f'(x) \leq q < 1$. Звідси

$$f'(x) \leq 1 + q < 2k\tau, \quad (19)$$

і

$$0 < \tau < \frac{2}{|f'(x)|}. \quad (20)$$

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

Підставивши $x_k = x + z_k$ в (16), отримаємо

$$z_{k+1} = z_k + \tau \cdot f(x + z_k). \quad (21)$$

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

$$f(\bar{x} + z_k) = f(\bar{x}) + z_k \cdot f'(\bar{x} + \theta \cdot z_k) = z_k \cdot f'(\bar{x} + \theta \cdot z_k) = z_k \cdot f'(\xi_k), \quad (22)$$

тобто

$$z_{k+1} = z_k + \tau \cdot f'(\xi_k) \cdot z_k. \quad (23)$$

Звідси

$$|z_{k+1}| \leq |1 + \tau \cdot f'(\xi_k)| \cdot |z_k| \leq \max_U |1 + \tau \cdot f'(\xi_k)| \cdot |z_k|. \quad (24)$$

А тому

$$|z_{k+1}| \leq \max \{|1 - \tau M_1|, |1 - \tau m_1|\} \cdot |z_k|, \quad (25)$$

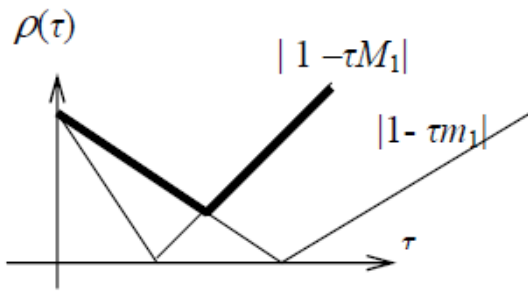
де

$$m_1 = \min_{[a,b]} |f'(x)|, \quad M_1 = \max_{[a,b]} |f'(x)| \quad (26)$$

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

$$q(\tau) = \max \{|1 - \tau M_1|, |1 - \tau m_1|\} \quad (27)$$

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



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

$$1 - \tau_0 m_1 = \tau_0 M_1 - 1 \implies \tau_0 = \frac{2}{M_1 - m_1} < \frac{2}{|f'(x)|}. \quad (28)$$

При цьому значенні τ маємо

$$q(\tau_0) = \rho_0 = \frac{M_1 - m_1}{M_1 + m_1}. \quad (29)$$

Тоді для похибки вірна оцінка

$$|x_n - \bar{x}| \leq \frac{\rho_0^n}{1 - \rho_0} \cdot (b - a) < \varepsilon. \quad (30)$$

Кількість ітерацій

$$n = n(\varepsilon) \geq \left\lceil \frac{\frac{\ln(\varepsilon(1-\rho_0))}{b-a}}{\ln \rho_0} \right\rceil + 1. \quad (31)$$

Задача 1: Дати геометричну інтерпретацію методу простої ітерації для випадків:

$$0 < \varphi'(x) < 1; \quad -1 < \varphi'(x) < 0; \quad \varphi'(x) < -1; \quad \varphi'(x) > 1. \quad (32)$$

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

2.4. Метод Ньютона (метод дотичних)

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

$$0 = f(\bar{x}) = f(x_k + \bar{x} - x_k) = f(x_k) + f'(\xi_k) \cdot (x - x_k), \quad (33)$$

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

$$f(x_k) + f'(x_k) \cdot (x_{k+1} - x_k) = 0. \quad (34)$$

Звідси маємо ітераційний процес

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad (35)$$

де $k = 0, 1, 2, \dots$; x_0 — 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

$$\varphi(x) = x - \frac{f(x)}{f'(x)}, \quad (36)$$

тобто

$$\tau(x) = -\frac{1}{f'(x)}. \quad (37)$$

Тому

$$\varphi'(x) = 1 - \frac{f'(x) \cdot f'(x) - f(x) \cdot f''(x)}{(f'(x))^2} = \frac{f(x) \cdot f''(x)}{(f'(x))^2}. \quad (38)$$

Якщо \bar{x} — корінь $f(x)$, то $\varphi'(\bar{x}) = 1$. знайдеться окіл кореня, \end{equation}

$$|\varphi'(x)| = \left| \frac{f(x) \cdot f''(x)}{(f'(x))^2} \right| < 1. \quad (39)$$

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:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)}, \quad k = 0, 1, 2, \dots \quad (40)$$

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

$$\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} \quad (41)$$

дає метод січних:

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \cdot f(x_k), \quad (42)$$

де $k = 1, 2, \dots$, x_0, x_1 — задані.

Задача 5: Дати геометричну інтерпретацію методу січних.

2.5. Збіжність методу Ньютона

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

$$f(x) = 0. \quad (43)$$

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

$$q = \frac{M_2 \cdot |x_0 - \bar{x}|}{2m_1} < 1, \quad (44)$$

де

$$m_1 = \min_{U_r} |f'(x)|, \quad M_2 = \max_{U_r} |f''(x)|, \quad (45)$$

то для $x_0 \in U_r$ метод Ньютона

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad (46)$$

збігається і має місце оцінка

$$|x_n - \bar{x}| \leq q^{2^n - 1} \cdot |x_0 - \bar{x}|. \quad (47)$$

З (46) маємо

$$x_{k+1} - \bar{x} = x_k - \frac{f(x_k)}{f'(x_k)} - \bar{x} = \frac{(x_k - \bar{x}) \cdot f'(x_k) - f(x_k)}{f'(x_k)} = \frac{F(x_k)}{f'(x_k)}, \quad (48)$$

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

- $F(\bar{x}) = 0$;

- $F'(x) = (x - \bar{x}) \cdot f''(x).$

Тоді

$$F(x_k) = F(\bar{x}) + \int_x^{x_k} F'(t) dt = \int_x^{x_k} (t - \bar{x}) \cdot f''(t) dt. \quad (49)$$

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

$$F(x_k) = f''(\xi_k) \int_x^{x_k} (t - \bar{x}) dt = \frac{(x_k - \bar{x})^2}{2} \cdot f''(\xi_k), \quad (50)$$

де $\xi_k = \bar{x} + \theta_k \cdot (x_k - \bar{x})$, де $0 < \theta_k < 1$. З (48), (50) маємо

$$x_{k+1} - \bar{x} = \frac{(x_k - \bar{x})^2}{2f'(\xi_k)} \cdot f''(\xi_k). \quad (51)$$

Доведемо оцінку (46) за індукцією. Так як $x_0 \in U_r$, то

$$|\xi_0 - \bar{x}| = |\theta_0 \cdot (x_0 - \bar{x})| < |\theta_0| \cdot |x_0 - \bar{x}| < r \quad (52)$$

звідси випливає $\xi_0 \in U_r$.

Тоді $f''(\xi_0) \leq M_2$, тому

$$|x_1 - \bar{x}| \leq \frac{(x_0 - \bar{x})^2 \cdot M_2}{2m_1} = \frac{M_2 \cdot |x_0 - \bar{x}|}{2m_1} \cdot |x_0 - \bar{x}| = q \cdot |x_0 - \bar{x}| < r, \quad (53)$$

тобто $x_1 \in U_r$.

Ми довели твердження (47) при $n = 1$. Нехай воно справджується при $n = k$

$$|x_k - \bar{x}| \leq q^{2^k - 1} \cdot |x_0 - \bar{x}| < r, \quad (54)$$

$$|\xi_k - \bar{x}| = |\theta_k \cdot (x_k - \bar{x})| < r. \quad (55)$$

Тоді $x_k, \xi_k \in U_r$.

Доведемо (47) для $n = k + 1$. З (51) маємо

$$\begin{aligned} |x_{k+1} - \bar{x}| &\leq \frac{|x_k - \bar{x}|^2 \cdot M_2}{2m_1} \leq \left(q^{2^k - 1}\right)^2 \cdot \frac{|x_0 - \bar{x}|^2 \cdot M_2}{2m_1} = \\ &= q^{2^{k+1} - 2} \cdot \frac{|x_0 - \bar{x}| \cdot M_2}{2m_1} \cdot |x_0 - \bar{x}| = q^{2^{k+1} - 1} \cdot |x_0 - \bar{x}|. \end{aligned} \quad (56)$$

Таким чином (47) справджується для $n = k + 1$. Значить (47) виконується і для довільного n . Таким чином $x_n \xrightarrow{n \rightarrow \infty} x$. \square

З (47) маємо оцінку кількості ітерацій для досягнення точності ε

$$n \geq \left\lceil \log_2 \left(1 + \frac{\ln \left(\frac{\varepsilon}{b-a} \right)}{\ln q} \right) \right\rceil + 1. \quad (57)$$

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

$$|x_{k+1} - \bar{x}| = O(|x_k - \bar{x}|^m). \quad (58)$$

Для методу Ньютона

$$|x_{k+1} - \bar{x}| = \frac{|x_k - \bar{x}|^2 \cdot |f''(\xi_k)|}{2|f'(x_k)|}. \quad (59)$$

Звідси випливає, що

$$|x_{k+1} - \bar{x}| = O(|x_k - \bar{x}|^2). \quad (60)$$

Значить ступінь 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 x простий корінь рівняння $f(x) = 0$ ($f'(x) \neq 0$). Якщо $f'(x) \cdot f''(x) > 0$ ($f'(x) \cdot f''(x) < 0$) то для методу Ньютона при $x_0 = b$ послідовність наближень $\{x_k\}$ монотонно спадає (монотонно зростає при $x_0 = a$).

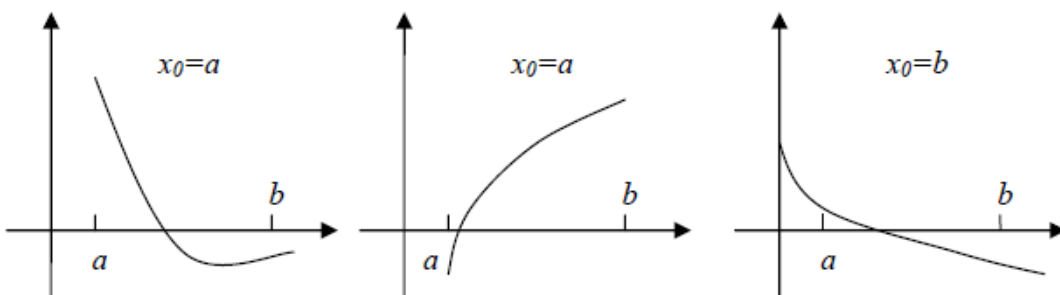
Задача 6: Довести [теорему 2](#) при

- $f'(x) \cdot f''(x) > 0$;
- $f'(x) \cdot f''(x) < 0$.

Задача 7: Знайти ступінь збіжності методу січних [Калиткин Н.Н., Численные методы, с. 145–146]

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

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



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

$$f^{(m)}(\bar{x}) = 0, \quad m = 0, 1, \dots, p-1; \quad f^{(p)}(\bar{x}) \neq 0, \quad (61)$$

то в методі Ньютона необхідна наступна модифікація

$$x_{k+1} = x_k - p \cdot \frac{f(x_k)}{f'(x_k)} \quad (62)$$

i

$$q = \frac{M_{p+1} \cdot |x_0 - \bar{x}|}{m_p \cdot p \cdot (p+1)} < 1. \quad (63)$$

Зауваження 2: Метод Ньютона можна застосовувати і для обчислення комплексного кореня

$$z_{k+1} = z_k - \frac{f(z_k)}{f'(z_k)} \quad (64)$$

В теоремі про збіжність

$$q = \frac{|x_0 - \bar{x}| M_2}{2m_1}, \quad (65)$$

де

$$m_1 = \min_{U_r} |f'(z)|, \quad M_2 = \max_{U_r} |f''(z)|. \quad (66)$$

Тут $|z|$ — модуль комплексного числа.

Переваги методу Ньютона:

- висока швидкість збіжності;
- узагальнюється на системи рівнянь;
- узагальнюється на комплексні корені.

Недоліки методу Ньютона:

- на кожній ітерації обчислюється не тільки $f(x_k)$, а і похідна $f'(x_k)$;
- збіжність залежить від початкового наближення x_0 , оскільки від нього залежить умова збіжності

$$q = \frac{M_2 |x_0 - \bar{x}|}{2m_1} < 1; \quad (67)$$

- 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 Гауса

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 $Q = O(n! \cdot n)$.

Запишемо СЛАР у вигляді

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \dots + a_{1,n}x_n = b_1 \equiv a_{1,n+1}, \\ a_{2,1}x_1 + a_{2,2}x_2 + \dots + a_{2,n}x_n = b_2 \equiv a_{2,n+1}, \\ \dots \\ a_{n,1}x_1 + a_{n,2}x_2 + \dots + a_{n,n}x_n = b_n \equiv a_{n,n+1}. \end{cases} \quad (2)$$

Якщо $a_{1,1} \neq 0$, то ділимо перше рівняння на нього і виключаємо x_1 з інших рівнянь:

$$\begin{cases} x_1 + a_{1,2}^{(1)}x_2 + \dots + a_{1,n}^{(1)}x_n = a_{1,n+1}^{(1)}, \\ a_{2,2}^{(1)}x_2 + \dots + a_{2,n}^{(1)}x_n = a_{2,n+1}^{(1)}, \\ \dots \\ a_{n,2}^{(1)}x_2 + \dots + a_{n,n}^{(1)}x_n = a_{n,n+1}^{(1)}. \end{cases} \quad (3)$$

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

$$\begin{cases} x_1 + a_{1,2}^{(1)}x_2 + \dots + a_{1,n}^{(1)}x_n = a_{1,n+1}^{(1)}, \\ x_2 + \dots + a_{2,n}^{(2)}x_n = a_{2,n+1}^{(2)}, \\ \dots \\ x_n = a_{n,n+1}^{(n)}. \end{cases} \quad (4)$$

Тобто

$$A^{(n)}\vec{x} = \vec{a}^{(n)}. \quad (5)$$

Це прямий хід методу Гауса. Формули прямого ходу

```

for k in range(1, n):
    for j in range(k + 1, n + 2):
        a[k, j][k] = a[k, j][k - 1] / a[k, k][k - 1]
    for i in range(k + 1, n + 1):
        a[i, j][k] = a[i, j][k - 1] - \
            a[i, j][k - 1] * a[k, j][k]

```

Звідси

$$x_n = a_{n,n+1}^{(n)}, \quad x_i = a_{i,n+1}^{(i)} - \sum_{j=i+1}^n a_{i,j}^{(n)} x_j, \quad (6)$$

для $i = \overline{n-1, 1}$. Це формули оберненого ходу.

