Чисельні методи в задачах механіки

Частина І Теоретична та прикладна механіка

Київ 2015

Чисельні методи в задачах механіки. Частина І Теоретична та прикладна механіка. Методичний посібник. Кафедра Теоретичної та прикладної механіки механіко-математичного факультету Київського національного університету імені Тараса Шевченка. Укладач доцент кафедри к.ф-м н. Зражевський Г.М. Київ 2015. Електронна версія. 99 стор.

Студенти молодших курсів спеціалізації ”механіка” вивчають нормативний курс ”Чисельні методи в задачах механіки” в якому викладаються основні чисельні методи, що базуються на аналітичних методах та підходах таких нормативних курсів, як ”Математичний аналіз”, ”Лінійна алгебра”, ”Аналітична геометрія”, ”Диференціальна геометрія”, ”Теорія звичайних диференціальних рівнянь”. На протязі прослуховування матеріалу, слухачі мають набути навичок, що дозволяють їм вільно орієнтуватись в класичних методах обчислень, класифікувати їх та використовувати як при подальшому прослуховуванні програмного матеріалу старших курсів, так і в науковій діяльності при розробці курсових та дипломних робіт, роботі над науковими проектами, тощо. В той же час, абстрактне викладання цього надзвичайно важливого для спеціаліста в галузі сучасної механіки предмета, відсторонене від безпосереднього практичного застосування до задач механіки є недостатнім, оскільки не дає змогу встановити когнітивний зв’язок між абстрактними знаннями, що має студент та їх використанням при вирішенні конкретних задач, що мають фізичний зміст та можуть бути застосовані на практиці. В той же час, на момент прослуховування курс, студент має достатньо фаховий рівень знань як в галузі теоретичної та прикладної механіки, так і достатні навички використання інформаційних технологій для формулювання та програмної реалізації чисельних алгоритмів.

Методичний посібник має на меті не лише короткий опис чисельних методів, що входять в програму курсу з прикладами та зауваженнями щодо їх формулювання та застосування, а перш за все, продемонструвати на

численних прикладах з різних розділів теоретичної та прикладної механіки використання чисельних методів для вирішення практично значимих задач та проблем. Вважається, що користувач посібника достатньо ознайомлений з інтегрованими математично орієнтованими програмними чисельно-аналітичними системами (Mathematica, Matlab, Maple, тощо).

1. Абсолютна та відносна похибки. Похибка функції.

При підрахунках, що стосуються наближеного визначення скалярних чи векторних величин оперують поняттями **абсолютної** та **відносної похибок**.

Озн. Абсолютною похи*бкою* наближеного значення a^* деякої величини називається величина $\Delta(a^*)$ така, що

$$\left|a^* - a\right| \le \Delta(a^*), \tag{1.1}$$

де a - точне значення цієї величини.

Озн. Відносною похибкою наближеного значення a^* деякої величини називається величина $\delta(a^*)$ така, що

$$\left|(a^* - a)/a^*\right| \le \delta(a^*), \tag{1.2}$$

Де a - точне значення цієї величини.

Абсолютна та відносна похибки можуть бути використані для задання меж знаходження наближених величин: $a = a^* \pm \Delta(a^*)$.

Приклад. Оскільки $\sqrt{21} = 4.58257...$ то $\sqrt{21} = 4.582 \pm 6 \cdot 10^{-4}$, тобто 4.582 є наближеним значенням $\sqrt{21}$ з абсолютною похибкою $6 \cdot 10^{-4}$. Відносна похибка в цьому прикладі буде $1.31 \cdot 10^{-4}$ оскільки $6 \cdot 10^{-4} / 4.582 \le 1.31 \cdot 10^{-4}$.

Похибка значення функції вимірюється граничною абсолютною похибкою.

Озн. Якщо y^* - наближене значення функції $y = y(a_1,...,a_n)$, то **граничною абсолютною похибкою** $A(y^*)$ значення функції називають

$$A(y^*) = \sup_{(a_1,...a_n) \in G} \left|y(a_1,...,a_n) - y^*\right|, \tag{1.3}$$

де G - область зміни аргументів функції.

Зауваження. Граничною відносною похибкою функції називають величину $A(y^*)/\left|y^*\right|$.

Якщо G - прямокутник: $a=(a_1,...,a_n)\in G \Leftrightarrow \left|a_j - a_j^*\right| \le \Delta\left(a_j^*\right), \quad j=\overline{1,n}$ та $y=y(a_1,...,a_n)\in D(G)$, згідно з формулою Лагранжа:

$$\left|y(a_1,...,a_n)-y^*\right| \le A_0(y^*)=\sum B_j \Delta\left(a_j^*\right),\ j=\overline{1,n}, \tag{1.4}$$

де $B_j=\sup\limits_G \left|y_{a_j}(a_1,...,a_n)\right|$

Очевидно, $A(y^*)\le A_0(y^*)$.

У випадку, коли $y=y(a_1,...,a_n)\in C^1(G)$ та $\rho=\sqrt{\sum_{j=1}^n \Delta(a_j^*)^2}=o(1)$, має місце

$B_j=\left|y_{a_j}(a_1^*,...,a_n^*)\right|+o(1)$ при $\rho\to 0$, та $A_0(y^*)$ можна записати як

$$\begin{aligned} &A_0(y^*)=A^0(y^*)+\varepsilon_1(\rho),\ \varepsilon_1(\rho)=o(\rho) \\ &A^0(y^*)=\sum_{j=1}^n \left|y_{a_j}(a_1^*,...,a_n^*)\right|\Delta\left(a_j^*\right) \end{aligned} \tag{1.5}$$

та

$$A^0(y^*)+\varepsilon_2(\rho)\le A(y^*),\ \varepsilon_2(\rho)=o(\rho)$$

а отже

$$A^0(y^*)+\varepsilon_2(\rho)\le A(y^*)\le A_0(y^*)=A^0(y^*)+\varepsilon_1(\rho) \tag{1.6}$$

Озн. $A^0(y^*)$, що визначається формулою (5) називають **лінійною оцінкою похибки**.

Зауваження. Відношення $\left|y(a_1,...,a_n)-y^*\right|\le A^0(y^*)$ взагалі кажучи невірне, але оскільки в більшості випадків побудова граничної абсолютної похибки технічно достатньо складна, лінійна оцінка похибки є достатньо корисною альтернативою для оцінки функції.

Приклад. $y=\ln(a), \quad a^*=1, \quad \Delta(a^*)=0.001 \quad \Rightarrow \quad y^*=0, \quad y_a(a^*)=1,$

$B=\sup\limits_{|a-1|\le 0.001}\left|1/a\right|=1/0.999=1.001001...,$

$A(y^*)=\sup\limits_{|a-1|\le 0.001}\left|\ln(a)\right|=-\ln(0.999)=0.0010005...,$

$A_0(y^*) = B\Delta(a^*) = 0.001001\ldots$,

$A^0(y^*) = |y_a(a^*)|\Delta(a^*) = 0.001$.

Тобто лінійна оцінка похибки є достатньо точною, хоча відношення $|y(a_1,...,a_n) - y^*| \le A^0(y^*)$ взагалі кажучи не виконується.

Якщо ж $\Delta(a^*) = 0.1 \Rightarrow A(y^*) = -\ln(0.9) = 0.10536...$, $A_0(y^*) = B\Delta(a^*) = 0.1111..$,

$A^0(y^*) = |y_a(a^*)|\Delta(a^*) = 0.1$ та відмінність оцінок похибки є суттєвою.

Приклад. Для оцінки похибки лінійної функції $y(a_1,...,a_n) = \sum_{j=1}^n \gamma_j a_j$ з відомими похибками аргументів $\Delta(a_j^*)$, $j = \overline{1,n}$ матимемо: $A(y^*) = \sup\limits_{|a_j - a_j^*| \le \Delta(a_j^*)} \left|\sum_{j=1}^n \gamma_j (a_j - a_j^*)\right|$, $A_0(y^*) = A^0(y^*) = \sum_{j=1}^n |\gamma_j| \Delta(a_j^*)$ та для випадку $|\gamma_j| = 1$, $j = \overline{1,n} \Rightarrow A(y^*) = A_0(y^*) = A^0(y^*) = \sum_{j=1}^n \Delta(a_j^*)$ тобто гранична оцінка похибки суми значень рівна сумі абсолютних значень, що сумуються.

Приклад. Для оцінки похибки степеневої функції: $y(a_1,...,a_n) = \prod_{j=1}^n a_j^{p_j}$ з відомими похибками аргументів $\Delta(a_j^*)$, $j = \overline{1,n}$ матимемо: $A(y^*) = \sup\limits_{|a_j - a_j^*| \le \Delta(a_j^*)} \left|\prod_{j=1}^n \left(a_j^{p_j} - a_j^{*p_j}\right)\right|$, $A_0(y^*) = |y^*| \sum_{j=1}^n |p_j| \Delta(a_j^*) / |a_j^*| - \Delta(a_j^*)|$,

$|y^*| \sum_{j=1}^n |p_j| \Delta(a_j^*) / |a_j^*|$, та очевидно:

$A(y^*)/|y^*| \approx A^0(y^*)/|y^*| = \sum_{j=1}^n |p_j| \Delta(a_j^*)/|a_j^*| = \sum_{j=1}^n |p_j| \delta(a_j^*)$. У випадку, коли $|p_j| = 1$, $j = \overline{1,n} \Rightarrow A(y^*)/|y^*| \approx A^0(y^*)/|y^*| = \sum_{j=1}^n \delta(a_j^*)$ та гранична відносна похибка добутку чи відношення значень наближено (а лінійна відносна оцінка похибки точно) дорівнює сумі відносних похибок значень.

Якщо функція задана неявним чином:

$$F(y, a_1,...,a_n) = 0 \qquad (1.7)$$

та відомо значення y^* таке, що $F(y^*, a_1^*,...,a_n^*) = 0$, лінійну оцінку похибки функції можна побудувати наступним чином:

$$A^0(y^*)=\sum_{j=1}^{n}\left|y_{a_j}(a_1^*,...,a_n^*)\right|\Delta\left(a_j^*\right),\ y_{a_j}(a_1^*,...,a_n^*)=-F_{a_j}\cdot\left(F_y\right)^{-1}\Big|_{(y^*,a_1^*,...,a_n^*)} \qquad (1.8)$$

Задача 1. Мостовий кран рухається згідно рівнянню $x(t)=v_1 t$, по крану котиться візок згідно рівнянню: $y(t)=v_2 t$, ціп з вантажем вкорочується зі швидкістю v_3 та в момент часу $t=0$ мав довжину l. Вісь Oz напрямлено вниз. Визначити абсолютну похибку положення вантажу в момент часу T, якщо абсолютні похибки вимірювання швидкостей відомі $\Delta(v_i)=\Delta v_i, i=\overline{1,3}$. Побудувати лінійну оцінку похибки шляху s, пройденого вантажем.

Розв'язання. При відомих значеннях швидкостей, координати вантажу в $t=T$ в системі координат $Oxyz$ мають значення $\{v_1T, v_2T, l-v_3T\}$. Оскільки вирази є лінійними (відносно v_i), абсолютні похибки положення вантажу рівні $\Delta(x)=\Delta v_1 T$, $\Delta(y)=\Delta v_2 T$, $\Delta(z)=\Delta v_3 T$. Шлях, пройдений вантажем (траєкторія вантажу – відрізок прямої) визначається формулою $s(v_1,v_2,v_3,t)=\sqrt{(v_1t)^2+(v_2t)^2+(l-v_3t)^2}$. Лінійна оцінка похибки може бути знайдена за (1.5): $A^0(s^*)=\sum_{j=1}^{3}\left|s_{v_j}(v_1^*,v_2^*,v_2^*,T)\right|\Delta\left(v_j^*\right)$. Отже $A^0(s^*)=\left(v_1T^2\Delta v_1+v_2T^2\Delta v_2+(l-v_3T)T\Delta v_3\right)/\sqrt{(v_1T)^2+(v_2T)^2+(l-v_3T)^2}$.

Задача 2.Плоский рух матеріальної точки, кинутої під кутом α до горизонту з початковою швидкістю v_0 в вертикально розташованій системі координат Oxy (вісь Ox - горизонтальна, вісь Oy напрямлено вертикально вгору) має вигляд: вісь $x(t)=v_0\cos\alpha\, t$, $y(t)=v_0\sin\alpha\, t-gt^2/2$. Побудувати лінійну оцінку похибки для максимальної висоти підйому H, дальності L та часу T польоту, якщо відомі похибки задання $\Delta(v_0)=\Delta v_0, \Delta(\alpha)=\Delta\alpha<<1$.

Розв'язання. Легко отримати $H=v_0{}^2/(2g)\sin^2\alpha$, $L=v_0{}^2/g\sin 2\alpha$, $T=2v_0/g\sin\alpha$. Отже, за (1.5): $A^0(H)=v_0\sin(\alpha)\left(\sin(\alpha)\Delta v_0+v_0\cos(\alpha)\Delta\alpha\right)/g$, $A^0(L)=2v_0\left(\sin(2\alpha)\Delta v_0+v_0\cos(2\alpha)\Delta\alpha\right)/g$, $A^0(T)=2\left(\sin(\alpha)\Delta v_0+v_0\cos(\alpha)\Delta\alpha\right)/g$

Завдання для самостійного розв'язання

1. Однорідний стрижень кінцями A та B може ковзати по горохуватій дузі кола з радіусом a. Відстань OC стрижня до центру дуги O, розташованого у вертикальній площині b. Коефіцієнт тертя між стрижнем та дугою f. Визначити межі зміни кута φ між прямою OC та вертикальним діаметром дуги, в положенні рівноваги, якщо коефіцієнт тертя, a та b вимірюються з похибками $\Delta f, \Delta a, \Delta b$. Нехай відоме значення кута φ в граничному положенні рівноваги з точністю вимірювання $\Delta\varphi$. Визначити коефіцієнт тертя f та побудувати для нього лінійну оцінку похибки, вважаючи a та b відомими.

2. Для Задачі 2 вважати, що рух є тривимірним завдяки можливої присутності третьої компоненти початкової швидкості $\vec{v} = \{v_0 \cos\alpha, v_0 \sin\alpha, v_z\}$. Повторити розв'язок задачі при додатковій умові $v_z = 0$, $\Delta(v_z) = \Delta v_z << 0$.

3. При розриві маховика деякі з його частин знайдено на максимальній відстані $s \pm \Delta s$. Нехтуючи опором повітря знайти вірогідне значення кутової швидкості маховика в момент катастрофи та побудувати для нього лінійну оцінку похибки. Як зміниться шукані характеристики при врахуванні опору повітря?

2. Інтерполювання та суміжні питання.

Інтерполювання – один з методів наближення функції.

Озн. Наближення функції $f(x) \approx g(x; a_1, \ldots a_n)$ де $a_1, \ldots a_n$ - параметри наближення називається **інтерполюванням** якщо параметри наближення знаходяться з умови

$$g(x_i; a_1, \ldots a_n) = f(x_i),\ i = \overline{1,n}, \qquad (2.1)$$

В цьому випадку $g(x; a_1, \ldots a_n)$ називається інтерполяційною функцією, а x_i називаються **вузлами інтерполювання**, n - **порядок інтерполювання**.

Озн. Якщо інтерполяційна функція є лінійною по відношення до параметрів інтерполювання:

$$g(x_i; a_1, \ldots a_n) = \sum_{i=1}^{n} a_i \varphi_i(x), \qquad (2.1)$$

то інтерполювання називається **лінійною**. В цьому випадку $a_1,...a_n$ називаються **коефіцієнтами інтерполяційної функції** (коефіцієнтами інтерполювання).

Озн. Методом невизначених коефіцієнтів розв'язання задачі лінійного інтерполювання називається метод визначення коефіцієнтів інтерполювання шляхом розв'язання системи лінійних алгебраїчних рівнянь (СЛАР):

$$\sum_{j=1}^{n} a_j \varphi_j(x_i) = f(x_i),\ i = \overline{1,n} \tag{2.2}$$

Озн. У випадку, коли $\varphi_i(x) = x^{i-1}$, інтерполювання називається **поліноміальним**, а інтерполяційна функція $g(x_i; a_1,...a_n) = \sum_{i=1}^{n} a_i x^{i-1}$ називається **інтерполяційним многочленом**.

Зауваження. Метод невизначених коефіцієнтів для випадку поліноміального інтерполювання завжди коректна та має єдиний розв'язок при виконання умови $x_i \neq x_j, i \neq j, i, j = \overline{1,n}$ (всі вузли інтерполювання мають кратність 1) оскільки в цьому випадку $\det\left\lfloor x_j^{\,i-1} \right\rfloor = \prod_{1 \le j < i \le n} (x_i - x_j)$ є визначником Вандермонда та завжди відмінний від 0.

2.1 Інтерполяційний многочлен Лагранжа

Озн. Многочлен степені $n-1$, що задовольняє умову

$$\Phi_i(x_j) = \delta_{ij},\ j = \overline{1,n} \wedge\ j \neq i \tag{2.3}$$

Називається **фундаментальним многочленом Лагранжа**.

Зауваження. У випадку, коли всі вузли інтерполювання різні (мають кратність 1), фундаментальним многочленом Лагранжа може бути легко побудований, оскільки відомі значення многочлену в n різних точках при n невідомих коефіцієнтах такого многочлену. Фундаментальний многочлен Лагранжа має вигляд:

$$\Phi_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j} \tag{2.4}$$

Озн. Інтерполяційним многочленом Лагранжа функції $f(x)$ з n заданими вузлами інтерполювання $x_i : x_i \neq x_j, i \neq j, i, j = \overline{1,n}$ називають інтерполяційний многочлен степені $n-1$:

$$L_n(x) = \sum_{i=1}^{n} f(x_i)\Phi_i(x), \text{ або } L_n(x) = \sum_{i=1}^{n} f(x_i)\prod_{j\neq i}\frac{x - x_j}{x_i - x_j} \qquad (2.5)$$

Зауваження. Інтерполяційним многочленом Лагранжа має сенс лише у випадку, коли всі вузли інтерполювання різні.

Зауваження. При умові задання вузлів інтерполювання, інтерполяційний многочлен визначається єдиним чином. Отже термін ”Інтерполяційний многочлен Лагранжа” скоріше має відношення до способу його побудови.

Озн. Залишковим членом інтерполяційного многочлену Лагранжа називають функцію:

$$f(x) - L_n(x) = f^{(n)}(\zeta)\omega_n(x)/n! \qquad (2.6)$$

де $\omega_n(x) = \prod_{i=1}^{n}(x - x_i)$ - канонічний многочлен з нулями в вузлах інтерполювання $\{x_i\}_{i=1}^{n} : x_i \in [y_1, y_2]$, а $\zeta \in [y_1, y_2]$ - деяка точка з області визначення інтерполяційного многочлену ($y_1 = \min(x_1,...x_n, x)$, $y_{21} = \max(x_1,...x_n, x)$).

Зауваження. Залишковий член інтерполяційного многочлену Лагранжа визначає похибку наближення функції на деякому відрізку її інтерполяційним многочленом в термінах многочлену степені n (порядок інтерполювання).

Зауваження. Може видатись дивним, що різниця деякої функції та многочлену поточкові визначається многочленом. Насправді це не так. Треба прийняти до уваги, що взагалі кажучи ζ також залежить від значення x. При цьому у випадку інтерполювання диференційованих функцій можна побудувати оцінку в нормі $C_{[y_1,y_2]}$:

$$\left\| f(x) - L_n(x) \right\|_{C_{[y1,y2]}} \leq \sup_{\zeta\in[y_1,y_2]} \left| f^{(n)}(\zeta) \right| \left\| \omega_n(x) \right\|_{C_{[y1,y2]}} / n! \qquad (2.7)$$

Формула (2.6) свідчить про те, що поліноміальна інтерполяція (зрештою, як і будь яка інтерполяція) не є апроксимацією функції з нормою $C_{[y_1,y_2]}$, оскільки оцінка

похибки залежить від гладкості функції, а також від норми канонічного многочлену. Підвищення порядку інтерполювання (тобто збільшення кількості вузлів інтерполювання) при збереженні відрізку інтерполювання може призвести до погіршення наближення як за рахунок втрати гладкості так і за рахунок галопування канонічного многочлену. Головна (та єдина) властивість наближення функції цілком визначається умовою інтерполювання (2.2).

2.2 Розділені різниці та їх властивості. Інтерполяційний многочлен Ньютона з розділеними різницями.

Озн. Розділеною різницею функції $f(x)$ порядку $n-1$, побудованою скінченній множині точок $x_i : x_i \neq x_j, i \neq j, i, j = \overline{1,n}$ називається функція, що задається рекурсивним правилом:

$$\begin{aligned} &f(x_i; x_j) = \big(f(x_i) - f(x_j)\big)/(x_j - x_i) \\ &f(x_i; x_j; x_k) = \big(f(x_j; x_k) - f(x_i; x_j)\big)/(x_k - x_i) \\ &... \\ &f(x_1; x_2; ...; x_n) = \big(f(x_2; ...; x_n) - f(x_1; ...; x_{n-1})\big)/(x_n - x_1) \end{aligned} \qquad (2.8)$$

де відповідно, $f(x_i; x_j)$ - розділена різниця 1-го порядку, $f(x_i; x_j; x_k)$ - другого, $f(x_1; x_2; ...; x_n)$ - $n-1$-го.

Лема. **Розділену різницю** функції $f(x)$ порядку $n-1$, побудованою скінченній множині точок $x_i : x_i \neq x_j, i \neq j, i, j = \overline{1,n}$ можна визначити формулою:

$$f(x_1; ...; x_n) = \sum_{i=1}^{n} f(x_i) / \prod_{j=1, j\neq i}^{n} (x_i - x_j) \qquad (2.9)$$

Зауваження. Лему можна достатньо просто довести за методом математичної індукції.

Розділені різниці, виходячи з їх означення мають наступні властивості.

1. Лінійність:

$$(\alpha f + \beta g)(x_1; x_2; ...; x_n) = \alpha f(x_1; x_2; ...; x_n) + \beta g(x_1; x_2; ...; x_n) \qquad (2.10)$$

2. Симетричність по відношенню до аргументів:

$$f(x_1;...;x_i;...;x_j;...;x_n) = f(x_1;...;x_j;...;x_i;...;x_n) \quad (2.11)$$

Ці властивості є очевидними з огляду на твердження Леми.

Рекурсивне правило (2.8) дає змогу ввести простий алгоритм побудови розділених різниць $n-1$-го порядку, який називається **таблиця розділених різниць.**

Таблиця 2.1

x	$f(.)$	$f(.;.)$	$f(.;.;.)$	…	$f(.;...;.)$
x_1	$f(x_1)$				
x_2	$f(x_2)$	$f(x_1;x_2)$			
x_3	$f(x_3)$	$f(x_2;x_3)$	$f(x_1;x_2;x_3)$		
…	…	…	…	…	
x_n	$f(x_n)$	$f(x_{n-1};x_n)$	$f(x_{n-2};x_{n-1};x_n)$	…	$f(x_1;...;x_n)$

Алгоритм побудови таблиці розділених різниць є наступним. Перші 2 стовпчики заповнюються значеннями вузлів $\{x_i\}_{i=1}^n$ та вузловими значеннями $\{f(x_i)\}_{i=1}^n$ (їх можна вважати розділеними різницями 0-го порядку). Наступні стовпчики містить розділені різниці 1-го, 2-го,… $n-1$-го порядків. Значення в таблиці вибудовуються послідовно по стовпчикам зліва направо. Так елемент, що відповідає розділеній різниці $k-1$-го поряду (розташовується в $k+1$ стовпчику) в одному рядку з x_i (тобто в i рядку) a_{ik+1}, згідно з (2.8) знаходиться за формулою $a_{ik+1} = (a_{ik} - a_{i-1k})/(a_{i1} - a_{i-k+11})$.

Зауваження. Функціональна частина таблиці розділених різниць є нижньо трикутною.

Сенс введення розділених різниць полягає в тому, що

$$f(x) - L_n(x) = f(x;x_1;...;x_n)\omega_n(x) \quad (2.12)$$

якщо $x \neq x_i, i = \overline{1,n}$ розглядається як один з вузлів розділеної різниці.

Окрім того, легко показати, що:

$$L_m(x) - L_{m-1}(x) = f(x_m;x_1;...;x_{m-1})\omega_{m-1}(x),\ 1 \le m \le n \quad (2.13)$$

де $\omega_{m-1}(x)$ як і раніше – канонічний многочлен степені $m-1$, побудований на точках $\{x_i\}_{i=1}^{m-1}$.

Формула (2.13) дає змогу побудувати відмінне від (2.5) представлення інтерполяційного многочлену, з вузлами інтерполювання $\{x_i\}_{i=1}^{n}$:

$$L_n(x) = f(x_1) + f(x_1;x_2)(x-x_1) + f(x_1;x_2;x_3)(x-x_1)(x-x_2) + \dots + f(x_1;\dots;x_n)(x-x_1)\dots(x-x_{n-1}) \quad (2.14)$$

Озн. Запис інтерполяційного многочлену у вигляді (2.14) називається **інтерполяційним многочленом Ньютона з розділеними різницями**.

Формула (2.12) в сукупності з (2.6) дає змогу визначити розділену різницю через характеристику функції для якої побудована розділена різниця:

$$f(x;x_1;\dots;x_n) = f^{(n)}(\zeta)/n!,\ \zeta \in [y_1, y_2] \quad (2.15)$$

В свою чергу, (2.12) є іншим записом залишкового члену інтерполяційного многочлену.

Інтерполяційний многочлен Ньютона з розділеними різницями зручно будувати використовуючи таблицю розділених різниць. Нехай функціональна частина таблиці розділених різниць складає нижньо трикутну матрицю $\{F_{ij}\}_{i,j=1}^{n}$, $F_{ij} = 0,\ j > i,\ i,j = \overline{1,n}$ тобто в j-му стовпчику розташовуються розділені різниці $j-1$-го порядку. Тоді (2.14) можна побудувати як

$$L_n(x) = F_{11} + \sum_{i=2}^{n} F_{ii}(x-x_1)\dots(x-x_{i-1}). \quad (2.16)$$

Окрім того, враховуючи властивість симетричності розділених різниць, можна побудувати кілька представлень виду (2.14) шляхом простого пере нумерування вузлів інтерполювання, хоча, звичайно, коефіцієнти інтерполяційного многочлену при цьому не зміняться.

Наприклад, для побудови інтерполяційного многочлену можна використати останній рядок таблиці розділених різниць:

$$L_n(x) = f(x_n) + f(x_{n-1};x_n)(x-x_n) + f(x_{n-2};x_{n-1};x_n)(x-x_n)(x-x_{n-1}) + \dots + f(x_1;\dots;x_n)(x-x_n)\dots(x-x_2) \quad (2.17)$$

або в термінах коефіцієнтів таблиці:

$$L_n(x) = F_{n1} + \sum_{i=2}^{n} F_{ni}(x - x_n)...(x - x_{n-i+1}). \qquad (2.18)$$

Задача 1.

Положення руху матеріальної точки фіксується на проміжку $0 \le t \le 10$ (с) та задається таблично:

Таблиця 2.2

i	1	2	3	4	5	6	7	8
t_i(с)	0	2	3	4	6	8	9	10
$x_i = x(t_i)$(м)	0	4.755	4.755	2.939	2.939	-4.755	-2.939	0
$y_i = y(t_i)$(м)	10	3.090	-3.090	-8.090	-8.090	3.090	8.090	10

Побудувати інтерполяційні многочлени $X_8(t) \approx x(t)$ та $Y_8(t) \approx y(t)$. Визначити величину швидкості та прискорення в моменти часу $t_j = t_0 + \tau\, j$, $t_0 = 0,\ \tau = 1(c),\ j = \overline{0,10}$. Знайти максимальний та мінімальний радіуси кривизни траєкторії.

Розв'язання.

