Ю. В. Прокопенко,
Д. Д. Татарчук,
В. А. Казміренко

ОБЧИСЛЮВАЛЬНА МАТЕМАТИКА

Затверджено Методичною радою НТУУ «КПІ»
як навчальний посібник для студентів
бакалаврату напряму «Електроніка»

Київ
«Політехніка»
2013

УДК 519.6 (075.8)

Рецензенти: *С. С. Забара*, докт. техн. наук, проф.
В. Г. Вербицький, докт. техн. наук, снс

Відповідальний редактор *В. І. Тимофєєв,* докт. техн. наук, проф.

Прокопенко Ю. В., Татарчук Д. Д., Казміренко В. А.

Обчислювальна математика: Навч. посіб. – К.: Видавництво, 2013. – 224 с.

Викладено навчальний матеріал про чисельне розв'язання рівнянь і систем, знаходження власних чисел і векторів, інтерполяції та апроксимації табличних залежностей, чисельного інтегрування та диференціювання функцій, а також розв'язання диференціальних рівнянь і задач оптимізації. Особливу увагу приділено питанням визначення та мінімізації похибки обчислень.

Для студентів бакалаврату з напряму «Електроніка».

УДК 519.6 (075.8)

Зміст

Вступ

Курс «Обчислювальна математика» – складова частина обов'язкових дисциплін у підготовці бакалаврів за напрямом «Електроніка». Він ґрунтується на знаннях, отриманих у результаті вивчення математики, персональних комп'ютерів, алгоритмічних мов і основ програмування, які забезпечують розуміння основ побудови методів обчислювальної математики та особливостей реалізації їх на комп'ютері. Підґрунтям для подальшого вивчення курсу є засвоєння методів моделювання та проектування в електроніці.

Мета курсу – оволодіння чисельними методами розв'язання задач аналізу і проектування в електроніці, отримання навичок реалізації чисельних алгоритмів на комп'ютері та інших обчислювальних системах.

Застосування методів обчислювальної математики до розв'язання задач із електроніки має свою специфіку. Так, наприклад, диференціальні рівняння, до яких зводяться задачі електроніки, виявляються жорсткими, а матриці коефіцієнтів систем лінійних алгебраїчних рівнянь (СЛАР) – погано обумовленими. Основну увагу в курсі приділяють тим методам обчислювальної математики, які пристосовані до задач електроніки.

Жоден з існуючих методів обчислювальної математики не задовольняє всіх висунутих до них вимог. Тому реальні обчислювальні алгоритми, як правило, будують комбінуванням різних методів. Тому сучасному спеціалісту з електроніки необхідно володіти широким колом обчислювальних методів та глибоко розуміти їх суть, обмеження та особливості застосування.

Автори не ставили за мету формулювання та доведення відомих теорем обчислювальної математики, що стали класичними і добре висвітлені в цитованій літературі. Основну увагу приділено роз'ясненню суті методів, що розглядаються, та особливостям їх застосування.

1. Похибки обчислень

1.1. Поняття обчислювального експерименту

У зв’язку зі швидким розвитком обчислювальної техніки обчислювальний експеримент набув широкого застосування для проведення наукових досліджень та інженерного проектування. Він ґрунтується на побудові та аналізі за допомогою комп’ютера математичних моделей досліджуваного об’єкта чи явища.

Розглянемо схему обчислювального експерименту (рис. 1.1).

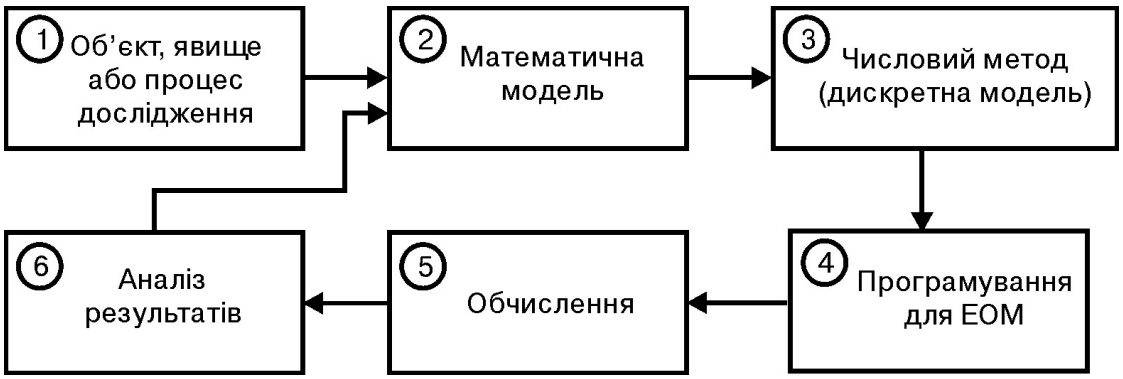


Рис. 1.1. Схема обчислювального експерименту

Нехай треба дослідити певний об’єкт, явище або процес (1). Тоді спочатку формулюють основні закони та взаємозв’язки, що описують цей об’єкт. На їх основі розробляють математичну модель (2), що становить собою, як правило, запис цих законів у вигляді системи рівнянь (алгебраїчних, диференціальних, інтегральних та ін.). Після того, як задачу сформульовано, її потрібно розв’язати. Тільки в досить простих випадках удається отримати розв’язок у явному вигляді. У більшості випадків виникає необхідність використання того чи іншого наближеного методу: обчислювального методу або дискретної моделі (3). На основі отриманої

дискретної моделі будують обчислювальний алгоритм, результатом реалізації якого є число або таблиця чисел.

Для реалізації обчислювального методу потрібно розробити програму для комп'ютера (4). Після розроблення та відладки програми наступає етап проведення обчислень (5). Отримані результати детально аналізують (6) з точки зору їх відповідності досліджуваному явищу і, за необхідності, вносять зміни в математичну модель або обирають інший обчислювальний метод. Цей цикл повторюють доти, доки не буде отримано результати з заданою точністю.

Обчислювальна математика забезпечує лише один з етапів обчислювального експерименту, а саме етап вибору (побудови) обчислювального методу, від якого значною мірою залежить ефективність усього експерименту.

1.2. Похибки обчислень

Сучасні обчислювальні системи оперують числами, записаними в одній із наведених нижче форм [1].

Перша форма запису — з фіксованою крапкою. Така форма відповідає позиційній системі з основою r і будь-який запис

$$a = \pm a_n a_{n-1} \ldots a_0 , a_{-1} a_{-2} \ldots a_{-m}$$

означає, що

$$a = \pm \sum_{k=-m}^{n} a_k r^k ,$$

де $0 \le a_k < r$ — цифра k-го розряду.

Найчастіше використовують запис чисел у формі з плаваючою крапкою, тобто у вигляді

$$a = \pm \mathrm{M} r^{p}, \qquad (1.1)$$

(як правило, $r = 2$).

Число M записують у формі числа з фіксованою крапкою і називають мантисою числа a, причому

$$r^{-1} \le \left|\mathrm{M}\right| < 1.$$

Таке подання чисел з плаваючою крапкою називається нормалізованим. Число p називають порядком числа a. Запис (1.1) означає, що

$$a = \pm r^{p} \sum_{k=-t}^{-1} a_k r^{k},$$

де t – кількість значущих цифр мантиси.

В обчислювальній системі для запису кожного числа виділяють фіксоване число розрядів (бітів). Число відведених розрядів залежить як від типу обчислювальної системи, так і від мови програмування. Діапазон чисел, з якими оперує обчислювальна система, внаслідок скінченності розрядної сітки для запису порядку обмежений:

$$\left|a\right| \in \left[\mathrm{M}_0, \mathrm{M}_\infty\right],$$

де M_0 — машинний нуль, M_∞ — машинна нескінченність. Через скінченність розрядної сітки для запису значення мантиси в обчислювальній системі можна подати точно не всі числа з діапазону $\mathrm{M}_0 \ldots \mathrm{M}_\infty$, а лише скінченну їх множину. Це пов'язано не тільки з існуванням ірраціональних чисел, але й з переведенням числа з однієї системи числення в іншу. Наприклад, число 0,7 не можна подати у вигляді скінченної суми за степенями $\frac{1}{2}$, тому 0,7 в обчислювальній системі подається як 0,6999…

Число a, яке не можна записати в обчислювальній системі точно, округлюється, тобто заміняється близьким йому числом $\tilde{a}$, яке подається точно в обчислювальній системі. Точність подання чисел із плаваючою крапкою в обчислювальній системі характеризується відносною похибкою наближеного значення [2, с. 22-25]

$$\varepsilon = \left| \frac{\tilde{a} - a}{\tilde{a}} \right|.$$

За найпростішого способу округлення – відкидання всіх розрядів мантиси, які виходять за межі розрядної сітки, відносну похибку округлення можна оцінити як [3]:

$$\varepsilon = \left| \frac{\tilde{a} - a}{\tilde{a}} \right| \le 2^{1-t}.$$

Оскільки у нормалізованій формі запису мантиси старший біт завжди дорівнює одиниці, то цей біт не запам’ятовують, хоча його присутність завжди мають на увазі під час обчислень. Це звільняє місце для збереження додаткового біта. Тому в більшості обчислювальних систем мантиса обмежується як

$$1 \le |\mathrm{M}| < r,$$

але старший біт не зберігається. Оскільки у цьому випадку кількість реально збережених бітів збільшується на одиницю, то реальна відносна машинна похибка оцінюється як:

$$\varepsilon = \left| \frac{\tilde{a} - a}{\tilde{a}} \right| \le 2^{-t}.$$

1.3. Обчислювальна похибка визначення значення функції

Нехай необхідно обчислити значення функції $y = f(x_1, x_2, \ldots, x_n)$ від n змінних. Кожен з аргументів функції x_i, $i = 1,...,n$ подано в обчислювальній системі як число $\tilde{x}_i$ з абсолютною похибкою округлення $|\Delta x_i| \geq |\tilde{x}_i - x_i|$, $i = \overline{1,n}$.

Позначимо $\tilde{y} = f(\tilde{x}_i, \tilde{x}_2, ..., \tilde{x}_n)$ як обчислене значення функції. Визначимо, наскільки обчислене значення $\tilde{y}$ відрізняється від точного y.

Якщо відомі всі похибки округлення Δx_i, то відома деяка область G, якій належать точні значення аргументів функції x_i, $i = 1,..., n$. Введемо означення:

якщо $\tilde{y}$ – наближене значення функції $y = f(x_1, x_2, ..., x_n)$, то граничною абсолютною похибкою Δy називають значення

$$\Delta y = \sup_{(x_1, x_2, \ldots, x_n) \in G} |y - \tilde{y}| = \sup_{(x_1, x_2, \ldots, x_n) \in G} \left| f(x_1, x_2, \ldots, x_n) - f(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n) \right|,$$

де $\sup_{x \in G}(g(x))$ - найбільше значення функції $g(x)$ в області G.

Граничною відносною похибкою називають значення

$$\delta y = \frac{\Delta y}{|\tilde{y}|}.$$

Якщо $y = f(x_1, x_2, ..., x_n)$ – неперервно диференційована функція своїх аргументів, а її частинні похідні обмежені

$$F_i = \sup_{(x_1, x_2, ..., x_n) \in G} \left| \frac{\partial f(\tilde{x}_1 + \theta(x_1 - \tilde{x}_1), ..., \tilde{x}_n + \theta(x_n - \tilde{x}_n))}{\partial x_i} \right|, \qquad (1.2)$$

то можна показати, що гранична абсолютна похибка обчислення функції оцінюється так:

$$\Delta y = \sum_{i=1}^{n} F_i \Delta x_i \,. \qquad (1.3)$$

Слід відзначити, що обчислення найбільших значень частинних похідних у виразі (1.3) може бути досить складною задачею, тому на практиці використовують більш просту оцінку:

$$\Delta y_0 = \sum_{i=1}^{n} \left| \frac{\partial f(\tilde{x}_1,\ \tilde{x}_2, ...,\ \tilde{x}_n)}{\partial x_i} \right| \Delta x_i \,,$$

яка називається *лінійною оцінкою похибки*.

З рівнянь (1.2) і (1.3) випливає, що вплив похибок округлення аргументів на точність обчислення функції залежить від частинних похідних функції за цими аргументами. Чим більше за абсолютною величиною значення частинної похідної функції за даним аргументом, тим більший внесок в обчислювальну похибку функції дає похибка округлення аргументу.

Контрольні завдання

1. Оцінити похибки обчислення функції $y = f(x_1,\ x_2)$ в точці $x_1 = 3,$ $x_2 = 5$, якщо всі обчислення виконують зі звичайним дійсним типом (Real, float та ін.), дійсним типом подвійної точності (double) та дійсним типом підвищеної точності (extended).

Варіант	$f(x_1, x_2)$	Варіант	$f(x_1, x_2)$
1	$e^{x_1^2} x_2^3$	9	$x_1^2 \left(1 + \ln x_2\right)$
2	$10^{x_1} \operatorname{tg}\left(x_2^5\right)$	10	$e^{x_1 + x_2} \sqrt{x_2^5}$
3	$x_1^{x_2} \cos\left(x_2^5\right)$	11	$e^{\sin(x_1)} \operatorname{tg}\left(x_2^2\right)$
4	$x_1^{\cos(x_2)} x_2^5$	12	$\dfrac{e^{x_1}}{\sin\left(x_2\right)}$

5	$x_1^{\log_2\left(5x_2^2\right)}\sin\left(x_2^3\right)$	13	$\frac{x_1^5}{1+\sin\left(x_2\right)}$
6	$e^{x_1x_2}\cos(x_1x_2)$	14	$e^{\frac{x_1}{1+\cos(x_2)}}$
7	$\mathrm{tg}\left(x_1x_2\right)x_2^{3x_1}$	15	$\ln\left(1+\cos\left(x_1\right)\right)x_2^3$
8	$e^{\cos^2(x_1)}x_2^5$		

2. Скільки бітів потрібно відвести для збереження мантис усіх змінних для того, щоб похибка обчислення функції $y = f(x_1, x_2)$ в точці $x_1 = 1,$ $x_2 = 2$ була меншою ніж 0,1 %.

Варіант	$f(x_1, x_2)$	Варіант	$f(x_1, x_2)$
16	$x_1^2\left(1+\ln x_2\right)$	24	$x_1^{x_2^2}\ln\left(x_2^3\right)$
17	$x_1^2\left(1+\sqrt{x_2}\right)$	25	$\cos\left(\sqrt{x_1^3}\right)e^{x_2}$
18	$\sin\left(x_1\right)\left(1+\cos\left(x_2\right)\right)$	26	$\mathrm{tg}\left(x_1^3\right)e^{2^{x_2}}$
19	$\mathrm{tg}\left(x_1\right)e^{x_2^2}$	27	$\frac{\mathrm{tg}\left(x_1^3\right)}{1+\cos(x_2)}$
20	$2^{x_1}\cos\left(x_2\right)$	28	$e^{x_1}\sqrt{5+x_2^5}$
21	$x_1^3e^{\sqrt{x_2}}$	29	$\ln\left(\cos\left(x_1\right)\right)\left(5+x_2^5\right)$
22	$\sqrt{x_1^3+3}\cos\left(e^{x_2}\right)$	30	$x_1^{\cos(x_2)}x_2^3$
23	$\ln\left(\cos\left(x_1\right)\right)x_2^3$		

2. Чисельні методи лінійної алгебри

2.1. Основні положення лінійної алгебри

2.1.1. Основні визначення

У прикладних задачах найпоширенішими є два типи задач лінійної алгебри: розв'язання систем лінійних алгебраїчних рівнянь (СЛАР) та обчислення власних чисел та векторів матриць.

Система лінійних алгебраїчних рівнянь у загальному випадку має вигляд:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1; \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2; \\ \qquad\qquad \ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n, \end{cases} \tag{2.1}$$

де a_{ij} – коефіцієнти системи; b_i – вільні члени системи; x_j – невідомі, які треба визначити.

Вводячи матрицю коефіцієнтів системи

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ & \ldots & & \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix},$$

та вектор-стовпець вільних членів $\mathbf{B} = \left[b_1, b_2, \ldots, b_n\right]^{\mathrm{T}}$, а також вектор-стовпець невідомих $\mathbf{X} = \left[x_1, x_2, \ldots, x_n\right]^{\mathrm{T}}$, систему (2.1) можна подати у матричній формі:

$$\mathbf{AX} = \mathbf{B}. \tag{2.2}$$

Якщо всі вільні члени $b_i = 0,\ i = \overline{1,n}$, то СЛАР називається однорідною. Інакше, СЛАР називають неоднорідною. Із курсу лінійної алгебри відомо, що неоднорідна СЛАР має єдиний розв'язок у випадку, коли $\det(\mathbf{A}) \neq 0$ [4, с. 269], де $\det(\mathbf{A})$ – визначник матриці **A**.

У багатьох наукових та інженерних задачах виникає потреба у знаходженні власних чисел і відповідних їм власних векторів. Наприклад, в аналізі динамічних систем власні числа визначають частоти коливань, а власні вектори характеризують їх форму.

Власні числа $\lambda_i, i = \overline{1,n}$ квадратної матриці **A** розмірності n є дійсними або комплексними числами, що задовольняють умові [5]:

$$(\mathbf{A} - \lambda_i \mathbf{E})\mathbf{\Psi}_i = 0, \qquad (2.3)$$

де **E** – одинична матриця; $\mathbf{\Psi}_i$ – власний вектор матриці **A**, що відповідає деякому власному числу λ_i.

Оскільки СЛАР (2.3) відносно елементів власного вектора є однорідною, то її ненульовий розв'язок можливий тільки за умови, що визначник матриці цієї системи дорівнює нулю:

$$\det(\mathbf{A} - \lambda \mathbf{E}) = 0.$$

З останньої умови випливає, що будь-яка матриця $n \times n$ має n власних чисел з урахуванням їх кратності. Множина власних чисел матриці називається її спектром.

2.1.2. Норми векторів та матриць

Часто виникає необхідність порівнювати вектори та матриці певною мірою близькості. З цією метою вводять норми векторів та матриць.

Нормою вектора **X** є дійсна функція $\|\mathbf{X}\|$, що задовольняє таким умовам:

1) Невід'ємності: $\|\mathbf{X}\| \geq 0$, причому $\|\mathbf{X}\| = 0$ тоді і тільки тоді, коли **X**=0.

2) Абсолютної однорідності: $\|a\mathbf{X}\| = |a|\|\mathbf{X}\|$, де a — скалярна величина.

3) Нерівність трикутника: $\|\mathbf{X} + \mathbf{Y}\| \leq \|\mathbf{X}\| + \|\mathbf{Y}\|$.

Найчастіше застосовуються такі норми:

1) $\|\mathbf{X}\|_\infty = \max\limits_{1 \leq i \leq n} \|x_i\|$ — (l_∞-норма, sup-норма, або норма Чебишова).

2) $\|\mathbf{X}\|_1 = \sum\limits_{i=1}^{n} |x_i|$ — (l_1-норма, або норма найменших абсолютних різниць).

3) $\|\mathbf{X}\|_2 = \sqrt{\sum\limits_{i=1}^{n} x_i^2}$ — (l_2 — норма, Евклідова норма, або норма найменших квадратів).

Всі вони є частковими випадками загального класу векторних l_q-норм:

$$\|\mathbf{X}\|_q = \left(\sum_{i=1}^{n} |x_i|^q \right)^{\frac{1}{q}}.$$

Нормою матриці **A** називається функція $\|\mathbf{A}\|$, що задовольняє умовам:

1) Невід'ємності: $\|\mathbf{A}\| > 0$, причому $\|\mathbf{A}\| = 0,$ тільки якщо **A**=0.

2) Абсолютної однорідності: $\|a\mathbf{A}\| = |a|\|\mathbf{A}\|$.

3) Нерівності трикутника: $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.

4) Мультиплікативності: $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$.

Норми матриць можуть уводитися різними способами. За аналогією до векторних норм можна ввести l_p-норму

$$\|\mathbf{A}\|_p = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^p \right)^{\frac{1}{p}}.$$

Матричну норму l_2 (аналог Евклідової) часто називають нормою Фробеніуса (а інколи так і називають Евклідовою).

Часто треба вміти вимірювати матриці згідно з їх роллю операторів. Добуток вектора **X** на матрицю **A** дає новий вектор **AX**, норма якого може суттєво відрізнятися від норми вектора **X**. Для оцінки образу **X** після перетворення оператором **A** вводять індуковану (операторну) норму матриці

$$\|\mathbf{A}\| = \max_{\mathbf{X} \neq 0} \left(\frac{\|\mathbf{AX}\|}{\|\mathbf{X}\|} \right), \tag{2.4}$$

де $\|\mathbf{AX}\|$ та $\|\mathbf{X}\|$ — векторні норми.

Для введених раніше векторних l_p-норм індукованими нормами матриць будуть такі [5, с. 245-249]:

$$\|\mathbf{A}\|_\infty = \max_i \left(\|a_{i\cdot}\|_1 \right) = \max_i \sum_{k=1}^{n} |a_{ik}|;$$

$$\|\mathbf{A}\|_1 = \max_j \left(\|a_{\cdot j}\|_1 \right) = \max_j \sum_{k=1}^{n} |a_{kj}|;$$

$$\|\mathbf{A}\|_2 = \sqrt{\mu},$$

де $a_{i\cdot}$ — i-й рядок матриці **A**, $a_{\cdot j}$ — j-й стовпчик матриці **A**; μ — найбільше власне число матриці $\mathbf{A}^\dagger\mathbf{A}$, де $\mathbf{A}^\dagger$ — матриця, ермітово спряжена з матрицею **A**, тобто матриця, елементи якої є комплексно спряженими з елементами матриці **A**, та транспонована ($\mathbf{A}^\dagger = \left(\overline{\mathbf{A}}\right)^{\mathrm{T}}$).

Якщо матриця **A** ермітова, тобто така, що $\mathbf{A}^{\dagger} = \mathbf{A}$ (у випадку дійсної матриці ермітова матриця є симетричною, тобто такою, що $\mathbf{A}^{\mathrm{T}}$=**A**), то $\|\mathbf{A}\|_2 = \max_i\left(|\lambda_i|\right)$, де λ_i — власне число матриці **A**.

Можна показати [6, с. 59-65], що індукована норма оберненої матриці дорівнює:

$$\left\|\mathbf{A}^{-1}\right\| = \frac{1}{\min\limits_{\mathrm{x}} \frac{\|\mathbf{AX}\|}{\|\mathbf{X}\|}}.$$

На сьогоднішній день розроблено багато методів розв'язування СЛАР. Особливе місце посідають методи факторизації матриці системи, тобто подання її у вигляді добутку матриць спеціального виду (трикутних, діагональних, блоково-діагональних). Найпопулярнішими є метод LU-факторизації, метод QR-факторизації та метод Холеського. Такий вибір зумовлений тим, що ці методи добре адаптовані до використання обчислювальної техніки для розв'язання задач електроніки [7].

2.1.3. Ортогональність векторів та унітарність матриць. Основні унітарні перетворення

Важливе місце в лінійній алгебрі посідають ортогональні вектори та унітарні матриці, а також відповідні унітарні перетворення.

Два вектора **X** і **Y** називаються ортогональними, якщо їх скалярний добуток дорівнює нулю, тобто

$$\mathbf{X}^{\dagger}\mathbf{Y} = 0. \qquad (2.5)$$

Вектор **X** називається нормованим, якщо

$$\mathbf{X}^{\dagger}\mathbf{X} = 1. \qquad (2.6)$$

Якщо вектори **X** та **Y** задовольняють умові (2.5) та умовам типу (2.6), то такі вектори називаються ортонормованими.

Квадратна матриця **Q** називається унітарною, якщо її стовпчики (рядки) попарно ортонормовані. Таким чином, матриця **Q** унітарна, якщо

$$\mathbf{Q}^{\dagger}\mathbf{Q} = \mathbf{Q}\mathbf{Q}^{\dagger} = \mathbf{E}, \qquad (2.7)$$

де **E** — одинична матриця.

Унітарну матрицю, всі елементи якої є дійсними, називають ортогональною.

Унітарні матриці мають такі властивості:

1) ермітово спряжена до унітарної матриці є теж унітарною;
2) det **Q** = ± 1;
3) $\mathbf{Q}^{-1} = \mathbf{Q}^{\dagger}$;
4) $\|\mathbf{QX}\|_2 = \|\mathbf{X}\|_2$;
5) $\|\mathbf{QA}\|_2 = \|\mathbf{AQ}\|_2 = \|\mathbf{A}\|_2$;
6) $\|\mathbf{Q}\|_2 = 1$;
7) добуток двох унітарних матриць теж є унітарною матрицею;
8) якщо матриця $\mathbf{Q}_\mathrm{N}$ унітарна, то матриця $\mathbf{Q}_{N+1} = \begin{bmatrix} 1 & 0 \\ 0 & \mathbf{Q}_N \end{bmatrix}$ теж унітарна;
9) модуль усіх власних чисел унітарної матриці дорівнює одиниці, тобто всі власні числа унітарної матриці знаходяться на одиничному колі комплексної площини;
10) для дійсної ортогональної матриці сума квадратів елементів кожного стовпчика (рядка) дорівнює 1.

Прикладами унітарних матриць є матриці відбивання та обертання.

Матриця відбивання (перетворення Хаусхолдера) має вигляд:

$$\mathbf{Q} = \mathbf{E} - (1+\omega)\mathbf{W}\mathbf{W}^{\dagger}, \qquad (2.8)$$

де $\mathbf{W}$ — нормований вектор-стовпчик, тобто такий, що $\mathbf{W}^{\dagger}\mathbf{W} = 1$, ω — скалярне комплексне число, таке, що $|\omega| = 1$, або $\bar{\omega}\omega = 1$.

Покажемо, що матриця відбивання (2.8) є унітарною.

$$\begin{aligned}
&\mathbf{Q}\mathbf{Q}^{\dagger} = \left(\mathbf{E} - (1+\omega)\mathbf{W}\mathbf{W}^{\dagger}\right)\left(\mathbf{E} - (1+\omega)\mathbf{W}\mathbf{W}^{\dagger}\right)^{\dagger} = \\
&= \left(\mathbf{E} - (1+\omega)\mathbf{W}\mathbf{W}^{\dagger}\right)\left(\mathbf{E} - (1+\bar{\omega})\left(\mathbf{W}\mathbf{W}^{\dagger}\right)^{\dagger}\right) = \\
&= \mathbf{E} - (1+\bar{\omega})\mathbf{W}\mathbf{W}^{\dagger} - (1+\omega)\mathbf{W}\mathbf{W}^{\dagger} + (1+\bar{\omega})(1+\omega)\mathbf{W}\left(\mathbf{W}^{\dagger}\mathbf{W}\right)\mathbf{W}^{\dagger} \\
&= \mathbf{E} - (2+\bar{\omega}+\omega)\mathbf{W}\mathbf{W}^{\dagger} + (1+\bar{\omega}+\omega+\bar{\omega}\omega)\mathbf{W}\mathbf{W}^{\dagger} = 1.
\end{aligned}$$

Тут враховано, що $\left(\mathbf{W}\mathbf{W}^{\dagger}\right)^{\dagger} = \left(\mathbf{W}^{\dagger}\right)^{\dagger}\mathbf{W}^{\dagger} = \mathbf{W}\mathbf{W}^{\dagger}$.

Перетворення Хаусхолдера називають перетворенням відбивання, тому що воно виконує перетворення довільного вектора за правилом дзеркального відбивання від заданої площини S. Дійсно, нехай $\mathbf{W}$ — нормований вектор, ортогональний до S. Нехай $\mathbf{Z}$ довільний вектор. Подамо його як суму

$$\mathbf{Z} = \mathbf{X} + \mathbf{Y},$$

де $\mathbf{X}$ — вектор, ортогональний до $\mathbf{W}$, тобто $\mathbf{W}^{\dagger}\mathbf{X} = 0$, а $\mathbf{Y}$ — вектор, паралельний до $\mathbf{W}$, тобто $\mathbf{Y} = \alpha\mathbf{W}$, де $\alpha \neq 0$ — деяка скалярна величина, або $\mathbf{W}^{\dagger}\mathbf{Y} = \alpha$.

Тоді

$$\begin{aligned}
&\tilde{\mathbf{Z}} = \mathbf{Q}\mathbf{Z} = \left(\mathbf{E} - (1+\omega)\mathbf{W}\mathbf{W}^{\dagger}\right)\mathbf{Z} = \mathbf{Z} - (1+\omega)\mathbf{W}\mathbf{W}^{\dagger}\mathbf{Z} = \\
&= \mathbf{Z} - (1+\omega)\mathbf{W}\mathbf{W}^{\dagger}\mathbf{X} - (1+\omega)\mathbf{W}\mathbf{W}^{\dagger}\mathbf{Y} = \mathbf{Z} - (1+\omega)\alpha\mathbf{W} = \\
&= \mathbf{Z} - (1+\omega)\mathbf{Y} = \mathbf{X} - \omega\mathbf{Y}.
\end{aligned}$$

В евклідовому просторі $\omega = 1$, а тому $\tilde{\mathbf{Z}} = \mathbf{X} - \mathbf{Y}$. Тобто, вектор $\mathbf{Z}$ перетворюється у вектор, відбитий від площини S (рис. 2.1).

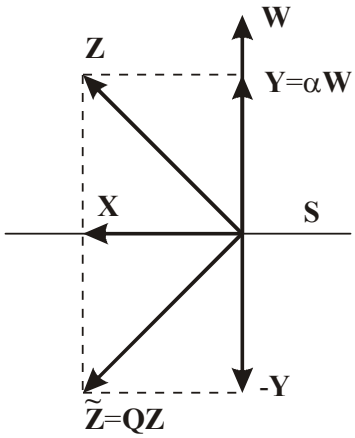


Рис. 2.1. Геометрична інтерпретація перетворення Хаусхолдера

Вектор **W** завжди можна підібрати так, щоб матриця **Q** переводила довільний вектор **Z** у вектор заданого напряму, наприклад, у паралельний деякому вектору **Y** одиничної довжини. Для цього слід взяти

$$\mathbf{W} = \frac{\mathbf{Z} - \alpha\mathbf{Y}}{\|\mathbf{Z} - \alpha\mathbf{Y}\|_2}, \tag{2.9}$$

де $\alpha = \pm\|\mathbf{Z}\|_2 = \pm\sqrt{\mathbf{Z}^\dagger\mathbf{Z}}$, $\|\mathbf{Z} - \alpha\mathbf{Y}\|_2 = \sqrt{2\alpha^2 - \alpha\mathbf{Y}^\dagger\mathbf{Z} - \alpha\mathbf{Z}^\dagger\mathbf{Y}}$, а $\omega = \dfrac{\mathbf{Z}^\dagger\mathbf{W}}{\mathbf{W}^\dagger\mathbf{Z}}$.

Тоді

$$\begin{aligned}
\mathbf{QZ} &= (\mathbf{E} - (1+\omega)\mathbf{W}\mathbf{W}^\dagger)\mathbf{Z} = \mathbf{Z} - \mathbf{W}\mathbf{W}^\dagger\mathbf{Z} - \frac{\mathbf{Z}^\dagger\mathbf{W}}{\mathbf{W}^\dagger\mathbf{Z}}\mathbf{W}\mathbf{W}^\dagger\mathbf{Z} = \\
&= \mathbf{Z} - \mathbf{W}\mathbf{W}^\dagger\mathbf{Z} - \mathbf{Z}^\dagger\mathbf{W}\mathbf{W} = \mathbf{Z} - \left(\mathbf{W}^\dagger\mathbf{Z} + \mathbf{Z}^\dagger\mathbf{W}\right)\mathbf{W} = \\
&= \mathbf{Z} - \left(\frac{(\mathbf{Z} - \alpha\mathbf{Y})^\dagger\mathbf{Z} + \mathbf{Z}^\dagger(\mathbf{Z} - \alpha\mathbf{Y})}{\|\mathbf{Z} - \alpha\mathbf{Y}\|_2}\right)\mathbf{W} = \\
&= \mathbf{Z} - \frac{2\alpha^2 - \alpha\mathbf{Y}^\dagger\mathbf{Z} - \alpha\mathbf{Z}^\dagger\mathbf{Y}}{\|\mathbf{Z} - \alpha\mathbf{Y}\|_2}\mathbf{W} = \mathbf{Z} - \frac{\|\mathbf{Z} - \alpha\mathbf{Y}\|_2^2}{\|\mathbf{Z} - \alpha\mathbf{Y}\|_2}\mathbf{W} = \\
&= \mathbf{Z} - \mathbf{Z} + \alpha\mathbf{Y} = \alpha\mathbf{Y}.
\end{aligned}$$

Формулу (2.9) легко поширити на випадок довільного вектора $\tilde{\mathbf{Y}}$, якщо покласти $\mathbf{Y} = \tilde{\mathbf{Y}}/\|\tilde{\mathbf{Y}}\|_2$.

Матрицею обертання (матрицею обертання Якобі, матрицею Гівенса) у двовимірному просторі називається матриця:

$$\mathbf{J}(\varphi)=\begin{bmatrix}\cos\varphi & -\sin\varphi\\ \sin\varphi & \cos\varphi\end{bmatrix},$$

або у еквівалентній формі для комплексної матриці

$$\mathbf{J}=\begin{bmatrix}c & -s\\ \overline{s} & \overline{c}\end{bmatrix}, \qquad (2.10)$$

де $c=\dfrac{a}{\sqrt{|a|^2+|b|^2}}$, $s=\dfrac{b}{\sqrt{|a|^2+|b|^2}}$, $\qquad |a|+|b|\neq 0$.

Легко переконатися, що матриця (2.10) є унітарною.

Матриця обертання (2.10) перетворює будь-який вектор $\mathbf{X}\neq 0$ у вектор, повернутий відносно $\mathbf{X}$ на кут φ. У багатовимірному випадку можна користуватися елементарними матрицями обертання виду

$$\mathbf{J}_{i,j}=\begin{array}{c}\\ \\ \\ i\\ \\ \\ \\ \\ j\\ \\ \\ \end{array}\begin{bmatrix}1 & 0 & \cdots & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0\\ 0 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 0\\ \vdots & & \ddots & \vdots & & & & \iddots & & \\ 0 & \cdots & 0 & c & 0 & \cdots & 0 & -s & 0 & 0\\ & & & 0 & 1 & \iddots & 0 & & & \\ & & & \vdots & \iddots & \ddots & & & & \\ & & & 0 & 0 & & 1 & & & \\ 0 & \cdots & 0 & \overline{s} & 0 & \cdots & 0 & \overline{c} & 0 & 0\\ & & & 0 & & & & & \ddots & \\ 0 & & & 0 & & & & & & 1\end{bmatrix}, \qquad (2.11)$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i \qquad\qquad\qquad\qquad j$$

які є одиничними матрицями, у яких на місці перетину i-рядка та i-стовпчика розміщений елемент c, j-рядка та j-стовпчика елемент $\overline{c}$, i-рядка та j-стовпчика елемент -s і j-рядка та i-стовпчика елемент $\overline{s}$. Елементи c та

s визначаються як і у (2.10). Очевидно, що в результаті множення матриці $\mathbf{A}$ зліва на матрицю $\mathbf{J}_{i,j}$ змінюються тільки i-й та j-й рядки матриці $\mathbf{A}$, а саме для матриці $\tilde{\mathbf{A}} = \mathbf{J}_{i,j}\mathbf{A}$ маємо

$$\begin{array}{l} \tilde{a}_{ik} = ca_{ik} - sa_{jk}; \\ \tilde{a}_{jk} = \overline{s}a_{ik} + \overline{c}a_{jk}; \end{array} \quad k = \overline{1,n}. \tag{2.12}$$

Якщо $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{J}_{i,j}$, то змінюються тільки i-й та j-й стовпчики за формулами

$$\begin{array}{l} \tilde{a}_{ki} = ca_{ki} + \overline{s}a_{kj}; \\ \tilde{a}_{kj} = -sa_{ki} + \overline{c}a_{kj}; \end{array} \quad k = \overline{1,n}. \tag{2.13}$$

Ясно, що коли хоч один з елементів a_{im} чи a_{jm} не дорівнює нулю, то можна підібрати в (2.12) c і s так, щоб для матриці $\tilde{\mathbf{A}} = \mathbf{J}_{ij}\mathbf{A}$ елемент $\tilde{a}_{im}$ дорівнював нулю. Для цього слід взяти

$$c = \frac{a_{jm}}{\sqrt{\left|a_{im}\right|^2 + \left|a_{jm}\right|^2}}; s = \frac{a_{im}}{\sqrt{\left|a_{im}\right|^2 + \left|a_{jm}\right|^2}}.$$

Отримання в матриці $\tilde{\mathbf{A}} = \mathbf{J}_{ij}\mathbf{A}$ на місці елемента $\tilde{a}_{im}$ нуля є корисною особливістю матриць обертання. Будь-яка невироджена матриця шляхом послідовних множень зліва на елементарні матриці обертання може бути перетворена на верхню трикутну матрицю $\mathbf{R}$:

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & r_{nn} \end{bmatrix}.$$

Аналогічно можна показати, що з використання послідовності будь-яких унітарних перетворень невироджену матрицю можна привести до верхньотрикутного виду

$$\mathbf{Q}_1\mathbf{Q}_2...\mathbf{Q}_m\mathbf{A} = \mathbf{Q}\mathbf{A} = \mathbf{R},$$

звідки

$$\mathbf{A} = \mathbf{Q}^{-1}\mathbf{R} = \mathbf{Q}^{\dagger}\mathbf{R}\,.$$

Оскільки матриця $\mathbf{Q}^{\dagger}$ унітарна, то, відповідно, будь-яку невироджену матрицю можна подати як добуток унітарної та верхньої трикутної матриці [9, с. 178-181]:

$$\mathbf{A} = \mathbf{Q}\mathbf{R}. \qquad (2.14)$$

2.2. Методи факторизації матриць розв'язання СЛАР

2.2.1. Метод LU-факторизації

У цьому методі матриця коефіцієнтів **A** подається у вигляді добутку матриць **L** та **U** [5, с. 40-43]:

$$\mathbf{A} = \mathbf{L}\mathbf{U}, \qquad (2.15)$$

де **L** – нижня трикутна матриця; а **U** – верхня трикутна матриця, усі діагональні елементи якої дорівнюють 1:

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ & \dots & & \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{bmatrix}, \qquad \mathbf{U} = \begin{bmatrix} 1 & u_{12} & \dots & u_{1n} \\ 0 & 1 & \dots & u_{2n} \\ & \dots & & \\ 0 & 0 & \dots & 1 \end{bmatrix}. \qquad (2.16)$$

Метод **LU**-факторизації складається з двох етапів:

1) етапу факторизації матриці **A**;

2) етапу отримання розв'язку **X**.

Вектор **B** у процесі факторизації не змінюється.

Із виразів (2.15) та (2.16) випливає, що для $i > j$ елементи матриці **A** можна виразити через елементи матриць **L** та **U**:

$$\begin{aligned} a_{ij} &= \left(\sum_{k=1}^{j-1} l_{ik} u_{kj} \right) + l_{ij}; \\ a_{ji} &= \left(\sum_{k=1}^{j-1} l_{jk} u_{ki} \right) + l_{jj} u_{ji}, \qquad i = \overline{1,n}, \qquad j = \overline{1,i}. \end{aligned} \tag{2.17}$$

Із рівнянь (2.17) випливає, що факторизація матриці **A** виконується за n стадій. На кожній j-й стадії послідовно обчислюють елементи l_{ij} чергового j-го стовпчика матриці **L** за формулою

$$l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj}, \quad i = \overline{j,n}, \tag{2.18}$$

та елементи u_{ji} чергового j-го рядка матриці **U** за формулою

$$u_{ji} = \frac{a_{ji} - \sum_{k=1}^{j-1} l_{jk} u_{ki}}{l_{jj}}, \; i = \overline{j+1,n}. \tag{2.19}$$

Із формул (2.18), (2.19) випливає, що на першій стадії елементи $l_{i1} = a_{i1}, i = \overline{1,n}$, а $u_{1j} = \dfrac{a_{1j}}{l_{11}}, \; j = \overline{2,n}$.

Слід відзначити, що суми у (2.18), (2.19) для шуканого елемента мають сенс скалярного добутку відповідних частин рядка та стовпчика, в яких він знаходиться, рис. 2.1.

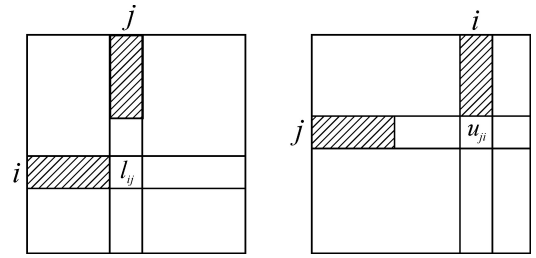


Рис. 2.1. Схема знаходження сум для обчислення елементів матриць **L** і **U**.

На практиці елементи матриць **L** та **U** в процесі факторизації розміщуються на місцях елементів матриці **A**, причому елементи $u_{jj} = 1,\ j = \overline{1,n}$ не запам'ятовуються. Таким чином, факторизована форма матриці **A** має вигляд

$$\mathbf{A} = \begin{bmatrix} l_{11} & u_{12} & \cdots & u_{1n} \\ l_{21} & l_{22} & \cdots & u_{2n} \\ & & \cdots & \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix}.$$

Після факторизації матриці **A** система буде мати вигляд

$$\mathbf{LUX} = \mathbf{B}. \tag{2.20}$$

Позначимо **Y** = **UX**, тоді із (2.20) маємо рівняння

$$\mathbf{LY} = \mathbf{B}. \tag{2.21}$$

Оскільки **L** – нижня трикутна матриця, то розв'язок системи (2.21) отримуємо прямою підстановкою:

$$y_i = \frac{b_i - \sum_{j=1}^{i-1} l_{ij} y_j}{l_{ii}},\quad i = \overline{1,n}.$$

Після цього залишається розв'язати ще систему з верхньою трикутною матрицею **U**, всі діагональні елементи якої дорівнюють 1:

$$\mathbf{UX} = \mathbf{Y}. \tag{2.22}$$

Розв'язавши систему (2.22), остаточно отримаємо шуканий вектор **X**:

$$x_i = y_i - \sum_{j=i+1}^{n} u_{ij} x_j,\quad i = \overline{n,1}.$$

Приклад:

$$\begin{bmatrix} 2 & 8 & 6 \\ 3 & 10 & 11 \\ 1 & 6 & 5 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 28 \\ 40 \\ 20 \end{bmatrix}.$$

$$j = 1:$$
$$u_{12} = \frac{8}{2} = 4; \quad \begin{bmatrix} 2 & 4 & 3 \\ 3 & 10 & 11 \\ 1 & 6 & 5 \end{bmatrix};$$
$$u_{13} = \frac{6}{2} = 3;$$

$$j = 2:$$
$$l_{22} = 10 - 3 \cdot 4 = -2; \; l_{32} = 6 - 1 \cdot 4 = 2; \quad \begin{bmatrix} 2 & 4 & 3 \\ 3 & -2 & -1 \\ 1 & 2 & 5 \end{bmatrix};$$
$$u_{23} = \frac{11 - 3 \cdot 3}{-2} = -1;$$

$$j = 3:$$
$$l_{33} = 5 - 1 \cdot 3 - 2 \cdot (-1) = 4; \quad \begin{bmatrix} 2 & 4 & 3 \\ 3 & -2 & -1 \\ 1 & 2 & 4 \end{bmatrix};$$

$$\mathbf{A_{LU}} = \begin{bmatrix} 2 & 4 & 3 \\ 3 & -2 & -1 \\ 1 & 2 & 4 \end{bmatrix} \Rightarrow \; \mathbf{L} = \begin{bmatrix} 2 & 0 & 0 \\ 3 & -2 & 0 \\ 1 & 2 & 4 \end{bmatrix}; \quad \mathbf{U} = \begin{bmatrix} 1 & 4 & 3 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix};$$

$$\mathbf{LY} = \mathbf{B} \Rightarrow \begin{bmatrix} 2 & 0 & 0 \\ 3 & -2 & 0 \\ 1 & 2 & 4 \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 28 \\ 40 \\ 20 \end{bmatrix},$$

$$y_1 = \frac{28}{2} = 14; \qquad y_2 = \frac{40 - 3 \cdot 14}{-2} = 1; \qquad y_3 = \frac{20 - 1 \cdot 14 - 2 \cdot 1}{4} = 1;$$

$$\mathbf{UX}=\mathbf{Y} \Rightarrow \begin{bmatrix} 1 & 4 & 3 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 14 \\ 1 \\ 1 \end{bmatrix}; \quad \begin{matrix} x_3 = 1; \\ x_2 = 1 - (-1) \cdot 1 = 2; \\ x_1 = 14 - 4 \cdot 2 - 3 \cdot 1 = 3; \end{matrix} \quad \mathbf{X} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}.$$

Описану схему розв'язку можна реалізувати лише у випадку, коли діагональні елементи $l_{jj} \neq 0$. Крім того, близькість цих елементів до нуля

може призвести до значної втрати точності. Щоб запобігти цьому, схему необхідно виконати з вибором (за допомогою перестановки рядків або стовпчиків) найбільшого за модулем ведучого елемента l_{jj}. Для цього матрицю **A** можна подати у вигляді **A** = **PLU**, де **P** – матриця перестановок.

Під час використання методу LU-факторизації з вибором головного елементу, крім матриць **L** та **U** необхідно зберігати матрицю перестановок **P**. Проте, якщо переставляються тільки рядки (або тільки стовпці), то замість матриці можна зберігати вектор, який містить поточні номери рядків (стовпців).