Складність, тобто кількість операцій, яку необхідно виконати для реалізації методу: $Q_f = 2/3n^2 + O(n^2)$ для прямого ходу, $Q_b = n^2 + O(n)$ для оберненого ходу.

Умова

$$a_{k,k}^{(k-1)} \neq 0 \quad (7)$$

не суттєва, оскільки знайдеться m , для якого

$$\left| a_{m,k}^{(k-1)} \right| = \max_i \left| a_{i,k}^{(k-1)} \right| \neq 0 \quad (8)$$

(оскільки $\det A \neq 0$). Тоді міняємо місцями рядки номерів k і m .

Означення: Елемент

$$a_{k,k}^{(k-1)} \neq 0 \quad (9)$$

називається *ведучим*.

Введемо матриці

$$M_k = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & m_{k,k} & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & m_{n,k} & \cdots & 1 \end{pmatrix} \quad (10)$$

елементи якої обчислюється так:

$$m_{k,k} = \frac{1}{a_{k,k}^{(k-1)}}, \quad m_{i,k} = -\frac{a_{i,k}^{(k-1)}}{a_{k,k}^{(k-1)}}. \quad (11)$$

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 Гауса можна записати у вигляді

$$M_n M_{n-1} \dots M_1 A \vec{x} = M_n M_{n-1} \dots M_1 \vec{b}. \quad (12)$$

Позначимо останню систему, яка співпадає з (5), так

$$U \vec{x} = \vec{c}, \quad U = (u_{i,j})_{i,j=1}^n, \quad (13)$$

причому

$$\begin{cases} u_{i,i} = 1, \\ u_{i,j} = 0, \quad i > j. \end{cases} \quad (14)$$

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

$$L_k = M_k^{-1} = \begin{pmatrix} 1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & a_{k,k}^{(k-1)} & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & a_{n,k}^{(k-1)} & \dots & 1 \end{pmatrix} \quad (15)$$

Тоді

$$A = L_1 \dots L_n U = LU; \quad L = L_1 \dots L_n, \quad (16)$$

де L — нижня трикутна матриця, U — верхня трикутна матриця. Таким чином метод Гауса 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 множенні на неї матриці A_{k-1} робимо ведучим елементом максимальний за модулем.

$$P_k = \begin{pmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{pmatrix} \quad (17)$$

За допомогою цих матриць перехід до трикутної системи (13) тепер має вигляд:

$$M_n M_{n-1} P_{n-1} \dots M_1 P_1 A \vec{x} = M_n M_{n-1} P_{n-1} \dots M_1 P_1 \vec{b}. \quad (18)$$

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.

Висновки про **переваги** трикутного розкладу:

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

3.2. Метод квадратних коренів

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

$$A\vec{x} = \vec{b}, \quad A^T = A. \quad (19)$$

Він оснований на розкладі матриці A в добуток:

$$A = S^T D S, \quad (20)$$

де S — верхня трикутна матриця, S^T — нижня трикутна матриця, D — діагональна матриця.

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

$$DS_{i,j} = \begin{cases} d_{i,i}s_{i,j}, & i \leq j, \\ 0, & i > j. \end{cases} \quad (21)$$

Далі

$$S^T D S_{i,j} = \sum_{l=1}^n s_{i,l}^T d_{l,l} s_{l,j} = \sum_{l=1}^{i-1} s_{l,i}^T s_{l,j} d_{l,l} + s_{i,i} s_{i,j} d_{i,i} + \underbrace{s_{i,i}^T \sum_{l=i+1}^n s_{l,i}^T s_{l,j} d_{l,l}}_{=0} = a_{i,j}, \quad (22)$$

для $i, j = \overline{1, n}$.

Якщо $i = j$, то

$$|s_{i,i}^2| d_{i,i} = a_{i,i} - \sum_{l=1}^{i-1} |s_{l,i}^2| d_{l,l} \equiv p_i. \quad (23)$$

Тому

$$d_{i,i} = \text{sign}(p_i), \quad s_{i,i} = \sqrt{|p_i|}. \quad (24)$$

Якщо $i < j$, то

$$s_{i,j} = \left(a_{i,j} - \sum_{l=1}^{i-1} s_{l,i}^T d_{l,j} s_{l,j} \right) / (s_{i,i} d_{i,i}), \quad (25)$$

де $i = \overline{1, n}$, а $j = \overline{i+1, n}$.

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

Знайдемо розв'язок рівняння (19). Враховуючи (20), маємо:

$$S^T D \vec{y} = \vec{b} \quad (26)$$

і

$$S \vec{x} = \vec{y} \quad (27)$$

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

$$y_i = \frac{b_i - \sum_{j=1}^{i-1} s_{j,i} d_{j,j} y_j}{s_{i,i} d_{i,i}}, \quad (28)$$

для $i = \overline{1, n}$ і

$$x_i = \frac{y_i - \sum_{j=1}^{i-1} s_{i,j} x_j}{s_{i,i}}, \quad (29)$$

для $i = \overline{n-1, 1}$, де $x_n = y_n / s_{n,n}$.

Метод застосовується лише для симетричних матриць. Його складність $Q = n^3/3 + O(n^2)$.

Переваги цього методу:

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

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

3.3. Обчислення визначника та оберненої матриці

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

$$\det P \det A = \det L \det U \quad (30)$$

звідки

$$\det A = (-1)^\ell \det L \det U = (-1)^\ell \prod_{k=1}^n a_{k,k}^{(k)}, \quad (31)$$

де ℓ — кількість перестановок. Ясно, що за методом Гауса

$$Q_{\det} = \frac{2}{3} \cdot n^3 + O(n^2) \quad (32)$$

В методі квадратного кореня $A = S^\top D S$. Тому

$$\det A = \det S^\top \det D \det S = \prod_{k=1}^n d_{k,k} \prod_{k=1}^n s_{k,k}^2. \quad (33)$$

Тепер $Q_{\det} = n^3/3 + O(n^2)$.

За означенням

$$A A^{-1} = E, \quad (34)$$

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

$$A^{-1} = (\alpha_{i,j})_{i,j=1}^n. \quad (35)$$

Тоді $\vec{\alpha}_j = (\alpha_{i,j})_{i=1}^n$ — вектор-стовпчик оберненої матриці. З (34) маємо

$$A \vec{\alpha}_j = \vec{e}_j, \quad j = \overline{1, n}. \quad (36)$$

де \vec{e}_j — стовпчики одиничної матриці: $\vec{e}_j = (\delta_{i,j})_{i=1}^n$,

$$\delta_{i,j} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \quad (37)$$

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

3.4. Метод прогонки

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

$$\begin{cases} -c_0 y_0 + b_0 y_1 = -f_0, \\ a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \\ a_N y_{N-1} - c_N y_N = -f_N. \end{cases} \quad (38)$$

Матриця системи

$$A = \begin{pmatrix} -c_0 & b_0 & & 0 \\ a_0 & \ddots & \ddots & \\ & \ddots & \ddots & b_N \\ 0 & & a_N & -c_N \end{pmatrix} \quad (39)$$

тридіагональна.

Розв'язок представимо у вигляді

$$y_i = \alpha_{i+1}y_{i+1} + \beta_{i+1}, \quad i = \overline{0, N-1}. \quad (40)$$

Замінімо в (40) і $i \mapsto i-1$ і підставимо в (33), тоді

$$(a_i\alpha_i - c_i) \cdot y_i + b_i y_{i+1} = -f_i - a_i\beta_i \quad (41)$$

Звідси

$$y_i = \frac{b_i}{c_i - a_i\alpha_i} \cdot y_{i+1} + \frac{f_i + a_i\beta_i}{c_i - a_i\alpha_i}. \quad (42)$$

Тому з (36)

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i\alpha_i}, \quad \beta_{i+1} = \frac{f_i + a_i\beta_i}{c_i - a_i\alpha_i}, \quad i = \overline{1, N-1}. \quad (43)$$

Умова розв'язності (38) — $c_i - a_i\alpha_i \neq 0$.

Щоб знайти всі α_i, β_i , треба задати перші значення. З (38):

$$\alpha_1 = \frac{b_0}{c_0}, \quad \beta_1 = \frac{f_0}{c_0}. \quad (44)$$

Після знаходження всіх α_i, β_i обчислюємо y_N з системи

$$\begin{cases} a_N y_N - c_N y_N = -f_N, \\ y_{N-1} = \alpha_N y_N + \beta_N. \end{cases} \quad (45)$$

Звідси

$$y_N = \frac{f_N + a_N\beta_N}{c_N - a_N\alpha_N}. \quad (46)$$

Алгоритм:

```
alpha[1], beta[1] = b[0] / c[0], f[0] / c[0]

for i in range(1, N):
    z[i] = c[i] - a[i] * alpha[i]
    alpha[i + 1], beta[i + 1] = b[i] / z[i], \
        (f[i] + a[i] * beta[i]) / z[i]

y[N] = (f[N] + a[N] * beta[N]) / \
    (c[N] - a[N] * alpha[N])

for i in range(N - 1, -1, -1):
    y[i] = alpha[i + 1] * y[i + 1] + beta[i + 1]
```

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 $a_i, b_i \neq 0$, та

$$|c_i| \geq |a_i| + |b_i|, \quad \forall i, \quad a_0 = b_N = 0, \quad (47)$$

та хоча би одна нерівність строга. Тоді $|\alpha_i| \leq 1$ та

$$z_i = c_i - a_i \alpha_i \neq 0, \quad i = \overline{1, N}. \quad (48)$$

Задача 8: Довести теорему про стійкість методу прогонки.

3.5. Обумовленість систем лінійних алгебраїчних рівнянь

Нехай задано СЛАР

$$A\vec{x} = \vec{b}. \quad (49)$$

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

$$B\vec{y} = \vec{h}, \quad (50)$$

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

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

$$A = \text{diag}(\varepsilon), \quad a_{i,j} = \varepsilon \delta_{i,j}. \quad (51)$$

Тоді $\det A = \varepsilon^n \ll 1$, але $x_i = b_i/\varepsilon$. Тому $\Delta x_i = \Delta b_i/\varepsilon \gg 1$.

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

$$(A + C)(\vec{x} + \vec{z}) = \vec{b} + \vec{\eta}. \quad (52)$$

Віднімемо від цієї рівності (49) у вигляді $A\vec{z} + C\vec{x} + C\vec{z} = \vec{\eta}$. Тоді

$$A\vec{z} = \vec{\eta} - C\vec{x} - C\vec{z}, \quad \vec{z} = A^{-1}(\vec{\eta} - C\vec{x} - C\vec{z}). \quad (53)$$

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

$$|\vec{z}|_1 = \sum_{i=1}^n |z_i|, \quad (54)$$

$$|\vec{z}|_2 = \left(\sum_{i=1}^n |z_i|^2 \right)^{1/2}, \quad (55)$$

$$|\vec{z}|_\infty = \max_i |z_i|. \quad (56)$$

Означення: Норми матриці, що відповідають нормам вектора, тобто

$$|A|_m = \sup_{|\vec{x}|_m \neq 0} \frac{|A\vec{x}|_m}{|\vec{x}|_m}, \quad m = 1, 2, \infty. \quad (57)$$

такі:

$$|A|_1 = \max_j \sum_{i=1}^n |a_{i,j}|, \quad (58)$$

$$|A|_2 = \max_i \sqrt{\lambda_i(A^T A)}, \quad (59)$$

$$|A|_\infty = \max_i \sum_{j=1}^n |a_{i,j}|, \quad (60)$$

де $\lambda_i(B)$ — власні значення матриці B .

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

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

Означення: Число обумовленості матриці A — $\text{cond}(A) = \|A\| \cdot \|A^{-1}\|$.

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

$$\delta(\vec{x}) \leq \frac{\text{cond}(A)}{1 - \text{cond}(A) \cdot \delta(A)} (\delta(A) + \delta(\vec{b})), \quad (61)$$

де $\text{cond}(A)$ — число обумовленості.

Доведення:

$$A\vec{z} = \vec{\eta} - C\vec{x} - C\vec{z}, \quad \vec{z} = A^{-1}\vec{\eta} - A^{-1}C\vec{x} - A^{-1}C\vec{z} \quad (62)$$

$$|\vec{z}| \leq |A^{-1}\vec{\eta}| + |A^{-1}C\vec{x}| + |A^{-1}C\vec{z}| \leq |A^{-1}| \cdot |\vec{\eta}| + |A^{-1}| \cdot |C| \cdot |\vec{x}| + |A^{-1}| \cdot |C| \cdot |\vec{z}|. \quad (63)$$

$$|\vec{z}| \leq \frac{|A^{-1}| \cdot (|\vec{\eta}| + |C| \cdot |\vec{x}|)}{1 - |A^{-1}| \cdot |C|} \quad (64)$$

Оцінка похибки

$$\begin{aligned} \delta(\vec{x}) &\leq \frac{|A^{-1}|}{1 - |A^{-1}| \cdot |C|} \left(\frac{|\vec{\eta}|}{|\vec{x}|} + |C| \right) = \frac{|A^{-1}| \cdot |A|}{1 - |A^{-1}| \cdot |A| \cdot \frac{|C|}{|A|}} \left(\frac{|\vec{\eta}|}{|A| \cdot |\vec{x}|} + \delta(A) \right) \leq \\ &\leq \frac{\text{cond}(A)}{1 - \text{cond}(A) \cdot \delta(A)} \left(\frac{|\vec{\eta}|}{|\vec{x}|} + \delta(A) \right). \quad \square \end{aligned} \quad (65)$$

Наслідок: Якщо $C \equiv 0$, то $\delta(\vec{x}) \leq \text{cond}(A) \cdot \delta(\vec{b})$.