Інтерполяційні многочлени з розділеними різницями:

$$\begin{cases} X_8(t) \approx x(t_1) + x(t_1;t_2)(t-t_1) + ... + x(t_1;...;t_8)(t-t_1)...(t-t_7) \\ Y_8(t) \approx y(t_1) + y(t_1;t_2)(t-t_1) + ... + y(t_1;...;t_8)(t-t_1)...(t-t_7) \end{cases}$$

Побудувавши таблиці розділених різниць (Таблиця 2.1)

t_i	$x(t_i)$	$x(;)$	$x(;;)$	$x(;;;)$	$x(;;;;)$	$x(;...;)$	$x(;...;)$	$x(;...;)$
0	0							
2	4.755	2.378						
3	4.755	0	-0.793					
4	2.939	-1.816	-0.908	-0.029				
6	-2.939	-2.939	-0.374	0.133	0.027			
8	-4.755	-0.908	0.508	0.176	0.007	-0.003		
9	-2.939	1.816	0.908	0.080	-0.016	-0.003	-0.0001	
10	0	2.939	0.561	-0.087	-0.027	-0.002	0.0002	0.0001

t_i	$y(t_i)$	$y(;)$	$y(;;)$	$y(;;;)$	$y(;;;;)$	$y(;...;)$	$y(;...;)$	$y(;...;)$
0	10							
2	3.091	-3.455						
3	-3.091	-6.180	-0.908					
4	-8.091	-5	0.590	0.375				
6	-8.091	0	1.667	0.269	-0.0176			
8	3.091	5.590	1.398	-0.054	-0.0538	-0.0045		
9	8.091	5	-0.197	-0.319	-0.0442	-0.0014	-0.0007	
10	10	1.910	-1.546	-0.337	-0.0030	0.0059	0.0006	0

Побудувавши (2.14) та зібравши коефіцієнти отримаємо:

$$\begin{cases} X_8(t) = 3.193t - 0.0964t^2 - 0.3117t^3 - 0.3133t^4 + 0.0116t^5 - 0.001t^6 + 0.00003t^7 \\ Y_8(t) = 10 - 0.474t - 1.238t^2 - 0.452t^3 + 0.206t^4 - 0.0235t^5 + 0.001t^6 \end{cases}$$

$$Y_8(t) = 10 - 0.474t - 1.238t^2 - 0.452t^3 + 0.206t^4 - 0.0235t^5 + 0.001t^6$$

Продиференціювавши інтерполяційні многочлени, отримаємо компоненти швидкості та прискорення руху:

$$\begin{cases} \dot{X}_8(t) = 3.193 - 0.193t - 0.395t^2 - 0.125t^3 + 0.582t^4 + 0.0062t^5 + 0.0002t^6 \\ \dot{Y}_8(t) = -0.474 - 2.477t - 1.357t^2 + 0.826t^3 - 0.117t^4 + 0.0058t^5 - 0.0001t^6 \end{cases}$$

$$\begin{cases} \ddot{X}_8(t) = -0.193 - 0.791t - 0.376t^2 + 0.233t^3 - 0.311t^4 + 0.0012t^5 \\ \ddot{Y}_8(t) = -2.477 - 2.714t + 2.477t^2 - 0.469t^3 + 0.029t^4 - 0.0004t^5 \end{cases}$$

Величину швидкості та прискорення знайдемо як $V(t) \approx \sqrt{\dot{X}^2(t) + \dot{Y}^2(t)}$, $W(t) \approx \sqrt{\ddot{X}^2(t) + \ddot{Y}^2(t)}$. Кривизна кривої, заданої параметрично: $K(t) \approx \left| \dot{Y}_8(t)\ddot{X}_8(t) - \ddot{Y}_8(t)\dot{X}_8(t) \right| \Big/ \left(\dot{X}_8^2(t) + \dot{Y}_8^2(t) \right)^{3/2}$. Порахувавши значення в моменти часу $t_j = t_0 + \tau\, j$, оформимо результат у вигляді Таблиці:

t_i	0	1	2	3	4	5	6	7	8	9	10
$V(t_i)$	3.23	4.40	6.03	6.06	4.48	3.14	4.47	6.05	6.07	4.46	3.17
$W(t_i)$	2.48	3.36	2.28	2.24	3.41	3.93	3.39	2.26	2.24	3.44	3.45
$K(t_i)$	0.24	0.14	0.056	0.056	0.14	0.39	0.15	0.06	0.06	0.14	0.34

Максимальна кривизна досягається в момент $t = 5$ секунд та дорівнює 0.39 (1/м).

Задача 2.

Шляхом спостереження радар зафіксував 3 положення мінометної міни в місцевій системі координат $Oxyz$ (вісь Oz напрямлена вертикально вгору). Нехтуючи опором повітря, знайти (в місцевій системі координат) координати місця пострілу, координати місця ураження та характеристики пострілу. Координати наведено в таблиці:

Таблиця 2.3

i	1	2	3
t_i(с)	10	25	35
$x_i = x(t_i)$(м)	2359.5	1248.76	508.27
$y_i = y(t_i)$(м)	1772.47	1131.19	703.66
$z_i = z(t_i)$(м)	1858.73	2807.45	2213.69

Розв'язання.

Введемо локальну систему координат, зв'язану з мінометом $O'x'y'z'$ таку, що вісь $O'z'$ напрямлена вертикально вгору, траєкторія міни лежить в площині $O'x'z'$, початок системи знаходиться в місці пострілу. В цьому випадку, вісь $O'x'$ лежить на прямій: $(x - x_1)/(x_3 - x_1) = (y - y_1)/(y_3 - y_1)$. Використовувати 3-ю точку (x_2, y_2) - не можна, оскільки координати визначаються з деякою похибкою. Приймаючи до уваги те, що постріл в наш бік (координати (x_i, y_i) зменшуються зі зростанням часу), кут між Ox та $O'x'$: $\varphi = \arctan(x_1 - x_3, y_1 - y_3) + \pi = 3.6652$. При цьому, з точністю до невідомої константи c, координати точок (x_i', z_i') в системі координат $O'x'z'$: $(x_i / \cos\varphi + c, z_i)$, або:

$(-2724.52+c, 1858.78), (-1441.95+c, 2807.45), (-586.89+c, 2213.69)$. Побудуємо інтерполяційний многочлен з розділеними різницями для траєкторії $z' = f(x')$: $f(x') \approx L_3(x') = f(x_1') + f(x_1'; x_2')(x' - x_1') + f(x_1'; x_2'; x_3')(x' - x_1')(x' - x_2')$ з вільним членом $a_0 = f(x_1') - f(x_1'; x_2')x_1' + f(x_1'; x_2'; x_3')x_1'x_2'$.

Таблиця розділених різниць для таких точок (2.4):

x_i'	$f(x_i')$	$f(x_i'; x_j')$	$f(x_i'; x_j'; x_k')$
-2043.39+c	1858.73		
-1081.46+c	2807.45	0.739701	
-440.17+c	2213.69	-0.694425	-0.000671

А отже $f(x') \approx L_3(x') = a_0 - (2.056 - 0.00134c)x' - 0.000671x'^2$ де $a_0 = 1238.37 + 2.0556c - 0.000671c^2$.

Знаходимо $c = 3579.57$ з умови $a_0 = 0$ - оскільки траєкторія в $O'x'z'$ має проходити через початок координат. Таким чином, координати точок спостереження (x_i', z_i') в системі координат $O'x'y'$: (855.05, 0) ,(2137.63, 0), (2992.68,0), а інтерполяційний многочлен: $L_3(x') = 2.7474x' - 0.000671x'^2$. Афінне перетворення між системами координат $O'x'y'$ та Oxy:

$\begin{cases} x - x_0 = x'\cos\varphi - y'\sin\varphi \\ y - y_0 = x'\sin\varphi + y'\cos\varphi \end{cases}$, отже $\begin{cases} x_0 = x_1 - x_1'\cos\varphi \\ y_0 = y_1 - x_1'\sin\varphi \end{cases}$, де $(x_0, y_0) = (3100, 2200)$ - координати точки пострілу (точки O'). Координати точки ураження легко знайти з рівняння траєкторії: $L_3(x_4') = 0 \Rightarrow x_4' = 0.000671/2.7474 = 4095.23$. Тоді

$\begin{cases} x_4 = x_0 + x_4'\cos\varphi = -446.572 \\ y_4 = y_0 + x_4'\sin\varphi = 152.384 \end{cases}$.

Визначимо характеристики пострілу. Закон руху в $O'x'z'$:

$\begin{cases} x' = V_0\cos\alpha\, t \\ z = V_0\sin\alpha\, t - gt^2/2 \end{cases}$, а отже $\begin{cases} \alpha = \arctan\left((z_i + gt_i^2/2)/x_i'\right) = 1.22 \approx 70^\circ \\ V_0 = \sin\alpha\, t_i - gt_i^2/2 = 250 \end{cases}$

Відповідь: координати точки пострілу (3100, 2200) (м), координати точки ураження (-446.57, 152.38) (м), початкова швидкість міни 250 (м/с), кут стріляння 70°.

Завдання для самостійного розв'язання

1. Положення руху матеріальної точки фіксується на проміжку $0 \le t \le 1$ (с) та задається таблично:

Таблиця 2.5

i	1	2	3	4	5	6	7	8
t_i(с)	0	0.1	0.2	0.3	0.5	0.7	0.9	1
$x_i = x(t_i)$(м)	0.0	0.927	1.7633	2.427	3.0	2.427	0.927	0.0
$y_i = y(t_i)$(м)	2.,	1.618	0.618	-0.618	-2.0	-0.618	1.618	2.0
$z_i = z(t_i)$(м)	0	0.1	0.2	0.3	0.5	0.7	0.9	1

Побудувати інтерполяційні многочлени $X_8(t) \approx x(t)$, $Y_8(t) \approx y(t)$ та $Z_8(t) \approx z(t)$. Переконатись, що траєкторія є плоскою кривою. Визначити величину швидкості та прискорення в моменти часу $t_j = t_0 + \tau\, j$, $t_0 = 0, \tau = 0.1(c), j = \overline{0,10}$. Знайти максимальний радіус кривизни траєкторії.

2. Політ ракети в перші 100 секунд спостерігається радіолокаційною системою з показами азимут (A_j-в градусах) відстань (R_j- в метрах) та висота (H_j - в метрах) на десяти секундній основі. Спостережені дані приведено в таблиці:

Таблиця 2.6

i	1	2	3	4	5	6	7	8	9	10	11
t_i	0	10	20	30	40	50	60	70	80	90	100
A_i	61.39	60.75	60.15	59.59	59.07	58.61	58.19	57.80	57.43	57.08	56.73
R_i	2506	2621	2955	3888	5604	8073	11232	15041	19478	24534	30201

H_i	0	300	1200	2700	4800	7500	10800	14700	19200	24300	30000

Побудувати закон руху ракети в географічній системі координат *Oxyz* (вісь *Ox* напрямлено на північ, вісь *Oz* - вертикально вгору). Побудувати таблиці швидкості та прискорення.

3. При посадці літака альтиметр фіксує висоту польоту (в метрах). Дані наведені в таблиці:

Таблиця 2.7

i	1	2	3	4	5	6	7	8	9
t_i	0	3	6	9	12	15	18	21	24
H_i	2025	1984.7	1866.24	1676.9	1428.84	1139.06	829.44	526.7	262.44

Побудувати таблиці швидкості та прискорення зниження літака. Визначити чи такий режим посадки є вдалим (швидкість зниження літака в момент дотику до злітної смуги дорівнює нулю).

2.3 Розділені різниці та інтерполяційний многочлен з кратними вузлами

В попередніх розділах, інтерполяційний многочлен будувався при умові, що вузли інтерполювання різні, а умова інтерполювання була задекларована, як збігання значень інтерпольованої функції та інтерполяційного многочлена в вузлах інтерполювання (2.1). В деяких задачах вказана умова інтерполювання є недостатньою. В вузлах інтерполювання збігатись мають не лише значення функції та многочлену, а також значення їх похідних до деякого порядку включно.

Озн. Кількість умов, накладених на інтерполяційний многочлен в вузлі інтерполювання називається **кратністю вузла інтерполювання**.

Таким чином, попередні методи побудови інтерполяційного многочлену справедливі для різних однократних вузлів інтерполювання.

Поставимо задачу побудови інтерполяційного многочлену, що задовольняє s умов:

$$\begin{aligned} &g_s(x_1) = f(x_1), g_s'(x_1) = f'(x_1), ..., g_s^{m_1-1}(x_1) = f^{m_1-1}(x_1) \\ &g_s(x_2) = f(x_2), g_s'(x_2) = f'(x_2), ..., g_s^{m_2-1}(x_2) = f^{m_2-1}(x_2) \\ &... \\ &g_s(x_n) = f(x_n), g_s'(x_n) = f'(x_n), ..., g_s^{m_n-1}(x_n) = f^{m_n-1}(x_n) \end{aligned} \qquad (2.19)$$

при цьому, очевидно, $s = \sum_{i=1}^{n} m_i$ - порядок інтерполювання, а інтерполяційний многочлен в загальному випадку має степінь $s-1$. Інтерпольована функція $f(x)$ має належати до класу $C_{[y_1,y_2]}^{\max\{m_i\}-1}$.

Коефіцієнти **інтерполяційного многочлену з кратними вузлами** $g_s(x)$ можна відшукати з використанням методу невизначених коефіцієнтів. Окрім того, $g_s(x)$ можна побудувати з використанням таблиці розділених різниць, узагальнивши розділені різниці на випадок кратних вузлів. Таке узагальнення можна провести за методом розщеплення вузлів.

Нехай $\varepsilon_0 = \min_{i \neq j}(|x_i - x_j|) / \max_{i=1,n}\{m_i - 1\}$. Введемо додаткові вузли інтерполювання так, що повний набір вузлів має вигляд:

$$\left\{x_{ij_i}^{\varepsilon}\right\}_{i=1,n,\ j_i=1,m_i}, x_{ij_i}^{\varepsilon} = x_i + (j_i - 1)\varepsilon,\ \varepsilon < \varepsilon_0 \qquad (2.20)$$

Кількість вузлів рівна s та всі вузли різні, а отже на них можна побудувати таблицю розділених різниць для $f(x)$ (формально, оскільки значення $f(x)$ невизначене в додаткових вузлах).

При цьому, очевидно $x_{ij_i}^{\varepsilon} \underset{\varepsilon \to 0}{\to} x_i, i = \overline{1,n},\ j_i = \overline{1,m_i}$.

Записавши розділену різницю $f(x_{ij_k}; x_{ij_{k+1}}; ...; x_{ij_{k+l}})$ порядку l через означення (2.8) та спрямувавши ε до нуля, легко отримати:

$$\lim_{\varepsilon \to 0} f(x_{ij_k}; x_{ij_{k+1}}; ...; x_{ij_{k+l}}) = f^{(l)}(x_i)/l! \qquad (2.21)$$

А отже, перехід до границі в таблиці розділених різниць при $\varepsilon \to 0$ дає підстави не змінюючи позначень розділених різниць ввести узагальнення розділеної різниці порядку l, побудованої на вузлі кратності $l+1$ через похідну функції порядку l:

$$f(x_i;...;x_i) = f^{(l)}(x_i)/l! \qquad (2.22)$$

В таблиці розділених різниць вузол порядку x_i та вузлове значення має повторюватись m_i разів, а невизначеність в підрахунку розділеної різниці на кратних точках – замінюватись виразом виду (2.20).

Таблиця 2.8

	x	$f(.)$	$f(.;.)$	$f(.;.;.)$	…	$f(.;...;.)$
1	x_1	$f(x_1)$				
2	x_1	$f(x_1)$	$f(x_1;x_1)$			
3	x_1	$f(x_1)$	$f(x_1;x_1)$	$f(x_1;x_1;x_1)$		
…	…	…	…	…	…	
m_1	x_1	$f(x_1)$	$f(x_1;x_1)$	$f(x_1;x_1;x_1)$	…	
m_1+1	x_2	$f(x_2)$	$f(x_1;x_2)$	$f(x_1;x_1;x_2)$	…	
…	…	…	…	…	…	
m_1+m_2	x_2	$f(x_2)$	$f(x_2;x_2)$	$f(x_2;x_2;x_2)$	…	
…	…	…	…	…	…	
s	x_n	$f(x_n)$	$f(x_n;x_n)$	$f(x_n;x_n;x_n)$	…	$f(x_1;...;x_1;...;x_n;...;x_n)$

Інтерполяційний многочлен з кратними вузлами з узагальненими розділеними різницями має вигляд:

$$\begin{aligned} g_s(x) &= f(x_1) + f(x_1;x_1)(x-x_1) + \\ &f(x_1;x_1;x_1)(x-x_1)^2 + ... + f(x_1;...;x_1)(x-x_1)^{m_1-1} + \\ &+ f(x_1;...;x_1;x_2)(x-x_1)^{m_1} + f(x_1;...;x_1;x_2;x_2)(x-x_1)^{m_1}(x-x_2) + ... \\ &+ f(x_1;...;x_1;...;x_n;...;x_n)(x-x_1)^{m_1}...(x-x_{n-1})^{m_{n-1}}(x-x_n)^{m_n-1} \end{aligned} \qquad (2.23)$$

Залишковий член інтерполювання має вигляд, аналогічний (2.12):

$$f(x) - g_s(x) = f(x; x_1; ...; x_1; ...; x_n; ...x_n)\omega_s(x) \qquad (2.24)$$

де $\omega_s(x) = \prod_{i=1}^{n}(x - x_i)^{m_i}$, або в термінах похідної функції:

$$f(x) - g_s(x) = f^{(s)}(\zeta)/s!\ \omega_s(x),\ \zeta \in [y_1, y_2] \qquad (2.25)$$

За необхідності, легко отримати оцінку похибки інтерполювання, аналогічну (2.7).

Зрозуміти інтерполювання з кратними вузлами можна, помітивши, що відрізок ряду Тейлора функції $f(x) \in C^s_{[y_1,y_2]}$ в деякій точці x_0 порядку s:

$$f(x) \approx f(x_0) + f'(x_0)/1!(x - x_0) + f''(x_0)/2!(x - x_0)^2 + ... + \\ + f^{(s-1)}(x_0)/(s-1)!(x - x_0)^{s-1} \qquad (2.26)$$

є ні чим іншим, як інтерполяційним многочленом функції з єдиним вузлом інтерполювання x_0 кратності s. В цьому легко переконатись з (2.23), враховуючи (2.22). В цьому випадку (2.23) є залишковим членом ряду Тейлора у формі Лагранжа.

Така аналогія дає також відповідь на питання, чи можуть бути умови інтерполювання (2.19) неповними. Тобто, чи всі похідні мають бути присутні. Так, оскільки для визначення відрізку ряду Тейлора мають бути задані всі похідні.

Задача 1. При визначенні характеристик двигуна при розгоні, кут повороту маховика (який можна вважати однорідним диском з масою m=50 кг та радіусом R=0.5 м) фіксувався за допомогою СКОТ (синусно-косинусний обертаючий ся трансформатор). Спостережені дані приведено в таблиці 2.9. Побудувати часову характеристику двигуна при розгоні.

Таблиця 2.9

i	1	2	3	4	5	6
t_i	0	0.2	0.4	0.6	0.8	1.0
$\sin(\varphi_i)$	0	0.616	0.487	-0.382	-0.846	-0.521
$\cos(\varphi_i)$	1.0	0.788	-0.873	0.924	-0.533	-0.853

Розв'язання.

Перерахувавши значення Таблиці 2.9: $\varphi_i = \arctan\left(\sin(\varphi_i)/\cos(\varphi_i)\right) + \pi k$ з врахуванням того, що кут не може зменшуватись: $\arctan\left(\sin(\varphi_{i+1})/\cos(\varphi_{i+1})\right) \le \varphi_i \Rightarrow \quad \varphi_{i+1} = \arctan\left(\sin(\varphi_{i+1})/\cos(\varphi_{i+1})\right) + \pi$, побудуємо таблицю кутів повороту:

i	1	2	3	4	5	6
t_i	0	0.2	0.4	0.6	0.8	1.0
φ_i(рад)	0	0.663	2.633	5.891	10.433	16.256

Приймаючи до уваги, що $\varphi(t=0)=0$ та $\omega(t=0)=0$, побудуємо інтерполяційний многочлен з 2-х кратним нулем в t_1:

$$\Phi_7(t) = \varphi(t_1) + \varphi(t_1;t_1)(t-t_1) + \varphi(t_1;t_1;t_2)(t-t_1)^2 + \varphi(t_1;t_1;t_2;t_3)(t-t_1)^2(t-t_2) +$$
$$\varphi(t_1;t_1;t_2;t_3;t_4)(t-t_1)^2(t-t_2)(t-t_3) + \varphi(t_1;t_1;t_2;t_3;t_4;t_5)(t-t_1)^2(t-t_2)(t-t_3)(t-t_4)$$
$$+ \varphi(t_1;t_1;t_2;t_3;t_4;t_5;t_6)(t-t_1)^2(t-t_2)(t-t_3)(t-t_4)(t-t_5)$$

Таблиця розділених різниць має вигляд:

t_i	φ_i	$\varphi_i(;)$	$\varphi_i(;;)$	$\varphi_i(;;;)$	$\varphi_i(;;;;)$	$\varphi_i(;;;;;)$	$\varphi_i(;;;;;;)$
0	0						
0	0	0					
0.2	0.663	3.318	16.589				
0.4	2.633	9.846	16.319	-0.672			
0.6	5.891	16.292	16.117	-0.336	0.559		
0.8	10.433	22.71	16.04	-0.123	0.266	-0.366	
1	16.256	29.116	16.015	-0.0455	0.097	-0.168	0.198

Відповідно, інтерполяційний многочлен для кута повороту:

$$\Phi_7(t) = 16.793t^2 - 1.2486t^3 + 1.2768t^4 - 0.7631t^5 + 0.1983t^6.$$

Продиференціювавши $\Phi_7(t)$ отримаємо інтерполяційний многочлен для кутової швидкості:

$$\Omega_7(t) = 33.586t - 3.745t^2 + 5.107t^3 - 3.816t^4 + 1.19t^5$$

Продиференціювавши $\Omega_7(t)$ та домноживши на $J = mR^2/2 = 2.5\,кг\,м^2$, отримаємо інтерполяційний многочлен для моменту двигуна:

$$M_7(t) = 83.965 - 18.728t + 38.303t^2 - 38.157t^3 + 14.873t^4$$

Протабулювавши $M_7(t)$ в моменти часу t_i отримаємо результуючу таблицю:

i	1	2	3	4	5	6
t_i	0	0.2	0.4	0.6	0.8	1.0
M_i(Нм)	83.965	81.47	80.541	80.203	80.052	80.256

Завдання для самостійного розв'язання

Задача 1. Двигун з попередньої задачі випробовується в режимі рекуперації на вільному пробігу на тому ж обладнанні. Дані кута повороту при зупинці приведено в Таблиці 2.13. Час відраховується від моменту початку гальмування, початковий кут відмінний від нуля. Приймаючи до уваги, що в момент зупинки кутова швидкість вала двигуна рівна нулеві, визначити гальмівний момент. Результати подати у вигляді інтерполяційного многочлену для моменту та таблиці його значень.

i	1	2	3	4	5	6					
t_i	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\sin_i$	-0.93	-0.99	-0.52	0.86	-0.12	-0.19	-0.09	0.83	-0.6	-0.99	-0.86
$\cos_i$	0.36	-0.02	-0.84	-0.49	0.99	-0.98	0.99	-0.54	-0.8	0.1	0.5

2.4 Скінченні різниці та інтерполяційні многочлени для рівних проміжків

В попередніх розділах на вузли інтерполювання не накладалось жодних умов (окрім скінченності та незбігання). У випадку накладання деяких умов, схеми побудови інтерполяційних многочленів можуть бути спрощеннями.

Вважаємо, що вузли інтерполювання утворюють рівномірну сітку. В цьому випадку, вони повністю визначені заданням деякої (взагалі кажучи - довільної) x_0 точки та кроком сітки h. Вузли таблиці задаються індексами: $x_i = x_0 + i\,h, \quad i = -\overline{m,n}, m,n \in N \cup \{0\}$. Вузлові значення функції також індексуються $f_i = f(x_i)$.

Озн. Скінченними різницями для вузла i називають величини:

$\Delta f_i = f_{i+1} - f_i$ - скінченна різниця 1-го порядку вперед;

$\nabla f_i = f_i - f_{i-1}$ - скінченна різниця 1-го порядку назад;

$\delta f_i = f_{i+1/2} = f_{i+1} - f_i$ - центральна скінченна різниця 1-го порядку на вузлі $i + 1/2$…. (2.27)

$\Delta^m f_i = \Delta(\Delta^{m-1} f_i) = \Delta^{m-1} f_{i+1} - \Delta^{m-1} f_i$ - скінченна різниця m-го порядку вперед;

$\nabla^m f_i = \nabla(\nabla^{m-1} f_i) = \nabla^{m-1} f_i - \nabla^{m-1} f_{i-1}$ - скінченна різниця m-го порядку назад;

$\delta^m f_i = \delta(\delta^{m-1} f_i) = \delta^{m-1} f_{i+1} - \delta^{m-1} f_i \equiv f^{m-1}_{i+1/2} - f^{m-1}_{i-1/2}$ - центральна скінченна різниця m-го порядку на вузлі i.

Аналогом таблиці розділених різниць в цьому випадку є таблиця скінченних різниць:

Таблиця 2.10

i	f	f^1	f^2	…	f^{n-1}
0	f_0				
1	f_1	$f^1_{1/2}$			
2	f_2	$f^1_{3/2}$	f_1^2		
…	…	…	…	…	
$n-1$	f_{n-1}	$f^1_{n-3/2}$	f^2_{n-2}	…	$f^{n-1}_{(n-1)/2}$

Таблиця скінченних різниць згідно з (2.27) будується аналогічно до таблиці скінченних різниць з тією різницею, що не потрібно різницю скінченних різниць попереднього порядку ділити на різницю значень вузлів.

Властивість лінійності для скінченних різниць залишається справедливою. Властивість симетричності аргументів втрачає сенс, оскільки в даному випадку вузли вважаються впорядкованими. Лема про характеризацію скінченних різниць набуває вигляду.

Лема. Скінченна різниця m-го порядку записується через вузлові значення функції згідно з формулою:

$$f_i^m = \sum (-1)^j C_m^j f_{i+m/2-j} \tag{2.28}$$

при цьому, i- ціле при парному m та i- напівціле при непарному m, так, що $i + m/2 - j$ - завжди ціле.

Справедливість (2.28) можна довести через метод математичної індукції.

Зв'язок скінченних та розділених різниць визначається лемою.

Лема. Якщо $x_i = x_0 + ih$, то

$$f(x_i;...;x_{i+l}) = f_{i+m/2}^m /(h^m m!) \tag{2.29}$$

Справедливість (2.29) також доводиться через метод математичної індукції.

Озн. Аналогом інтерполяційного многочлену з розділеними різницями (2.14) є **інтерполяційний многочлен Ньютона для рівних проміжків (з скінченними різницями) вперед**:

$$\begin{aligned} L_n(x_0 + ht) = N_n^+(t) = f_0 + f_{1/2}^1 t + f_1^2 t(t-1)/2! + ... \\ + f_{(n-1)/2}^{n-1} t(t-1)...(t-(n-2))/(n-1)! \end{aligned} \tag{2.30}$$

Побудова многочлену зводиться до перетворення (2.14) за формулою (2.29).

Озн. Інтерполяційним многочленом Ньютона для рівних проміжків (з скінченними різницями) назад називається інтерполяційний многочлен, побудований на вузлах $x_i = x_0 + ih,\ i = \overline{0,-(n-1)}$:

$$\begin{aligned} L_n(x_0 + ht) = N_n^-(t) = f_0 + f_{-1/2}^1 t + f_{-1}^2 t(t+1)/2! + ... \\ + f_{-(n-1)/2}^{n-1} t(t+1)...(t+(n-2))/(n-1)! \end{aligned} \tag{2.31}$$

Зауваження. Многочлени $N_n^+(t), N_n^-(t)$, є функціями від $t = (x - x_0)/h$. Якщо в Таблиці 2.3 формально перенумерувати вузли, змінивши їх знак, (2.30) та (2.31) будуть збігатись. Такого ж результату можна досягти, якщо проіндексувати вузли

в зворотному напрямку $x_i = x_{n-1} + ih, i = \overline{0,-(n-1)}$. В цьому випадку, (2.31) буде аналогом (2.17). Якщо ж x_0 для (2.30) та (2.31) є одним вузлом, многочлени $N_n^+(t), N_n^-(t)$ різняться, оскільки вони будуть будуватися на різних вузлах. Головним фактом є те, що на n вузлах можна побудувати лише один многочлен степені $n-1$. Різні назви інтерполяційних многочленів означають лише різні схеми їх побудови.

Залишковий член інтерполювання в змінній $t = (x - x_0)/h$ має вигляд:

$$f(x_0 + ht) - N_n^{\pm}(t) = f^{(n)}(\zeta)\omega_n^{\pm}(\zeta)h^n / n! \qquad (2.32)$$

де $\omega_n^{\pm}(t) = t(t \pm 1)...(t \pm (n-1))$ - канонічні многочлени, що легко отримати з (2.6).

2.5 Інтерполяційні многочлени Гауса та Беселя

У випадку фіксації порядку інтерполювання але наявної можливості вибору вузлів інтерполювання, оптимальним є максимальна симетризація вузлів відносно точки, в якій підраховується значення інтерполяційного многочлену, оскільки, в цьому випадку (при рівних проміжках) $|\omega_n(x)| = \prod_{i=1}^{n}|x - x_i|$ приймає мінімальне значення (див. (2.7)). Саме цей принцип лежить в основі інтерполяційних формул Гауса та Беселя.

Озн. Якщо $x \in (x_0, x_0 + h/2]$ природнім є вибір вузлів виду $x_0, x_0 + h, x_0 - h, ..., x_0 + lh, x_0 - lh, ...$. Інтерполяційний многочлен, побудований в таких припущеннях називається **інтерполяційним многочленом Гауса вперед**. Для випадку, коли $x \in [x_0 - h/2, x_0)$ вузлів вибираються у послідовності: $x_0, x_0 - h, x_0 + h, ..., x_0 - lh, x_0 + lh, ...$ Відповідно, на них будується **інтерполяційним многочленом Гауса назад**.