Для вибору головного елемента по стовпцю на кожному кроці k виконуються наступні операції. Обчислюються елементи l_{ik} з (2.18) та розміщуються на місцях $a_{ik}, i = \overline{k,n}$. Якщо $\max\limits_{i \geq k} |l_{ik}| = |l_{im}|$ і $m \neq k$, то рядки k та m обмінюються місцями, тобто для всіх елементів індекс рядки m замінюють індексом k та навпаки. У цьому випадку запам'ятовуються номери переставлених рядків (якщо матриця перестановок **P** зберігається як вектор з поточними номерами рядків, то в ньому також переставляються елементи з індексами k та m). Обчислюються елементи u_{ki} за виразами (2.19) і розміщуються на місці елементів $a_{ki}, i = \overline{k+1,n}$.

2.2.2. Метод QR-факторизації

У підрозділі 2.1.3 показано, що будь-яку невироджену матрицю можна подати у вигляді добутку унітарної матриці **Q** та верхньої трикутної матриці **R** (див. формулу (2.14)).

Як матрицю **Q** найчастіше використовують матрицю відбиття [9, с.80-82] або матрицю обертання [5, с. 55-58]. У першому випадку метод

QR-факторизації називають методом відбиття, а в другому – методом обертання.

У методі відбиття матрицю R отримують шляхом множення матриці A зліва на $n-1$ матрицю відбиття $\mathbf{Q}_j$:

$$\mathbf{A}_{n-1} = \mathbf{Q}_{n-1}\mathbf{Q}_{n-2}\ldots\mathbf{Q}_1\mathbf{A}.$$

Кожна матриця $\mathbf{Q}_j$ у процесі множення обнуляє елементи *j*-го стовпчика, що знаходяться нижче головної діагоналі, залишаючи без змін перші *j*-1 стовпчиків.

Для того, щоб обнулити елементи *j*-го стовпця матриці $\mathbf{A}_{j-1}$, необхідно щоб перетворення Хаусхолдера переводило вектор $\mathbf{Z}$

$$\mathbf{Z} = \left[z_1,\ z_2,\ldots\ z_{j-1},\ z_j,\ldots\ z_n \right]^{\mathrm{T}} = \left[a_{1j}^{(j-1)},\ a_{2j}^{(j-1)},\ \ldots\ a_{j-1\,j}^{(j-1)}, a_{jj}^{(j-1)},\ \ldots\ a_{nj}^{(j-1)} \right]^{\mathrm{T}},$$

який є *j*-м стовпчиком матриці $\mathbf{A}_{j-1,}$ у вектор, паралельний вектору $\mathbf{Y} = \left[z_1, z_2, \ldots z_{j-1}, \alpha_j, 0 \ldots 0 \right]^{\mathrm{T}}$, де

$$\alpha_j = \pm\sqrt{\sum_{i=j}^{n} a_{ij}^{(j-1)} a_{ij}^{(j-1)*}}\,. \qquad (2.23)$$

Введемо вектор

$$\mathbf{S}_j = \mathbf{Z} - \alpha_j \frac{\mathbf{Y}}{\|\mathbf{Y}\|_2} = \left[0, 0, \ldots, z_j - \alpha_j, z_{j+1}, \ldots z_n \right]^{\mathrm{T}}. \qquad (2.24)$$

Відповідно до (2.9)

$$\mathbf{W} = \frac{\mathbf{S}_j}{\|\mathbf{S}_j\|_2}.$$

Знак α_j у (2.23) вибирається так, щоб максимізувати $\left\|\mathbf{S}_j\right\|_2$. Оскільки α_j є дійсним числом, це досягається, коли $\alpha_j \operatorname{Re}\left(z_j\right) < 0$. Враховуючи (2.24), коефіцієнт ω з перетворення (2.9) можна визначити як

$$\omega = \frac{\mathbf{Z}^\dagger \mathbf{W}}{\mathbf{W}^\dagger \mathbf{Z}} = \frac{\mathbf{Z}^\dagger \dfrac{\mathbf{S}_j}{\left\|\mathbf{S}_j\right\|_2}}{\dfrac{\mathbf{S}_j^\dagger}{\left\|\mathbf{S}_j^\dagger\right\|_2}\mathbf{Z}} = \frac{\mathbf{Z}^\dagger \mathbf{S}_j}{\mathbf{S}_j^\dagger \mathbf{Z}} = \frac{\alpha_j^2 - \alpha_j z_j^*}{\alpha_j^2 - \alpha_j z_j}.$$

Відповідно до (2.24)

$$\mathbf{S}_j^\dagger \mathbf{S}_j = 2\alpha_j^2 - \alpha_j z_j^* - \alpha_j z_j .$$

Тоді матрицю $\mathbf{Q}_j$ можна обчислити з виразу

$$\mathbf{Q}_j = \mathbf{E} - k_j \mathbf{S}_j \mathbf{S}_j^\dagger,$$

де

$$k_j = \frac{1+\omega}{\mathbf{S}_j^\dagger \mathbf{S}_j} = \frac{1 + \dfrac{\alpha_j^2 - \alpha_j z_j^*}{\alpha_j^2 - \alpha_j z_j}}{2\alpha_j^2 - \alpha_j z_j^* - \alpha_j z_j}.$$

Матрицю **Q** у виразі (2.14) можна знайти через добуток елементарних матриць $\mathbf{Q}_j$. Однак на практиці матрицю **Q** явно не обчислюють, а за n-1 крок зводять систему до вигляду

$$\mathbf{A}_{n-1}\mathbf{X} = \mathbf{B}_{n-1}, \qquad (2.25)$$

де на кожному j-му кроці матрицю $\mathbf{A}_j$ та вектор $\mathbf{B}_j$ знаходять з попередньо розрахованих за рекурентними формулами:

$$\mathbf{A}_j = \mathbf{A}_{j-1} - k_j \mathbf{S}_j (\mathbf{S}_j^\dagger \mathbf{A}_{j-1});$$

$$\mathbf{B}_j = \mathbf{B}_{j-1} - k_j \mathbf{S}_j (\mathbf{S}_j^\dagger \mathbf{B}_{j-1}), \quad j = \overline{1, n-1}.$$

Розв'язок системи (2.25) знаходять з виразу

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

Для СЛАР з дійсними коефіцієнтами розрахункові формули спрощуються. Так, коефіцієнт $\alpha_j = \pm\sqrt{\sum_{i=j}^{n} \left(a_{ij}^{(j-1)} \right)^2}$, а знак α_j вибирається з умови $\alpha_j z_j < 0$. Оскільки $\omega = 1$, то

$$k_j = \frac{2}{\mathbf{S}_j^{\mathrm{T}} \mathbf{S}_j} = \frac{2}{2\alpha_j^2 - 2\alpha_j z_j} = \frac{1}{\alpha_j^2 - \alpha_j z_j}.$$

Тоді у формулах перерахунку операцію ермітового спряження можна замінити операцією транспонування:

$$\mathbf{A}_j = \mathbf{A}_{j-1} - k_j \mathbf{S}_j (\mathbf{S}_j^{\mathrm{T}} \mathbf{A}_{j-1});$$

$$\mathbf{B}_j = \mathbf{B}_{j-1} - k_j \mathbf{S}_j (\mathbf{S}_j^{\mathrm{T}} \mathbf{B}_{j-1}),\ \ j = \overline{1, n-1}.$$

Приклад:

$$\begin{bmatrix} 1 & 3 & -11 \\ 2 & 17 & 10 \\ 2 & 16 & -12 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -16 \\ 43 \\ -2 \end{bmatrix}.$$

$$j = 1:\ \alpha_1^2 = 1^2 + 2^2 + 2^2 = 9;\ z_1 = 1 > 0 \Rightarrow \alpha_1 = -\sqrt{9} = -3;$$

$$k_1 = \frac{1}{9 - (-3) \cdot 1} = \frac{1}{12}; \qquad \mathbf{S}_1^{\mathrm{T}} = \begin{bmatrix} 4 & 2 & 2 \end{bmatrix};$$

$$\mathbf{S}_1^{\mathrm{T}} \mathbf{A} = \begin{bmatrix} 4 & 2 & 2 \end{bmatrix} \begin{bmatrix} 1 & 3 & -11 \\ 2 & 17 & 10 \\ 2 & 16 & -12 \end{bmatrix} = \begin{bmatrix} 12 & 78 & -48 \end{bmatrix};$$

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 3 & -11 \\ 2 & 17 & 10 \\ 2 & 16 & -12 \end{bmatrix} - \frac{1}{12}\begin{bmatrix} 4 \\ 2 \\ 2 \end{bmatrix}\begin{bmatrix} 12 & 78 & -48 \end{bmatrix} = \begin{bmatrix} -3 & -23 & 5 \\ 0 & 4 & 18 \\ 0 & 3 & -4 \end{bmatrix}.$$

$$j = 2: \quad \alpha_2{}^2 = 4^2 + 3^2 = 25; \;\; z_2 = 4 > 0 \Rightarrow \alpha_2 = -\sqrt{25} = -5;$$

$$k_2 = \frac{1}{25 - (-5)\cdot 4} = \frac{1}{45}; \quad \mathbf{S}_2^{\mathrm{T}} = \begin{bmatrix} 0 & 9 & 3 \end{bmatrix}.$$

$$\mathbf{S}_2^{\mathrm{T}}\,\mathbf{A}_1 = \begin{bmatrix} 0 & 9 & 3 \end{bmatrix}\begin{bmatrix} -3 & -23 & 5 \\ 0 & 4 & 18 \\ 0 & 3 & -4 \end{bmatrix} = \begin{bmatrix} 0 & 45 & 150 \end{bmatrix}.$$

$$\mathbf{A}_2 = \begin{bmatrix} -3 & -23 & 5 \\ 0 & 4 & 18 \\ 0 & 3 & -4 \end{bmatrix} - \frac{1}{45}\begin{bmatrix} 0 \\ 9 \\ 3 \end{bmatrix}\begin{bmatrix} 0 & 45 & 150 \end{bmatrix} = \begin{bmatrix} -3 & -23 & 5 \\ 0 & -5 & -12 \\ 0 & 0 & -14 \end{bmatrix};$$

$$\mathbf{S}_1^{\mathrm{T}}\,\mathbf{B} = \begin{bmatrix} 4 & 2 & 2 \end{bmatrix}\begin{bmatrix} -16 \\ 43 \\ -2 \end{bmatrix} = 18, \quad \mathbf{B}_1 = \begin{bmatrix} -16 \\ 43 \\ -2 \end{bmatrix} - \frac{1}{12}\begin{bmatrix} 4 \\ 2 \\ 2 \end{bmatrix}18 = \begin{bmatrix} -22 \\ 40 \\ -5 \end{bmatrix};$$

$$\mathbf{S}_2^{\mathrm{T}}\,\mathbf{B}_1 = \begin{bmatrix} 0 & 9 & 3 \end{bmatrix}\begin{bmatrix} -22 \\ 40 \\ -5 \end{bmatrix} = 345, \quad \mathbf{B}_2 = \begin{bmatrix} -22 \\ 40 \\ -5 \end{bmatrix} - \frac{1}{45}\begin{bmatrix} 0 \\ 9 \\ 3 \end{bmatrix}345 = \begin{bmatrix} -22 \\ -29 \\ -28 \end{bmatrix};$$

$$\begin{bmatrix} -3 & -23 & 5 \\ 0 & -5 & -12 \\ 0 & 0 & -14 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -22 \\ -29 \\ -28 \end{bmatrix} \Rightarrow x_3 = \frac{-28}{-14} = 2; \quad x_2 = \frac{-29 - (-12)\cdot 2}{-5} = 1;$$

$$x_1 = \frac{-22 - (-23)\cdot 1 - 5\cdot 2}{-3} = 3 \Rightarrow \mathbf{X} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}.$$

2.2.3. Метод Холеського

Метод Холеського [10,11] використовують для розв'язування СЛАР із ермітовою матрицею **A**. Він ґрунтується на факторизації матриці **A** у добуток

$$\mathbf{A} = \mathbf{LDL}^{\dagger}, \tag{2.26}$$

де **L** – нижня трикутна матриця; **D** – діагональна матриця.

Із виразу (2.26) маємо

$$a_{ij} = \sum_{k=1}^{i} l_{ik} d_{kk} \overline{l}_{jk}\ , \qquad i \geq j\ ,$$

де $\overline{l}_{jk}$ - елемент, комплексно-спряжений до l_{jk}

Для $i = j$ одержуємо

$$d_{jj} l_{jj}^2 = a_{jj} - \sum_{k=1}^{j-1} \left(l_{jk}^2 d_{kk} \right). \tag{2.27}$$

Поклавши $l_{ii} = 1,\ i = \overline{1,n}$, в виразі (2.27), маємо

$$d_{jj} = a_{jj} - \sum_{k=1}^{j-1} (l_{jk}^2 d_{kk}), \qquad j = \overline{1,n}\,. \tag{2.28}$$

Якщо $i > j$, то

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} d_{kk} \overline{l}_{jk}}{d_{jj}}, \qquad j = \overline{1,n}\,,\ i = \overline{j+1,n}\,. \tag{2.29}$$

Таким чином, $\mathbf{LDL}^{\dagger}$ – факторизація матриці виконується за n стадій. На кожній j-й стадії вважаємо, що $l_{jj} = 1$. За формулою (2.28) спочатку обчислюємо елемент d_{jj}, а потім елементи l_{ij} чергового j-го стовпчика матриці **L** за виразом (2.29).

Після факторизації матриці **A** розв'язуємо СЛАР **LY** = **B** за допомогою прямої підстановки

$$y_i = b_i - \sum_{j=1}^{i-1}\left(l_{ij} y_j\right),\ \ i = \overline{1,n}\,.$$

Потім розв'язуємо СЛАР **DZ** = **Y:**

$$z_i = \frac{y_i}{d_{ii}}, \qquad i = \overline{1,n}\,.$$

Розв'язок **X** знаходимо із системи $\mathbf{L}^{\dagger}\mathbf{X} = \mathbf{Z}$ зворотною підстановкою:

$$x_i = z_i - \sum_{j=i+1}^{n}\left(\overline{l}_{ji}\, x_j\right), i = \overline{n,1}\,.$$

Приклад:

$$\begin{bmatrix} 9 & -2 & -6 \\ -2 & 16 & -12 \\ -6 & -12 & 18 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -13 \\ -6 \\ 24 \end{bmatrix},$$

$$j = 1 : l_{11} = 1;\ \ d_{11} = 9;\ \ l_{21} = -\frac{2}{9};\ \ \ l_{31} = -\frac{6}{9} = -\frac{2}{3};$$

$$j = 2 : l_{22} = 1;\ \ \ d_{22} = 16 - \frac{4}{81}9 = \frac{140}{9};\ \ \ l_{32} = \frac{-12 + \frac{2}{3}\cdot 9 \cdot \left(-\frac{2}{9}\right)}{\frac{140}{9}} = -\frac{6}{7};$$

$$j = 3 : l_{33} = 1;\ \ d_{33} = 18 - \frac{4}{9}9 - \frac{36}{49}\frac{140}{9} = \frac{18}{7}.$$

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ -2/9 & 1 & 0 \\ -2/3 & -6/7 & 1 \end{bmatrix};\qquad \mathbf{D} = \begin{bmatrix} 9 & 0 & 0 \\ 0 & \frac{140}{9} & 0 \\ 0 & 0 & 18/7 \end{bmatrix}.$$

$$\mathbf{LY}=\mathbf{B}\Rightarrow\begin{bmatrix}1&0&0\\-\frac{2}{9}&1&0\\-\frac{2}{3}&-\frac{6}{7}&1\end{bmatrix}\times\begin{bmatrix}y_1\\y_2\\y_3\end{bmatrix}=\begin{bmatrix}-13\\-6\\24\end{bmatrix}\Rightarrow$$

$$y_1=-13;\ y_2=-6-\frac{2}{9}13=-\frac{80}{9};\ y_3=24-\frac{2}{3}13-\frac{6}{7}\frac{80}{9}=\frac{54}{7}.$$

$$\mathbf{DZ}=\mathbf{Y}\Rightarrow\begin{bmatrix}9&0&0\\0&\frac{140}{9}&0\\0&0&\frac{18}{7}\end{bmatrix}\times\begin{bmatrix}z_1\\z_2\\z_3\end{bmatrix}=\begin{bmatrix}-13\\-\frac{80}{9}\\\frac{54}{7}\end{bmatrix}\Rightarrow$$

$$z_1=-\frac{13}{9};\ z_2=-\frac{4}{7};\ z_3=3.$$

$$\mathbf{L}^{\dagger}\mathbf{X}=\mathbf{Z}\Rightarrow\begin{bmatrix}1&-\frac{2}{9}&-\frac{2}{3}\\0&1&-\frac{6}{7}\\0&0&1\end{bmatrix}\times\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}=\begin{bmatrix}-\frac{13}{9}\\-\frac{4}{7}\\3\end{bmatrix}\Rightarrow$$

$$x_3=3;\ x_2=-\frac{4}{7}+\frac{6}{7}3=2;\ x_1=-\frac{13}{9}+\frac{2}{9}2+\frac{2}{3}3=1\ .$$

Часто задачі електроніки приводять до СЛАР не тільки з ермітовою (симетричною у випадку дійсної матриці), а ще й додатньо-визначеною матрицею **A** [7], тобто такою, що для довільного вектора **X**≠0 справедлива нерівність $\mathbf{X}^{\dagger}\mathbf{AX}>0$, де $\mathbf{X}^{\dagger}$ - комплексно-спряжений транспонований вектор [8]. В цьому випадку $d_{jj}>0, j=\overline{1,n}$ і схему розв'язку можна спростити, використовуючи $\mathbf{L}^{\dagger}\mathbf{L}$-факторизацію (метод квадратного кореня) [2]:

$$\mathbf{A} = \mathbf{L}^{\dagger}\mathbf{L}\,. \qquad (2.30)$$

Тоді з виразу (2.30) знаходимо

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}\,, \quad j = \overline{1,n}\,;$$

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}\overline{l}_{jk}}{l_{jj}}, \qquad j = \overline{1,n}\,, \quad i = \overline{j+1,n}\,.$$

Після факторизації матриці **A** розв'язок задачі зводимо до розв'язку двох СЛАР з трикутними матрицями **LY**=**B**, $\mathbf{L}^{\dagger}\mathbf{X} = \mathbf{Y}$ за формулами:

$$y_i = \frac{b_i - \sum_{j=1}^{i-1} l_{ij} y_j}{l_{ii}}, \quad i = \overline{1,n}\;;$$

$$x_i = \frac{y_i - \sum_{j=i+1}^{n} \overline{l}_{ji} x_j}{l_{ii}}, \quad i = \overline{n,1}\,.$$

Розв'яжемо цим методом СЛАР із попереднього прикладу:

$$\begin{bmatrix} 9 & -2 & -6 \\ -2 & 16 & -12 \\ -6 & -12 & 18 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -13 \\ -6 \\ 24 \end{bmatrix}.$$

$$l_{11} = \sqrt{9} = 3;$$

$$l_{21} = -\frac{2}{3};$$

$$l_{31} = -\frac{6}{3} = -2;$$

$$l_{22} = \sqrt{16 - \frac{4}{9}} = \sqrt{\frac{140}{9}} = \frac{\sqrt{140}}{3};$$

$$l_{32} = \frac{-12 - \frac{4}{3}}{\frac{\sqrt{140}}{3}} = -\frac{40}{\sqrt{140}};$$

$$l_{33} = \sqrt{18 - 4 - \frac{1600}{140}} = \sqrt{\frac{98}{7} - \frac{80}{7}} = \sqrt{\frac{18}{7}};$$

$$\mathbf{L} = \begin{bmatrix} 3 & 0 & 0 \\ -\frac{2}{3} & \frac{\sqrt{140}}{3} & 0 \\ -2 & -\frac{40}{\sqrt{140}} & \sqrt{\frac{18}{7}} \end{bmatrix}.$$

$$\mathbf{LY} = \mathbf{B} \Rightarrow \quad \begin{bmatrix} 3 & 0 & 0 \\ -\frac{2}{3} & \frac{2\sqrt{35}}{3} & 0 \\ -2 & -\frac{20}{\sqrt{35}} & 3\sqrt{\frac{2}{7}} \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -13 \\ -6 \\ 24 \end{bmatrix} \Rightarrow$$

$$y_1 = -\frac{13}{3};\ y_2 = \frac{-6 - \frac{2}{3}\frac{13}{3}}{\frac{2\sqrt{35}}{3}} = -\frac{40}{3\sqrt{35}};\ y_3 = \frac{24 - 2\frac{13}{3} - \frac{20}{\sqrt{35}}\frac{40}{3\sqrt{35}}}{3\sqrt{\frac{2}{7}}} = 9\sqrt{\frac{2}{7}}.$$

$$\mathbf{L}^{\dagger}\mathbf{X} = \mathbf{Y} \Rightarrow \quad \begin{bmatrix} 3 & -\frac{2}{3} & -2 \\ 0 & \frac{2\sqrt{35}}{3} & -\frac{20}{\sqrt{35}} \\ 0 & 0 & 3\sqrt{\frac{2}{7}} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -\frac{13}{3} \\ -\frac{40}{3\sqrt{35}} \\ 9\sqrt{\frac{2}{7}} \end{bmatrix}.$$

$$x_3 = \frac{9\sqrt{\frac{2}{7}}}{3\sqrt{\frac{2}{7}}} = 3;\ \ x_2 = \frac{-\frac{40}{3\sqrt{35}} + \frac{20}{\sqrt{35}}3}{\frac{2\sqrt{35}}{3}} = 2;\ \ x_1 = \frac{-\frac{13}{3} + \frac{2}{3}2 + 2 \cdot 3}{3} = 1.$$

2.3. Похибка розв'язання СЛАР

Часто компоненти матриці **A** і вектора **B** СЛАР є наближеними значеннями. Крім того, в процесі розв'язання СЛАР виникає похибка округлення. Розглянемо, як сильно ці похибки впливають на точність розв'язку СЛАР.

Можна показати [5, с. 21-23], що відношення $\|\Delta\mathbf{X}\|/\|\mathbf{X}\|$ пропорційне добутку $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$, де $\|\Delta\mathbf{X}\|$ – норма похибки розв'язку; $\|\mathbf{X}\|$ – норма точного розв'язку; $\|\mathbf{A}\|$ – норма матриці **A**; $\|\mathbf{A}^{-1}\|$ – норма матриці, оберненої до **A**.

Добуток $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ називають числом обумовленості матриці **A** [5, с. 22-23]:

$$\mathrm{cond}(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|. \qquad (2.31)$$

Використовуючи в (2.31) евклідову норму, маємо

$$cond(\mathbf{A}) = \sqrt{\frac{\max\limits_i |\mu_i|}{\min\limits_i |\mu_i|}},$$

де μ_i – власні числа матриці $\mathbf{A}^{\dagger}\mathbf{A}$.

Якщо матриця **A** – ермітова, то

$$cond(\mathbf{A}) = \frac{\max\limits_i |\lambda_i|}{\min\limits_i |\lambda_i|}$$

де λ_i – власне число матриці **A**.

Число обумовленості є мірою виродженості матриці. Якщо воно велике, то похибки округлення можуть суттєво вплинути на результат

розв'язування СЛАР. Матриця з великим числом обумовленості називається погано обумовленою. На практиці вважають, що коли $\mathrm{cond}(\mathbf{A}) \geq \frac{1}{\sqrt{\varepsilon}}$ (ε – відносна похибка округлення), то обчислений розв'язок є ненадійним.

У такому випадку необхідно провести регуляризацію системи [12, с. 431-440], наприклад, методом Тихонова [13, с. 235-256]

$$\left(\mathbf{A}^{\dagger}\mathbf{A} + \alpha\mathbf{E}\right)\mathbf{X} = \mathbf{A}^{\dagger}\mathbf{B}, \qquad (2.32)$$

де $\mathbf{E}$ – одинична матриця; α – параметр регуляризації, деяке досить мале значення. Число α повинно бути, з одного боку, досить великим, щоб суттєво зменшити число обумовленості матриці СЛАР, а з другого боку, досить малим, щоб СЛАР (2.32) мало відрізнялася від початкової системи.

Для практичного знаходження оптимального значення α може бути застосовано такий метод [14, с. 137-138]. Для деякого обраного α знаходять розв'язок $\mathbf{X}_{\alpha}$ СЛАР (2.32), обчислюють нев'язку $\mathbf{R}_{\alpha} = \mathbf{A}\mathbf{X}_{\alpha} - \mathbf{B}$ та порівнюють її норму з нормами відомої похибки вектора правої частини (2.2) $\Delta\mathbf{B}$ та впливу на точність розв'язку похибки коефіцієнтів матриці $\Delta\mathbf{A}\mathbf{X}_{\alpha}$. Якщо α значно більше за оптимальне, то нев'язка помітно більша за ці похибки, якщо значно менше, то помітно менше. Проводять серію розрахунків з різними α і оптимальним рахують те значення α, за якого $\|\mathbf{R}_{\alpha}\| \approx \|\Delta\mathbf{B}\| + \|\Delta\mathbf{A}\mathbf{X}_{\alpha}\|$.

Слід зазначити, що матриця системи (2.32) є ермітовою. Тому для її розв'язку можна застосовувати метод квадратного кореня.

2.4. Ітераційні методи розв’язання СЛАР

Досить часто в інженерних задачах доводиться розв’язувати СЛАР великої розмірності, у яких матриця коефіцієнтів системи містить велику кількість нульових елементів (матриця коефіцієнтів “розріджена”). За наявності у матриці коефіцієнтів системи великої кількості нульових елементів, використання методів факторизації може виявитись неефективним з точки зору використання ресурсів та часу обчислювальної системи. В цьому випадку більш ефективними можуть виявитися ітераційні методи розв’язку СЛАР.

Розглянемо побудову ітераційних методів у загальному випадку. Домножимо праву та ліву частини (2.2) на відому матрицю **C** та додамо в обидві частини вектор **X.** Тоді (2.2) можна переписати у вигляді

$$\mathbf{X} = \left(\mathbf{CA} + \mathbf{E}\right)\mathbf{X} - \mathbf{CB},$$

де **E** – одинична матриця.

Або

$$\mathbf{X} = \tilde{\mathbf{A}}\mathbf{X} + \tilde{\mathbf{B}}, \qquad (2.33)$$

де $\tilde{\mathbf{A}} = \mathbf{CA} + \mathbf{E}$, а $\tilde{\mathbf{B}} = -\mathbf{CB}$.

Якщо $\mathbf{X}^*$ є розв’язком (2.2), то цей вектор перетворює (2.33) у тотожність. Але оскільки точний розв’язок (2.2) невідомий, то можна вибрати початкове наближення $\mathbf{X}_0$, та підставити його в праву частину (2.33). В результаті отримаємо вектор $\mathbf{X}_1$. Далі можна підставити $\mathbf{X}_1$ в праву частину (2.33) та отримати наступне наближення $\mathbf{X}_2$. Цю процедуру повторюють до тих пір, поки наступне наближення не буде повторювати попереднє з заданою похибкою. Такий процес називають ітераційним, і його загальна формула на *k*-му ітераційному кроці має вигляд:

$$\mathbf{X}_{k+1} = \tilde{\mathbf{A}}\mathbf{X}_k + \tilde{\mathbf{B}}. \qquad (2.34)$$

Якщо існує границя послідовності $\mathbf{X} = \lim\limits_{k\to\infty}\left(\tilde{\mathbf{A}}\mathbf{X}_k + \tilde{\mathbf{B}}\right)$, то ітераційний процес (2.34) збігається, а $\mathbf{X}$ є розв'язком системи (2.2).

Можна показати [5, с. 63-69], що для збіжності методу (2.34) необхідно, щоб

$$\left\|\tilde{\mathbf{A}}\right\| < 1. \qquad (2.35)$$

Матрицю $\mathbf{C}$ вибирають таким чином, щоб задовольнити умову (2.35). В залежності від вибору матриці $\mathbf{C}$ існує багато різновидів ітераційних методів розв'язку СЛАР [15]. Нижче будуть розглянуті найпростіші методи.

2.4.1. Метод простої ітерації

Розв'яжемо перше рівняння системи (2.1) відносно x_1, друге відносно x_2 і так далі. В результаті отримаємо систему рівнянь, еквівалентну системі (2.1):

$$\begin{cases} x_1 = \beta_1 + \alpha_{12}x_2 + \alpha_{13}x_3 + \ldots + \alpha_{1n}x_n; \\ x_2 = \beta_2 + \alpha_{21}x_1 + \alpha_{23}x_3 + \ldots + \alpha_{2n}x_n; \\ \qquad\qquad \ldots \\ x_n = \beta_n + \alpha_{n1}x_1 + \alpha_{n2}x_2 + \ldots + \alpha_{nn-1}x_{n-1}, \end{cases} \qquad (2.36)$$

де $\beta_i = b/a_{ii}$; $\alpha_{ij} = -a_{ij}/a_{ii}$ якщо $i \neq j$, $\alpha_{ij} = 0$ у противному випадку.

Використовуючи позначення

$$\mathbf{X}=\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \\ \end{bmatrix},\ \tilde{\mathbf{B}}=\begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_n \\ \end{bmatrix},\ \tilde{\mathbf{A}}=\begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ & & \cdots & \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{bmatrix},$$

систему (2.36) можна записати у матричному вигляді (2.33). Таким чином, метод простої ітерації є частковим випадком методу (2.34) коли матриця **C** дорівнює:

$$\mathbf{C}=\begin{bmatrix} -\frac{1}{a_{11}} & 0 & \cdots & 0 \\ 0 & -\frac{1}{a_{22}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & -\frac{1}{a_{nn}} \end{bmatrix}.$$

Для розв'язання системи за допомогою формули (2.34) необхідно вибрати початкове наближення $\mathbf{X}_0$. В якості такого наближення можна вибрати $\mathbf{X}_0=\tilde{\mathbf{B}}$.

На практиці у разі застосування методу простої ітерації за критерій зупинки обчислень обирають збіжність за аргументом:

$$\left\|\Delta\mathbf{X}_k\right\|=\left\|\mathbf{X}_{k+1}-\mathbf{X}_k\right\|<\Delta, \tag{2.37}$$

де Δ — задана абсолютна похибка обчислень, або:

$$\frac{\left\|\Delta\mathbf{X}_k\right\|}{\left\|\mathbf{X}_{k+1}\right\|}<\varepsilon, \tag{2.38}$$

де ε — задана відносна похибка обчислень.

Для збіжності ітераційного процесу методу простої ітерації достатньо, щоб матриця системи була діагонально-домінуючою [16]:

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| < |a_{ii}|, i = 1, 2, \ldots, n. \qquad (2.39)$$

Приклад

Нехай необхідно розв'язати систему рівнянь методом простої ітерації з абсолютною похибкою не більше 1%:

$$\begin{bmatrix} 10 & 3 & 0 \\ 0 & 15 & 3 \\ 2 & 0 & 20 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 16 \\ 39 \\ 62 \end{bmatrix}. \qquad (2.40)$$

Перевіримо виконання умови (2.39):

$$\begin{aligned} &i = 1: \ a_{11} = 10, \quad a_{12} + a_{13} = 3, \quad 10 > 3; \\ &i = 2: \ a_{22} = 15, \quad a_{21} + a_{23} = 3, \quad 15 > 3; \\ &i = 3: \ a_{33} = 20, \quad a_{31} + a_{32} = 2, \quad 20 > 2. \end{aligned}$$

Умова (2.39) виконується, отже система (2.40) може бути розв'язана методом простої ітерації. Ітераційна формула відповідно до (2.34) має вигляд:

$$\tilde{\mathbf{B}} = \mathbf{X}_0 = \begin{bmatrix} 16/10 \\ 39/15 \\ 62/20 \end{bmatrix} = \begin{bmatrix} 1,6 \\ 2,6 \\ 3,1 \end{bmatrix}; \qquad \mathbf{X}_{k+1} = \begin{bmatrix} 0 & -0,3 & 0 \\ 0 & 0 & -0,2 \\ -0,1 & 0 & 0 \end{bmatrix} \mathbf{X}_k + \begin{bmatrix} 1,6 \\ 2,6 \\ 3,1 \end{bmatrix}.$$

Виконаємо ітерацію.

$$\mathbf{X}_1 = \begin{bmatrix} 0 & -0,3 & 0 \\ 0 & 0 & -0,2 \\ -0,1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1,6 \\ 2,6 \\ 3,1 \end{bmatrix} + \begin{bmatrix} 1,6 \\ 2,6 \\ 3,1 \end{bmatrix} = \begin{bmatrix} 0,82 \\ 1,98 \\ 2,94 \end{bmatrix}.$$

Використовуючи норми l_∞, перевіримо досягнуту похибку розв'язку:

$$\|\Delta\mathbf{X}_0\|_\infty = \begin{Vmatrix} 0,82-1,6 \\ 1,98-2,6 \\ 2,94-3,1 \end{Vmatrix}_\infty = 0,78;\ \|\mathbf{X}_1\|_\infty = \begin{Vmatrix} 0,82 \\ 1,98 \\ 2,94 \end{Vmatrix}_\infty = 2,94;\ \frac{\|\Delta\mathbf{X}_0\|}{\|\mathbf{X}_1\|} = \frac{0,78}{2,94} \cong 0,265.$$

Потрібна точність не досягнута, виконуємо наступну ітерацію:

$$\mathbf{X}_2 = \begin{bmatrix} 0 & -0,3 & 0 \\ 0 & 0 & -0,2 \\ -0,1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0,82 \\ 1,98 \\ 2,94 \end{bmatrix} + \begin{bmatrix} 1,6 \\ 2,6 \\ 3,1 \end{bmatrix} = \begin{bmatrix} 1,006 \\ 2,012 \\ 3,018 \end{bmatrix}.$$

$$\|\Delta\mathbf{X}_1\| = \begin{Vmatrix} 1,006-0,82 \\ 2,012-1,98 \\ 3,018-2,94 \end{Vmatrix} = 0,186;\ \|\mathbf{X}_2\| = \begin{Vmatrix} 1,006 \\ 2,012 \\ 3,018 \end{Vmatrix} = 3,018;\ \frac{\|\Delta\mathbf{X}_1\|}{\|\mathbf{X}_2\|} = \frac{0,186}{3,018} \cong 0,062.$$

Потрібна точність не досягнута, виконуємо наступну ітерацію:

$$\mathbf{X}_3 = \begin{bmatrix} 0 & -0,3 & 0 \\ 0 & 0 & -0,2 \\ -0,1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1,006 \\ 2,012 \\ 3,018 \end{bmatrix} + \begin{bmatrix} 1,6 \\ 2,6 \\ 3,1 \end{bmatrix} = \begin{bmatrix} 0,9964 \\ 1,9964 \\ 2,9994 \end{bmatrix}.$$

$$\|\Delta\mathbf{X}_2\| = \begin{Vmatrix} 0,9964-1,006 \\ 1,9964-2,012 \\ 2,9994-3,018 \end{Vmatrix} = 0,0186;\ \|\mathbf{X}_3\| = \begin{Vmatrix} 0,9964 \\ 1,9964 \\ 2,9994 \end{Vmatrix} = 2,9994;$$

$$\frac{\|\Delta\mathbf{X}_2\|}{\|\mathbf{X}_3\|} = \frac{0,0186}{2,9994} \cong 0,006.$$

Необхідна точність досягнута. Для порівняння, точний розв'язок системи $\mathbf{X}^* = [1;2;3]^T$.

2.4.2. Метод Зейделя

Метод Зейделя – це наближений метод розв'язку систем рівнянь, який є модифікацією метода простої ітерації.

Основна ідея цього методу полягає в тому, що для покращення збіжності ітераційного процесу, для розрахунку $(k+1)$ -го наближення i-ї змінної $x_i^{(k+1)}$ вектора $\mathbf{X}_{k+1}$ використовуються не лише компоненти вектора $\mathbf{X}_k$, але й всі обчислені до цього моменту значення компоненти вектора $\mathbf{X}_{k+1}$ - $\left[x_1^{(k+1)}, x_2^{(k+1)}, ..., x_{i-1}^{(k+1)}\right]^T$:

$$\begin{cases} x_1^{(k+1)} = \beta_1 + \alpha_{12}x_2^{(k)} + \alpha_{13}x_3^{(k)} + ... + \alpha_{1n}x_n^{(k)}; \\ x_2^{(k+1)} = \beta_2 + \alpha_{21}x_1^{(k+1)} + \alpha_{23}x_3^{(k)} + ... + \alpha_{1n}x_n^{(k)}; \\ \qquad\qquad \dots \\ x_n^{(k+1)} = \beta_n + \alpha_{n1}x_1^{(k+1)} + \alpha_{n2}x_2^{(k+1)} + ... + \alpha_{nn-1}x_{n-1}^{(k+1)}; \end{cases}$$

В усьому іншому метод не відрізняється від методу простої ітерації.

Розв'яжемо методом Зейделя систему із попереднього прикладу.

$$\tilde{\mathbf{B}} = \mathbf{X}_0 = \begin{bmatrix} 16/10 \\ 39/15 \\ 62/20 \end{bmatrix} = \begin{bmatrix} 1,6 \\ 2,6 \\ 3,1 \end{bmatrix}; \qquad \tilde{\mathbf{A}} = \begin{bmatrix} 0 & -0,3 & 0 \\ 0 & 0 & -0,2 \\ -0,1 & 0 & 0 \end{bmatrix}.$$

Виконаємо ітерацію:

$$\mathbf{X}_1 = \begin{bmatrix} 1,6 - 0\cdot 1,6 - 0,3\cdot 2,6 - 0\cdot 3,1 \\ 2,6 - 0\cdot \underline{0,82} - 0\cdot 2,6 - 0,2\cdot 3,1 \\ 3,1 - 0,1\cdot \underline{0,82} - 0\cdot \underline{\underline{1,98}} - 0\cdot 3,1 \end{bmatrix} = \begin{bmatrix} \underline{0,82} \\ \underline{\underline{1,98}} \\ 3,01 \end{bmatrix}.$$

Перевіримо досягнуту похибку розв'язку:

$$\left\|\Delta\mathbf{X}_0\right\|_\infty = \begin{Vmatrix} 0,82 - 1,6 \\ 1,98 - 2,6 \\ 3,01 - 3,1 \end{Vmatrix}_\infty = 0,78; \quad \left\|\mathbf{X}_1\right\|_\infty = \begin{Vmatrix} 0,82 \\ 1,98 \\ 3,01 \end{Vmatrix}_\infty = 3,01; \quad \frac{\left\|\Delta\mathbf{X}_0\right\|}{\left\|\mathbf{X}_1\right\|} = \frac{0,78}{3,01} \cong 0,260.$$

Потрібна точність не досягнута, виконуємо наступну ітерацію:

$$\mathbf{X}_2 = \begin{bmatrix} 1,6 - 0 \cdot 0,82 - 0,3 \cdot 1,98 - 0 \cdot 3,01 \\ 2,6 - 0 \cdot \underline{1,006} - 0 \cdot 1,98 - 0,2 \cdot 3,01 \\ 3,1 - 0,1 \cdot \underline{1,006} - 0 \cdot \underline{\underline{1,998}} - 0 \cdot 3,01 \end{bmatrix} = \begin{bmatrix} \underline{1,006} \\ \underline{\underline{1,998}} \\ 2,994 \end{bmatrix}.$$

Перевіримо досягнуту похибку розв'язку:

$$\left\|\Delta \mathbf{X}_1\right\| = \left\| \begin{matrix} 1,006 - 0,82 \\ 1,998 - 1,98 \\ 2,994 - 3,01 \end{matrix} \right\| = 0,186; \quad \left\|\mathbf{X}_2\right\| = \left\| \begin{matrix} 1,006 \\ 1,998 \\ 2,994 \end{matrix} \right\| = 2,994; \quad \frac{\left\|\Delta \mathbf{X}_1\right\|}{\left\|\mathbf{X}_2\right\|} = \frac{0,186}{2,994} \cong 0,062.$$

Потрібна точність не досягнута, виконуємо наступну ітерацію:

$$\mathbf{X}_3 = \begin{bmatrix} 1,6 - 0 \cdot 1,006 - 0,3 \cdot 1,998 - 0 \cdot 2,994 \\ 2,6 - 0 \cdot 1,0006 - 0 \cdot 1,998 - 0,2 \cdot 2,994 \\ 3,1 - 0,1 \cdot 1,006 - 0 \cdot 2,0012 - 0 \cdot 2,994 \end{bmatrix} = \begin{bmatrix} 1,0006 \\ 2,0012 \\ 2,9994 \end{bmatrix}.$$

Перевіримо досягнуту похибку розв'язку:

$$\left\|\Delta \mathbf{X}_2\right\| = \left\| \begin{matrix} 1,0006 - 1,006 \\ 2,0012 - 1,998 \\ 2,9994 - 2,994 \end{matrix} \right\| = 0,014; \quad \left\|\mathbf{X}_3\right\| = \left\| \begin{matrix} 1,0006 \\ 2,0012 \\ 2,9994 \end{matrix} \right\| = 2,9994;$$

$$\frac{\left\|\Delta \mathbf{X}_2\right\|}{\left\|\mathbf{X}_3\right\|} = \frac{0,014}{2,9994} \cong 0,005.$$

Потрібна точність досягнута.

Слід зазначити, що метод Зейделя збігається не для всіх систем. Разом з тим, для систем з симметричною та додатньо-визначеною матрицею $\mathbf{A}$ метод Зейделя збігається завжди [4]. Будь-яка лінійна система може бути приведена до системи з симметричною та додатньо-визначеною матрицею за допомогою множення лівої та правої частини системи на матрицю $\mathbf{A}^{\dagger}$ зліва, тобто система, що розв'язується зводиться до вигляду:

$$\mathbf{A}^{\dagger}\mathbf{A}\mathbf{X} = \mathbf{A}^{\dagger}\mathbf{B}.$$

Контрольні завдання 2.1

1. Вибрати СЛАР відповідно до свого варіанта.
2. Розв’язати вибрану СЛАР методом **LU**-факторизації.
3. Розв’язати вибрану СЛАР методом **QR**-факторизації.
4. Розв’язати вибрану СЛАР методом Холеського.
5. Розв’язати вибрану СЛАР методом Зейделя з похибкою не більше 1%.
6. Порівняти результати розв’язування різними методами. Оцінити відхил розв’язку.