Властивості $\text{cond}(A)$:

- $\text{cond}(A) \geq 1$;
- $\text{cond}(A) \geq \max_i |\lambda_i(A)| / \min_i |\lambda_i(A)|$;
- $\text{cond}(AB) \leq \text{cond}(A) \cdot \text{cond}(B)$;
- $A^\top = A^{-1} \implies \text{cond}(A) = 1$.

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

$$|A^{-1}| \geq \max \frac{1}{|\lambda_A|} = \frac{1}{\min |\lambda_A|}. \quad (66)$$

Якщо $1 \ll \text{cond}(A)$, 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 (61) витікає: $\delta(\vec{x}) = \text{cond}(A) \cdot O(n\beta^{-t})$. 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.

Приклад погано обумовленої системи — системи з матрицею Гільберта

$$H_n = \left(\frac{1}{i+j-1} \right)_{i,j=1}^n, \quad (67)$$

наприклад $\text{cond}(H_8) \approx 10^9$.

4. Ітераційні методи для систем

4.1. Ітераційні методи розв'язання СЛАР

Систему

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

зводимо до вигляду

$$\vec{x} = B\vec{x} + \vec{f}. \quad (2)$$

Будь яка система

$$\vec{x} = \vec{x} - C \cdot (A\vec{x} - \vec{b}) \quad (3)$$

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

$$\vec{x} = \vec{x} - \tau \cdot (A\vec{x} - \vec{b}). \quad (3')$$

4.1.1. Метод простої ітерації

Цей метод застосовується до рівняння (2)

$$\vec{x}^{(k+1)} = B\vec{x}^{(k)} + \vec{f}, \quad (4)$$

де $\vec{x}^{(0)}$ — початкове наближення, задано.

Теорема: Ітераційний процес збігається, тобто

$$\left| \vec{x}^{(k)} - \vec{x} \right| \xrightarrow[k \rightarrow \infty]{} 0, \quad (5)$$

якщо

$$|B| \leq q < 1. \quad (6)$$

При цьому має місце оцінка

$$\left| \vec{x}^{(n)} - \vec{x} \right| \leq \frac{q^n}{1 - q} \cdot \left| \vec{x}^{(1)} - \vec{x}^{(0)} \right|. \quad (7)$$

4.1.2. Метод Якобі

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

$$x_i = - \sum_{j=1}^{i-1} \frac{a_{i,j}}{a_{i,i}} \cdot x_j - \sum_{j=i+1}^n \frac{a_{i,j}}{a_{i,i}} \cdot x_j + \frac{b_i}{a_{i,i}}, \quad (8)$$

де $i = \overline{1, n}$.

Ітераційний процес запишемо у вигляді

$$x_i^{(k+1)} = - \sum_{j=1}^{i-1} \frac{a_{i,j}}{a_{i,i}} \cdot x_j^{(k)} - \sum_{j=i+1}^n \frac{a_{i,j}}{a_{i,i}} \cdot x_j^{(k)} + \frac{b_i}{a_{i,i}}, \quad (9)$$

де $k = 0, 1, \dots$, а $i = \overline{1, n}$.

Теорема: Ітераційний процес збігається до розв'язку, якщо виконується умова

$$\forall i : \sum_{\substack{j=1 \\ i \neq j}}^n |a_{i,j}| \leq |a_{i,i}|. \quad (10)$$

Це умова діагональної переваги матриці A .

Теорема: Якщо ж

$$\forall i : \sum_{\substack{j=1 \\ i \neq j}}^n |a_{i,j}| \leq q \cdot |a_{i,i}|, \quad 0 \leq q < 1. \quad (11)$$

то має місце оцінка точності:

$$|\vec{x}^{(n)} - \vec{x}| \leq \frac{q^n}{1 - q} \cdot |\vec{x}^{(0)} - \vec{x}|. \quad (12)$$

4.1.3. Метод Зейделя

В компонентному вигляді ітераційний метод Зейделя записується так:

$$x_i^{(k+1)} = - \sum_{j=1}^{i-1} \frac{a_{i,j}}{a_{i,i}} \cdot x_j^{(k+1)} - \sum_{j=i+1}^n \frac{a_{i,j}}{a_{i,i}} \cdot x_j^{(k)} + \frac{b_i}{a_{i,i}}, \quad (13)$$

де $k = 0, 1, \dots$, а $i = \overline{1, n}$.

На відміну від методу Якобі на k -му-кроці попередні компоненти розв'язку беруться з $(k + 1)$ -ої ітерації.

Теорема: Достатня умова збіжності методу Зейделя — $A^T = A > 0$.

4.1.4. Матрична інтерпретація методів Якобі і Зейделя

Подано матрицю A у вигляді

$$A = A_1 + D + A_2, \quad (14)$$

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

$$D\vec{x} = A_1\vec{x} + A_2\vec{x} + \vec{b}, \quad (15)$$

або

$$\vec{x} = D^{-1}A_1\vec{x} + D^{-1}A_2\vec{x} + D^{-1}\vec{b}, \quad (16)$$

Матричний запис методу Якобі:

$$\vec{x}^{(k+1)} = D^{-1}A_1\vec{x}^{(k)} + D^{-1}A_2\vec{x}^{(k)} + D^{-1}\vec{b}, \quad (17)$$

методу Зейделя:

$$\vec{x}^{(k+1)} = D^{-1}A_1\vec{x}^{(k+1)} + D^{-1}A_2\vec{x}^{(k)} + D^{-1}\vec{b}, \quad (18)$$

Теорема: Необхідна і достатня умова збіжності методу Якобі: всі корені рівняння

$$\det(D + \lambda(A_1 + A_2)) = 0 \quad (19)$$

по модулю більше 1.

Теорема: Необхідна і достатня умова збіжності методу Зейделя: всі корені рівняння

$$\det(A_1 + D + \lambda A_2) = 0 \quad (20)$$

по модулю більше 1.

4.1.5. Однокрокові (двошарові) ітераційні методи

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

$$B_k \frac{\vec{x}^{(k+1)} - \vec{x}^{(k)}}{\tau_{k+1}} + A\vec{x}^{(k)} = \vec{b}, \quad (21)$$

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

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

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \tau_{k+1} (A\vec{x}^{(k)} + \vec{b}). \quad (22)$$

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

$$B_k \vec{x}^{(k+1)} = F^k. \quad (23)$$

У цьому випадку на кожній ітерації необхідно розв'язувати СЛАР.

Означення: Якщо $\tau_{k+1} \equiv \tau$, $B_k \equiv B$, то ітераційний процес називається *стаціонарним*; інакше — *нестационарним*.

Методам, що розглянуті вище відповідають:

- методу простої ітерації: $B_k = E$, $\tau_{k+1} = \tau$;
- 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

$$\gamma_1 E \leq A \leq \gamma_2 E, \quad \gamma_1, \gamma_2 > 0. \quad (24)$$

Тоді при виборі $\tau = \tau_0 = \frac{2}{\gamma_1 + \gamma_2}$ ітераційний процес збігається. Найбільш точним значенням γ_1, γ_2 при яких виконуються обмеження (24) є $\gamma_1 = \min \lambda_i(A)$, $\gamma_2 = \max \lambda_i(A)$. Тоді

$$q = q_0 = \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1} = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \quad (25)$$

і справедлива оцінка

$$\left| \vec{x}^{(n)} - \vec{x} \right| \leq \frac{q^n}{1-q} \cdot \left| \vec{x}^{(0)} - \vec{x} \right|. \quad (26)$$

Зауваження: аналогічно обчислюється q і для методу релаксації розв'язання нелінійних рівнянь, де $\gamma_1 = m = \min |f'(x)|$, $\gamma_2 = M_1 = \max |f'(x)|$.

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

$$B \equiv E, \quad \tau_k : \min_{\tau} q(\tau), \quad n = n(\varepsilon) \rightarrow \min, \quad (27)$$

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:

$$(D + \omega A_1) \cdot \frac{\vec{x}^{(k+1)} + \vec{x}^{(k)}}{\omega} + A\vec{x}^{(k)} = \vec{b}, \quad (28)$$

де D — діагональна матриця з елементами $a_{i,i}$ 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 апіорної інформації про γ_1 , γ_2 в (24).

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

$$\left| \vec{r}^{(k+1)} \right|^2 = \left| \vec{r}^{(k)} \right|^2 - 2\tau_{k+1} \cdot \left\langle \vec{r}^{(k)}, A\vec{r}^{(k)} \right\rangle + \tau_{k+1}^2 \cdot \left| A\vec{r}^{(k)} \right|^2 \rightarrow \min. \quad (29)$$

Тому

$$\tau_{k+1} = \frac{\left\langle A\vec{r}^{(k)}, \vec{r}^{(k)} \right\rangle}{\left| \vec{r}^{(k)} \right|^2}, \quad (30)$$

де $\vec{r}^{(k)} = A\vec{x}^{(k)} - \vec{b}$ — нев'язка.

Умова для завершення ітераційного процесу:

$$\left| \vec{r}^{(n)} \right| < \varepsilon. \quad (31)$$

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

Аналогічно будуються методи з $B \neq E$. Матриця B називається переобумовлювачем і дозволяє підвищити швидкість збіжності ітераційного процесу. Його вибирають з умов

- легко розв'язувати СЛАР $B\vec{x}^{(k)} = F_k$ (діагональний, трикутний, добуток трикутних та інше);
- зменшення числа обумовленості матриці $B^{-1}A$ у порівнянні з A .

4.2. Методи розв'язання нелінійних систем

Розглянемо систему рівнянь

$$\begin{cases} f_1(x_1, \dots, x_n) = 0, \\ \dots \\ f_n(x_1, \dots, x_n) = 0. \end{cases} \quad (32)$$

Перепишемо її у векторному вигляді:

$$\vec{f}(\vec{x}) = 0. \quad (33)$$

4.2.1. Метод простої ітерації

В цьому методі рівняння (33) зводиться до еквівалентного вигляду

$$\vec{x} = \vec{\Phi}(\vec{x}). \quad (34)$$

Ітераційний процес представимо у вигляді:

$$\vec{x}^{(k+1)} = \vec{\Phi}(\vec{x}^{(k)}). \quad (35)$$

початкове наближення $\vec{x}^{(0)}$ — задано.

Нехай оператор $\vec{\Phi}$ визначений на множині H . За теоремою про стискуючі відображення ітераційний процес (35) сходиться, якщо виконується умова

$$\left| \vec{\Phi}(\vec{x}) - \vec{\Phi}(\vec{y}) \right| \leq q \cdot |\vec{x} - \vec{y}|, \quad 0 < q < 1, \quad (36)$$

або

$$\left| \vec{\Phi}'(\vec{x}) \right| \leq q < 1, \quad (37)$$

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

$$\left| \vec{x}^{(m)} - \vec{x} \right| \leq \frac{q^n}{1-q} \cdot \left| \vec{x}^{(0)} - \vec{x} \right|. \quad (38)$$

Частинним випадком методу простої ітерації є метод релаксації для рівняння (33):

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \tau \cdot \vec{F}(\vec{x}^{(k)}), \quad (39)$$

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

4.2.2. Метод Ньютона

Розглянемо рівняння

$$\vec{F}(\vec{x}) = 0. \quad (40)$$

Представимо його у вигляді

$$\vec{F}(\vec{x}^{(k)}) + \vec{F}'(\vec{\xi}^{(k)}) \cdot (\vec{x} - \vec{x}^{(k)}) = 0, \quad (41)$$

де

$$\vec{\xi}^{(k)} = \vec{x}^{(k)} + \theta_k \cdot (\vec{x}^{(k)} - \vec{x}), \quad (42)$$

де $0 < \theta_k < 1$. Тут $\vec{F}'(\vec{x}) = \left(\frac{\partial f_i}{\partial x_j} \right)_{i,j=1}^n$ — матриця Якобі для $\vec{F}(\vec{x})$. Можемо наближено вважати $\vec{\xi}^{(k)} \approx \vec{x}^{(k)}$. Тоді з (41) матимемо

$$\vec{F}(\vec{x}^{(k)}) + \vec{F}'(\vec{x}^{(k)}) \cdot (\vec{x}^{(k+1)} - \vec{x}^{(k)}) = 0. \quad (43)$$

Ітераційний процес представимо у вигляді:

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \vec{F}'(\vec{x}^{(k)})^{-1} \cdot \vec{F}(\vec{x}^{(k)}). \quad (44)$$

Для реалізації методу Ньютона потрібно, щоб існувала обернена матриця

$$\vec{F}'(\vec{x}^{(k)})^{-1}. \quad (45)$$

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

$$\begin{aligned} A_k \vec{z}^{(k)} &= \vec{F}(\vec{x}^{(k)}), \\ \vec{x}^{(k+1)} &= \vec{x}^{(k)} - \vec{z}^{(k)}, \end{aligned} \quad (46)$$

де $k = 0, 1, 2, \dots$, і $\vec{x}^{(0)}$ — задано, а матриця $A_k = \vec{F}'(\vec{x}^{(k)})$.

Метод має квадратичну збіжність, якщо добре вибрано початкове наближення. Складність методу (при умові використання методу Гаусса розв'язання СЛАР (46) на кожній ітерації $Q_n = \frac{2}{3}n^3 + O(n^2)$, де n — розмірність системи (33).