Для прикладу, побудуємо інтерполяційний многочлен Гауса вперед по $2n+2$ вузлам $x_0, x_1, x_{-1}, x_2 ..., x_{-(n-1)}, x_n, x_{-n}, x_{n+1}$. Задля цього модифікуємо Таблицю 2.10:

Таблиця 2.11

i	f	f^1	f^2	…	f^{2n+1}
$-n$	f_{-n}				
$-(n-1)$	$f_{-(n-1)}$	$f^1_{-n+1/2}$			
…	…	…	…		
-1	f_{-1}	$f^1_{-3/2}$	f^2_{-2}		
0	f_0	$f^1_{-1/2}$	f^2_{-1}	…	
1	f_1	$f^1_{1/2}$	f^2_0		
2	f_2	$f^1_{3/2}$	f^2_1		
…	…	…	…	…	
$n+1$	f_{n+1}	$f^1_{n+1/2}$	f^2_n	…	$f^{2n+1}_{1/2}$

та використаємо ланцюжок, що починається з 0: $\left\{f_0, f^1_{1/2}, f^2_0, f^3_{1/2},..., f^{2n}_0, f^{2n+1}_{1/2}\right\}$.

Інтерполяційний многочлен Гауса вперед матиме вигляд:

$$L_{2n+2}(x_0+ht) = G^+_{2n+2}(t) = f_0 + f^1_{1/2}t + f^2_0 t(t-1)/2! + f^3_{1/2}t(t^2-1)/3! + ... + f^{2n+1}_{1/2}t(t^2-1)...(t^2-n^2)/(2n+1)! \quad (2.33)$$

Для того, щоб не перебудовувати Таблицю 2.11, вважатимемо, що $x \in [x_1, x_1 - h/2)$ (при цьому набір вузлів залишається незмінним). Інтерполяційний многочлен Гауса назад по $2n+2$ вузлам $x_1, x_0, x_2, x_{-1}, x_3 ..., x_n, x_{-(n-1)}, x_{n+1}, x_{-n}$ буде вибудовуватись по ланцюжку Таблиці 2.4 $\left\{f_1, f^1_{1/2}, f^2_1, f^3_{1/2},..., f^{2n}_1, f^{2n+1}_{1/2}\right\}$ та матиме вигляд:

$$L_{2n+2}(x_0+ht) = G^-_{2n+2}(t) = f_1 + f^1_{1/2}(t-1) + f^2_1 t(t-1)/2! + f^3_{1/2}t(t-1)(t-2)/3! + ... + f^{2n+1}_{1/2}t(t^2-1)...(t^2-(n-1)^2)(t-n)(t-(n+1))/(2n+1)! \quad (2.34)$$

Звичайно, в даному прикладі $G^+_{2n+2}(t)$ та $G^-_{2n+2}(t)$ збігаються. Але якщо $x \in [x_0 - h/2, x_0)$, потрібно було б будувати многочлен по іншому набору вузлів $x_0, x_{-1}, x_1, x_{-2}, x_2 ..., x_n, x_{-(n+1)}$. При цьому Таблиця 2.11 мала б бути перебудована та $G^+_{2n+2}(t)$ і $G^-_{2n+2}(t)$ були б різними.

Озн. **Многочленом Беселя** є середнє для $G^{+}_{2n+2}(t)$ і $G^{-}_{2n+2}(t)$:

$$B_{2n+2}(t)=\left(G^{+}_{2n+2}(t)+G^{-}_{2n+2}(t)\right)/2=\mu f_{1/2}+f^{1}_{1/2}(t-1/2)+\mu f^{2}_{1/2}t(t-1)/2!...+ \\ +f^{3}_{1/2}t(t-1)(t-1/2)/3!+...+f^{2n+1}_{1/2}t(t^2-1)...(t^2-(n-1)^2)(t-n)(t-1/2)/(2n+1)! \quad (2.35)$$

де $\mu f^{j}_{i+1/2}=\left(f_i{}^{j}+f^{j}_{i+1}\right)/2$

Многочлен (2.35) є інтерполяційним многочленом, оскільки він збігається як з $G^{+}_{2n+2}(t)$ так і з $G^{-}_{2n+2}(t)$. Він є оптимальним, якщо $x\in[x_0,x_1]$. У випадку, коли $x\in[x_0-h/2,x_0+-h/2]$ та вузли для побудови $G^{+}_{2n+2}(t)$ і $G^{-}_{2n+2}(t)$ різні, $B_{2n+2}(t)$ взагалі кажучи не є інтерполяційним многочленом, оскільки в точках $x_{-(n+1)}$ та x_{n+1} умови інтерполювання не виконуються.

2.6 Чисельне диференціювання

Озн. Під **чисельним диференціюванням** функції розуміють наближення похідної функції через задання її точкових значень.

Одним зі способів чисельного диференціювання є побудова та диференціювання її інтерполяційного многочлена:

$$f^{(k)}(x)\approx L_n{}^{(k)}(x) \quad (2.36)$$

Точність такого наближення можна оцінити за залишковим членом інтерполювання:

$$f^{(k)}(x)-L_n{}^{(k)}(x)=\left(f(x;x_1;...;x_n)\omega_n(x)\right)^{(k)}= \\ =\sum_{i=0}^{k}C^{i}_{k}\left(f(x;x_1;...;x_n)\right)^{(i)}\omega_n{}^{(k-i)}(x) \quad (2.37)$$

Згідно з інтерпретацією (2.22), диференціювання функції в точці, з точністю до факторіального множника еквівалентне побудові на цій точці кратної розділеної різниці також порядку. Тоді:

$$f^{(k)}(x)-L_n{}^{(k)}(x)=\sum_{i=0}^{k}k!/(k-i)!\,f(x;...x;x_1;...;x_n)\omega_n{}^{(k-i)}(x) \quad (2.38)$$

де $f(x;...x;x_1;...;x_n)$ - розділена різниця $n+j$ порядку з $j+1$ кратним вузлом x. (2.38) дає змогу записати оцінку похибки формули чисельного диференціювання:

$$\left| f^{(k)}(x) - L_n{}^{(k)}(x) \right| \le \sum_{i=0}^{k} k!/(k-i)!/(n+i)!\ \max \left| f^{(n+i)}(x) \right\| \omega_n{}^{(k-i)}(x) \Big| \quad (2.39)$$

В частинних випадках (2.36) дає змогу отримати формули наближеного диференціювання.

Оскільки $L_n(x)$ - многочлен степені $n-1$, то очевидно: $f^{(n-1)}(x) \approx (n-1)!\, f(x_1;...;x_n) = const$. Якщо похідна вираховується в вузлі інтерполювання, а вузли складають рівномірну сітку, формули спрощуються.

Запишемо широковживану формулу дискретного диференціювання 2-го порядку в центральній точці рівномірної сітки. Нехай сітка складає $2n+1$ вузол $x_0, x_{-1}, x_1, x_{-2}, x_2 ..., x_n, x_{-n}$ так, що $x_i - x_{i-1} = h,\ i = \overline{-(n+1), n}$. Аналогічно до (2.33):

$$L_{2n+1}(x_0 + ht) = N_{2n+1}(t) =$$
$$\{ f_0 + f_0^2 t^2/2! + f_0^4 t^2(t^2-1)/4! + ... + f_0^{2n} t^2(t^2-1)...(t^2-(n-1)^2)/(2n-1)!\} +$$
$$\{ f_{1/2}^1 t - f_0^2 t/2! + f_{1/2}^3 t(t^2-1)/3! + f_0^4 t(t^2-1)(-2)/4! + ...$$
$$+ f_{1/2}^{2n-1} t^2(t^2-1)...(t^2-(n-1)^2)/(2n-1)! + f_0^{2n} t(t^2-1)...(t^2-(n-1)^2)(-n)/(2n)!\}$$

Вираз розбито на парну та непарну по t частини. Отже, друга похідна (по t) другої частини в $t=0$ рівна 0. Очевидно також, що $f''(x_0) \approx N_{2n+1}''(0)/h^2$. Порахувавши другу похідну по t другої частини в $t=0$ отримаємо:

$$f''(x_0) \approx 2/h^2 \sum_{j=1}^{n} (-1)^j ((j-1)!)^2/(2j)!\, f_0^{2j} \quad (2.40)$$

Оцінка похибки може бути отримана з (2.38). Очевидно,

$$f''(x_0) - N_{2n+1}''(0)/h^2 = h^{2n+1}(f^{2n+4}(\zeta_1)/(2n+4)!\, \omega_{2n+1}^*(0) +$$
$$+ 2 f^{2n+3}(\zeta_2)/(2n+3)!\, \omega_{2n+1}^*{}'(0)/h + f^{2n+2}(\zeta_3)/(2n+2)!\, \omega_{2n+1}^*{}''(0)/h^2)$$

де $\omega_{2n+1}^*(t) = t\prod_{i=1}^{n}(t^2 - i^2)$ - канонічний многочлен в змінних t. Очевидно, $\omega_{2n+1}^*(0) = \omega_{2n+1}^*{}''(0) = 0$ та $\omega_{2n+1}^*{}'(0) = (-1)^n (n!)^2$, а отже, справедлива оцінка:

$$\left\| f(x) - N_{2n+1}''(x) \right\|_{C_{[y_1,y_2]}} \le h^{2n} \sup_{\zeta \in [y_1,y_2]} \left| f^{(2n+2)}(\zeta) \right| 2(n!)^2/(2n+2)! \quad (2.41)$$

Зауваження. Згідно з (2.41) наближення другої похідної з використанням симетричних вузлів є найкращим. Найкращим в розумінні оцінки похибки є

також формула чисельного диференціювання першого порядку, побудована на $2n$ вузлах $x_0, x_{-1}, x_1, x_{-2}, x_2 ..., x_{n-1}, x_{-(n-1)}, x_n$.

$$f'(x_0) \approx L_{2n}'(x_0) = N_{2n}'(0)/h = \\ (f_{1/2}^1 - f_0^2/2! + f_{1/2}^3(-1)/3! + f_0^4(-1)(-2)/4! + ... \\ + f_{1/2}^{2n-1}(-1)...(-(n-1)^2)/(2n-1)! + f_0^{2n}(-1)...(-(n-1)^2)(-n)/(2n)!)/h \quad (2.42)$$

2.7 Ортогональні системи. Поліноми Чебишева та інші ортогональні многочлени.

Озн. Система n функцій $\{\varphi_i\}_{i=1}^n$ в гільбертовому просторі H зі скалярним добутком (f, g) та нормою $\| f \| = \sqrt{(f, f)}$ називається **лінійно незалежною**, якщо рівняння $\sum_{i=1}^n C_i \varphi_i = 0$ має лише тривіальний розв'язок ($C_i = 0, i = \overline{1,n}$).

Озн. Система n функцій $\varphi_{(n)} = \{\varphi_i\}_{i=1}^n$ в гільбертовому просторі H називається **ортогональною**, якщо $(\varphi_i, \varphi_j) = 0, \forall i, j = \overline{1,n}, i \neq j$. Якщо окрім того $\|\varphi_i\| = 1, \forall i = \overline{1,n}$, система називається **ортонормальною**.

Лема. Будь яка лінійно незалежна система функцій $\{\varphi_i\}_{i=1}^n$ може бути ортогоналізована за ітераційною процедурою:

$$\psi_j = \sum_{i=1}^j b_{ji} \varphi_i \text{ де } b_{jj} = 1, j = \overline{1,n} \quad (2.43)$$

Побудована ортогональна система функцій $\psi_{(n)} = \{\psi_i\}_{i=1}^n$ також є лінійно незалежною.

Таким чином, ортогоналізація системи функцій зводиться до лінійного відображення:

$$\psi_{(n)} = \mathbf{B_n} \varphi_{(n)} \quad (2.44)$$

з нижньотрикутною матрицею $\mathbf{B_n}$ на головній діагоналі якої стоять 1. Отже існує зворотнє лінійне відображення:

$$\varphi_{(n)} = \mathbf{A_n} \psi_{(n)} \quad (2.45)$$

з нижньотрикутною матрицею $\mathbf{A_n} = \mathbf{B_n}^{-1}$ на головній діагоналі якої стоять 1.

Важливими з точки зору базових чисельних методів гільбертових просторів є:

1. H - простір комплексно значних інтегрованих за квадратом модуля з додатньо визначеною ваговою функцією $p(x) > 0, x \in [a,b]$ на відрізку $[a,b]$ зі скалярним добутком:

$$(f,g) = \int_a^b p(x)\, f(x)\, \overline{g}(x) dx \qquad (2.46)$$

2. Простір комплексно значних n-вимірних векторів $(x_1, x_2, ..., x_n)$ зі скалярним добутком:

$$(x,y) = \sum_{j=1}^{n} p_{ij} x_i \overline{y}_j \qquad (2.47)$$

де $P = \{p_{ij}\}_{i,j=1}^{n}$ - дійсно значна, додатньо визначена вагова матриця.

З огляду на Лему, оскільки $x_{(n)} = \left\{x^{i-1}\right\}_{i=1}^{n}$ є лінійно незалежною в сенсі скалярного добутку (2.46) для довільних скінченних a та b, для різних вагових функцій $p(x) > 0, x \in [a,b]$ (з точністю до множини точок міри 0), можна побудувати системи ортогональних на $[a,b]$ многочленів за рекурентним правилом:

$$P_k{}^p(x) = x^{k-1} - \sum_{i=1}^{k} a_{ki} P_{k-1}{}^p(x),\ k = \overline{1,n} \qquad (2.48)$$

де $P_k{}^p(x)$ - многочлен степені k, ортогональний з вагою $p(x)$, а

$$a_{ki} = \int_a^b p(x) x^{k-1} x^{i-1} dx,\ k = \overline{1,n},\ i \le k \qquad (2.49)$$

Озн. Многочлени, ортогональні на відрізку $[-1,1]$ з вагою $p(x) = (1-x)^{\alpha}(1-x)^{\beta}$, $\alpha, \beta > -1$ називають **ортогональними многочленами Якобі.**

Для побудови многочленів Якобі існує декілька загальних алгоритмів. (2.48) – лише один з них. Поширеними з них є використання генераторної функції:

$$P_k^{(\alpha,\beta)}(x) = (-1)^n / (2^n n!)(1-x)^{-\alpha}(1+x)^{-\beta} d^n((1-x)^{\alpha+n}(1+x)^{\beta+n}) / dx^n \qquad (2.50)$$

та реалізація рекурентних формул побудови.

Важливими з точки зору інтерполяційного аналізу є частинні випадки ортогональних многочленів:

1. Многочлени Лежандра – ортогональні многочлени Якобі з вагою $p(x) \equiv 1$.

Генерація таких многочленів має вигляд:

$$L_n(x) = 1/(2^n n!) d^n((x^2-1)^n)/dx^n \quad (2.51)$$

Рекурентна формула:

$$(n+1)L_{n+1}(x) - (2n+1)xL_n(x) + nL_{n-1}(x) = 0 \quad (2.52)$$

2. Многочлени Чебишева - многочлени Якобі з вагою $p(x) = (1-x)^{-1/2}(1+x)^{-1/2}$.

Рекурентна формула:

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0 \quad (2.53)$$

3. Многочлени Ерміта – ортогональні многочлени на $[-\infty,+\infty]$ з вагою $p(x) = \exp\left(-x^2\right)$ можуть бути згенеровані:

$$H_n(x) = (-1)^n \exp\left(x^2\right) d^n\left(\exp\left(-x^2\right)\right)/dx^n \quad (2.54)$$

або побудовані за рекурентними формулами:

$$H_{n+1}(x) - 2xH_n(x) + 2nH_{n-1}(x) = 0 \quad (2.55)$$

4. Многочлени Лагерра – ортогональні многочлени на $[0,+\infty]$ з вагою $p(x) = x^\alpha \exp\left(-x^2\right)$ можуть бути згенеровані:

$$L_n^\alpha(x) = (-1)^n x^{-\alpha} \exp(x) d^n\left(x^{\alpha+n}\exp(-x)\right)/dx^n \quad (2.56)$$

або побудовані за рекурентними формулами:

$$L_{n+1}^\alpha(x) - (x-\alpha-2n-1)L_n^\alpha(x) + n(\alpha+n)L_{n-1}^\alpha(x) = 0 \quad (2.57)$$

Ортогональні многочлени мають ряд важливих властивостей.

Лема 1. Нехай $\{P_i(x)\}_{i=0}^n$ - система ортогональних на $[a,b]$ многочленів. Тоді кожен з таких многочленів має в точності n різних нулів на інтервалі (a,b).

Лема може бути легко доведена від супротивного.

Лема 2. Нехай $a < x_1^{(n)} < x_2^{(n)} < \ldots < x_n^{(n)} < b$ - нулі многочлена $P_n(x)$. Тоді нулі многочленів $P_n(x)$ та $P_{n-1}(x)$ перемежаються:

$$a < x_1^{(n)} < x_1^{(n-1)} < \ldots < x_{n-1}^{(n-1)} < x_n^{(n)} < b$$

Лема 3. Всі многочлени парної степені $P_{2l}(x)$ - парні функції, непарної степені - $P_{2l+1}(x)$ - непарні функції: $P_n(-x) = (-1)^n P_n(x)$

Особливу роль в теорії інтерполювання відіграють многочлени Чебишева - многочлени Якобі з вагою $p(x)=(1-x)^{-1/2}(1+x)^{-1/2}$. Многочлени Чебишева будуються за рекурентною формулою (2.53) та мають цікаву функціональну інтерпретацію, що спрощує аналіз їх властивостей:

$$T_n(x)=\cos(n\arccos(x)) \qquad (2.58)$$

Згідно з (2.53) головний коефіцієнт $T_n(x)$ (при $n\neq 0$) рівний $T_n(x)$ 2^{n-1}. Нулі $T_n(x)$ згідно з (2.58) $x_m=\cos\left(\pi(2m+1)/(2n)\right)$, $m=\overline{0,n-1}$ є однократними, дійсними та розташовані в $(-1,1)$ (що підтверджується Лемою 1.). Екстремуми $T_n(x)$ перемежаються з нулями: $x_{(m)}=\cos\left(\pi m/n\right)$, $m=\overline{0,n}$, та задають рівні за модулем значення 1: $T_n(x_{(m)})=(-1)^m$.

Озн. Многочлени Чебишева, приведені до канонічного вигляду $\overline{T}_n(x)=2^{1-n}T_n(x)$ називають **многочленами, що найменше відхиляються від нуля.**

Така назва пояснюється важливою властивістю:

Лема 4. Для всіх многочленів степені n, приведених до одиничного головного коефіцієнта, виконується нерівність:

$$\max_{[-1,1]}\left|P_n(x)\right|\geq\max_{[-1,1]}\left|\overline{T}_n(x)\right|=2^{1-n} \qquad (2.59)$$

Найбільше відхилення $\overline{T}_n(x)$ від нуля на $[-1,1]$ рівне 2^{1-n}. Нулі $\overline{T}_n(x)$ збігаються з нулями $T_n(x)$.

Задання многочленів Чебишева на відрізку $[-1,1]$ є універсальним. Вочевидь, для довільних скінченних a та b, шляхом заміни змінної $x=(a+b)/2+(b-a)/2t$ (зі зворотнім перетворенням $t=2/(b-a)(x-(b+a)/2)$) можна перейти від $t\in[-1,1]$ до $x\in[a,b]$. Отже, многочлен $T_n^{[a,b]}(x)=T_n(2/(b-a)(x-(b+a)/2))$ є ортогональним многочленом Чебишева на $[a,b]$ з вагою $p(x)=(b-x)^{-1/2}(x-a)^{-1/2}$. Очевидно, що головний коефіцієнт $T_n^{[a,b]}(x)$ дорівнює $2^{n-1}\left(2/(b-a)\right)^n$. Отже, многочлен, що найменше відхиляється від нуля (серед усіх канонічних многочленів степені n) є многочлен

$\overline{T}_n^{[a,b]}(x) = (b-a)^n 2^{1-2n} T_n\big(2/(b-a)(x-(b+a)/2)\big)$. Найбільше відхилення $\overline{T}_n^{[a,b]}(x)$ від нуля на $[a,b]$ рівне $(b-a)^n 2^{1-2n}$. Нулі $\overline{T}_n^{[a,b]}(x)$:. $x_n = (a+b)/2 + (b-a)/2\cos\big(\pi(2m+1)/(2n)\big),\ m = \overline{0,n-1}.$

З точки зору теорії інтерполювання, властивості мінімального відхилення многочлена Чебишева вказують на існування інтерполяційного многочлену (при заданному порядку інтерполювання), що є найкращим в розумінні (2.7).

Дійсно, при заданній функції та порядку інтерполювання, мінімізація похибки в (2.7) повністю визначається можливістю мінімізації $\sup\limits_{[y_1,y_2]} |\omega_n(x)|$. На заданій рівномірній сітці, така мінімізація можлива (як було показано (2.33)- (2.35)) шляхом симетризації вузлів інтерполювання. Нехай, існує можливість вибору вузлів інтерполювання на $[a,b]$. Якщо за вузли інтерполювання обрати нулі $T_n^{[a,b]}(x)$ (це можливо за лемою 1 або ж за їх явним виглядом x_n), то $\omega_n(x)$ буде приведеним до канонічного виду ортогональним на $[a,b]$ многочленом Чебишева, а тобто, многочленом, що найменше відхиляється від нуля на $[a,b]$ $\overline{T}_n^{[a,b]}(x)$. За Лемою 4 не буде існувати жодного іншого канонічного многочлена степені n з меншим відхиленням. Оцінка похибки в цьому випадку має вигляд:

$$\| f(x) - L_n(x) \| \le \| f^{(n)}(x) \| (b-a)^n 2^{1-2n} / n! \qquad (2.60)$$

2.8 Чисельне інтегрування. Квадратурні формули Ньютона-Котеса.

Озн. **Квадратурною формулою** називається вираз, що дозволяє наближено відшукати інтегральну характеристику функції через її вузлові значення.

Озн. Квадратурна формула, що базується на інтерполюванні підінтегральної функції називається **квадратурною формулою Ньютона-Котеса.**

Отже під квадратурною формулою Ньютона-Котеса скоріше розуміється спосіб побудови такої формули:

$$S_n^{[a,b]}[f] \equiv \int_a^b p(x) L_n(x)\,dx \approx I_{[a,b]}^{p}[f] \equiv \int_a^b p(x)\, f(x)\,dx \qquad (2.61)$$

де $p(x) > 0$ - **вага квадратурної формули**, $L_n(x)$ - інтерполяційний многочлен $f(x)$ на $[a,b]$.

Озн. **Порядком квадратурної формули** називається порядок інтерполяційного многочлену, на якому вона побудована, вузли інтерполяційного многочлен називаються **вузлами квадратурної формули**.

Озн. **Залишком квадратурної формули** називається інтегральне значення залишку інтерполювання:

$$R_n[f] \equiv \int_a^b p(x)(f(x) - L_n(x))\,dx \qquad (2.62)$$

Техніка Ньютона-Котеса передбачає зведення інтегралу до стандартного проміжку інтегрування $[-1,1]$, що надає універсалізму як при побудові квадратурної формули, оцінки її точності так і практичному використанні.

$$\begin{aligned} &\int_a^b p(x)\, f(x)\,dx = \left[x = (b+a)/2 + (b-a)/2\,t,\ dx = (b-a)/2\,dt\right] = \\ &= (b-a)/2 \int_{-1}^{1} p(x(t))\, f(x(t))\,dt = (b-a)/2 \int_{-1}^{1} p^0(t)\, \tilde{f}(t)\,dt \end{aligned} \qquad (2.63)$$

Нехай $\{d_j\}_{j=1}^{n} \in [-1,1]$ - вузли інтерполювання (вузли квадратурної формули). Записавши інтерполяційний многочлен у формі Лагранжа:

$$\begin{aligned} &L_n(t) = \sum_{j=1}^{n} f(x_j) \prod_{i \neq j} (t - d_j)/(d_i - d_j) \\ &x_j = (b+a)/2 + (b-a)/2\,d_j,\ j = \overline{1,n} \end{aligned} \qquad (2.64)$$

отримаємо:

$$\begin{aligned} &S_n^{[a,b]}[f] \equiv \int_a^b p(x) L_n(x)\,dx = (b-a)/2 \sum_{j=1}^{n} D_j\, f((b+a)/2 + (b-a)/2\,d_j) \\ &D_j = \int_{-1}^{1} p^0(t) \prod_{i \neq j} (t - d_i)/(d_j - d_i)\,dt \end{aligned} \qquad (2.65)$$

де $p^0(t) = p((b+a)/2 + (b-a)/2\,t)$.

Озн. Коефіцієнти лінійної комбінації вузлових значень функції $D_j, j=\overline{1,n}$ в квадратурній формулі Ньютона-Котеса (2.65) називаються **вагами квадратурної формули**.

Зауваження. Для побудови квадратурної формули достатньо задати вузли квадратурної формули на стандартному проміжку [–1,1], після чого, ваги квадратурної формули знаходяться за правилом (2.65), а квадратурна формула є лінійна форма побудована на вузлових значеннях функції з ваговими коефіцієнтами:

$$\int_a^b p(x)\, f(x)\, dx \approx (b-a)/2 \sum_{j=1}^{n} D_j\, f\left((b+a)/2 + (b-a)/2\, d_j\right) \qquad (2.66)$$

При побудові квадратурних формул доцільно використовувати наступний очевидний результат:

Лема 1. Якщо вагова функція $p(x)$ - парна відносно $(b+a)/2$ $p((b+a)/2+x) = p((b+a)/2-x)$ та вузли квадратурної формули d_j - симетричні відносно 0 ($d_j = -d_{n+1-j}$), тоді ваги квадратурної формули симетричні ($D_j = D_{n+1-j}$).

Зауваження. Наслідком леми є зокрема те, що при виконанні умов леми, (2.66) є точною для будь якої функції, непарної відносно $(b+a)/2$ (в цьому випадку, очевидно, $\int_a^b p(x)\, f(x)\, dx = S_n^{[a,b]}[f] = 0$).

Зауваження. За методом побудови квадратурної формули, очевидно, що (2.66) є точною для випадку інтегрування довільного многочлену степені не вищої за $n-1$.

Зауваження. При побудові квадратурної формули, можна використовувати інтерполяційні многочлени з кратними вузлами. В цьому випадку замість (2.64) потрібно використовувати (2.23).

Оцінку залишку квадратурної формули можна побудувати, скориставшись оцінкою залишку інтерполювання (2.7):

$$|R_n[f]| = \left|\int_a^b p(x)(f(x) - L_n(x))\, dx\right| \le \max_{[a,b]} \left|f^{(n)}(x)\right| \int_a^b |p(x)|\,|\omega_n(x)|\, dx / n! \qquad (2.67)$$

Виконавши в (2.62) заміну змінних $x=(a+b)/2+(b-a)/2t$ та ввівши позначення:

$$D(d_1,...,d_n)=\int_{-1}^{1}\left|p^0(t)\omega^0(t)\right|dt/n!,\quad \omega^0(t)=\prod_{i=1}^{n}(t-d_i) \tag{2.68}$$

отримаємо:

$$\left|R_n[f]\right|\le\left((b-a)/2\right)^{n+1}D(d_1,...,d_n)\max_{[a,b]}\left|f^{(n)}(x)\right| \tag{2.69}$$

Приклади квадратурних формул.

1. Формула прямокутників. $n=1,\ d_1=0,\ D_1=\int_{-1}^{1}dt=2,\ D(d_1)=\int_{-1}^{1}|t|\,dt=1.$ Квадратурна формула: $\int_a^b f(x)\,dx\approx S_1^{[a,b]}[f]=(b-a)f\left((a+b)/2\right)$ з оцінкою залишку:

$\left|R_1[f]\right|\le\max_{[a,b]}\left|f'(x)\right|(b-a)^2/4$. Зауважимо, що формулу прямокутників можна отримати є простих міркувань: $f(x)\approx L_1(x)=f\left((a+b)/2\right)$ та

$\int_a^b f(x)\,dx\approx f\left((a+b)/2\right)\int_a^b dx=(b-a)f\left((a+b)/2\right)$

В той же час, якщо $n=2,\ d_1=d_2=0$ - вузол є симетричним та кратним, очевидно:

$f(x)\approx L_2(x)=f\left((a+b)/2\right)+f'\left((a+b)/2\right)\left(x-(a+b)/2\right),$

$\int_a^b f(x)\,dx\approx f\left((a+b)/2\right)\int_a^b dx+f'\left((a+b)/2\right)\int_a^b\left(x-(a+b)/2\right)dx=(b-a)f\left((a+b)/2\right)$

Тобто, квадратурна формула не змінилась, але залишок квадратурної формули: $\left|R_2[f]\right|\le\max_{[a,b]}\left|f''(x)\right|(b-a)^3/24$. Окрім того, формула є точною для будь яких многочленів степені не вище за 1 (а не 0). Цей результат є прямим наслідком Леми 1.