Варіанти завдань

1. $\begin{cases} 19x_1 - 8x_2 - 4x_3 = -27; \\ -8x_1 + 16x_2 - 7x_3 = -6; \\ -4x_1 - 7x_2 + 11x_3 = 26. \end{cases}$

2. $\begin{cases} 16x_1 - 8x_2 - 2x_3 = -28; \\ -8x_1 + 15x_2 - 7x_3 = 9; \\ -2x_1 - 7x_2 + 9x_3 = 13. \end{cases}$

3. $\begin{cases} 16x_1 - 8x_2 - 2x_3 = -12; \\ -8x_1 + 19x_2 + -5x_3 = 9; \\ -2x_1 - 5x_2 + 7x_3 = 9. \end{cases}$

4. $\begin{cases} 16x_1 - 8x_2 - 2x_3 = -14; \\ -8x_1 + 14x_2 - 3x_3 = 5; \\ -2x_1 - 3x_2 + 5x_3 = 12. \end{cases}$

5. $\begin{cases} 16x_1 - 8x_2 - 2x_3 = -22; \\ -8x_1 + 9x_2 - x_3 = 15; \\ -2x_1 - x_2 + 3x_3 = 7. \end{cases}$

6. $\begin{cases} 18x_1 - 6x_2 - 9x_3 = -21; \\ -6x_1 + 15x_2 - 8x_3 = 0; \\ -9x_1 - 8x_2 + 17x_3 = 26. \end{cases}$

7. $\begin{cases} 12x_1 - 6x_2 - 4x_3 = -16; \\ -6x_1 + 18x_2 - 3x_3 = 18; \\ -4x_1 - 3x_2 + 7x_3 = 18. \end{cases}$

8. $\begin{cases} 12x_1 - 6x_2 - 4x_3 = -22; \\ -6x_1 + 15x_2 - 2x_3 = 31; \\ -4x_1 - 2x_2 + 6x_3 = 14. \end{cases}$

9. $\begin{cases} 12x_1 - 6x_2 - 4x_3 = -10; \\ -6x_1 + 12x_2 - x_3 = 20; \\ -4x_1 - x_2 + 5x_3 = 9. \end{cases}$

10. $\begin{cases} 18x_1 - 6x_2 - x_3 = 13; \\ -6x_1 + 15x_2 - 8x_3 = -7; \\ -x_1 - 8x_2 + 9x_3 = 19. \end{cases}$

11. $\begin{cases} 18x_1 - 4x_2 - 12x_3 = -40; \\ -4x_1 + 18x_2 - 12x_3 = 4; \\ -12x_1 - 12x_2 + 24x_3 = 48. \end{cases}$

12. $\begin{cases} 18x_1 - 4x_2 - 12x_3 = -22; \\ -4x_1 + 11x_2 - 6x_3 = 2; \\ -12x_1 - 6x_2 + 18x_3 = 30. \end{cases}$

13. $\begin{cases} 19x_1 - 4x_2 - 8x_3 = -7; \\ -4x_1 + 12x_2 - 5x_3 = 6; \\ -8x_1 - 5x_2 + 13x_3 = 34. \end{cases}$

14. $\begin{cases} 12x_1 - 4x_2 - 6x_3 = -20; \\ -4x_1 + 10x_2 - 3x_3 = 20; \\ -6x_1 - 3x_2 + 9x_3 = 21. \end{cases}$

15. $\begin{cases} 20x_1 - 4x_2 - 5x_3 = 30; \\ -4x_1 + 16x_2 - 12x_3 = -8; \\ -5x_1 - 12x_2 + 17x_3 = 22. \end{cases}$

16. $\begin{cases} 20x_1 - 4x_2 - 5x_3 = 25; \\ -4x_1 + 10x_2 - 3x_3 = 13; \\ -5x_1 - 3x_2 + 8x_3 = 21. \end{cases}$

17. $\begin{cases} 12x_1 - 3x_2 - 4x_3 = 2; \\ -3x_1 + 15x_2 - 4x_3 = 50; \\ -4x_1 - 4x_2 + 8x_3 = 16. \end{cases}$

18. $\begin{cases} 6x_1 - 3x_2 - 2x_3 = -2; \\ -3x_1 + 11x_2 - 8x_3 = -5; \\ -2x_1 - 8x_2 + 10x_3 = 12. \end{cases}$

19. $\begin{cases} 16x_1 - 2x_2 - 8x_3 = 4; \\ -2x_1 + 16x_2 - 8x_3 = 22; \\ -8x_1 - 8x_2 + 16x_3 = 40. \end{cases}$

20. $\begin{cases} 16x_1 - 2x_2 - 8x_3 = 2; \\ -2x_1 + 9x_2 - 4x_3 = 21; \\ -8x_1 - 4x_2 + 12x_3 = 28. \end{cases}$

21. $\begin{cases} 9x_1 - 2x_2 - 6x_3 = -8; \\ -2x_1 + 9x_2 - 6x_3 = 3; \\ -6x_1 - 6x_2 + 12x_3 = 18. \end{cases}$

22. $\begin{cases} 6x_1 - 2x_2 - 3x_3 = -5; \\ -2x_1 + 18x_2 - 9x_3 = 33; \\ -3x_1 - 9x_2 + 12x_3 = 27. \end{cases}$

23. $\begin{cases} 18x_1 - x_2 - 6x_3 = 46; \\ -x_1 + 18x_2 - 6x_3 = 84; \\ -6x_1 - 6x_2 + 12x_3 = 24. \end{cases}$

24. $\begin{cases} 8x_1 - x_2 - 4x_3 = 12; \\ -x_1 + 15x_2 - 8x_3 = 41; \\ -4x_1 - 8x_2 + 12x_3 = 16. \end{cases}$

25. $\begin{cases} 8x_1 - x_2 - 4x_3 = 8; \\ -x_1 + 8x_2 - 4x_3 = 17; \\ -4x_1 - 4x_2 + 8x_3 = 20. \end{cases}$

26. $\begin{cases} 8x_1 + 4x_2 + 8x_3 = 176; \\ 4x_1 + 7x_2 + 3x_3 = 117; \\ 8x_1 + 3x_2 + x_3 = 91. \end{cases}$

27. $\begin{cases} 3x_1 + 6x_2 + 2x_3 = 98; \\ 6x_1 + x_2 + 2x_3 = 77; \\ 2x_1 + 2x_2 + x_3 = 44. \end{cases}$

28. $\begin{cases} 3x_1 + 6x_2 + 2x_3 = 100; \\ 6x_1 + 9x_2 + 6x_3 = 195; \\ 2x_1 + 6x_2 + x_3 = 81. \end{cases}$

29. $\begin{cases} 4x_1 + 2x_2 + 4x_3 = 96; \\ 2x_1 + 5x_2 + 6x_3 = 132; \\ 4x_1 + 6x_2 + x_3 = 103. \end{cases}$

30. $\begin{cases} 6x_1 + 3x_2 + 2x_3 = 106; \\ 3x_1 + 5x_2 + 8x_3 = 165; \\ 2x_1 + 8x_2 + x_3 = 109. \end{cases}$

2.5. Зведення СЛАР з комплексними коефіцієнтами до СЛАР з дійсними коефіцієнтами

Розглянуті методи можуть бути застосовані для розв'язання СЛАР як з дійсними, так і з комплексними коефіцієнтами. Однак практична реалізація комплексного алгоритма на ЕОМ може бути утруднена через відсутність підтримки комплексної арифметики обраною мовою програмування. У такому випадку СЛАР з комплексними коефіцієнтами можна звести до СЛАР з дійсними коефіцієнтами [2, с. 265]. Нехай $\mathbf{AX} = \mathbf{B}$ – СЛАР з комплексними коефіцієнтами. Нехай

$$\mathbf{A} = \mathbf{A}_1 + i\mathbf{A}_2;$$
$$\mathbf{B} = \mathbf{B}_1 + i\mathbf{B}_2;$$
$$\mathbf{X} = \mathbf{X}_1 + i\mathbf{X}_2.$$

Тоді вихідна система рівносильна системі з дійсними коефіцієнтами

$$\mathbf{CY} = \mathbf{D},$$

де

$$\mathbf{C} = \begin{pmatrix} \mathbf{A}_1 & -\mathbf{A}_2 \\ \mathbf{A}_2 & \mathbf{A}_1 \end{pmatrix}; \quad \mathbf{D} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}; \quad \mathbf{Y} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}.$$

Цей підхід не є рівноцінною альтернативою розв'язанню СЛАР безпосередньо в комплексних числах через збільшення витрат. Так, трудоємність факторизації матриці системи з дійсними коефіцієнтами розмірності n оцінюється як $N = O\left(n^3\right)$ операцій з плаваючою комою [2, с. 253]. Факторизація системи такої ж розмірності, але з комплексними коефіцієнтами потребує в 4 рази більше дійсних арифметичних операцій, тобто $M = 4N$ [17, с. 233]. Разом з тим, зведення комплексної СЛАР до

системи з дійсними коефіцієнтами збільшує її розмірність в 2 рази. Таким чином, трудоємність її факторизації становить $K = O\left(\left(2n\right)^3\right) = 8N = 2M$. На практиці виграш від розв'язання систем безпосередньо в комплексних числах може бути меншим через повільнішу реалізацію комплексних операцій.

2.6. Методи розв'язання проблеми власних значень

Визначивши власні числа, з умови (2.3) можна знайти власні вектори. Однак, у разі великих розмірностей матриці **A**, складність розрахунку значно зростає. Тому розроблено ряд спеціальних методів, що полегшують розв'язання задачі на власні числа і власні вектори [8].

2.6.1. Степеневий метод

Під час розв'язання багатьох задач потрібне знання не всього спектра матриці, а тільки кількох власних чисел (найчастіше максимального чи мінімального за модулем) і відповідних власних векторів. Таку задачу називають частковою проблемою власних чисел.

Розглянемо типову задачу відшукання максимального за модулем власного числа. Нехай $\left\{\mathbf{\Psi}_i\right\}$ – система власних векторів, а $\{\lambda_i\}$ – система власних чисел матриці **A**:

$$\mathbf{A}\mathbf{\Psi}_i = \lambda_i \mathbf{\Psi}_i . \qquad (2.41)$$

Нехай власні числа розташовані з урахуванням їх кратності в порядку зростання:

$$\left|\lambda_1\right| \le \left|\lambda_2\right| \le \left|\lambda_3\right| \le \ldots \le \left|\lambda_{n-1}\right| \le \left|\lambda_n\right| .$$

Задамо деяке початкове наближення $\mathbf{X}_0$ до вектора $\mathbf{\Psi}_n$ і будемо послідовно обчислювати вектори

$$\mathbf{X}_k = \mathbf{A}\,\mathbf{X}_{k-1},\ k = 1, 2\ldots \qquad (2.42)$$

Оскільки система власних векторів $\{\mathbf{\Psi}_i\}$ повна, то $\mathbf{X}_0$ можна подати у вигляді

$$\mathbf{X}_0 = \sum_{i=1}^{n} c_i \mathbf{\Psi}_i,$$

де c_i – коефіцієнти розвинення. Тоді з (2.41), (2.42) маємо

$$\mathbf{X}_k = \sum_{i=1}^{n} c_i \mathbf{A}^k \mathbf{\Psi}_i = \sum_{i=1}^{n} c_i \lambda_i^k \mathbf{\Psi}_i. \qquad (2.43)$$

З огляду на те, що $|\lambda_n| > |\lambda_{n-1}|$, для досить великих k з (2.43) маємо

$$\mathbf{X}_k = \lambda_n^k \left(c_n \mathbf{\Psi}_n + O\left(\frac{\lambda_{n-1}}{\lambda_n} \right)^k \right). \qquad (2.44)$$

Із виразу (2.44) випливає, що якщо $c_n \neq 0$, то для досить великого k

$$|\lambda_n| \approx \frac{\|\mathbf{X}_k\|}{\|\mathbf{X}_{k-1}\|}.$$

Отже, максимальне за модулем власне число λ_n можна знайти з ітераційного процесу (2.42). Для цього модуль k-того наближення $\lambda_n^{(k)}$ до власного числа λ_n оцінюють як

$$\left|\lambda_n^{(k)}\right| = \frac{\|\mathbf{X}_k\|}{\|\mathbf{X}_{k-1}\|},$$

а k-те наближення до власного вектора $\mathbf{\Psi}_n^{(k)}$:

$$\mathbf{\Psi}_n^{(k)} = \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|}. \qquad (2.45)$$

Критерієм зупинки ітераційного процесу (2.42) є виконання умови

$$\frac{\left\|\left|\lambda_n^{(k)}\right| - \left|\lambda_n^{(k-1)}\right\|\right|}{\left|\lambda_n^{(k)}\right|} < \varepsilon,$$

де ε – задане значення відносної похибки.

Після виконання умови збіжності ітераційного процесу (2.42) власне число λ_n може бути знайдене за формулою:

$$\lambda_n = \frac{x_i^{(k)}}{x_i^{(k-1)}}, \quad \forall i:\ x_i^{(k-1)} \neq 0.$$

Із виразів (2.3) та (2.43) випливає, що швидкість збіжності ітераційного процесу лінійна і визначається відношенням $\left|\frac{\lambda_n}{\lambda_{n-1}}\right|$, оскільки

$$\lambda_n^{(k)} = \lambda_n + O\left(\frac{\lambda_{n-1}}{\lambda_n}\right)^k.$$

Для стійкості чисельного процесу необхідно час від часу нормувати вектор $\mathbf{X}_k$, щоб $\|\mathbf{X}_k\| = 1$. Інакше, якщо $|\lambda_n| > 1$, то, як випливає з виразу (2.44), $\|\mathbf{X}_k\| \to \infty$, якщо $k \to \infty$. Тому за досить великого k може відбутися переповнення розрядної сітки обчислювальної системи. Якщо $|\lambda_n| < 1$, то $\|\mathbf{X}_k\| \to 0$ і внаслідок скінченності розрядної сітки обчислювальної системи може статися, що починаючи з деякого k $\mathbf{X}_k = 0$.

Для пошуку мінімального за модулем власного числа використовують той факт, що мінімальне за модулем власне число матриці **A** дорівнює величині, оберненій до максимального за модулем власного числа оберненої матриці $\mathbf{A}^{-1}$ [5]. Це дозволяє обчислити максимальне за модулем власне число $\lambda_{\max\ inv}$ матриці $\mathbf{A}^{-1}$, а потім знайти шукане мінімальне за модулем власне число матриці **A** з виразу $\lambda_{\min} = 1/\lambda_{\max inv}$.

Проте обчислення оберненої матриці – складна операція, яка у разі чисельної реалізації призводить до значної похибки внаслідок округлення. Тому на практиці під час пошуку мінімального за модулем власного числа чергове наближення вектора $\mathbf{X}_k$ знаходять не з ітераційного процесу $\mathbf{X}_k = \mathbf{A}^{-1}\mathbf{X}_{k-1}$, а з розв'язку системи лінійних рівнянь

$$\mathbf{A}\mathbf{X}_k = \mathbf{X}_{k-1}.$$

Модуль k-того наближення $\lambda_{\min}^{(k)}$ до власного числа $\lambda_{\min}$ можна оцінити так:

$$\left|\lambda_{\min}^{(k)}\right| = 1 \Big/ \left|\lambda_{\max\, inv}^{(k)}\right| = \frac{\left\|\mathbf{X}_{k-1}\right\|}{\left\|\mathbf{X}_k\right\|}.$$

Описаний підхід називають методом зворотних ітерацій [18].

Приклад. Знайдемо максимальне і мінімальне власні числа та відповідні їм власні вектори для матриці $\mathbf{A} = \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix}$, яка має власні числа $\lambda_1 = 2$, $\lambda_2 = 20$ та відповідні їм власні вектори $\mathbf{\Psi}_1 = \begin{bmatrix} 1 & 1 \end{bmatrix}^{\mathrm{T}}, \mathbf{\Psi}_2 = \begin{bmatrix} 1 & -1 \end{bmatrix}^{\mathrm{T}}$. Спочатку знаходимо максимальне за модулем власне число та відповідний вектор.

Нехай $\mathbf{X}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Тоді згідно з виразом (2.42) маємо:

$$\mathbf{X}_1 = \mathbf{A}\mathbf{X}_0 = \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 11 \\ -9 \end{bmatrix},$$

$$\lambda_{\max}^{(1)} = \frac{\left\|\mathbf{X}_1\right\|_2}{\left\|\mathbf{X}_0\right\|_2} = \frac{\sqrt{11^2 + (-9)^2}}{\sqrt{1^2 + 0^2}} \approx 14{,}2127;$$

$$\mathbf{X}_2 = \mathbf{A}\mathbf{X}_1 = \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix} \times \begin{bmatrix} 11 \\ -9 \end{bmatrix} = \begin{bmatrix} 202 \\ -198 \end{bmatrix},$$

$$\lambda_{\max}^{(2)} = \frac{\|\mathbf{X}_2\|_2}{\|\mathbf{X}_1\|_2} = \frac{\sqrt{202^2 + (-198)^2}}{\sqrt{11^2 + (-9)^2}} \approx 19{,}9017;$$

$$\mathbf{X}_3 = \mathbf{A}\mathbf{X}_2 = \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix} \times \begin{bmatrix} 202 \\ -198 \end{bmatrix} = \begin{bmatrix} 4\,004 \\ -3\,996 \end{bmatrix},$$

$$\lambda_{\max}^{(3)} = \frac{\|\mathbf{X}_3\|_2}{\|\mathbf{X}_2\|_2} = \frac{\sqrt{4004^2 + (-3996)^2}}{\sqrt{202^2 + (-198)^2}} \approx 19{,}9990,$$

$$\varepsilon \approx \left| \frac{\lambda_{\max}^{(3)} - \lambda_{\max}^{(2)}}{\lambda_{\max}^{(3)}} \right| = \left| \frac{19{,}9990 - 19{,}9017}{19{,}9990} \right| \approx 0{,}0049\ ,$$

$$\boldsymbol{\Psi}_{\max} \approx \frac{\mathbf{X}_3}{\|\mathbf{X}_3\|_\infty} = \frac{\begin{bmatrix} 4\,004 \\ -3\,996 \end{bmatrix}}{4\,004} \approx \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Далі обчислимо мінімальне за модулем власне число і відповідний власний вектор.

Нехай $\mathbf{X}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Тоді:

$$\mathbf{A}\mathbf{X}_1 = \mathbf{X}_0 \Rightarrow \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix} \mathbf{X}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow \mathbf{X}_1 = \begin{bmatrix} 11/40 \\ 9/40 \end{bmatrix},$$

$$\lambda_{\min}^{(1)} = \frac{\|\mathbf{X}_0\|_2}{\|\mathbf{X}_1\|_2} = \frac{\sqrt{1^2 + 0^2}}{\sqrt{(11/40)^2 + (9/40)^2}} \approx 2{,}8144;$$

$$\mathbf{A}\mathbf{X}_2 = \mathbf{X}_1 \Rightarrow \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix} \mathbf{X}_2 = \begin{bmatrix} 11/40 \\ 9/40 \end{bmatrix} \Rightarrow \mathbf{X}_2 = \begin{bmatrix} 101/800 \\ 99/800 \end{bmatrix},$$

$$\lambda_{\min}^{(2)} = \frac{\|\mathbf{X}_1\|_2}{\|\mathbf{X}_2\|_2} = \frac{\sqrt{\left(11/40\right)^2 + \left(9/40\right)^2}}{\sqrt{\left(101/800\right)^2 + \left(99/800\right)^2}} \approx 2,0099;$$

$$\mathbf{A}\mathbf{X}_3 = \mathbf{X}_2 \Rightarrow \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix} \mathbf{X}_3 = \begin{bmatrix} 101/800 \\ 99/800 \end{bmatrix} \Rightarrow \mathbf{X}_3 = \begin{bmatrix} 1\,001/16\,000 \\ 999/16\,000 \end{bmatrix},$$

$$\lambda_{\min}^{(3)} = \frac{\|\mathbf{X}_2\|_2}{\|\mathbf{X}_3\|_2} = \frac{\sqrt{\left(101/800\right)^2 + \left(99/800\right)^2}}{\sqrt{\left(1\,001/16\,000\right)^2 + \left(999/16\,000\right)^2}} \approx 2,0001,$$

$$\varepsilon \approx \left| \frac{\lambda_{\min}^{(3)} - \lambda_{\min}^{(2)}}{\lambda_{\min}^{(3)}} \right| = \left| \frac{2,0001 - 2,0099}{2,0001} \right| \approx 0,0049,$$

$$\mathbf{\Psi}_{\min} \approx \frac{\mathbf{X}_3}{\|\mathbf{X}_3\|_\infty} = \frac{\begin{bmatrix} 1\,001/16\,000 \\ 999/16\,000 \end{bmatrix}}{1\,001/16\,000} \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

2.6.2. Метод скалярного добутку

Для ермітової матриці **A** всі власні числа є дійсними. Максимальне за модулем власне число доцільно шукати з ітераційного процесу [4]:

$$\mathbf{X}_k = \mathbf{A}\mathbf{X}_{k-1},\ \lambda_{\max}^{(k)} = \frac{\mathbf{X}_k^\dagger \mathbf{X}_k}{\mathbf{X}_{k-1}^\dagger \mathbf{X}_k}, \qquad \mathbf{\Psi}_{\max}^{(k)} = \frac{\mathbf{X}_k}{\|\mathbf{X}_k\|}, \tag{2.46}$$

де k – номер ітерації; $\lambda_{\max}^{(k)}$ – k-те наближення максимального за модулем власного числа; $\mathbf{\Psi}_{\max}^{(k)}$ – k-те наближення відповідного власного вектора. Для початку ітераційного процесу необхідно задати початкове наближення вектора $\mathbf{X}_0$.

Найменше за модулем власне число ермітової матриці **A** шукають з ітераційного процесу:

$$\mathbf{A}\mathbf{X}_k = \mathbf{X}_{k-1},\ \lambda_{\min}^{(k)} = \frac{\mathbf{X}_{k-1}^{\dagger}\mathbf{X}_k}{\mathbf{X}_k^{\dagger}\mathbf{X}_k}, \qquad \mathbf{\Psi}_{\min}^{(k)} = \frac{\mathbf{X}_k}{\left\|\mathbf{X}_k\right\|},$$

де k – номер ітерації; $\lambda_{\min}^{(k)}$ – k-те наближення мінімального за модулем власного числа, $\mathbf{\Psi}_{\min}^{(k)}$ – k-те наближення відповідного власного вектора.

Можна показати [4], що метод скалярного добутку має квадратичну збіжність, тобто $\lambda_n^{(k)} = \lambda_n + O\left(\frac{\lambda_{n-1}}{\lambda_n}\right)^{2k}$. Тому для його реалізації потрібно майже в два рази менше ітерацій, ніж для використання степеневого методу.

Приклад. Знайдемо максимальне і мінімальне власні числа та відповідні їм власні вектори для матриці з попереднього прикладу $\mathbf{A} = \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix}$. Спочатку знайдемо максимальне за модулем власне число та відповідний власний вектор.

Нехай $\mathbf{X}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Тоді згідно з (2.46) маємо:

$$\mathbf{X}_1 = \mathbf{A}\mathbf{X}_0 = \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 11 \\ -9 \end{bmatrix},$$

$$\lambda_{\max}^{(1)} = \frac{\mathbf{X}_1^{\dagger}\mathbf{X}_1}{\mathbf{X}_0^{\dagger}\mathbf{X}_1} = \frac{\begin{bmatrix} 11 & -9 \end{bmatrix} \times \begin{bmatrix} 11 \\ -9 \end{bmatrix}}{\begin{bmatrix} 1 & 0 \end{bmatrix} \times \begin{bmatrix} 11 \\ -9 \end{bmatrix}} = \frac{202}{11} = 18{,}3636;$$

$$\mathbf{X}_2 = \mathbf{A}\mathbf{X}_1 = \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix} \times \begin{bmatrix} 11 \\ -9 \end{bmatrix} = \begin{bmatrix} 202 \\ -198 \end{bmatrix},$$

$$\lambda_{\max}^{(2)} = \frac{\mathbf{X}_2^{\dagger}\mathbf{X}_2}{\mathbf{X}_1^{\dagger}\mathbf{X}_2} = \frac{\begin{bmatrix}202 & -198\end{bmatrix} \times \begin{bmatrix}202 \\ -198\end{bmatrix}}{\begin{bmatrix}11 & -9\end{bmatrix} \times \begin{bmatrix}202 \\ -198\end{bmatrix}} = \frac{80\,008}{4\,004} = 19,9820;$$

$$\mathbf{X}_3 = \mathbf{A}\mathbf{X}_2 = \begin{bmatrix}11 & -9 \\ -9 & 11\end{bmatrix} \times \begin{bmatrix}202 \\ -198\end{bmatrix} = \begin{bmatrix}4\,004 \\ -3\,996\end{bmatrix},$$

$$\lambda_{\max}^{(3)} = \frac{\mathbf{X}_3^{\dagger}\mathbf{X}_3}{\mathbf{X}_2^{\dagger}\mathbf{x}_3} = \frac{\begin{bmatrix}4\,004 & -3\,996\end{bmatrix} \times \begin{bmatrix}4\,004 \\ -3\,996\end{bmatrix}}{\begin{bmatrix}202 & -198\end{bmatrix} \times \begin{bmatrix}4\,004 \\ -3\,996\end{bmatrix}} = \frac{3\,200\,032}{1\,600\,016} = 19,9998;$$

$$\varepsilon \approx \left| \frac{\lambda_{\max}^{(3)} - \lambda_{\max}^{(2)}}{\lambda_{\max}^{(3)}} \right| = \left| \frac{19,9998 - 19,9820}{19,9998} \right| \approx 0,00089,$$

$$\mathbf{\Psi}_{\max} \approx \frac{\mathbf{X}_3}{\|\mathbf{X}_3\|_{\infty}} = \frac{\begin{bmatrix}4\,004 \\ -3\,996\end{bmatrix}}{4\,004} \approx \begin{bmatrix}1 \\ -1\end{bmatrix}.$$

Далі обчислимо мінімальне за модулем власне число і відповідний власний вектор.

Нехай $\mathbf{X}_0 = \begin{bmatrix}1 \\ 0\end{bmatrix}$. Тоді:

$$\mathbf{A}\mathbf{X}_1 = \mathbf{X}_0 \Rightarrow \begin{bmatrix}11 & -9 \\ -9 & 11\end{bmatrix}\mathbf{X}_1 = \begin{bmatrix}1 \\ 0\end{bmatrix} \Rightarrow \mathbf{X}_1 = \begin{bmatrix}11/40 \\ 9/40\end{bmatrix},$$

$$\lambda_{\min}^{(1)} = \frac{\mathbf{X}_0^{\dagger}\mathbf{X}_1}{\mathbf{X}_1^{\dagger}\mathbf{X}_1} = \frac{\begin{bmatrix}1 & 0\end{bmatrix} \times \begin{bmatrix}11/40 \\ 9/40\end{bmatrix}}{\begin{bmatrix}11/40 & 9/40\end{bmatrix} \times \begin{bmatrix}11/40 \\ 9/40\end{bmatrix}} = \frac{11/40}{101/800} \approx 2,1782;$$

$$\mathbf{A}\mathbf{X}_2 = \mathbf{X}_1 \Rightarrow \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix} \mathbf{X}_2 = \begin{bmatrix} 11/40 \\ 9/40 \end{bmatrix} \Rightarrow \mathbf{X}_2 = \begin{bmatrix} 101/800 \\ 99/800 \end{bmatrix},$$

$$\lambda_{\min}^{(2)} = \frac{\mathbf{X}_1^{\dagger}\mathbf{X}_2}{\mathbf{X}_2^{\dagger}\mathbf{X}_2} = \frac{\begin{bmatrix} 11/40 & 9/40 \end{bmatrix} \times \begin{bmatrix} 101/800 \\ 99/800 \end{bmatrix}}{\begin{bmatrix} 101/800 & 99/800 \end{bmatrix} \times \begin{bmatrix} 101/800 \\ 99/800 \end{bmatrix}} = \frac{2\,002/32\,000}{20\,002/640\,000} \approx 2{,}0018;$$

$$\mathbf{A}\mathbf{X}_3 = \mathbf{X}_2 \Rightarrow \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix} \mathbf{X}_3 = \begin{bmatrix} 101/800 \\ 99/800 \end{bmatrix} \Rightarrow \mathbf{X}_3 = \begin{bmatrix} 1\,001/16\,000 \\ 999/16\,000 \end{bmatrix},$$

$$\lambda_{\min}^{(3)} = \frac{\mathbf{X}_2^{\dagger}\mathbf{X}_3}{\mathbf{X}_3^{\dagger}\mathbf{X}_3} = \frac{\begin{bmatrix} 101/800 & 99/800 \end{bmatrix} \times \begin{bmatrix} 1\,001/16\,000 \\ 999/16\,000 \end{bmatrix}}{\begin{bmatrix} 1\,001/16\,000 & 999/16\,000 \end{bmatrix} \times \begin{bmatrix} 1\,001/16\,000 \\ 999/16\,000 \end{bmatrix}} =$$

$$= \frac{200\,002/12\,800\,000}{2\,000\,002/256\,000\,000} \approx 2{,}0000;$$

$$\varepsilon \approx \left| \frac{\lambda_{\min}^{(3)} - \lambda_{\min}^{(2)}}{\lambda_{\min}^{(3)}} \right| = \left| \frac{2{,}0000 - 2{,}0018}{2{,}0000} \right| = 0{,}0009,$$

$$\boldsymbol{\Psi}_{\min} \approx \frac{\mathbf{X}_3}{\|\mathbf{X}_3\|_{\infty}} = \frac{\begin{bmatrix} 1\,001/16\,000 \\ 999/16\,000 \end{bmatrix}}{1\,001/16\,000} \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

2.6.3. Метод Віландта

Часто в практичних задачах виникає необхідність знаходження власного числа, яке є найближчим до деякого заданого числа μ. У такому випадку можна розв'язати задачу методом зворотних ітерацій зі зсувом [14, с. 191] (також відомий як метод Віландта [19, с. 100]). Суть методу полягає в тому, що спочатку знаходять деяку додаткову матрицю $\tilde{\mathbf{A}} = \left(\mathbf{A} - \mu\mathbf{E}\right)^{-1}$, де $\mathbf{E}$ – одинична матриця. Після цього визначають максимальне за модулем власне число матриці ν. Це можна зробити будь-яким відомим методом (наприклад, степеневим методом або методом скалярних добутків, якщо матриця **A** ермітова). Остаточно шукане власне число λ обчислюють за формулою

$$\lambda = \mu + \frac{1}{\nu}.$$

Слід зазначити, що на практиці немає потреби шукати обернену матрицю для обчислення $\tilde{\mathbf{A}}$, а можна скористатися методом зворотних ітерацій.

Приклад. Знайдемо власне число матриці $\mathbf{A} = \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix}$, яке є найближчим до числа μ = 4.

$$\tilde{\mathbf{A}}^{-1} = \left(\mathbf{A} - \mu\mathbf{E}\right) = \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix} - 4\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 7 & -9 \\ -9 & 7 \end{bmatrix}; \mathbf{X}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

$$\left(\mathbf{A} - \mu\mathbf{E}\right)\mathbf{X}_1 = \mathbf{X}_0 \Rightarrow \begin{bmatrix} 7 & -9 \\ -9 & 7 \end{bmatrix}\mathbf{X}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow \mathbf{X}_1 = \begin{bmatrix} -\frac{7}{32} \\ -\frac{9}{32} \end{bmatrix}$$

$$\nu^{(1)} = \frac{\mathbf{X}_1^{\dagger}\mathbf{X}_1}{\mathbf{X}_0^{\dagger}\mathbf{X}_1} = \frac{\begin{bmatrix} -7/32 & -9/32 \end{bmatrix} \times \begin{bmatrix} -\frac{7}{32} \\ -\frac{9}{32} \end{bmatrix}}{\begin{bmatrix} 1 & 0 \end{bmatrix} \times \begin{bmatrix} -\frac{7}{32} \\ -\frac{9}{32} \end{bmatrix}} = \frac{130/1024}{-7/32} \approx -0,5804\,.$$

$$\left(\mathbf{A} - \mu\mathbf{E}\right)\mathbf{X}_2 = \mathbf{X}_1 \Rightarrow \begin{bmatrix} 7 & -9 \\ -9 & 7 \end{bmatrix}\mathbf{X}_2 = \begin{bmatrix} -\frac{7}{32} \\ -\frac{9}{32} \end{bmatrix} \Rightarrow \mathbf{X}_2 = \begin{bmatrix} \frac{65}{512} \\ \frac{63}{512} \end{bmatrix}$$

$$\nu^{(2)} = \frac{\mathbf{X}_2^{\dagger}\mathbf{X}_2}{\mathbf{X}_1^{\dagger}\mathbf{X}_2} = \frac{\begin{bmatrix} 65/512 & 63/512 \end{bmatrix} \times \begin{bmatrix} \frac{65}{512} \\ \frac{63}{512} \end{bmatrix}}{\begin{bmatrix} -7/32 & -9/32 \end{bmatrix} \times \begin{bmatrix} \frac{65}{512} \\ \frac{63}{512} \end{bmatrix}} = \frac{4\,097/131\,072}{-511/8\,192} \approx -\,0,5011.$$

$$\left(\mathbf{A} - \mu\mathbf{E}\right)\mathbf{X}_3 = \mathbf{X}_2 \Rightarrow \begin{bmatrix} 7 & -9 \\ -9 & 7 \end{bmatrix}\mathbf{X}_3 = \begin{bmatrix} \frac{65}{512} \\ \frac{63}{512} \end{bmatrix} \Rightarrow \mathbf{X}_3 = \begin{bmatrix} -\frac{511}{8192} \\ -\frac{513}{8192} \end{bmatrix}$$

$$\nu^{(3)} = \frac{\mathbf{X}_3^{\dagger}\mathbf{X}_3}{\mathbf{X}_2^{\dagger}\mathbf{X}_3} = \frac{\begin{bmatrix} -511/8192 & -513/8192 \end{bmatrix} \times \begin{bmatrix} -\dfrac{511}{8192} \\ -\dfrac{513}{8192} \end{bmatrix}}{\begin{bmatrix} 65/512 & 63/512 \end{bmatrix} \times \begin{bmatrix} -\dfrac{511}{8192} \\ -\dfrac{513}{8192} \end{bmatrix}} =$$

$$= \frac{262\,145/33\,554\,432}{-32\,767/2\,097\,152} \approx -0,5000, \quad \lambda = \mu + \frac{1}{\nu} \approx \mu + \frac{1}{\nu^{(3)}} = 4 + \frac{1}{-0,5000} = 2.$$

Контрольне завдання 2.2

1. Вибрати матрицю **A** як матрицю коефіцієнтів СЛАР контрольного завдання 2.1 відповідно до свого варіанта.

2. Знайти максимальне і мінімальне власні числа матриці та відповідні власні вектори степеневим методом та методом скалярного добутку з похибкою не більше 1 %.

3. Знайти власне число матриці, найближче до заданого числа μ з похибкою не більше 1 %. Значення числа μ вибрати з табл. 2.1 відповідно до варіанта.

4. Оцінити число обумовленості матриці.

5. Порівняти результати розв'язку різними методами.

Таблиця 2.1. **Значення числа μ**

Варіант	μ	Варіант	μ	Варіант	μ	Варіант	μ	Варіант	μ
1	18	2	14	3	13	4	10	5	4
6	22	7	13	8	11	9	9	10	16
11	21	12	16	13	16	14	12	15	22
16	12	17	14	18	8	19	17	20	11
21	10	22	9	23	18	24	10	25	8
26	–6	27	–4	28	–2	29	–4	30	–5

3. Чисельні методи розв'язання нелінійних рівнянь

Нехай $f(x)$ – функція дійсного чи комплексного аргументу. Задача розв'язку нелінійного рівняння полягає в тому, щоб знайти один чи більше нулів (коренів) функції $f(x)$:

$$f(x) = 0. \qquad (3.1)$$

Рівняння (3.1), як правило, не має аналітичного розв'язку, і його розв'язують чисельними методами [10,6]. Задачу знаходження коренів рівняння (3.1) розв'язують у два етапи.

На першому етапі вивчають розташування коренів (у загальному випадку – на комплексній площині) і проводять їх розділення, тобто виділяють області на комплексній площині, що містять тільки один корінь. Цей етап важко формалізувати. Якщо $f(x)$ є дійсною функцією дійсного аргументу, то найпростіший спосіб розв'язання задачі на цьому етапі – обчислення таблиці значень функції $f(x)$ у заданих точках $x_i \in [a,b], i = 0,1,\ldots,k$. Якщо виявиться, що за деякого i числа $f(x_i)$, $f(x_{i+1})$ мають різні знаки і функція $f(x)$ неперервна на відрізку $[x_i, x_{i+1}]$, то це означає, що на цьому відрізку рівняння (3.1) має, принаймні, один дійсний корінь. Якщо похідна монотонна, то на вказаному проміжку існує лише один корінь. Потім відрізок $[x_i, x_{i+1}]$ розбивають на дрібніші відрізки і за допомогою аналогічної процедури уточнюють розташування кореня. Тим самим знаходять деякі початкові наближення для коренів рівняння (3.1).

На другому етапі, для уточнення значення шуканого кореня, будують ітераційний процес, тобто послідовність $x_0, x_1, \ldots$, причому кожне нове

значення аргументу обчислюють на підставі $m \geq 1$ попередніх, тобто в загальному випадку

$$x_{k+1} = \psi_k (x_{k-m+1}, x_{k-m+2}, ..., x_k), \; k \geq m\text{-}1, \tag{3.2}$$

де ψ_k – деяка функція, що залежить від методу розв'язання. Ітераційний процес (3.2) називають m-кроковим [20]. Для початку ітераційного процесу (3.2) необхідно знати m попередніх наближень: $x_0, x_1, \ldots, x_{m-1}$. Початкове наближення x_0 задається з тих чи інших міркувань в залежності від характеру функції $f(x)$ та сенсу невідомого x. Так, наприклад, якщо x є напругою в якомусь вузлі електронної схеми, то x_0 може бути деяка величина між нулем та напругою живлення схеми. Якщо $m>1$, то наближення $x_1, \ldots, x_{m-1}$, як правило знаходять l-кроковим методом, де $l<m$.

Поширеною обчислювальною задачею є знаходження окремих чи всіх розв'язків системи n нелінійних рівнянь [21]:

$$\begin{cases} f_1 (x_1, x_2, ..., x_n) = 0; \\ f_2 (x_1, x_2, ..., x_n) = 0; \\ \quad \ldots \\ f_n (x_1, x_2, ..., x_n) = 0. \end{cases} \tag{3.3}$$

Позначимо через $\mathbf{X}$ вектор-стовпчик $\mathbf{X} = [x_1, x_2, \ldots, x_n]^T$, а вектор-стовпець функцій $[f_1, f_2, ..., f_n]^T$ через $\mathbf{F}$. Тоді систему (3.3) можна подати у векторній формі

$$\mathbf{F}(\mathbf{X}) = 0. \tag{3.4}$$

На відміну від систем лінійних рівнянь не існує прямих методів розв'язування нелінійних систем. Тому для розв'язання системи рівнянь (3.4), як і у випадку одного рівняння (3.1), будують m-кроковий ітераційний процес

$$\mathbf{X}_{k+1} = \mathbf{\Psi}_k(\mathbf{X}_{k-m+1}, \mathbf{X}_{k-m+2}, ..., \mathbf{X}_k), \; k \geq m\text{-}1,$$

де $\mathbf{\Psi}_k$ – вектор-стовпець функцій, що залежить від методу розв'язування; $\mathbf{X}_i$ – i-те наближення до розв'язку.

Критерієм зупинки ітераційного процесу є збіжність за аргументом:

$$\left\|\Delta \mathbf{X}_k\right\| = \left\|\mathbf{X}_{k+1} - \mathbf{X}_k\right\| \le \Delta, \text{ чи} \tag{3.5}$$

$$\frac{\left\|\Delta \mathbf{X}_k\right\|}{\left\|\mathbf{X}_{k+1}\right\|} \le \varepsilon, \tag{3.6}$$

де Δ та ε – задані абсолютна та відносна похибки розв'язку, або збіжність за значенням функції:

$$\left\|\mathbf{F}\left(\mathbf{X}_{k+1}\right)\right\| \le \delta, \tag{3.7}$$

де δ – задана абсолютна нев'язка розв'язку. На практиці корисно поєднувати перевірки за всіма критеріями.

3.1. Метод бісекції (поділу навпіл, дихотомії)

Нехай $f(x)$ – дійсна функція однієї змінної x. Щоб почати пошук нуля $f(x)$, припустимо, що можна знайти відрізок $[a, b]$, на якому $f(x)$ змінює знак. У цьому випадку, якщо $f(x)$ неперервна, то має на $[a, b]$ принаймні один корінь. Слід зазначити, що оскільки розрядна сітка обчислювальної системи скінченна, то функція, що обчислюється, набуває лише дискретну множину значень, серед яких існує малий відрізок $[\alpha, \beta]$, на якому $f(x)$ змінює знак. Такий відрізок можна знайти і звузити настільки, наскільки дозволяє система чисел з плаваючою крапкою, тобто так, щоб кінцями цього відрізку були два сусідні числа цієї системи.

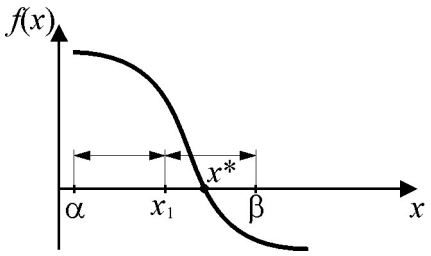


Рис. 3.1. Пошук кореня методом бісекції

У методі бісекції точка x_k розміщується у центрі відрізку $[\alpha, \beta]$ (рис.3.1) і як новий відрізок вибирають той з відрізків $[\alpha, x_k]$ і $[x_k, \beta]$, який містить корінь. Поділ відрізка навпіл продовжують доти, доки довжина відрізка не буде меншою від заданої похибки обчислення. За корінь вибирають середину відрізку, що залишився. Таким чином, якщо відомі такі a і b, що $f(a)f(b) < 0$, то за заданої відносної похибки ε для визначення кореня $f(x)$ методу бісекції використовують алгоритм:

1) покласти $\alpha = a$, $\beta = b$;

2) покласти $x_k = \dfrac{\alpha + \beta}{2}$;

3) якщо $f(\alpha)f(x_k) < 0$, то $\beta = x_k$, інакше $\alpha = x_k$;

4) якщо $|\beta - \alpha| > \varepsilon \left|\dfrac{\alpha + \beta}{2}\right|$, перейти до кроку 2, інакше $x^* \approx \dfrac{\alpha + \beta}{2}$;

кінець.

У разі машинної реалізації цього алгоритму слід ураховувати проблему машинних нулів і переповнень. Так, наприклад, якщо значення функції будуть дуже великими, то на кроці 3 у перевірці $f(\alpha) f(x_k) < 0$ відбувається переповнення і зупинка виконання програми. Якщо ж значення функції дуже малі, то перевірку буде виконано неправильно і відбудеться збій у роботі алгоритму. Тому на кроці 3 замість наведеної перевірки слід використовувати умову $\dfrac{f(\alpha)}{|f(\alpha)|} f(x_k) < 0$.

Метод бісекції працює теоретично завжди [22]. Тому його відносять до класу методів, які збігаються глобально. Дійсно, якщо тільки виконана

умова $f(\alpha)f(x_k)<0$ і $f(x)$ неперервна, то незалежно від поводження $f(x)$ метод бісекції теоретично завжди дозволить визначити значення кореня із заданою похибкою.

3.2. Метод Ньютона розв'язання рівнянь з однією змінною

Нехай x_k – k-те наближення розв'язання задачі (3.1). Розкладемо функцію $f(x)$ у ряд Тейлора в околі точки $x = x_k$ і обмежимося першими двома членами ряду. Тоді

$$f(x) \approx f(x_k) + (x - x_k) f'(x_k). \qquad (3.8)$$

Відповідно до виразу (3.1), прирівнюючи вираз (3.8) до нуля, знаходимо нове наближення до розв'язку

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}. \qquad (3.9)$$

Метод Ньютона будується за ітераційною формулою (3.9) [5,6].

Для установлення геометричного змісту методу Ньютона відзначимо, що вираз (3.8) є рівнянням дотичної до графіка функції $y = f(x)$ у точці x_k. Отже, відповідно до формули (3.9) нове наближення x_{k+1} шукається як точка перетину дотичної до $y = f(x)$ у точці x_k з віссю Ox (рис. 3.2). Тому метод Ньютона іноді називають методом дотичних.

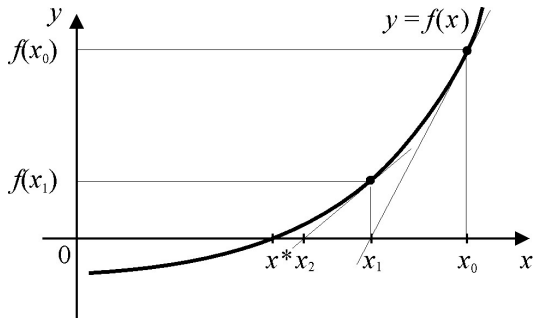


Рис. 3.2. Пошук кореня методом Ньютона

Можна довести, що метод Ньютона збігається, якщо [6]:

$$\left|x_0 - x^*\right| c \le 1, \tag{3.10}$$

де c характеризує ступінь відносної нелінійності функції $f(x)$ і визначається як $c = \dfrac{\sup\limits_x \left|f''(x)\right|}{2\inf\limits_x \left|f'(x)\right|}$, $\inf\limits_G \left(f\left(x\right)\right)$ – найменше значення функції $f(x)$ в області G, x^*– точний розв'язок рівняння (3.1).

Як випливає з виразу (3.10), у разі невдалого вибору початкового наближення метод Ньютона може не збігатися. Тому його необхідно включати в більш надійний метод, який успішно працював би і за віддаленіших початкових точок. Тому методу Ньютона часто передує який-небудь алгоритм типу бісекції, який сходиться глобально. Потім можна переключатися на ітерації методу Ньютона, які швидко збігаються.

Як приклад такого підходу розглянемо наступний метод. Припустимо, що метод Ньютона генерує не тільки крок $x_{k+1} = x_k - \dfrac{f\left(x_k\right)}{f'\left(x_k\right)}$, але й напрям, на який цей крок указує (передбачається, що $f'\left(x_k\right) \ne 0$). Хоча ньютонівський крок може призвести до збільшення абсолютного значення функції, його напрям такий, що вздовж нього абсолютне значення функції зменшується (рис. 3.3).

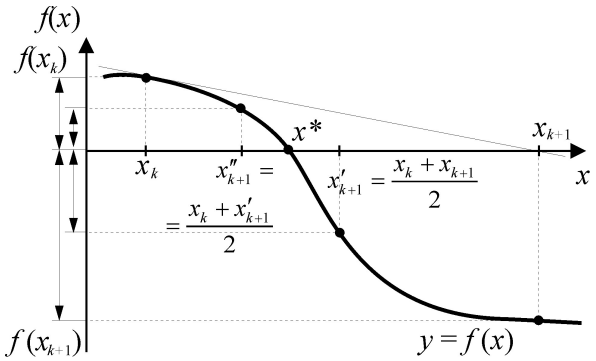


Рис. 3.3. Лінійний пошук

Тоді, якщо ньютонівська точка x_{k+1} не приводить до зменшення $|f(x)|$, то розумна стратегія полягає в дробленні кроку з рухом у зворотньому напрямі від x_{k+1} до x_k, доки не зустрінеться така точка x''_{k+1}, для якої $|f(x''_{k+1})| < |f(x_k)|$ [6]. Ітерація тут може бути такою:

1. Обчислити $x_{k+1} = x_k - \dfrac{f(x_k)}{f'(x_k)}$.
2. Якщо $|f(x_{k+1})| \geq |f(x_k)|$, то перейти до кроку 3; інакше перейти до наступної ньютонівської ітерації, тобто, якщо не виконується критерій зупинки ітераційного процесу, то перейти до кроку 1.
3. Обчислити $x_{k+1} = \dfrac{x_k + x_{k+1}}{2}$. Перейти до кроку 2.

Розглянутий метод є прикладом гібридного алгоритму, у якому робиться спроба поєднати глобальну і швидку локальну збіжності, перевіряючи спочатку на кожній ітерації ньютонівський крок, але завжди домагаючись, щоб у результаті ітерації поліпшувалася деяка міра близькості до розв'язку.

3.3. Квазіньютонівські методи розв'язання рівнянь з однією змінною

У багатьох практичних задачах $f(x)$ не задається формулою, а є результатом деякої обчислювальної чи експериментальної процедури. У таких випадках значення похідної $f'(x)$ недоступне. Крім цього, часто затрати на обчислення $f'(x)$ можуть виявитися набагато більшими, ніж обчислення $f(x)$. Тому метод Ньютона для практичного застосування потрібно модифікувати. Групу методів, що використовують апроксимуючий вираз замість $f'(x)$, називають *квазіньютонівськими* [6].