4.2.3. Модифікований метод Ньютона

Ітераційний процес має вигляд:

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \vec{F}'(\vec{x}^{(0)})^{-1} \cdot \vec{F}(\vec{x}^{(k)}). \quad (47)$$

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

Запишемо метод у вигляді системи лінійних рівнянь (аналог (46))

$$\begin{aligned} A_0 \vec{z}^{(k)} &= \vec{F}(\vec{x}^{(k)}), \\ \vec{x}^{(k+1)} &= \vec{x}^{(k)} - \vec{z}^{(k)}, \end{aligned} \quad (48)$$

де $k = 0, 1, 2, \dots$

Оскільки матриця 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 $A \in \mathbb{R}^{n \times n}$. Тоді задача на власні значення ставиться так: знайти число λ та вектор $x \neq 0$, що задовольняють рівнянню

$$Ax = \lambda x. \quad (1)$$

Означення: λ називається *власним значенням* A , а x — *власним вектором*.

З (1)

$$\det(A - \lambda E) = P_n(\lambda) = (-1)^n \lambda^n + a_n \lambda^{n-1} + \dots + a_0 = 0. \quad (2)$$

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

Для розв'язання багатьох задач механіки, фізики, хімії потрібне знаходження всіх власних значень $\lambda_i, i = \overline{1, n}$, а іноді й всіх власних векторів x_i , що відповідають λ_i .

Означення: 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.

Означення: Останні дві задачі називають *частковими проблемами власних значень*.

5.1. Степеневий метод

1. Знаходження λ_{\max} : $\lambda_{\max} \equiv |\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \dots$

Нехай $x^{(0)}$ — заданий вектор, будемо послідовно обчислювати вектори

$$x^{(k+1)} = Ax^{(k)}, \quad k = 0, 1, \dots \quad (3)$$

Тоді $x^{(k)} = A^k x^{(0)}$. Нехай $\{e_i\}_{i=1}^n$ — система власних векторів. Представимо $x^{(0)}$ у вигляді:

$$x^{(0)} = \sum_{i=1}^n c_i e_i. \quad (4)$$

Оскільки $Ae_i = \lambda_i e_i$, то $x^{(k)} = \sum_{i=1}^n c_i \lambda_i^k e_i$. При великих k : $x^{(k)} \approx c_1 \lambda_1^k e_1$. Тому

$$\mu_1^{(k)} = \frac{x_m^{(k+1)}}{x_m^{(k)}} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right). \quad (5)$$

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

Якщо матриця $A = A^T$ симетрична, то існує ортонормована система векторів $\langle e_i, e_j \rangle = \delta_{ij}$. Тому

$$\begin{aligned} \mu_1^{(k)} &= \frac{\langle x^{(k+1)}, x^{(k)} \rangle}{\langle x^{(k)}, x^{(k)} \rangle} = \frac{\left\langle \sum_i c_i \lambda_i^{k+1} e_i, \sum_j c_j \lambda_j^k e_j \right\rangle}{\left\langle \sum_i c_i \lambda_i^k e_i, \sum_j c_j \lambda_j^k e_j \right\rangle} = \frac{\sum_i c_i^2 \lambda_i^{2k+1}}{\sum_i c_i^2 \lambda_i^{2k}} = \\ &= \frac{c_1^2 \lambda_1^{2k+1} + c_2^2 \lambda_2^{2k+1} + \dots}{c_1^2 \lambda_1^{2k} + c_2^2 \lambda_2^{2k} + \dots} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right) \xrightarrow[k \rightarrow \infty]{} \lambda_1. \end{aligned} \quad (6)$$

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

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

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

```
e[0] = x[0] / norm(x[0])

k = 0
while True:
    k += 1

    x[k + 1] = A * x[k]
    μ[k][1] = scalar_product(x[k + 1], e[k])
    e[k + 1] = x[k + 1] / norm(x[k + 1])

    if abs(μ[k + 1][1] - μ[k][1]) < ε:
        break

λ[1] = μ[k + 1][1]
```

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

2. Знаходження λ_2 : $|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \dots$. Нехай λ_1, e_1 відомі.

Задача 10: Довести, що якщо $|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \dots$ то

$$\mu_2^{(k)} = \frac{x_m^{(k+1)} - \lambda_1 x_m^{(k)}}{x_m^{(k)} - \lambda_1 x_m^{(k-1)}} \xrightarrow{k \rightarrow \infty} \lambda_2, \quad (7)$$

де $x^{(k+1)} = Ax^{(k)}$, $x_m^{(k)}$ — m -та компонента $x^{(k)}$.

Задача 11: Побудувати алгоритм обчислення λ_2, 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 $\lambda_i(a) > 0$ то відоме λ_{\max} . Розглянемо матрицю $B = \lambda_{\max}E - A$.
Маємо

$$\forall i: \quad \lambda_i(B) = \lambda_{\max} - \lambda_i(A). \quad (8)$$

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

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

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

$$Ax^{(k+1)} = x^{(k)}, \quad k = 0, 1, \dots \quad (9)$$

Але цей метод вимагає більшої кількості арифметичних операцій: складність методу на основі формули (3): $Q = O(n^2)$, а на основі (9) — $Q = O(n^3)$, оскільки треба розв'язувати СЛАР, але збігається метод (9) швидше.

5.2. Ітераційний метод обертання

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

$$A = U\Lambda U^T, \quad (10)$$

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

З (1) маємо

$$\Lambda = U^T A U. \quad (11)$$

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

Тоді діагональні елементи мало відрізняються від власних значень

$$|\tilde{\lambda}_{ij} - \lambda_i(A)| < \varepsilon = \varepsilon(\delta). \quad (12)$$

Введемо

$$t(A) = \sum_{\substack{i,j=1 \\ i \neq j}}^n a_{ij}^2. \quad (13)$$

З малості величини $t(A)$ випливає, що діагональні елементи малі. По $A = A_0$ за допомогою матриць обертання U_k :

$$U_k = \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cos \phi & \cdots & -\sin \phi & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sin \phi & \cdots & \cos \phi & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}. \quad (14)$$

що повертають систему векторів на кут φ , побудуємо послідовність $\{A_k\}$ таку, що $A_k \xrightarrow[k \rightarrow \infty]{} \Lambda$.

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

Послідовно будуюмо:

$$A_{k+1} = U_k^T A_k U_k, \quad (15)$$

Означення: Процес (15) називається *монотонним*, якщо: $t(A_{k+1}) < t(A_k)$.

Задача 13: Довести, що для матриці (15) виконується:

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} \cos(2\varphi) + \frac{1}{2} \left(a_{j,j}^{(k)} - a_{i,i}^{(k)} \right) \sin(2\varphi), \quad (16)$$

Показати, що

$$t(A_{k+1}) = t(A_k) - 2\left(a_{i,j}^{(k)}\right)^2 \quad (17)$$

якщо вибрати φ з умови $a_{i,j}^{(k+1)} = 0$.

Звідси

$$\varphi = \varphi_k = \frac{1}{2} \arctan\left(p^{(k)}\right), \quad (18)$$

тобто

$$p^{(k)} = \frac{2a_{i,j}^{(k)}}{a_{i,i}^{(k)} - a_{j,j}^{(k)}}, \quad (19)$$

де

$$\left|a_{i,j}^{(k)}\right| = \max_{\substack{m,l \\ m \neq l}} \left|a_{m,l}^{(k)}\right|. \quad (20)$$

Тоді $t(A_k) \xrightarrow[k \rightarrow \infty]{} 0$. Чим більше n тим більше ітерацій необхідно для зведення A до Λ .

Якщо матриця несиметрична, то застосовують QR- або QL-методи.

6. Інтерполявання функцій

6.1. Постановка задачі інтерполявання

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 $x_i \neq x_j$ для $i \neq j$.

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

Задача інтерполявання функції має не єдиний розв'язок.

Означення: Виберемо систему лінійно незалежних функцій $\{\varphi_k(x)\}_{k=0}^n$, $\varphi_k(x) \in C([a, b])$ і побудуємо лінійну комбінацію

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

яка називається *узагальненим багаточленом*.

Умови інтерполявання дають СЛАР

$$\sum_{k=0}^n c_k \cdot \varphi_k(x_i) = y_i, \quad i = \overline{1, n} \quad (2)$$

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

Якщо

$$D(x_0, \dots, x_n) = \begin{vmatrix} \varphi_0(x_0) & \cdots & \varphi_n(x_0) \\ \vdots & \ddots & \vdots \\ \varphi_0(x_n) & \cdots & \varphi_n(x_n) \end{vmatrix} \neq 0, \quad (3)$$

то система (2) має єдиний розв'язок.

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

Приклади систем Чебишова:

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

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

$$\begin{aligned} D(x_0, \dots, x_n) &= \begin{vmatrix} 1 & x_0 & \cdots & x_0^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^n \end{vmatrix} = \\ &= \prod_{0 \leq k < m \leq n} (x_k - x_m) \neq 0, \end{aligned} \quad (4)$$

2. $\varphi_k(x) = L_k(x)$ — ортогональні багаточлени Лежандра;

3. $\varphi_k(x) = T_k(x)$ — ортогональні багаточлени Чебишова.

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

Тоді

$$\begin{aligned}\Phi_n(x) &= T_n(x) = \\ &= a_0 + \sum_{k=1}^n (a_k \cdot \cos(kx) + b_k \cdot \sin(kx))\end{aligned}\quad (5)$$

— тригонометричний багаточлен.

6.2. Інтерполяційна формула Лагранжа

Якщо $\varphi_k(x) = x^k$, то

$$\Phi_n(x) = P_n(x) = \sum_{k=0}^n c_k \cdot x^k. \quad (6)$$

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

$$P_n(x) = L_n(x) = \sum_{k=0}^n f(x_k) \cdot \Phi_k^{(n)}(x). \quad (7)$$

Означення: Тут $L_n(x)$ — інтерполяційний 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

$$\Phi_k^{(n)}(x_i) = \delta_{i,k}. \quad (8)$$

Оскільки $\Phi_k^{(n)}(x)$ — багаточлен степеня n , то він має вигляд

$$\Phi_k^{(n)}(x) = A_k(x - x_0) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_n), \quad (9)$$

де A_k — число.

Знайдемо його з умови $\Phi_k^{(n)}(x_k) = 1$:

$$A_k = \frac{1}{(x_k - x_0) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_n)}. \quad (10)$$

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

$$\Phi_k^{(n)}(x) = \frac{(x - x_0) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_n)}{(x_k - x_0) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_n)} \quad (11)$$

Позначивши

$$\omega_n(x) = \prod_{i=0}^n (x - x_i), \quad (12)$$

маємо

$$\Phi_k^{(n)}(x) = \frac{\omega_n(x)}{(x - x_k) \cdot \omega'_n(x_k)}. \quad (13)$$

Остаточно формула Лагранжа має вигляд

$$L_n(x) = \sum_{k=0}^n f(x_k) \cdot \frac{\omega_n(x)}{(x - x_k) \cdot \omega'_n(x_k)} \quad (14)$$

6.3. Залишковий член інтерполяційного полінома

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

Означення: Заміняючи функцію $f(x)$ на $L_n(x)$ ми допускаємо похибку $r_n(x) = f(x) - L_n(x)$. Це *залишковий член* інтерполявання.

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:

$$g(t) = f(t) - L_n(t) - K \cdot \omega_n(t), \quad (15)$$

де $t \in [a, b]$, і $g(x_i) = 0$ для $i = \overline{0, n}$.

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

$$K = \frac{f(x) - L_n(x)}{\omega_n(x)}. \quad (16)$$

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, отримуємо, що існує хоча б одна $\xi \in [a, b]$ така, що $g^{(n+1)}(\xi) = 0$. Оскільки

$$g^{(n+1)}(t) = f^{(n+1)} - 0 - K \cdot (n+1)!, \quad (17)$$

то $\exists \xi$, що

$$g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - (n+1)! \cdot \frac{f(x) - L_n(x)}{\omega_n(x)} = 0. \quad (18)$$

Звідси

$$r_n(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \cdot \omega_n(x). \quad (19)$$

Оскільки ξ невідомо, то використовують оцінку залишкового члена:

$$|r_n(x)| = |f(x) - L_n(x)| \leq \frac{M_{n+1}}{(n+1)!} \cdot |\omega_n(x)|, \quad (20)$$

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

6.4. Багаточлени Чебишова. Мінімізація залишкового члена інтерполяційного полінома

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

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

$$T_{n+1}(x) - 2x \cdot T_n(x) + T_{n-1}(x) = 0, \quad (21)$$

де початкові значення

$$T_0(x) = 1, \quad T_1(x) = x. \quad (22)$$

Знайдемо явний вигляд багаточлена Чебишова. Будемо шукати розв'язок рівняння (21) у вигляді $T_n(x) = q^n$, де $q = q(x)$. Підставивши в (21), отримуємо характеристичне рівняння $q^2 - 2xq + 1 = 0$. Тоді при $|x| \geq 1 \implies q_{1,2} = x \pm \sqrt{x^2 - 1}$, а при $|x| < 1 \implies q_{1,2} = \cos(\varphi) \pm i \sin(\varphi)$, $\varphi = \arccos(x)$.

Розглянемо обидва випадки детальніше:

1. при $|x| \leq 1$: $T_n(x) = A \cdot \cos(n\varphi) + B \cdot \sin(n\varphi)$. З (21) випливає $A = 1$, $B = 0$ і тому

$$T_n(x) = \cos(n \arccos(x)). \quad (23)$$

2. при $|x| > 1$:

$$T_n(x) = \frac{1}{2} \left(\left(x + \sqrt{x^2 - 1} \right)^n + \left(x - \sqrt{x^2 - 1} \right)^n \right). \quad (24)$$

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:

$$x_k = \cos\left(\frac{(2k+1)\pi}{2n}\right) \in [-1, 1], \quad k = \overline{0, n-1}. \quad (25)$$

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

$$x'_k = \cos\left(\frac{k\pi}{2n}\right), \quad k = \overline{0, n}. \quad (26)$$

Коефіцієнт при старшому члені багаточлена дорівнює 2^{n-1} .