2. Формула трапецій. $n=2,\ d_1=-1,d_2=1,\ D_1=D_2=1,\ D(d_1,d_2)=2/3.$

Квадратурна формула: $\int_a^b f(x)\,dx\approx S_2^{[a,b]}[f]=(b-a)/2\left(f(a)+f(b)\right)$ з оцінкою залишку:

$\left|R_2[f]\right|\le\max_{[a,b]}\left|f''(x)\right|(b-a)^3/12.$

3. Формула Сімпсона. $n=3$, $d_1=-1, d_2=1, d_3=0$, $D_1=D_3=1/3, D_2=4/3$, $D(d_1,d_2,d_3)=1/12$. Квадратурна формула: $\int_a^b f(x)\,dx \approx S_3^{[a,b]}[f]=(b-a)/6\big(f(a)+4f((a+b)/2)+f(b)\big)$ з оцінкою залишку: $|R_3[f]| \le \max\limits_{[a,b]}|f^{(3)}(x)|(b-a)^4/192$. В той же час, аналогічно до формули прямокутників з кратним вузлом, для $n=4$, $d_1=-1, d_2=1, d_3=d_4=0$, згідно з Лемою 1:

$$f(x)\approx L_4(x)=f(a)+f(a;b)(x-a)+f(a;b;(a+b)/2)(x-a)(x-b)+$$
$$+f(a;b;(a+b)/2;(a+b)/2)(x-a)(x-b)(x-(a+b)/2)$$

Оскільки інтеграл від останнього члену рівний 0, вираз для квадратурної формули не зміниться. Але залишковий член має вигляд: $|R_4[f]| \le \max\limits_{[a,b]}|f^{(4)}(x)|(b-a)^5/2880$.

2.9 Квадратурні формули Гауса.

При побудові квадратурних формул Ньютона-Котеса, вважалось, що вузли квадратурної формули задаються наперед. Як бачимо, чисельне інтегрування в цьому випадку, зводиться до інтегрування інтерполяційного многочлену, що будується на вузлах квадратурної формули. Саме цей факт дає змогу стверджувати, що квадратурна формула порядку n є точною для всіх многочленів степені не вищою за $n-1$. Техніка Гауса дозволяє будувати квадратурні формули, що є точними для многочленів більш високої степені (при фіксованому n) за рахунок спеціального вибору вузлів квадратурної формули. Ця техніка також використовує наближення функції інтерполяційним многочленом, а отже, формули отримання вагових коефіцієнтів та залишку не відрізняються від методу Ньютона-Котеса.

Коректність такого підходу пояснюється введенням додаткових невідомих в рівняння занулення залишку. Дійсно, нехай будується квадратурна формула порядку N, що є точною для всіх многочленів виду $P_n(x)=\sum_{i=0}^{n} a_i x^i$, степені не

вищої за n. В силу лінійності залишку квадратурної формули $R_N[P_n]=\sum_{i=0}^{n}a_iR_N[x^i]$. В силу довільності коефіцієнтів $\{a_i\}_{i=0}^{n}$, $R_N[P_n]=0$, $\Leftrightarrow R_N[x^i]=0, i=\overline{0,n}$, а отже, для побудови такої квадратурної формули необхідно та достатньо задовольнити рівняння $R_N[x^i]=\int_a^b p(x)x^i dx-(b-a)/2\sum_{j=1}^{N}D_j x_j{}^i=0, i=\overline{0,n}$.

У випадку, коли вузли x_j, $j=\overline{1,N}$ задано (як в техніці Ньютона-Котеса), достатньо відшукати N значень D_j, для чого необхідно використати в точності N рівнянь, а отже, $n=N-1$, що доводить коректність техніки Ньютона-Котеса. У випадку, коли x_j, $j=\overline{1,N}$ також є невідомими, маємо $n+1$ рівняння з $2N$ невідомими $D_{j,}\,x_j$, $j=\overline{1,N}$. Отже, в цьому випадку, $n=2N-1$ і можна сподіватись на побудову квадратурних формул порядку n , точних для довільних многочленів степені, не вищої за $2n-1$. Звичайно, всі знайдені вузли мають розташовуватись в $[a,b]$.

Побудова квадратурних формул Гауса базується на Лемах.

Лема 1. Якщо $\{x_i\}_{i=0}^{n}$ - вузли квадратурної формули, точної для всіх многочленів степені $2n-1$, то $\int_a^b p(x)\omega_n(x)P_{n-1}(x)dx=0$, де $\omega_n(x)=\prod_{i=1}^{n}(x-x_i)$ - канонічний многочлен, побудований на вузлах, а $P_{n-1}(x)$ - довільний многочлен степені $n-1$.

Доведення Леми є очевидним: оскільки $\omega_n(x)P_{n-1}(x)$ є многочленом степені $2n-1$, то за умовою $\int_a^b p(x)\omega_n(x)P_{n-1}(x)dx=(b-a)/2\sum_{j=1}^{n}D_j\omega_n(x_j)P_{n-1}(x_j)=0$.

Лема 1 з необхідністю стверджує, що якщо вдасться побудувати квадратурну формулу порядку n, точну для всіх многочленів степені $2n-1$, то вузлами такої формули мають бути нулі многочлена степені n з системи ортогональних на $[a,b]$ многочленів з вагою $p(x)$ (вважаємо $p(x)>0$ на $[a,b]$). Це слідує з того, що кожен многочлен степені $n-1$ можна представити як лінійну комбінацію ортогональних з $p(x)$ многочленів степенем нижчих за n (довільність коефіцієнтів такого розкладу обумовлюється довільністю $P_{n-1}(x)$). З теорії

ортогональних многочленів слідує, що така система многочленів є єдиною, а отже для заданих n та $p(x)$ на $[a,b]$, $\{x_i\}_{i=0}^{n}$ будуть однозначно визначеними. Окрім того, $x_i \in [a,b], i=\overline{1,n}$, а отже така квадратурна формула (якщо її вдасться побудувати) буде коректною.

Лема 2. Якщо $\{x_i\}_{i=0}^{n}$ - нулі $\psi_n(x)$ (елементу системи ортогональних на $[a,b]$ многочленів з вагою $p(x)$) та квадратурна формулу є точною для всіх многочленів степені $n-1$, то вона є точною для всіх многочленів степені $2n-1$.

Доведення Леми ґрунтується на розкладі довільного многочлену степені $2n-1$ $Q_{2n-1}(x)=\psi_n(x)g_{n-1}(x)+r_{n-1}(x)$, який можна отримати шляхом ділення з залишком $Q_{2n-1}(x)$ на $\psi_n(x)$. Тоді, очевидно: $R_n[Q_{2n-1}]=R_n[\psi_n g_{n-1}]+R_n[r_{n-1}]$. Останній залишок в правій части дорівнює 0 за умовою Леми, а перший: $R_n[\psi_n g_{n-1}]=\int_a^b p(x)\psi_n(x)g_{n-1}(x)dx-(b-a)/2\sum_{j=1}^{n}D_j\psi_n(x_j)g_{n-1}(x_j)=0$ в силу ортогональності $\psi_n(x)$ та вибору вузлів x_j.

Лема 2 фактично визначає можливість побудови такої квадратурної формули за стандартною технікою Ньютона-Котеса. Дійсно, для побудови квадратурної формули Гауса порядку n на $[a,b]$ з ваговою функцією $p(x)$, необхідно і достатньо обрати в якості вузлів нулі $\psi_n(x)$ - елементу системи ортогональних на $[a,b]$ многочленів з вагою $p(x)$ та скористатись формулами Ньютона-Котеса (2.68) для побудови вагових коефіцієнтів. Квадратурна формула Гауса не відрізняється по вигляду від формули Ньютона-Котеса:

$$\int_a^b p(x)\,f(x)\,dx \approx (b-a)/2\sum_{j=1}^{n}G_j\,f\left((b+a)/2+(b-a)/2\,d_j\right) \qquad (2.70)$$

(тут d_j - нулі ортогонального на $[-1,1]$ з вагою $p^0(t)$).

Залишок квадратурної формули Гауса слідує з її властивостей. Дійсно, побудувавши інтерполяційний многочлен $f(x)$ порядку $2n$ з 2-ох кратними вузлами на вузлах квадратурної формули: $g_{2n}(x): g_{2n}(x_i)=f(x_i), s_{2n}'(x_i)=f'(x_i), i=\overline{1,n}$, згідно з (2.24) маємо:

$f(x)-g_{2n}(x)=f(x;x_1;x_1;...;x_n;x_n)\psi_{2n}{}^2(x)$, а отже

$$R_n[f]=R_n\left[f(x;x_1;x_1;...;x_n;x_n)\psi_{2n}{}^2(x)\right]=\int_a^b p(x)f(x;x_1;x_1;...;x_n;x_n)\psi_{2n}{}^2(x)dx-$$
$$-(b-a)/2\sum_{j=1}^n D_j f(x_j;x_1;x_1;...;x_n;x_n)\psi_{2n}{}^2(x_j)$$

де другий доданок дорівнює нулю. Застосувавши до інтегралу теорему про середнє, отримаємо:

$$R_n[f]=f(\zeta_1;x_1;x_1;...;x_n;x_n)\int_a^b p(x)\psi_n{}^2(x)dx=$$
$$=f^{2n}(\zeta_2)\int_a^b p(x)\psi_{2n}{}^2(x)/n!dx,\ \zeta_{1,2}\in[a,b] \qquad (2.71)$$

Вагові коефіцієнти Гауса завжди додатні. Окрім того, $(b-a)/2\sum_{j=1}^n G_j=\int_a^b p(x)dx$. У випадку, коли $p(x)$ - парна відносно $(b+a)/2$, вузли квадратури (нулі відповідного ортогонального многочлена) симетричні відносно $(b+a)/2$, а вагові коефіцієнти симетричні $G_j=G_{n+1-j}$. Для випадку $p(x)\equiv 1$, можлива універсалізація квадратурної формули оскільки в цьому випадку $d_j=2/(b-a)(x_j-(b+a)/2)\in[-1,1]$ не залежать від $[a,b]$ та є нулями многочлена Лежандра. З (2.71) отримаємо:

$$|R_n[f]|\le((b-a)/2)^{2n+1}\max_{x\in[a,b]}\left|f^{2n}(x)\right|D(d_1,...,d_n),$$
$$D(d_1,...,d_n)=\int_{-1}^1 p^0(t)\psi_n^{0\,2}(t)dt/(2n)!,\ \psi_n^0(t)=\prod_{i=1}^n(t-d_i) \qquad (2.72)$$

2.10 Деякі спеціальні випадки побудови квадратурних формул

При практичній побудові та реалізації квадратурних формул виявляються деякі особливості, що потребують додаткових досліджень та побудов. Найбільш широко вживані з них розглянуто в цьому параграфі.

2.11 Складені квадратурні формули

Оцінка залишку (2.69) вказує на очевидний спосіб підвищення точності квадратурної формули шляхом зменшення $(b-a)$ при подрібненні проміжку інтегрування.

Озн. У випадку, коли $p(x)\equiv 1$, вагові коефіцієнти не залежать від меж інтегрування (а для випадку формул Гауса, від меж інтегрування не залежать також і вузли). Комплекса квадратурна формула, побудована в результаті розбиття проміжку інтегрування називається **складеною квадратурною формулою**.

Розглянемо квадратурну формулу порядку n з вузлами $d_j \in [-1,1]$ та вагами D_j: $\int_a^b f(x)\,dx \approx S_n^{[a,b]}[f] = (b-a)/2 \sum_{j=1}^n D_j f((b+a)/2 + (b-a)/2\, d_j)$ з оцінкою залишку $\left|R_n[f]\right| \le \left((b-a)/2\right)^{n+1} D(d_1,...,d_n) \max\limits_{[a,b]} \left| f^{(n)}(x)\right|$.

Виконавши розбиття $[a,b] = \bigcup_{i=1}^N [a_{i-1}, a_i]$ так, що $a_0 = a,\ a_N = b$ маємо:

$$\begin{aligned} &\int_a^b f(x)\,dx = \sum_{i=1}^N \int_{a_{i-1}}^{a_i} f(x)\,dx \approx \tilde{S}_n^{[a,b]}[f] = \sum_{i=1}^N S_n^{[a_{i-1},a_i]}[f] = \\ &\sum_{i=1}^N (a_i - a_{i-1})/2 \sum_{j=1}^n D_j f((a_{i-1}+a_i)/2 + (a_i - a_{i-1})/2\, d_j) \end{aligned} \tag{2.72}$$

з оцінкою залишкового члену:

$$\begin{aligned} &\left|\tilde{R}_n^{[a,b]}[f]\right| = \left|\sum_{i=1}^N R_n^{[a_{i-1},a_i]}[f]\right| \le \sum_{i=1}^N \left|R_n^{[a_{i-1},a_i]}[f]\right| = \\ &= D(d_1,...,d_n) \sum_{i=1}^N \left((a_i - a_{i-1})/2\right)^{n+1} \max_{[a_{i-1},a_i]} \left| f^{(n)}(x)\right| \le \\ &\le D(d_1,...,d_n) \max_{[a,b]} \left| f^{(n)}(x)\right| \sum_{i=1}^N \left((a_i - a_{i-1})/2\right)^{n+1} \end{aligned} \tag{2.73}$$

Як бачимо з (2.73), похибка може бути суттєво зменшена при зменшенні $\max\limits_{i=1,N}(a_i - a_{i-1})$. Зокрема, при рівномірному розбитті $a_i - a_{i-1} = (b-a)/N,\ i = \overline{1,N}$, оцінка похибки має вигляд:

$$\left|\tilde{R}_n^{[a,b]}[f]\right| \le \left((b-a)/2\right)^{n+1} D(d_1,...,d_n) \max_{[a,b]} \left| f^{(n)}(x)\right| N^{-n} \tag{2.74}$$

тобто в N^n разів краща за вихідну.

Приклад 1. Складена формула трапецій.

Формула трапецій має вигляд: $\int_a^b f(x)\,dx \approx S_2^{[a,b]}[f] = (b-a)/2\left(f(a) + f(b)\right)$ з оцінкою $\left|R_2[f]\right| \le \max\limits_{[a,b]} \left| f''(x)\right| (b-a)^3/12$.

Складена формула трапецій згідно з (2.72), (2.74) матиме вигляд:

$\int_a^b f(x)\,dx \approx \widetilde{S}_2^{[a,b]}[f] = (b-a)/2/N\big(f(a) + 2f(a_1) + ... + 2f(a_{N-1}) + f(b)\big)$ з оцінкою залишку: $\left|\widetilde{R}_2^{[a,b]}[f]\right| \le \max\limits_{[a,b]}\left|f''(x)\right|(b-a)^3/12N^{-2}$.

Приклад 2. Складена формула Сімпсона.

Формула Сімпсона з кратним вузлом має вигляд: $\int_a^b f(x)\,dx \approx S_4^{[a,b]}[f] = (b-a)/6\big(f(a) + 4f((a+b)/2) + f(b)\big)$ з оцінкою залишку: $\left|R_4[f]\right| \le \max\limits_{[a,b]}\left|f^{(4)}(x)\right|(b-a)^5/2880$. Розбиваємо $[a,b]$ на N рівних відрізків так, що:

$\int_{a_{i-1}}^{a_i} f(x)\,dx \approx S_4^{[a_{i-1},a_i]}[f] = (a_i - a_{i-1})/6\big(f(a_{i-1}) + 4f((a_{i-1}+a_i)/2) + f(a_i)\big)$.

Отже:

$\int_a^b f(x)\,dx \approx \widetilde{S}_4^{[a,b]}[f] = (b-a)/6/N(f(a) + 4f((a_1+a_2)/2) + 2f(a_2) + ... +$
$+ 2f((a_{N-2}+a_{N-1})/2) + 4f(a_{N-1}) + f(b))$

з оцінкою залишку: $\left|\widetilde{R}_4^{[a,b]}[f]\right| \le \max\limits_{[a,b]}\left|f^{(4)}(x)\right|(b-a)^5/2880\ N^{-4}$.

Приклад 3. Складена формула Гауса.

Нехай маємо формулу Гауса:

$\int_a^b p(x)\,f(x)\,dx \approx (b-a)/2\sum_{j=1}^n G_j\,f\big((b+a)/2 + (b-a)/2\,d_j\big)$ з оцінкою залишку: $\left|R_n[f]\right| \le ((b-a)/2)^{2n+1}\max\limits_{x\in[a,b]}\left|f^{2n}(x)\right|D(d_1,...,d_n)$. Розбиваємо $[a,b]$ на N рівних відрізків так, що:

$\int_{a_{i-1}}^{a_i} f(x)\,dx \approx S_n^{[a_{i-1},a_i]}[f] = (a_i - a_{i-1})/2\sum_{j=1}^n G_j\,f(a_i^j)$,

$a_i^j = (a_{i-1}+a_i)/2 + (a_i - a_{i-1})/2\,d_j, i = \overline{1,N},\ j = \overline{1,n},\ a_i^1 = a_{i-1},\ a_i^n = a_i$.

$\int_a^b f(x)\,dx \approx \widetilde{S}_n^{[a,b]}[f] = (b-a)/2/N\sum_{i=1}^N\sum_{j=1}^n G_j\,f(a_1^j)$ з оцінкою залишку:

$\left|\widetilde{R}_n[f]\right| \le ((b-a)/2)^{2n+1}N^{-2n}\max\limits_{x\in[a,b]}\left|f^{2n}(x)\right|D(d_1,...,d_n)$

2.12 Квадратурні формули для сильно осцилюючих функцій

При підрахунку інтегралів Фур'є необхідно підраховувати функції виду: $\int_a^b f(x)e^{i\omega x}dx$. Оскільки чисельне інтегрування базується на інтерполюванні підінтегральної функції, точність квадратурної формули прямим чином залежить від якості інтерполювання. Очевидно, що кількість нулів дійсної та уявної частини підінтегральної функції в даному випадку має порядок $\omega(b-a)/\pi$, о отже, порядок квадратурної формули n має бути $n >> \omega(b-a)/\pi$. Для великих значень ω ця умова не може бути виконана.

Якщо вважати $e^{i\omega x}$ аналогом вагової функції, можна побудувати спеціальну квадратурну формулу для інтегрування функцій з регулярною осциляцією за методом Ньютона-Котеса.

$$S_n^{\omega}[f] \equiv \int_a^b e^{i\omega x} L_n(x)\,dx = \left[x = (a+b)/2 + (b-a)/2t, dx = (b-a)/2dt\right] =$$
$$= (b-a)/2e^{i\omega(a+b)/2}\int_{-1}^{1} e^{i\omega(b-a)t} L_n((a+b)/2 + (b-a)/2t)\,dt$$

якщо $x_j = (a+b)/2 + (b-a)/2d_j,\ j = \overline{1,n},\quad x_j \in [a,b],\quad d_j \in [-1,1]$ - вузли інтегральної квадратури, то

$$S_n^{\omega}[f] = (b-a)/2e^{i\omega(a+b)/2}\sum_{j=1}^{n} D_j(\omega(b-a)/2)\,f(x_j)$$
$$D_j(p) = \int_{-1}^{1} e^{ipt}\prod_{i\neq j}(t-d_i)/(d_j-d_i)dt \qquad (2.75)$$

і є шукана квадратура. Як бачимо, точність такої квадратурної формули визначається якістю інтерполювання не осцилюючої функції $f(x)$.

Оцінка залишку квадратурної формули визначається як і раніше (2.69):

$$\left|R_n[f]\right| \le \left((b-a)/2\right)^{n+1} D(d_1,...,d_n)\max_{[a,b]}\left|f^{(n)}(x)\right|, \qquad D(d_1,...,d_n) = \int_{-1}^{1}\left|\omega^0(t)\right|dt/n!,$$

$\omega^0(t) = \prod_{i=1}^{n}(t-d_i)$, оскільки, $\left|e^{i\omega x}\right| = 1$.

Приклад 1. Формула Філона. Нехай $n=3,\ d_1=-1, d_2=0, d_3=-1$.

$$D_1(p) = \overline{D_3}(p) = \int_{-1}^{1} e^{ipt}t(t-1)dt = (2p\cos p + (p^2-2)\sin p)/p^3 + i(p\cos p - \sin p)/p^2$$

$$D_2(p) = \int_{-1}^{1} e^{ipt}(1 - t^2 dt = (-2p\cos p + 2\sin p)/p^3$$

Зауваження. Очевидно, для значень $p \approx 0$ для підрахунку вагових коефіцієнтів необхідно скористатись правилом Лопіталя.

2.13 Квадратурні формули для функцій з особливостями

Проблема інтерполювання функції виникає також у випадках, коли підінтегральна функція має степеневі особливості на проміжку інтегрування. Як приклад, можна вказати інтегрування функції з інтегрованою особливістю $f(x) \underset{x \to a}{\approx} A(x-a)^{\alpha} \quad \alpha > -1$. Очевидно, при $-1 < \alpha < 0$ такий інтеграл є невласним та підінтегральна функція не може бути проінтельпольована многочленом скінченної степені. Але, навіть у випадку, $0 < \alpha$, наближення $f(x)$ в околі $x = a$ цілостепеневим многочленом є неефективним. Наприклад, нехай має місце розклад в ряд Тейлора: $f(x)(x-a)^{-\alpha} = A + a_1(x-a) + ...,$ тоді $f(x) = (x-a)^{\alpha}(A + a_1(x-a) + ...)$.

Вирішити вказану проблему можна різними методами.

Так у вказаному випадку, очевидно, $\int_a^b f(x)dx = \int_a^b p(x)\tilde{f}(x)dx$ де $\tilde{f}(x) = f(x)(x-a)^{-\alpha}$, $p(x) = (x-a)^{\alpha}$. А отже, за методом Ньютона-Котеса:

$$\begin{gathered}\int_a^b f(x)\,dx \approx S_n^{[a,b]}[f] = (b-a)/2\sum_{j=1}^{n} D_j (t+1)^{-\alpha} f((b+a)/2 + (b-a)/2\,d_j) \\ D_j = \int_{-1}^{1} (t+1)^{\alpha} \prod_{i \neq j} (t-d_i)/(d_j - d_i)dt\end{gathered} \tag{2.76}$$

При цьому, слід пам'ятати, що в (2.76) член суми для $d_j = -1$ слід розуміти як $\lim_{t \to -1} D_j (t+1)^{-\alpha} f((b+a)/2 + (b-a)/2t)$.

Оцінка похибки, згідно з (2.69):

$$|R_n[f]| \le ((b-a)/2)^{n+\alpha+1} D(d_1,...,d_n) \max_{[a,b]} \left| \left(f(x)(x-a)^{-\alpha} \right)^{(n)} \right| \tag{2.77}$$

$$D(d_1,...,d_n)=\int_{-1}^{1}\left|(t+1)^{\alpha}\omega^0(t)\right|dt/n!,\quad \omega^0(t)=\prod_{i=1}^{n}(t-d_i)$$

Приклад 1. Формула трапецій для функцій зі степеневою особливістю. $n=2$, $d_1=-1, d_2=1.$

$$\int_a^b f(x)\,dx\approx S_2^{[a,b]}[f]=(b-a)/2\left(D_1\lim(x-a)^{-\alpha}f(x)+D_2 f(b)\right)$$

$$D_1=1/2\int_{-1}^{1}(t+1)^{\alpha}(1-t)dt=-2^{2+\alpha}/(2+3\alpha+\alpha^2)$$

$$D_2=1/2\int_{-1}^{1}(t+1)^{\alpha}(1+t)dt=2^{2+\alpha}/(2+\alpha)$$

Оцінка похибки:

$$\left|R_2[f]\right|\le\max_{[a,b]}\left|\left(f(x)(x-a)^{-\alpha}\right)^{(2)}\right|(b-a)^{3+\alpha}/4/(6+5\alpha+\alpha^2).$$

Для підрахунку інтегралу з особливістю можна також застосувати метод подрібнення відрізку інтегрування. Дійсно, $\int_a^b f(x)dx=\int_a^{a+\varepsilon}f(x)dx+\int_{a+\varepsilon}^b f(x)dx,\ \varepsilon<<1$. При цьому, перший інтеграл можна відшукати за (2.76) (з межами інтегрування $[a,a+\varepsilon]$), а другий інтеграл вже не містить особливості, а отже, може бути підрахований за стандартними методами. При цьому, слід пам'ятати, що хоча при $\varepsilon\to 0$ точність визначення 1-го інтегралу прямує до 0 з порядком $\varepsilon^{n+\alpha+1}$, але оцінка точності 2-го інтегралу містить співмножник $\max\limits_{[a+\varepsilon,b]}\left|f^{(n)}(x)\right|$ який при достатньо великому значенні n прямує до нескінченності, а отже оцінка точності може бути завищеною.

Якщо функція має степеневі особливості на обох кінцях відрізку інтегрування: $f(x)\underset{x\to a}{\approx}A(x-a)^{\alpha}$, $f(x)\underset{x\to a}{\approx}B(b-x)^{\beta}$, $\alpha,\beta>-1$, можна скористатись розібраним раніше підходом. Існує і інший шлях, який базується на побудові квадратурної формули Гауса. Очевидно: $\int_a^b f(x)dx=\int_a^b p(x)\widetilde{f}(x)dx$ де $\widetilde{f}(x)=f(x)(x-a)^{-\alpha}(b-x)^{-\beta}$, $p(x)=(x-a)^{\alpha}(b-x)^{\beta}$. Розглянемо ортогональний на $[-1,1]$ з вагою $p(t)=(1-t)^{\alpha}(t-1)^{\beta}$ многочлен Якобі степені $n: P_n^{(\alpha,\beta)}(t)$ (2.50). Згідно з властивостями ортогональних многочленів він має n різних нулів

$d_j \in (-1,1)$. Згідно з методом Гауса, виберемо в якості вузлів квадратурної формули Гауса нулі $P_n^{(\alpha,\beta)}(t)$ та побудуємо за технікою Ньютона-Котеса квадратурну формулу:

$$\int_a^b f(x)\,dx \approx (b-a)/2 \sum_{j=1}^{n} G_j f\left((b+a)/2 + (b-a)/2\, d_j\right)(d_j+1)^{-\alpha}(1-d_j)^{-\beta}$$
$$G_j = \int_{-1}^{1}(t+1)^{\alpha}(1-t)^{\beta}\prod_{i\neq j}(t-d_i)/(d_j-d_i)dt,\ \overline{j=1,n} \qquad (2.78)$$

Оцінка точності знаходиться згідно з (2.72):

$$|R_n[f]| \le ((b-a)/2)^{2n+\alpha+\beta+1} \max_{x\in[a,b]} \left|\left((x-a)^{-\alpha}(x-a)^{-\beta} f(x)\right)^{2n}\right| D(d_1,...,d_n),$$
$$D(d_1,...,d_n) = \int_{-1}^{1}(t+1)^{\alpha}(1-t)^{\beta}\psi_n^{0\,2}(t)dt/(2n)!,\ \psi_n^0(t) = \prod_{i=1}^{n}(t-d_i) \qquad (2.79)$$

При цьому, оскільки нулі ортогональних многочленів знаходяться в $(-1,1)$, (2.78) є коректною формулою.

2.14 Елементи теорії наближення функцій

В попередньому розділі розібрано теорію інтерполювання. На її основі отримано ряд важливих алгоритмів, зокрема для дискретного диференціювання та інтегрування функцій. Ідея побудови таких алгоритмів, вочевидь, полягала в заміні об'єкту операції деяким іншим, більш зручним для чисельного аналізу об'єктом (в даному випадку, функція була замінена її інтерполяційним многочленом). Наближення функції її інтерполяцією базується на виконанні умови інтерполювання. Воно є досить простим, але не завжди зручним для застосування. В цьому розділі буде розглянуто більш загальний підхід до наближення функцій.

Основою теорії наближення функцій в лінійному нормованому просторі є загальна теорема про існування та єдність найкращого наближення.

Озн. Якщо $f \in \Phi$ - елемент лінійного нормованого простору Φ та $\{g_i\}_{i=1}^n \in \Phi$ - система лінійно незалежних функцій, **найкращим лінійним наближенням** f по $\{g_i\}_{i=1}^n$ називається лінійна комбінація $\sum_{i=1}^n c_i^0 g_i \in \Phi$ така, що

$$\left\| f - \sum_{i=1}^{n} c_i^0 g_i \right\| = \Delta = \inf_{c_i, i=\overline{1,n}} \left\| f - \sum_{i=1}^{n} c_i g_i \right\| \qquad (2.80)$$

Теорема. Найкраще лінійне наближення існує. У випадку, коли Φ є строго нормованим простором, таке наближення є єдиним.

Випадок, коли Φ є гільбертовим простором - найбільш цікавий з точки зору наближення саме функцій буде основним предметом подальших досліджень.

В гільбертовому просторі коефіцієнти найкращого наближення можуть бути отримані за методом найменших квадратів (МНК). Згідно з МНК задача наближення еквівалентна мінімізації функціоналу:

$$\delta^2(a_1,...,a_n) = \left\| f - \sum_{i=1}^{n} a_i g_i \right\|^2 = \left(f - \sum_{i=1}^{n} a_i g_i,\ f - \sum_{i=1}^{n} a_i g_i \right) \qquad (2.81)$$

Використавши необхідну умову існування мінімуму функціоналу, можна прирівняти до нуля частинні похідні (3.2), отримавши систему лінійних алгебраїчних рівнянь (СЛАР), розв'язком якої і будуть шукані коефіцієнти.

До такого ж результату можна прийти з більш загальних позицій. Нехай $H \in \Phi$ є лінійний підпростір Φ. Якщо $h_0 \in H$ є найкраще наближення $f \in \Phi$, тобто, $\|f - h_0\| = \inf_{h \in H} \|f - h\|$, тоді:

Лема 1. Якщо $h_0 \in H$ існує, то $f - h_0$ ортогональне до всіх елементів H: $(f - h_0, h) = 0, \forall h \in H$.