У скінченнорізницевому методі Ньютона похідна $f'(x)$ апроксимується виразом

$$f'(x) \approx a_k = \frac{f(x_k + h_k) - f(x_k)}{h_k}.$$

Тоді квазіньютонівський ітераційний крок має вигляд

$$x_{k+1} = x_k - \frac{f(x_k)}{a_k}. \tag{3.11}$$

Для збільшення швидкості збіжності скінченнорізницевого методу Ньютона необхідно зменшувати h_k. Однак на практиці, через наявність арифметики скінченної точності, значення h_k обмежено знизу [12]. Розумний компроміс полягає в тому, щоб збалансувати похибку дискретизації, пов'язану з вибором занадто великих h_k з похибками арифметики скінченної точності. Тому на практиці значення h_k вибирають таким, щоб внести зміни приблизно в половину розрядів мантиси x_k [6]:

$$|h_k| = \sqrt{\varepsilon_{\text{маш}}}\, \max\left(|x_{\text{тип}}|, |x_k|\right), \tag{3.12}$$

де $\varepsilon_{\text{маш}}$ – відносна похибка округлення дійсних чисел; $x_{\text{тип}}$ – характерне значення змінної x.

Недоліком скінченнорізницевого методу Ньютона є обчислення на одному кроці відразу двох значень функції $f(x)$. Якщо обчислення $f(x)$ виявляється трудомістким, то додаткове обчислення функції є небажаним. У цьому випадку h_k вважається рівним $x_k - x_{k-1}$, тому

$$a_k = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$

Тоді з виразу (3.11)

$$x_{k+1} = x_k - \left(x_k - x_{k-1}\right)\frac{f\left(x_k\right)}{f\left(x_k\right) - f\left(x_{k-1}\right)}. \qquad (3.13)$$

Таким чином, у ітераційному процесі (3.13) на кожному кроці x_{k+1} одержують з x_k і x_{k-1} як єдиний нуль лінійної функції, що набуває значення $f(x_k)$ у x_k і $f(x_{k-1})$ у x_{k-1}. Ця лінійна функція є січною до кривої $y = f(x)$, що проходить через її точки з абсцисою x_k і x_{k-1}. Тому цей метод називають *методом січних* (рис. 3.4) [20].

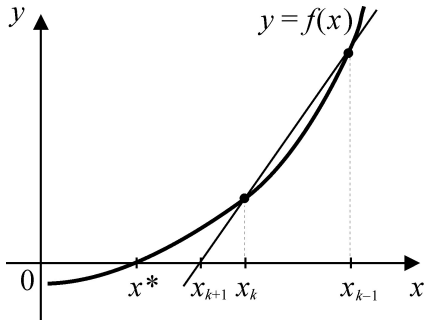


Рис. 3.4. Метод січних

Метод січних належить до двокрокових ітераційних методів, оскільки нове наближення до кореня залежить від двох попередніх наближень (формула (3.13)). Оскільки на початковій ітерації методу січних також потрібні знання двох початкових наближень і значень функції в них, то, як правило, методу січних передує одна ітерація однокрокового методу, наприклад, скінченнорізницевого методу Ньютона. Метод січних збігається трохи повільніше, ніж метод Ньютона і скінченнорізницевий метод. Однак, завдяки необхідності обчислення на кожній ітерації лише одного значення функції, метод січних може перевершувати їх за обчислювальною ефективністю.

3.4. Метод Ньютона розв'язання систем нелінійних рівнянь

Розглянемо систему нелінійних рівнянь (3.3). Нехай $\mathbf{X}_k = \left[x_1^{(k)},\, x_2^{(k)},\ldots,\, x_n^{(k)} \right]^{\mathrm{T}}$ – k-те наближення до розв'язку задачі (3.3).

Розкладемо кожну функцію $f_i(x_1,\, x_2,\ldots,\, x_n),\ i = \overline{1,n}$ у ряд Тейлора в околі точки $\mathbf{X} = \mathbf{X}_k$ і обмежимося лінійними членами рядів. Тоді замість системи (3.3) маємо наближену систему рівнянь [21]:

$$f_i(\mathbf{X}_k) + \sum_{j=1}^{n} \left(x_j - x_j^{(k)} \right) \frac{\partial f_j(\mathbf{X}_k)}{\partial x_i} = 0\ ,\ i = \overline{1,n}\,, \qquad (3.14)$$

лінійну відносно $\left(x_j - x_j^{(k)} \right)$. Розв'язок системи (3.14) вважатимемо за наступне наближення до розв'язку $\mathbf{X}_{k+1} = \left[x_1^{(k+1)},\, x_2^{(k+1)},\ldots,\, x_n^{(k+1)} \right]$. Тоді

$$f_i(\mathbf{X}_k) + \sum_{j=1}^{n} \left(x_j^{(k+1)} - x_j^{(k)} \right) \frac{\partial f_j(\mathbf{X}_k)}{\partial x_i} = 0\ ,\ i = \overline{1,n}\,. \qquad (3.15)$$

Систему (3.15) записують у векторній формі

$$\mathbf{F}(\mathbf{X}_k) + \mathbf{J}(\mathbf{X}_k) \Delta \mathbf{X}_k = 0\,,$$

де $\Delta \mathbf{X}_k = \mathbf{X}_{k+1} - \mathbf{X}_k$, $\mathbf{J}(\mathbf{X}_k)$ – матриця Якобі

$$\mathbf{J}(\mathbf{X}_k) = \begin{bmatrix} \dfrac{\partial f_1(\mathbf{X}_k)}{\partial x_1} & \dfrac{\partial f_1(\mathbf{X}_k)}{\partial x_2} & \cdots & \dfrac{\partial f_1(\mathbf{X}_k)}{\partial x_n} \\ \dfrac{\partial f_2(\mathbf{X}_k)}{\partial x_1} & \dfrac{\partial f_2(\mathbf{X}_k)}{\partial x_2} & \cdots & \dfrac{\partial f_2(\mathbf{X}_k)}{\partial x_n} \\ & \cdots & & \\ \dfrac{\partial f_n(\mathbf{X}_k)}{\partial x_1} & \dfrac{\partial f_n(\mathbf{X}_k)}{\partial x_2} & \cdots & \dfrac{\partial f_n(\mathbf{X}_k)}{\partial x_n} \end{bmatrix}.$$

Тоді процес розв'язання нелінійної системи подають у вигляді ітераційної процедури [6]:

$$\mathbf{J}_k\left(\mathbf{X}_k\right)\Delta\mathbf{X}_k = -\mathbf{F}\left(\mathbf{X}_k\right);$$
$$\mathbf{X}_{k+1} = \mathbf{X}_k + \Delta\mathbf{X}_k.$$

Можна показати, що для метода Ньютона розв'язання СНР справедлива оцінка

$$\left\|\mathbf{X}_{k+1} - \mathbf{X}^*\right\| \le c\left\|\mathbf{X}_k - \mathbf{X}^*\right\|^2, \qquad (3.16)$$

де $c = \sup_{\mathbf{X}}\left\|\mathbf{J}(\mathbf{X})^{-1}\right\| \sup_{\mathbf{X}}\left\|\mathbf{H}(\mathbf{X})\right\|$, $\mathbf{H}\left(\mathbf{X}\right)$ — матриця Гессе, $\mathbf{X}^*$— точний розв'язок системи (3.4). Таким чином, якщо матриця Якобі — невироджена, а другі похідні — неперервні, то за умови достатньої близькості початкового наближення до кореня, метод Ньютона має квадратичну збіжність.

Для прикладу знайдемо розв'язок системи рівнянь $\begin{cases} x_1 + x_2 - 3 = 0, \\ x_1^2 + x_2^2 - 9 = 0. \end{cases}$

Для даної системи

$$\mathbf{F}\left(\mathbf{X}\right) = \begin{bmatrix} x_1 + x_2 - 3 \\ x_1^2 + x_2^2 - 9 \end{bmatrix},\ \mathbf{J}\left(\mathbf{X}\right) = \begin{bmatrix} 1 & 1 \\ 2x_1 & 2x_2 \end{bmatrix}.$$

Виберемо початкове наближення $\mathbf{X}_0 = \begin{bmatrix} 1 & 5 \end{bmatrix}^{\mathrm{T}}$. Для оцінювання похибки у (3.6) та (3.7) використаємо евклідові норми $\left\|\mathbf{X}\right\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$. Тоді маємо:

$$\mathbf{J}(\mathbf{X}_0)\Delta\mathbf{X}_0 = -\mathbf{F}(\mathbf{X}_0) \Rightarrow \begin{bmatrix} 1 & 1 \\ 2 & 10 \end{bmatrix}\Delta\mathbf{X}_0 = -\begin{bmatrix} 3 \\ 17 \end{bmatrix} \Rightarrow \Delta\mathbf{X}_0 = \begin{bmatrix} -\frac{13}{8} \\ -\frac{11}{8} \end{bmatrix} \Rightarrow$$

$$\Rightarrow \mathbf{X}_1 = \begin{bmatrix} 1 \\ 5 \end{bmatrix} + \begin{bmatrix} -\frac{13}{8} \\ -\frac{11}{8} \end{bmatrix} = \begin{bmatrix} -0,625 \\ 3,625 \end{bmatrix}, \quad \frac{\|\Delta\mathbf{X}_0\|}{\|\mathbf{X}_1\|} \approx \frac{2,13}{3,7} \approx 0,58;$$

$$\mathbf{J}(\mathbf{X}_1)\Delta\mathbf{X}_1 = -\mathbf{F}(\mathbf{X}_1) \Rightarrow \begin{bmatrix} 1 & 1 \\ -\frac{5}{4} & \frac{29}{4} \end{bmatrix}\Delta\mathbf{X}_1 = -\begin{bmatrix} 0 \\ \frac{145}{32} \end{bmatrix} \Rightarrow \Delta\mathbf{X}_1 = \begin{bmatrix} \frac{145}{272} \\ -\frac{145}{272} \end{bmatrix} \Rightarrow$$

$$\Rightarrow \mathbf{X}_2 = \begin{bmatrix} -0,625 \\ 3,625 \end{bmatrix} + \begin{bmatrix} \frac{145}{272} \\ -\frac{145}{272} \end{bmatrix} = \begin{bmatrix} -\frac{25}{272} \\ \frac{841}{272} \end{bmatrix}, \ \frac{\|\Delta\mathbf{X}_1\|}{\|\mathbf{X}_2\|} \approx 0,24;$$

$$\mathbf{J}(\mathbf{X}_2)\Delta\mathbf{X}_2 = -\mathbf{F}(\mathbf{X}_2) \Rightarrow \begin{bmatrix} 1 & 1 \\ -\frac{25}{136} & \frac{841}{136} \end{bmatrix}\Delta\mathbf{X}_2 = -\begin{bmatrix} 0 \\ \frac{21025}{36992} \end{bmatrix} \Rightarrow \Delta\mathbf{X}_2 = \begin{bmatrix} \frac{21\,025}{235\,552} \\ -\frac{21\,025}{235\,552} \end{bmatrix} \Rightarrow$$

$$\Rightarrow \mathbf{X}_3 = \begin{bmatrix} -\frac{25}{272} \\ \frac{841}{272} \end{bmatrix} + \begin{bmatrix} \frac{21\,025}{235\,552} \\ -\frac{21\,025}{235\,552} \end{bmatrix} = \begin{bmatrix} -\frac{625}{235\,552} \\ \frac{707\,281}{235\,552} \end{bmatrix} \approx \begin{bmatrix} -0,0027 \\ 3,0027 \end{bmatrix}, \ \frac{\|\Delta\mathbf{X}_2\|}{\|\mathbf{X}_3\|} \approx 0,04.$$

Для порівняння, точні розв'язки цієї системи – $\mathbf{X}_1 = [3 \ \ 0]^{\mathrm{T}}$, $\mathbf{X}_2 = [0 \ \ 3]^{\mathrm{T}}$.

3.5. Квазіньютонівські методи

Під час розв'язання систем нелінійних рівнянь методом Ньютона може виникнути ситуація, коли аналітичне обчислення похідних є досить

складним, а іноді навіть неможливим. У такому випадку виникає необхідність модифікації методу Ньютона з метою спростити розрахунок якобіана системи [21].

Одним з можливих способів розрахунку якобіана є заміна похідних скінченними різницями

$$\frac{\partial f_j\left(x_1,\, x_2,\ldots,\, x_n\right)}{\partial x_i} \approx \frac{f_j\left(x_1,\, x_2,\ldots,\, x_i + h_i,\ldots,\, x_n\right) - f_j\left(x_1,\, x_2,\ldots,\, x_i,\ldots,\, x_n\right)}{h_i},$$

де h_i – приріст i-го аргументу цієї функції. Оптимальне значення кроку скінченних різниць визначають аналогічно (3.12).

Іншим способом модифікації методу Ньютона є використання методу січних, для якого якобіан системи на кожному k + 1-му кроці обчислюється за формулою

$$\mathbf{J}_{k+1} = \mathbf{J}_k + \frac{\mathbf{F}\left(\mathbf{X}_{k+1}\right)\mathbf{G}_k^T}{\mathbf{G}_k^T \Delta \mathbf{X}_k},$$

де залежно від вибору вектора $\mathbf{G}_k$ маємо різні методи розв'язання системи нелінійних рівнянь [6]:

$\mathbf{G}_k = \Delta\mathbf{X}_k$ – метод Бройдена;

$\mathbf{G}_k = \mathbf{J}_k^{\mathrm{T}}\Delta\mathbf{F}_k,\ \Delta\mathbf{F}_k = \mathbf{F}\left(\mathbf{X}_{k+1}\right) - \mathbf{F}\left(\mathbf{X}_k\right)$ – модифікований метод Бройдена;

$\mathbf{G}_k = \Delta\mathbf{F}_k$ – метод Пірсона; $\mathbf{G}_k = \Delta\mathbf{F}_k - \mathbf{J}_k\Delta\mathbf{X}_k$ – симетричний метод першого рангу.

Оскільки для першого кроку алгоритму січних необхідно мати початкову матрицю-якобіан $\mathbf{J}_0$, то на практиці на першому кроці спочатку один раз обчислюють матрицю Якобі через апроксимацію похідних скінченними різницями або аналітичним способом.

Для прикладу зробимо кілька ітерацій методом Бройдена для системи із попереднього прикладу:

$$\mathbf{F}(\mathbf{X})=\begin{bmatrix} x_1+x_2-3 \\ x_1^2+x_2^2-9 \end{bmatrix}=0\ ;\ \mathbf{X}_0=\begin{bmatrix}1 & 5\end{bmatrix}^{\mathrm{T}};$$

$$\mathbf{J}_0=\mathbf{J}(\mathbf{X}_0)=\begin{bmatrix}1 & 1\\ 2 & 10\end{bmatrix}-\text{обчислено аналітично};\ \mathbf{F}_0=\mathbf{F}(\mathbf{X}_0)=\begin{bmatrix}3\\ 17\end{bmatrix};$$

$$\mathbf{J}_0\Delta\mathbf{X}_0=-\mathbf{F}_0\Rightarrow\begin{bmatrix}1 & 1\\ 2 & 10\end{bmatrix}\Delta\mathbf{X}_0=\begin{bmatrix}-3\\ -17\end{bmatrix}\Rightarrow\Delta\mathbf{X}_0=\begin{bmatrix}-1{,}625\\ -1{,}375\end{bmatrix};$$

$$\mathbf{X}_1=\mathbf{X}_0+\Delta\mathbf{X}_0=\begin{bmatrix}1\\ 5\end{bmatrix}+\begin{bmatrix}-1{,}625\\ -1{,}375\end{bmatrix}=\begin{bmatrix}-0{,}625\\ 3{,}625\end{bmatrix};\ \ \mathbf{F}_1=\mathbf{F}(\mathbf{X}_1)=\begin{bmatrix}0\\ 4{,}53125\end{bmatrix};$$

$$\mathbf{J}_1=\mathbf{J}_0+\frac{\mathbf{F}_1\Delta\mathbf{X}_0^{\mathrm{T}}}{\Delta\mathbf{X}_0^{\mathrm{T}}\Delta\mathbf{X}_0}=\begin{bmatrix}1 & 1\\ 2 & 10\end{bmatrix}+\frac{\begin{bmatrix}0\\ 4{,}53125\end{bmatrix}\times\begin{bmatrix}-1{,}625 & -1{,}375\end{bmatrix}}{\begin{bmatrix}-1{,}625 & -1{,}375\end{bmatrix}\times\begin{bmatrix}-1{,}625\\ -1{,}375\end{bmatrix}}=\begin{bmatrix}1 & 1\\ 0{,}375 & 8{,}625\end{bmatrix};$$

$$\mathbf{J}_1\Delta\mathbf{X}_1=-\mathbf{F}_1\Rightarrow\begin{bmatrix}1 & 1\\ 0{,}375 & 8{,}625\end{bmatrix}\Delta\mathbf{X}_1=\begin{bmatrix}0\\ -4{,}53125\end{bmatrix}\Rightarrow\Delta\mathbf{X}_1=\begin{bmatrix}0{,}549\\ -0{,}549\end{bmatrix}$$

$$\mathbf{X}_2=\mathbf{X}_1+\Delta\mathbf{X}_1=\begin{bmatrix}-0{,}625\\ 3{,}625\end{bmatrix}+\begin{bmatrix}0{,}549\\ -0{,}549\end{bmatrix}=\begin{bmatrix}-0{,}076\\ 3{,}076\end{bmatrix}.$$

3.6. Модифікації методу Ньютона, що збігаються глобально

Найскладнішим завданням у практичному застосуванні методу Ньютона є вибір початкового наближення. Для кожного конкретного класу задач завдання вибору $\mathbf{X}_0$ вирішується індивідуально. Однак умови локальної теореми збіжності (3.16) підказують надійний критерій області квадратичної збіжності. На практиці, якщо ітерації не збігаються за 6…7 кроків, то слід вибрати інше початкове наближення. Також існують

модифікації методу Ньютона, які забезпечують його збіжність до розв'язку **X***.

Найпростіший спосіб забезпечення збіжності — лінійний пошук [6]. Нехай на k-му кроці знайдено ньютонівський або квазіньютонівський напрям $\Delta\mathbf{X}_k$. Покладемо довжину кроку t_k=1 і перевіримо нерівність

$$\left\|\mathbf{F}\left(\mathbf{X}_k + t_k\Delta\mathbf{X}_k\right)\right\| \le \left\|\mathbf{F}\left(\mathbf{X}_k\right)\right\|. \qquad (3.17)$$

Якщо (3.17) не виконується, то довжина кроку t_k зменшується (наприклад, в два рази) до тих пір, поки не виконається умова (3.17). На цьому лінійний пошук завершують і обчислюють нове значення за формулою:

$$\mathbf{X}_{k+1} = \mathbf{X}_k + t_k\Delta\mathbf{X}_k. \qquad (3.18)$$

Таким чином, вдалий лінійний пошук забезпечує монотонне зменшення норми нев'язки $\left\|\mathbf{F}(\mathbf{X}_k)\right\|$ з ростом k. Однак у разі використання методів січних квазіньютонівський напрям пошуку $\Delta\mathbf{X}_k$ може суттєво відрізнятися від ньютонівського. Тоді лінійний пошук може бути невдалим. В таких випадках слід відновити матрицю Якобі за допомогою, наприклад, апроксимації скінченними різницями.

Приклад.

$$\mathbf{F}(\mathbf{X}) = \begin{bmatrix} e^{x_1} & -1 \\ e^{x_2} & -1 \end{bmatrix} = 0;$$

$$\mathbf{X}_0 = \left[-10, -10\right]^T;$$

$$\mathbf{J}(\mathbf{X}) = \begin{bmatrix} e^{x_1} & 0 \\ 0 & e^{x_2} \end{bmatrix};$$

$$\Delta\mathbf{X}_0 = \left[-1 + e^{10}, -1 + e^{10}\right]^T;$$

$\left\|\mathbf{F}(\mathbf{X}_0)\right\| = \left(1 - e^{-10}\right)^2 \approx 1.$

Умова $\left\|\mathbf{F}\left(\mathbf{X}_0 + t\Delta\mathbf{X}_0\right)\right\| < \left\|\mathbf{F}\left(\mathbf{X}_0\right)\right\|$ виконується якщо

$$e^{-10+t(-1+e^{10})} < 2; \qquad \Rightarrow \qquad t < \frac{10 + \ln 2}{e^{10} - 1} \le 0.0004854.$$

Нехай t=2^{-12}=0.0002441:

$$\mathbf{X}_1 = \mathbf{X}_0 + t\Delta\mathbf{X}_0 = \begin{bmatrix} -10 + 0.0002441(-1 + e^{10}) \\ -10 + 0.0002441(-1 + e^{10}) \end{bmatrix} = \begin{bmatrix} -4.623586 \\ -4.623586 \end{bmatrix}.$$

$\left\|\mathbf{F}\left(\mathbf{X}_1\right)\right\| = 0.9804613.$

$$\Delta\mathbf{X}_1 = \left[\underbrace{-1 + e^{4.623586}}_{100.8586}, -1 + e^{4.623586}\right], \qquad \text{умова} \qquad \left\|\mathbf{F}\left(\mathbf{X}_1 + t\Delta\mathbf{X}_1\right)\right\| < \left\|\mathbf{F}\left(\mathbf{X}_1\right)\right\|$$

виконується якщо

$$e^{-4.623586+t(-1+e^{4.623586})} < 2 \qquad \Rightarrow \qquad t < \frac{4.623586 + \ln 2}{e^{4.623586} - 1} = 0.0527143.$$

Нехай $t = 2^{-5} = 0.03125 < 0.0527141.$

$$\mathbf{X}_2 = \mathbf{X}_1 + t\Delta\mathbf{X}_1 = \begin{bmatrix} -4.623586 + 0.03125\left(-1 + e^{4.623586}\right) \\ -4.623586 + 0.03125\left(-1 + e^{4.623586}\right) \end{bmatrix} = \begin{bmatrix} -1.4717548 \\ -1.4717548 \end{bmatrix}.$$

3.6.1. Метод продовження по параметру

Метод продовження по параметру є найбільш універсальним [23]. Позначимо через t параметр, що змінюється від 0 до 1. Введемо систему нелінійних рівнянь

$$\mathbf{G}\left(\mathbf{X}, t\right) = 0, \tag{3.19}$$

таку, що за t=0 система $\mathbf{G}\left(\mathbf{X}, 0\right)$ має відомий розв'язок $\mathbf{X}_0$, а у разі t=1 система $\mathbf{G}\left(\mathbf{X}, 1\right)$ має розв'язок $\mathbf{X}^*$, що відповідає шуканому розв'язку вихідної системи (3.4).

Систему (3.19) можна побудувати різними способами. Головною вимогою є неперервність $\mathbf{G}(\mathbf{X},t)$ від $t\in[0,1]$. Тоді, змінюючи t від 0 до 1 і розв'язуючи для кожного $t=t_i$ систему (3.19) методом Ньютона, можна знайти послідовність $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2,..., \mathbf{X}^*$.

Оскільки $\mathbf{X}_0$ для $t_0=0$ відоме, то завжди можна знайти t_1, досить близьке до t_0, за якого будуть виконуватися умови збіжності методу Ньютона. Аналогічно можна забезпечити умови збіжності методу Ньютона і для $t_2, t_3,..., t=1$.

Один з можливих способів побудови $\mathbf{G}(\mathbf{X},t)$ дає наступну систему нелінійних рівнянь:

$$\mathbf{G}(\mathbf{X},t)=\mathbf{F}(\mathbf{X})+(t-1)\mathbf{F}(\mathbf{X}_0)=0,$$

де $\mathbf{X}_0$ — фіксоване значення $\mathbf{X}$.

Оскільки $\dfrac{d\mathbf{G}(\mathbf{X},t)}{d\mathbf{X}}=\dfrac{d\mathbf{F}(\mathbf{X})}{d\mathbf{X}}=\mathbf{J}(\mathbf{X})$, то для кожного значення t_i ітераційна формула буде мати вигляд

$$\mathbf{X}_{k+1}=\mathbf{X}_k-\mathbf{J}(\mathbf{X}_k)^{-1}\mathbf{G}(\mathbf{X}_k,t_i)$$

за початкового наближення $\mathbf{X}_0(t_i)=\mathbf{X}^*(t_{i-1})$, де $\mathbf{X}^*(t_{i-1})$ — розв'язок (3.19) для $t=t_{i-1}$.

Для покращення збіжності методу Ньютона, після кількох кроків по параметру t можна використовувати прогноз $\mathbf{X}_0(t_i)$ за формулами екстраполяції. Якщо в ході обчислення матриці Якобі, функції чи розв'язання СЛАР виникають переповнення, обчислення коренів з від'ємних чисел та зупинки з інших подібних причин, то крок Δt слід одразу зменшити. Метод продовження по параметру універсальний і завжди приводить до розв'язку $\mathbf{X}^*$, однак для системи з кількома стійкими станами може бути отриманий розв'язок $\mathbf{X}^*$, що відповідає точці нестійкої

рівноваги. В такому випадку слід задавати інше значення $\mathbf{X}_0(0)$ або використовувати інші способи побудови $\mathbf{G}(\mathbf{X},t)$.

Інший спосіб побудови $\mathbf{G}(\mathbf{X},t)$ полягає у масштабуванні нелінійності. Введемо параметр t і домножимо на нього кожну нелінійну залежність вихідної системи $\mathbf{F}(\mathbf{X})$. Тоді для t=0 отримаємо СЛАР, яка розв'язується за одну ітерацію по методу Ньютона, і отримаємо $\mathbf{X}_0(0)$. Потім, змінюючи t від 0 до 1, отримаємо розв'язок системи $\mathbf{G}(\mathbf{X},1)=0$.

Третій спосіб ґрунтується на тому, що в багатьох практичних випадках задача природним чином залежить від деякого S, причому для $S=S_{ном}$ отримуємо систему $\mathbf{F}(\mathbf{X})=0$, а у разі S = 0 — систему $\mathbf{F}_0(\mathbf{X})=0$, що має відомий розв'язок $\mathbf{X}_0(0)$ ($S_{ном}$ — номінальне значення S). Тоді $\mathbf{G}(\mathbf{X},t)=\mathbf{F}(\mathbf{X},tS_{ном})$ для $0<t<1$. Наприклад, для електронних схем $S_{ном}$ може бути напруга живлення схеми. Оскільки у випадку S=0 схема буде відключена від живлення, то струми і напруги на компонентах дорівнюють нулю, тобто $\mathbf{X}_0(0)=0$. Для $S=S_{ном}$ отримаємо розв'язок $\mathbf{X}^*$, що відповідає стійкому стану схеми.

3.6.2. Метод диференціювання по параметру

В методі диференціювання по параметру алгебраїчна задача зводиться до диференціальної [24]. Розглянемо вектор-функцію $\mathbf{G}(\mathbf{X},t)$ як функцію параметра $t\in[0,1]$, тобто $\mathbf{\Phi}(t)=\mathbf{G}(\mathbf{X}(t),t)$. Припустимо, що $\mathbf{\Phi}(t)$ неперервно диференційована по t на відрізку [0,1], тобто $\frac{d\mathbf{\Phi}(t)}{dt}=\frac{\partial\mathbf{G}}{\partial\mathbf{X}}\frac{d\mathbf{X}}{dt}+\frac{\partial\mathbf{G}}{\partial t}$. Оскільки $\mathbf{X}=\mathbf{X}(t)$ задовольняє рівнянню $\mathbf{G}(\mathbf{X}(t),t)=0$, то $\frac{d\mathbf{\Phi}(t)}{dt}=0$ для всіх t, і, відповідно, $\mathbf{X}$ задовольняє диференціальному рівнянню

$$\frac{\partial \mathbf{G}}{\partial \mathbf{X}} \mathbf{X}'(t) = -\frac{\partial \mathbf{G}}{\partial t},$$

або

$$\mathbf{X}'(t) = -\left[\frac{\partial \mathbf{G}}{\partial \mathbf{X}}\right]^{-1} \frac{\partial \mathbf{G}}{\partial t}. \qquad (3.20)$$

Таким чином, розв'язок системи звичайних диференціальних рівнянь (ЗДР) (3.20) для t=1 з початковими умовами $\mathbf{G}(\mathbf{X}(0),0)$=0 для t=0 дасть розв'язок вихідної системи $\mathbf{F}(\mathbf{X}) = 0$.

Зокрема, якщо $\mathbf{G}(\mathbf{X},t)=\mathbf{F}(\mathbf{X})+(t-1)\mathbf{F}(\mathbf{X}_0)$, то система ЗДР має вигляд

$$\mathbf{X}'(t)= -\mathbf{J}^{-1}(\mathbf{X}(t))\mathbf{F}(\mathbf{X}_0) \qquad (3.21)$$

з початковою умовою $\mathbf{X}(0)=\mathbf{X}_0$.

Розв'язок системи (3.21) є лише наближенням до $\mathbf{X}^*$ через накопичення похибки інтегрування, тому метод диференціювання по параметру завершується циклом ітерацій методом Ньютона з початкового наближення $\mathbf{X}_0=\mathbf{X}(1)$.

Приклад.

$x^2 - 1 = 0, \quad x^* = \pm 1.$

$f(x) = x^2 - 1;$

$x_0 = 2;$

$G(x,t) = f(x) + (t-1) f(x_0) = x^2 - 1 + (t-1)3 = x^2 + 3t - 4;$

$\frac{\partial G}{\partial t} = 3; \qquad \frac{\partial G}{\partial x} = 2x;$

$\frac{dx}{dt} = -\left(\frac{\partial G}{\partial x}\right)^{-1} \frac{\partial G}{\partial t} = -\frac{3}{2x}, \qquad x(0) = x_0 = 2, \qquad x(1) = ?$

Таким чином, розв'язання вихідного нелінійного рівняння відповідає розв'язанню задачі Коші

$$\frac{dx}{dt} = -\frac{3}{2x}, \qquad x(0) = x_0 = 2, \qquad x \in [0,1].$$

Покажемо, що розв'язок задачі Коші для t=1 співпадає з шуканим розв'язком нелінійного рівняння.

$$2xdx = -3t \quad \Rightarrow \quad x^2 = -3t + c \quad \Rightarrow \quad x = \pm\sqrt{-3t + c};$$

$$x(0) = 2 \quad \Rightarrow \quad c = 4.$$

Таким чином,

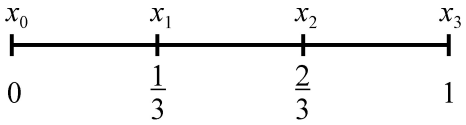
$$x(t) = \pm\sqrt{4 - 3t} \qquad x(1) = \pm 1.$$

Застосуємо для чисельного розв'язання задачі Коші явний метод Ейлера:

$$\frac{dx}{dt} = \varphi(t,x).$$

В нашому випадку $\varphi(t,x) = -\frac{3}{2x}$

$x_{i+1} = x_i + \Delta t\varphi(t_i, x_i)$, $\Delta t = \frac{1}{3}$:



$$x_1 = x_0 + \frac{1}{3}\left(-\frac{3}{2x_0}\right) = 2 - \frac{1}{3}\frac{3}{2\cdot 2} = \frac{7}{4};$$

$$x_2 = x_1 + \frac{1}{3}\left(-\frac{3}{2x_1}\right) = \frac{7}{4} - \frac{1}{3}\frac{3}{2\cdot\frac{7}{4}} = \frac{41}{28};$$

$$x_3 = x_2 + \frac{1}{3}\left(-\frac{3}{2x_2}\right) = \frac{41}{28} - \frac{1}{3}\frac{3}{2\cdot\frac{41}{28}} = \frac{41}{28} - \frac{14}{41} \approx 1.123.$$

Контрольні завдання

1. Вибрати з табл. 3.1 нелінійне рівняння відповідно до варіанта.
2. Визначити відрізки, що містять корені. На одному із визначених відрізків знайти корінь рівняння методом бісекції, Ньютона та січних з похибкою, не більшою 1 %.
3. Порівняти розв'язки, знайдені різними методами.
4. Вибрати із табл. 3.2 систему рівнянь відповідно до свого варіанта.
5. Розв'язати систему методом Ньютона з похибкою, не більшою 1 %.
6. Розв'язати систему методом Бройдена з похибкою, не більшою 1 %.
7. Порівняти результати розв'язків різними методами.

Варіанти завдань

Таблиця 3.1. **Нелінійні рівняння**

Варіант	Рівняння	Варіант	Рівняння
1	$3\operatorname{arctg}(2x-7)-2\ln(x)=0$	2	$7\operatorname{arctg}(x)-\frac{4}{x^3}+\frac{1}{x}=0$
3	$x^3+x^2-10x+1=0$	4	$x^3-2x^2-3x+5=0$
5	$x^5+x^3-7x+3=0$	6	$9\operatorname{arctg}(x)-3x-\frac{2}{x}-1=0$
7	$2\ln(x^2-x+1)-x=0$	8	$x^3-4x^2+x+2=0$
9	$3\ln(x^2-x)-3x+7=0$	10	$8\operatorname{arctg}(2x)-5x^3-8x-1=0$
11	$x^2-5x+1=0$	12	$3^x-2x-3=0$
13	$x^5-5x^3+4x-1=0$	14	$x^4+x^3-2x^2-3x-2=0$
15	$2\operatorname{arctg}(x)+x+\frac{3}{x}+8=0$	16	$\operatorname{arctg}(3x-2)-\ln(x+1)=0$
17	$\operatorname{arctg}(3x)+2x^3-5x+1=0$	18	$e^x-x-6=0$

19	$x^4+2x-3=0$	20	$2\ln\left(-(x+2)(x+5)\right)+x+3=0$
21	$3x^4+8x^3+6x^2-10=0$	22	$\operatorname{arctg}(4x)-\ln(x+2)=0$
23	$x^4+x^3-3x^2-3=0$	24	$x^4+x^3-4x^2+1=0$
25	$x^4+x^3-2x^2+3x+1=0$	26	$3x^4-3x^3+2x^2-3x-5=0$
27	$\operatorname{arctg}(x)-2x^3+2x-1=0$	28	$3\operatorname{arctg}(4x)+\frac{9}{x^3}-\frac{15}{x}+2=0$
29	$\operatorname{arctg}(3x)+x-\frac{1}{x}+3=0$	30	$5\operatorname{arctg}(2x)+\frac{3}{x^3}-\frac{2}{x}-7=0$

Таблиця 3.2. **Системи нелінійних рівнянь**

Варіант	Система рівнянь	Варіант	Система рівнянь
1	$\begin{cases}\sin(x)+\cos(y)=1;\\ \sin(3x)-2\cos(y)=-1.\end{cases}$	2	$\begin{cases}\sin^2(x)+\cos(y)=1;\\ \cos(x)-\sin(y)=0.\end{cases}$
3	$\begin{cases}e^{2x}+\sqrt{y}=3;\\ e^x-3\sqrt{y}=-1.\end{cases}$	4	$\begin{cases}\sin(x)+\sqrt{y}=2;\\ 2\cos(x)+\sqrt{y}=3.\end{cases}$
5	$\begin{cases}3\sin(x)+y^2=2;\\ \sin(x)-y^2=0.\end{cases}$	6	$\begin{cases}\sin^2(x)+\sqrt{y}=1;\\ \sin(x)-2\sqrt{y}=-1.\end{cases}$
7	$\begin{cases}e^{2x}+y=2;\\ e^x-2y=1.\end{cases}$	8	$\begin{cases}e^x+\cos(y)=3;\\ \cos(x)+e^y=4.\end{cases}$
9	$\begin{cases}2\cos(x)+\sin(y)=1;\\ \sin(x)-\cos(y)=0.\end{cases}$	10	$\begin{cases}\cos^2(x)+\sqrt{y}=1;\\ \sqrt{x}-\cos(y)=0.\end{cases}$
11	$\begin{cases}\cos(x)+y=1;\\ \sqrt{x}-y=0.\end{cases}$	12	$\begin{cases}e^x+e^y=3;\\ x+2y=1.\end{cases}$
13	$\begin{cases}e^{2x}+\sin(y)=3;\\ \sin(x)+e^y=2.\end{cases}$	14	$\begin{cases}e^x+e^y=3;\\ \cos(2x)+\cos(y)=1.\end{cases}$
15	$\begin{cases}e^x+e^{2y}=5;\\ 2\sqrt{x}+\sqrt{y}=2.\end{cases}$	16	$\begin{cases}e^{3x}-e^y=4;\\ \sqrt{x}+2\sqrt{y}=2.\end{cases}$

17	$\begin{cases} \operatorname{tg}(x)+\sin(y)=2; \\ \sqrt{x}+\sqrt{y}=2. \end{cases}$	18	$\begin{cases} \sin(x)+\operatorname{tg}(y)=2; \\ \sin(x)+2\cos(y)=1. \end{cases}$
19	$\begin{cases} \operatorname{tg}(x)+e^{y}=2; \\ e^{2x}+\sqrt{y}=2. \end{cases}$	20	$\begin{cases} \operatorname{tg}(x)+\operatorname{tg}(y)=3; \\ e^{x}+e^{2y}=6. \end{cases}$
21	$\begin{cases} \sqrt{x}+\operatorname{tg}(y)=2; \\ x+2y=3. \end{cases}$	22	$\begin{cases} 3\operatorname{tg}(2x)+\sqrt{y}=2; \\ x-2y=0. \end{cases}$
23	$\begin{cases} \operatorname{tg}^{2}(x)+y=0, \\ 2x+\sqrt{y}=0 \end{cases}$	24	$\begin{cases} e^{x}+\operatorname{tg}^{2}(y)=2; \\ 2x+e^{y}=y. \end{cases}$
25	$\begin{cases} xe^{y}=2; \\ \sin(x)+y=2. \end{cases}$	26	$\begin{cases} x\sqrt{y}+y=3; \\ \cos^{2}(x)+xy=2. \end{cases}$
27	$\begin{cases} xe^{x}+\sin(xy)=1; \\ \sin(x)+2\cos(y)=2. \end{cases}$	28	$\begin{cases} x^{2}+\cos(y)=2; \\ e^{x}\cos(x)+y=2. \end{cases}$
29	$\begin{cases} \operatorname{tg}(xy)+e^{xy}=2; \\ x-y=1. \end{cases}$	30	$\begin{cases} 3\operatorname{tg}^{2}(x)+xy=1; \\ xe^{x}+y=1. \end{cases}$

4. Інтерполяція функцій

Нехай деяку функцію задано табл. 4.1.

Таблиця 4.1. **Деяка таблична функція**

x	x_1	x_2	…	x_n
$f(x)$	y_1	y_2	…	y_n

Під задачею інтерполяції функції розуміють побудову такої функції $f(x)$, щоб для кожного i виконувалась рівність:

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

Точки (x_i, y_i) називають вузлами інтерполяції.

Інакше кажучи, завдання інтерполяції полягає в тому, щоб за значеннями функції, заданої в декількох точках відрізка, відновити її значення в інших точках цього відрізка [25].

Геометричний зміст інтерполяції полягає в побудові такої функції $y = f(x)$, графік якої проходив би через усі задані точки (x_i, y_i),

Інтерполяційну функцію будують, як правило, у вигляді лінійної комбінації деяких елементарних функцій:

$$f(x) = \sum_{j=1}^{n} c_j \varphi_j(x).$$

Вид інтерполяції визначається функцією $\varphi_j(x)$. Якщо $\varphi_j(x)$ є поліномом n-1 степеня, тобто $\varphi_j(x) = b_{j0} + b_{j1}x + b_{j2}x^2 + \ldots + b_{jn-1}x^{n-1}$, то інтерполяція називається *поліноміальною*. Якщо $\varphi_j(x) = \cos(b_j x + a_j)$, то така інтерполяція називається *тригонометричною*. Кусково-поліноміальну інтерполяцію називають сплайн-інтерполяцією.

4.1. Інтерполяційна формула Лагранжа

Шукатимемо поліноміальну інтерполяцію табличної функції у вигляді полінома n-1 степеня $P_{n-1}(x)$.

Такий поліном можна подати у вигляді

$$P_{n-1}(x)=\sum_{i=1}^{n}Q_i(x)y_i\,, \tag{4.1}$$

де $Q_i(x)$ – поліноми n-1 порядку, а вагові коефіцієнти дорівнюють y_i. Із умови $f\left(x_j\right)=y_j$ маємо

$$\sum_{i=1}^{n}y_iQ_i\left(x_j\right)=y_j,\ \ j=\overline{1,n}. \tag{4.2}$$

Співвідношення (4.2) виконуються за умови, що

$$Q_i\left(x_j\right)=\begin{cases}0, & i\neq j\\ 1, & i=j\end{cases}\quad j=\overline{1,n}. \tag{4.3}$$

Враховуючи, що $Q_i(x)$ – поліном степеня n-1, із умови (4.3) маємо

$$Q_i\left(x\right)=\prod_{\substack{j=1\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}\ . \tag{4.4}$$

Тоді, використовуючи (4.1) і (4.4), отримаємо шукану інтерполяційну формулу

$$f\left(x\right)\approx y\left(x\right)=\sum_{i=1}^{n}y_i\prod_{\substack{j=1\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}. \tag{4.5}$$

Формулу (4.5) називають інтерполяційною формулою Лагранжа.

В такому вигляді інтерполяційна формула Лагранжа досить незручна для практичного використання, але за допомогою простих перетворень її можна представити у більш зручному вигляді [26]. Для цього розглянемо окремий випадок, коли $y\left(x\right)=1$. Тоді для всіх i $y_i=1$.

Розділивши (4.5) на $1=\sum_{i=1}^{n}\prod_{\substack{j=1\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}$ та поділивши чисельник та знаменник правої частини на $\prod_{j=1}^{n}x-x_j$, отримаємо

$$f(x)=\frac{\sum_{i=1}^{n}y_i\frac{a_i}{x-x_i}}{\sum_{i=1}^{n}\frac{a_i}{x-x_i}}, \qquad (4.6)$$

де $a_i=\prod_{\substack{j=1\\ j\neq i}}^{n}\frac{1}{x_i-x_j}$.

Формулу (4.6) називають *барицентричною* [26].

Основним недоліком формули Лагранжа є необхідність перерахунку всіх коефіцієнтів полінома у разі зміни кількості вузлів інтерполяції. Разом з тим, формулою Лагранжа зручно користуватись, коли на одній й тій же сітці аргументів вузлів треба інтерполювати декілька функцій.

Приклад. Для табличної функції, заданої в табл. 4.2, побудувати інтерполяційний поліном Лагранжа.

Таблиця 4.2. **Задана таблична функція**

x	0	1	2
y	0	2	10

Згідно з формулою (4.5)

$$f(x)=0\frac{(x-1)(x-2)}{(0-1)(0-2)}+2\frac{(x-0)(x-2)}{(1-0)(1-2)}+10\frac{(x-0)(x-1)}{(2-0)(2-1)}=3x^2-x.$$

4.2. Інтерполяційна формула Ньютона

Формула Ньютона дає змогу виразити інтерполяційний поліном $P_{n\text{-}1}(x)$ через значення y_i в одному з вузлів та через розділені різниці функції $f(x)$,

побудованими за іншими вузлами x_1, x_2,…, x_n. Вона є аналогом формули Тейлора [4].

Розділеними різницями першого порядку називають відношення

$$\Delta(x_i,\ x_j)=\frac{y_i-y_j}{x_i-x_j},\ i\neq j,\ i=\overline{1,n}.$$

Розглянемо розділені різниці, побудовані за сусідніми вузлами, тобто вирази $\Delta(x_1,\ x_2)$, $\Delta(x_2,\ x_3)$, …,$\Delta(x_{n-1},\ x_n)$. За цими розділеними різницями можна побудувати різниці другого порядку:

$$\Delta(x_1,\ x_2,\ x_3)=\frac{\Delta(x_1,\ x_2)-\Delta(x_2,\ x_3)}{x_1-x_3},$$

…

$$\Delta(x_{n-2},\ x_{n-1},\ x_n)=\frac{\Delta(x_{n-2},\ x_{n-1})-\Delta(x_{n-1},x_n)}{x_{n-2}-x_n}.$$

За різницями другого порядку будують різниці третього порядку і т. д. В загальному випадку розділені різниці k-го порядку розраховують за формулами:

$$\Delta(x_j,\ x_{j+1},\ldots,\ x_{j+k})=\frac{\Delta(x_j,\ x_{j+1},\ldots,\ x_{j+k-1})-\Delta(x_{j+1},x_{j+2}\ldots,\ x_{j+k})}{x_j-x_{j+k}},\ j=\overline{1,n-k},$$

де $\Delta(x_j,\ x_{j+1},\ldots,\ x_{j+k-1}),\Delta(x_{j+1},x_{j+2}\ldots,\ x_{j+k})$ – розділені різниці (k-1)-го порядку.

Процес розрахунку розділених різниць зручно подати схемою рис. 4.1 [25].

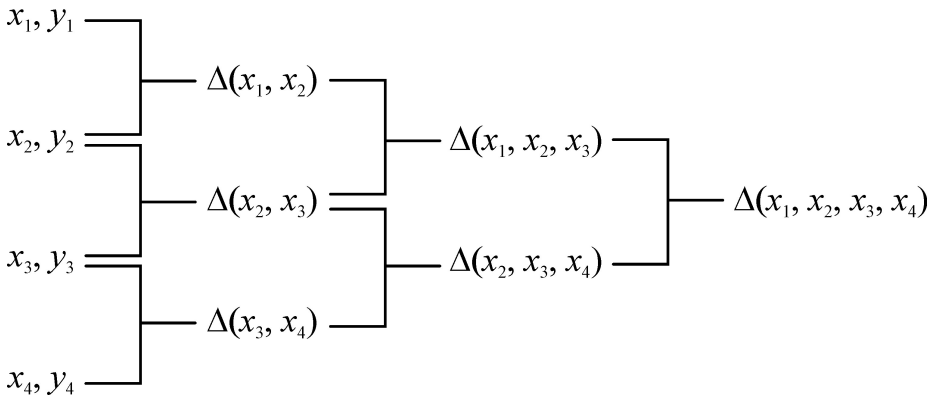


Рис. 4.1. Схема розрахунку розділених різниць.