Означення: Введемо нормований багаточлен Чебишова $\bar{T}_n(x) = 2^{1-n} T_n(x) = x^n + \dots$

Тоді

$$\left| \bar{T}_n \right|_{C([-1,1])} = \max_{x \in [-1,1]} |T_n(x)| = 2^{1-n}. \quad (27)$$

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

$$\Delta(f, \Phi) = |f(x) - \Phi(x)|_{C([a,b])}. \quad (28)$$

Теорема (Чебишова): Серед усіх багаточленів n -го степеня з коефіцієнтом 1 при старшому степені $\overline{T}_n(x)$ найменше відхиляється від 0 на $[-1, 1]$, тобто

$$\left| \overline{T}_n(x) - 0 \right|_{C([-1,1])} = \inf_{\overline{P}_n(x)} \left| \overline{P}_n(x) \right|_{C([-1,1])} = 2^{1-n}. \quad (29)$$

Доведення. Будемо доводити від супротивного: припустимо, що існує багаточлен, такий, що

$$\overline{Q}_n(x) < 2^{1-n}. \quad (30)$$

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:

$$\operatorname{sgn}(Q_{n-1}(x'_k)) = \operatorname{sgn}(\overline{T}_n(x'_k) - \overline{Q}_n(x'_k)) = \operatorname{sgn}(\overline{T}_n(x'_k)) = \alpha \cdot (-1)^k, \quad (31)$$

де $\alpha = \pm 1$.

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 $\leq n - 1$. \square

Тепер узагальнимо наш багаточлен Чебишова на довільний проміжок. Нагадаємо $T_n(t) = \cos(n \arccos t)$, $-1 \leq t \leq 1$. Від змінної $t \in [-1, 1]$ перейдемо до $x \in [a, b]$. Запровадимо заміну

$$t = \frac{2x}{b-a} - \frac{b+a}{b-a}, \quad x = \frac{b+a}{2} + \frac{b-a}{2}t. \quad (32)$$

Тоді

$$T_n^{[a,b]}(t) = \overline{T}_n\left(\frac{2x}{b-a} - \frac{b+a}{b-a}\right) = 2^{1-n} \cos\left(n \arccos\left(\frac{2x - (b+a)}{b-a}\right)\right). \quad (33)$$

Побудований нами багаточлен Чебишова на $[a, b]$ не є нормованим.

Нормований багаточлен Чебишова на $[a, b]$:

$$\overline{T}_n^{[a,b]}(x) = \frac{(b-a)^n}{2^{2n-1}} \cos\left(n \arccos\left(\frac{2x - (b+a)}{b-a}\right)\right). \quad (34)$$

Відповідно його нулі

$$x_k = \frac{a+b}{2} - \frac{b-a}{2} \cdot t_k, \quad t_k = \cos\left(\frac{(2k+1)\pi}{2n}\right), \quad (35)$$

де $k = \overline{0, n-1}$, а точки екстремуму

$$x'_k = \frac{a+b}{2} - \frac{b-a}{2} \cdot t'_k, \quad t'_k = \cos\left(\frac{k\pi}{n}\right), \quad k = \overline{0, n}. \quad (36)$$

Теорема Чебишова вірна і у випадку $[a, b]$. Тепер

$$\left| \overline{T}_n^{[a,b]} \right|_{C([a,b])} = \frac{(b-a)^n}{2^{2n-1}}. \quad (37)$$

Перейдемо до питання мінімізації залишкового члена. Нагадаємо, що

$$|r_n(x)| = |f(x) - L_n(x)| \leq \frac{M_{n+1}}{(n+1)!} \cdot |\omega_n(x)|, \quad (38)$$

$$\text{де } M_{n+1} = \max_{x \in [a,b]} |f^{(n+1)}(x)|, \omega_n(x) = \prod_{i=0}^n (x - x_i) = x^{n+1} + \dots$$

Поставимо задачу

$$\inf_{\bar{P}_n(x)} \max_{x \in [a,b]} |\omega_n(x)|. \quad (39)$$

З теоремою Чебишова $\omega_n(x) = \bar{T}_{n+1}^{[a,b]}(x)$ поліном Чебишова. Якщо співпадають поліноми, то співпадають їх нулі. Отже: x_k — вузли інтерполяції співпадають з нулями багаточлена Чебишова

$$x_k = \frac{a+b}{2} - \frac{b-a}{2} \cdot t_k, \quad t_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right), \quad (40)$$

де $k = \overline{0, n}$.

В цьому випадку

$$|r_n(x)| = |f(x) - L_n(x)| \leq \frac{M_{n+1}}{(n+1)!} \cdot \frac{(b-a)^{n+1}}{2^{2n+1}}. \quad (41)$$

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

$$\max_{x \in [a,b]} |f(x) - L_n(x)| = \frac{(b-a)^{n+1}}{2^{2n+1}}. \quad (42)$$

6.5. Розділені різниці

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

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

$$f(x_i; x_j) = \frac{f(x_i) - f(x_j)}{x_i - x_j}. \quad (43)$$

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

$$f(x_i; x_j; x_k) = \frac{f(x_i; x_j) - f(x_j; x_k)}{x_i - x_k}. \quad (44)$$

Аналогічно визначаються розділені різниці довільного порядку.

Наведемо деякі властивості розділених різниць:

$$1. \quad f(x_0; \dots; x_n) = \sum_{i=0}^n \frac{f(x_i)}{\prod_{i \neq j} (x_i - x_j)}. \quad (45)$$

2. Розділена різниця — лінійний функціонал:

$$(\alpha_1 f_1 + \alpha_2 f_2)(x_0; x_1) = \alpha_1 f_1(x_0; x_1) + \alpha_2 f_2(x_0; x_1). \quad (46)$$

3. Розділена різниця — симетричний функціонал:

$$f(x_1; \dots; x_i; \dots; x_j; \dots; x_n) = f(x_1; \dots; x_j; \dots; x_i; \dots; x_n). \quad (47)$$

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

Задача 14: Довести першу властивість розділених різниць.

Таблиця розділених різниць має вигляд:

x_i	f_i	р.р.1	р.р.2	...	р.р.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	$f(x_2)$				
\vdots	\vdots	\vdots	\vdots	\vdots	
...	$f(x_0; \dots; x_n)$
\vdots	\vdots	\vdots	\vdots	\vdots	
x_{n-2}	$f(x_{n-2})$				
		$f(x_{n-2}; x_{n-1})$			
x_{n-1}	$f(x_{n-1})$		$f(x_{n-2}; x_{n-1}; x_n)$		
		$f(x_{n-1}; x_n)$			
x_n	$f(x_n)$				

6.6. Інтерполяційна формула Ньютона

Запишемо формулу Лагранжа інтерполяційного багаточлена

$$L_n(x) = \sum_{i=0}^n f(x_i) \cdot \frac{\omega_n(x)}{(x - x_i) \cdot \omega'_n(x_i)}, \quad (48)$$

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

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

$$L_n(x) = L_0(x) + (L_1(x) - L_0(x)) + \dots + (L_n(x) - L_{n-1}(x)), \quad (49)$$

і

$$L_j(x_i) = L_{j-1}(x_i) = f(x_i), \quad i = \overline{0, j-1}, \quad (50)$$

то

$$\Phi_j(x_i) = A_j \cdot (x - x_0) \cdot \dots \cdot (x - x_{j-1}), \quad (51)$$

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

$$\Phi_j(x) = \frac{f(x_j) - L_{j-1}(x_j)}{(x_j - x_0) \dots (x_j - x_{j-1})} \cdot (x - x_0) \dots (x - x_{j-1}). \quad (52)$$

Тоді

$$\begin{aligned} A_j &= \frac{f(x_j) - L_{j-1}(x_j)}{(x_j - x_0) \dots (x_j - x_{j-1})} = \frac{f(x_j)}{(x_j - x_0) \dots (x_j - x_{j-1})} - \\ &\quad - \sum_{i=0}^{j-1} \frac{f(x_i)}{(x_i - x_0) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_j - x_i)} = \\ &= \frac{f(x_j)}{(x_j - x_0) \dots (x_j - x_{j-1})} + \sum_{i=0}^{j-1} \frac{f(x_i)}{(x_i - x_0) \dots (x_i - x_j)} = \\ &= \sum_{i=0}^j \frac{f(x_i)}{(x_i - x_0) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_j)} = f(x_0; \dots; x_j). \end{aligned} \quad (53)$$

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

$$L_n(x) = f(x_0) + f(x_0; x_1)(x - x_0) + \dots + f(x_0; \dots; x_n)(x - x_0) \dots (x - x_{n-1}). \quad (54)$$

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

$$L_n(x) = f(x_n) + f(x_n; x_{n-1})(x - x_n) + \dots + f(x_n; \dots; x_0)(x - x_n) \dots (x - x_1). \quad (55)$$

Маємо рекурсію за степенем багаточлена

$$L_n(x) = L_{n-1}(x) + f(x_0; \dots; x_n)(x - x_0) \dots (x - x_1). \quad (56)$$

Звідси

$$L_n(x) = f(x_0) + (x - x_0)(f(x_0; x_1) + (x - x_1)(\dots + (x - x_{n-1})f(x_0; x_1; \dots; x_n) \dots)) \quad (57)$$

і цю формулу розкриваємо починаючи з середини (це аналог схеми Горнера обчислення значення багаточлена).

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

$$f(x; x_0; \dots; x_n) = \frac{f(x)}{(x - x_0) \dots (x - x_n)} + \sum_{k=0}^n \frac{f(x_k)}{\prod_{i \neq k} (x - x_i)}. \quad (58)$$

Звідси

$$\begin{aligned} f(x) &= f(x_0) \cdot \frac{(x - x_1) \dots (x - x_n)}{(x_0 - x_1) \dots (x_0 - x_n)} + \dots + f(x_n) \cdot \frac{(x - x_1) \dots (x - x_{n-1})}{(x_n - x_1) \dots (x_n - x_{n-1})} + \\ &\quad + f(x; x_0; \dots; x_n)(x - x_0) \dots (x - x_n) = L_n(x) + f(x; x_0; \dots; x_n) \cdot \omega_n(x). \end{aligned} \quad (59)$$

Тоді похибка має вигляд

$$r_n(x) = f(x) - L_n(x) = f(x; x_0; \dots; x_n) \cdot \omega_n(x). \quad (60)$$

Це нова форма для залишкового члена.

Порівнюючи з формулою залишкового члена в пункті 6.3, маємо

$$f(x; x_0; \dots; x_n) = \frac{f^{(n+1)}(\xi)}{(n+1)!}, \quad (61)$$

що доводить останню властивість розділених різниць.

Нехай маємо сітку рівновіддалених вузлів: $x_i = a + ih$, $h = \frac{b-a}{n}$, $i = \overline{0, n}$, $x_0 = a$, $x_n = b$. Розначимо $\Delta f_0 = f_1 - f_0$, $\Delta^2 f_0 = \Delta f_1 - \Delta f_0 = f_2 - 2f_1 + f_0$, \dots — скінченні різниці.

Запишемо формули Ньютона у нових позначеннях:

$$L_n(x) = L_n(x_0 + th) = f_0 + t\Delta f_0 + \dots + \frac{t(t-1)\dots(t-n+1)}{n!} \cdot \Delta^n f_0, \quad (62)$$

де $t = \frac{x-x_0}{h}$.

Це інтерполяційна формула Ньютона вперед по рівновіддалених вузлах.

Задача 15: Побудувати інтерполяційну формулу Ньютона назад по рівновіддалених вузлах.

6.7. Інтерполювання з кратними вузлами

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 кратним вузлах, де

$$m = \sum_{i=1}^n k_i - 1. \quad (63)$$

Якщо $k_i = 1$, то $H_m(x) = L_n(x)$.

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

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

$$\begin{aligned} L_m^\varepsilon = & f(x_{0,0}^\varepsilon) + f(x_{0,0}^\varepsilon; x_{0,1}^\varepsilon) (x - x_{0,0}^\varepsilon) + \dots \\ & + f(x_{0,0}^\varepsilon; \dots; x_{n,k_n-1}^\varepsilon) (x - x_{0,0}^\varepsilon) \dots (x - x_{n,k_n-1}^\varepsilon). \end{aligned} \quad (64)$$

При $\varepsilon \rightarrow 0$ маємо $x_{i,j}^\varepsilon \rightarrow x_i$. Крім того

$$f(x_{i,0}^\varepsilon; \dots; x_{i,k_i-1}^\varepsilon) = f(x_i; \dots; x_i) = \frac{f^{(k_i)}(x_i)}{k_i!}. \quad (65)$$

Тому $L_m^\varepsilon(x) \rightarrow H_m(x)$ та

$$R_m(x) = f(x) - H_m(x) = \frac{f^{(m+1)}(\xi)}{(m+1)!} \cdot \Omega_m(x), \quad (66)$$

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

6.8. Збіжність процесу інтерполювання

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

Введемо норму

$$|f(x) - L_n|_{C([a,b])} = \max_{x \in [a,b]} |f(x) - L_n(x)|. \quad (67)$$

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

$$|f(x) - L_n(x)|_{C([a,b])} \leq \frac{M_{n+1}}{(n+1)!} |\omega_n(x)|_{C([a,b])}, \quad (68)$$

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

А яка оцінка буде для довільної неперервної функції?