Лема 2. Якщо $(f - h_0, h) = 0, \forall h \in H$, то $h_0 \in H$ є найкраще наближення $f \in \Phi$.

Очевидно, $\{g_i\}_{i=1}^{n}$ утворює лінійну оболонку $H \in \Phi$. Якщо $\sum_{i=1}^{n} a_i g_i$ є найкраще наближення $f \in \Phi$, то згідно з Лемою 1:

$$\left(f - \sum_{i=1}^{n} a_i g_i, g_i \right) = 0,\ j = \overline{1,n} \qquad (2.82)$$

Оскільки, згідно з теоремою про існування найкращого наближення воно існує, то (3.3) має розв'язок та він єдиний. (3.3) можна переписати у вигляді СЛАР:

$$\sum_{j=1}^{n} a_j (g_j, g_i) = (f, g_i), i = \overline{1,n} \qquad (2.83)$$

У випадку, коли $\{g_i\}_{i=1}^{n}$ утворюють ортонормовану систему $(g_i, g_j) = \delta_{ij}, i, j = \overline{1,n}$ (якщо ні, то її завжди можна ортогоналізувати (2.44) та віднормувати), матриця

(3.4) стає одиничною, та її розв'язок: $a_i = (f, g_i), i = \overline{1,n}$ і найкраще наближення має вигляд:

$$f \approx g = \sum_{j=1}^{n} (f, g_j) g_j \qquad (2.84)$$

Такий підхід побудови наближення функції називають методом Гальоркіна.

Слід зауважити, що отримані результати лише декларують існування найкращого наближення та вказують шлях його побудови, але жодним чином не дозволяють оцінити якість такого наближення. Очевидно, для різних наборів $\{g_i^{(1)}\}_{i=1}^{n}$ та $\{g_i^{(2)}\}_{i=1}^{n}$ величина похибки може суттєво різнитись. Окрім того, принциповим є питання збіжності наближення, тобто формулювання умов на наближення, при виконанні яких похибка наближення прямуватиме до нуля (в деякому сенсі) при $n \to \infty$. Єдиним результатом в цьому напрямку є рівність:

$$\|f - g\|^2 = (f, f) - \sum_{j=1}^{n} |(f, g_j)|^2 \qquad (2.85)$$

Причиною цього є загальність отриманих результатів, оскільки, при конкретизації задачі наближення необхідно приймати до уваги конкретні особливості.

Як приклад, розглянемо Φ - простір обмежених дійсно значних на $[a,b]$ функцій з нормою: $\|f\| = \sup_{[a,b]} |f(x)|$. Будемо шукати найкраще наближення многочленом так, що $g_i = x^i, i = \overline{0,n}$: $Q_n(x) = \sum_{i=1}^{n} a_i x^i$. Згідно з теоремо, такий многочлен існує і він є єдиним.

Озн. Многочлен $Q_n^0(x) = \sum_{i=1}^{n} a_i^0 x^i : \sup_{[a,b]} |f(x) - Q_n^0(x)| = \inf_{a_i, i=\overline{0,n}} \sup_{[a,b]} |f(x) - Q_n(x)|$ називається **многочленом найкращого рівномірного наближення**.

Побудова многочлена найкращого рівномірного наближення визначається теоремою:

Теорема Чебишева. Для того, щоб $Q_n(x)$ був многочленом найкращого рівномірного наближення, необхідно та достатньо, щоб існував набір принаймні $m = n + 2$ різних точок $x_0 < ... < x_{n+1}$ на $[a,b]$, такий, що:

$$f(x_i) - Q_n(x_i) = \alpha(-1)^i \|f - Q_n\|, \alpha = -1|1, i = \overline{0, n+1} \qquad (2.86)$$

Озн. Набір точок $\{x_i\}_{i=0}^{n}$, що задовольняють умови теореми називається **чебишевським альтернантом.**

Окрім того, многочлен найкращого рівномірного наближення є єдиним.

Приклад. Якщо $f^{(n+1)}(x) \geq 0, x \in [a,b]$ то можна показати, що $Q_n(x)$ незначно відхиляється від інтерполяційного многочлена, побудованого на нулях многочлена Чебишева: $x_k = (b+a)/2 + (b-a)/2\cos(\pi(2k-1)/2/(n+1)), k = \overline{1,n+1}$ з оцінкою похибки:

$$\left| f(x) - P_n(x) \right| \leq ((b-a)/2)^{n+1} \max_{x \in [a,b]} \left| f^{(n+1)}(x) \right| / 2^n /(n+1)!$$

3 Чисельні методи лінійної алгебри

До чисельних методів лінійної алгебри відносять чисельні методи розв'язання систем лінійних алгебраїчних рівнянь (СЛАР), обернення матриць, розв'язання задачі на власні значення для лінійних відображень, розв'язання алгебраїчних рівнянь.

3.1 Розв'язання СЛАР

Класифікацію СЛАР можна провести згідно з розміром її матриці, на малі, середні та великі. Така класифікація пов'язана з вибором методів розв'язання та необхідних для цього ресурсів. Оскільки матеріальна база широкодоступної обрахункової техніки стрімко зростає, така класифікація актуальна лише на короткому проміжку часу. Але у відносному виміру вона є достатньо важливою. Відповідно можна про класифікувати методи розв'язання СЛАР як точні методи, ітераційні методи та ймовірності методи. Останні два класи є методами отримання наближеного розв'язку. Відповідно, точні методи застосовуються для розв'язання СЛАР з малими матрицями, ітераційні – з середніми та ймовірності –

з великими. Окрім того, методи розв'язання СЛАР можна прокласифікувати за принципом універсальні та спеціалізовані методи.

Озн. Точним називається метод розв'язання, який дає точний результат при виконанні скінченної кількості операцій за умови відсутності похибки округлення.

Озн. Ітераційним називається метод, який передбачає розв'язання у вигляді нескінченного впорядкованого ланцюга задач, умова кожної з яких залежить від результатів попередніх розв'язків.

Ітераційний метод дає лише наближений результат при виконанні скінченної кількості операцій та завжди передбачає наявність критерію припинення ітерацій. Як правило, таким критерієм є умова досягнення заданої точності розв'язання.

Озн. Універсальним називається метод розв'язання, який може бути застосований до розв'язання усіх задач певного класу.

Озн. Спеціалізованим називається метод розв'язання, який може бути застосований до розв'язання усіх задач певного підкласу заданого класу.

3.2 Метод послідовного виключення змінних

Метод послідовного виключення змінних розв'язання СЛАР базується на приведенні матриці задачі до стандартного діагонального виду шляхом виконання еквівалентних операцій на розширеній матриці.

Це типовий точний універсальний метод знаходження розв'язку СЛАР. Він може також бути застосований до задачі обернення матриці та визначення рангу (визначника) матриці.Розглянемо СЛАР $\mathbf{A}\vec{x} = \vec{b}$ з матрицею $\mathbf{A} = \{a_{ij}\}_{i\,j=1}^{n}$ та вектором правої частини $\vec{b}^{T} = \{b_i\}_{i=1}^{n}$:

$$\sum_{j=1}^{n} a_{ij} x_j = b_i,\ i = \overline{1,n} \qquad (3.1)$$

Вважаючи, що $a_{11} \neq 0$, розділимо перше рівняння на a_{11} та віднімемо від інших рівнянь $i = \overline{2,n}$ перетворене перше рівняння помножене на коефіцієнт першого члену a_{i1}. Отримаємо еквівалентну систему рівнянь:

$$\begin{cases} x_1 + \sum_{j=2}^{n} a_{1j}^1 x_j = b_1^1 \\ \sum_{j=2}^{n} a_{ij}^1 x_j = b_i^1, i = \overline{2,n} \end{cases}, \tag{3.2}$$

де $a_{1j}^1 = a_{1j} / a_{11}, j = \overline{2,n}, b_1^1 = b_1 / a_{11}, \quad a_{ij}^1 = a_{ij} - a_{1j}^1 a_{i1}, j = \overline{2,n}, \; b_i^1 = b_i - b_1^1 a_{i1}, i = \overline{2,n}.$

При цьому, очевидно, перший стовпчик $\mathbf{A}^1 = \left\{a_{ij}^1\right\}_{i\,j=1}^{n}$ містить лише перший ненульовий елемент рівний 1. Розглянемо редукцію матриці $\mathbf{A}^1$ з $i, j = \overline{2,n}$ та застосуємо до неї вказану процедуру (вважаючи, що $a_{22}^1 \neq 0$):

$$\begin{cases} x_2 + \sum_{j=3}^{n} a_{2j}^2 x_j = b_2^2 \\ \sum_{j=3}^{n} a_{ij}^2 x_j = b_i^2, i = \overline{3,n} \end{cases}, \tag{3.3}$$

де $a_{2j}^2 = a_{2j}^1 / a_{22}^1, j = \overline{3,n}, b_2^2 = b_2^1 / a_{22}^1, \; a_{ij}^2 = a_{ij}^1 - a_{2j}^2 a_{i2}^1, j = \overline{3,n}, \; b_i^2 = b_i^1 - b_2^2 a_{i2}^1, i = \overline{3,n}.$

При цьому, очевидно, перший стовпчик $\mathbf{A}^2 = \left\{a_{ij}^2\right\}_{i\,j=2}^{n}$ містить лише перший ненульовий елемент рівний 1.

Повторюємо вказану процедуру $n-1, \; k = \overline{1,n-1}$ раз з загальним правилом:

$$\begin{cases} x_k + \sum_{j=k+1}^{n} a_{kj}^k x_j = b_k^k \\ \sum_{j=k+1}^{n} a_{ij}^k x_j = b_i^k, i = \overline{k+1,n} \end{cases}, \tag{3.4}$$

де $a_{kj}^k = a_{kj}^{k-1} / a_{kk}^{k-1}, j = \overline{k+1,n}, b_k^k = b_k^{k-1} / a_{kk}^{k-1},$

$a_{ij}^k = a_{ij}^{k-1} - a_{kj}^k a_{ik}^{k-1}, j = \overline{k+1,n}, \; b_i^k = b_i^{k-1} - b_k^k a_{ik}^{k-1}, i = \overline{k+1,n},$ провівши $n-1$ редукцію, отримаємо лінійне лінійне рівняння: $a_{nn}^{n-1} x_n = b_n^{n-1}$, або $x_n = b_n^n$ де $b_n^n = b_n^{n-1} / a_{nn}^{n-1}$. А отже, еквівалентними операціями, СЛАР (3.1) зведена до СЛАР з верхньотрикутною матрицею:

$$\sum_{j=i}^{n} a_{ij}^i x_j = b_i^i, \; i = \overline{1,n}, \; a_{ii}^i = 1, i = \overline{1,n} \tag{3.5}$$

Озн. Вказана процедура (що визначена в (3.4)) називається **прямим ходом** методу послідовного виключення змінних (метода Гауса). **Зворотнім ходом** методу послідовного виключення змінних (метода Гауса) називається розв'язання СЛАР з трикутною матрицею:

$$\begin{aligned} & x_n = b_n^n \\ & x_{n-1} = b_{n-1}^{n-1} - a_{n-1\,n}^{n-1}\, x_n \\ & \ldots \\ & x_k = b_k^k - \sum_{j=k+1}^{n} a_{k\,j}^k\, x_j \\ & \ldots \\ & x_1 = b_1^1 - \sum_{j=2}^{n} a_{1\,j}^1\, x_j \end{aligned} \qquad (3.6)$$

Зауваження. Повне виконання прямого проходу методу послідовного виключення змінних можливий тоді та лише тоді, коли $\det \mathbf{A}^k \neq 0 \;\; k = \overline{1, n-1}$, а отже, коли $\det \mathbf{A} \neq 0$ і (3.1) має розв'язок (і він єдиний) для довільного $\vec{b}$. Якщо це не так і $\operatorname{rank} \mathbf{A} = m < n$, виконання (3.4) $m-1$ разів призведе до матриці $m \times n$, верхньо трикутного виду. Якщо при цьому останні $n-m$ коефіцієнтів правої частини не всі рівні 0, система не сумісна. Якщо, рівні 0, існує $n-m$ параметричний розв'язок, який можна отримати задавши $x_k = C_k$, $k = \overline{m-1, n}$ та застосувавши зворотній хід методу послідовного виключення змінних.

Зауваження. Необхідною умовою (3.3) є $a_{kk}^{k-1} \neq 0$, оскільки, (3.4) починається з операції ділення k-го рядка розширеної матриці на a_{kk}^{k-1}. Навіть якщо $a_{kk}^{k-1} \neq 0$, але $\left|a_{kk}^{k-1}\right| << \max\limits_{j=k+1,n} \left|a_{kj}^{k-1}\right|$, операція ділення може призвести до значної об рахункової похибки. Уникнути цього можна шляхом перестановки стовпчиків редукованої матриці $\mathbf{A}^{k-1}$ так, щоб індекс (k,k) завжди мав максимальний за модулем в коефіцієнт в рядку. Такий метод називається методом з вибором найбільшого коефіцієнта в рядку. Якщо задля цієї мети використовується перестановка рядків, метод називається з вибором максимального елементу в стовпчику.

Зауваження. Метод послідовного виключення змінних для системи з матрицею $n \times n$ при реалізації потребує $\approx n^3 / 3$ операцій. При цьому, зворотній хід потребує $\approx n^2$ операцій.

3.2 Метод квадратного кореня

Метод квадратного кореня – точний спеціалізований метод, який може бути використаний для СЛАР з ермітово самоспряженими матрицями.

Метод базується на розкладі Холєцького: $\mathbf{A} = S^* D S$, де S - верхньотрикутна матриця (S^* - як ермітово спряжен до S - нижньотрикутна), $D = \{\pm d_{ii}\,\delta_{ij}\}_{i,j=1}^{n}, |d_{ii}| = 1$ - одно діагональна матриця.

Коефіцієнти матриць $S = \{s_{ij}\}_{i,j=1}^{n}$ та D знаходяться шляхом підрахунку рекурентних формул, які можуть бути отримані з прямої реалізації розкладу:

$$\begin{aligned} d_{ii} &= sign\left(a_{ii} - \textstyle\sum_{k=1}^{i-1} |\, s_{ik}\,|^2\, d_{kk}\right), \\ s_{ii} &= \sqrt{\left|a_{ii} - \textstyle\sum_{k=1}^{i-1} |\, s_{ik}\,|^2\, d_{kk}\right|}, \\ s_{ij} &= \left(a_{ij} - \textstyle\sum_{k=1}^{i-1} \bar{s}_{ki} s_{kj} d_{kk}\right) / (s_{ii}\, d_{ii}) \end{aligned} \qquad (3.7)$$

Нехай маємо СЛАР $\mathbf{A}\vec{x} = \vec{b}$, так, що $\mathbf{A} = \mathbf{A}^*$ ($a_{ij} = \bar{a}_{ji}, i, j = \overline{1,n}$). Виконавши розклад Холєцького: $\mathbf{A} = S^* D S$, отримаємо $S^* D S \vec{x} = \vec{b}$. Якщо ввести проміжний невідомий вектор $\vec{y}^T = \{y_i\}_{i=1}^{n}$: $DS\vec{x} = \vec{y}$, отримаємо системи рівнянь, які розв'язуються послідовно:

$$\begin{cases} S^* \vec{y} = \vec{b} \\ DS\vec{x} = \vec{y} \end{cases} \qquad (3.8)$$

Кожна з таких систем має трикутньовизначену матрицю, а отже може бути розв'язана за методом зворотного ходу метода Гауса (3.6), реалізація якого потребує лише $\approx n^2$ операцій .

3.3 Ітераційні методи розв'язання СЛАР

Озн. Узгодженою з векторною нормою $\|\vec{x}\|$ **матричною нормою** називають: $\|\mathbf{A}\| = \sup_{\vec{x}\neq 0} \|\mathbf{A}\vec{x}\| / \|\vec{x}\|$

Найбільш широко вживані векторні та узгоджені з ними матричні норми:

$$\|\vec{x}\|_1 = \max_{1\le i\le n} |x_i|,\ \|\mathbf{A}\|_1 = \max_{1\le i\le n}\left(\sum_{j=1}^{n} |a_{ij}|\right) \tag{3.9}$$

$$\|\vec{x}\|_2 = \sum_{i=1}^{n} |x_i|,\ \|\mathbf{A}\|_2 = \max_{1\le j\le n}\left(\sum_{i=1}^{n} |a_{ij}|\right) \tag{3.10}$$

$$\|\vec{x}\|_3 = \sqrt{\sum_{i=1}^{n} |x_i|^2}\ ,\ \|\mathbf{A}\|_3 = \sqrt{\max_{1\le i\le n} \lambda^i_{A^*A}} \tag{3.11}$$

Особливістю всіх ітераційних методів розв'язання СЛАР є вибірковість, тобто необхідність виконання додаткових умов, накладених на компоненти системи, що забезпечують збіжність ітераційного процесу до точного розв'язку. Як правило такі умови формулюються у вигляді теорем збіжності.

3.4 Метод простої ітерації розв'язання СЛАР

Метод простої ітерації – базовий ітераційний універсальний метод розв'язання СЛАР.

Нехай СЛАР:

$$\mathbf{A}\,\vec{x} = \vec{b}. \tag{3.12}$$

Перетворимо її до виду:

$$\vec{x} = B\vec{x} + \vec{c}. \tag{3.13}$$

Процедура перетворення базується на введенні довільної не виродженої матриці $D : \det D \neq 0$. Очевидно, (3.12) еквівалентна $\vec{x} = \vec{x} - D\left(A\vec{x} - \vec{b}\right)$. Позначивши $B = -D\mathbf{A}, \vec{c} = D\vec{b}$, отримаємо (3.13).

Озн. Методом простої ітерації розв'язання СЛАР (3.13) називається ітераційний процес отримання наближеного розв'язку за процедурою:

$$
\begin{aligned}
&\vec{x}^{0} = \vec{x}_0,\\
&\vec{x}^{1} = B\vec{x}^{0} + \vec{c},\\
&\ldots \qquad\qquad .\\
&\vec{x}^{k+1} = B\vec{x}^{k} + \vec{c}\\
&\ldots
\end{aligned}
\tag{3.14}
$$

При цьому, збіжність $\vec{x}^{k} \underset{k\to\infty}{\to} \vec{X}$ (де $\vec{X}$ - точний розв'язок (3.11): $\mathbf{A}\vec{X} = \vec{b}$) не є автоматичною та потребує дослідження та обґрунтування. $\vec{x}_0$ називається початковою точкою (або початковим наближенням), $\vec{x}^{k}$ - наближення розв'язку на k-й ітерації.

Теорема про достатню умову збіжності методу простої ітерації.

Якщо $\|B\| < 1$, то (3.13) має єдиний розв'язок, який може бути знайдений за методом простої ітерації (3.14). При цьому, наближений розв'язок прямує до точного з швидкістю геометричної прогресії:

$$
s_k = \sup_{\vec{x}^0 \neq \vec{X}} \left\|\vec{x}^{k} - \vec{X}\right\| / \left\|\vec{x}^{0} - \vec{X}\right\| = \left\|B^{k}\right\| \tag{3.15}
$$

Ітераційний процес для знаходження точного розв'язку, при умові його збігання, формально має продовжуватись до нескінченності. На практиці, кількість ітерацій залежить від необхідної точності наближення розв'язку та є скінченною. При виконанні умови теореми можна показати, що $\forall \varepsilon > 0, \|B\|^{k} \le \varepsilon, \Rightarrow s_k \le \varepsilon$. Отже, при заданому ε, достатня кількість ітерацій: $n \ge \ln(1/\varepsilon)/\ln(1/\|B\|)$. Однак, при реалізації частіше використовують практичний критерій зупинки ітерування: $\left\|\mathbf{A}\vec{x}^{n} - \vec{b}\right\| < \varepsilon_1$, або $\left\|\vec{x}^{n} - \vec{x}^{n-1}\right\| < \varepsilon_2$.

Теорема вказує лише достатні умови збіжності методу простої ітерації, але вона достатньо зручна для використання, оскільки визначення узгодженої норми матриці достатньо просте. Слід також приймати до уваги наступну нерівність $\|\vec{x}\|_1 \le \|\vec{x}\|_3 \le \|\vec{x}\|_2 \le A\|\vec{x}\|_1$.

Найбільш строгий результат щодо збіжності методу простої ітерації визначений в наступній теоремі.

Теорема про необхідну та достатню умову збіжності методу простої ітерації.

Нехай (3.11) має єдиний розв'язок. (3.13) збігається до точного розв'язку при довільній початковій умові тоді і лише тоді, коли всі власні значення матриці B за модулем менші за 1.

Ця теорема має скоріше теоретичну цінність, оскільки процедура визначення максимального за модулем власного числа матриці (часткова проблема власних значень матриці) як правило, більш затратна, ніж розв'язання СЛАР.

3.5 Метод Зейделя

Метод Зейделя – універсальний ітераційний метод розв'язання СЛАР, що є частинним випадком методу простої ітерації. Це один з найпопулярніших ітераційних методів для чисельного розв'язання СЛАР з матрицями середнього розміру.

Для СЛАР (3.12) метод Зейделя базується на ідеї визначення наближення на кожній ітерації так, що i-а невідома знаходиться з розв'язання i-го рівняння.

Задавши початкові значення для змінних $x_i^0, i=\overline{2,n}$, на 1-й ітерації визначаємо x_1^1 з 1-го рівняння: $a_{11}x_1^1+\sum_{j=2}^{n}a_{1j}x_j^0=b_1$. x_2^1 - з 2-го рівняння: $a_{21}x_1^1+a_{22}x_2^1+\sum_{j=2}^{n}a_{2j}x_j^0=b_2,\ldots,$ x_i^1 - з i-го рівняння: $\sum_{j=1}^{i-1}a_{kj}x_j^1+a_{ii}x_i^1+\sum_{j=i+1}^{n}a_{ij}x_j^0=b_i,\ k=\overline{3,n}$. Отже, алгоритм для k-ї ітерації має вигляд:

$$\begin{cases} a_{11}x_1^k+a_{12}x_2^{k-1}+a_{13}x_3^{k-1}+\ldots+a_{1n}x_n^{k-1}=b_1 \\ a_{21}x_1^k+a_{22}x_2^k+a_{23}x_3^k+\ldots+a_{2n}x_n^{k-1}=b_2 \\ \ldots \\ a_{i1}x_1^k+a_{22}x_2^k+\ldots+a_{ii}x_i^k+\ldots+a_{2n}x_n^{k-1}=b_i \\ a_{n1}x_1^k+a_{n2}x_2^k+a_{n3}x_3^k+\ldots+a_{nn}x_n^k=b_n \end{cases} \qquad (3.16)$$

(3.16) в матричному вигляді записується як:

$$B\vec{x}^k + C\vec{x}^{k-1} = \vec{b} \tag{3.17}$$

де

$$B = \{b_{ij}\}_{i,j=1}^{n},\ b_{ij} = \begin{cases} 0, & j > i \\ a_{ij}, & j \le i \end{cases},\ C = \{c_{ij}\}_{i,j=1}^{n},\ b_{ij} = \begin{cases} a_{ij}, & j > i \\ 0, & j \le i \end{cases} \tag{3.18}$$

Якщо вважати (необхідна умова застосування алгоритму методу Зейделя), що $a_{ii} \neq 0, i = \overline{1,n}$, очевидно, існує B^{-1} та (3.17) еквівалентна:

$$\vec{x}^k = -B^{-1}C\vec{x}^{k-1} + B^{-1}\vec{b} \tag{3.19}$$

Як бачимо, (3.17) співпадає з (3.14) – ітераційним правилом методу простої ітерації. Це дає змогу сформулювати необхідну та достатню умову збіжності методу Зейделя. Характеристичне рівняння матриці методу має вигляд: $\det\left(-B^{-1}C - \lambda I\right) = 0$, або:

$$\det\left(C + \lambda B\right) = 0 \tag{3.20}$$

Отже, необхідною та достатньою умовою збіжності (3.18) є умова того, що всі корені алгебраїчного рівняння (3.19) за модулем менші за 1.

Більш зручна для використання наступна теорема.

Теорема про достатні умови збіжності методу Зейделя.

Якщо виконується умова:

$$\sum_{j=1,\, j\neq i}^{n} |\, a_{ij}\,| \le q\,|\, a_{ii}\,|, i = \overline{1,n},\ 0 < q < 1 \tag{3.21}$$

то метод Зейделя збігається та $\left\|\vec{x}^k - \vec{X}\right\| \le q^k \left\|\vec{x}^0 - \vec{X}\right\|$, тобто, збіжність має геометричний характер.

Метод Зейделя можна як метод по координатного спуску мінімізації квадратичного функціоналу.

Нехай **A** - дійсно значна, симетрична та додатньо визначена матиця.. Розглянемо функціонал:

$$F(\vec{x}) = \left(\mathbf{A}(\vec{x} - \vec{X}), \vec{x} - \vec{X}\right) - \left(\mathbf{A}\vec{X}, \vec{X}\right) = \left(\mathbf{A}\vec{x}, \vec{x}\right) - 2\left(\mathbf{A}\vec{X}, \vec{x}\right) = \left(\mathbf{A}\vec{x}, \vec{x}\right) - 2\left(\vec{b}, \vec{x}\right) \tag{3.22}$$

Очевидно, $F(\vec{x})$ - квадратичний, додатньо визначений (а, отже – опуклий) функціонал, що приймає на точному розв'язку (3.12) мінімальне значення рівне 0. Отже, розв'язок (3.12) існує для довільної правої частини $\vec{b}$, а розв'язання (3.12) еквівалентне знаходженню аргументу мінімуму (3.22).

Реалізація методу по кординатного спуску мінімізації $F(\vec{x})$ виглядає наступним чином. Візьмемо деяке початкове наближення $x_i^0, i = \overline{2,n}$ та підставимо в аргумент (3.21). Отримаємо функцію однієї змінної x_1 (квадратна опукла вниз функція):

$$f_1(x_1) = F(x_1, x_2^0, ..., x_n^0) =$$
$$= a_{11}x_1^{\,2} + 2x_1\sum_{j=2}^{n} a_{1j}x_j^0 + \sum_{i,\,j=2}^{n} a_{ij}x_i^0 x_j^0 - 2x_1 b_1 - 2\sum_{i=2}^{n} x_i^0 b_i \,.$$

Мінімальне значення $f_1(x_1)$ в силу того, що $a_{11} > 0$, досягається в точці x_1^1: $f_1'(x_1^1)/2 = a_{11}x_1^1 + \sum_{j=2}^{n} a_{1j}x_j^0 - b_1 = 0$, що збігається з першим рівнянням 1-ї ітерації методу Зейделя. Побудувавши для знаходження i-ї змінної функцію

$$f_i(x_i) = F(x_1^1, x_2^1, ..., x_i, x_{i+1}^0 ... x_n^0) =$$
$$= a_{ii}x_i^{\,2} + 2x_i\sum_{j=1}^{i-1} a_{ij}x_j^1 + 2x_i\sum_{j=i+1}^{n} a_{ij}x_j^0 + \sum_{k,\,j=1,k,\,j\neq i}^{n} a_{kj}x_k^0 x_j^0 -$$
$$- 2x_i b_i - 2\sum_{j=1}^{i-1} x_j^1 b_j - 2\sum_{j=i+1}^{n} x_j^0 b_j,\ i = \overline{2,n}$$

Легко записати рівняння для знаходження мінімального значення:

$$f_i'(x_i^1)/2 = a_{ii}x_i^1 + \sum_{j=1}^{i-1} a_{ij}x_j^1 + \sum_{j=i+1}^{n} a_{ij}x_j^0 - b_i = 0,\ i = \overline{2,n},$$

що збігається з іншими рівняннями 1-ї ітерації методу Зейделя. Знайшовши наближений розв'язок на першій ітерації, вибудовуємо на ньому другу ітерацію, і так далі.

Зауваження. Слід пам'ятати, що умовою такої інтерпретації є додаткові властивості СЛАР. Отже в даному випадку, метод Зейделя має класифікуватись як спеціалізований метод.

3.5 Метод найшвидшого градієнтного спуску

Еквівалентність задач розв'язання СЛАР та мінімізації квадратичного функціоналу може бути покладена в основу низки методів лінійної алгебри. Одним з таких методів є метод найскорішого градієнтного спуску. Це спеціалізований ітераційний метод.

На відміну від методу Зейделя, при отриманні наближення на певній ітерації розшукується не в координатному напрямку $((x_1^k,...,x_k^k,...,x_n^{k-1}) = (x_1^k,...,x_k^{k-1} + \Delta_k,...,x_n^{k-1}))$, а в градієнтному напрямку:

$$\vec{x}^{k+1} = \vec{x}^k - \delta_k \text{grad}\, F(\vec{x}^k) \qquad (3.23)$$

Озн. Ітераційні методи мінімації функціонала, що використовують градієнтний напрямок для знаходження наближень (3.23) називаються **методами градієнтного спуску**.

Градієнтні методи різняться способом визначення скалярного параметру наближення δ_k. Знак мінус в (3.23) обрано, щоб підкреслити, що для опуклого вниз знаковизначеного функціоналу, градієнтний напрямок, взагалі кажучи, визначає напрямок зростання (а не спадання).