Розглянемо розділену різницю $\Delta(x,\, x_1) = \dfrac{f(x) - f(x_1)}{x - x_1}$. Звідси знаходимо

$$f(x) = y_1 + (x - x_1)\Delta(x, x_1).$$

Аналогічно:

$$\Delta(x,\, x_1) = \Delta(x_1,\, x_2) + (x - x_2)\Delta(x,\, x_1,\, x_2);$$
$$\Delta(x,\, x_1,\, x_2) = \Delta(x_1,\, x_2,\, x_3) + (x - x_3)\Delta(x,\, x_1,\, x_2,\, x_3).$$

Скориставшись цим виразом, записуємо інтерполяційну формулу

$$\begin{aligned} f(x) \approx\ & y_1 + (x - x_1)\Delta(x_1,\, x_2) + (x - x_1)(x - x_2)\Delta(x_1,\, x_2,\, x_3) + \dots \\ & + (x - x_1)(x - x_2)\dots(x - x_{n-1})\Delta(x_1,\, x_2, \dots,\, x_n) = \\ & = y_1 + \sum_{k=2}^{n}\left(\Delta(x_1,\, x_2, \dots,\, x_k)\prod_{j=1}^{k-1}(x - x_j)\right). \end{aligned} \qquad (4.7)$$

Формула (4.7) називається формулою Ньютона інтерполювання вперед, оскільки формула містить значення y, що знаходяться справа від x_1. Аналогічним чином можна побудувати формулу Ньютона, використовуючи вузли, що знаходяться зліва від x_n [25]:

$$f(x) \approx y_n + (x - x_n)\Delta(x_{n-1},\, x_n) + (x - x_n)(x - x_{n-1})\Delta(x_{n-2},\, x_{n-1},\, x_n) + \ldots$$
$$+ (x - x_n)(x - x_{n-1})\ldots(x - x_1)\Delta(x_1,\, x_2, \ldots,\, x_n) =$$
$$= y_n + \sum_{k=2}^{n}\left(\Delta(x_{n-k+1},\, x_{n-k+2}, \ldots,\, x_n) \prod_{j=n-k+2}^{n} (x - x_j) \right).$$

(4.8)

Формула (4.8) називають формулою Ньютона інтерполювання назад. Її зручно використовувати для інтерполяції функції у точках, близьких до кінця таблиці.

Зручність використання інтерполяційних формул Ньютона полягає в тому, що у разі доповнення таблиці новим вузлом не виникає необхідності перерахунку всіх коефіцієнтів поліному, а достатньо розрахувати лише однин коефіцієнт біля старшого степеня поліному.

Побудуємо інтерполяційний поліном Ньютона для табличної функції, заданої в табл. 4.2. Використовуючи формулу Ньютона інтерполювання вперед (4.7), маємо:

$$\Delta(x_1,\, x_2) = \frac{0-2}{0-1} = 2, \quad \Delta(x_2,\, x_3) = \frac{2-10}{1-2} = 8, \quad \Delta(x_1,\, x_2,\, x_3) = \frac{2-8}{0-2} = 3;$$

$$f(x) \approx 0 + (x - 0)2 + (x - 0)(x - 1)3 = 2x + 3(x^2 - x) = 3x^2 - x.$$

Побудуємо многочлен, користуючись формулою Ньютона інтерполювання назад.

$$\Delta(x_2,\, x_3) = \frac{2-10}{1-2} = 8,\ \Delta(x_1,\, x_2) = \frac{0-2}{0-1} = 2,\ \Delta(x_1,\, x_2,\, x_3) = \frac{2-8}{0-2} = 3;$$

$$f(x) \approx 10 + (x - 2)8 + (x - 2)(x - 1)3 = 3x^2 - x.$$

4.3. Похибка поліноміальної інтерполяції

Оцінимо похибку поліноміальної інтерполяції. Якщо відомий аналітичний вираз для функції $y(x)$, то похибку інтерполяції можна оцінити за виразом

$$R(x) = y(x) - f(x). \tag{4.9}$$

Вважатимемо, що функція $y(x)$ має всі похідні до n-го порядку включно. Оскільки у вузлах інтерполяції похибка дорівнює нулю, тобто $R(x_i) = 0,\ i = \overline{1,n}$, то функцію $R(x)$ можна подати у вигляді полінома степеня n [14]:

$$R(x) = m\prod_{j=1}^{n}(x - x_j), \tag{4.10}$$

де m – обмежена функція.

Враховуючи (4.9) і (4.10), отримаємо вираз для $y(x)$:

$$y(x) = f(x) + m\prod_{j=1}^{n}(x - x_j).$$

Обчислимо похідну порядку n функції $y(x)$:

$$y^{(n)}(x) = n!m. \tag{4.11}$$

З виразу (4.11) маємо

$$m = \frac{y^{(n)}(x)}{n!}. \tag{4.12}$$

На підставі виразів (4.12) та (4.10) записуємо

$$|R(x)| \le \left| \frac{\max\limits_{x\in[x_1;x_n]}\left(y^{(n)}(x)\right)}{n!}\prod_{j=1}^{n}(x - x_j) \right|.$$

Для табличної функції, коли $y(x)$ невідома, замість похідної підставляємо розділену різницю n-го порядку. У результаті маємо

$$\left|R(x)\right| \le \left| \frac{\max\limits_{x\in[x_1;x_n]} \left(\Delta\left(x,\, x_1,\, x_2,\ldots,x_n\right)\right)}{n!} \prod_{j=1}^{n} \left(x - x_j\right) \right|. \qquad (4.13)$$

Якщо відстань між суміжними точками не перевищує деякої величини h, то $\left|\prod\limits_{j=1}^{n}\left(x-x_j\right)\right| \le c_n h^n$, де c_n – деяка додатня константа, що залежить від способу розбиття відрізка $[x_1; x_n]$. Враховуючи цю нерівність та (4.13), маємо [23]:

$$\left|R(x)\right| \le \left| \frac{\max\limits_{x\in[x_1;x_n]} \left(\Delta\left(x,\, x_1,\, x_2,\ldots,x_n\right)\right) c_n h^n}{n!} \right|.$$

4.4. Інтерполяція сплайнами

За великої кількості вузлів інтерполяції збільшується порядок інтерполяційного полінома, що, з одного боку, робить їх незручними для використання, а з другого – призводить до значного збільшення похибки інтерполяції. Крім того, часто виникає необхідність обчислення не лише значень функції, але і значень похідних. Точність значення похідної, обчисленої шляхом диференціювання інтерполяційного полінома, може виявитися незадовільною.

Виходом з такої ситуації є використання сплайн-інтерполяції [27]. Сплайн – це функція, що на кожному частковому відрізку є поліномом певного степеня, а на всьому заданому відрізку – неперервна разом з кількома своїми похідними. Найчастіше використовують сплайни третього степеня.

Нехай функцію задано таблицею табл. 4.1 і треба побудувати сплайн-інтерполяцію цієї функції. За допомогою сплайнів функцію можна подати у вигляді

$$f(x)=\begin{cases} S_1(x)=a_{01}+a_{11}(x-x_1)+\ldots+a_{m1}(x-x_1)^m,\ x\in[x_1;\ x_2]; \\ S_2(x)=a_{02}+a_{12}(x-x_2)+\ldots+a_{m2}(x-x_2)^m,\ x\in[x_2;\ x_3]; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ S_{n-1}(x)=a_{0n-1}+a_{1n-1}(x-x_{n-1})+\ldots+a_{mn-1}(x-x_{n-1})^m,\ x\in[x_{n-1};\ x_n]. \end{cases}$$

Для побудови сплайн-інтерполяції визначають невідомі коефіцієнти a_{01}, a_{02},…$a_{m(n-1)}$. Якщо функція інтерполюється на n–1 відрізках за допомогою поліномів порядку m, то кількість таких невідомих буде $(m+1)(n-1)$. Для їх визначення потрібно $(m+1)(n-1)$ рівнянь. Щоб отримати необхідну кількість рівнянь, можна скористатися умовами рівності функції табличним значенням у вузлах інтерполяції. Ці умови дадуть $2(n-1)$ рівнянь, чого вистачить для інтерполяції сплайнами першого порядку (лінійними сплайнами, $m = 1$). Для інтерполяції сплайнами другого порядку (параболічними сплайнами, $m = 2$) ці рівняння необхідно доповнити умовою неперервності першої похідної. Це дасть ще $(n-2)$ рівняння. Отриману систему треба доповнити ще одним рівнянням. Наприклад, можна задати поведінку функції на одному з кінців відрізку (наприклад, рівність $S_1'(x_1)=f'(x_1)$, причому похідна функції обчислюється чисельним методом на основі табличних значень функції). У разі інтерполяції кубічними сплайнами систему рівнянь необхідно доповнити ще $(n-1)$ рівняннями. Для цього можна поставити вимогу неперервності другої похідної функції, що дасть $(n-2)$ рівняння.

Останнє рівняння отримують, задавши поведінку функції на другому (незадіяному) кінці відрізка. На практиці часто задають однакові умови на кінцях відрізку. Як правило, це умова гладкості функції на кінцях ($S_1''(x_1)=S_{n-1}''(x_n)=0$).

Для прикладу проведемо кубічну інтерполяцію функції, заданої в табл. 4.2.

На кожному *i*-му відрізку:

$$S_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i,$$

$$S_i'(x) = 3a_i(x - x_i)^2 + 2b_i(x - x_i) + c_i$$

$$S_i''(x) = 6a_i(x - x_i) + 2b_i.$$

Побудуємо систему рівнянь для визначення невідомих коефіцієнтів. З умови рівності функції табличним значенням у вузлових точках маємо:

$$\begin{aligned}
&i = 1;\\
&x = 0:\ a_1(0-0)^3 + b_1(0-0)^2 + c_1(0-0) + d_1 = 0;\\
&x = 1:\ a_1(1-0)^3 + b_1(1-0)^2 + c_1(1-0) + d_1 = 2;\\
&i = 2;\\
&x = 1:\ a_2(1-1)^3 + b_2(1-1)^2 + c_2(1-1) + d_2 = 2;\\
&x = 2:\ a_2(2-1)^3 + b_2(2-1)^2 + c_2(2-1) + d_2 = 10.
\end{aligned} \qquad (4.14)$$

З умови неперервності першої та другої похідних функцій:

$$\begin{aligned}
&x = 1:\\
&3a_1(1-0)^2 + 2b_1(1-0) + c_1 = c_2;\\
&6a_1(1-0) + 2b_1 = 6a_2(1-1) + 2b_2.
\end{aligned} \qquad (4.15)$$

З умови гладкості функції:

$$\begin{aligned}
&x = 0:\ 6a_1(0-0) + 2b_1 = 0;\\
&x = 2:\ 6a_2(2-1) + 2b_2 = 0.
\end{aligned} \qquad (4.16)$$

Розв’язавши рівняння (4.14) – (4.16), отримаємо:

$$a_1 = \frac{3}{2},\ b_1 = 0,\ c_1 = \frac{1}{2},\ d_1 = 0;$$

$$a_2 = -\frac{3}{2},\ b_2 = \frac{9}{2},\ c_2 = 5,\ d_2 = 2.$$

Тоді функція має вигляд

$$f(x)=\begin{cases}\dfrac{3}{2}x^3+\dfrac{1}{2}x,\ x\in[0;\ 1];\\ -\dfrac{3}{2}(x-1)^3+\dfrac{9}{2}(x-1)^2+5(x-1)+2,\ x\in[1;\ 2].\end{cases}$$

Контрольні завдання

1. Вибрати табличну функцію відповідно до свого варіанта.

2. Побудувати інтерполяційний поліном Лагранжа для заданої функції.

3. Побудувати інтерполяційний поліном Ньютона для заданої функції.

4. Побудувати кубічну інтерполяційну формулу заданої функції.

5. Побудувати графіки, нанести на них точки з таблиці та порівняти отримані результати.

Варіанти завдань

1	x	-3	-2	-1	0
	y	-27,0	-12,0	-5,0	-3,0

2	x	-3	-2	-1	1
	y	11,0	3,0	0	-3,0

3	x	-3	-2	-1	2
	y	-6,5	0	1,5	6,0

4	x	-3	-2	-1	3
	y	19,0	5,0	-1,0	-5,0

5	x	-3	-2	0	1
	y	-9,5	-3,0	-2,0	-1,5

6	x	-3	-2	0	2
	y	-5,5	-8,0	-4,0	-8,0

7	x	-3	-2	0	3
	y	-3,5	0	-2,0	17,5

8	x	-3	-2	1	2
	y	9,0	3,0	3,0	-1,0

9	x	-3	-2	1	3
	y	-0,5	2,0	-2,5	14,5

10	x	-3	-2	2	3
	y	-5,5	-8,0	-8,0	-2,5

11	x	-3	-1	0	1
	y	-1,0	3,0	2,0	3,0

12	x	-3	-1	0	2
	y	15,0	2,0	3,0	5,0

13	x	-3	-1	0	3
	y	-12,0	3,0	3,0	9,0

14	x	-3	-1	1	2
	y	-9,5	-2,5	0,5	8,0

15	x	-3	-1	1	3
	y	-2,5	-4,5	-4,5	3,5

16	x	-3	-1	2	3
	y	-17,0	-2,0	-2,0	4,0

17	x	-3	0	1	2
	y	0,0	3,0	2,0	5,0

18	x	-3	0	1	3
	y	-17,0	4,0	3,0	7,0

19	x	-3	0	2	3
	y	7,5	4,0	0,0	-1,5

20	x	-3	1	2	3
	y	6,5	2,5	-1,0	-11,5

21	x	-2	-1	0	1
	y	-7,0	-3,0	-3,0	-4,0

22	x	-2	-1	0	2
	y	5,0	3,5	4,0	-1,0

23	x	-2	-1	0	3
	y	1,0	1,0	-1,0	11,0

24	x	-2	-1	1	2
	y	7,0	5,5	2,5	7,0

25	x	-2	-1	1	3
	y	-4,0	-0,5	-2,5	3,5

26	x	-2	-1	2	3
	y	8,0	5,0	-4,0	-17,0

27	x	-2	0	1	2
	y	6,0	3,0	3,0	8,0

28	x	-2	0	1	3
	y	-1,0	-3,0	-2,5	7,5

29	x	-2	0	2	3
	y	-2,0	-2,0	-6,0	-17,0

30	x	-2	1	2	3
	y	-8,0	-6,5	-8,0	-5,5

5. Чисельне інтегрування функцій

Під час моделювання різноманітних фізичних процесів та електричних кіл досить часто виникає необхідність обчислення інтегралів [7]. Далеко не завжди вдається обчислити інтеграли аналітично, тому використовують методи обчислювальної математики.

Задача чисельного інтегрування полягає в знаходженні наближеного значення інтеграла

$$I(f)=\int_a^b f(x)dx,$$

де $f(x)$ – задана функція. Для розв'язання цієї задачі відрізок $[a, b]$ зазвичай розбивають на n елементарних відрізків точками $a = x_0 < x_1 < x_2 < ... < x_n = b$ і шукане значення інтеграла замінюють сумою

$$I_n(f) = \sum_{i=0}^{n-1} I_i(f) = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)dx. \qquad (5.1)$$

На кожному елементарному відрізку $[x_i, x_{i+1}]$ вводиться сітка $x_i = \xi_{i,0} < \xi_{i,1} < ... < \xi_{i,m} = x_{i+1}$. Як наближене значення інтеграла розглядають число

$$I_i = h_i \sum_{j=0}^{m} c_j f(\xi_{i,j}), \qquad (5.2)$$

де $f(\xi_{ij})$ – значення функції $f(x)$ у вузлах $x = \xi_{ij}$; c_j – вагові множники, що залежать тільки від вузлів, а не від вибору $f(x)$, $h_i = x_{i+1} - x_i$. Для чисельного наближення визначених інтегралів часто використовують термін «квадратура», щоб уникнути плутанини з чисельним інтегруванням звичайних диференціальних рівнянь. Тому формулу (5.2) називають квадратурною формулою. Число m називають порядком квадратурної

формули. Точки $\xi_{i,j}$ називають вузлами, а числа c_j – коефіцієнтами квадратурної формули.

Задача чисельного інтегрування за допомогою квадратур полягає у відшуканні таких вузлів $\xi_{i,j}$ і таких вагових множників c_j, щоб похибка квадратурної формули

$$R_i(f) = I_i(f) - I_i = \int_{x_i}^{x_{i+1}} f(x)dx - h_i \sum_{j=0}^{m} c_j f(\xi_{i,j})$$

була мінімальною для функції із заданого класу (значення $R_i(f)$ залежить від гладкості $f(x)$). Якщо $R_i = 0(h_i^{p+1})$, то кажуть, що квадратурна формула має порядок точності p.

Оскільки квадратурна формула (5.2) повинна бути справедливою для будь-якої функції $f(x)$, у тому числі й для $f(x) \equiv 1$, одержуємо

$$\sum_{j=0}^{m} c_j = 1. \qquad (5.3)$$

Для побудови формули чисельного інтегрування на всьому відрізку $[a; b]$ досить побудувати квадратурну формулу для інтеграла $\int_{x_i}^{x_{i+1}} f(x)dx$ на елементарному відрізку $[x_i, x_{i+1}]$ і скористатися формулою (5.1), яку називають складеною квадратурною формулою. Похибка складеної квадратурної формули дорівнює сумі похибок квадратурної формули на елементарних відрізках

$$R_n(f) = I(f) - I_n(f) = \sum_{i=0}^{n-1} R_i .$$

Різні групи методів відрізняються за способом вибору вузлів та вагових коефіцієнтів.

5.1. Загальна похибка чисельного інтегрування

Нехай значення інтеграла визначається за складеною квадратурною формулою (5.1), яка характеризується похибкою $R_n(f)$. Тоді загальна похибка обчислення інтеграла дорівнює сумі похибки квадратурної формули $R_n(f)$ і обчислювальної похибки квадратури ΔI_n [1]:

$$\Delta I_\Sigma = \Delta I_n + R_n(f).$$

Обчислювальна похибка квадратурної формули пов'язана з похибкою запису вагових коефіцієнтів Δc_i і похибкою обчислення функції $f(x)$ у вузлах. Тому обчислювальна похибка квадратури (5.1) дорівнює

$$\Delta I_n = \sum_{i=0}^{n-1}\left(\max\left|f(x_i)\right|\Delta c_i + c_i \max\left|\frac{df(x_i)}{dx}\right|\Delta x_i \right), \qquad (5.4)$$

де Δc_i – абсолютна похибка вагових коефіцієнтів; Δx_i – абсолютна похибка вузлів.

Нехай відносна похибка округлення дійсних чисел дорівнює $\varepsilon_{маш}$. Тоді, якщо похибка запису вагових коефіцієнтів і вузлів пов'язана тільки з похибкою округлення, то $\Delta c_i = \varepsilon_{маш} c_i$, $\Delta x_i = \varepsilon_{маш} x_i$. Отже, для похибки (5.4) можна записати

$$\Delta I_n \le n\varepsilon_{маш}(A_1 + A_2) = nA\varepsilon_{маш},$$

де $A_1 \ge c_i \max\left|f(x_i)\right|$, $\forall\, i = \overline{0,n}$, $A_2 \ge c_i x_i \max\left|\frac{df(x_i)}{dx}\right|$, $\forall\, i = \overline{0,n}$, $A = A_1 + A_2$.

Таким чином, обчислювальна похибка квадратурної формули лінійно зростає зі збільшенням кількості вузлів.

Похибка квадратурної формули $R_n(f)$ оцінюється як

$$R_n(f) \le B_1 h^p,$$

де p – стала, що визначається порядком квадратурної формули; h – крок сітки. З огляду на те, що $h = \dfrac{b-a}{n}$, маємо

$$R_n(f) \le \frac{B_1(b-a)^p}{n^p} = \frac{B}{n^p}.$$

Таким чином, похибка дискретизації квадратурної формули зменшується зі зростанням n.

Отже, існує абсолютна похибка $\Delta I_{\Sigma\min}$ для будь-якої квадратурної формули, яку не можна зменшити, збільшуючи число кроків n (рис. 5.1).

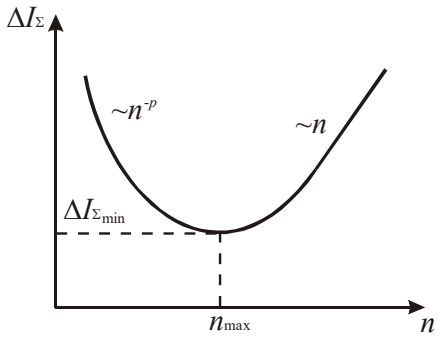


Рис. 5.1. Загальна похибка чисельного інтегрування

Конкретне значення $\Delta I_{\min}$ залежить від значення A_1, A_2, B_1, a, b, $\varepsilon_{маш}$. Якщо останні можна оцінити, то значення $n_{\max}$ і $\Delta I_{\Sigma\min}$ можна знайти з умови

$$\frac{dI_\Sigma}{dn} = \frac{d}{dn}\left(A\varepsilon_{маш} n + \frac{B}{n^p}\right) = 0.$$

Звідси $n_{\max} = \sqrt[p+1]{\dfrac{pB}{A\varepsilon_{маш}}}$.

5.2. Формули Ньютона-Котеса

Формули Ньютона-Котеса одержують інтегруванням інтерполяційних поліномів, побудованих за рівномірною сіткою [4, 27]. Розрізняють формули відкритого і закритого типів. У формулах закритого типу у розрахунках використовують значення функції, обчислені в обох кінцях елементарного відрізка, а у формулах відкритого типу, принаймні, одне зі значень участі у розрахунках не бере.

Розглянемо побудову формули Ньютона-Котеса m-го порядку закритого типу. Для цього розіб'ємо елементарний відрізок $[x_i, x_{i+1}]$ на m рівних частин і обчислимо значення функції у вузлах $f(\xi_{i,j})$, де $\xi_{i,j} = x_i + j\frac{h_i}{m}$, $j = \overline{0,m}$. Побудуємо інтерполяційний поліном Лагранжа на відрізку $[x_i, x_{i+1}]$ по вузлах $(\xi_{i,j}, f(\xi_{i,j}))$, $j = \overline{0,m}$:

$$f(x) \approx P_m(x) = \sum_{j=0}^{m} \prod_{k=0,k\neq j}^{m} \frac{x-\xi_{i,k}}{\xi_{i,j}-\xi_{i,k}} f_{i,j} = \sum_{j=0}^{m} \frac{(-1)^{m-j}}{j!(m-j)!} \prod_{k=0,k\neq j}^{m} (t-k) f_{i,j}, \quad (5.5)$$

де $f_{i,j} = f(\xi_{i,j}) = f(x_i + j\frac{h_i}{m})$, $t = \dfrac{x-x_i}{\frac{h_i}{m}}$.

Тоді, інтегруючи (5.5) і виконуючи заміну змінних $x \to t$, $dx \to \frac{h_i}{m}dt$, одержимо

$$\int_{x_i}^{x_{i+1}} f(x)dx = h_i \sum_{j=0}^{m} c_j f_{i,j}, \quad (5.6)$$

де c_j – коефіцієнти Котеса, дорівнюють

$$c_j = \frac{(-1)^{m-j}}{j!(m-j)!}\frac{1}{m}\int_0^m \prod_{k=0,k\neq j}^{m}(t-k)dt. \qquad (5.7)$$

Формулу (5.6) називають квадратурною формулою Ньютона-Котеса. З виразу (5.7) випливає, що коефіцієнти Котеса c_j не залежать від функції $f(x)$. Тоді, вважаючи, що $f(x)\equiv 1$, з (5.6) маємо

$$\sum_{j=0}^{m} c_j = 1.$$

Крім цього, з (5.7) випливає, що

$$c_j = c_{m-j}.$$

Отже, коефіцієнти Котеса симетричні відносно центра елементарного відрізка. Значення коефіцієнтів Котеса до десятого порядку включно наведено у табл. 5.1.

Похибку формули Ньютона-Котеса оцінюють як

$$R_i = \begin{cases} a_i f^{(m+2)}(\xi_i)h_i^{m+3}, & m = 2l \\ a_i f^{(m+1)}(\xi_i)h_i^{m+2}, & m = 2l-1 \end{cases}, \ l = 1,\ 2,\ldots \qquad (5.8)$$

Коефіцієнти Котеса розраховують за формулою

$$c_i = \frac{\alpha_i}{\sum_{j=1}^{m}\alpha_j}.$$

Зазначимо, що формули Ньютона-Котеса парного порядку за рахунок симетрії розбиття елементарного відрізка мають додатковий порядок точності. З цієї причини найчастіше використовують формули парного порядку.

Таблиця 5.1. **Сталі для розрахунку коефіцієнтів Котеса і похибка формули Ньютона-Котеса порядку *m*** [4, 25, 27]

M	α_0	α_1	α_2	α_3	α_4	α_5	$\sum_{j=0}^{m}\alpha_j$	R_i
1	1	1					2	$-\frac{1}{12}h_i^3 f''(\xi)$
2	1	4	1				6	$-\frac{1}{90\cdot 2^5}h_i^5 f^{IV}(\xi)$
3	1	3	3	1			8	$-\frac{3}{80\cdot 3^5}h_i^5 f^{IV}(\xi)$
4	7	32	12	32	7		90	$-\frac{8}{945\cdot 4^7}h_i^7 f^{VI}(\xi)$
5	19	75	50	50	75	19	288	$-\frac{275}{12096\cdot 5^7}h_i^7 f^{VI}($
6	41	216	27	272	27	216	840	$-\frac{9}{1\,400\cdot 6^9}h_i^9 f^{VIII}($
7	751	3 577	1 323	2 989	2 989	1 323	17 280	$-\frac{8\,183}{518400\cdot 7^9}h_i^9 f^{VI}$
8	989	5 888	– 928	10 496	– 4 540	10 496	28 350	$\sim h_i^{11} f^{X}(\xi)$
9	2 857	15 741	1 080	19 344	5 778	5 778	89 600	$\sim h_i^{11} f^{X}(\xi)$
10	16 067	106 300	– 48 525	272 400	– 260 550	427 368	598 752	$\sim h_i^{13} f^{XII}(\xi)$

Із формули (5.8) випливає, що для зменшення похибки квадратурної формули Ньютона-Котеса необхідно збільшувати її порядок. Однак формули з $m \geq 10$ рідко використовують через їх чисельну нестійкість, що призводить до різкого зростання обчислювальної похибки. Причиною такої нестійкості є те, що коефіцієнти c_j у формулі (5.6) для великих m мають різні знаки. Деякі коефіцієнти стають від'ємними для $m = 8$. У разі $m = 9$ вони всі додатні, але для $m \geq 10$ існують як додатні, так і від'ємні значення.

Приклад. Обчислимо інтеграл $I=\int\limits_0^1 \frac{dx}{1+x^2}$. Точне значення інтеграла $I=\frac{\pi}{4}\approx 0,785398$. За формулою Ньютона-Котеса 4-го порядку згідно з (5.6) для одного елементарного відрізка маємо

$$I\approx \frac{1}{90}\left(7f\left(0\right)+32f\left(0,25\right)+12f\left(0,5\right)+32f\left(0,75\right)+7f\left(1\right)\right)=0,785529.$$

5.3. Формули Чебишова

Як вже було показано, у формулах Ньютона-Котеса вузли інтерполяції розміщуються рівномірно, а вагові коефіцієнти обчислюють так, щоб формула була точною для полінома степеня, не нижчого m [28]. У формулах Чебишова всі вагові коефіцієнти задають однаковими, а вузли вибирають так, щоб формула була точною для полінома степеня, не нижчого від $m+1$. Ураховуючи, що формула має бути точною для $f(x)\equiv 1$, із (5.3) маємо:

$$c_j=c=\frac{1}{m+1},$$

$$\int\limits_{x_i}^{x_{i+1}} f(x)dx=\frac{h_i}{m+1}\sum_{j=0}^{m} f(\xi_{i,j}). \qquad (5.9)$$

Для визначення вузлів $\xi_{i,j}$ проведемо в (5.9) заміну змінних

$$x=\frac{x_{i+1}+x_i}{2}+\frac{h_i}{2}t,$$

тоді

$$\xi_{i,j}=\frac{x_{i+1}+x_i}{2}+\frac{h_i}{2}t_j. \qquad (5.10)$$

Із рівняння (5.9) одержуємо

$$\int_{-1}^{1} f\left(\frac{x_{i+1}+x_i}{2}+\frac{h_i}{2}t\right)dt=\frac{2}{m+1}\sum_{j=0}^{m} f\left(\frac{x_{i+1}+x_i}{2}+\frac{h_i}{2}t_j\right). \qquad (5.11)$$

Введемо позначення $f\left(\frac{x_{i+1}+x_i}{2}+\frac{h_i}{2}t\right)=g(t)$. Тоді з (5.11)

$$\int_{-1}^{1} g(t)dt=\frac{2}{m+1}\sum_{j=0}^{m} g(t_j). \qquad (5.12)$$

Для визначення вузлів t_j використовуємо умову, що формула (5.12) є точною для поліномів $g(t)=t,\ t^2,\ldots,\ t^{m+1}$. Підставляючи ці функції в (5.12), одержуємо систему рівнянь:

$$\begin{gathered}
t_0+t_1+\ldots+t_m=0;\\
t_0^2+t_1^2+\ldots+t_m^2=\frac{m+1}{3};\\
t_0^3+t_1^3+\ldots+t_m^3=0;\\
t_0^4+t_1^4+\ldots+t_m^4=\frac{m+1}{5};\\
\ldots\\
t_0^{m+1}+t_1^{m+1}+\ldots+t_m^{m+1}=\frac{(m+1)(1-(-1)^{m+1})}{2(m+2)}.
\end{gathered} \qquad (5.13)$$

Із виразів (5.13) можна знайти t_j , $j=\overline{0,m}$, а потім з (5.10) – $\xi_{i,j}$. Нижче наведено t_j для порядків $m=\overline{1,4}$:

$$\begin{gathered}
m=1;\ -t_0=t_1=\frac{1}{\sqrt{3}};\\
m=2;\ -t_0=t_2=0,707107;\ \ t_1=0;\\
m=3;\ -t_0=t_3=0,794654;\ -t_1=t_2=0,187592;\\
m=4;\ -t_0=t_4=0,832498;\ \ -t_1=t_3=0,374541;\ t_2=0.
\end{gathered}$$

Методи Чебишова мають порядок точності m + 2.

Для прикладу обчислимо інтеграл $I = \int_0^1 \frac{dx}{1+x^2}$ з попереднього прикладу методом Чебишова третього порядку за одним елементарним відрізком.

Згідно з (5.10) знайдемо розташування вузлів формули Чебишова:

$$\xi_{0,0} = \frac{0+1}{2} + \frac{1-0}{2} \cdot \left(-0.794654\right) = 0.102673;$$

$$\xi_{0,1} = \frac{0+1}{2} + \frac{1-0}{2} \cdot \left(-0.187592\right) = 0.406204;$$

$$\xi_{0,2} = \frac{0+1}{2} + \frac{1-0}{2} \cdot 0.187592 = 0.593796;$$

$$\xi_{0,3} = \frac{0+1}{2} + \frac{1-0}{2} \cdot 0.794654 = 0.897327.$$

За формулою (5.9) обчислимо наближене значення інтегралу:

$$I \approx \frac{1}{4}\left(\frac{1}{1+0{,}102673^2} + \frac{1}{1+0{,}406204^2} + \frac{1}{1+0{,}593796^2} + \frac{1}{1+0{,}897327^2} \right) \approx$$
$$\approx 0{,}785303.$$

5.4. Формули Гауса

На відміну від формул Ньютона-Котеса та Чебишова, у виведеній з виразу (5.2) формулі Гауса вузли і вагові коефіцієнти не задаються, а визначаються так, щоб формула була точною для полінома найвищого можливого степеня [28]. Можна показати, що ця вимога виконується, якщо вузли обчислюються за формулою (5.10), причому t_j є коренями полінома Лежандра $P_{m+1}(t)$ степеня $m+1$, а вагові коефіцієнти обчислюються за формулою [29]:

$$c_j = \frac{\left[P_{m+1}(1)\right]^2}{\left(1-t_j^2\right)\left[P'_{m+1}\left(t_j\right)\right]^2}.$$

Поліноми Лежандра знаходять за виразом

$$P_m(x) = \frac{1}{2^m m!}\frac{d^m}{dx^m}(x^2-1)^m,$$

або за рекурентною формулою

$$P_{m+1}(x) = \frac{2m+1}{m+1}xP_m(x) - \frac{m}{m+1}P_{m-1}(x) = xP_m(x) + \frac{x^2}{m+1}\frac{dP_m(x)}{dx};$$

$$P_0(x) = 1, \quad P_1(x) = x;$$

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

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3).$$

Деякі значення параметрів формули Гауса для методів 1 – 4-го порядків наведено нижче:

$m = 1$: $-t_0 = t_1 = \frac{1}{\sqrt{3}}$; $\quad c_0 = c_1 = \frac{1}{2}$;

$m = 2$: $-t_0 = t_2 = \sqrt{\frac{3}{5}}; t_1 = 0$; $\quad c_0 = c_2 = \frac{5}{18}; c_1 = \frac{4}{9}$;

$m = 3$; $-t_0 = t_3 \cong 0{,}861136$; $\quad c_0 = c_3 \cong 0{,}173928$;

$-t_1 = t_2 \cong 0{,}339981$; $\quad c_1 = c_2 \cong 0{,}326072$;

$m = 4$; $-t_0 = t_4 \cong 0{,}9061798$; $\quad c_0 = c_4 \cong 0{,}1184634$;

$-t_1 = t_3 = 0{,}5384693$; $\quad c_1 = c_3 \cong 0{,}2393143$;

$t_2 = 0$; $\quad c_2 \cong 0{,}2844444$.

Методи Гауса мають порядок точності $2m + 1$ [27].

Приклад. Обчислимо інтеграл $I = \int_0^1 \frac{dx}{1+x^2}$ методом Гауса 3-го порядку за одним елементарним відрізком, $x_0 = 0, x_1 = 1$.

$$m = 3;\ -t_0 = t_3 \cong 0{,}861136;\ c_0 = c_3 \cong 0{,}173928;$$
$$-t_1 = t_2 \cong 0{,}339981;\ c_1 = c_2 \cong 0{,}326072.$$

Згідно з (5.10) знайдемо розташування вузлів формули Гауса:

$$\xi_{0,0} = \frac{0+1}{2} + \frac{1-0}{2} \cdot \left(-0.861136\right) = 0.069432;$$
$$\xi_{0,1} = \frac{0+1}{2} + \frac{1-0}{2} \cdot \left(-0.339981\right) = 0.330010;$$
$$\xi_{0,2} = \frac{0+1}{2} + \frac{1-0}{2} \cdot 0.339981 = 0.669991;$$
$$\xi_{0,3} = \frac{0+1}{2} + \frac{1-0}{2} \cdot 0.861136 = 0.930568.$$

Тоді згідно з формулою (5.2) маємо

$$I \approx 1 \cdot \left(\begin{array}{l} 0{,}173928 \cdot \dfrac{1}{1+0{,}069432^2} + 0{,}326072 \cdot \dfrac{1}{1+0{,}330010^2} + \\ +0{,}326072 \cdot \dfrac{1}{1+0{,}669991^2} + 0{,}173928 \cdot \dfrac{1}{1+0{,}930568^2} \end{array} \right) =$$
$$= 0{,}785403$$

5.5. Апостеріорна оцінка похибки інтегрування

Відомо, що похибка інтегрування залежить від довжини елементарного відрізка, на якому застосовують квадратурну формулу. Виходячи із цього, для зменшення похибки довжину елементарного відрізка слід зменшувати, але таке зменшення значно збільшує обсяг розрахунків і, відповідно, обчислювальну похибку. Тому важливо вміти оцінити похибку інтегрування в процесі розрахунку.

Зробити це можна за допомогою першого правила Рунге [22], згідно з яким

$$R = \frac{I_{h/q} - I_h}{q^p - 1}, \qquad (5.14)$$

де p – порядок точності методу; q – ціле число, як правило 2.

Таким чином, для оцінки похибки треба розрахувати інтеграл два рази. Спочатку розраховують інтеграл, розбивши відрізок інтегрування на елементарні відрізки довжиною h, потім на елементарні відрізки довжиною h/q. Після цього за виразом (5.14) оцінюють похибку.

Використовуючи правило Рунге, можна не лише оцінити похибку, але й уточнити значення інтеграла

$$I \approx I_{h/q} + R = I_{h/q} + \frac{I_{h/q} - I_h}{q^p - 1}.$$

Такий спосіб називається екстраполяцією Річардсона. Він дозволяє підвищити точність квадратурної формули не менше, ніж на порядок. Якщо з якоїсь причини порядок точності квадратурної формули невідомий, то його можна оцінити за допомогою другого правила Рунге:

$$p = \frac{\ln\left(\dfrac{I_h - I_{h/q}}{I_{h/q} - I_{h/q^2}}\right)}{\ln(q)}.$$

Основним параметром, за допомогою якого впливають на величину похибки інтегрування, є довжина елементарного відрізка h_i. Для забезпечення високої точності крок інтегрування слід зменшувати. Однак для ефективного розв’язання задачі крок інтегрування слід вибирати найбільшим із можливих. Таким чином, величина кроку інтегрування визначається із суперечливих вимог похибки та ефективності. Величина оптимального кроку інтегрування залежить від характеру підінтегральної функції і може суттєво змінюватися на відрізку [a, b].

Можливість апостеріорної оцінки похибки інтегрування за правилами Рунге дозволяє обчислювати шуканий інтеграл із заданою похибкою шляхом автоматичного вибору кроку інтегрування h_i. Методи, в яких крок інтегрування визначається автоматично з умови заданої точності, називаються адаптивними.

Нехай використовується складена квадратурна формула

$$I_n = \sum_{i=0}^{n-1} I_{h,i},$$

де $I_{h,i}$ — квадратурна сума на елементарному відрізку, причому на кожному елементарному відрізку використовується одна і та ж квадратурна формула. Нехай відомий порядок квадратурної формули p, і на кожному елементарному відрізку $[x_i,x_{i+1}]$ усі обчислення виконано двічі з кроками h_i і h_i/q, а похибка оцінена за правилом Рунге. Якщо на кожному елементарному відрізку для заданого $\Delta>0$ виконується нерівність

$$\left|R_i\right| = \frac{\left|I_{h/q,i} - I_{h,i}\right|}{q^p - 1} \le \frac{\Delta\, h_i}{b-a},\ i = \overline{0,n-1}, \qquad (5.15)$$

то похибка складеної квадратурної формули на всьому відрізку:

$$\left|R_n\right| \le \frac{\Delta}{b-a}\sum_{i=0}^{n-1} h_i = \Delta\,.$$

Виконання умови (5.15) на кожному елементарному відрізку гарантує отримання розв'язку із заданою абсолютною похибкою Δ. Якщо задається відносна похибка ε, то Δ можна оцінити як $\Delta = \varepsilon\left|I_n\right|$.

Якщо на деякому з елементарних відрізків умова (5.15) не виконується, то на цьому відрізку крок інтегрування слід зменшити ще в q раз і повторити розрахунки. Зменшення кроку слід продовжувати до тих пір, поки не буде досягнуто виконання умови (5.15).

Навпаки, якщо оцінка (5.15) виконується, то слід перевірити, чи не можна рухатися з більшим кроком. Якщо умова (5.15) виконується із значним запасом, то наступний крок слід збільшити. Якщо на *i*-му елементарному відрізку досягнуто заданої похибки, то наступний крок вибирають за правилом зон:

$$h_{i+1} = \begin{cases} h_i, & \dfrac{\Delta}{q^{p+1}} \le R \le \Delta; \\ q h_i, & R < \dfrac{\Delta}{q^{p+1}}. \end{cases}$$

Існують і інші методи вибору кроку інтегрування.

Таким чином, автоматичний вибір кроку інтегрування призводить до того, що інтегрування ведеться з крупним кроком на ділянках плавної зміни функції $f(x)$ і з дрібним кроком — на ділянках швидкої зміни $f(x)$. Це дозволяє для заданої абсолютної Δ або відносної ε похибки зменшити загальну кількість обчислення значень $f(x)$ порівняно з методами з фіксованим кроком.

Іноді подрібнення кроку інтегрування може тривати надто довго. Тому у практичних випадках слід передбачити обмеження на кількість подрібнень чи довжину елементарного відрізка.

Як правило, у разі чисельного розв’язання задається відносна похибка. Однак існують ситуації, коли оцінка тільки відносної похибки за правилом Рунге може призвести до зациклювання програми. Тому поряд з оцінкою відносної похибки слід перевіряти також і абсолютну похибку і закінчувати розрахунки на поточному кроці після досягнення однієї з них. Величина абсолютної похибки задається виходячи з фізичного змісту розв’язуваної задачі.

Контрольні завдання

1. Вибрати функцію та межі інтегрування відповідно до свого варіанта.

2. Обчислити інтеграл методом Ньютона-Котеса 4-го порядку з похибкою, не більшою 1 %.

3. Використовуючи результати виконання другого завдання, уточнити значення інтеграла за допомогою екстраполяції Річардсона.

4. Обчислити інтеграл методом Чебишова 4-го порядку з похибкою, не більшою 1 %.

5. Обчислити інтеграл методом Гауса 4-го порядку з похибкою, не більшою 1 %.

6. Порівняти результати виконання пп. 1 – 5.

Варіанти завдань

1. $0{,}37e^{(\sin(x))}$; $x \in [0;1]$.

2. $0{,}5x + x\lg(x)$; $x \in [1;2]$.

3. $(x+1{,}9)\sin\left(\frac{x}{3}\right)$; $x \in [1;2]$.

4. $\ln(x+2)/x$; $x \in [2;3]$.

5. $3\cos(x)/(2x+1{,}7)$; $x \in [0;1]$.

6. $(2x+0{,}6)\cos\left(\frac{x}{2}\right)$; $x \in [1;2]$.

7. $2{,}6x^2\ln(x)$; $x \in [1{,}2;2{,}2]$.

8. $(x^2+1)\sin(x-0{,}5)$; $x \in [0{,}5;1{,}5]$.

9. $x^2\cos\left(\frac{x}{4}\right)$; $x \in [2;3]$.

10. $\sin(0{,}2x+3)(x^2+1)$; $x \in [3;4]$.

11. $3x + \ln(x)$; $x \in [1;2]$.

12. $4xe^x$; $x \in [-1;0]$.

13. $3x^2 + tg(x)$; $x \in [-0{,}5;0{,}5]$.

14. $3x^2 + \sin(x)$; $x \in [0;1]$.

15. $3xe^{\cos(x)}$; $x \in [0{,}2;1{,}2]$.

16. $x^2 tg\left(\frac{x}{2}\right)$; $x \in [0{,}5;1{,}5]$.

17. $x^2 \Big/ (1+0,25x)$; $x \in [1,1; 2,1]$.

18. $(x^3 - 0,3x) \Big/ \sqrt{(1+2x)}$; $x \in [2; 3]$.

19. $2e^{-x}(2+x^3)$; $x \in [1; 2]$.

20. $\cos(x^2)x$; $x \in [0; 1]$.

21. $\sqrt{(1+x)}\sin(x)$; $x \in [2; 3]$.

22. $e^x + x^2 - 1$; $x \in [0; 1]$.

23. $(e^x + x)\sin(x)$; $x \in [0; 1]$.

24. $\sqrt{(3+x)}\lg(x)$; $x \in [1; 2]$.

25. $(4+x)\sin(x^2)$; $x \in [1; 2]$.

26. $xe^{\sin(x)}$; $x \in [2; 3]$.

27. $\sin(x)\cos(x)x$; $x \in [2; 3]$.

28. $\sin(x)\ln(x)$; $x \in [1; 2]$.

29. $\cos(x)\ln(x)$; $x \in [3; 4]$.

30. $x^2 \lg(x)$; $x \in [2; 3]$.

6. Чисельне інтегрування звичайних диференціальних рівнянь

Значна частина інженерних задач зводиться до розв'язання звичайних диференціальних рівнянь (ЗДР). Розрізняють три типи задач для ЗДР: задачі Коші, крайові задачі та задачі на власні числа.

Для ЗДР першого порядку задача Коші полягає в знаходженні такого розв'язку рівняння [27]:

$$\frac{dy}{dt} = f\left(t,\, y\right), \tag{6.1}$$

що задовольняє початкову умову

$$y\left(0\right) = y_0, \tag{6.2}$$

де $f\left(t,\, y\right)$ – задана неперервна функція двох аргументів, y_0 – задана константа.

Для диференціального рівняння *m*-го порядку

$$\frac{d^m y}{dt^m} = f\left(t, y, \frac{dy}{dt}, \ldots \frac{d^{m-1} y}{dt^{m-1}}\right) \tag{6.3}$$

задача Коші полягає в знаходженні функції $y = y(t)$, що задовольняє рівняння (6.3) і початкові умови:

$$y\left(0\right) = y_0;$$
$$\frac{dy\left(0\right)}{dt} = y_0^{(1)};$$
$$\ldots$$
$$\frac{d^{m-1} y\left(0\right)}{dt^{m-1}} = y_0^{(m-1)}.$$

Розв’язуючи деякі задачі, слід знайти m функцій $y_1(t), y_2(t),\ldots, y_m(t)$, що задовольняють систему диференціальних рівнянь:

$$\begin{aligned}
\frac{dy_1}{dt} &= f_1\left(t,\, y_1,\, y_2,\ldots,\, y_m\right);\\
\frac{dy_2}{dt} &= f_2\left(t,\, y_1,\, y_2,\ldots, y_m\right);\\
&\ldots\\
\frac{dy_m}{dt} &= f_m\left(t,\, y_1,\, y_2,\ldots,\, y_m\right)
\end{aligned} \tag{6.4}$$

і початкові умови

$$\begin{aligned}
&y_1(0) = y_{10};\\
&y_2(0) = y_{20};\\
&\ldots\\
&y_m(0) = y_{m0}.
\end{aligned} \tag{6.5}$$

Систему диференціальних рівнянь (6.4) називають нормальною [22].