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

$$\forall \{x_i\}_{i=1}^n : h = \max_{i=1,n} \rightarrow 0 : \lim_{n \rightarrow \infty} L_n(x) = f(x), \quad (69)$$

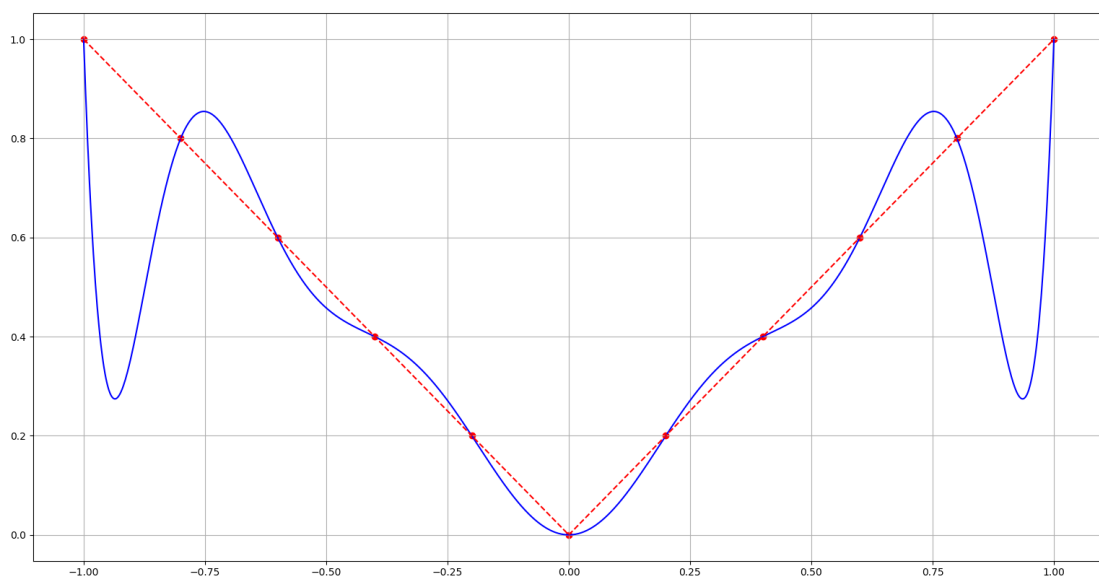
де, як завжди, $h_i = x_i - x_{i-1}$.

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

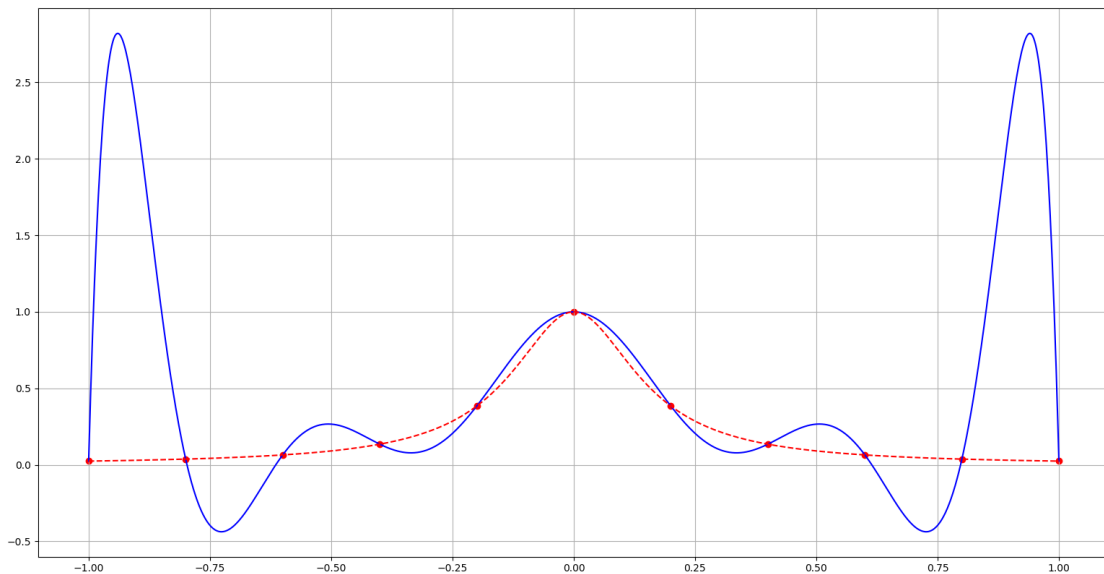
Розглянемо приклади поведінки інтерполяційних багаточленів при $n \rightarrow \infty$ для деяких функцій.

Приклад 1: Послідовність інтерполяційних багаточленів (сітка рівномірна), побудованих для неперервної функції $f(x) = |x|$, $-1 \leq x \leq 1$ (функція неперервна, але негладка), не збігається на $x \in [-1, 1]$, крім точок $x = -1, 0, 1$.

На рисунку дано графіки самої функції (штрихова лінія) та інтерполяційного багаточлена (суцільна лінія) на рівномірній сітці $x_i = -1 + ih$, $h = 2/n$, $i = 0, n$ для $n = 10$:

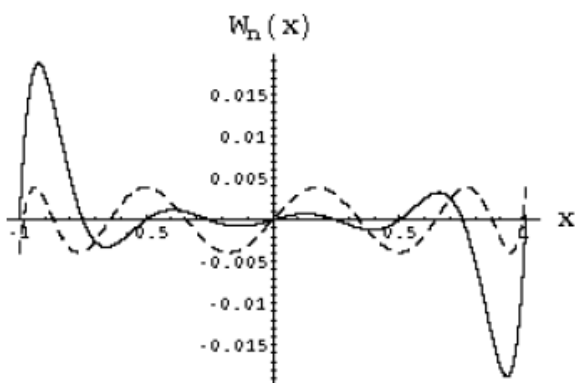


Приклад 2: Функція Рунге $f(x) = \frac{1}{1+40x^2}$, $-1 \leq x \leq 1$ (функція аналітична!). Для рівномірної сітки $x_i = -1 + ih$, $h = 2/n$, $i = \overline{0, n}$ маємо графіки: суцільна лінія — інтерполяційного багаточлена; пунктирна — самої функції для $n = 10$:



Пояснити чому рівномірна сітка дає великі похибки інтерполювання біля кінців інтервалу інтерполювання допомагає наступний рисунок. На цьому рисунку суцільною лінією представлено графік функції

$\omega_n = \prod_{i=0}^n (x - x_i)$ ($n = 8$) для рівномірної сітки. Як бачимо максимальні за модулем значення цієї функції припадають на кінці інтервалу.



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

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

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

Теорема: Стала Лебега

$$|P_n| = \max_j \sum_{j=0}^n |\varphi_j^{(n)}(x)|, \quad (70)$$

де

$$\varphi_j^{(n)}(x) = \frac{\omega_n(x)}{(x - x_j) \cdot \omega'_n(x_j)}. \quad (71)$$

Теорема: Для $f(x) \in C([a, b])$:

$$|f(x) - L_n(x)|_{C([a,b])} \leq (1 + |P_n|) \cdot E_n(f), \quad (72)$$

де

$$E_n(f) = \inf_{Q_n(x)} |f(x) - Q_n(x)|_{C([a,b])} \quad (73)$$

— відхилення багаточлена n -го степеня найкращого рівномірного наближення від $f(x)$.

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

$$|P_n^E| \approx C_1 \cdot 2^n, \quad |P_n^T| \approx C_2 \cdot \ln(n). \quad (74)$$

Останні оцінки пояснюють розбіжність процесу інтерполювання при великих n .

6.9. Кульково-лінійна інтерполяція

Інтерполяція багаточленом Лагранжа або Ньютона на відрізок $[a, b]$ 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:

$$\Phi_1(x) = L_1^i(x) = f(x_{i-1}) \cdot \frac{x - x_{i-1}}{x_i - x_{i-1}} + f(x_i) \cdot \frac{x_i - x}{x_i - x_{i-1}}, \quad (75)$$

де $x \in [x_{i-1}, x_i]$.

Представимо її у вигляді

$$\Phi_1(x) = \sum_{i=0}^n f(x_i) \cdot \varphi_i(x). \quad (76)$$

З умов інтерполювання маємо

$$\Phi_1(x_j) = \sum_{i=0}^n f(x_i) \cdot \varphi_i(x_j) = f(x_j). \quad (77)$$

Звідси

$$\varphi_i(x_j) = \delta_{i,j} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}. \quad (78)$$

Значить

$$\varphi_i(x) = \begin{cases} 0, & a \leq x \leq x_{i-1} \\ \frac{x-x_{i-1}}{x_i-x_{i-1}}, & x_{i-1} \leq x \leq x_i \\ \frac{x_{i+1}-x}{x_{i+1}-x_i}, & x_i \leq x \leq x_{i+1} \\ 0, & x_{i+1} \leq x \leq b \end{cases} \quad (79)$$

Теорема: Для довільної $f(x) \in C^{(2)}([a, b])$ справедлива оцінка

$$|f(x) - \Phi_1(x)|_{C([a,b])} \leq \frac{M_2}{8} \cdot |h|^2, \quad (80)$$

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

Доведення. Маємо для $x \in [x_{i-1}, x_i]$:

$$z(x) = f(x) - \Phi_1(x) = f(x) - L_1^i(x) = \frac{f''(\xi_i)}{2!} \cdot (x - x_{i-1}) \cdot (x - x_i). \quad (81)$$

Звідси

$$|f(x) - \Phi_1(x)| \leq \frac{M_2^i}{2} \cdot |(x - x_{i-1})(x - x_i)| \leq \frac{M_2^i \cdot h_i^2}{8}, \quad (82)$$

де

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

Остання оцінка отримана з нерівності

$$\max_{[x_{i-1}, x_i]} |(x - x_{i-1}) \cdot (x - x_i)| = \frac{h_i^2}{4}. \quad (83)$$

Тоді

$$\max_{i=\overline{1, n}} \max_{x \in [x_{i-1}, x_i]} |z(x)| \leq \frac{h^2 M_2}{8}, \quad (84)$$

де $M_2 = \max_{x \in [a, b]} |f''(x)|$, $h_i = \max_i h_i$, що доводить (80). \square

Задача 16: Довести оцінку $|f'(x) - \Phi'_1(x)| \leq |h| \cdot M_2$.

Отже маємо збіжність процесу інтерполявання за допомогою кусково-лінійної функції

$$\left| f(x) - \Phi_1^{(n)}(x) \right|_{C([a,b])} \xrightarrow{h \rightarrow 0, n \rightarrow \infty} 0, \quad (85)$$

тобто

$$\left\{ \Phi_1^{(n)}(x) \right\} \implies f(x). \quad (86)$$

Розглянемо деякі простори:

1. $H_0 = L_2([a, b])$ — гільбертів простір, в якому скалярний добуток визначається так:

$$\langle u, v \rangle = \int_a^b (u(x) \cdot v(x)) \, dx \quad (87)$$

а норма $\|u\|_0 = \sqrt{\langle u, u \rangle}$.

2. $H_k = W_2^k([a, b])$. Тепер скалярний добуток

$$\langle u, v \rangle_k = \sum_{m=0}^k \int_a^b (u^{(m)}(x) \cdot v^{(m)}(x)) \, dx, \quad (88)$$

а норма $\|u\|_k = \sqrt{\|u^{(0)}\|^2 + \dots + \|u^{(k)}\|^2}$.

Теорема: Нехай $f(x) \in H_2 = W_2^2([a, b])$. Тоді

$$\left| f^{(k)} - \Phi_1^{(k)} \right|_0 \leq |h|^{2-k} \cdot |f|_2, \quad (89)$$

де $k = 1, 2$.

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: $y_i = f(x_i)$, $y'_i = f'(x_i)$, $i = 0, n$. Позначимо через $\Phi_3(x)$ функцію, яка буде інтерполювати задану. Тоді

$$\Phi_3(x) = H_3^i(x), \quad x \in [x_{i-1}, x_i]. \quad (90)$$

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

x_i	y_i			
		y'_i		
x_i	y_i		$\frac{y_{i-1,i} - y'_i}{h_i}$	
		$y_{i-1,i}$		$\frac{y'_i - 2y_{i-1,i} + y'_{i-1}}{h_i^2}$
x_{i-1}	y_{i-1}		$\frac{y'_{i-1} - y_{i-1,i}}{h_i}$	
		y'_{i-1}		
x_{i-1}	y_{i-1}			

$$H_3^i(x) = y_i + y'_i(x - x_i) + \frac{y_{i-1,i} - y'_i}{h_i} \cdot (x - x_i)^2 + \frac{y'_i - 2y_{i-1,i} + y'_{i-1}}{h_i^2} \cdot (x - x_i)^2 \cdot (x - x_{i-1}) \quad (91)$$

Можна представити кусково-кубічну функцію і в такому вигляді:

$$\Phi_3(x) = \sum_{i=0}^n (y_i \cdot \varphi_i^0(x) + y'_i \cdot \varphi_i^1(x)). \quad (92)$$

Умови інтерполявання: $\Phi_3(x_i) = y_i$, $\Phi'_3(x_i) = y'_i$, $i = \overline{0, n}$. Якщо ці умови підставити в \u{yiref{eq:6.10.2}}, то отримаємо умови на базисні функції:

$$\varphi_i^0(x_j) = \delta_{i,j}, \quad (93)$$

$$(\varphi_i^0)'(x_j) = 0, \quad (94)$$

$$\varphi_i^1(x_j) = 0, \quad (95)$$

$$(\varphi_i^1)'(x_j) = \delta_{i,j}. \quad (96)$$

для $i, j = \overline{0, n}$.

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 $s = \frac{x - x_i}{h}$, $x \in [x_{i-1}, x_i] \implies s \in [-1, 0]$.