Озн. Методи градієнтного спуску в яких параметр наближення визначається з умови мінімального значення функціоналу називаються **методами найскорішого градієнтного спуску**.

Відповідно, метод Зейделя має в основі метод найскорішого координатного спуску. Методи градієнтного та координатного спуску є загальними стандартними методами, що використовуються в математичному програмуванні та опуклій безумовній оптимізації. Метод, що буде далі сформульовано є лише застосуванням такого загального підходу до розв'язання СЛАР.

Прийнявши до уваги умови застосування методу координатного спуску до розв'язання СЛАР ($\mathbf{A}$ - дійсно значна, симетрична та додатньо визначена матиця з квадратичним функціоналом (3.21)) та, оскільки, $\text{grad}\, F(\vec{x}^k) = 2\mathbf{A}\vec{x}^k - 2\vec{b}$, (3.23) набуває вигляду:

$$\vec{x}^{k+1} = \vec{x}^k - \Delta_k (\mathbf{A}\vec{x}^k - \vec{b}),\ \Delta_k = 2\delta_k > 0 \qquad (3.24)$$

Необхідна умова мінімуму $F(\vec{x}^k - \Delta_k \operatorname{grad} F(\vec{x}^k))$ як функції від Δ_k в силу симетрії $\mathbf{A}$ має вигляд:

$$\left(\mathbf{A}\vec{x}^k - \vec{b} - \Delta_k \mathbf{A}\left(\mathbf{A}\vec{x}^k - \vec{b}\right), \mathbf{A}\vec{x}^k - \vec{b}\right) = 0 \qquad (3.25)$$

Отже:

$$\Delta_k = \left(\mathbf{A}\vec{x}^k - \vec{b}, \mathbf{A}\vec{x}^k - \vec{b}\right) / \left(\mathbf{A}\left(\mathbf{A}\vec{x}^k - \vec{b}\right), \mathbf{A}\vec{x}^k - \vec{b}\right) \qquad (3.26)$$

На відміну від методу координатного спуску з геометричною (степеневою) збіжністю наближень до точного розв'язку, метод найскорішого градієнтного спуску має експоненційну (показникову) збіжність, що визначається теоремою збіжності.

Теорема. Наближення ітераційного процесу (3.23), (3.25) задовольняють співвідношення:

$$\left(\mathbf{A}(\vec{x}^k - \vec{X}), \vec{x}^k - \vec{X}\right) \le \left((M-m)/(M+m)\right)^{2k} \left(\mathbf{A}(\vec{x}^0 - \vec{X}), \vec{x}^0 - \vec{X}\right) \qquad (3.27)$$

де $\vec{x}^0$ - початкове наближення, а M та m - найбільше та найменше власні значення матриці $\mathbf{A}$ (в силу властивостей $\mathbf{A}$, її власні значення дійсні та додатні).

3.6 Обумовленість матриць. Регуляризація

Чисельні методи розв'язання СЛАР, приведені в попередніх розділах можуть бути неефективними, тобто давати велику похибку розв'язку. Не слід забувати, що конкретна реалізація розрахунків на ЕОМ завжди проходить в умовах скінченної розрядності операндів арифметичних операцій. Отже сам процес розв'язання неминуче супроводжується збуренням коефіцієнтів матриці та вектору правої частини. В деяких випадках цю проблему можна мінімізувати шляхом підвищення порядку розрядної сітки операндів. Але при великих розмірах матриць цей підхід не є ефективним. Прикладом є розв'язання СЛАР з не виродженими матрицями, визначники яких малі в порівнянні з нормою розв'язку. Такі задачі належать до класу **некоректних задач**, в яких за

означенням як завгодно малі збурення параметрів можуть призвести до скінченних змін розв'язку.

Мірами коректності СЛАР є **міра обумовленості матриці** та **міра обумовленості СЛАР**.

Нехай розв'язується СЛАР

$$\mathbf{A}\vec{x} = \vec{b}. \tag{3.28}$$

Надамо матриці збурення Δ та вектору правої частини $\vec{\eta}$. Якщо точним розв'язком (3.28) є $\vec{X}$, то розв'язком $(\mathbf{A}+\Delta)\vec{x} = (\vec{b}+\vec{\eta})$ буде $\vec{X}+\vec{r}$. Якщо вважати, що $\| \Delta \|, \| \vec{\eta} \| << 1$, а отже і $\| \vec{r} \|$ (з умови $\| \mathbf{A}+\Delta \| \neq 0$ слідує існування та єдність розв'язку, а отже і неперервна залежність розв'язку від правої частини), можемо отримати: $\vec{r} \approx \mathbf{A}^{-1}(\vec{\eta} - \Delta \vec{X})$ з оцінкою:

$$\| \vec{r} \| \leq \| \mathbf{A}^{-1} \| (\| \vec{\eta} \| + \| \Delta \| \| \vec{X} \|) \tag{3.29}$$

Якщо знехтувати збуренням матриці, можна ввести міру обумовленості СЛАР:

$$\tau = \| \vec{b} \| / \| \vec{X} \| \sup_{\vec{\eta}} (\| \vec{r} \| / \| \vec{\eta} \|) \tag{3.30}$$

що характеризує відносну похибку розв'язання СЛАР зі збуреною правою частиною в термінах відношення до правої частини та точного розв'язку незбуреної системи. Оскільки $\vec{r} \approx \mathbf{A}^{-1}\vec{\eta}$, в наближеному сенсі можна вважати, що

$$\tau = \| \vec{b} \| / \| \vec{X} \| \| \mathbf{A}^{-1} \| \tag{3.31}$$

Озн. Кількісна характеристика (3.31) називається **мірою обумовленості СЛАР** (3.28).

Чим менша ця величина, тим обумовленість є кращою та обрахункові похибки є менш значущими при розв'язанні СЛАР.

Озн. У випадку, коли (3.28) розв'язується при різних (чи початково невизначених) векторах правої частини, вводиться міра **міра обумовленості матриці** як $\nu(\mathbf{A}) = \sup_{\vec{b}} \tau$. Оскільки за означенням, $\sup_{\vec{x} \neq 0} (\| \mathbf{A}\vec{x} \| / \| \vec{x} \|) = \| \mathbf{A} \|$, то

$$\nu(\mathbf{A}) = \| \mathbf{A} \| \| \mathbf{A}^{-1} \| \tag{3.32}$$

Якщо відомий спектр матриці $\mathbf{A}$, в силу відомих оцінок: $\| \mathbf{A} \| \geq \max | \lambda_A |$, $\| \mathbf{A}^{-1} \| \geq 1 / \min | \lambda_A |$, маємо:

$$\nu(\mathbf{A}) = \max | \lambda_A | / \min | \lambda_A | \qquad (3.33)$$

Методи знаходження стійких розв'язків для СЛАР з погано обумовленими матрицями основуються на параметричній модифікації матриці зі знаходженням оптимального значення параметра. Слід зазначити, що регуляризація ніколи не призводить до знаходження точного розв'язку СЛАР а лише до знаходження "найкращого" в деякому сенсі, що обумовлюється конкретною процедурою регуляризації розв'язку. Як правило, регуляризуючий параметр вводиться в головну діагональ матриці.

Нехай $\mathbf{A}$ - дійсно значна та симетрична, зі спектром: $S_A = \{\lambda_i\}_{i=1}^n$ та власними нормованими векторами $\vec{S}_A = \{\vec{e}_i\}_{i=1}^n$. Тоді, розв'язок (3.28) можна представити у вигляді: $\vec{X} = \sum_{j=1}^n (\vec{b}, \vec{e}_j) / \lambda_j \vec{e}_j$. Якщо вектор правої частини набуває збурення $\vec{\eta}$, збуреним розв'язком буде $\tilde{\vec{X}} = \sum_{j=1}^n (\vec{b}, \vec{e}_j) / \lambda_j \vec{e}_j + \sum_{j=1}^n (\vec{\eta}, \vec{e}_j) / \lambda_j \vec{e}_j$. Для малих значень λ_i, додатковий член може мати скінченні значення навіть для малих збурень $\vec{\eta}$.

Збуримо параметричним чином матрицю $\mathbf{A}$ з параметром α, перейшовши до системи $(\mathbf{A} + \alpha I)\vec{x}^{\alpha} = (\vec{b} + \vec{\eta})$ з розв'язком $\tilde{\vec{X}}^{\alpha} = \sum_{j=1}^n (\vec{b}, \vec{e}_j) / (\lambda_j + \alpha) \vec{e}_j + \sum_{j=1}^n (\vec{\eta}, \vec{e}_j) / (\lambda_j + \alpha) \vec{e}_j$. Очевидно, для доданків, що відповідають $\lambda_i >> \alpha$, така модифікаціє буде мати несуттєве значення в порівнянні з доданками для $\lambda_i << \alpha$, вплив на розв'язок яких суттєво послабиться. Отже буде існувати деяке значення α, яке несуттєво змінить розв'язок, але призведе до значного підвищення стійкості розв'язання до арифметичних похибок.

Вказану процедуру регуляризації можно застосовувати як до розв'язання СЛАР з пагано обумовленими матрицями точними методами, так і ітераційними.

3.7 Чисельні методи розв'язання проблеми власних значень

Озн. Повною проблемою власних значень називають задачу визначення всього набору (повного спектру матриці) власних значень та відповідних до них власних векторів:

$$\mathbf{A}\vec{x} = \lambda\vec{x}. \tag{3.34}$$

Для розв'язання повної проблеми власних значень необхідно побудувати характеристичний многочлен

$$D(\lambda) = \det(\mathbf{A} - \lambda I), \tag{3.35}$$

коренями якого є власні значення $S_A = \{\lambda_i\}_{i=1}^n$, та знайти (з певною умовою нормування) нетривіальні розв'язки СЛАР:

$$\mathbf{A}\vec{e}_i = \lambda_i\vec{e}_i,\ i = \overrightarrow{1,n} \tag{3.36}$$

що є власними векторами $\mathbf{A}$.

В тому випадку, коли достатньо знайти лише деякі обумовлено визначені частини спектру, кажуть про **часткову проблему власних значень**. Методи розв'язання проблеми власних значень поділяються на точні методи (коли точне розв'язання можливе при виконання скінченної кількості операцій) та ітераційні, в яких наближення будуються через деяку ітераційну процедуру. Вони також поділяються на універсальні та спеціалізовані.

3.8 Метод Крилова

Це точний універсальний метод розв'язання повної проблеми власних значень, що базується на теоремі Гамільтона-Келі:

$$\mathbf{A}^n + p_1\mathbf{A}^{n-1} + ... + p_{n-1}\mathbf{A} + p_n I = 0 \tag{3.37}$$

де $\{p_i\}_{i=1}^n$ - коефіцієнти характеристичного многочлену для $\mathbf{A}$:

$$D(\lambda) = (-1)^n(\lambda^n + p_1\lambda^{n-1} + ... + p_{n-1}\lambda + p_n), \tag{3.38}$$

Оберемо довільний ненульовий вектор $\vec{c}_0{}^T = \{c_{0i}\}_{i=1}^n$ та побудуємо сукупність n векторів за правилом: $\vec{c}_j = \mathbf{A}\vec{c}_{j-1},\ j = \overline{1,n}$ (очевидно, $\vec{c}_j = \mathbf{A}^j\vec{c}_0$) з координатами

$$\vec{c}_j{}^T = \{c_{ji}\}_{i=1}^n = \left\{\sum_{k=1}^n a_{ik}c_{j-1k}\right\}_{i=1}^n \tag{3.39}$$

При цьому, як показав Крилов, якщо ввести позначення:

$$\tilde{D}(\lambda) = \det\begin{pmatrix} 1 & \lambda & ... & \lambda^n \\ c_{01} & c_{11} & ... & c_{n1} \\ ... & ... & ... & ... \\ c_{0n} & c_{1n} & ... & c_{nn} \end{pmatrix} \tag{3.40}$$

то

$$\tilde{D}(\lambda) / D(\lambda) = C = const \tag{3.41}$$

Нехай $C \neq 0$. Домноживши (3.37) справа на $\vec{c}_0$, отримаємо:

$$\vec{c}_n + p_1\vec{c}_{n-1} + ... + p_{n-1}\vec{c}_1 + p_n\vec{c}_0 = 0 \tag{3.42}$$

а отже, $\{\vec{c}_i\}_{i=0}^n$ є системою лінійно залежних векторів (при цьому, очевидно, $\{\vec{c}_i\}_{i=0}^{n-1}$ є лінійно незалежна), а коефіцієнтами лінійної залежності є коефіцієнти характеристичного рівняння $\vec{p}^T = \{p_i\}_{i=1}^n$, які можна відшукати як розв'язок СЛАР:

$$(\vec{c}_{n-1},...,\vec{c}_1,\vec{c}_0)\vec{p} = -\vec{c}_n \tag{3.43}$$

У випадку, коли $C = 0$ (це можливо, лише коли (3.38) має кратні корені), система векторів $\{\vec{c}_i\}_{i=0}^{n-1}$ не є лінійно незалежною, що зокрема означає, що існує многочлен $\phi(\lambda)$ степені меншої за n такий, що $\phi(\mathbf{A}) = 0$. Такий многочлен називається **мінімальним многочленом**. Коренями мінімального многочлена є всі характеристичні числа матриці $\mathbf{A}$ з кратністю, не вищою за кратність в $D(\lambda) = 0$. В цьому випадку, потрібно знайти таке $m < n$, що $\{\vec{c}_i\}_{i=0}^{m-1}$ - є лінійно незалежною системою, в той час, як $\{\vec{c}_i\}_{i=0}^{m}$ є лінійно залежною. Для цього можна використати, наприклад, матод Гауса. Тоді розв'язком СЛАР:

$$(\vec{c}_{m-1},...,\vec{c}_1,\vec{c}_0)\vec{q} = -\vec{c}_m \tag{3.44}$$

Буде вектор $\vec{q}^T = \{q_i\}_{i=1}^m$ коефіцієнтів мінімального многочлену:

$$\phi(\lambda) = \lambda^m + q_1\lambda^{m-1} + ... + q_{m-1}\lambda + q_m, \tag{3.45}$$

Отже, в будь якому випадку, метод Крилова дозволяє побудувати алгебраїчне рівняння, коренями якого є всі власні значення матриці.

Розв'язавши характеристичне (або мінімальне) рівняння, та визначивши всі власні значення, можна побудувати відповідні їм власні вектори.

Нехай $m \le n$ - степінь мінімального (чи характеристичного многочлену). Тоді система $\{\vec{c}_i\}_{i=0}^{m-1}$- є лінійно незалежною і власний вектор $\vec{e}_i$, що відповідає $\lambda_i,\ i = \overline{1,m}$: $\mathbf{A}\vec{e}_i = \lambda_i\vec{e}_i,\ i = \overline{1,m}$ (3.46)

можна шукати у вигляді розкладу:

$$\vec{e}_i = \sum_{i=0}^{m-1}\gamma_i\vec{c}_i \tag{3.47}$$

Підставивши (3.46) в (3.45) отримаємо: $\sum_{i=0}^{m-1}\gamma_i\mathbf{A}\vec{c}_i = \lambda_i\sum_{i=0}^{m-1}\gamma_i\vec{c}_i$, або $\gamma_{m-1}\vec{c}_m + \sum_{i=0}^{m-2}\gamma_i\vec{c}_{i+1} = \lambda_i\sum_{i=0}^{m-1}\gamma_i\vec{c}_i$. В той же час, за теоремою Гамільтона-Келі: $\vec{c}_m = -\sum_{i=0}^{m-1}q_{m-i}\vec{c}_i$. Отже

$$-\gamma_{m-1}\sum_{i=0}^{m-1}q_{m-i}\vec{c}_i + \sum_{i=0}^{m-2}\gamma_i\vec{c}_{i+1} - \lambda_i\sum_{i=0}^{m-1}\gamma_i\vec{c}_i = 0 \tag{3.48}$$

Система $\{\vec{c}_i\}_{i=0}^{m-1}$- є лінійно незалежною. Прирівнявши в (3.47) до 0 коефіцієнти при $\vec{c}_i, i = \overline{0,m-1}$, отримаємо вирази для послідовного визначення $\gamma_i,\ i = \overline{0,m-1}$:

$$\begin{cases} \gamma_{m-2} = \lambda_i\gamma_{m-1} + q_0\gamma_{m-1} \\ \gamma_{m-3} = \lambda_i\gamma_{m-2} + q_1\gamma_{m-1} \\ \qquad\qquad ... \\ \gamma_0 = \lambda_i\gamma_1 + q_m\gamma_{m-1} \end{cases} \tag{3.49}$$

При цьому γ_{m-1} є довільною величиною та може вибиратись з умови $\| \vec{e}_i \| = 1$.

3.9 Метод Данилевського

Метод Данілевського універсальний точний метод розв'язання проблеми власних значень, що дозволяє побудувати характеристичний многочлен шляхом зведення матриці еквівалентними перетвореннями до нормальної форми Фробеніуса:

$$\tilde{\mathbf{A}} = \begin{pmatrix} p_1 & p_2 & \dots & p_{n-1} & p_{n1} \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix} \tag{3.50}$$

Легко бачити, що характеристичні многочлени $\mathbf{A}$ та $\tilde{\mathbf{A}}$ збігаються:

$$D(\lambda) = (-1)^n(\lambda^n + p_1\lambda^{n-1} + \dots + p_{n-1}\lambda + p_n), \tag{3.51}$$

Для зведення $\mathbf{A}$ до нормальної форми можна скористатись еквівалентними (щодо спектру) операціями перетворення: $\mathbf{A}_1 = B_1\mathbf{A}B_1^{-1}$, $\mathbf{A}_2 = B_2\mathbf{A}_1B_2^{-1}$, …, $\tilde{\mathbf{A}} = \mathbf{A}_{n-1} = B_{n-1}\mathbf{A}_{n-2}B_{n-1}^{-1}$, де кожне перетворення нормалізує 1 рядок матриці (n-й, $n-1$-й, …, 2-й).

Матриці B_i та B_i^{-1}, $i = \overline{1, n-1}$ будуються по правилу:

$$B_1 = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -a_{n1}/a_{nn-1} & -a_{n2}/a_{nn-1} & \dots & 1/a_{nn-1} & -a_{nn}/a_{nn-1} \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}, \tag{3.52}$$

$$B_1^{-1} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn-1} & a_{nn} \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix} \tag{3.53}$$

В цьому випадку, останній рядок матриці $\mathbf{A}_1 = B_1 \mathbf{A} B_1^{-1}$ має вигляд: $(0 \quad 0 \quad ... \quad 1 \quad 0)$.

Провівши подібні перетворення $n-1$ раз, отримаємо (3.50), в якій перший рядок складено з коефіцієнтів характеристичного многочлену (3.51). Знайшовши корені характеристичного многочлену (власні числа), власні вектори, що відповідають власним числам можна відшукати як однорідні розв'язки (3.46).

3.10 Метод Левер'є-Фаддєєва

Метод Левер'є-Фаддєєва – універсальний точний метод, що базується на побудові приєднаної до $\mathbf{A} - \lambda I$ матриці.

За формулами Ньютона, симетричні форми коренів характеристичного многочлену $S_k = \sum_{i=1}^{n} \lambda_i^{\,k}$ зв'язані з коефіцієнтами характеристичного многочлену співвідношеннями (ці формули справедливі для будь-яких многочленів з дійсними коефіцієнтами):

$$S_k + S_{k-1} p_1 + ... + S_1 p_{k-1} + k p_k = 0, \; k = \overline{1, n} \qquad (3.54)$$

Тобто, коефіцієнти многочлену можна відшукати, якщо відомі симетричні форми коренів. В той же час, відомо, що $S_k = \mathrm{Sp}(\mathbf{A}^k)$. Описана процедура належить Левер'є. За модифікацією Фаддєєва, будуємо матричний поліном $C(\lambda) = \sum_{i=0}^{n-1} C_i \lambda^{n-1-i}$, що є приєднаним до $\mathbf{A} - \lambda I$:

$$C(\lambda)(\mathbf{A} - \lambda I) = D(\lambda) I \qquad (3.55)$$

де $D(\lambda)$ - характеристичний многочлен (3.50).

Прирівнявши в (3.55) коефіцієнти при однакових степенях λ, отримаємо рекурентні вирази для отримання C_i, $i = \overline{0, n-1}$:

$$\begin{cases} -C_0 = (-1)^n I \\ C_1 - C_0\mathbf{A} = (-1)^{n-1} p_1 I \\ \dots \\ C_{n-1} - C_{n-2}\mathbf{A} = (-1)^{n-1} p_{n-1} I \\ C_{n-1}\mathbf{A} = (-1)^n p_n I \end{cases} \tag{3.56}$$

Перше рівняння визначає $C_0 = (-1)^{n-1} I$, друге - $C_1 = (-1)^{n-1}(\mathbf{A} + p_1 I)$ де згідно з (3.54) $p_1 = -S_1$. Помноживши останній вираз справа на $\mathbf{A}$ та порахувавши слід від правої та лівої частин, враховуючи (3.54) отримаємо значення p_2: $Sp(C_1\mathbf{A}) = (-1)^{n-1}\left(Sp(\mathbf{A}^2) + p_1 Sp(\mathbf{A})\right) = (-1)^{n-1}(S_2 + p_1 S_1) = (-1)^n 2p_2$, що дає змогу з (3.55) визначити $C_2 = C_1\mathbf{A} + (-1)^{n-1} p_2 I$. Множимо цей вираз на $\mathbf{A}$ та визначаємо слід, що дає змогу знайти $SP(C_2\mathbf{A}) = (-1)^n 3p_3$. Повторивши описану процедуру $n-1$ раз отримаємо: $Sp(C_{n-1}\mathbf{A}) = (-1)^n np_n$. При цьому, останнє співвідношення з (3.56) є перевірочним. В результаті отримаємо коефіцієнти характеристичного многочлену $\{p_k\}_{k=1}^n$ (3.51) та матриці-коефіцієнти прєднаного матричного многочлену $\{C_k\}_{k=0}^{n-1}$. Слід зауважити, що процедура Левер'є-Фаддєєва за кількістю операцій еквівалентна методу Левер'є. Але вона дозволяє отримати додаткову спектральну інформацію. Окрім того, що поклавши в (3.51) $\lambda = 0$ отримаємо $\det(\mathbf{A}) = (-1)^n p_n$ (що взагалі кажучи є універсальним виразом для всіх методів побудови характеристичного многочлену), поклавши в (3.55) $\lambda = 0$ отримаємо: $C_{n-1}\mathbf{A} = (-1)^n I$, а отже, з точністю до знака C_{n-1} є оберненою до матриці $\mathbf{A}$. Поклавши в (3.55) $\lambda = \lambda_i$ (λ_i - власне число $\mathbf{A}$, тобто корінь характеристичного рівняння), якщо $C(\lambda_i) \neq 0$, матимемо: $C(\lambda_i)(\mathbf{A} - \lambda_i I) = (\mathbf{A} - \lambda_i I)C(\lambda_i) = D(\lambda_i)I = 0$. А отже, вектори-стовпчики $C(\lambda_i)$ є власними векторами $\mathbf{A}$, що відповідають власному числу λ_i.

3.11 Ітераційний метод розв'язання часткової задачі власних значень

Часткова проблема власних значень полягає в визначенні умовної спектральної інформації. Описаний в цьому параграфі метод полягає в побудові ітераційного процесу для наближеного визначення максимального по модулю власного числа та відповідного власного вектора.

Для простоти вважаємо, що максимальне по модулю власне число (головне власне число) відділено від спектру: $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$. Нехай $\{\vec{e}_i\}_{i=1}^n$ - повний набір нормованих власних векторів ($\mathbf{A}\vec{e}_i = \lambda_i \vec{e}_i, \|\vec{e}_i\| = 1, i = \overline{1,n}$). Вибравши початкове значення $\vec{x}_0 = \sum_{i=1}^n c_i \vec{e}_i$, побудуємо ітераційну процедуру:

$$\vec{x}^{k+1} = \mathbf{A}\vec{x}^k \quad k = k = 0,1,... \tag{3.57}$$

Очевидно,

$$\vec{x}^k = \mathbf{A}\vec{x}^{k-1} = \sum_{i=1}^n c_i \mathbf{A}^k \vec{e}_i = \sum_{i=1}^n c_i \lambda^k \vec{e}_i \tag{3.58}$$

А отже, оскільки за припущенням $\vec{x}^k = c_1 {\lambda_1}^k \vec{e}_i + O(|\lambda_2|^k)$, $(\vec{x}^k, \vec{x}^k) = |c_1|^2 |\lambda_1|^{2k} + O(|\lambda_1|^k |\lambda_2|^k)$ та $(\vec{x}^k, \vec{x}^{k+1}) = \lambda_1 |c_1|^2 |\lambda_1|^{2k} + O(|\lambda_1|^k |\lambda_2|^k)$. Будемо визначати наближене значення λ_1 на k-й ітерації як:

$${\lambda_1}^{(k)} = (\vec{x}^k, \vec{x}^{k+1}) / (\vec{x}^k, \vec{x}^k) = \lambda_1 + O(|\lambda_2 / \lambda_1|^k) \tag{3.59}$$

Отже, згідно з умовою відділення головного власного вектора, ітераційний процес (3.57) дозволяє відшукати λ_1 з геометричною швидкістю збіжності.

Наближення головного власного вектора можна знайти, прийнявши до уваги $\|\vec{x}^k\| = \sqrt{(\vec{x}^k, \vec{x}^k)} = |c_1| \|\lambda_1|^k + O(|\lambda_2|^k)$:

$$\vec{e}_1^{(k)} = \vec{x}^k / \|\vec{x}^k\| = e^{i\varphi^{(k)}} \vec{e}_1 + O(|\lambda_2 / \lambda_1|^k) \tag{3.60}$$

де комплексний множник $e^{i\varphi^{(k)}}$ - за модулем дорівнює 1.

4 Чисельні методи розв'язання алгебраїчних та трансцендентних рівнянь

В механіці часто виникає необхідність чисельного розв'язання нелінійних рівнянь різного типу. Тобто, знаходження таких значень $\vec{x}$, що $\vec{F}(\vec{x}) = 0$ ($\vec{x} = \{x_i\}_{i=1}^n, \vec{F} = \{f_i(x_1,...,x_n)\}_{i=1}^n$). У випадку, коли $n = 1$ - рівняння є скалярним, в зворотному – векторним. Якщо скалярне рівняння визначається для многочленів $f(x) = \sum_{j=0}^n a_{n-j} x^j$ - рівняння називається алгебраїчним рівнянням степені n.

4.1 Чисельні методи розв'язання алгебраїчних рівнянь

Чисельні методи розв'язання алгебраїчних рівнянь в силу їх простоти зазвичай виділяють в для окремого розгляду. Хоча, універсальні методи розв'язання нелінійних рівнянь також можуть бути застосовані до розв'язання алгебраїчних рівнянь.

За основною теоремою алгебри, алгебраїчне рівняння степені n з дійсними коефіцієнтами має в точності n розв'язків (коренів), які можуть бути кратними, дійсними або комплексно спряженими.

Розв'язання алгебраїчних рівнянь (як і в цілому – нелінійних рівнянь) слід починати з **відділення коренів**, тобто визначення областей розташування коренів рівняння. Для цього можна скористатись відповідними теоремами.

Теорема 1. Всі корені рівняння:

$$f(z) = \sum_{j=0}^n a_{n-j} z^j = 0 \qquad (4.1)$$

розташовані в кільці комплексної площини:

$$|a_n| / (\bar{a} + |a_n|) \le |z| \le 1 + \underline{a} / |a_0| \qquad (4.2)$$

де $\bar{a} = \max\{|a_1|,...,|a_n|\}$, $\underline{a} = \max\{|a_0|,...,|a_{n-1}|\}$.

Теорема 2 (теорема Штурма – стосується лише дійсних коренів). Число дійсних коренів рівняння

$$f(x) = \sum_{j=0}^{n} a_{n-j} x^j = 0 \qquad (4.3)$$

на відрізку $[a,b]$ $(a < b)$ дорівнює різниці чисел переміни знаку послідовності Штурма в точках a та b.

Зауваження. Послідовність Штурма $\{f(x), f_1(x),..., f_m(x)\}$ будується за наступним правилом:

$f(x)$ - многочлен (4.3), $f_1(x) = f'(x)$, $f_2(x) = -f(x) \bmod f_1(x)$, $f_3(x) = -f_1(x) \bmod f_2(x)$, …, $f_m(x) = const$.

4.2 Метод Лобачевського розв'язання алгебраїчних рівнянь (метод квадрувань)

Метод Лобачевського базується на симетричних формах коренів алгебраїчного рівняння (4.3):

$$\begin{cases} x_1 + x_1 + ... + x_n = -a_1 / a_0 \\ x_1x_2 + x_1x_3 + ... + x_{n-1}x_n = a_2 / a_0 \\ \qquad\qquad ... \\ x_1x_2x_3...x_n = (-1)^n a_n / a_0 \end{cases} \qquad (4.4)$$

та надає алгоритм відділення різних коренів.