Задачу Коші (6.4) з початковими умовами (6.5) можна подати у векторній формі

$$\begin{aligned}
\frac{d\mathbf{Y}}{dt} &= F(t, \mathbf{Y});\\
\mathbf{Y}(0) &= \mathbf{Y}_0,
\end{aligned} \tag{6.6}$$

де $\mathbf{Y} = \left[y_1(t),\, y_2(t),\ldots, y_m(t)\right]^{\mathrm{T}}$, $F(t,\, \mathbf{Y}) = \left[f_1(t,\, \mathbf{Y}), f_2(t,\, \mathbf{Y}),\ldots, f_m(t,\, \mathbf{Y})\right]^{\mathrm{T}}$,
$\mathbf{Y}_0 = \left[y_{10},\, y_{20},\ldots,\, y_{m0}\right]^{\mathrm{T}}$.

Систему диференціальних рівнянь

$$\frac{dy_1}{dt} = a_{11}y_1 + a_{12}y_2 + \cdots + a_{1m}y_m + b_1;$$
$$\frac{dy_2}{dt} = a_{21}y_1 + a_{22}y_2 + \cdots + a_{2m}y_m + b_2;$$
$$\cdots$$
$$\frac{dy_m}{dt} = a_{m1}y_1 + a_{m2}y_2 + \cdots + a_{mm}y_m + b_m,$$

у якій $a_{ij},\ b_j,\ i, j = \overline{1,m}$ – функції від t, називають лінійною.

Якщо ввести позначення

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ & \ldots & & \\ a_{m1} & a_{m2} & \ldots & a_{mm} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_m \end{bmatrix},$$

то систему лінійних диференціальних рівнянь записують у матричній формі:

$$\frac{d\mathbf{Y}}{dt} = \mathbf{AY} + \mathbf{B}.$$

Рівняння (6.3) за допомогою заміни змінних

$$y_1(t) = \frac{dy}{dt};$$
$$y_2(t) = \frac{d^2y}{dt^2} = \frac{dy_1}{dt};$$
$$\cdots$$
$$y_{m-1}(t) = \frac{d^{m-1}y}{dt^{m-1}} = \frac{dy_{m-2}}{dt}$$

можна звести до вигляду (6.6). Наприклад, рівняння

$$\frac{d^2y}{dt^2} = f\left(t,\ y,\ \frac{dy}{dt}\right)$$

можна записати як систему

$$\frac{dy}{dt} = y_1;$$
$$\frac{dy_1}{dt} = f(t, y, y_1).$$

Отже, задачу Коші будь-якого порядку можна звести до задачі (6.6). Слід зазначити, що методи розв'язання задачі Коші (6.1) з початковою умовою (6.2) з однаковим успіхом можна застосувати до систем (6.6). Тому надалі зупинимося на методах розв'язання задачі Коші для одного рівняння, хоча усі формули застосовні до систем диференціальних рівнянь з урахуванням того, що всі скалярні операції заміняються операціями над векторами, матрицями і тензорами.

Для чисельного розв'язання задачі Коші по змінній t вводять сітку $0<t_1<t_2<...<t_n$ та шукають значення невідомої функції в вузлах сітки. Позначимо як $y(t)$ — точний розв'язок задачі (6.1), (6.2), а y_{i+1} — наближений розв'язок в точці $t=t_{i+1}$. Локальною похибкою розв'язку називають

$$R_i = y\left(t_{i+1}\right) - y_{i+1} \qquad (6.7)$$

У разі використання наближених методів основним є питання про збіжність. Найбільше розповсюдження отримало поняття про збіжність за умови $h_i \to 0$, де

$$h_i = t_{i+1} - t_i.$$

Говорять, що метод збігається в точці t_{i+1}, якщо $\lim\limits_{h_i \to 0} \left|R_i\right| = 0$. Метод збігається на відрізку $[0,t_n]$, якщо він збігається в кожній точці $t \in [0,t_n]$.

Говорять, що метод має p-ий порядок точності, якщо існує таке число $p>0$, що $R_i = O\left(h_i^{p+1}\right)$, за умови $h_i \to 0$.

6.1. Метод Ейлера

Для чисельного розв'язання задачі Коші (6.1), (6.2) розіб'ємо відрізок $[0,\ t_n]$, на якому шукається розв'язок, на елементарні відрізки $[t_i,t_{i+1}]$, $i=\overline{0,n-1}$, $t_0=0$. На кожному елементарному відрізку будемо шукати розв'язок у вигляді полінома першого степеня. Нехай значення функції на границях елементарного відрізку становить y_i, y_{i+1}. За інтерполяційною формулою Лагранжа

$$y(t)=\frac{t-t_i}{t_{i+1}-t_i}y_{i+1}+\frac{t-t_{i+1}}{t_i-t_{i+1}}y_i=\frac{t-t_i}{h_i}y_{i+1}-\frac{t-t_{i+1}}{h_i}y_i.$$

Тоді

$$\frac{dy}{dt}=\frac{y_{i+1}-y_i}{h_i}. \tag{6.8}$$

Підставляючи (6.8) в (6.1) та замінюючи функцію $f(t,y)$ її значенням на початку елементарного відрізка, отримаємо

$$\frac{y_{i+1}-y_i}{h_i}=f(t_i,y_i). \tag{6.9}$$

В методі Ейлера рівняння (6.1) замінюється різницевим рівнянням (6.9). Геометрично така заміна означає, що на кожному елементарному відрізку розв'язок задачі Коші шукається як пряма, тангенс кута нахилу якої дорівнює значенню функції на початку елементарного відрізка. З (6.9) випливає, що наближений розв'язок задачі Коші знаходиться явно за рекурентною формулою

$$y_{i+1}=y_i+h_i f\left(t_i,y_i\right) \tag{6.10}$$

Оцінимо локальну похибку методу Ейлера. Для цього припустимо, що в точці $t=t_i$ значення функції $y(t)$ обчислено точно, тобто $y(t_i)=y_i$.

Розкладаючи функцію $y(t)$ в ряд Тейлора в околі точки $t=t_i$ і враховуючи (6.1), отримаємо

$$y(t) = y_i + f(y_i, t_i)(t - t_i) + \frac{y''(t_i)}{2}(t - t_i)^2 + \ldots$$

Звідки

$$y(t_{i+1}) = y_i + f(y_i, t_i)h_i + \frac{y''(t_i)}{2}h_i^2 + \ldots \qquad (6.11)$$

Враховуючи (6.10), вважаючи крок інтегрування h_i малим та нехтуючи малими величинами вищих порядків із (6.7) та (6.11) отримуємо

$$R_i = \frac{y''(t_i)}{2}h_i^2 .$$

Таким чином, метод Ейлера має перший порядок точності. Тому в методі Ейлера зменшення кроку h_i в два рази в чотири рази зменшує локальну похибку розв'язку.

Існують інші різновиди методу. Так, підставляючи (6.8) до (6.1) і замінюючи функцію $f(t,y)$ її значенням у кінці елементарного відрізку, отримаємо

$$\frac{y_{i+1} - y_i}{h_i} = f(y_{i+1}, t_{i+1}) .$$

Звідки

$$y_{i+1} = y_i + h_i f(t_{i+1}, y_{i+1}) . \qquad (6.12)$$

Для довільної функції $f(t,y)$ значення y_{i+1} не може бути виражене з (6.12) явно. Тому метод (6.12) називають неявним методом Ейлера. У разі використання неявного методу Ейлера, нове значення y_{i+1} визначається на основі попереднього y_i шляхом розв'язання нелінійного рівняння (6.12). Неявний метод Ейлера також має перший порядок точності. Можна показати, що похибка неявного методу Ейлера дорівнює

$$R_i = -\frac{y''(t_{i+1})}{2} h_i^2 .$$

Приклад. Нехай дано задачу Коші

$$\begin{cases} \dfrac{dy}{dt} = -(t+1)y^2 \\ y(0) = 1 \end{cases}.$$

Точним розв'язком цієї задачі є функція $y(t) = \dfrac{2}{(t+1)^2 + 1}$.

Необхідно обчислити таблицю значень $y(t)$ на відрізку $t \in [0,1]$ з кроком h=0,25, тобто для t_0=0, t_1=0,25, t_2=0,5, t_3=0,75, t_4=1.

Скористаємося явним методом Ейлера. Оскільки $f(t,y)$=$-y^2(t$+1), то згідно з формулою (6.10)

$$y_{i+1} = y_i + h \cdot \left(-(t_i + 1) y_i^2\right) = y_i - h(t_i + 1) y_i^2 .$$

Тоді

$$y_0 = y(0) = 1;$$

$$y_1 = y(0,25) = y_0 - h(t_0 + 1) y_0^2 = 1 - 0,25 \cdot (0+1) \cdot 1^2 = 0,75;$$

$$y_2 = y(0,5) = y_1 - h(t_1 + 1) y_1^2 = 0,75 - 0,25 \cdot (0,25+1) \cdot 0,75^2 \cong 0,574219;$$

$$y_3 = y(0,75) = y_2 - h(t_2 + 1) y_2^2 = 0,574219 - 0,25 \cdot (0,5+1) \cdot 0,574219^2 \cong 0,450571;$$

$$y_4 = y(1) = y_3 - h(t_3 + 1) y_3^2 = 0,450571 - 0,25 \cdot (0,75+1) \cdot 0,450571^2 \cong 0,361752;$$

Розв'яжемо цю ж задачу неявним методом Ейлера. Згідно з формулою (6.12) маємо

$$y_{i+1} = y_i + h \cdot \left(-(t_{i+1} + 1) y_{i+1}^2\right) = y_i - h(t_{i+1} + 1) y_{i+1}^2,$$

що приводить до рівняння відносно y_{i+1}:

$$(t_{i+1}+1)hy_{i+1}^2 + y_{i+1} - y_i = 0.$$

Тоді

$$y_0 = y(0) = 1;$$

$$(t_1+1)hy_1^2 + y_1 - y_0 = 0 \Rightarrow (0,25+1)\cdot 0,25\cdot y_1^2 + y_1 - 1 = 0 \Rightarrow y_1 = 0.8;$$

$$(t_2+1)hy_2^2 + y_2 - y_1 = 0 \Rightarrow (0,5+1)\cdot 0,25\cdot y_2^2 + y_2 - 0,8 = 0 \Rightarrow y_2 \cong 0,644320;$$
$$(t_3+1)hy_3^2 + y_3 - y_2 = 0 \Rightarrow (0,75+1)\cdot 0,25\cdot y_3^2 + y_3 - 0.644320 = 0 \Rightarrow y_3 \cong 0,524132;$$
$$(t_4+1)hy_4^2 + y_4 - y_3 = 0 \Rightarrow (1+1)\cdot 0,25\cdot y_4^2 + y_4 - 0,524132 = 0 \Rightarrow y_4 \cong 0,431175.$$

Як вже зазначалося, методи розв'язку задачі Коші (6.1), (6.2) можна застосовувати і для розв'язання систем ЗДР (6.6). Для прикладу покажемо, як застосувати до розв'язання систем ЗДР явний метод Ейлера. Нехай задано систему ЗДР з початковими умовами

$$\begin{cases} \dfrac{dx}{dt} = x + t; \\ \dfrac{dy}{dt} = x + y; \end{cases} \quad x(0) = x_0 = 1,\ y(0) = y_0 = 2.$$

Розв'язок будемо шукати на відрізку $t \in [0;\ 0,2]$ з кроком $h = 0,1$. Використовуючи формулу (6.10) для системи ЗДР матимемо:

$$\begin{cases} x_{i+1} = x_i + h\cdot(x_i + t_i) \\ y_{i+1} = y_i + h\cdot(x_i + y_i) \end{cases}$$

$$t_1 = 0,1:$$

$$\begin{cases} x_1 = x_0 + h\cdot(x_0 + t_0) \\ y_1 = y_0 + h\cdot(x_0 + y_0) \end{cases} \Rightarrow \begin{cases} x_1 = 1 + 0,1\cdot(1+0) \\ y_1 = 2 + 0,1\cdot(1+2) \end{cases} \Rightarrow \begin{cases} x_1 = 1,1; \\ y_1 = 2,3. \end{cases}$$

$$t_2 = 0,2:$$

$$\begin{cases} x_2 = x_1 + h\cdot(x_1 + t_1) \\ y_2 = y_1 + h\cdot(x_1 + y_1) \end{cases} \Rightarrow \begin{cases} x_2 = 1,1 + 0,1\cdot(1,1+0,1) \\ y_2 = 2,3 + 0,1\cdot(1,1+2,3) \end{cases} \Rightarrow \begin{cases} x_2 = 1,22; \\ y_2 = 2,64. \end{cases}$$

6.2. Методи Рунге – Кутта

У найзагальнішому випадку метод Рунге-Кутта порядку точності p будується за рекурентною формулою [16]:

$$y_{i+1} = y_i + h_i \sum_{j=1}^{m} c_j k_j \,, \tag{6.13}$$

де

$$k_j = f\left(t_i + a_j h_i, y_i + h_i \sum_{l=1}^{j-1} b_{jl} k_l \right), \tag{6.14}$$

a_1=0, a_j та c_j – константи, b_{jl} – елементи нижньотрикутної матриці, такої, що кожне k_j отримується з попередніх значень k_l.

Формули (6.13) та (6.14) містять $\frac{m(m-1)}{2} + 2m - 1$ коефіцієнтів b_{jl}, a_j та c_j, що підлягають визначенню. Для того, щоб знайти ці коефіцієнти всі функції k_j розкладають в ряд Тейлора в околі точки (t_i, y_i). Ці розклади підставляють в (6.13) і результат порівнюють з рядом Тейлора для функції $y(t_{i+1})$. Оскільки локальна похибка визначена згідно з (6.7), то накладається умова, щоб коефіцієнти біля всіх $h_i^l, l = \overline{0, p}$ в розкладах для $y(t_{i+1})$ та y_{i+1} були рівні. Ця вимога приводить до системи рівнянь відносно коефіцієнтів b_{jl}, a_j та c_j.

Як приклад, розглянемо побудову методів Рунге-Кутта другого порядку точності (p=2). В цьому випадку можна отримати розв'язок для коефіцієнтів b_{jl}, a_j та c_j для m=2. Тоді формула (6.13) перетвориться:

$$y_{i+1} = y_i + h_i \left(c_1 k_1 + c_2 k_2 \right), \tag{6.15}$$

де $k_1 = f\left(t_i, y_i\right)$, $k_2 = f\left(t_i + a h_i, y_i + h_i b k_1\right)$.

Розкладемо k_1 і k_2 в ряд Тейлора в околі точки (t_i,y_i). Для цього відмітимо, що

$$f(t,y)=f(t_i,y_i)+(y-y_i)\frac{\partial f(t_i,y_i)}{\partial y}+(t-t_i)\frac{\partial f(t_i,y_i)}{\partial t}+\cdots.$$

Тоді

$$k_1=f(t_i,y_i), \tag{6.16}$$

$$k_2=f(t_i,y_i)+(y_i+h_ibk_1-y_i)\frac{\partial f(t_i,y_i)}{\partial y}+(t_i+ah_i-t_i)\frac{\partial f(t_i,y_i)}{\partial t}+\cdots=$$
$$=f(t_i,y_i)+h_ibf(t_i,y_i)\frac{\partial f(t_i,y_i)}{\partial y}+ah_i\frac{\partial f(t_i,y_i)}{\partial t}+\cdots. \tag{6.17}$$

Підставляючи (6.16) та (6.17) в (6.15), отримаємо

$$y_{i+1}=y_i+h_if(t_i,y_i)(c_1+c_2)+h_i^2c_2\left(bf(t_i,y_i)\frac{\partial f(t_i,y_i)}{\partial y}+a\frac{\partial f(t_i,y_i)}{\partial t}\right)+\cdots.$$

Розкладемо функцію $y(t)$ в ряд Тейлора в околі точки $(y_i,\ t_i)$ та знайдемо $y(t_{i+1})$. Тоді

$$y(t_{i+1})=y(t_i)+h_i\frac{dy(t_i)}{dt}+\frac{h_i^2}{2}\frac{d^2y(t_i)}{dt^2}+\cdots. \tag{6.18}$$

Врахуємо, що

$$\frac{dy(t_i)}{dt}=f(t_i,y_i), \tag{6.19}$$

$$\frac{d^2y(t_i)}{dt^2}=\frac{d}{dt}\left(\frac{dy(t_i)}{dt}\right)=\frac{df(t_i,y_i)}{dt}=\frac{\partial f(t_i,y_i)}{\partial y}\frac{dy(t_i)}{dt}+\frac{\partial f(t_i,y_i)}{\partial y}=$$
$$=\frac{\partial f(t_i,y_i)}{\partial y}f(t_i,y_i)+\frac{\partial f(t_i,y_i)}{\partial y}. \qquad (6.20)$$

Підставляючи (6.19) та (6.20) в (6.18), маємо

$$y(t_{i+1})=y(t_i)+h_i f(t_i,y_i)+\frac{h_i^2}{2}\left(f(t_i,y_i)\frac{\partial f(t_i,y_i)}{\partial y}+\frac{\partial f(t_i,y_i)}{\partial y}\right)+\cdots \qquad (6.21)$$

Для того, щоб метод мав другий порядок точності ($R_i=0\left(h_i^3\right)$), вимагатимемо, щоб в різниці рівнянь (6.18) та (6.21) були відсутні доданки, які містять h_i^0, h_i^1 та h_i^2. Виходячи з цієї умови, коефіцієнти a, b, c_1 та c_2 повинні задовольняти систему рівнянь:

$$\begin{cases} c_1+c_2=1; \\ c_2 b=\dfrac{1}{2}; \\ c_2 a=\dfrac{1}{2}. \end{cases} \qquad (6.22)$$

Рівняння (6.22) складають систему з 3 рівнянь відносно 4 невідомих. Виразимо b, c_1 та c_2 через $a\neq 0$. Отримаємо:

$$b=a,\quad c_2=\frac{1}{2a},\quad c_1=1-\frac{1}{2a}. \qquad (6.23)$$

Підставляючи (6.23) в (6.15) отримаємо такий узагальнений метод Рунге-Кутта другого порядку точності

$$y_{i+1} = y_i + h_i\left(\left(1-\frac{1}{2a}\right)k_1 + \frac{1}{2a}k_2\right), \qquad (6.24)$$

де $k_1 = f\left(t_i, y_i\right)$, $k_2 = f\left(t_i + ah_i, y_i + h_i a k_1\right)$.

Широковідомим методом 2 порядку є окремий випадок методу (6.24) якщо $a = \frac{1}{2}$:

$$y_{i+1} = y_i + h_i k_2,$$

де $k_1 = f\left(t_i, y_i\right)$, $k_2 = f\left(t_i + \frac{h_i}{2}, y_i + \frac{h_i}{2}k_1\right)$.

Аналогічно будуються розрахункові формули методів Рунге-Кутта вищих порядків точності. Відповідні системи рівнянь відносно b_{jl}, a_j та c_j для p=2,3,4 можуть бути розв'язані відповідно для m=2,3,4. Таким чином, для того, щоб побудувати метод порядку точності p для p=2,3 та 4 достатньо $m=p$ викликів функції $f(t,y)$. Проте для p=5 ця система може бути розв'язана тільки якщо $m\geq 6$ [22]. Таким чином, для методу Рунге-Кутта 5 порядку точності потрібно щонайменше 6 викликів функції на кожному кроці. Аналогічно для p=6 m повинно бути не менше 7. Це, зокрема, пояснює популярність класичного методу четвертого порядку, оскільки для того, щоб отримати додатковий порядок точності, потрібні ще два обчислення функції.

Слід відзначити, що методи Рунге-Кутта з p>6 чисельно нестійкі і тому на практиці не використовуються.

Як видно з вище наведеного прикладу, вибір коефіцієнтів b_{jl}, a_j та c_j не є однозначним. Тому в літературі можна знайти формули методів Рунге-Кутта, які відрізняються одна від одної, наприклад, формули Фелберга та Кеша-Карпа [22], Кутта-Мерсона [27], або більш стійкі формули методів 4-

5 порядку для m=7 [30]. Для чисельного розв'язання задачі Коші, особливий інтерес становлять такі формули, коли на одному й тому самому наборі функцій $k_j, j=\overline{1,m}$ можна одразу побудувати методи різних порядків точності. Це, як буде показано далі, дає змогу досить ефективно оцінити похибку розв'язку.

Серед всього різноманіття методів Рунге-Кутта 4 порядку слід також відмітити найпоширенішу модифікацію методу з найбільш простим набором коефіцієнтів b_{jl}, a_j та c_j:

$$\begin{gathered} y_{i+1} = y_i + \frac{h_i}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right); \\ k_1 = f\left(t_i,\, y_i\right); \quad k_2 = f\left(t_i + \frac{h_i}{2},\, y_i + \frac{h_i k_1}{2}\right); \\ k_3 = f\left(t_i + \frac{h_i}{2},\, y_i + \frac{h_i k_2}{2}\right); \\ k_4 = f\left(t_i + h_i,\, y_i + h_i k_3\right). \end{gathered} \tag{6.25}$$

Проте, слід зазначити, що використання формул (6.25) не дає змогу побудувати досить ефективні алгоритми розв'язку задачі Коші. Формули (6.25) слід використовувати тільки для напів-аналітичних розрахунків або на початкових етапах розробки програмного забезпечення для отримання первинних оцінок.

Приклад:

$$\begin{cases} \dfrac{dy}{dt} = -\left(t+1\right)y^2 \\ y\left(0\right) = 1 \end{cases};$$

$$t \in \left[0,1\right], \qquad h = 0{,}5.$$

t_0=0;

$y_0 = y\left(0\right) = 1$;

$$t_1 = 0,5;$$

$$k_1 = f\left(t_0, y_0\right) = -\left(t_0 + 1\right) y_0^2 = -\left(0 + 1\right) \cdot 1^2 = -1;$$

$$k_2 = f\left(t_0 + \frac{h}{2}, y_0 + \frac{hk_1}{2}\right) = -\left(t_0 + \frac{h}{2} + 1\right)\left(y_0 + \frac{hk_1}{2}\right)^2 =$$

$$= -\left(0 + \frac{0,5}{2} + 1\right)\left(1 - \frac{0,5 \cdot 1}{2}\right)^2 = -0,703125;$$

$$k_3 = f\left(t_0 + \frac{h}{2}, y_0 + \frac{hk_2}{2}\right) = -\left(t_0 + \frac{h}{2} + 1\right)\left(y_0 + \frac{hk_2}{2}\right)^2 =$$

$$= -\left(0 + \frac{0,5}{2} + 1\right)\left(1 - \frac{0,5 \cdot 0,703125}{2}\right)^2 \cong -0,849171;$$

$$k_4 = f\left(t_0 + h, y_0 + hk_3\right) = -\left(t_0 + h + 1\right)\left(y_0 + hk_3\right)^2 =$$

$$= -\left(0 + 0,5 + 1\right)\left(1 - 0,5 \cdot 0,849171\right)^2 \cong -0,496653;$$

$$y_1 = y_0 + \frac{h}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) =$$

$$= 1 + \frac{0,5}{6}\left(-1 - 2 \cdot 0,703125 - 2 \cdot 0,849171 - 0,496653\right) \cong 0,616563;$$

$$t_2 = 1;$$

$$k_1 = f\left(t_1, y_1\right) = -\left(t_1 + 1\right) y_1^2 = -\left(0,5 + 1\right) \cdot 0,616563^2 = -0,570225;$$

$$k_2 = f\left(t_1 + \frac{h}{2}, y_1 + \frac{hk_1}{2}\right) = -\left(t_1 + \frac{h}{2} + 1\right)\left(y_1 + \frac{hk_1}{2}\right)^2 =$$

$$= -\left(0,5 + \frac{0,5}{2} + 1\right)\left(0,616563 - \frac{0,5 \cdot 0,570225}{2}\right)^2 = -0,393194;$$

$$k_3 = f\left(t_1 + \frac{h}{2}, y_1 + \frac{hk_2}{2}\right) = -\left(t_1 + \frac{h}{2} + 1\right)\left(y_1 + \frac{hk_2}{2}\right)^2 =$$

$$= -\left(0,5 + \frac{0,5}{2} + 1\right)\left(0,616563 - \frac{0,5 \cdot 0,393194}{2}\right)^2 \cong -0,470047;$$

$$k_4 = f\left(t_1 + h, y_1 + hk_3\right) = -\left(t_1 + h + 1\right)\left(y_1 + hk_3\right)^2 =$$

$$= -\left(0,5 + 0,5 + 1\right)\left(0,616563 - 0,5 \cdot 0,470047\right)^2 \cong -0,291145;$$

$$y_2 = y_1 + \frac{h}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) =$$

$$= 0,616563 + \frac{0,5}{6}\left(-0,570225 - 2 \cdot 0,393194 - 2 \cdot 0,470047 - 0.291145\right) \cong$$

$$\cong 0,400909;$$

Методи Ейлера та Рунге – Кутта називаються одноточковими. Це пояснюється тим, що для обчислення результату в $(i+1)$-й точці достатньо знати значення лише в одній попередній *i*-й точці.

6.3. Багатоточкові методи

В одноточкових методах значення y_{i+1} обчислюють на основі значення в одній попередній точці y_i. Логічно припустити, що можна підвищити точність, якщо в розрахунку значення в точці y_{i+1} на кожному кроці бере участь не одна, а кілька попередніх точок – $y_i, y_{i-1}, \ldots$

В загальному випадку лінійним *m*-кроковим різницевим методом з постійним кроком називається система різницевих рівнянь [3]:

$$\frac{a_0 y_{i+1} + a_1 y_i + \ldots + a_m y_{i-m+1}}{h} = b_0 f_{i+1} + b_1 f_i + \ldots + b_m f_{i-m+1}; i = \overline{m-1, n-1}, (6.26)$$

де a_k, b_k — чисельні коефіцієнти, що не залежать від i, $k=\overline{0,m}$, причому $a_0\neq 0$, $f_k=f(t_k,y_k)$, $h=t_{i+1}-t_i$, $i=\overline{0,n-1}$.

Метод називається лінійним, тому що кожне f_k входить до формули лінійно.

Рівняння (6.26) є рекурентним співвідношенням, за яким нове значення y_{i+1} виражається через раніше знайдені y_i, y_{i-1}, y_{i-2},..., y_{i-m+1}. Обчислення починається з $i=m-1$, тобто з рівняння

$$\frac{a_0 y_m + a_1 y_{m-1} + \ldots + a_m y_0}{h} = b_0 f_m + b_1 f_{m-1} + \ldots + b_m f_0 .$$

Звідси видно, що для початку обчислень слід задати m початкових значень y_0, y_1, ... y_{m-1}. Значення y_0 задається умовою (6.2). Величини y_1, y_2, ... y_{m-1} можна обчислити використовуючи однокрокові методи, наприклад, методи Рунге-Кутта.

Метод (6.26) називається явним, якщо $b_0=0$, і, відповідно, шукане значення y_{i+1} виражається через попередні y_i, y_{i-1}, ..., y_{i-m+1}:

$$y_{i+1} = \frac{1}{a_0}\sum_{k=1}^{m}\left[-a_k y_{i-k+1} + h b_k f_{i-k+1}\right] \qquad (6.27)$$

Якщо $b_0\neq 0$, то метод називається неявним. Тоді для знаходження y_{i+1} доводиться розв'язувати нелінійне рівняння

$$y_{i+1}a_0 - b_0 h f(y_{i+1}, t_{i+1}) = \sum_{k=1}^{m}\left[-a_k y_{i-k+1} + h b_k f_{i-k+1}\right]. \qquad (6.28)$$

Зазвичай це рівняння розв'язують методом Ньютона і задають початкове наближення $y_{i+1}^{(0)}$, що дорівнює y_i.

Слід відзначити, що коефіцієнти рівняння (6.26) задані з точністю до множника. Щоб усунути цю неоднозначність, вимагають виконання умови нормування

$$\sum_{k=0}^{m} b_k = 1. \tag{6.29}$$

Можна показати, що для того, аби різницевий метод (6.26) мав порядок точності p, необхідно щоб були виконані умови [3]:

$$\sum_{k=0}^{m} a_k = 0; \tag{6.30}$$

$$\sum_{k=0}^{m} k^{l-1}(ka_k + lb_k) = 0, l = \overline{1, p}. \tag{6.31}$$

Разом з умовою нормування (6.29), рівняння (6.30) і (6.31) становлять систему з p+2 лінійних алгебраїчних рівнянь відносно 2m+2 невідомих $a_0, a_1, ..., a_m, b_0, b_1, ..., b_m$.

Для того, щоб існував розв'язок цієї системи, необхідно, щоб кількість рівнянь була не більшою від кількості невідомих, тому $p \le 2m$. Ця вимога означає, що лінійні m-крокові різницеві методи мають порядок точності не більше, ніж 2m. Тому порядок точності неявних m-крокових методів дорівнює 2m, коли $a_i \neq 0$, $b_i \neq 0, i = \overline{0, m}$, а явних — 2$m$-1, оскільки b_0=0 і невідомих в (6.29), (6.30), (6.31) на одну менше.

У практиці обчислень найбільше поширення отримали методи Адамса, які становлять частковий випадок багатокрокових методів (6.26), коли похідна $\frac{dy}{dt}$ апроксимується тільки по двох точках t_{i+1}, t_i, тобто

$$a_0 = -a_1 = 1, \quad a_k = 0, k = \overline{2, m}.$$

Таким чином, методи Адамса мають вигляд

$$\frac{y_{i+1} - y_i}{h} = \sum_{k=0}^{m} b_k f_{i-k+1}. \tag{6.32}$$

У випадку b_0=0 методи Адамса називають явними, а якщо $b_0 \neq 0$ — неявними.

Для методів Адамса умови (6.29), (6.31) приймають вигляд

$$l\sum_{k=1}^{m} k^{l-1} b_k = 1; \qquad l = \overline{1, p}; \tag{6.33}$$

$$b_0 = 1 - \sum_{k=1}^{m} b_k . \tag{6.34}$$

Звідси видно, що найвищий порядок точності m-крокового неявного методу Адамса дорівнює m+1 (кількість рівнянь p, а невідомих — m+1), а найвищий порядок точності явного методу дорівнює m (b_0=0).

Знайдемо коефіцієнти b_k для кількох m. Розглянемо явні методи Адамса. Коефіцієнти знаходять із (6.33), (6.34) прийнявши, що p=m. У випадку m=1 отримуємо метод Ейлера

$$\frac{y_{i+1} - y_i}{h} = f_i .$$

Для m=2 (6.33) приймає вигляд:

$$b_1 + b_2 = 1;$$
$$2(b_1 + 2b_2) = 1 .$$

Звідки $b_1 = \frac{3}{2}; b_2 = -\frac{1}{2}$.

Таким чином, для m=2, отримаємо явний метод Адамса другого порядку

$$\frac{y_{i+1} - y_i}{h} = \frac{3}{2} f_i - \frac{1}{2} f_{i-1} .$$

Звідки

$$y_{i+1} = y_i + h\left(\frac{3}{2} f\left(t_i, y_i\right) - \frac{1}{2} f\left(t_{i-1}, y_{i-1}\right)\right).$$

Розв'язки для інших m наведено у таблиці.

Таблиця 6.1. Коефіцієнти явного методу Адамса порядку m

m	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}
1	1									
2	3/2	-1/2								
3	23/12	-4/3	5/12							
4	55/24	-59/24	37/24	-3/8						
5	1901/720	-1387/360	109/30	-637/360	251/720					
6	4277/1440	-2641/480	4991/720	-3649/720	959/480	-95/288				
7	28392/ 8641	-18637/ 2520	235183/ 20160	-10754/ 945	135713/ 20160	-5603/ 2520	6054/ 19183			
8	16083/ 4480	-2281287/ 239500	242653/ 13440	-24605/ 1117	28060547/ 1614563	-2384104/ 276831	32863/ 13440	-5257/ 17280		
9	2787217/ 717463	-294002/ 24739	1929453/ 73333	-7436324/ 192949	135961/ 3576	-11884/ 473	261790/ 24463	50246/ 18867	-994/ 3371	
10	173234/ 41525	-17722/ 1225	1971119/ 53794	-4119436/ 65757	10176373/ 137186	-6963789/ 113632	843419/ 24231	-6835/ 526	122585/ 42599	-130/ 453

Коефіцієнти неявного методу Адамса знаходяться із системи (6.33), (6.34) з $p=m+1$. Наведемо формули неявних m-точкових методів Адамса для деяких значень m:

$m = 1$:

$$y_{i+1} = y_i + \frac{h}{2}\left(f\left(t_{i+1}, y_{i+1}\right) + f\left(t_i, y_i\right)\right);$$

$m = 2$:

$$y_{i+1} = y_i + \frac{h}{12}\left(5 f\left(t_{i+1}, y_{i+1}\right) + 8 f\left(t_i, y_i\right) - f\left(t_{i-1}, y_{i-1}\right)\right);$$

$m = 3$:

$$y_{i+1} = y_i + \frac{h}{24}\left(9f\left(t_{i+1}, y_{i+1}\right) + 19f\left(t_i, y_i\right) - 5f\left(t_{i-1}, y_{i-1}\right) + f\left(t_{i-2}, y_{i-2}\right)\right).$$

Перевагою багатоточкових методів є висока точність розрахунку, а основними недоліками – складність оцінки похибки на кожному кроці, а щоб почати розрахунки цими методами, слід мати значення шуканої функції в кількох попередніх точках. Ці значення можна попередньо обчислити одним із одноточкових методів.

Часто, на кожному кроці розв'язання спільно використовуються два багатокрокових методи. Спочатку обчислюють прогнозоване значення y_{i+1}^n за допомогою явного методу, що називається предиктором. Знайдене значення використовується як початкове наближення для його уточнення неявним методом, що називається коректором. Такі методи називають методами предиктор-коректор. Для $m = 3$ обчислювальна схема матиме вигляд:

$$y_{i+1}^n = y_i + \frac{h}{12}\left(23f\left(t_i, y_i\right) - 16f\left(t_{i-1}, y_{i-1}\right) + 5f\left(t_{i-2}, y_{i-2}\right)\right);$$

$$y_{i+1} = y_i + \frac{h}{24}\left(9f\left(t_{i+1}, y_{i+1}^n\right) + 19f\left(t_i, y_i\right) - 5f\left(t_{i-1}, y_{i-1}\right) + f\left(t_{i-2}, y_{i-2}\right)\right).$$

Схему предиктор-коректор застосовують тоді, коли застосувати неявну схему неможливо (наприклад, якщо нелінійне рівняння, яке виникає під час застосування неявної схеми надто складне). Слід однак зазначити, що такий підхід є, по суті, виконанням тільки однієї ітерації розв'язку нелінійного рівняння методом простої ітерації.

Приклад. Нехай дано задачу Коші

$$\begin{cases} \dfrac{dy}{dt} = -\left(t+1\right)y^2 \\ y\left(0\right) = 1 \end{cases}$$

Необхідно обчислити таблицю значень $y(t)$ на відрізку $t \in [0,1]$, якщо h=0,25.

Застосуємо трьохточковий явний метод Адамса. Використовуючи таблицю 6.1, отримаємо загальну формулу для цього методу:

$$y_{i+1} = y_i + \frac{h}{12}\left(23f\left(t_i, y_i\right) - 16f\left(t_{i-1}, y_{i-1}\right) + 5f\left(t_{i-2}, y_{i-2}\right)\right).$$

Спочатку обчислимо значення y_1, y_2 явним методом Ейлера:

$$y_0 = y(0) = 1;$$

$$y_1 = y(0,25) = y_0 - h\left(t_0 + 1\right)y_0^2 = 1 - 0,25 \cdot \left(0 + 1\right) \cdot 1^2 = 0,75;$$

$$y_2 = y(0,5) = y_1 - h\left(t_1 + 1\right)y_1^2 = 0,75 - 0,25 \cdot \left(0,25 + 1\right) \cdot 0,75^2 \cong 0,574219.$$

Після цього продовжимо розв'язок заданим методом:

$$t_3 = 0,75;$$

$$y_3 = y_2 + \frac{h}{12}\left(23\left(-\left(t_2 + 1\right)y_2^2\right) - 16\left(-\left(t_1 + 1\right)y_1^2\right) + 5\left(-\left(t_0 + 1\right)y_0^2\right)\right) =$$

$$= 0,574219 + \frac{0,25}{12}\left(\begin{array}{l} -23\left(0,5 + 1\right) \cdot 0,574219^2 + \\ +16\left(0,25 + 1\right) \cdot 0,75^2 - 5\left(0 + 1\right) \cdot 1^2 \end{array}\right) \cong$$

$$\cong 0,467436;$$

$$t_4 = 1;$$

$$y_4 = y_3 + \frac{h}{12}\left(23\left(-\left(t_3 + 1\right)y_3^2\right) - 16\left(-\left(t_2 + 1\right)y_2^2\right) + 5\left(-\left(t_1 + 1\right)y_1^2\right)\right) =$$

$$= 0,467436 + \frac{0,25}{12}\left(\begin{array}{l} -23\left(0,75 + 1\right) \cdot 0,467436^2 + 16\left(0,5 + 1\right) \cdot 0,574219^2 - \\ -5\left(0,25 + 1\right) \cdot 0,75^2 \end{array}\right) \cong$$

$$\cong 0.375839.$$

6.4. Апостеріорна оцінка похибки розв'язання задачі Коші. Автоматичний вибір кроку інтегрування.

У разі чисельного розв'язання задачі Коші використовують різні методи апостеріорної оцінки похибки в залежності від обраного методу та інших факторів. Найбільш широко застосовується правило Рунге:

$$R_i \approx \frac{y_{i+1,\frac{h}{q}} - y_{i+1,h}}{q^p - 1}, \qquad (6.35)$$

де $y_{i+1,\frac{h}{q}}$ та $y_{i+1,h}$ розв'язки задачі Коші в точці $t=t_{i+1}$, отримані відповідно з кроками $\frac{h}{q}$ та h, q – ціле число, як правило рівне 2, p – порядок точності методу, що використовується, R_i – апостеріорна оцінка похибки для результату, отриманого з кроком $\frac{h}{q}$.

Приклад. Нехай дано задачу Коші

$$\begin{cases} \frac{dy}{dt} = -(t+1)y^2 \\ y(0) = 1 \end{cases}.$$

Необхідно обчислити значення $y(t)$ в точці $t=1$ та оцінити похибку результату. Скористаємося явним методом Ейлера. Виберемо крок $h=0{,}5$. Тоді

$$y_0 = y(0) = 1;$$

$$y_1 = y(0{,}5) = y_0 - h(t_0+1)y_0^2 = 1 - 0{,}5\cdot(0+1)\cdot 1^2 = 0{,}5;$$

$$y_2 = y(1)_h = y_1 - h(t_1+1)y_1^2 = 0{,}5 - 0{,}5\cdot(0{,}5+1)\cdot 0{,}5^2 \cong 0{,}3125.$$

Для того, щоб скористатись правилом Рунге (6.35), необхідно обчислення виконати ще раз, але з кроком вдвічі меншим (q=2), тобто h=0,25. Такі обчислення вже були проведені в параграфі 6.1. Результатом є $y(1)_{\frac{h}{2}} = 0,361752$. Тоді, враховуючи, що метод Ейлера має перший порядок точності (p=1), з формули (6.35) маємо

$$R_i \approx \frac{y(1)_{\frac{h}{2}} - y(1)_h}{2^1 - 1} = \frac{0,361752 - 0,3125}{2 - 1} = 0,049252.$$

Точним розв'язком цієї задачі Коші є функція $y(t) = \frac{2}{(t+1)^2 + 1}$. Тому точним розв'язком в точці t=1 є $y(1)$=0.4. Як видно, отримана апостеріорна оцінка похибки обчислення $y(1)_{\frac{h}{2}}$ є досить близькою до дійсної. Слід відзначити, що апостеріорна оцінка похибки тим ближча до дійсної, чим менший крок h.

Під час розв'язання задачі Коші неявними методами використання правила Рунге є неефективним, оскільки у разі зменшення кроку доводиться додатково розв'язувати q нелінійних рівнянь. Тому за використання неявних методів застосовують інші методи оцінки похибки. Так, наприклад, з порівняння формул для похибок явного та неявного методів Ейлера видно, що за умови $h \to 0$ модулі цих похибок є однаковими, але вони мають протилежний знак. Ця особливість дає змогу апостеріорно оцінити похибку неявного методу за формулою:

$$R_i \approx \frac{\tilde{y}_{i+1} - y_{i+1}}{2} \qquad (6.36)$$

де $\tilde{y}_{i+1}$ та y_{i+1} – розв'язки, отримані відповідно явним та неявним методом.

Приклад. Для вищенаведеної задачі Коші треба оцінити похибку неявного методу Ейлера в точці t=1 у разі розв'язання задачі з кроком h=0,25. Використовуючи приклад з параграфу 6.1, маємо $\tilde{y}_4$=0,361752, а y_4=0,431175. Тоді

$$R_i \approx \frac{0{,}361752\text{-}0{,}431175}{2} \cong -0{,}034712.$$

Як видно, ця оцінка близька до дійсної похибки.

Під час застосування методів Рунге-Кутта використовують інший підхід. В ньому оцінка похибки розраховується як різниця результатів отриманих методами з різними порядками точності:

$$R_i \approx y_{i+1} - y_{i+1}^*, \tag{6.37}$$

де y_{i+1}– розв'язок, отриманий методом з порядком точності p, а y_{i+1}^*- з порядком точності p+1. Як було відмічено в параграфі 6.2, існує цілий ряд методів Рунге-Кутта, коли для одних й тих самих функцій k_j можна на наборі коефіцієнтів c_j в (6.13) отримати метод 4 порядку точності, а на наборі c_j^*- 5 порядок точності. Тому використання (6.37) практично не приводить до додаткових витрат, а похибка оцінюється за формулою:

$$R_i \approx h_i \sum_{j=1}^{m} \left(c_j - c_j^*\right) k_j .$$

Існують і інші підходи для оцінки похибки розв'язання задачі Коші [18]. Так, наприклад, можна оцінити похідну, яка входить в формулу похибки. Для явного методу Ейлера достатньо оцінити $\frac{d^2 y(t_i)}{dt^2}$. Для цього можна скористатися апроксимацією похідної функції $f(t,y)$ скінченними різницями:

$$y''(t_i)=\frac{d}{dt}\left(\frac{dy(t_i)}{dt}\right)=\frac{d}{dt}\left(f(t_i,y_i)\right)=\frac{\partial f}{\partial y}\frac{dy}{dt}\bigg|_{t=t_i}+\frac{\partial f}{\partial t}\bigg|_{t=t_i}=f(t_i,y_i)\frac{\partial f}{\partial y}\bigg|_{t=t_i}+\frac{\partial f}{\partial t}\bigg|_{t=t_i},$$

де $\left.\frac{\partial f}{\partial y}\right|_{t=t_i}\approx\frac{f(t_i,y_i+\Delta y_i)-f(t_i,y_i)}{\Delta y_i}$, $\left.\frac{\partial f}{\partial t}\right|_{t=t_i}\approx\frac{f(t_i+\Delta t_i,y_i)-f(t_i,y_i)}{\Delta t_i}$,

$\Delta y_i=\sqrt{\varepsilon_{маш}}\,y_i,\qquad \Delta t_i=\sqrt{\varepsilon_{маш}}\,t_i.$

Інший спосіб оцінки похідної ґрунтується на диференціюванні інтерполяційних поліномів. Так, наприклад, для оцінки $\frac{d^2y(t_i)}{dt^2}$ можна використовувати отримані значення y_{i-1}, y_i, y_{i+1}. Двічі диференціюючи інтерполяційний поліном Лагранжа, що проходить через точки (t_{i-1},y_{i-1}), (t_i,y_i), (t_{i+1},y_{i+1}) отримаємо

$$y''(t_i)=2\left(\frac{y_{i-1}}{h_{i-1}(h_i+h_{i-1})}-\frac{y_i}{h_{i-1}h_i}+\frac{y_{i+1}}{h_i(h_i+h_{i-1})}\right).$$

Наявність методів оцінки похибки на кожному кроці дає можливість підбирати крок інтегрування залежно від необхідної точності розв'язку. Для цього після знаходження оцінки похибки R_i перевіряють виконання умови:

$$|R_i|\le\varepsilon|y_{i+1}|, \tag{6.38}$$

де ε – задана відносна похибка розв'язку.

Якщо умова (6.38) не виконується, то крок зменшують в q разів або в $\sqrt[p+1]{\frac{|R_i|}{\varepsilon|y_{i+1}|}}$ раз. Зменшення кроку виконують до тих пір, поки не виконається умова (6.38). Після цього переходять до наступного елементарного відрізку.

Крок на наступному відрізку вибирають за формулою [18]:

$$h_{i+1} = \left(\frac{\varepsilon |y_{i+1}|}{|R_i|} \right)^{\frac{1}{p+1}} h_i,$$

або за правилом зон [22]:

$$h_{i+1} = \begin{cases} h_i \ , \ \dfrac{\varepsilon}{q^{p+1}} \le \left| \dfrac{R_i}{y_{i+1}} \right| \le \varepsilon \ ; \\ q h_i \ , \ \left| \dfrac{R_i}{y_{i+1}} \right| < \dfrac{\varepsilon}{q^{p+1}}. \end{cases}$$

Якщо необхідно обчислити значення функції в наперед заданих вузлах, можна використовувати метод екстраполяції нульового кроку [18]. Для цього на кожному елементарному відрізку знаходять значення y_{i+1} з різними кроками, які, як правило, відрізняються один від одного в два рази. На j-му етапі екстраполяції будують послідовність розв'язків за формулами:

$$y^{(j)}_{i+1,\,h/2^k} = y^{(j-1)}_{i+1,\,h/2^k} + \frac{y^{(j-1)}_{i+1,\,h/2^k} - y^{(j-1)}_{i+1,\,h/2^{k-1}}}{2^{p+j} - 1}, \quad k = \overline{j,n}.$$

Крок послідовно зменшують доти, доки нова екстраполяція не буде відрізнятися від попередньої не більше, ніж на задану величину похибки розв'язку. Остання оцінка є екстраполяцією нульового кроку. В цьому методі автоматичний вибір кроку не потрібний.