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

$$\bar{\varphi}_1^0(0) = 1, \quad (97)$$

$$\bar{\varphi}_1^0(1) = 0, \quad (98)$$

$$(\bar{\varphi}_1^0)'(0) = 0, \quad (99)$$

$$(\bar{\varphi}_1^0)'(1) = 0. \quad (100)$$

Її явний вигляд отримаємо за допомогою таблиці розділених різниць:

0	1			
		0		
0	1		-1	
		-1		2
1	0		1	
		0		
1	0			

$$H_3(s) = 1 + 0 \cdot s - 1 \cdot s^2 + 2s^2(s - 1) = 2s^3 - 3s^2 + 1 \equiv \bar{\varphi}_1^0(s). \quad (101)$$

Аналогічно

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

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

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

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

$$\varphi_i^0(x) = \begin{cases} 0, & a \leq x \leq x_{i-1}, \\ -2s^3 - 3s^2 + 1, & x_{i-1} \leq x \leq x_i, \\ 2s^3 - 3s^2 + 1, & x_i \leq x \leq x_{i+1}, \\ 0, & x_{i+1} \leq x \leq b, \end{cases} \quad (102)$$

i

$$\varphi_i^1(x) = \begin{cases} 0, & a \leq x \leq x_{i-1}, \\ hs(s+1)^2, & x_{i-1} \leq x \leq x_i, \\ hs(s-1)^2, & x_i \leq x \leq x_{i+1}, \\ 0, & x_{i+1} \leq x \leq b, \end{cases} \quad (103)$$

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

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

$$f(x) - \Phi_3(x) = f(x) - H_3^i(x) = \frac{f^{(4)}(\xi)}{4!} \cdot (x - x_{i-1})^2(x - x_i)^2. \quad (104)$$

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 $\bar{x}_i = \frac{x_i + x_{i-1}}{2}$ тому для модуля похибки маємо:

$$|f(x) - \Phi_3(x)| \leq \frac{M_4^i}{24} \left(\frac{h^2}{4}\right)^2 = \frac{M_4^i h^4}{384}, \quad (105)$$

$$|f(x) - \Phi_3(x)|_{C([a,b])} \leq \frac{M_4 h^4}{384}. \quad (106)$$

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

$$\Phi_3(x) = \sum_{i=0}^n \left(y_i \varphi_i^{(0)}(x) + y'_i \varphi_i^{(1)}(x) \right) \quad (107)$$

має місце оцінка

$$|f(x) - \Phi_3(x)|_{C([a,b])} \leq \frac{M_4 h^4}{384}. \quad (108)$$

А для похідної

$$|f'(x) - \Phi'_3(x)|_{C([a,b])} \leq M \cdot M_4 h^3, \quad (109)$$

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

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

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

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

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

1. Кубічність:

$$s(x) \in \pi_3, \quad x \in [x_{i-1}, x_i], \quad i = \overline{1, n} \quad (110)$$

2. Дефект 1:

$$s(x) \in C^{(2)}([a, b]) \quad (111)$$

3. Інтерполуює $f(x)$:

$$s(x_i) = f(x_i), \quad i = \overline{0, n} \quad (112)$$

4. Природній:

$$s''(a) = s''(b) = 0. \quad (113)$$

Умови (113), так звані *умови природності*, необхідні, щоб разом було $4n$ умов для знаходження $4n$ коефіцієнтів сплайну. Замість них можуть бути такі умови:

$$s''(a) = A, \quad s''(b) = B \quad (4.a)$$

$$s'(a) = A, \quad s'(b) = B \quad (4.b)$$

$$s(a) = s(b), \quad s'(a) = s'(b), \quad s''(a) = s''(b) \quad (4.c)$$

Умови (4.c) — це так звані умови періодичності.

Побудуємо природній сплайн. З (110) та (111) маємо

$$s''(x) = m_{i-1} \cdot \frac{x_i - x}{h_i} + m_i \cdot \frac{x - x_{i-1}}{h_i}, \quad (114)$$

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

Двічі інтегруючи (114), маємо

$$s(x) = m_{i-1} \cdot \frac{(x_i - x)^3}{6h_i} + m_i \cdot \frac{(x - x_{i-1})^3}{6h_i} + \left(f_{i-1} - \frac{m_{i-1}h_i^2}{6}\right) \cdot \frac{x_i - x}{h_i} + \left(f_i - \frac{m_i h_i^2}{6}\right) \cdot \frac{x - x_{i-1}}{h_i}, \quad (115)$$

для $x \in [x_{i-1}, x_i]$.

З (113) маємо $m_0 = m_n = 0$.

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

$$\begin{cases} \frac{h_i m_{i-1}}{6} + \frac{(h_i + h_{i+1})m_i}{3} + \frac{h_{i+1} m_{i+1}}{6} = \\ = \frac{f_{i+1} - f_i}{h_i} - \frac{f_i - f_{i-1}}{h_i}, \quad i = \overline{1, n-1}, \\ m_0 = m_n = 0. \end{cases} \quad (116)$$

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

$$|f^{(k)}(x) - s^{(k)}(x)|_{C([a,b])} \leq M_4 |h|^{4-k}, \quad (117)$$

де $k = 0, 1, 2$ і $M_4 = \max_{[a,b]} |f^{(4)}(x)|$, $|h| = \max_i h_i$.

Введемо клас функцій

$$U = \left\{ u(x) : u(x) \in W_2^2([a, b]), u(x_i) = f_i, i = \overline{0, n} \right\} \quad (118)$$

— це функції досить гладкі і приймають задані значення. Якщо ввести такий функціонал

$$\Phi(u) = \int_a^b (u''(x))^2 dx, \quad (119)$$

то

$$\Phi(s) = \inf_{u \in U} \Phi(u), \quad (120)$$

де $s(x)$ — кубічний природний 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 інтерполують значення $f(x_i)$.

Для того, щоб не розв'язувати СЛАР (116) інколи будують наближення до сплайну $\tilde{s}(x)$, яке отримується заміною $m_i = s''(x_i)$ на

$$f_{\tilde{x}, \hat{x}, i} \equiv \frac{1}{\bar{h}_i} \left(\frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i} \right) \approx f''(x_i) \approx s''(x_i), \quad (121)$$

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

Зауваження 1: Складемо матрицю A розмірності $(n-1) \times (n-1)$:

$$A = \begin{pmatrix} \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & \dots & 0 \\ \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & \ddots & \vdots \\ 0 & \frac{h_3}{6} & \frac{h_3+h_4}{3} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \dots & 0 & \frac{h_{n-1}}{6} & \frac{h_{n-1}+h_n}{3} \end{pmatrix} \quad (122)$$

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

$$H = \begin{pmatrix} \frac{1}{h_1} & -\left(\frac{1}{h_1} + \frac{1}{h_2}\right) & \frac{1}{h_2} & 0 & \dots & 0 \\ 0 & \frac{1}{h_2} & -\left(\frac{1}{h_2} + \frac{1}{h_3}\right) & \frac{1}{h_3} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{h_{n-1}} & -\left(\frac{1}{h_{n-1}} + \frac{1}{h_n}\right) & \frac{1}{h_n} \end{pmatrix} \quad (123)$$

Тоді можна записати СЛАР (116) відносно моментів $\vec{m} = (m_1, m_2, \dots, m_{n-1})$ вигляді:

$$A\vec{m} = F\vec{f}, \quad (124)$$

де

$$\vec{f} = (f_0, f_1, \dots, f_n)^T \quad (125)$$

Зауваження 2: Нагадаємо формулу для інтерполяційного багаточлена Лагранжа

$$L_n(x) = \sum_{i=0}^n f(x_i) \Phi_i^{(n)}(x), \quad (126)$$

де $\Phi_i^{(n)}$ — множники Лагранжа. Це представлення інтерполяційного багаточлена Лагранжа по системі функцій $\{\Phi_i^{(n)}\}$. Для

$$\Phi_1 = \sum_{i=1}^n f(x_i) \varphi_i(x) \quad (127)$$

маємо представлення по системі кусково-лінійних функцій $\{\varphi_i(x)\}$. Для

$$\Phi_3(x) = \sum_{i=1}^n (f(x_i) \varphi_i^0(x) + f'(x_i) (\varphi_i^1)'(x)) \quad (128)$$

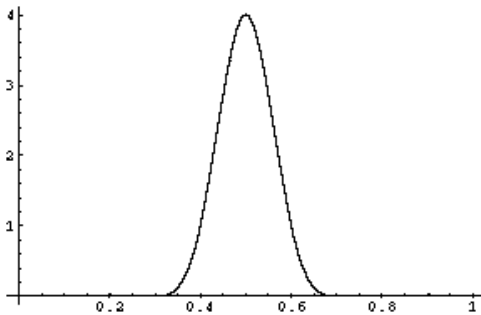
— представлення по системі $\{\varphi_i^0, (\varphi_i^1)'\}$.

Аналогічно, якщо представити кубічний сплайн у вигляді

$$s_3(x) = \sum_{i=0}^n c_i B_3^i(x), \quad (129)$$

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.:

$$B_3^i(z) = \frac{1}{6h} \begin{cases} \left(\frac{z-x_{i-2}}{h}\right)^3, & z \in [x_{i-2}, x_{i-1}]; \\ -3\left(\frac{z-x_{i-1}}{h}\right)^3 + 3\left(\frac{z-x_{i-1}}{h}\right)^2 + 3\left(\frac{z-x_{i-1}}{h}\right) + 1, & z \in [x_{i-1}, x_i]; \\ -3\left(\frac{x_{i+1}-z}{h}\right)^3 + 3\left(\frac{x_{i+1}-z}{h}\right)^2 + 3\left(\frac{x_{i+1}-z}{h}\right) + 1, & z \in [x_i, x_{i+1}]; \\ \left(\frac{x_{i+2}-z}{h}\right)^3, & z \in [x_{i+1}, x_{i+2}]; \\ 0, & z < x_{i-2} \vee x_{i+2} < z. \end{cases} \quad (130)$$



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

Для знаходження коефіцієнтів c_i записується СЛАР з умов інтерполявання.

6.12. Параметричні сплайни

На практиці часто виникає задача побудови кривої по заданим точкам $(x_i, y_i)_{i=1}^n$. В цьому випадку використовують сплайни. Якщо відповідна функція $y = f(x)$ однозначна, то сплайн будується за алгоритмом з попереднього пункту.

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

В цьому випадку відповідна функція задається параметрично

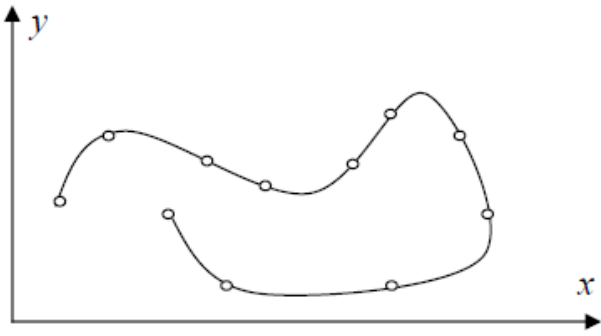
$$x = x(t), \quad y = y(t), \quad t \in [A, B]. \quad (131)$$

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 $s_y(t_i) = y_i, i = \overline{1, n}$.

Означення: Тоді параметрична функція

$$(s_x(t), s_y(t)), \quad t \in [A, B]. \quad (132)$$

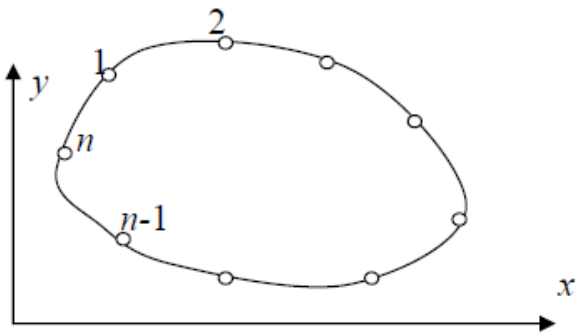
називається *параметричним сплайном* для функції (132).



Стає питання про вибір параметру t . Нехай $t_i = i, i = \overline{1, n}$, тобто для табличних даних $(x_i, y_i)_{i=1}^n$ 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 попередньому пункті 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:



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

$$\forall t: \quad x(t) = x(t + n), \quad y(t) = y(t + n). \quad (133)$$

Наприклад, для значень в точках маємо:

$$x_1 = x_{n+1}, \quad y_1 = y_{n+1}; \quad x_0 = x_n, \quad y_0 = y_n. \quad (134)$$

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

$$\begin{aligned} s(t) = & \frac{m_{i-1}(t_i - t)^3}{6h_i} + \frac{m_i(t - t_{i-1})^3}{6h_i} + \\ & + \left(f_{i-1} - \frac{m_{i-1}h_i^2}{6} \right) \frac{t_i - t}{h_i} + \\ & + \left(f_i - \frac{m_ih_i^2}{6} \right) \frac{t - t_{i-1}}{h_i}, \end{aligned} \quad (135)$$

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

$$\begin{cases} \frac{h_i m_{i-1}}{6} + \frac{(h_i + h_{i+1})m_i}{3} + \frac{h_{i+1}m_{i+1}}{6} = \\ = \frac{f_{i+1} - f_i}{h_i} - \frac{f_i - f_{i-1}}{h_i}, \quad i = \overline{1, n}, \\ m_0 = m_n, \quad m_1 = m_{n+1}, \end{cases} \quad (136)$$

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

Системі (74) відповідає матриця розмірності $(n \times n)$:

$$A = \begin{pmatrix} \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & \cdots & 0 & \left\langle \frac{h_1}{6} \right\rangle \\ \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & \ddots & & 0 \\ 0 & \frac{h_3}{6} & \frac{h_3+h_4}{3} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & \ddots & \ddots & \frac{h_n}{6} \\ \left\langle \frac{h_1}{6} \right\rangle & 0 & \cdots & 0 & \frac{h_n}{6} & \frac{h_n+h_1}{3} \end{pmatrix} \quad (137)$$

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:

$$\begin{cases} a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \quad i = \overline{1, n}, \\ y_0 = y_n, \quad y_{n+1} = y_1, \quad a_1 = a_n, \quad b_{n+1} = b_1, \end{cases} \quad (138)$$

Формули методу [ЛМС, стор. 391–392]:

1. $\alpha_2 = b_1/c_1, \beta_2 = f_1/c_1, \gamma_2 = a_1/c_1$;
2. $z_i = c_i - a_i \alpha_i; \alpha_{i+1} = b_i/z_i; \beta_{i+1} = (f_i + a_i \beta_i)/z_i; \gamma_{i+1} = a_i \gamma_i/z_i, i = \overline{2, n}$;
3. $p_{n-1} = \beta_n; q_{n-1} = \alpha_n + \gamma_n$;
4. $p_i = \alpha_{i+1} p_{i+1} + \beta_{i+1}; q_i = \alpha_{i+1} q_{i+1} + \gamma_{i+1}, i = \overline{n-2, 1}$;
5. $y_n = (\beta_{n+1} + a_{n+1} p_1)/(1 - \alpha_{n+1} q_1 - \gamma_{n+1})$;
6. $y_i = p_i + y_n q_i, i = \overline{1, n-1}$

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 (74) 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

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

$$r_1^i(x) = f(x) - L_1^i(x) = \frac{f''(\xi)}{2!} \cdot (x - x_{i-1})(x - x_i). \quad (139)$$

Тоді

$$|r_1^i(x)| \leq \frac{M_2^i}{2} \cdot |(x - x_{i-1})(x - x_i)| \leq \frac{M_2^i h^2}{8}, \quad x \in [x_{i-1}, x_i]. \quad (140)$$

Таким чином

$$|f(x) - L_1^i(x)|_{C([a,b])} \leq \frac{M_2 h^2}{8}. \quad (141)$$

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

Для того, щоб похибка була меншою за ε потрібно вибрати

$$h \leq \sqrt{\frac{8\varepsilon}{M_2}}. \quad (142)$$

Аналогічно, для квадратичного інтерполювання маємо

$$|f(x) - L_2^i(x)|_{C([a,b])} \leq \frac{M_3 h^3}{9\sqrt{3}} < \varepsilon. \quad (143)$$

Звідси

$$h \leq \sqrt[3]{\frac{9\sqrt{3}\varepsilon}{M_3}}. \quad (144)$$

2. Розв'язування рівнянь. Нехай необхідно розв'язати рівняння

$$f(x) = \bar{y}. \quad (145)$$

При $\bar{y} = 0$ маємо рівняння $f(x) = 0$. Нехай \bar{x} корінь рівняння (145).