Дійсно, нехай $|x_1| >> |x_2| >> ... >> |x_n|$. Тобто корені різні та достатньо відділені. Тоді, очевидно, згідно з (4.4), з деякою точністю:

$$\begin{cases} x_1 \approx x_1(1 + x_2 / x_1 + ... + x_n / x_1) = -a_1 / a_0 \\ x_1x_2 \approx x_1x_2(1 + x_1x_3 /(x_1x_2) + ... + x_{n-1}x_n /(x_1x_2)) = a_2 / a_0 \\ \qquad\qquad ... \\ x_1x_2x_3...x_n = (-1)^n a_n / a_0 \end{cases}$$

що дає змогу знайти наближені значення коренів як:

$$x_i \approx -a_i / a_{i-1},\ i = \overline{1,n} \qquad (4.5)$$

В методі Лобачевського пропонується ітераційна процедура кадрування яка дозволяє будувати алгебраїчні рівняння, корені яких на кожній наступній ітерації підносяться до квадрату (квадруються).

Рівняння (4.3) з коренями $\{x_i\}_{i=1}^n$ можна записати у вигляді $a_0\prod_{i=1}^n(x-x_i)=0$. В той же час, рівняння з коренями $\{-x_i\}_{i=1}^n$ має вигляд $a_0\prod_{i=1}^n(x+x_i)=0$. Їхній добуток: $a_0{}^2\prod_{i=1}^n(x^2-x_i{}^2)=0$. Якщо ввести нову змінну $x^{(1)}=-x^2$, можна вказати процедуру знаходження коефіцієнтів цього рівняння

$$f^{(1)}(x^{(1)})=\sum_{j=0}^n a^{(1)}{}_{n-j}x^{(1)\,j}=0 \tag{4.6}$$

$$a^{(1)}{}_j=a_j{}^2-2a_{j-1}a_{j+1}+2a_{j-2}a_{j+2}+\ldots \tag{4.7}$$

де індекси в добутку коефіцієнтів збільшуються та зменшуються симетрично відносно j поки не буде досягнуто граничних значень (0 чи n). Знаки добутків передуються. Згідно з (4.5): $x^{(1)}{}_i\approx -a^{(1)}{}_i/a^{(1)}{}_{i-1}$, $i=\overline{1,n}$ буде мати вищу точність наближення ніж (4.5).

Провівши квадрування k разів, отримаємо рівняння

$$f^{(k)}(x^{(k)})=\sum_{j=0}^n a^{(k)}{}_{n-j}x^{(k)\,j}=0 \tag{4.8}$$

корені якого

$$-x_i{}^{2k}=x^{(k)}{}_i\approx -a^{(k)}{}_i/a^{(k)}{}_{i-1},\ i=\overline{1,n} \tag{4.9}$$

Квадрування можна припиняти при виконанні умови:

$$\max_i\left(\left|a^{(k)}{}_i/a^{(k)}{}_{i-1}-\left(a^{(k-1)}{}_i/a^{(k-1)}{}_{i-1}\right)^2\right|\right)<\varepsilon \tag{4.10}$$

Метод може бути легко поширений на випадок комплексно спряжених та кратних коренів.

4.3 Універсальні ітераційні методи розв'язання нелінійних рівнянь

Універсальних алгоритмів відділення коренів нелінійних рівнянь в загальному випадку не існує. Для конкретних випадків можна використати метод заміни рівняння іншим, що не є еквівалентним, але зберігає структуру вихідного рівняння.

Приклад. Відділити корені рівняння $tg\,x = x$.

Розв'язання. Оскільки $y(x) = tg\,x$ - монотонно зростаюча та має зліченну кількість гілок, визначених на $x \in (-\pi/2 + \pi k, \pi/2 + \pi k)$, $k \in Z$, так, що $tg\,\pi k = 0$, $tg\,x \underset{x \to \pm\pi/2+\pi k}{\to} \pm\infty$ а $y(x) = x$ - монотонно зростає на $x \in (-\infty, +\infty)$, очевидно, що рівняння має зліченну кількість коренів. Окрім того, оскільки функція рівняння є непарною, кожному кореню $x_k > 0$ буде відповідати симетричний корінь $-x_k$. Отже достатньо відшукати лише додатні корені рівняння (окрім тривіального кореня $x_0 = 0$). Зі вказаних властивостей функції рівняння, очевидно, $x_k = x_k^0 - \varepsilon_k, x_k^0 = \pi/2 + \pi k$, $k = 1,2,...$ де $\varepsilon_k \underset{k\to\infty}{\to} 0$. Запишемо рівняння у вигляді $\sin x - x\cos x = 0$, розкладемо функцію рівняння для точки шуканого кореня $x_k = x_k^0 - \varepsilon_k$ в ряд Тейлора в околі $x_k^0 = \pi/2 + \pi k$, утримавши члени порядку $O(\varepsilon_k)$ знайдемо перше наближення для $\varepsilon_k \underset{k\to\infty}{=} \varepsilon_k^1 + o(1/k)$:

$$\sin(x_k^0)\cos(\varepsilon_k) - \cos(x_k^0)\sin(\varepsilon_k) - (x_k^0 - \varepsilon_k)\left(\cos(x_k^0)\cos(\varepsilon_k) - \sin(x_k^0)\sin(\varepsilon_k)\right) = 0,$$

$(-1)^k - x_k^0(-1)^k\varepsilon_k^1 = 0$. А отже, $\varepsilon_k^1 = 1/x_k^0 = 1/(\pi/2 + \pi k)$. Повторивши процедуру для $x_k = x_k^0 - \varepsilon_k^1 + \varepsilon_k^2 + o(k^2)$ можна отримати друге наближення ε_k^2. Отже корені віідлені та знаходяться в інтервалах $x_k \in (\pi/2 + \pi k - 1/(\pi/2 + \pi k), \pi/2 + \pi k)$, $k = 1,2,...$

4.4 Метод простої ітерації розв'язання нелінійних рівнянь

Метод простої ітерації використовує поняття стиснутого відображення та базується на теоремі Банаха про існування нерухомої точки відображення.

Озн. Якщо $y = g(x)$ - відображення повного метричного простору в себе з метрикою $\rho(x_1, x_2) \geq 0$ задовольняє умову $\rho(g(x_1), g(x_2)) \leq q\,\rho(x_1, x_2)$, $0 < q < 1$ для довільних x_1, x_2, то таке відображення називається **стиснутим**.

Зауваження. g - в загальному випадку є оператором, в частинному випадку – функцією (векторною чи скалярною).

Озн. Ітераційний процес:

$$x^{k+1} = g(x^k),\ k = 0,1,2,... \tag{4.11}$$

розв'язання рівняння $x = g(x)$ називається **методом простої ітерації**.

Збіжність методу до точного розв'язку X визначається теоремою про достатні умови збіжності методу простої ітерації.

Теорема. Якщо $y = g(x)$ - стиснуте відображення, рівняння $x = g(x)$ має єдиний розв'язок X, який може бути знайдений за методом простої ітерації (4.11) для довільного початкового наближення x^0. При цьому:

$$\rho(X, x^k) \le q^n /(1-q)\rho(x^1, x^0) \tag{4.12}$$

За умову припинення ітерацій можна використати одне з правил:

$$\rho(x^k, x^{k-1}) < \varepsilon_1, \text{ або } \rho(x^k, g(x^k)) < \varepsilon_2 \tag{4.13}$$

Для функціональних рівнянь $\vec{x} = \vec{G}(\vec{x})$, умова стисну тості є умовою Ліпшиця з коефіцієнтом меншим за 1: $\left|\vec{G}(\vec{x}_1) - \vec{G}(\vec{x}_2)\right| < q\,|\,\vec{x}_1 - \vec{x}_2\,|$. У випадку, коли $\vec{G}(\vec{x}) \in C^1_\Omega$, для аналізу стисну тості можна використати теорему Лагранжа: $\exists \xi \in \Omega:\ \vec{G}(\vec{x}_1) - \vec{G}(\vec{x}_2) = \vec{G}'(\xi)(\vec{x}_1 - \vec{x}_2)$. Отже, умова стисну тості $\vec{G}(\vec{x})$ в деякій області в цьому випадку еквівалентна умові $\left|\vec{G}'(\vec{x})\right| < 1$.

Метод простої ітерації аналогічно до методу Зейделя розв'язання СЛАР можна сформулювати в по координатній формі:

$$\begin{cases} x_1^{k+1} = g_1(x_1^k, x_2^k, ..., x_n^k) \\ x_2^{k+1} = g_1(x_1^{k+1}, x_2^k, ..., x_n^k) \\ \qquad ... \\ x_n^{k+1} = g_n(x_1^{k+1}, x_2^{k+1}, ..., x_n^k) \end{cases} \tag{4.14}$$

в якій на кожній ітерації наближення для значення x_i^{k+1} знаходиться з i-го рівняння.

Для випадку системи нелінійних рівнянь $\vec{F}(\vec{x}) = 0$, $\vec{F}(\vec{x}) = \{f_i(x_1,...,x_n)\}_{i=1}^n$, метод по координатного розв'язання має вигляд:

$$\begin{cases} f_1(x_1^{k+1}, x_2^k,..., x_n^k) = 0 \\ f_2(x_1^{k+1}, x_2^{k+1},..., x_n^k) = 0 \\ \quad ... \\ f_n(x_1^{k+1}, x_2^{k+1},..., x_n^{k+1}) = 0 \end{cases} \tag{4.15}$$

Тобто, на кожній ітерації розв'язуються n скалярних рівнянь, причому, наближення для значення x_i^{k+1} знаходиться з розв'язання i-го рівняння.

4.5 Метод Ньютона (Ньютона-Рафсона) та пов'язані з ним методи розв'язання нелінійних рівнянь.

Метод Ньютона – універсальний ітераційний метод розв'язання нелінійних рівнянь

$$\vec{F}(\vec{x}) = 0 \tag{4.16}$$

градієнтного типу, що забезпечує показникові збіжність до точного розв'язку.

Озн. Якщо $\vec{F}(\vec{x})$ - відображення повного нормованого (з нормою $\|.\|$) простору в себе, та існує лінійний оператор $\mathbf{P}$ такий, що

$$\left\|\vec{F}(\vec{x}+\vec{\eta}) - \vec{F}(\vec{x}) - \mathbf{P}\vec{\eta}\right\| \underset{\vec{\eta}\to 0}{=} o(\|\vec{\eta}\|) \tag{4.17}$$

то $\mathbf{P}$ називається **похідною** $\vec{F}(\vec{x})$ в точці $\vec{x}$.

У функціональному випадку: $\vec{F}(\vec{x}) = \{f_i(x_1,...,x_n)\}_{i=1}^n$, очевидно,

$$\mathbf{P} = \vec{F}'(\vec{x}) = \{\partial f_i / \partial x_j\}_{i,j=1}^n \tag{4.18}$$

Для скалярного випадку ($y = f(x)$), $\mathbf{P} = f'(x)$.

Якщо в (4.17) покласти $\vec{x} = \vec{x}^k$, $\vec{\eta} = \vec{X} - \vec{x}^k$, де $\vec{X}$ - точний розв'язок (4.16), $\vec{x}^k$ - наближений розв'язок: $\left\|\vec{F}(\vec{X}) - \vec{F}(\vec{x}^k) - \vec{F}'(\vec{x}^k)(\vec{X} - \vec{x}^k)\right\| \underset{\vec{x}^k\to\vec{X}}{=} o(\|\vec{X} - \vec{x}^k\|)$. Якщо в цьому виразі знехтувати правою частиною, та прийняти до уваги, що $\vec{X}$ - точний

розв'язок, отримаємо: $\vec{F}(\vec{x}^k)+\vec{F}'(\vec{x}^k)(\vec{X}-\vec{x}^k)\approx 0$. Обираємо за покращення наближення $\vec{x}^k$ розв'язок рівняння $\vec{F}(\vec{x}^k)+\vec{F}'(\vec{x}^k)(\vec{x}^{k+1}-\vec{x}^k)=0$ та отримаємо ітераційну процедуру методу Ньютона:

$$\vec{x}^{k+1}=\vec{x}^k-\left(\vec{F}'(\vec{x}^k)\right)^{-1}\vec{F}(\vec{x}^k) \qquad (4.19)$$

Для випадку (4.18), $\left(\vec{F}'(\vec{x}^k)\right)^{-1}$ - обернена матриця до (4.18), для скалярного випадку, $\left(\vec{F}'(\vec{x}^k)\right)^{-1}=1/\ f'(x^k)$.

Збіжність ітераційного процесу методу Ньютона (4.19) визначається теоремою достатності.

Теорема 1. Нехай $\Omega_a=\left\{\vec{x}:\left\|\vec{x}-\vec{X}\right\|<a\right\}$. Якщо існують такі $0\le a_1,a_2\le\infty$, для яких виконуються умови:

$\left\|\left(\vec{F}(\vec{x})\right)^{-1}\right\|\le a_1,\ \wedge\ \left\|\vec{F}(\vec{x}_1)-\vec{F}(\vec{x}_2)-\vec{F}'(\vec{x}_2)(\vec{x}_1-\vec{x}_2)\right\|\le a_2\parallel\vec{x}_1-\vec{x}_2\parallel,\ \vec{x},\vec{x}_1,\vec{x}_2\in\Omega_a$, тоді для довільних початкових значень $\vec{x}^0\in\Omega_b$ ітераційний процес (4.19) збігається до точного розв'язку (4.15) з показниковою швидкістю:

$$\parallel\vec{x}^k-\vec{X}\parallel\le 1/c\left(c\parallel\vec{x}^0-\vec{X}\parallel\right)^{2^k} \qquad (4.20)$$

де $c=a_1\cdot a_2$, $b=\min(a,\cdot 1/c)$.

Для скалярного рівняння $f(x)=0$, (4.19) має вигляд:

$$x^{k+1}=x^k-f(x^k)/\ f'(x^k) \qquad (4.21)$$

Аналогом теореми 1 є теорема Кантаровича:

Теорема 2. Якщо $f(x)\in D^2_{[X-a,X+a]}$, $1/\left|f'(x)\right|\le a_1$, $\left|\left|f(x)/\ f'(x)\right|\le a_2$, $\left|f''(x)\right|\le c<1/(2a_1\cdot a_2)$, $x\in[X-a,X+a]$, $(0\le a_1,a_2,c\le\infty)$ тоді для довільних початкових значень $x^0\in[X-a,X+a]$ (4.21) збігається з показниковою швидкістю:

$$\parallel x^k-X\parallel\le a_2/2^{k-1}\left(2a_1a_2c\right)^{2^{k-1}} \qquad (4.22)$$

(4.21) має просту геометричну інтерпретацію. Дотична до графіку функції $y=f(x)$, проведена в точці x^k має вигляд: $y-f(x^k)=f'(x^k)(x-x^k)$. Отже в

x^{k+1} є точкою перетину дотичною вісі абсцис. Таким чином, (4.21) можна інтерпретувати, як ітераційне покращення розв'язку нелінійного рівняння шляхом побудови розв'язків лінійних наближень за градієнтним методом.

Лінійні (або в більш загальному випадку - поліноміальні) наближення можна будувати з використанням інтерполювання. Найпростішим таким методом є **метод січних**. Нехай визначено 2 точки x_1^k та x_2^k, які задовольняють умову $f(x_1^k)\cdot f(x_2^k)<0$, $x_1^k<x_2^k$, а отже на (x_1^k, x_2^k) знаходиться (вважаємо, що відділенням коренів забезпечується існування та єдність кореня на (x_1^0, x_2^0)) корінь рівняння. Побудуємо в якості наближення функції лінійний інтерполяційний многочлен на точках $(x_1^k, f(x_1^k))$, $(x_2^k, f(x_2^k))$: $y-f(x^k)=\left(f(x_2^k)-f(x_1^k)\right)/(x_2^k-x_1^k)(x-x_1^k)$. Корінь наближеного рівняння має вигляд:

$$x^{k+1}=x_1^k-f(x_1^k)(x_2^k-x_1^k)/\left(f(x_2^k)-f(x_1^k)\right) \qquad (4.23)$$

Очевидно, що $x^{k+1}\in(x_1^k, x_2^k)$ та перевизначимо (x_1^{k+1}, x_2^{k+1}) з умови розташування на ньому кореня вихідного рівняння:

$$\left[\begin{array}{l} f(x_1^k)\cdot f(x^{k+1})<0 \Rightarrow X\in(x_1^k, x^{k+1}) \Rightarrow x_1^{k+1}=x_1^k, x_2^{k+1}=x^{k+1} \\ f(x^{k+1})\cdot f(x_2^k)<0 \Rightarrow X\in(x^{k+1}, x_2^k) \Rightarrow x_1^{k+1}=x^{k+1}, x_2^{k+1}=x_2^k \end{array}\right. \qquad (4.24)$$

Оскільки $(x_2^{k+1}-x_1^{k+1})<(x_2^k-x_1^k)$, ітераційний процес (4.23), (4.24) – метод січних є збіжним (за вкладеністю інтервалів).

За іншою інтерпретацією, (4.23) є аналогом (4.21), в якому $f'(x^k)$ наближається розділеною різницею 1-го порядку $f'(x_1^k)\approx f(x_1^k; x_2^k)=\left(f(x_2^k)-f(x_1^k)\right)/(x_2^k-x_1^k)$.

Метод січних можна легко узагальнити шляхом наближення рівняння поліноміальним рівнянням. Наприклад, на 3 точках (x_1^k, x_2^k, x_3^k) таких, що $x_1^k<x_2^k<x_3^k$, $f(x_1^k)\cdot f(x_3^k)<0$, можна наблизити $f'(x^k)$ комбінацією розділених

різниць до 2-го порядку $f'(x_2^k) \approx f(x_1^k; x_2^k) + f(x_1^k; x_2^k; x_3^k)(x_2^k - x_1^k)$ та побудувати 3-точковий ітераційний метод січних (**метод Мюлєра**)

$$x^{k+1} = x_1^k - f(x_1^k) / \left(f(x_1^k; x_2^k) + f(x_1^k; x_2^k; x_3^k)(x_2^k - x_1^k) \right) \quad (4.25)$$

що використовує наближення вихідного рівняння квадратним рівнянням.

Найпростішим методом розв'язання скалярного нелінійного рівняння є метод **поділу відрізку навпіл** (**метод бісекції**). Він збігається з методом січних, за тією відмінністю, що для знаходження наступного наближення $x^{k+1} \in (x_1^k, x_2^k)$ не використовується лінійне рівняння, а воно береться, як середня точка (x_1^k, x_2^k): $x^{k+1} = (x_1^k + x_2^k)/2$. Процедура (4.24) залишається незмінною.

Узагальненням метода Ньютона є **метод Чебишева побудови ітерацій вищих порядків**, що основується на розкладі в ряд Тейлора оберненої до $f(x)$ функції в околі нуля.

Нехай $X \in [a,b]$ та $f(x) \in C^r_{[a,b]}$ та $f'(x) \neq 0$ для всіх $x \in [a,b]$. В цьому випадку, існує $x = F(y),\ y \in [c,d] : x \equiv F(f(x))$ - обернена до $f(x)$ функція з такими ж властивостями. Для кореня рівняння X має місце $X = F(0)$. Отже, для розв'язання рівняння достатньо побудувати обернену функцію та визначити її значення в нулі. Розкладемо $F(y)$ в ряд Тейлора в околі точки y зі значенням в $y = 0$: $X = F(0) \approx F(y) + \sum_{i=1}^{r} (-1)^i F^{(i)}(y) / i! y^i$, або:

$$X = F(0) \approx x + \sum_{i=1}^{r} (-1)^i F^{(i)}(f(x)) / k! \left(f(x) \right)^k \quad (4.26)$$

(4.26) дозволяє побудувати сімейство ітераційних методів (з параметром r):

$$x^{k+1} = x^k + \sum_{i=1}^{r} (-1)^i F^{(i)}(f(x^k)) / k! \left(f(x^k) \right)^i \quad (4.27)$$

Оскільки $F(f(x)) \equiv x$, то

$$\begin{cases} F'(f(x)) f'(x) = 1 \\ F''(f(x)) f'^2(x) + F'(f(x)) f''(x) = 0 \\ \dots \end{cases} \quad (4.28)$$

Якщо $r=1$ то $F'(f(x))=1/f'(x)$ та $x^{k+1}=x^k-f(x^k)/f'(x^k)$ - метод Ньютона (4.19). Для випадку $r=2$: $F''(f(x))=-F'(f(x))f''(x)/f'^2(x)=-f''(x)/f'^3(x)$ та

$$x^{k+1}=x^k-f(x^k)/f'(x^k)-f''(x^k)f^2(x^k)/f'^3(x^k) \qquad (4.29)$$

5 Чисельні методи розв'язаня задачі Коші для звичайних диференціальних рівнянь

Розв'язання задачі Коші - необхідний крок при отриманні розв'язку більшої частини задач динаміки та аналітичної механіки. Існує багато методів розв'язання задачі Коші, а, отже, і систем їх класифікації. Одна з них - поділ методів розв'язання на однокрокові та багатокрокові. Більшість чисельних методів розв'язання задачі Коші використовують наближення розв'язку, що так чи інакше спирається на дискретне вузлове представлення функції.

Озн. Однокроковим методом розв'язання задачі Коші називається метод, який дозволяє отримати наближене значення функції в дискретному вузлі, використовуючи при цьому властивості розв'язку в одному попередньому вузлі. У випадку, коли для знаходження вузлового значення використовуються властивості розв'язку в k $(k\geq 2)$ попередніх вузлах, називається k**-кроковим** методом.

5.1 Метод розкладу в ряд Тейлора

Озн. Метод розкладу в ряд Тейлора- однокроковий метод, що використовує інтерполяційний многочлен по однократному вузлу.

Нехай треба розв'язати задачу Коші:

$$\begin{cases} y'(x)=f(x,y(x)), x\in[x_0,X] \\ y(x_0)=y_0 \end{cases} \qquad (5.1)$$

У випадку, коли $f(x,y(x)) \in C^{(n)}[x_0, x_0+\Delta)$ ($y = y(x)$- шуканий розв'язок (5.1)), очевидно, за правилом диференціювання складеної функції для $x \in [x_0, x_0+\Delta)$:

$$\begin{cases} y'(x) = f(x,y(x)) \\ y''(x) = f_x(x,y(x)) + f_y(x,y(x))y'(x) \\ y'''(x) = f_{xx}(x,y(x)) + 2f_{xy}(x,y(x))y'(x) + f_{yy}(x,y(x))(y'(x))^2 \\ \dots \\ y^{(n)}(x) = f_{x\dots x}(x,y(x)) + \dots + f_y(x,y(x))y^{(n-1)}(x) \end{cases} \qquad (5.2)$$

Особливість (5.2) полягає в тому, що для визначення $y^{(k)}(x)$ потрібно визначити $y(x), y'(x), \dots, y^{(k-1)}(x)$. А, отже, для відповідного значення $(x_0, y(x_0))$, (5.2) дає змогу визначити $y'(x_0), \dots, y^{(n)}(x_0)$, на яких можна побудувати інтерполяційний многочлен степені n (x_0 - кратний вузол з кратністю $n+1$). Цей многочлен, як було зазначене, є відрізком ($n+1$-членним) ряду Тейлора:

$$y(x) \approx \sum_{i=0}^{n} y^{(i)}(x_0)/i!\,(x-x_0)^i, \; x \in [x_0, x_0+\Delta) \qquad (5.3)$$

(5.3) в сукупності з рекурсивним правилом (5.2) складають метод розкладу в ряд Тейлора.

Нехай $f(x,y(x)) \in C^{(n)}[x_0, x_0+X]$. Тоді, розбивши $[x_0, x_0+X]$ на N проміжків $[x_0, x_0+X] = \bigcup_{j=1}^{N} [x_{j-1}, x_j]$ так, що $x_N = x_0 + X$, та, позначивши за $y_j^{(n)} = y^{(n)}(x_j)$, $j = \overline{0,N}$, маємо з (5.2):

$$\begin{cases} y_0^{(1)} = f(x_0,y_0) \\ y_0^{(2)} = f_x(x_0,y_0) + f_y(x_0,y_0)y_0^{(1)} \\ \dots \\ y_0^{(n)} = f_{x\dots x}(x_0,y_0) + \dots + f_y(x_0,y_0)y_0^{(n-1)} \end{cases}$$

та з (5.3):

$y(x_1) \approx \sum_{i=0}^{n} y^{(i)}(x_0)/i!\,(x-x_0)^i$. Отже, знаходимо $y_1 \approx y(x_1)$.

Процедуру легко поширити на інші вузли. Нехай знайдено $y_k \approx y(x_k)$. Тоді,очевидно:

$$\begin{cases} y_k{}^{(1)} = f(x_k, y_k) \\ \dots \\ y_k{}^{(n)} = f_{x\dots x}(x_k, y_k) + \dots + f_y(x_k, y_k) y_k{}^{(n-1)} \end{cases} \qquad (5.4)$$

$$y(x_{k+1}) \approx y_{k+1} = \sum_{i=0}^{n} y_k{}^{(n)}/i!(x_{k+1}-x_k)^i$$

(5.4) при $k = \overline{0, N-1}$ і є алгоритм методу розкладу в ряд Тейлора n-го порядку.

Точність наближення вузлових значень $y(x_j) \approx y_j$ можна визначиити за формулою залишку ряду Тейлора або за формулою залишку інтерполювання з кратним вузлом.

Якщо $h = \max_{j=\overline{1,N}} |x_j - x_{j-1}|$, то, очевидно,:

$$\max_{0 \le j \le N} |y_j - y(x_j)| \underset{h \to 0}{=} \underset{=}{O}(h^{n+1}). \qquad (5.5)$$

5.2 Методи Рунге-Кутта

Іншим методом побудови методів розв'язання задачі Коші є використання квадратурних формул при прямому інтегруванні диференціального виразу (4.1). Дійсно: $y(x+h) = y(x) + \int_0^h y'(x+t)dt$. Замінивши інтеграл в правій частині наближеним значенням $h\,y'(x)$ з точністю до $\underset{=}{O}(h^2)$, матимемо: $y(x+h) = y(x) + hy'(x) + \underset{=}{O}(h^2)$. Зауважимо, що цей вираз в точності збігається з виразом, що лежить в основі методу розкладу в ряд Тейлора 1-го порядку, а,

отже, може бути отриманий шляхом наближення функції $y(x+h)$ інтерполяційним многочленом з 2-х кратним вузлом x степені 1.

Якщо інтеграл наблизити формулою трапеції , то, очевидно:

$$y(x+h) = y(x) + \frac{h}{2}(y'(x) + y'(x+h)) + \underset{=}{O}(h^3) \tag{5.6}$$

якщо прямокутників з кратним вузлом, то:

$$y(x+h) = y(x) + hy'\left(x + \frac{h}{2}\right) + \underset{=}{O}(h^3) \tag{5.7}$$

Альтернативою такої побудови є використання побудови інтерполяційного многочлену з кратними вузлами.

Дійсно, побудуємо $g_3(t)$, $t \in [x, x+h]: g_3(x) = y(x)$, $g_3'(x) = y'(x)$, $g_3'(x+h) = y'(x+h)$. Такий многочлен $g_3(t) = a_0 + a_1 t + a_2 t^2$ легко побудувати за методом невизначених коефіцієнтів. Очевидно, що $g_3(x+h) = y(x) + \frac{h}{2}(y'(x) + y'(x+h))$ та $| y(x+h) - g_3(x+h) | = \underset{=}{O}(h^3)$.

Аналогічно, для отримання (5.7) достатньо побудувати інтерполяційний многочлен $\tilde{g}_3(x)$ за умови $\tilde{g}_3(x) = g(x)$, $\tilde{g}_3\left(x + \frac{h}{2}\right) = y\left(x + \frac{h}{2}\right)$, $\tilde{g}_3'\left(x + \frac{h}{2}\right) = y'\left(x + \frac{h}{2}\right)$.

Такий підхід не є надто зручним, але він явно демонструє той факт, що методи Рунге-Кутта мають в основі теорію інтерполювання розв'язку задачі Коші многочленами.

З (5.5) та (5.7) отримаємо алгоритм

$$y_{j+1} = y_j + \frac{h}{2}(f(x_j, y_j) + f(x_{j+1}, y_{j+1})) \tag{5.8}$$

та

$$y_{j+1} = y_j + hf(x_{j+1/2}, y_{j+1/2}),\ j = \overline{0, N-1}. \tag{5.9}$$

(5.8) та (5.9) мають суттєвий недолік: $y_{j+1}(y_{j+1/2})$ входять в праві частини виразів алгоритмів. Отже, алгоритм потребує розв'язання рівнянь. Цього можна позбутись, помітивши, що

$$\begin{cases} y(x+h) = y(x) + hy'(x) + \underset{=}{O}(h^2) \\ y(x+h/2) = y(x) + h/2y'(x) + \underset{=}{O}(h^2) \end{cases} \tag{5.10}$$

Підстановка (4.10) в праві частини (5.6) та (5.7) дасть:

$$\begin{cases} y(x+h) = y(x) + h/2(y'(x) + (y(x) + \underset{=}{hy'}(x))) + O(h^3) \\ y(x+h) = y(x) + h\big(y(x) + h/2y'(x)\big) + \underset{=}{O}(h^3) \end{cases} \tag{5.11}$$

Очевидно, в (5.11) точність визначення $y(x+h)$ є такою ж, що і в (5.6) та (5.7). Отже, явний алгоритм, що дає еквівалентний по точності розв'язок (5.8), (5.9), має вигляд:

$$y_{j+1} = y_j + \frac{h}{2}(f(x_j, y_j) + f(x_{j+1}, y_j + hf(x_j, y_j))) \tag{5.12}$$

$$y_{j+1} = y_j + hf(x_{j+1/2}, y_j + h/2\, f(x_j, y_j)) \tag{5.13}$$

Озн. Алгоритм знаходження наближених розв'язків задачі Коші (5.12) та (5.13) називаються **модифікованими методами Ейлера**.