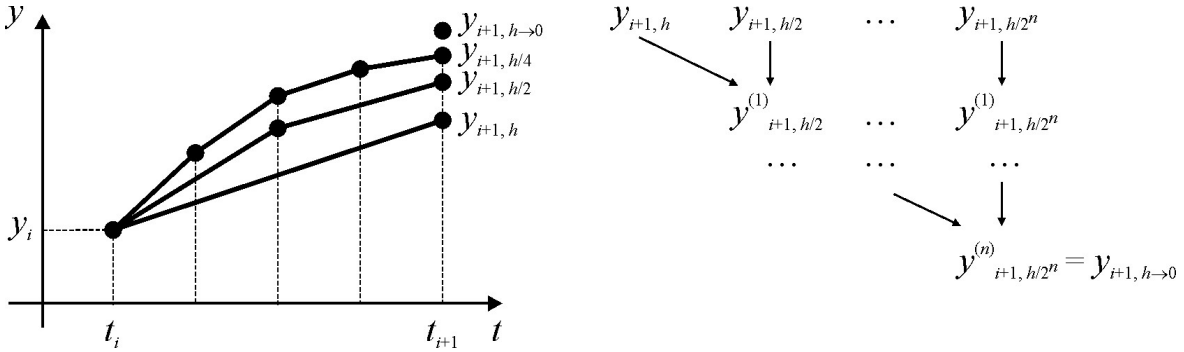


Рис. 6.1. Екстраполяція нульового кроку.

6.5. Жорсткі рівняння

Поняття жорсткого рівняння пов'язано з жорсткою умовою стійкості для різницевих схем розв'язання цих рівнянь [3,16,27]. Розглянемо задачу Коші:

$$\frac{dy}{dt} = \lambda y,\ \ y(0) = y_0, \tag{6.39}$$

де λ – досить велике від'ємне число. Точний розв'язок задачі (6.39) має вигляд

$$y(t) = y_0 e^{\lambda t}. \tag{6.40}$$

Із виразу (6.40) випливає, що для будь-якого t_i і $h_i > 0$:

$$\left|y(t_{i+1})\right| = \left|y(t_i + h_i)\right| < \left|y(t_i)\right|. \tag{6.41}$$

Застосуємо для розв'язання задачі (6.39) явний метод Ейлера:

$$y_{i+1} = (1 + \lambda h_i) y_i = (1 + \lambda h_i)^i y_0. \tag{6.42}$$

Із виразу (6.42) випливає, що умова (6.41) виконується, якщо

$$\left|1 + \lambda h_i\right| < 1.$$

Отже, процес (6.42) буде стійким для виконання співвідношення (6.41) тільки за умови, що

$$h_i \le \frac{2}{|\lambda|}. \tag{6.43}$$

За великих значень λ крок інтегрування жорстко обмежують і вибирають, виходячи не стільки з умови забезпечення заданої похибки розв'язку, а так, щоб задовольнити вираз (6.43).

Жорсткі рівняння виникають зазвичай в задачах, характерною особливістю яких є повільна зміна їх розв'язків за наявності швидко

згасаючих збурень. Наприклад, задачі на визначення перехідних характеристик електронних схем часто є жорсткими.

Для кількісної оцінки жорсткості рівняння вводять коефіцієнт жорсткості *S*. Коефіцієнт жорсткості для задачі (6.39) залежить не тільки від $|\lambda|$, але й від довжини відрізку інтегрування:

$$S=|\lambda|\,t_n,$$

де t_n – права границя відрізку інтегрування.

Для системи диференціальних рівнянь жорсткість можна оцінити як

$$S=\frac{\max\limits_{i,\ t\in[0,t_n]}\left(-\mathrm{Re}\lambda_i(t)\right)}{\min\limits_{i,\ t\in[0,t_n]}\left|\mathrm{Re}\lambda_i(t)\right|},$$

де λ_i — власні числа матриці Якобі системи диференціальних рівнянь.

Обмеження (6.43) під час використання явних методів призводить до великих обчислювальних затрат через малий крок інтегрування, тому загальним підходом до розв’язання жорстких задач є використання неявних схем. Незважаючи на потребу розв’язання систем нелінійних рівнянь, за рахунок можливості значного збільшення кроку інтегрування загальний обсяг обчислень може бути значно меншим, ніж для явних методів.

На сьогодні часто використовують метод Гіра у вигляді неявної схеми [3]:

$$\sum_{k=0}^{m}a_k y_{i-k+1}=hf\left(y_{i+1},t_{i+1}\right), \qquad (6.44)$$

яка є окремим випадком багатоточкового методу (6.26) якщо

$$b_i=0, i=\overline{1,m}, b_0=1. \qquad (6.45)$$

Для метода (6.44), який має порядок точності $p=m$, сталі $a_i, i=\overline{0,m}$ можуть бути знайдені з системи рівнянь (6.30), (6.31) з урахуванням (6.45). Розв'язки цієї системи наведені в таблиці:

Таблиця 6.2. Коефіцієнти методу Гіра порядку m

m	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
1	1	-1									
2	3/2	-2	1/2								
3	11/6	-3	3/2	-1/3							
4	25/12	-4	3	-4/3	1/4						
5	137/60	-5	5	-10/3	5/4	-1/5					
6	49/20	-6	15/2	-20/3	15/4	-6/5	1/6				
7	363/140	-7	21/2	-35/3	35/4	-21/5	7/6	-1/7			
8	761/280	-8	14	-56/3	35/2	-56/5	14/3	-8/7	1/8		
9	7129/2520	-9	18	-28	63/2	-126/5	14	-36/7	9/8	-1/9	
10	7381/2520	-10	45/2	-40	105/2	-252/5	35	-120/7	45/8	-10/9	1/10

Приклад. Нехай дано задачу Коші

$$\begin{cases} \dfrac{dy}{dt} = -(t+1)y^2; \\ y(0) = 1. \end{cases}$$

Необхідно обчислити таблицю значень $y(t)$ на відрізку $t \in [0,1]$ методом Гіра 3-го порядку, якщо h=0,25.

Використовуючи таблицю 6.2 отримаємо з (6.44) загальну формулу для цього методу:

$$\frac{11}{6}y_{i+1} - 3y_i + \frac{3}{2}y_{i-1} - \frac{1}{3}y_{i-2} - hf\left(t_{i+1}, y_{i+1}\right) = 0.$$

Спочатку обчислимо значення y_1, y_2 неявним методом Ейлера:

$$y_0 = y(0) = 1;$$

$$(t_1+1)hy_1^2+y_1-y_0=0 \Rightarrow (0,25+1)\cdot 0,25\cdot y_1^2+y_1-1=0 \Rightarrow y_1=0.8;$$

$$(t_2+1)hy_2^2+y_2-y_1=0 \Rightarrow (0,5+1)\cdot 0,25\cdot y_2^2+y_2-0,8=0 \Rightarrow y_2\cong 0,644320.$$

Після цього продовжимо розв'язок заданим методом:

$$t_3=0,75;$$

$$\frac{11}{6}y_3-3y_2+\frac{3}{2}y_1-\frac{1}{3}y_0-h\left(-(t_3+1)y_3^2\right)=0;\Rightarrow$$

$$\Rightarrow 0,25(0,75+1)y_3^2+\frac{11}{6}y_3-3\cdot 0,644320+\frac{3}{2}\cdot 0,8-\frac{1}{3}\cdot 1=0;\Rightarrow$$

$$\Rightarrow y_3\cong 0,517665;$$

$$t_4=1;$$

$$\frac{11}{6}y_4-3y_3+\frac{3}{2}y_2-\frac{1}{3}y_1-h\left(-(t_4+1)y_4^2\right)=0;\Rightarrow$$

$$\Rightarrow 0,25(1+1)y_4^2+\frac{11}{6}y_4-3\cdot 0,517665+\frac{3}{2}\cdot 0,644320-\frac{1}{3}\cdot 0,8=0;\Rightarrow$$

$$\Rightarrow y_4\cong 0,417772;$$

6.6. Крайові задачі

Часто фізичні задачі зводять до диференціального рівняння m-го порядку

$$\frac{d^m y}{dx^m}=f\left(x,\ y,\ \frac{dy}{dx},\ldots,\ \frac{d^{m-1}y}{dx^m}\right), \qquad (6.46)$$

де $y(x)$ – деяка фізична величина, визначена на відрізку $[a,\ b]$, що описує певний процес. Часто відома поведінка процесу на кінцях відрізка $[a,\ b]$. Тому замість початкових умов у загальному випадку задають умови типу

$$\sum_{i=0}^{m-1}\left[\alpha_{ij}\frac{d^i y(a)}{dx^i}+\beta_{ij}\frac{d^i y(b)}{dx^i}\right]=\gamma_j,\ \ j=\overline{1,m}, \qquad (6.47)$$

де α_{ij}, β_{ij}, γ_j – сталі. Наприклад, для диференціального рівняння 2-го порядку задають умови на кінцях $y(a)=y_a$, $y(b)=y_b$. Таку задачу називають крайовою задачею першого класу або задачею із закріпленими кінцями.

Аналогічно поставлено крайову задачу для системи диференціальних рівнянь. Методи розв'язання крайових задач відрізняються від методів розв'язання задачі Коші і поділені на два загальних класи – методи зведення крайових задач до задач Коші [1, 27] та спеціальні методи, призначені для розв'язання власне крайових задач [27].

6.6.1. Зведення крайових задач до задач Коші

Подамо крайову задачу із виразів (6.46) і (6.47) для диференціального рівняння *m*-го порядку у вигляді:

$$Ly = f(x); \tag{6.48}$$

$$\sum_{i=0}^{m-1}\left[\alpha_{ij}\frac{d^i y(0)}{dx^i}+\beta_{ij}\frac{d^i y(l)}{dx^i}\right]=\gamma_j, \qquad j=\overline{0,m-1}, \tag{6.49}$$

де L – диференціальний оператор, а α_{ij}, β_{ij}, γ_j – сталі.

Задачу (6.48), з умовою (6.49) найпростіше звести до задачі Коші, якщо L – лінійний диференціальний оператор

$$L=a_m(x)\frac{d^m}{dx^m}+a_{m-1}(x)\frac{d^{m-1}}{dx^{m-1}}+\ldots+a_1(x)\frac{d}{dx}+a_0(x).$$

У цьому випадку розв'язок шукають у вигляді

$$y(x)=z_0(x)+\sum_{k=1}^{m}c_k z_k(x), \tag{6.50}$$

де $z_k(x)$ знаходять з $(m+1)$ задачі Коші

$$Lz_0 = f(x),\ \ z_0^{(i)}(0) = 0,\ \ i = \overline{0, m-1};$$
$$Lz_k = 0,\ z_k^{(i)}(0) = \begin{cases} 0\,, & i \neq k-1 \\ 1\,, & i = k-1 \end{cases},\ i = \overline{0, m-1},\quad k = \overline{1, m}. \tag{6.51}$$

Після розв'язання задач Коші (6.51) на відрізку [0, *l*], вираз (6.50) підставляють у вираз (6.49) і одержують СЛАР з *m* рівнянь відносно *m* невідомих $c_k,\ k = \overline{1, m}$. Розв'язавши СЛАР з (6.50), знаходять розв'язок задач (6.48) з умовою (6.49).

Приклад:

$$\frac{d^2y}{dx^2} + 4y = x;$$
$$y(0) = 1;$$
$$y(1) = 2.$$

Шукаємо розв'язок у вигляді

$$y(x) = z_0(x) + c_1 z_1(x) + c_2 z_2(x),$$

де $z_0(x)$ – розв'язок неоднорідної задачі Коші з нульовими початковими умовами, а $z_1(x)$, $z_2(x)$– розв'язки однорідних задач Коші з ненульовими початковими умовами відповідно до (6.51).

Із розв'язків задач Коші на відрізку [0, 1]

$$\frac{d^2z_0}{dx^2} + 4z_0 = x, \qquad z_0(0) = 0, \qquad z_0'(0) = 0;$$
$$\frac{d^2z_1}{dx^2} + 4z_1 = 0, \qquad z_1(0) = 1, \qquad z_1'(0) = 0;$$
$$\frac{d^2z_2}{dx^2} + 4z_2 = 0, \qquad z_2(0) = 0, \qquad z_2'(0) = 1.$$

отримуємо $z_0(x),\ z_1(x),\ z_2(x)$. Далі знаходимо сталі c_1 та c_2.

З першої крайової умови маємо

$$0 + c_1 \cdot 1 + c_2 \cdot 0 = 1 \quad \Rightarrow \qquad c_1 = 1 .$$

З другої умови:

$$z_0(1) + c_1 z_1(1) + c_2 z_2(1) = 2 \qquad \Rightarrow \qquad c_2 = \frac{2 - z_0(1) - z_1(1)}{z_2(1)} .$$

Якщо L – нелінійний диференціальний оператор, то для зведення крайової задачі до задачі Коші застосовують метод «стрільби». Для цього розв'язок $y(x)$ шукають у вигляді

$$y = z(x),$$

де $z(x)$ – розв'язок задачі Коші

$$Lz = f(x); \tag{6.52}$$

$$z^{(i)}(0) = z_i, \; i = \overline{0, m-1}. \tag{6.53}$$

Значення z_i у початковій умові (6.53) підбирають такими, щоб задовольнити (6.49). У цьому випадку крайові умови (6.49) разом з (6.52) і (6.53) утворюють систему m нелінійних рівнянь щодо m невідомих z_i, $i = \overline{0, m-1}$. Останній розв'язок (6.52) і є розв'язком задачі (6.48) з умови (6.49).

Приклад. Припустимо, що потрібно розв'язати задачу

$$\frac{d^2 y}{dx^2} = f\left(x, y, y'\right)$$

за умов:

$$y(0) = y_0;$$
$$y(l) = y_l.$$

Можна застосувати такий ітераційний метод [24]:

1) вибирати ξ, що апроксимує $y'(0)$, тобто покласти

$$z'(0) = \xi;$$

2) розв'язати задачу Коші:

$$z'' = f(x,\ z,\ z');$$
$$z(0) = y_0;$$
$$z'(0) = \xi.$$

3) якщо $\left|z(l) - y_l\right| < \varepsilon$, де ε – задана похибка, то покласти $y(x) \approx z(x)$, в іншому випадку – змінити ξ (виконати ітерацію розв'язку нелінійного рівняння) і повернутися до пункту 2).

6.6.2. Метод скінченних різниць

Поширеним методом розв'язання крайових задач є метод скінченних різниць (МСР). В основі МСР лежить апроксимація похідних в операторі L (6.48) і в крайових умовах (6.49) скінченними різницями, а неперервної функції $f(x)$ – сітковою функцією, визначеною у вузлах відрізка $[0, l]$ [27]. Розглянемо найпростіший випадок, коли вузли сіткової функції – рівновіддалені. Для цього відрізок $[0, l]$ ділять на $n \geq m+1$ рівних відрізків вузлами $x_0 = 0,\ x_1 = x_0 + h,\ x_2 = x_0 + 2h, \ldots,\ x_n = x_0 + nh$. Розв'язання задачі (6.48) за (6.49) зводиться до визначення значень функцій у вузлах $y_i = y(x_i)$.

Один зі способів різницевої апроксимації похідних побудований на диференціюванні інтерполяційних формул. Так, для одержання скінченнорізницевої апроксимації похідної $y_i^{(j)} = \dfrac{d^j y(x_i)}{dx^j}$ будуємо інтерполяційний поліном Лагранжа $P_k(x), k \geq j$, який проходить через $k + 1$ точку $y_{i-\left[\frac{k+1}{2}\right]},\ y_{i-\left[\frac{k+1}{2}\right]+1}, \ldots,\ y_i, \ldots,\ y_{i+\left[\frac{k+1}{2}\right]-1},\ y_{i+\left[\frac{k+1}{2}\right]}$.

Тоді

$$y_i^{(j)} = \frac{d^j}{dx^j} P_k(x_i).$$

Похибка такої апроксимації оцінюється як

$$R_i^{(j)} \sim f^{(k)}(\xi) h^{k+1},$$

де $\xi \in \left[x_{i-\left\lceil \frac{k+1}{2} \right\rceil}, x_{i+\left\lceil \frac{k+1}{2} \right\rceil} \right]$.

Як приклад, знайдемо скінченно різницеву апроксимацію першої і другої похідних функції $y(x)$ у точці $x = x_i$, побудовану в трьох точках. Побудуємо поліном Лагранжа, що проходить через точки (x_{i-1}, y_{i-1}), (x_i, y_i), (x_{i+1}, y_{i+1}):

$$P_2(x) = \frac{(x - x_i)(x - x_{i+1})}{2h^2} y_{i-1} - \frac{(x - x_{i-1})(x - x_{i+1})}{h^2} y_i + \frac{(x - x_{i-1})(x - x_i)}{2h^2} y_{i+1}.$$

Тоді

$$y_i' = \frac{d}{dx} P_2(x_i) = \frac{-y_{i-1} + y_{i+1}}{2h};$$

$$y_i'' = \frac{d^2}{dx^2} P_2(x_i) = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}.$$

Похибкою цих апроксимацій є $0(h^2)$.

У методі скінченних різниць диференціальне рівняння (6.48) заміняється скінченно різницевими рівняннями, записаними для $n - m + 1$ внутрішнього вузла відрізка $[0, l]$. Граничні умови (6.49) заміняються m скінченно різницевими рівняннями. У підсумку одержуємо СНУ (СЛАУ, якщо L – лінійний диференціальний оператор) з $n + 1$ рівняння відносно $n + 1$ невідомих y_i, $i = \overline{0,n}$. Похибка розв'язку методом скінченних різниць відповідає найбільшій похибці апроксимації похідних у (6.48) і (6.49).

Приклад:

$$\frac{d^2y}{dx^2} + y = x;$$
$$y(0) = 0;$$
$$y(1) = \frac{1}{3}.$$

Точний розв'язок:

$$y(x) = -\frac{2\sin x}{3\sin 1} + x\,.$$

Поділимо відрізок [0, 1] на три рівні відрізки. Тоді $h = \frac{1}{3}$, $x_0 = 0$, $x_1 = \frac{1}{3}$, $x_2 = \frac{2}{3}$, $x_3 = 1$.

Шукаємо $y_0 = y(0)$, $y_1 = y\left(\frac{1}{3}\right)$, $y_2 = y\left(\frac{2}{3}\right)$, $y_3 = y(1)$.

Використовуючи скінченно різницеву апроксимацію $y_i'' = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}$, запишемо скінченно різницеве рівняння для вузлів x_1, x_2:

$$\frac{y_0 - 2y_1 + y_2}{\left(\frac{1}{3}\right)^2} + y_1 = x_1; \qquad -17y_1 + 9y_2 = \frac{1}{3};$$

$$\frac{y_1 - 2y_2 + y_3}{\left(\frac{1}{3}\right)^2} + y_2 = x_2; \qquad 9y_1 - 17y_2 = -\frac{7}{3}.$$

З крайових умов:

$$y_0 = 0;$$
$$y_3 = \frac{1}{3}.$$

Одержимо систему чотирьох рівнянь стосовно чотирьох невідомих y_0, y_1, y_2, y_3. Розв'язуючи її, одержимо:

$$y_1 = \frac{23}{312} \cong 0{,}0737; \qquad y\left(\frac{1}{3}\right) \cong 0{,}0741;$$
$$y_2 = \frac{55}{312} \cong 0{,}1763; \qquad y\left(\frac{2}{3}\right) \cong 0{,}1768.$$

6.6.3. Проекційні методи розв'язання крайових задач

Подамо крайові умови (6.49) у вигляді

$$L_s y = B_s, \tag{6.54}$$

де L_s – векторний лінійний диференціальний оператор, j-й елемент якого дорівнює

$$L_{sj} = \sum_{i=0}^{m-1} \left[\alpha_{ij} \frac{d^i}{dx^i}\Big|_{x=0} + \beta_{ij} \frac{d^i}{dx^i}\Big|_{x=l} \right];$$

$B_s = \left[\gamma_0, \gamma_1, ..., \gamma_{m-1}\right]^{\mathrm{T}}$ – вектор сталих величин.

Для побудови проекційних методів уведемо дві повні системи лінійно-незалежних функцій [31] – базисну (координатну) $\{\varphi_i(x)\}$ і проекційну (вагову) $\{\psi_i(x)\}$ [25]. Подамо розв'язок задачі (6.48) з умовою (6.54) у вигляді розвинення за системою базисних функцій:

$$y(x) = \sum_{i=1}^{\infty} c_i \varphi_i(x), \tag{6.55}$$

де c_i – сталі, які треба визначити.

Підставивши вираз (6.55) у (6.48) і (6.54), одержимо:

$$L\sum_{i=1}^{\infty} c_i \varphi_i(x) - f(x) = R(x);$$

$$L_s \sum_{i=1}^{\infty} c_i \varphi_i(x) - B_s = 0, \qquad (6.56)$$

де $R(x)$ – відхил розв'язку, що залежить від різниці точного розв'язку і подання (6.55). Сталі c_i вибирають так, щоб відхил дорівнював нулю. Тотожне обернення відхилу в нуль під час чисельного розв'язання задачі, очевидно, неможливе. Алгоритм проекційних методів передбачає мінімізацію відхилу з тим, щоб точний і наближений розв'язок якнайменше відрізнялися один від одного.

Замість обернення відхилу в нуль вимагають, щоб оберталися в нуль усі проекції відхилу на множину проекційних функцій $\{\psi_i(x)\}$. Якщо цю вимогу виконано, усі скалярні добутки $\left(R(x),\ \psi_j(x)\right),\ \ j=\overline{1,\infty}$ дорівнюють нулю. Звідси

$$\left(L\sum_{i=1}^{\infty} c_i \varphi_i(x),\ \psi_j(x)\right) - \left(f(x),\ \psi_j(x)\right) = 0, \qquad (6.57)$$

де $j=\overline{1,\infty}$.

У загальному випадку рівняння (6.56) і (6.57) є системою нелінійних рівнянь відносно невідомих c_i. Розв'язавши цю систему, з (6.55) знаходять розв'язок крайової задачі. Описаний алгоритм відомий як «метод моментів» (метод Гальоркіна, метод Гальоркіна – Петрова) і є найважливішим із проекційних методів. Відомі окремі випадки методу моментів, що мають спеціальні назви. Так, у методі Гальоркіна – Бубнова базисна система функцій збігається з проекційною, тобто $\psi_i(x) = \varphi_i(x)$, $i=\overline{1,\infty}$. У методі найменших квадратів (МНК) $\psi_i(x) = L\varphi_i(x),\ \ i=\overline{1,\infty}$.

Нескінченну систему (6.56) з умовою (6.57) зазвичай розв'язують методом редукції. Для цього в розвиненні (6.55) утримують скінченне число $n > m$ членів:

$$y_n(x) = \sum_{i=1}^{n} c_i \varphi_i(x)$$

і потребують обернення в нуль проекційного відхилу на $n - m$ перших проекційних функцій. У підсумку одержують систему з n рівнянь щодо n невідомих c_i, $i = \overline{1,n}$. Потім n поступово збільшують, доки максимальне відхилення нового розв'язку від попереднього не буде перевищувати заданої похибки розв'язку

$$\max_{x \in [0,l]} \left| y_{n1}(x) - y_{n2}(x) \right| < \xi. \qquad (6.58)$$

Якщо для деяких n_1, n_2 умова (6.58) виконується, то кажуть, що проекційний метод збігається. Якщо з ростом n $\left| y_{n1}(x) - y_{n2}(x) \right|$ збільшується, то кажуть, що метод розбігається і є нестійким.

У багатьох задачах задають однорідні крайові умови, тобто такі, що одержані з (6.54) за умови $B_s = 0$. Крім того, багато крайових умов можна звести до однорідних за допомогою заміни шуканих функцій [32]. У цьому випадку підбирають таку систему базисних функцій, у якій кожна з $\varphi_i(x)$ задовольняє (6.54). Отже, і (6.55) буде задовольняти крайові умови. Тоді проекційний метод зводиться тільки до розв'язку (6.57).

Якщо L – лінійний диференціальний оператор, то $L\sum_{i=1}^{\infty} c_i \varphi_i(x) = \sum_{i=1}^{\infty} c_i L\varphi_i(x)$ і (6.57) зводиться до СЛАР відносно c_i, $i = \overline{1,\infty}$:

$$\sum_{i=1}^{\infty} c_i a_{ij} = b_j,\ j = \overline{1,\infty}, \qquad (6.59)$$

де $a_{ij} = (L\varphi_i(x),\ \psi_j(x)),\quad b_j = \left(f(x),\ \psi_j(x)\right)$.

Розв'язують СЛАР (6.59) методом редукції.

Інший спосіб побудови проекційних методів оснований на методі Рітца. У теорії варіаційного числення доводиться, що якщо $y(x)$ є розв'язком рівняння (6.48) з симетричним додатньо-визначеним оператором L, то ця функція зводить до мінімуму функціонал енергії [25]:

$$F(x) = (Ly,\ y) - 2\,\mathrm{Re}\left(y, f(x)\right). \tag{6.60}$$

Підставляючи рівняння (6.55) у (6.60), одержимо

$$F(y) = F(c_1,\ c_2, \ldots,\ c_n, \ldots). \tag{6.61}$$

Коефіцієнти c_i визначають з умови існування мінімуму функціонала (6.61)

$$\frac{\partial F(c_1,\ c_2, \ldots, c_n, \ldots)}{\partial c_j} = 0,\ j = \overline{1,\infty}, \tag{6.62}$$

яка приводить до системи рівнянь щодо невідомих c_i. Якщо L – лінійний диференціальний оператор, то (6.62) являє собою СЛАР, що збігається з методом Гальоркіна – Бубнова.

Приклад:

$$\frac{d^2y}{dx^2} + y = x;\ y(0) = 0;\ y(1) = \frac{1}{3}.$$

Зведемо задачу до однорідних крайових умов. Для цього шукаємо функцію $y(x)$ у вигляді

$$y(x) = \tilde{y}(x) + \frac{x}{3}.$$

Тоді крайова задача щодо $\tilde{y}(x)$ набуває вигляду

$$\frac{d^2\tilde{y}}{dx^2}+\tilde{y}=\frac{2}{3}x;\ \tilde{y}(0)=0;\ \tilde{y}(1)=0.$$

Як координатний базис використаємо повну систему функцій $\{\sin(i\pi x)\}$.

Отже, будемо шукати розв'язок у вигляді

$$\tilde{y}(x)=\sum_{i=1}^{\infty}c_i\sin(i\pi x).$$

Оскільки кожна з функцій $\sin(i\pi x)$ задовольняє крайові умови, то необхідно підібрати коефіцієнти c_i так, щоб задовольнити диференціальне рівняння. Використовуємо метод Гальоркіна – Бубнова. Підставляючи розвинення для $\tilde{y}(x)$ у вихідне диференціальне рівняння, одержуємо:

$$\sum_{i=1}^{\infty}\left[c_i\left(1-i^2\pi^2\right)\sin(i\pi x)\right]-\frac{2}{3}x=0.$$

Оскільки $\int_0^1\sin(i\pi x)\sin(j\pi x)dx=0$ якщо $i\neq j$, з умови рівності нулю проекції відхилу на систему функцій $\{\sin(i\pi x)\}$ одержуємо:

$$c_i\left(1-i^2\pi^2\right)\int_0^1\sin^2(i\pi x)dx-\int_0^1\frac{2}{3}x\sin(i\pi x)dx=0,\quad i=\overline{1,\infty}.$$

Обчислюючи інтеграли, знаходимо:

$$a_{ij}c_i=b_i,\ i=\overline{1,\infty};$$

$$a_{ij}=\frac{1}{2}\left(1-i^2\pi^2\right),\ b_i=\frac{2}{3}\frac{(-1)^{i+1}}{i\pi};$$

$$c_i\left(1-i^2\pi^2\right)\frac{1}{2}=\frac{2}{3}\frac{(-1)^{i+1}}{i\pi},\ i=\overline{1,\infty}.$$

Розв'язуючи СЛАР, маємо

$$c_i = \frac{b_i}{a_{ij}} = \frac{4(-1)^{i+1}}{3i\pi\left(1-i^2\pi^2\right)}.$$

Тоді:

$$\tilde{y}_n(x) \approx \sum_{i=1}^{n} \frac{4(-1)^{i+1}}{3i\pi\left(1-i^2\pi^2\right)} \sin(i\pi x);$$

$$y_n(x) \approx \sum_{i=1}^{n} \frac{4(-1)^{i+1}}{3i\pi\left(1-i^2\pi^2\right)} \sin(i\pi x) + \frac{x}{3},$$

де $\tilde{y}_n(x)$ та $y_n(x)$ – n-ті наближення до розв`язку.

n	$y_n(1/3)$	$y_n(2/3)$
1	0,06967	0,18078
2	0,07445	0,17601
3	0,07445	0,17601
4	0,07386	0,17659
5	0,074161	0,17689

6.6.4. Метод скінченних елементів розв’язання крайових задач

Метод скінченних елементів (МСЕ) належить до групи проекційних методів. Однак як базисні функції в ньому використовують спеціальні фінітні функції [33, 34], відмінні від нуля тільки на невеликому відрізку, на якому шукається розв’язок задачі. Фінітні функції будують у такий спосіб:

1) відрізок [0, l] розбивають на скінченне число n відрізків D_i, $i = \overline{1,n}$, які не перетинаються. Ці відрізки називаються скінченними елементами;

2) у кожному скінченному елементі фіксуються $r \geq m+1$ точок (вузлів) x_{ij}, $j = \overline{1,r}$, що можуть розташовуватися як усередині елемента, так і на його кінцях;

3) у кожному скінченному елементі D_i вводиться r базисних функцій $\varphi_{ij}(x)$, $\overline{1,r}$ (рис. 6.1), які дорівнюють нулю поза цим елементом і мають властивість

$$\varphi_{ij}\left(x_{ik}\right)=\begin{cases}0, & j\neq k,\\ 1, & j=k,\end{cases} \tag{6.63}$$

яка дозволяє записати наближений розв'язок у межах елемента D_i у вигляді

$$y(x)=\sum_{j=1}^{r} y_{ij}\varphi_{ij}(x),$$

де y_{ij} – значення $\varphi_{ij}(x)$ у вузлах x_{ij}.

Множину точок, у яких $\varphi_{ij}(x) \neq 0$, називають носієм функції. Оскільки поза елементом D_i $\varphi_{ij}(x) \equiv 0$ (носії сусідніх елементів не перетинаються), то скінченно-елементна апроксимація функції $y(x)$ на усьому відрізку $[0, l]$ має вигляд

$$y(x)=\sum_{i=1}^{n}\sum_{j=1}^{r} y_{ij}\varphi_{ij}(x). \tag{6.64}$$

Вираз (6.64) забезпечує неперервність функції $y(x)$ і, можливо, її похідних до деякого порядку у вузлах, розташованих на кінцях суміжних елементів. Точність його збільшується як за рахунок використання більшого числа базисних функцій, так і за рахунок дрібнішого розбиття області на скінченні елементи зі збереженням постійного числа базисних функцій у кожному з них.

У виразі (6.64) значення функції в тому самому вузлі, що належить двом сусіднім кінцевим елементам, повторюється. Тому (6.64) можна подати так:

$$y(x)=\sum_{i=1}^{n}\sum_{j=1}^{r} y_{ij}\varphi_{ij}(x)-\frac{1}{2}\sum_{i=1}^{n-1} y_{ir}\left(\varphi_{ir}(x)+\varphi_{i+1,1}(x)\right). \tag{6.65}$$

Підставляючи формулу (6.65) у (6.54) і враховуючи, що в розвиненні (6.65) залишаться тільки базисні функції, носіями яких є елементи D_1 і D_n, що лежать на межах, одержимо m рівнянь:

$$L_s\left(\sum_{j=1}^{r}\left(y_{1j}\varphi_{1j}(x)+y_{nj}\varphi_{nj}(x)\right)-\frac{1}{2}\left(y_{1r}\varphi_{1r}(x)+y_{n1}\varphi_{n1}(x)\right)\right)=B_s. \quad (6.66)$$

Далі використаємо стандартну процедуру проекційних методів, причому за проекційний базис візьмемо базисні функції, носіями яких є всі $n(r-2)-2$ внутрішні вузли всіх скінченних елементів D_2, D_3,…, $D_{n\text{-}1}$, $2r-2-m$ базисні функції, носіями яких є внутрішні вузли скінченних елементів D_1 та D_n, а також n–1 базисна функція ($\varphi_{ir}(x)+\varphi_{i+1,1}(x)$) з носіями у вузлах, що одночасно належать сусіднім елементам. m базисних функцій, носіями яких є крайні точки скінченних елементів D_1 та D_n, у проекційний базис не включають, оскільки за їх допомогою вже визначено значення функції на кінцях відрізка (із граничних умов). У цьому випадку враховують, що всі скалярні добутки базисних функцій на носіях, що не збігаються, дорівнюють нулю. У підсумку одержимо систему рівнянь:

$$\left(L\left(\sum_{j=1}^{r-1}y_{1j}\varphi_{1j}(x)+\frac{1}{2}y_{1r}\varphi_{1r}(x)\right),\ \varphi_{1s}(x)\right)-\left(f(x),\varphi_{1s}(x)\right)=0,\ \ s=\overline{\left[\frac{m}{2}\right]+1,r-1};$$

$$\left(L\left(\sum_{j=2}^{r-1}y_{ij}\varphi_{ij}(x)+\frac{1}{2}\left(y_{i1}\varphi_{i1}(x)+y_{ir}\varphi_{ir}(x)\right)\right),\ \varphi_{is}(x)\right)-\left(f(x),\varphi_{is}(x)\right)=0,\ \ i=\overline{2,n-1},\ s\neq 1,\ s\neq r;$$

$$\left(L\left(\sum_{j=2}^{r}y_{nj}\varphi_{nj}(x)+\frac{1}{2}y_{n1}\varphi_{n1}(x)\right),\ \varphi_{ns}(x)\right)-\left(f(x),\varphi_{ns}(x)\right)=0,\ \ s=\overline{2,r-\left[\frac{m}{2}\right]};$$

$$\left(L\left(\sum_{j=2}^{r-1}y_{ij}\varphi_{ij}(x)+\frac{1}{2}\left(y_{i1}\varphi_{i1}(x)+y_{ir}\varphi_{ir}(x)\right)\right),\ \ \varphi_{ir}(x)\right)+$$

$$+\left(L\left(\sum_{j=2}^{r-1}y_{i+1,j}\varphi_{i+1,j}(x)+\frac{1}{2}\left(y_{i+1,1}\varphi_{i+1,1}(x)+y_{i+1,r}\varphi_{i+1,r}(x)\right)\right),\ \ \varphi_{i+1,1}(x)\right)- \quad (6.67)$$

$$-\left(f(x),\varphi_{ir}(x)+\varphi_{i+1,1}(x)\right)=0,\ \ i=\overline{2,n-1}.$$

Таким чином, одержали систему $n(r-1)+1$ нелінійних рівнянь щодо $n(r-1)+1$ невідомих y_{ij}.

Якщо L – лінійний диференціальний оператор, то вирази (6.66) і (6.67) являють собою СЛАР щодо невідомих значень функції y_{ij}.

Приклад:

$$\begin{cases} \dfrac{d^2y}{dx^2}+y=x; \\ y(0)=0; \\ y(1)=\dfrac{1}{3}; \end{cases} \quad \Rightarrow \quad \begin{cases} Ly=f(x); \\ L=\dfrac{d^2}{dx^2}+1; \\ f(x)=x. \end{cases}$$

Розіб'ємо відрізок [0, 1] на два скінченні елементи – $D_1: x\in\left[0,\dfrac{1}{2}\right]$, $D_2: x\in\left[\dfrac{1}{2},1\right]$. У кожному скінченному елементі зафіксуємо по три вузли і введемо по три базисні функції, які дорівнюють нулю поза скінченним елементом. Для задоволення умови (6.63) побудуємо базисні функції у вигляді інтерполяційних поліномів Лагранжа:

$$\varphi_{11}(x)=\frac{\left(x-\frac{1}{3}\right)\left(x-\frac{1}{2}\right)}{\left(-\frac{1}{3}\right)\left(-\frac{1}{2}\right)}=6\left(x-\frac{1}{3}\right)\left(x-\frac{1}{2}\right);$$

$$\varphi_{21}(x)=\frac{\left(x-\frac{2}{3}\right)(x-1)}{\left(\frac{1}{2}-\frac{2}{3}\right)\left(\frac{1}{2}-1\right)}=12\left(x-\frac{2}{3}\right)(x-1);$$

$$\varphi_{12}(x)=\frac{x\left(x-\frac{1}{2}\right)}{\frac{1}{3}\left(\frac{1}{3}-\frac{1}{2}\right)}=-18x\left(x-\frac{1}{2}\right);$$

$$\varphi_{22}(x)=\frac{\left(x-\frac{1}{2}\right)(x-1)}{\left(\frac{2}{3}-\frac{1}{2}\right)\left(\frac{2}{3}-1\right)}=-18\left(x-\frac{1}{2}\right)(x-1);$$

$$\varphi_{13}(x)=\frac{x\left(x-\frac{1}{3}\right)}{\frac{1}{2}\left(\frac{1}{2}-\frac{1}{3}\right)}=12x\left(x-\frac{1}{3}\right);$$

$$\varphi_{23}(x)=\frac{\left(x-\frac{1}{2}\right)\left(x-\frac{2}{3}\right)}{\left(1-\frac{1}{2}\right)\left(1-\frac{2}{3}\right)}=6\left(x-\frac{1}{2}\right)\left(x-\frac{2}{3}\right).$$

$$y(x)=y_{11}\varphi_{11}(x)+y_{12}\varphi_{12}(x)+\frac{1}{2}y_{13}\left(\varphi_{13}(x)+\varphi_{21}(x)\right)+y_{22}\varphi_{22}(x)+y_{23}\varphi_{23}(x).$$

Побудуємо проекційну схему, причому за проекційний базис використаємо базисні функції з носіями у всіх внутрішніх вузлах $\varphi_{12}(x)$ і $\varphi_{22}(x)$ і функцію $\varphi_{13}(x)+\varphi_{21}(x)$ з носієм у вузлі, що належить одночасно елементам D_1 і D_2:

$$y_{11}\underbrace{\int_0^{1/2}L\varphi_{11}(x)\varphi_{12}(x)dx}_{\frac{183}{40}}+y_{12}\underbrace{\int_0^{1/2}L\varphi_{12}(x)\varphi_{12}(x)dx}_{-\frac{1053}{80}}+\frac{1}{2}y_{13}\underbrace{\int_0^{1/2}L\varphi_{13}(x)\varphi_{12}(x)dx}_{\frac{717}{80}}=$$

$$=\underbrace{\int_0^{1/2}f(x)\varphi_{12}(x)dx}_{\frac{3}{32}};$$

$$\frac{1}{2}y_{21}\underbrace{\int_{1/2}^{1} L\varphi_{21}(x)\varphi_{22}(x)dx}_{\frac{717}{80}} + y_{22}\underbrace{\int_{1/2}^{1} L\varphi_{22}(x)\varphi_{22}(x)dx}_{-\frac{1053}{80}} + y_{23}\underbrace{\int_{1/2}^{1} L\varphi_{23}(x)\varphi_{22}(x)dx}_{\frac{183}{40}} =$$

$$= \underbrace{\int_{1/2}^{1} f(x)\varphi_{22}(x)dx}_{\frac{9}{32}};$$

$$y_{11}\underbrace{\int_{0}^{1/2} L\varphi_{11}(x)\varphi_{13}(x)dx}_{-\frac{7}{240}} + y_{12}\underbrace{\int_{0}^{1/2} L\varphi_{12}(x)\varphi_{13}(x)dx}_{-\frac{3}{80}} + \frac{1}{2}y_{13}\times$$

$$\times\left[\underbrace{\int_{0}^{1/2} L\varphi_{13}(x)\varphi_{12}(x)dx}_{\frac{1}{30}} + \underbrace{\int_{1/2}^{1} L\varphi_{21}(x)\varphi_{21}(x)dx}_{\frac{1}{30}}\right] +$$

$$+y_{22}\underbrace{\int_{1/2}^{1} L\varphi_{22}(x)\varphi_{21}(x)dx}_{-\frac{3}{80}} + y_{23}\underbrace{\int_{1/2}^{1} L\varphi_{23}(x)\varphi_{21}(x)dx}_{-\frac{7}{240}} = \underbrace{\int_{0}^{1/2} f(x)\varphi_{13}(x)dx}_{\frac{1}{48}} + \underbrace{\int_{1/2}^{1} f(x)\varphi_{21}(x)dx}_{-\frac{1}{48}}.$$

Враховуючи, що

$\varphi_{12}(0) = \varphi_{13}(0) = 0;\ \varphi_{2i}(0) = 0,\ i = \overline{1,3};\ \varphi_{21}(1) = \varphi_{22}(1) = 0;\ \varphi_{1i}(1) = 0,\ i = \overline{1,3},$

із крайових умов маємо:

$$y_{11} = 0,\quad y_{23} = 1/3.$$

Одержуємо СЛАР:

$$-\frac{1053}{80}y_{12} + \frac{717}{160}y_{13} = \frac{3}{32},\ \frac{717}{160}y_{13} - \frac{1053}{80}y_{22} = -\frac{199}{160},$$
$$-\frac{3}{80}y_{12} + \frac{1}{30}y_{13} - \frac{3}{80}y_{22} = \frac{7}{720}.$$

Розв’язуючи систему, одержуємо:

y_{12} =0,0745, $y(1/3)$=0,0741,
$y_{13} = y_{21}$ =0,1209, $y(1/2)$=0,1202,
y_{22} = 0,1784, $y(2/3)$=0,1768.

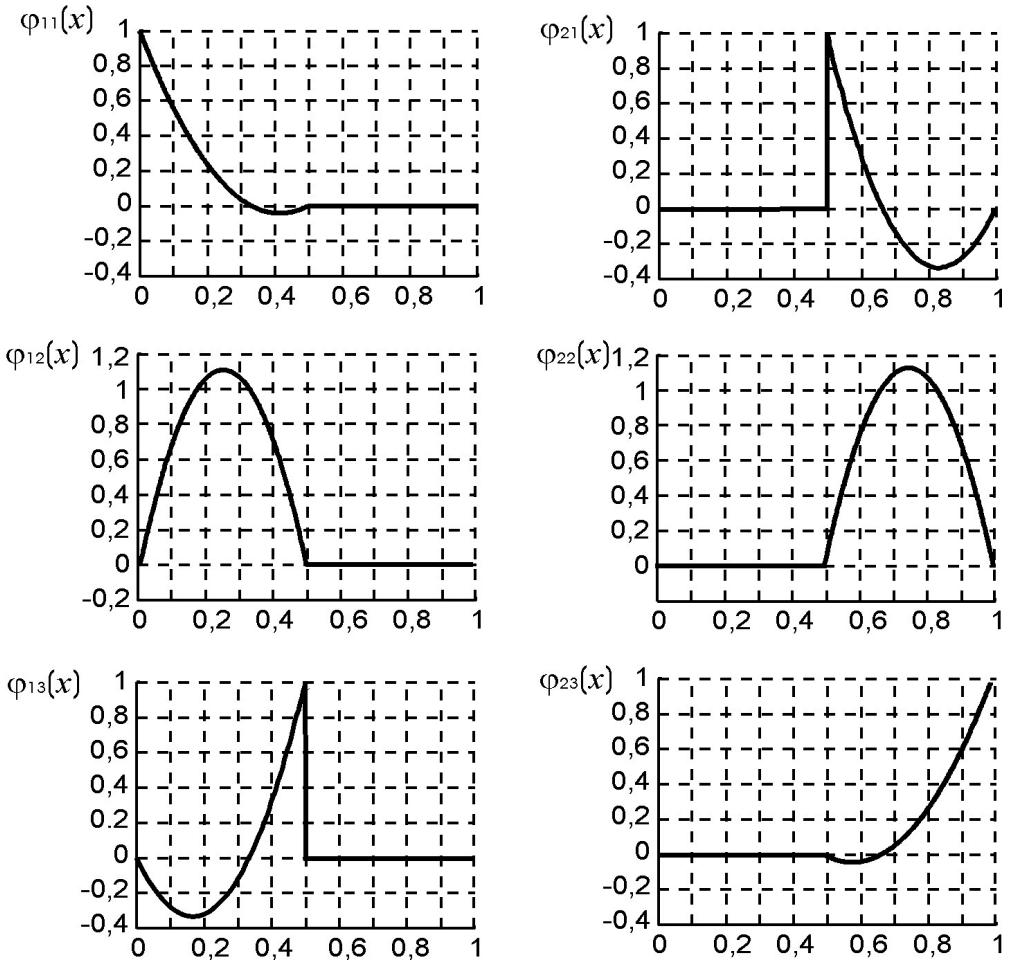


Рис. 6.2. Базис фінітних функцій

Контрольні завдання

1. Вибрати задачі Коші відповідно до свого варіанта.
2. Розв’язати одновимірну задачу Коші методом Ейлера з похибкою не більше 5 %.
3. Розв’язати одновимірну задачу Коші методом Рунге – Кутта 4-го порядку з похибкою не більше 5 %.
4. Розв’язати одновимірну задачу Коші явним методом Адамса 4-го порядку з похибкою не більше 5 %.
5. Розв’язати одновимірну задачу Коші методом Гіра 4-го порядку з похибкою не більше 5 %.