1. 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 $\{y_i, x_i\}_{i=0}^n$ де y_i вважаються значеннями аргументу, а x_i — значеннями оберненої функції. Тоді наближення до кореня є $x^* = L_n(y)$.

Оцінимо похибку:

$$|\bar{x} - x^*| = |x(\bar{y}) - L_n(\bar{y})| \leq \frac{M_{n+1}}{(n+1)!} |\omega_n(\bar{y})|, \quad (146)$$

$$\text{де } M_{n+1} = \max_{y_{\min} \leq y \leq y_{\max}} \left| \frac{d^{n+1}}{dy^{n+1}} x(y) \right|, \quad |\omega_n(y)| = (y - y_0) \dots (y - y_n).$$

Недоліком методу є складність обчислення похідних старших порядків оберненої функції.

2. Пряме інтерполювання. Нехай знову відомі $y_i = f(x_i)$, $x_i \in [a, b]$. Тоді замість рівняння (145) розв'язуємо рівняння

$$L_n(x^*) = y, \quad (147)$$

де $L_n(x)$ інтерполяційний багаточлен по значенням $\{x_i, y_i\}_{i=0}^n$.

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:

$$f(x^*) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \cdot \omega_n(x). \quad (148)$$

Далі $f(x^*) - y = f(x^*) - f(\bar{x})$, звідки

$$|f(x^*) - f(\bar{x})| \leq \frac{M_{n+1}}{(n+1)!} \cdot |\omega_n(x)|. \quad (149)$$

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

За теоремою Лагранжа $f(x^*) - f(\bar{x}) = f'(\eta)(x^* - \bar{x})$.

Припустимо, що $f'(x) \neq 0$. Це означає, що на проміжку $[a, b]$ функція $f(x)$ монотонна. Звідси

$$|x^* - \bar{x}| \leq \frac{|f(x^*) - f(\bar{x})|}{\min_{x \in [a, b]} |f'(x)|} \leq \frac{M_{n+1}}{\min_{x \in [a, b]} |f'(x)|} \cdot \frac{|\omega_n(x)|}{(n+1)!}. \quad (150)$$

3. 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, $\lambda \in [-\|A\|_k, \|A\|_k]$, де $k = 1$ або $k = \infty$) деякі точки $\lambda_i, i = \overline{0, n}$. За допомогою методу Гауса 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 розмірностей $n \leq 10$ матриці A .

6.14. Тригонометрична інтерполяція

Інтерполяція відбувається за системою функцій

$$1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots, \sin(kx), \cos(kx), \dots \quad (151)$$

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

$$T_n(x) = \sum_{i=0}^{2n} f(x_i) \Phi_i(x), \quad (152)$$

де

$$\Phi_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{2n} \frac{\sin\left(\frac{x-x_j}{2}\right)}{\sin\left(\frac{x_i-x_j}{2}\right)}, \quad (153)$$

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

$$\Phi_i(x_k) = \prod_{\substack{j=0 \\ j \neq i}}^{2n} \frac{\sin\left(\frac{x_k-x_j}{2}\right)}{\sin\left(\frac{x_i-x_j}{2}\right)}, \quad (154)$$

для $i \neq k$, а

$$\Phi_i(x_i) = \prod_{\substack{j=0 \\ j \neq i}}^{2n} \frac{\sin\left(\frac{x_i-x_j}{2}\right)}{\sin\left(\frac{x_i-x_j}{2}\right)} = 1. \quad (155)$$

Таким чином за допомогою формули (152) 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:

$$T_n(x) = \sum_{i=0}^n f(x_i) \prod_{\substack{j=0 \\ j \neq i}}^n \frac{\cos(x) - \cos(x_i)}{\cos(x_i) - \cos(x_j)}. \quad (156)$$

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:

$$T_n(x) = \sum_{i=1}^n f(x_i) \frac{\sin(x)}{\sin(x_i)} \prod_{\substack{j=1 \\ j \neq i}}^n \frac{\cos(x) - \cos(x_i)}{\cos(x_i) - \cos(x_j)}. \quad (157)$$

Задача 20: Показати, що тригонометричні багаточлени (156), (157) є інтерполуючими для функції $f(x)$. Яке значення функції інтерполуює (157) при $x = 0$? Чому?

6.15. Двовимірна інтерполяція

Побудова багаточлена для функції від двох змінних $z = f(x, y)$, що інтерполуює 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

$$P_n(x, y) = \sum_{0 \leq k+m \leq n} a_{k,m} x^k y^m, \quad (158)$$

має $N = \frac{(n+1)(n+2)}{2}$ коефіцієнтів $a_{k,m}$. Тому необхідно задати значення в точках $A_i(x_i, y_i)_{i=1}^N$.

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

$$P_n(x_i, y_i) = \sum_{0 \leq k+m \leq n} a_{k,m} x_i^k y_i^m = z_i, \quad (159)$$

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

$\begin{equation}$

$\end{equation}$

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:

$$\bar{\Omega} = (x, y) : 0 \leq x \leq L_1, 0 \leq y \leq L_2. \quad (160)$$

Введемо сітку

$$x_i = ih_x, \quad h_x = L_1/N, \quad i = \overline{0, N}; \quad y_j = jh_y, \quad h_y = L_2/M, \quad j = \overline{0, M} \quad (161)$$

Тоді інтерполяційний багаточлен має вигляд

$$P(x, y) = \sum_{i=0}^N \sum_{j=0}^M f(x_i, y_j) \prod_{\substack{p=0 \\ p \neq i}}^N \prod_{\substack{q=0 \\ q \neq j}}^M \left(\frac{x - x_i}{x_p - x_i} \cdot \frac{y - y_j}{y_q - y_j} \right) \quad (162)$$

Розглянемо випадок, коли $N = M = 1$. Тоді

$$\begin{aligned} P_{1,1}(x, y) &= f(x_0, y_0) \cdot \frac{x - x_1}{x_0 - x_1} \cdot \frac{y - y_1}{y_0 - y_1} + f(x_0, y_1) \cdot \frac{x - x_1}{x_0 - x_1} \cdot \frac{y - y_0}{y_1 - y_0} + \\ &+ f(x_1, y_0) \cdot \frac{x - x_0}{x_1 - x_0} \cdot \frac{y - y_1}{y_0 - y_1} + f(x_1, y_1) \cdot \frac{x - x_0}{x_1 - x_0} \cdot \frac{y - y_0}{y_1 - y_0} = \\ &= a_0 + a_1 x + a_2 y + a_{1,2} xy. \end{aligned} \quad (163)$$

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

Формула (162) являє собою приклад інтерполювання на всій області. В одновимірному випадку при великих степенях багаточлена отримують погане наближення через розбіжність процесу інтерполювання. Так же 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 $\Omega \subset \mathbb{R}^2$ — багатокутник в площині. Представимо її у вигляді

$$\Omega = \bigsqcup_{i=1}^n K_i, \quad (164)$$

де $K_i \in T_h$.

Означення: T_h називається «триангуляцією» області Ω ,

а K_i — багатокутники з непорожньою внутрішністю що не мають спільних внутрішніх точок, причому $\text{diam } K_i \leq h$, де 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 $c_h \in X_h$. Причому звуження цієї функції на область K_i , тобто $v_h|_{K_i}$ є поліномом.

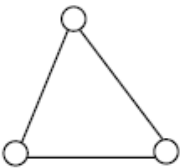
Позначимо $\Pi_k, k \geq 0$ — простір багаточленів степеня k по сукупності змінних x, y ; його розмірність $\dim \Pi_k = \frac{(k+1)(k+2)}{2}$. Нехай $\Theta_k, k \geq 0$ — простір багаточленів степеня по кожній окремії змінній x, y ; його розмірність $\dim \Theta_k = (k+1)^2$.

Наприклад, $P_1(x, y) = a_0 + a_1x + a_2y \in \Pi_1$ — поліном степеня 1 по x, y , а $Q_1(x, y) = a_0 + a_1x + a_2y + a_{1,2}xy \in \Theta_1$ лінійна по кожній окремії змінній.

Позначимо через $X_h^k = \{v_h \in C^0(\Omega) : v_h|_k, \forall k \in T_k\}$ — простір інтерполантів при розбитті області на трикутники, а $Y_h^k = \{v_h \in C^0(\Omega) : v_h|_k \in \Theta_k, \forall k \in T_h\}$ — 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 :

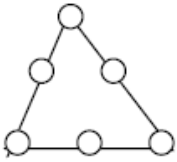
$$\begin{vmatrix} 1 & x & y & z \\ 1 & x_1 & y_1 & z_1 \\ 1 & x_2 & y_2 & z_2 \\ 1 & x_3 & y_3 & z_3 \end{vmatrix} = 0. \quad (165)$$

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 $A_i = (x_i, y_i), i = 1, 2, 3$.

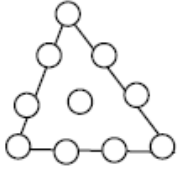
Приклад 2: Для $X_h^2, k = 2, \dim \Pi_2 = 6$.

Треба задати 6 значень, щоб забезпечити однозначність наближення. Тому вибираємо точки інтерполювання так:



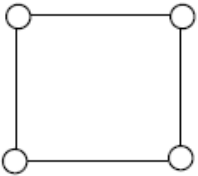
Приклад 3: X_h^2 , $k = 3$, $\dim \Pi_3 = 10$.

Потрібно задати 10 точок, як на наступному малюнку:

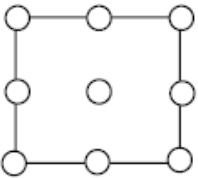


Приклад 4: Y_h^1 , $k = 1$, $\dim \Theta_1 = 4$.

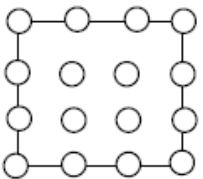
Формула для $Q_1(x, y)$ наведена в (??). Точки:



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



Приклад 6: Y_h^3 , $k = 3$, $\dim \Theta_4 = 16$.



Нехай $X = W_2^m(\Omega) = H^m(\Omega)$ — це простір з нормою

$$\|v\|_m^2 = \sum_{k=0}^m \|v_k\|^2, \quad (166)$$

де

$$|v_k|^2 = \int_{\Omega} (D^m v)^2 d\Omega = \int_{\Omega} \left((D_{x^k}^k)^2 + (D_{x^{k-1}y}^k)^2 + \dots + (D_{y^k}^k)^2 \right) d\Omega \quad (167)$$

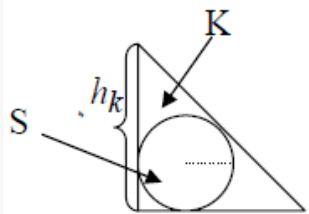
$$D_{x^k}^k v = \frac{\partial^k v}{\partial x^k}; \quad D_{x^{k-1}y}^k v = \frac{\partial^k v}{\partial x^{k-1} \partial y}; \quad \dots \quad (168)$$

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

Розглянемо розбиття на трикутники. Накладемо обмеження на них.

Означення: Розбиття T_h називається *регулярним*, якщо $\exists \sigma \geq 1$ таке, що

$$\max_{k \in T_h} \frac{h_k}{\rho_k} \leq \sigma, \quad h_k = \text{diam } K, \quad S \subset K, \quad \rho_k = \mu(S): \quad (169)$$



Якщо $h_k/\rho_k \gg 1$, то K вироджується в пряму і це погано.

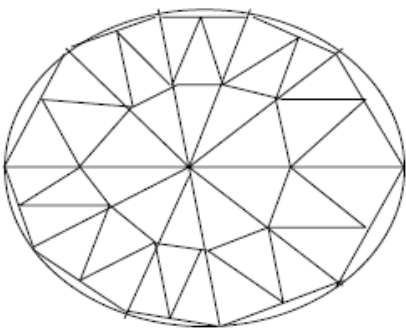
Теорема: Нехай $v \in W_2^{l+1}(\Omega)$, $1 \leq l \leq k$, T_h — регулярна триангуляція $v_h \in X_h = \{v_h : v_h|_K = P_k\}$. Тоді

$$|v - v_h|_m \leq Ch^{l+1-m} |v|_{l+1}, \quad m = 0, 1, \quad k \geq 1. \quad (170)$$

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

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:



Для цього вибираємо точки на границі і будуємо вписаний багатогранник. Його триангулюємо. Далі на кожному трикутнику будуємо інтерполянт. В результаті отримуємо $v_h \in X_h^k$.

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