Озн. Методом Ейлера називається алгоритм

$$y_{j+1} = y_j + hf(x_j, y_j),\ j = \overline{0, N-1}. \tag{5.14}$$

Застосування даного підходу в альтернативному способі побудови алгоритмів через використання інтерполяційні многочлени виглядає наступним чином. Нехай треба побудувати

$g_3(t) = a_0^{(3)} + a_1^{(3)}t + a_2^{(3)}t^2,\ t \in [x, x+h]$

$g_3(x) = y(x),\ g_3'(x) = y'(x),\ g_3'(x+h) = y'(x+h).$

Такий многочлен побудовано раніше, але його коефіцієнти залежать від $y'(x+h)$. При цьому, як показано, $|y(x+h) - g_3(x+h)| = \underset{=}{O}(h^3)$. Побудуємо

$g_2(t) = a_0^{(2)} + a_1^{(2)}t,\ t \in [x, x+h],\ g_2(x) = y(x),\ g_2'(x) = y'(x).$

При цьому, очевидно, $a_0^{(2)} = a_0^{(3)}$ та $a_1^{(2)} = a_1^{(3)}$. Дійсно, побудова $g_3(t)$ зводится до відшукання $a_2^{(3)}$ з умови $a_1^{(2)} + 2a_1^{(3)}(x+h) = y'(x+h)$. Побудуємо інший многочлен (він насправді не є інтерполяційним в повному сенсі), такий, що

$\tilde{g}_3(t) = a_0^{(2)} + a_1^{(2)}t + \tilde{a}_3^{(3)}t^2$, $\tilde{g}_3(x) = y(x)$, $\tilde{g}_3'(x) = y(x)$,

$\tilde{g}_3'(x+h) = f(x+h, g_2(x+h))$.

При цьому,

$|y(x+h) - \tilde{g}_3(x+h)| \le |y(x+h) - g_3(x+h)| + |g_3(x+h) - \tilde{g}_3(x+h)| = \underline{\underline{O}}(h^2)$ оскільки $\tilde{g}_3(t)$ та $g_3(t)$ відрізняються лише старшими членами.

Ця ідея може бути покладена в основу підвищення порядку інтерполювання без введення додаткових вузлових значень.

Нехай $L_n(x) = \sum_{i=1}^{n} a_i x^{i-n} \approx f(x)$, $x \in [a,b]$ так, що $L_n(x_j) = f(x_j)$, $j = \overline{1,n}$, $x_j \in [a,b]$. Побудуємо $\tilde{L}_{n+1}(x_j) = L_n(x) + \tilde{a}_{n+1}x^n$ з умови $\tilde{L}_{n+1}(x_{n+1}) = L_n(x_{n+1})$, $x_{n+1} \in [a,b]$, $x_{n+1} \ne x$, $j = \overline{1,n}$. Многочлен $\tilde{L}_{n+1}(x)$ не є інтерполяційним многочленом порядку $n+1$ для $f(x)$ по $\{x_i\}_{i=1}^{n+1}$, оскільки $\tilde{L}_{n+1}(x_{n+1}) \ne f(x_{n+1})$. Але він не потребує визначення $f(x_{n+1})$ та має такий же порядок залишку, що і $L_{n+1}(x)$.

Підхід побудови однокрокових методів типу (5.12)-(5.14) можна узагальнити наступним чином. За стандартною процедурою, вводимо невизначені параметри $\{\alpha_i\}_{i=2}^{q}$, $\{p_i\}_{i=1}^{q}$, $\{\beta_{ij}\}_{0<j<i\le q}$ та будуємо рекурентно задані вирази:

$$\begin{cases} k_1(h) = hf(x,y) \\ k_2(h) = hf(x,+\alpha_2 h, y + \beta_{21}k_1(h)) \\ \dots \\ k_q(h) = hf(x+\alpha_{q1}h, y + \beta_{q1}k_1(h) + \dots + \beta_{qq-1}k_{q-1}(h)) \end{cases} \qquad (5.15)$$

Покладемо

$$y(x+h) \approx z(h) = y(x) + \sum_{i=1}^{p} p_i k_i(h). \qquad (5.16)$$

Визначивши $\phi(h) = y(x+h) - z(h)$ - функцію h, накладемо наступні умови: $\phi(0) = \phi'(0) = \ldots = \phi^{(s)}(0) = 0$, де s - максимально можливе значення при фіксованому q.

Озн. Алгоритм розвязання задачі Коші (5.15), (5.16) називається **методом Рунге-Кутта.**

Найбільш широко вживані методи Рунге-Кутта 3-го порядку:

$$\begin{cases} K_1 = hf(x, y) \\ K_2 = hf(x + h/2, y + K_1/2) \\ K_3 = hf(x, +h, y - K_1 + 2K_2) \\ y(x+h) = y(x) + 1/6(K_1 + 4K_2 + K_3) \end{cases} \tag{5.17}$$

та метод Рунге кута 4-го порядку. У вигляді алгоритму він записується так

$$\begin{cases} k_1 = hf(x_j, y_j) \\ k_2 = hf\left(x_{j+1/2}, y_j + k_1/2\right) \\ k_3 = hf\left(x_{j+1/2}, y_j + k_2/2\right) \\ k_4 = hf(x_{j+1}, y_j + k_3) \\ y_{j+1} = y_j + 1/6(k_1 + 2k_2 + 2k_3 + k_4) \end{cases} \tag{5.18}$$

Оцінка точності (5.15), (5.16) може бути отримана з побудови методу. Дійсно, $\phi(h) = y(x+h) - z(h) = \phi(0) + \ldots + \phi^{(s)}(0)h^s + \underset{=}{O}(h^{s+1})$. Отже, очевидно, $| y(x_{j+1}) - y_{j+1} | \approx \underset{=}{O}(h^{s+1})$. (5.16) будується для $s = 3$, отже похибка має порядок h^4. Для формули (5.17) $s = 4$ (похибка порядку h^5).

5.3 Багатокрокові методи розв'язання задачі Коші.

Озн. Нехай наближене розв'язання задачі Коші зведене до знаходження наближень вузлових значень $y_j = y(x_j)$, $j = \overline{0,N}$, $x_0 < \ldots < x_N = x_0 + X$. Якщо алгоритм розв'язання зводиться до виду

$$y_j = F(f; x_j, \ldots, x_{j-k}; h; y_{j-1}, \ldots, y_{j-k}), \tag{5.19}$$

то такий метод називається k**-кроковим методом**.

Очевидно, методи Рунге-Кутта в незалежності від порядку є однокроковими методами.

Найбільш поширеними серед k-крокових методів інтегрування задачі Коші є скінченно-різницеві методи (на сітці з постійним кроком).

Озн. Методи інтегрування задачі Коші з алгоритмом

$$\sum_{i=0}^{n} a_{-i} y_{j-1} - h\sum_{i=0}^{k} b_{-i} f(x_{n-i}, y_{n-i}) = 0 \tag{5.20}$$

називаються **скінченно-різницевими**.

Скінченно-різницеві схеми, як і метод Рунге-Кутта, можна побудувати на основі квадратурних формул.

Дійсно, якщо існує квадратурна формула виду

$\int_{-ph}^{0} f(x)dx \approx h\sum_{i=0}^{m} b_{-i} f(x_{-i})$, то, очевидно,

$$\int_{-ph}^{0} y'(x + x_j)dx = y(x_j) - y(x_{j-p}) \approx h\sum_{i=0}^{m} b_{-i} y'(x_{j-i}) = h\sum_{i=0}^{m} b_{-i} f(x_{j-i}, y_{j-i}).$$

Отже, відповідна скінченно-різницева схема має вигляд:

$$y_j - y_{j-p} - h\sum_{i=0}^{m} b_{-i} f(x_{j-i}, y_{j-i}) = 0 \tag{5.21}$$

Вид інтерполяції (екстраполяції) при отриманні (5.21) лежить в основі класифікації скінченно-різницевих схем.

Озн. Якщо в (5.21) $a_0 \neq 0$, $b_0 = 0$,то схема називається **екстраполяційною**. У випадку $a_0 \neq 0$, $b_0 \neq 0$ схема називається **інтерполяційною**. У випадку $a_0 = 0$, $b_0 \neq 0$ - схемою **із забіганням вперед**.

Очевидно, лише екстраполяційні схеми є явними схемами, в яких визначення наближених вузлових значень не потребує розв'язання додаткових рівнянь.

Приклад. Квадратурна формула Сімпсона має вигляд

$$\int_{-2h}^{0} f(x)dx \approx \frac{h}{3}(f(0)+4f(-h)+f(-2h)).$$

Їй відповідає скінченно-різницева схема виду

$$y_j - y_{j-2} = \frac{h}{3}(f_j + 4f_{j-1} + f_{j-2}).$$

Це інтерполяційна 2-х крокова схема. Тобто, визначення y_j на кожному кроці потребує розв'язання деякого рівняння , оскільки y_j присутнє як в лівій частині, так і в правій частині (5.23). В той же час, можна побудувати екстраполяційну формулу Сімпсона:

$$\int_{0}^{h} f(x)dx \approx \frac{h}{12}(23f(0)-4f(-h)+f(-2h)).$$

Відповідна 3-х крокова екстраполяційна скінченно-різницева схема:

$$y_j - y_{j-1} = \frac{h}{12}(23f_{j-1} - 4f_{j-2} + 5f_{j-3}) \qquad (5.22)$$

Озн. (5.21) та (5.22) отримані в припущенні $p=1$. Такі скінченно-різницеві схеми називають **схемами Адамса.**

В загальному випадку, для побудови інтерполяційних та екстраполяційних схем Адамса, можна використати інтерполяційний многочлен Ньютона зі скінченними різницями назад.

Нехай $x_i = ih,\ ,i=\overline{0,-m}$ - рівномірна сітка з кроком h. Тоді, очевидно, $\int_0^h f(x)dx = h\int_0^1 f(ht)dt,\quad \int_{-h}^{0} f(x)dx = h\int_{-1}^{0} f(ht)dt.$ Побудуємо інтерполяційний многочлен Ньютона зі скінченними різницями назад функції $f(x)$ на вказаній сітці

$$f(ht) \approx N_{m+1}^{-}(t) = f_0 + f_{-1/2}^{1}t + f_{-1}^{2}/2!t(t+1) + \ldots + f_{-m/2}^{m}/m!t\ldots(t+(m-1)) \qquad (5.23)$$

із залишковим членом

$f(ht) - N_{m+1}^{-}(t) = \dfrac{f^{(m+1)}(\zeta)}{h!} h^{m+1} \omega_{n+1}^{*}(t)$, де $\zeta \in [-mh, \max(ht, 0)]$ та $\omega_{n+1}^{*}(t) = \prod_{i=0}^{m} (t+i)$.

Проінтегрувавши (5.23) на проміжку [0,1], отримаємо екстраполяційну квадратурну формулу (інтегрування виконується поза проміжком інтерполювання)

$$\int_0^h f(x)dx \approx S_{m+1}^{+}[f] = h\sum_{i=0}^{m} \gamma_i^{+} f_{-i/2}^{i} \tag{5.24}$$

з оцінкою залишку

$$\left| \int_0^h f(x)dx - S_{m+1}^{+}[f] \right| \le \gamma_{m+1}^{+} h^{m+2} \max_{x \in [-mh,1]} | f^{(m+1)}(x) |$$

де $\gamma_i^{+} = \dfrac{1}{i!} \int_0^1 \prod_{j=0}^{i-1} (t+j)$. (5.24) записано через скінченні різниці $f_{-i/2}^{i}$. Якщо згадати формулу (2.28), то (5.24) можна переписати через вузлові значення $S_{m+1}^{+}[f] = h\sum_{i=0}^{m} d_i^{+} f_{-i}$. Відповідна схема Адамса буде мати вигляд:

$$y_j - y_{j-1} = h\sum_{i=0}^{m} d_i^{+} f_{j-i-1} \tag{5.25}$$

з оцінкою похибки $| y(x_j) - y_j | \le \gamma_{m+1}^{+} h^{m+2} \max_{x \in [-mh,1]} | y^{(m+1)}(x) |$.

Приклад. Нехай $m = 4$. Побудуємо екстраполяційну формулу Адамса. Очевидно, $\{\gamma_i^{+}\}_{i=0}^{5} = \{1; 1/2; 5/12; 3/8; 25/720; 475/12\}$. При цьому

$f_0^0 = f_0$, $\quad f_{-1/2}^1 = f_0 + f_{-1}$, $\quad f_{-1}^2 = f_0 - 2f_{-1} + f_{-2}$, $\quad f_{-3/2}^3 = f_0 - 3f_{-1} + 3f_{-2} - f_{-3}$,

$f_{-2}^4 = f_0 - 4f_{-1} + 6f_{-2} - 4f_{-3} + f_{-4}$.

В результаті отримаємо

$$y_j - y_{j-1} = \frac{h}{720}(1901 f_{j-1} - 2774 f_{j-2} + 2616 f_{j-3} - 1274 f_{j-4} + 251 f_{j-5}) \tag{5.26}$$

з оцінкою похибки

$$| y(x_j) - y_j | \le \frac{95}{288} h^6 \max_{x \in [-5h,0]} | y^{(5)}(x) |. \tag{5.27}$$

Озн. Екстраполяційні формули Адамса називають **скінченно-різницевими схемами Адамса-Бошфорта**.

Інтерполяційні квадратурні формули можна отримати аналогічно, при інтегруванні (5.23) на $[-1,0]$. Наприклад:

$$\int_{-h}^{0} f(x)dx \approx S_{m+1}^{-}[f] = h\sum_{i=0}^{m} \gamma_i^{-} f_{-i/2}^{i}$$

з оцінкою похибки

$$\left| \int_{-h}^{0} f(x)dx - S_{m+1}^{-}[f] \right| \le \gamma_{m+1}^{-} h^{m+2} \max_{x\in[-mh,0]} |f^{(m+1)}(x)|, \text{ де } \gamma_i^{-} = \frac{1}{i!}\int_{-1}^{0}\prod_{j=0}^{i-1}(t+j).$$

Аналогічно до попереднього, квадратурну формулу можна записати через вузлові значення $S_{m+1}^{-}[f] = h\sum_{i=0}^{m} d_i^{-} f_{-i}$. Відповідна інтерполяційна схема Адамса буде мати вигляд

$$y_j - y_{j-1} = h\sum_{i=0}^{m} d_i^{-} f_{j-i} \qquad (5.28)$$

з оцінкою похибки

$$|y(x_j) - y_j| \le |\gamma_{m+1}^{-}| h^{m+2} \max_{x\in[-mh,0]} |y^{(m+1)}(x)| \qquad (5.29)$$

Озн. Інтерполяційні формули Адамса називають **скінченно-різницевими схемами Адамса-Моултона**.

Приклад. Побудуємо 5-ти крокову схему Адамса-Моултона. Очевидно,

$$\{\gamma_i^{-}\}_{i=0}^{6} = \{1;-1/2;-1/12;-1/24;19/720;-3/160;-863/50480\}.$$

При цьому,

$$f_0^0 = f_0, \quad f_{-1/2}^1 = f_0 - f_{-1}, \quad f_{-1}^2 = f_0 - 2f_{-1} + f_{-2}, \quad f_{-3/2}^3 = f_0 - 3f_{-1} + 3f_{-2} - f_{-3},$$

$$f_{-2}^4 = f_0 - 4f_{-1} + 6f_{-2} - 4f_{-3} + f_{-4}, \quad f_{-5/2}^3 = f_0 - 5f_{-1} + 10f_{-2} - 10f_{-3} + 5f_{-4} - f_{-5}.$$

Скінченно-різницева схема

$$y_j - y_{j-1} = \frac{h}{1440}(475f_j + 1427f_{j-1} - 798f_{j-2} + 482f_{j-3} - 173f_{j-4} + 27f_{j-5}) \qquad (5.30)$$

з оцінкою похибки

$$| y(x_j) - y_j | \le | \frac{863}{60480} h^7 \max_{x \in [-5h, 0]} | y^{(6)}(x) | . \qquad (5.31)$$

(5.26) та (5.30) мають свої недоліки. (5.30) має вищу точність (5.31) в порівнянні з (5.26) за рахунок звуження відрізку визначення характеристики гладкості функції. В свою чергу, (5.26) на відміну від (5.30) є екстраполяційна. Як правило, (5.26) та (5.30) об'єднують в один алгоритм.

Озн. Методи, в яких екстраполяційні скінченно-різницеві схеми використовуються для довизначення неявних схем називаються **методами прогнозу та корекції**.

Метод прогнозу та корекції в явній та неявній схемах Адамса має вигляд:

$y_j^* - y_{j-1} = h \sum_{i=0}^{m-2} d_i^+ f_{j-i-1}$ - $m-1$-крокова схема Бошфорта,

$$y_j - y_{j-1} = h \left\{ d_0^- f_j^* + \sum_{i=1}^{m} d_i^- f_{j-i} \right\}, \qquad (5.32)$$

де $f_j^* = f(x_j, y_j^*)$.

Зауваження. Згідно з означенням, модифіковані методи Ейлера та Рунге-Кутта є однокроковими методами прогнозу та корекції.

Зауваження. Легко бачити, що порядок оцінки точності порядок точності $O(h^{m+2})$ для методу прогнозу та корекції (5.32) залишається незмінним (5.29).

Приклад. Побудуємо 5-ти кроковий метод Адамса з прогнозом та корекцією. Використовуючи попередні приклади, маємо

$$y_j^* = y_{j-1} + \frac{h}{720}(1901 f_{j-1} - 2774 f_{j-2} + 2616 f_{j-3} - 1274 f_{j-4} + 251 f_{j-5}),$$

$$f_j^* = f(x_j, y_j^*),$$

$$y_j - y_{j-1} = \frac{h}{1440}(475 f_j + 1427 f_{j-1} - 798 f_{j-2} + 482 f_{j-3} - 173 f_{j-4} + 27 f_{j-5}).$$

5.4 Інтегрування задачі Коші для систем диференціальних рівнянь

У випадку, коли система диференціальних рівнянь 1-го порядку розв'язана відносно похідної

$$\vec{y}' = \vec{f}(x, \vec{y}),\ \vec{y}(x_0) = \vec{y}_0, \tag{5.33}$$

більшість методів переноситься з врахуванням векторності задачі.

Так модифікований метод Ейлера має вигляд:

$$\vec{y}_{j+1} = \vec{y}_j + \frac{h}{2}\left(\vec{f}(x_j, \vec{y}_j) + \vec{f}(x_{j+1}, \vec{y}_j + h\vec{f}(x_j, \vec{y}_j))\right) \tag{5.34}$$

Метод Рунге-Кутта для систем диференціальних рівнянь будується з тих же міркувань, що і в скалярному випадку. Так, метод Рунге-Кутта 4-го порядку має вигляд:

$$\begin{cases} \vec{k}_1 = h\vec{f}(x_j, \vec{y}_j) \\ \vec{k}_2 = h\vec{f}\left(x_j + h/2, \vec{y}_j + h/2\vec{k}_1\right) \\ \vec{k}_3 = h\vec{f}\left(x_j + h/2, y_j + h/2\vec{k}_2\right) \\ \vec{k}_4 = h\vec{f}(x_j + h, \vec{y}_j + h\vec{k}_3) \\ \vec{y}_{j+1} = \vec{y}_j + 1/6(\vec{k}_1 + 2\vec{k}_2 + 2\vec{k}_3 + \vec{k}_4) \end{cases} \tag{5.35}$$

Скінченно-різницеві методи також можуть бути легко переформульовані. Так, розглянутий в Прикладі 5-ти кроковий метод Адамса прогнозу та корекції має вигляд:

$$\begin{aligned} &\vec{y}^*_{j+1} - \vec{y}_j = 1/720(1901\vec{f}_j - 2774\vec{f}_{j-1} + 2616\vec{f}_{j-2} - 1274\vec{f}_{j-3} + 251\vec{f}_{j-4}), \\ &\vec{y}_{j+1} - \vec{y}_j = 1/1440(475\vec{f}(x_{j+1}, \vec{y}^*_{j+1}) + 1427\vec{f}_j - 798\vec{f}_{j-1} + \\ &482\vec{f}_{j-2} - 173\vec{f}_{j-3} + 27\vec{f}_{j-4}) \end{aligned}. \tag{5.36}$$

5.5 Розв'язання задачі Коші для диференціальних рівнянь 2-го порядку

Якщо диференціальне рівняння 2-го порядку (не має значення скалярне чи векторне) розв'язане відносно старшої похідної $\vec{y}'' = \vec{f}(x, \vec{y}, \vec{y}')$, то простою заміною змінних $\vec{y}' = \vec{z}$ його, очевидно, можна звести до системи диференціальних рівнянь 1-го порядку:

$$\begin{cases} \vec{z}' = \vec{f}(x, \vec{y}, \vec{z}) \\ \vec{y}' = \vec{z} \end{cases}$$

Тим не менше, оскільки велика кількість задач механіки (зокрема всі задачі динаміки) потребують розв'язання рівнянь 2-го порядку, та, окрім того, спеціалізовані алгоритми завжди більш ефективні та універсальні, розглянемо побудову таких методів на базі метода Штермера.

$$\vec{y}'' = \vec{f}(x, \vec{y}), \; x \in (x_0, x_0 + X) \tag{5.37}$$

$$\begin{cases} \vec{y}(x_0) = \vec{y}_0 \\ \vec{y}'(x_0) = \vec{v}_0 \end{cases}$$

Загальний вигляд k-крокового явного методу Штермера:

$$\vec{y}_{j+1} - 2\vec{y}_j + \vec{y}_{j-1} = h^2 \sum_{i=0}^{k} b_{-i}^{-} \vec{f}(x_{j-i}, \vec{y}_{j-i}) \tag{5.38}$$

та неявного:

$$\vec{y}_{j+1} - 2\vec{y}_j + \vec{y}_{j-1} = h^2 \sum_{i=0}^{k} b_{-i}^{+} \vec{f}(x_{j-i}, \vec{y}_{j-i}) \tag{5.39}$$

Отримати b_{-i}^{-} та b_{-i}^{+} можна як і для методу Адамса, базуючись на застосуванні інтегральної квадратури до виразу:

$$\vec{y}(x) = \vec{y}(x_j) + \vec{y}'(x_j)(x - x_j) + \int_{x_j}^{x} (x - t)\vec{y}''(t)dt.$$

Очевидно

$$\begin{cases} \vec{y}(x_{j-1}) = \vec{y}(x_j) + \vec{y}'(x_j)h - \int\limits_{x_{j-1}}^{x_j} (x_{j-1} - t)\vec{y}''(t)dt \\ \vec{y}(x_{j+1}) = \vec{y}(x_j) + \vec{y}'(x_j)h + \int\limits_{x_j}^{x_{j+1}} (x_{j+1} - t)\vec{y}''(t)dt \end{cases}$$

а, отже,

$$\vec{y}(x_{j+1}) - 2\vec{y}(x_j) + \vec{y}(x_{j-1}) = \int\limits_{x_{j-1}}^{x_j} (t - x_{j-1})\vec{y}''(t)dt + \int\limits_{x_j}^{x_{j+1}} (x_{j+1} - t)\vec{y}''(t)dt \ . \qquad (5.40)$$

Замінивши в лівій частині значення функції на наближені вузлові значення, а в правій частині $\vec{y}''(t)$ на інтерполяційний многочлен назад, отримаємо (5.38) та (5.39). Як і для методу Адамса, (5.38) можна отримати, побудувавши інтерполяційний многочлен на точках $x_j, x_{j-1}, \ldots, x_{j-k}$. Відповідно, для отримання (5.39) - на точках $x_{j+1}, x_{j+21}, \ldots, x_{j-k}$.

Приклад. Побудуємо 4 крокову явну та неявну схеми Штермера та об'єднаємо їх в схему метода прогнозу та корекції.

$$\int\limits_{x_{j-1}}^{x_j} (x - x_{j-1}) f(x)dx + \int\limits_{x_j}^{x_{j+1}} (x_{j+1} - x_j) f(x)dx \approx h^2 \sum_{i=0}^{k} d_{-i}^{-} f(x_{-i}) \text{ при } k = 4.$$

Маємо $d_0 = 299/240$, $d_{-1} = -176/240$, $d_{-2} = 194/240$, $d_{-3} = -96/240$, $d_{-4} = 19/240$. Отже, 4 крокова явна схема Штермера має вигляд

$$\vec{y}_{j+1} - 2\vec{y}_j + \vec{y}_{j-1} = h^2/240\left(299\vec{f}_j - 176\vec{f}_{j-1} + 194\vec{f}_{j-2} - 96\vec{f}_{j-3} + 19\vec{f}_{j-4}\right). \qquad (5.41)$$

Аналогічно, для явної схеми маємо

$$\int\limits_{x_{j-1}}^{x_j} (x - x_{j-1}) f(x)dx + \int\limits_{x_j}^{x_{j+1}} (x_{j+1} - x_j) f(x)dx \approx h^2 \sum_{i=-1}^{k} d_{-i}^{+} f(x_{-i}), \ k = 4.$$

$$\vec{y}_{j+1} - 2\vec{y}_j + \vec{y}_{j-1} = \frac{h^2}{240}\left(18\vec{f}_{j+1} + 209\vec{f}_j + 4\vec{f}_{j-1} + 14\vec{f}_{j-2} - 6\vec{f}_{j-3} + \vec{f}_{j-4}\right). \qquad (5.42)$$

Отже, метод прогнозу та корекції має вигляд:

$$\begin{cases} \vec{y}^*_{j+1} = 2\vec{y}_j - \vec{y}_{j-1} + h^2/240\left(299\vec{f}_j - 176\vec{f}_{j-1} + 194\vec{f}_{j-2} - 96\vec{f}_{j-3} + 19\vec{f}_{j-4}\right), \\ \vec{f}^*_{j+1} = \vec{f}(x_{j+1}, y^*_{j+1}), \\ \vec{y}_{j+1} = 2\vec{y}_j - \vec{y}_{j-1} + h^2/240\left(18\vec{f}^*_{j+1} + 209\vec{f}_j + 4\vec{f}_{j-1} + 14\vec{f}_{j-2} - 6\vec{f}_{j-3} + \vec{f}_{j-4}\right) \end{cases} \qquad (5.43)$$

Для розв'язання задачі Коші для рівняння $\vec{y}'' = \vec{f}(x, \vec{y}, \vec{y}')$ можна рекурентним чином використовувати схеми розв'язання задач Коші для рівнянь першого та другого порядків. Дійсно, маємо:

$$\begin{cases} \vec{z}' = \vec{f}(x, \vec{y}, \vec{z}) \\ \vec{y}' = \vec{z} \end{cases}, \quad x \in (x_0, x_0 + X)$$

$$\begin{cases} \vec{y}(x_0) = \vec{y}_0 \\ \vec{z}(x_0) = \vec{v}_0 \end{cases}$$

Отже, наприклад, явна k-крокова схема Адамса

$$\begin{cases} \vec{z}_{j+1} - \vec{z}_j = h\sum_{i=0}^{k} d^-_{-i}\vec{f}(x_{j-i}, \vec{y}_{j-i}, \vec{z}_{j-i}) \\ \vec{y}_{j+1} - \vec{y}_j = h\sum_{i=0}^{k} d^-_{-i}\vec{z}_{j-i} \end{cases} \qquad (5.44)$$

Література

1. Н.С. Бахвалов. Численные методы. т. 1. М. Наука 1973.
2. Математический практикум /Под ред. Г.Н. Положего/. Гос. Изд-во Физ-Мат Лит-ры, 1960.
3. И.С. Березин, Н.П. Жидков. Методы вычислений. т. 1, 2. 1966.
4. А.А. Самарский. Введение в численные методы. 1984.
5. И.И. Крылов, В.В. Бобков, П.И. Монастырский. Вычислительные методы. т. 1, 2. 1977.
6. И.И. Ляшко, В.Л. Макаров, А.А. Скоробагатько. Методы вычислений. 1977.
7. І.П. Гаврилюк, В.Л. Макаров. Методи обчислень. т. 1, 2. Вища школа. 1995.
8. Сборник задач по методам вычислений /Под ред. П.И. Монастырского/. Изд-во БГУ, 1983.
9. Г.И. Марчук. Методы вычислительной математики. 1977.
10. Н.С. Бахвалов, Н.П. Жидков, Г.И. Кобельков. Численные методы. 1977.
11. А.А. Самарский, А.В. Гулин. Численные методы. М. Наука, 1989.
12. Н.А. Кильчевский. Курс теоретической механики. Т. 1, 2, М.: Наука, 1972г.
13. Гантмахер Ф.Р. Лекции по аналитической механике. М. Наука. 1966.
14. И.В. Мещерский. Сборник задач по теоретической механике. М.: Наука, 1980-1986

Зміст