6. Порівняти результати пп. 2 – 5.
7. Розв’язати систему ЗДР методом Ейлера. Оцінити похибку розв’язку.

Варіанти завдань

1. $\dfrac{dy}{dt} = t^2 - 3y;$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = \dfrac{y}{2}, \\ \dfrac{dy}{dt} = -x, \end{cases}$$
$x(0) = 1,\ y(0) = 1,\ t \in [0;\ 1].$

2. $\dfrac{dy}{dt} = y(1 - t);$ $y(0) = 1,\quad t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = x - y, \\ \dfrac{dy}{dt} = xy, \end{cases}$$
$x(0) = 1,\ y(0) = 1,\ t \in [0;\ 1].$

3. $\dfrac{dy}{dt} = ty - t;$ $y(0) = 0,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = \sin(t) - y, \\ \dfrac{dy}{dt} = \cos(t) - x, \end{cases}$$
$x(0) = 0,\ y(0) = 1,\ t \in [0;\ 1].$

4. $\dfrac{dy}{dt} = 4t^2 - y;$ $y(0) = 0,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = x - y, \\ \dfrac{dy}{dt} = x + y, \end{cases}$$
$x(0) = 1,\ y(0) = 0,\ t \in [0;\ 1].$

5. $\dfrac{dy}{dt} = \cos(t + y);$ $y(0) = 0,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y, \\ \dfrac{dy}{dt} = -t, \end{cases}$$
$x(0) = 0,\ y(0) = 1,\ t \in [0;\ 1].$

6. $\dfrac{dy}{dt} = \sin(t - y);$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y - x, \\ \dfrac{dy}{dt} = t, \end{cases}$$
$x(0) = 1,\ y(0) = 0,\ t \in [0;\ 1].$

7. $\dfrac{dy}{dt} = -\sin(t + y - 1);$ $y(0) = 0,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = x - y, \\ \dfrac{dy}{dt} = t, \end{cases}$$

$x(0) = 1,\ y(0) = 0,\ t \in [0;\ 1].$

8. $\dfrac{dy}{dt} = \cos(2t) + y;$ $y(0) = -\dfrac{1}{2},\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y - x, \\ \dfrac{dy}{dt} = -t^2, \end{cases}$$

$x(0) = 0,\ y(0) = 1,\ t \in [0;\ 1].$

9. $\dfrac{dy}{dt} = \sin(t - 1) + y,$ $y(0) = \dfrac{1}{2},\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y + t, \\ \dfrac{dy}{dt} = x - 2t, \end{cases}$$

$x(0) = 0,\ y(0) = \dfrac{1}{2},\ t \in [0;\ 1].$

10. $\dfrac{dy}{dt} = \sin(t) - y;$ $y(0) = 0,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y + t^2, \\ \dfrac{dy}{dt} = x - t^2, \end{cases}$$

$x(0) = \dfrac{1}{2},\ y(0) = -1,\ t \in [0;\ 1].$

11. $\dfrac{dy}{dt} = \cos(t) - y;$ $y(0) = 0,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = t - y, \\ \dfrac{dy}{dt} = x - t, \end{cases}$$

$x(0) = 0,\ y(0) = 1,\ t \in [0; 1].$

12. $\dfrac{dy}{dt} = (1 - 2y)\sqrt{1 + t};$ $y(0) = 0,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y + \sin(t), \\ \dfrac{dy}{dt} = x + t, \end{cases}$$

$x(0) = -1,\ y(0) = 0,\ t \in [0;\ 1].$

13. $\dfrac{dy}{dt} = t\sqrt{y};$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y + \cos(t), \\ \dfrac{dy}{dt} = x + t, \end{cases}$$
$x(0) = -1,\ y(0) = 0,\ t \in [0;\ 1].$

14. $\dfrac{dy}{dt} = ty;$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y - x^2, \\ \dfrac{dy}{dt} = x + t, \end{cases}$$
$x(0) = 1,\ y(0) = 0,\ t \in [0;\ 1].$

15. $\dfrac{dy}{dt} = -y^2;$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y - \cos(t), \\ \dfrac{dy}{dt} = x + t, \end{cases}$$
$x(0) = 1,\ y(0) = 0,\ t \in [0;\ 1].$

16. $\dfrac{dy}{dt} = -ty^2;$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y - \cos(t), \\ \dfrac{dy}{dt} = t - x, \end{cases}$$
$x(0) = 0,\ y(0) = 0,\ t \in [0;\ 1].$

17. $\dfrac{dy}{dt} = t^2 y;$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y - \sin(t), \\ \dfrac{dy}{dt} = t - x, \end{cases}$$
$x(0) = 1,\ y(0) = 0,\ t \in [0;\ 1].$

18. $\dfrac{dy}{dt} = e^t + y;$ $y(0) = 0,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y - e^t, \\ \dfrac{dy}{dt} = t - x, \end{cases}$$
$x(0) = 1,\ y(0) = 0,\ t \in [0;\ 1].$

19. $\dfrac{dy}{dt} = e^t - y;$ $y(0) = 0,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y + e^t, \\ \dfrac{dy}{dt} = t - x, \end{cases}$$
$x(0) = -1,\ y(0) = -1,\ t \in [0;\ 1].$

20. $y' = -\sqrt{t}\,y^2;$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y + e^t, \\ \dfrac{dy}{dt} = t + x, \end{cases}$$
$x(0) = -1,\ y(0) = 0,\ t \in [0;\ 1].$

21. $\dfrac{dy}{dt} = t^2\sqrt{y};$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y - e^t, \\ \dfrac{dy}{dt} = x + t, \end{cases}$$
$x(0) = 1,\ y(0) = 0,\ t \in [0;\ 1].$

22. $\dfrac{dy}{dt} = \sqrt{ty};$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = x - 2y - t, \\ \dfrac{dy}{dt} = x + y + t, \end{cases}$$
$x(0) = 1,\ y(0) = 1,\ t \in [0;\ 1].$

23. $\dfrac{dy}{dt} = -t^2y^2;$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = 2x + y - t, \\ \dfrac{dy}{dt} = y - x + t, \end{cases}$$
$x(0) = 0,\ y(0) = 1,\ t \in [0;\ 1].$

24. $\dfrac{dy}{dt} = y\sin(t);$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = 2y - t, \\ \dfrac{dy}{dt} = t - x, \end{cases}$$
$x(0) = 1,\ y(0) = 1,\ t \in [0;\ 1].$

25. $\dfrac{dy}{dt} = y\cos(t);$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y + \cos^2(t), \\ \dfrac{dy}{dt} = x, \end{cases}$$
$x(0) = -1,\ y(0) = 0,\ t \in [0;\ 1].$

26. $\dfrac{dy}{dt} = -y^2\cos(t);$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y + \sin^2(t), \\ \dfrac{dy}{dt} = x, \end{cases}$$
$x(0) = -1,\ y(0) = 1,\ t \in [0;\ 1].$

27. $\dfrac{dy}{dt} = -y^2\sin(t);$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y + e^{2t}, \\ \dfrac{dy}{dt} = x, \end{cases}$$
$x(0) = -1,\ y(0) = -1,\ t \in [0;\ 1].$

28. $\dfrac{dy}{dt} = y\sin^2(t);$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y + e^{(t/2)}, \\ \dfrac{dy}{dt} = -x, \end{cases}$$
$x(0) = 1,\ y(0) = 1,\ t \in [0;\ 1].$

29. $\dfrac{dy}{dt} = y\cos^2(t);$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y + \sin(2t), \\ \dfrac{dy}{dt} = x, \end{cases}$$
$x(0) = 1,\ y(0) = -1,\ t \in [0;\ 1].$

30. $\dfrac{dy}{dt} = -yt^3;$ $y(0) = 1,\ t \in [0;\ 1].$

$$\begin{cases} \dfrac{dx}{dt} = y + \cos(2t), \\ \dfrac{dy}{dt} = x, \end{cases}$$
$x(0) = 0,\ y(0) = 0,\ t \in [0;\ 1].$

7. Чисельне розв’язання задач оптимізації

Часто наукові та інженерно-технічні обчислення полягають у визначенні максимуму чи мінімуму (і відповідних аргументів) дійсної функції $f(x_1, x_2, \ldots x_n)$ від n дійсних змінних у області D n-мірного простору [35]. Слово «оптимізація» означає або мінімізацію, або максимізацію функції. Іноді D збігається з усім n-вимірним простором; у цьому випадку задача називається безумовною. Якщо ні, то задача має обмеження, тобто умови, що визначають область D. Зазвичай D визначається сукупністю нелінійних функцій, що задовольняють певні рівняння чи нерівності. Іншими словами, точка **X** n-мірного простору з координатами $(x_1, x_2, \ldots x_n)$ належить D тоді й тільки тоді, коли **X** задовольняє нерівності

$$g_i \geq 0,\ i = \overline{1,n},$$

де g_i – задані функції від **X**.

Функцію $f(x_1, x_2, \ldots x_n)$, що підлягає мінімізації, називають *цільовою функцією* [6]. Методи обчислення мінімумів тривіально переносяться на задачу максимізації (оскільки мінімуми функції f є максимумами для $-f$).

Якщо функція f і всі обмеження g_i є лінійними функціями, то говорять про задачу лінійного програмування. У цьому випадку розв’язання лежить у вершині опуклого многогранника, описаного обмеженнями в n-мірному просторі. Звичайний метод розв’язання полягає в пошуку потрібної вершини переміщенням від наступної вершини до суміжної [16]. Якщо функція f або яке-небудь з обмежень нелінійні, то кажуть про задачу нелінійного програмування.

Нехай функція $f(x_1, x_2,..., x_n)$ — неперервно диференційована в області D. Нехай відомо, що в D існує точка $\mathbf{X} = \left(x_1, x_2, \ldots x_n\right)$, в якій виконуються співвідношення:

$$\begin{aligned} &\frac{\partial}{\partial x_1} f(x_1, x_2,..., x_n) = 0 \\ &\frac{\partial}{\partial x_2} f(x_1, x_2,..., x_n) = 0 \\ &..... \\ &\frac{\partial}{\partial x_n} f(x_1, x_2,..., x_n) = 0 \end{aligned} \qquad (7.1)$$

Розв'язання системи (7.1) дозволяє визначити невідомі характеристичні точки. Потім, використовуючи достатні умови мінімуму, варто виділити ті точки області D, у яких досягається локальний мінімум [35]. Порівнявши значення цільової функції в точках локального мінімуму, треба вибрати точку з мінімальним значенням локального мінімуму, тобто визначити глобальний мінімум $f(x_1, x_2,..., x_n)$ у області D. Однак надійні й одночасно економні методи пошуку глобального мінімуму на цей час невідомі.

Необхідну умову екстремуму (7.1) – умову стаціонарності можна записати у вигляді

$$\nabla f(\mathbf{X}) = 0,$$

де $\nabla f(\mathbf{X}) = \left[\frac{\partial f(\mathbf{X})}{\partial x_1}, \frac{\partial f(\mathbf{X})}{\partial x_2},..., \frac{\partial f(\mathbf{X})}{\partial x_n} \right]^{\mathrm{T}}$ – градієнт цільової функції в точці $\mathbf{X} = \left[x_1, x_2,..., x_n\right]^{\mathrm{T}}$.

Дослідимо максимально можливу точність розв'язання задачі оптимізації. Нехай x^* — точка мінімуму функції $f(x)$. Розкладемо в околі цієї точки функцію $f(x)$ в ряд Тейлора. Тоді

$$f(x^* + \Delta x) = f(x^*) + \Delta x f'(x^*) + \Delta x^2 \frac{f''(x^*)}{2} + \cdots .$$

Оскільки в точці екстремума $f'\left(x^*\right) = 0$, то

$$f(x^* + \Delta x) \approx f(x^*) + c\Delta x^2 ,$$

де $c = \frac{f''(x^*)}{2}$. Звідси випливає, що для того, аби відхилення функції $f(x)$ від значення в стаціонарній точці $f(x^*)$ були помітними, необхідно, щоб

$$\Delta x \sim \sqrt{\varepsilon} ,$$

де ε — відносна похибка округлення дійсних чисел. Таким чином, якщо нелінійне рівняння $f(x) = 0$ можна розв'язати з похибкою ε, то задачу оптимізації $\min\left\{f\left(x\right)\right\}$ — тільки з похибкою $\sqrt{\varepsilon}$ [11, с. 242] (див. рис. 7.1).

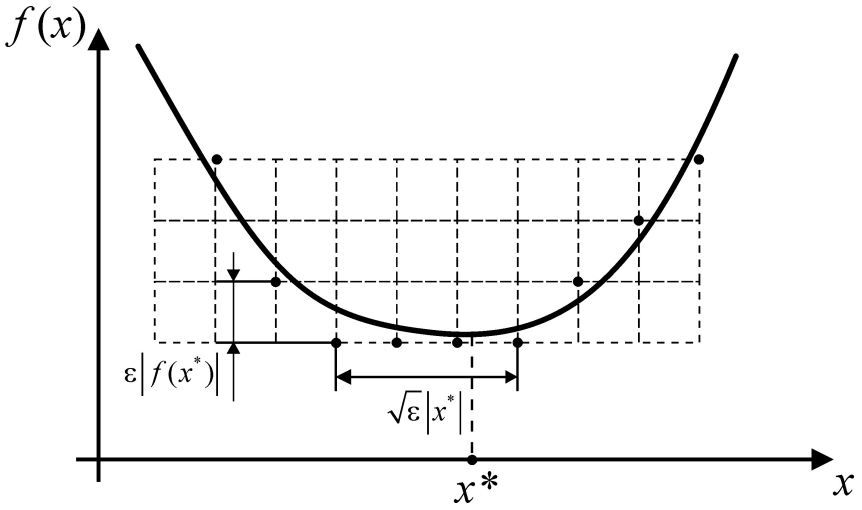


Рис. 7.1. Гранична похибка розв'язання задачі оптимізації

Найчисленнішу групу методів розв’язання задач оптимізації складають локальні методи безумовної оптимізації [6]. Деяке уявлення про методи, які широко застосовуються цією групою, ілюструє рис. 7.2.

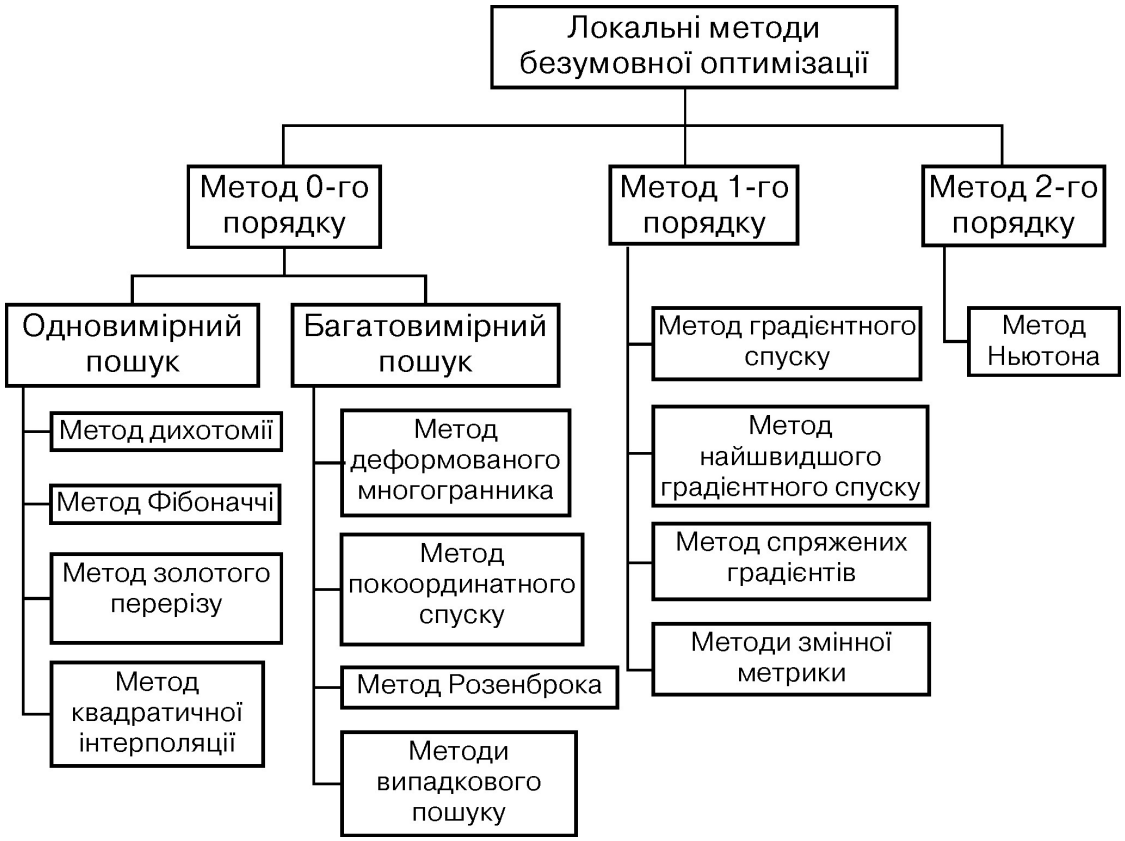


Рис. 7.2. Класифікація методів локальної безумовної оптимізації

Залежно від порядку використовуваних похідних цільової функції методи безумовної оптимізації поділяють на методи 0-го, 1-го і 2-го порядків.

Практично всі методи оптимізації прагнуть побудувати таку послідовність значень $\mathbf{X}_0$, $\mathbf{X}_1$, $\mathbf{X}_2$…, за якої $f(\mathbf{X}_0) > f(\mathbf{X}_1) > f(\mathbf{X}_2) >$… У цьому випадку метод забезпечує збіжність і можна сподіватися, що мінімум функції буде знайдений.

Схему алгоритму пошуку для загального випадку показано на рис. 7.3. Суть методу оптимізації визначають 2 і 3 етапами алгоритму, на яких

вибирається напрям подальшого пошуку й обчислюються координати наступної точки $\mathbf{X}_{i+1}$ на траєкторії пошуку.

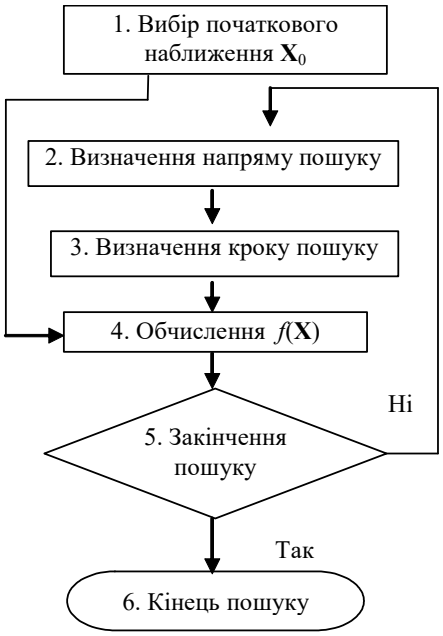


Рис. 7.3. Узагальнена схема оптимізації

7.1. Методи одновимірного пошуку

Нехай $f(x)$ – дійсна функція одного аргументу, визначена на відрізку $[a, b]$. Припустимо, що існує єдине значення x^*, таке, що $f(x^*)$ – мінімум $f(x)$ на відрізку $[a, b]$, і що $f(x)$ строго спадає для $x < x^*$ та строго зростає для $x > x^*$. Така функція називається унімодальною [11, с. 239]. Для її графіка існують три можливі форми, показані на рис. 7.4.

Зауважимо, що унімодальна функція не зобов'язана бути гладкою чи навіть неперервною.

Методи одновимірного пошуку будують на припущенні унімодальності функції $f(x)$ на відрізку $[a, b]$. Їх розділяють на методи послідовного пошуку (метод дихотомії, Фібоначчі та золотого перерізу) і

методи, що використовують апроксимацію функції (методи квадратичної і кубічної інтерполяції) [6, 11, с. 259].

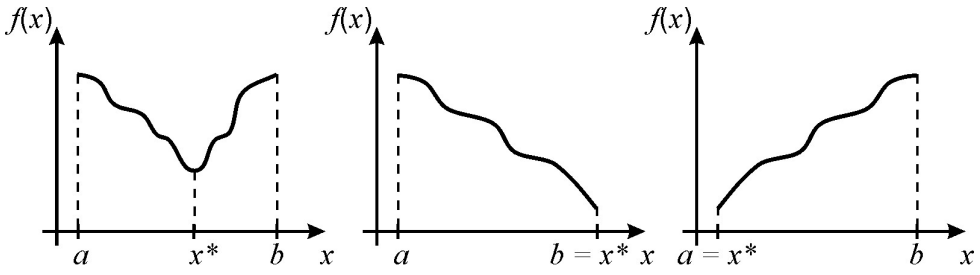


Рис. 7.4. Унімодальні функції

Розглянемо методи послідовного пошуку. Нехай відомо, що $f(x)$ унімодальна на $[a, b]$. Тоді за будь-якими двома значеннями $f(x_1)$, $f(x_2)$ можна вказати відрізок, у якому знаходиться точка x^*, яка відповідає мінімуму $f(x)$, причому цей відрізок має довжину меншу, ніж вихідний. Нехай $x_1 < x_2$. Можливі варіанти показано на рис. 7.5.

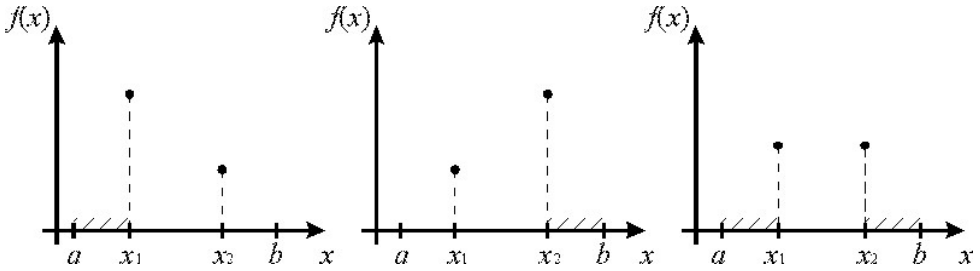


Рис. 7.5. Послідовний пошук мінімуму

У першому випадку слід відкинути відрізок $[a, x_1]$, у другому випадку – $[x_2, b]$, у третьому – $[a, x_1]$, $[x_2, b]$, оскільки в цих відрізках не може знаходитися x^*, за інших випадків порушується припущення про унімодальність $f(x)$. Завдання полягає в тому, щоб знайти таку множину абсцис $x_1, x_2, \ldots, x_n$, у якій мінімум функції $f(x)$ лежить у деякому відрізку $x^* \in [x_{i-1}, x_i]$. Такий відрізок називають *інтервалом невизначеності*.

Стратегія вибору значень x_{i-1} і x_i для проведення дослідів з урахуванням попередніх результатів визначає сутність різних методів послідовного пошуку.

У методі дихотомії (половинного поділу) на кожному кроці аргументи x_1 і x_2 вибирають на відстані $\delta/2$ праворуч і ліворуч від середини відрізку:

$$x_1 = \frac{a+b}{2} - \frac{\delta}{2}, \; x_2 = \frac{a+b}{2} + \frac{\delta}{2}.$$

Обчислюючи $f(x_1)$ і $f(x_2)$, знаходять новий інтервал невизначеності:

– якщо $f(x_1) < f(x_2)$, то $b = x_2$;

– якщо $f(x_1) = f(x_2)$, то $a = x_1$, $b = x_2$;

– якщо $f(x_1) > f(x_2)$, то $a = x_1$.

Потім знову обчислюють x_1 і x_2 та продовжують пошук, який припиняють, коли довжина інтервалу невизначеності $|b–a|$ стає меншою від заданої похибки визначення x^*.

Золотий переріз, відкритий Евклідом, полягає в розбитті відрізку $[a, b]$ точкою x_1 на дві частини таким чином, щоб відношення довжини всього відрізку до більшої частини дорівнювало відношенню більшої частини до меншої:

$$\frac{b-a}{b-x_1} = \frac{b-x_1}{x_1-a}.$$

Легко перевірити, що золотий переріз утворюють дві точки

$$x_1 = a + (1 - \delta)(b - a);$$
$$x_2 = a + \delta(b - a),$$

де $\delta = \frac{\sqrt{5}-1}{2} \approx 0.618$.

Точка x_1 утворює золотий переріз відрізка $[a, x_2]$, а точка x_2 – відрізку $[x_1, b]$. Тому на відрізку, що залишився, потрібно визначити тільки одну точку, що утворює золотий переріз. На кожному кроці довжина першого

інтервалу невизначеності дорівнює приблизно 0,618 довжини попереднього інтервалу.

Алгоритм методу золотого перерізу:

1. Обчислити:

$x_1 := a + (1 - \delta)(b - a)$;
$x_2 := a + \delta(b - a)$.

2. Обчислити $f(x_1)$, $f(x_2)$.

3. Якщо $f(x_1) < f(x_2)$, то покласти $b := x_2$, $x_2 := x_1$, $f(x_2) := f(x_1)$, $x_1 := a + (1 - \delta)(b - a)$ і обчислити $f(x_1)$, інакше покласти $a := x_1$, $x_1 := x_2$, $f(x_1) := f(x_2)$, $x_2 := a + \delta(b - a)$ і обчислити $f(x_2)$.

4. Якщо $|b - a| > \varepsilon \left| \frac{b + a}{2} \right|$, то перейти до кроку 3, інакше $x^* \approx \frac{b + a}{2}$.

З методів, що використовують апроксимацію функції, розглянемо алгоритм методу квадратичної інтерполяції. У цьому методі серед чотирьох рівновіддалених вузлів вибирають три найближчі до очікуваного мінімуму, за якими будують інтерполяційний поліном – квадратичну параболу, мінімум якої знаходиться у її вершині.

Алгоритм методу квадратичної інтерполяції:

1. Обчислити $f(x)$ у початковій точці x_0.

2. Вибрати крок h. Якщо $f(x_0 + h) > f(x_0)$, то $h := -h$.

3. Обчислити $x_{i+1} = x_i + h$ та $f(x_{i+1})$.

4. Якщо $f(x_{i+1}) \le f(x_i)$, то $h := 2h$ і перейти до етапу 3.

5. Якщо $f(x_{i+1}) > f(x_i)$, то $x_m := x_{i+1}$, $x_{m-1} := x_i$, $h := h/2$ і перейти до етапу 3 востаннє.

6. З чотирьох рівновіддалених значень x_{m+1}, x_m, x_{m-1}, x_{m-2} виключити x_{m+1}, або x_{m-2}, залежно від того, яка з цих точок знаходиться далі від точки x,

у якій $f(x)$ має найменше значення. Нехай x_a, x_b, x_c – три точки, що залишилися, де x_c – центральна, а $x_a = x_c - h$, $x_b = x_c + h$.

7. Провести квадратичну інтерполяцію для визначення координат точки

$$\chi \approx x_c + \frac{h(f(x_a) - f(x_b))}{2(f(x_a) - 2f(x_c) + f(x_b))}.$$

Приклад.

Знайти мінімум цільової функції $f(x) = (x-1)^2$ методом золотого перерізу.

Легко бачити, що точним розв'язком цієї задачі оптимізації є точка x=1.

Оберемо інтервал пошуку $x \in [0,3]$, тобто $a = 0$, $b = 3$. Обчислимо x_1 та x_2.

$$x_1 = a + (1-\delta)(b-a) = 0 + \left(1 - \frac{\sqrt{5}-1}{2}\right)(3-0) \simeq 1{,}14590;$$

$$x_2 = a + \delta(b-a) = 0 + \frac{\sqrt{5}-1}{2}(3-0) \simeq 1{,}85410.$$

Знайдемо значення цільової функції в точках x_1 та x_2.

$$f(x_1) = (x_1 - 1)^2 = (1{,}14590 - 1)^2 \simeq 0{,}02129,$$

$$f(x_2) = (x_2 - 1)^2 = (1{,}85410 - 1)^2 \simeq 0{,}729490.$$

Оскільки $f(x_1) < f(x_2)$, то $b = x_2 \simeq 1{,}85410$. Таким чином інтервал пошуку звузився до $x \in [0, 1{,}85410]$.

Для нового інтервалу значення x_2 та $f(x_2)$ обчислювати не треба. Їх можна взяти з попередньої ітерації:

$$x_2 = x_1 = 1{,}14590, \quad f(x_2) = 0{,}02129.$$

Обчислимо нове значення x_1 та $f(x_1)$:

$$x_1 = a + (1-\delta)(b-a) = 0 + \left(1 - \frac{\sqrt{5}-1}{2}\right)(1{,}85410 - 0) \simeq 0{,}708204;$$

$$f(x_1) = (x_1 - 1)^2 = (0{,}708204 - 1)^2 \simeq 0{,}08514.$$

Оскільки $f(x_1) > f(x_2)$, то $a = x_1 \simeq 0{,}708204$. Тобто інтервал пошуку звузився до $x \in [0{,}708204,\ 1{,}85410]$. Таким чином з кожною ітерацією інтервал пошуку звужується в околі точного розв'язку.

7.2. Методи багатовимірного пошуку

У методі покоординатного спуску [11, с. 270] в області D (рис. 7.6) задають початкове наближення $\mathbf{X}_0 = \left[x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\right]^{\mathrm{T}}$ і послідовно проводять одновимірні пошуки вздовж різних напрямів, паралельних координатним осям.

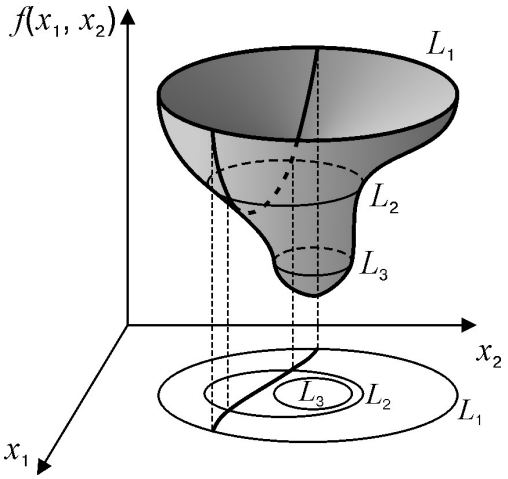


Рис. 7.6. Пошук мінімуму покоординатним спуском

Для цього розглядають функцію однієї змінної $f\left(x_1, x_2^{(0)}, \ldots, x_n^{(0)}\right)$ і методом одновимірного пошуку знаходять $\min\limits_{x_1 \in D} f\left(x_1, x_2^{(0)}, \ldots, x_n^{(0)}\right)$. Значення x_1, що відповідає мінімуму, позначають через $x_1^{(1)}$:

$$f\left(x_1^{(1)},\, x_2^{(0)},...,\, x_n^{(0)}\right) \le f\left(x_1^{(0)},\, x_2^{(0)},\ldots,\, x_n^{(0)}\right).$$

Далі розглядають функцію однієї змінної $f\left(x_1^{(1)},\, x_2,\ldots,\, x_n^{(0)}\right)$ і знаходять $\min\limits_{x_2 \in D} f\left(x_1^{(1)},\, x_2,\ldots,\, x_n^{(0)}\right)$. Значення x_2, що відповідає мінімуму функції, позначають через $x_2^{(1)}$. У цьому разі

$$f\left(x_1^{(1)},\, x_2^{(1)},\ldots,\, x_n^{(0)}\right) \le f\left(x_1^{(1)},\, x_2^{(0)},\ldots,\, x_n^{(0)}\right).$$

Після *n* кроків одержують

$$f\left(x_1^{(1)},\, x_2^{(1)},\ldots,\, x_n^{(1)}\right) \le f\left(x_1^{(0)},\, x_2^{(0)},\ldots,\, x_n^{(0)}\right).$$

У результаті *n* кроків покоординатного спуску відбувається перехід із точки $\mathbf{X}_0$ у точку $\mathbf{X}_1 = \left[x_1^{(1)},\, x_2^{(1)},\ldots,\, x_n^{(1)}\right]^{\mathrm{T}}$.

Якщо $f(\mathbf{X}_0) = f(\mathbf{X}_1)$, то $\mathbf{X}_0$ є точкою мінімуму $f(\mathbf{X})$. Якщо $f(\mathbf{X}_0) > f(\mathbf{X}_1)$, то виконують наступний крок покоординатного спуску, в якому за початкову точку взято $\mathbf{X}_1$. У результаті знаходять точку $\mathbf{X}_2$ і так далі. Цей процес продовжують доти, доки не виконають умову закінчення ітерацій, наприклад:

$$\frac{\left\|\mathbf{X}_{i+1} - \mathbf{X}_i\right\|}{\left\|\mathbf{X}_{i+1}\right\|} < \varepsilon,$$

де ε – задана відносна точність розв'язку.

Ідея методу випадкового пошуку полягає в тому, щоб перебором сукупностей випадкових послідовностей змінних знайти найменше значення цільової функції. Найпростішим є метод Монте-Карло (сліпий пошук). На $i + 1$ кроці в допустимій області вибирають випадкову точку $\mathbf{X}_{i+1}$, обчислюють значення $f(\mathbf{X}_{i+1})$ і порівнюють зі знайденим на

попередньому кроці. Якщо $f(\mathbf{X}_{i+1}) < f(\mathbf{X}_i)$, то запам'ятовується нове значення цільової функції і відповідної йому координати.

Теоретично за досить великої кількості вибірок можна досягти якої завгодно високої точності визначення мінімуму. Однак на практиці це потребує великих обчислювальних витрат і точність, як правило, недостатня. Тому в практично використовуваних методах випадкового пошуку обчислення виконують відповідно до загальної схеми оптимізації.

Координати нової точки знаходять за виразом

$$\mathbf{X}_{i+1} = \mathbf{X}_i + h\mathbf{A}_i,$$

де h – крок пошуку; $\mathbf{A}_i$ – випадковий напрямок.

Точка $\mathbf{X}_{i+1}$ вважається обчисленою, якщо $f\left(\mathbf{X}_{i+1}\right) < f\left(\mathbf{X}_i\right)$. Інакше роблять спробу досягти успіху або за рахунок зміни кроку, або за рахунок зміни напряму на протилежний. Якщо всі спроби виявляються невдалими, то вибирають новий випадковий напрям пошуку. Випадковий напрям пошуку знаходять за допомогою випадкової точки $\mathbf{A} = [a_1, a_2, \ldots, a_n]$, компоненти якої є випадковим значенням, рівномірно розподіленим на відрізку [–1, 1].

Приклад. Знайти мінімум цільової функції

$$f\left(\mathbf{X}\right) = f\left(x_1, x_2\right) = 4\left(x_1 - 2\right)^2 + \frac{\left(x_1 x_2 - 2\right)^2}{2} \qquad (7.2)$$

методом покоординатного спуску. Поверхня, що описується цільовою функцією (7.2), та області однакового рівня зображені на рис. 7.7.

Як видно, цільова функція (7.2) досягає мінімуму в точці $\mathbf{X}^* = \begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$.

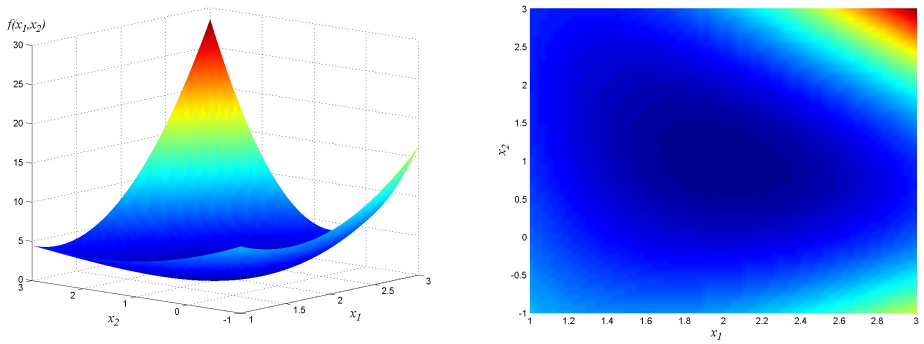


Рис. 7.7. Поверхня, що описується цільовою функцією, та області однакового рівня

Оберемо початкове наближення $\mathbf{X}_0 = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ та розглянемо функцію однієї змінної

$$g_1(x_1) = f\left(x_1, x_2^{(0)}\right) = 4(x_1 - 2)^2 + \frac{\left(x_1 x_2^{(0)} - 2\right)^2}{2} = 4(x_1 - 2)^2 + 2(x_1 - 1)^2.$$

Знайдемо мінімум функції $g_1(x_1)$ використовуючи один із методів одномірного пошуку, наприклад, метод золотого перерізу. Результатом є $x_1^{(1)} = \frac{5}{3}$. Тепер розглянемо функцію

$$g_2(x_2) = f\left(x_1^{(1)}, x_2\right) = 4\left(x_1^{(1)} - 2\right)^2 + \frac{\left(x_1^{(1)} x_2 - 2\right)^2}{2} = \frac{4}{9} + \frac{\left(\frac{5}{3} x_2 - 2\right)^2}{2}.$$

Мінімум цієї функції є $x_2^{(1)} = 1,2$. В результаті першої ітерації отримуємо наступне наближення: $\mathbf{X}_1 = \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \end{bmatrix} = \begin{bmatrix} \frac{5}{3} \\ 1,2 \end{bmatrix}$. Легко переконатись, що $f(\mathbf{X}_1) < f(\mathbf{X}_0)$, а також в тому, що точка $\mathbf{X}_1$ ближче розташована до $\mathbf{X}^*$ ніж $\mathbf{X}_0$.

Проведемо ще одну ітерацію.

$$g_3(x_1) = f\left(x_1, x_2^{(1)}\right) = 4(x_1 - 2)^2 + \frac{\left(x_1 x_2^{(1)} - 2\right)^2}{2} = 4(x_1 - 2)^2 + \frac{(1{,}2x_1 - 2)^2}{2};$$

$$\min g_3(x_1) \quad \Rightarrow x_1^{(2)} \simeq 1{,}82207.$$

$$g_4(x_2) = f\left(x_1^{(2)}, x_2\right) = 4\left(x_1^{(2)} - 2\right)^2 + \frac{\left(x_1^{(2)} x_2 - 2\right)^2}{2} \simeq 0{,}126636 + \frac{\left(1{,}82207 x_2 - 2\right)^2}{2}$$

$$\min g_4(x_2) \quad \Rightarrow x_2^{(2)} \simeq 1{,}09766.$$

Тоді $\mathbf{X}_2 = \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \end{bmatrix} = \begin{bmatrix} 1{,}82207 \\ 1{,}09766 \end{bmatrix}$.

Як видно, точка $\mathbf{X}_2$ ще більш наблизилась до $\mathbf{X}^*$.

7.3. Градієнтні методи

Відомо, що градієнт ортогональний до поверхні рівня цільової функції в точці його визначення і збігається з напрямом найшвидшого зростання цільової функції. Тому напрям, протилежний градієнту, вказує напрямок зменшення цільової функції, тобто напрямок, де може бути розташований мінімум функції.

Усі градієнтні методи використовують вказані особливості поведінки градієнта, а стратегія їх пошуку будується на рекурентному виразі типу:

$$\mathbf{X}_{i+1} = \mathbf{X}_i + h\mathbf{S}_i,$$

де h – величина кроку, а $\mathbf{S}_i$ – одиничний вектор напряму пошуку на i-му кроці.

Спосіб вибору кроку та напряму пошуку визначає суть методу.

У разі пошуку мінімуму цільової функції потрібно рухатися в напрямі, протилежному градієнту функції $f(\mathbf{X})$. Тому напрям

$$\mathbf{S}_i = -\frac{\nabla f(\mathbf{X}_i)}{\left\|\nabla f(\mathbf{X}_i)\right\|_2}$$

дозволяє одержати таку рекурентну формулу для методу градієнтного спуску:

$$\mathbf{X}_{i+1} = \mathbf{X}_i - h\frac{\nabla f(\mathbf{X}_i)}{\left\|\nabla f(\mathbf{X}_i)\right\|_2}. \qquad (7.3)$$

Особливістю методу найшвидшого спуску є рух з оптимальним кроком, розрахованим за допомогою одновимірної мінімізації цільової функції по h уздовж антиградієнтного напряму. Дійсно, якщо в точці $\mathbf{X}_i$ напрям пошуку визначений, то значення цільової функції в наступній точці $\mathbf{X}_{i+1}$ є функцією кроку спуску:

$$f(\mathbf{X}_{i+1}) = f\left(\mathbf{X}_i - h\frac{\nabla f(\mathbf{X}_i)}{\left\|\nabla f(\mathbf{X}_i)\right\|_2}\right) = g(h).$$

Тому крок h можна вибрати так, щоб $f(\mathbf{X}_{i+1})$ максимально зменшила своє значення:

$$f(\mathbf{X}_{i+1}) = \min_h f\left(\mathbf{X}_i - h\frac{\nabla f(\mathbf{X}_i)}{\left\|\nabla f(\mathbf{X}_i)\right\|_2}\right). \qquad (7.4)$$

Вибір оптимального кроку зводиться до розв’язання одновимірної оптимізації функції $g(h)$.

Алгоритм методу найшвидшого градієнтного спуску:

1. Обчислити всі частинні похідні цільової функції.

2. Знайти одним з методів одновимірного пошуку оптимальний крок уздовж антиградієнтного напряму. Крок h визначають з умови мінімуму функції $f\left(\mathbf{X}_i - h\dfrac{\nabla f(\mathbf{X}_i)}{\left\|\nabla f(\mathbf{X}_i)\right\|_2}\right)$ по h.

3. Обчислити координати нової точки $\mathbf{X}_{i+1}$ за формулою (7.3).

4. Якщо умова припинення пошуку не виконується, то перейти до кроку 1.

Якщо крок h був знайдений з вимоги (7.4), то виконана необхідна умова існування мінімуму: $\dfrac{d}{dh}f\left(\mathbf{X}_i - h\dfrac{\nabla f(\mathbf{X}_i)}{\left\|\nabla f(\mathbf{X}_i)\right\|_2}\right)=0$. Враховуючи, що

$$\frac{df\left(\mathbf{X}_i - h\dfrac{\nabla f(\mathbf{X}_i)}{\left\|\nabla f(\mathbf{X}_i)\right\|_2}\right)}{dh} = -\frac{\nabla f(\mathbf{X}_i)^{\mathrm{T}}}{\left\|\nabla f(\mathbf{X}_i)\right\|_2}\nabla f\left(\mathbf{X}_i - h\frac{\nabla f(\mathbf{X}_i)}{\left\|\nabla f(\mathbf{X}_i)\right\|_2}\right) = -\frac{\nabla f(\mathbf{X}_i)^{\mathrm{T}}}{\left\|\nabla f(\mathbf{X}_i)\right\|_2}\nabla f(\mathbf{X}_{i+1}),$$

отримаємо:

$$\nabla f(\mathbf{X}_i)^{\mathrm{T}}\nabla f(\mathbf{X}_{i+1})=0, \qquad (7.5)$$

тобто градієнти $\nabla f(\mathbf{X}_i)$ и $\nabla f(\mathbf{X}_{i+1})$ є ортогональними.

Траєкторію пошуку цим методом показано на рис. 7.8, з якого видно, що рух уздовж одного напряму припиняється, коли лінія напряму пошуку стає дотичною до якої-небудь лінії рівня. Кожен новий напрям руху до екстремуму ортогональний попередньому.

Розглянуті методи збігаються до локального мінімуму зі швидкістю геометричної прогресії, тобто лінійно.

Приклад. Знайти мінімум цільової функції (7.2) методом найшвидшого градієнтного спуску.

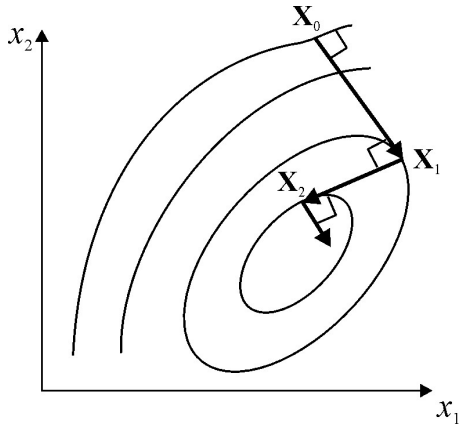


Рис. 7.8. Метод найшвидшого градієнтного спуску

Знайдемо градієнт цільової функції

$$\nabla f(\mathbf{X}) = \begin{bmatrix} \dfrac{df(\mathbf{X})}{dx_1} \\ \dfrac{df(\mathbf{X})}{dx_2} \end{bmatrix} = \begin{bmatrix} 8(x_1 - 2) + x_2(x_1 x_2 - 2) \\ x_1(x_1 x_2 - 2) \end{bmatrix}. \qquad (7.6)$$

Оберемо початкове наближення $\mathbf{X}_0 = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

Тоді

$$\nabla f(\mathbf{X}_0) = \begin{bmatrix} 8\left(x_1^{(0)} - 2\right) + x_2^{(0)}\left(x_1^{(0)} x_2^{(0)} - 2\right) \\ x_1^{(0)}\left(x_1^{(0)} x_2^{(0)} - 2\right) \end{bmatrix} = \begin{bmatrix} -8 \\ 0 \end{bmatrix};$$

$$\left\|\nabla f(\mathbf{X}_0)\right\|_2 = \sqrt{(-8)^2 + 0^2} = 8;$$

$$g(h) = f\left(\mathbf{X}_0 - h\frac{\nabla f(\mathbf{X}_0)}{\left\|\nabla f(\mathbf{X}_0)\right\|_2}\right) = f\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix} - \frac{h}{8}\begin{bmatrix} -8 \\ 0 \end{bmatrix}\right) = f(1+h, 2) =$$

$$= 4(1 + h - 2)^2 + \frac{\left((1+h)2 - 2\right)^2}{2} = 4(h-1)^2 + 2h^2.$$

Знайдемо мінімум функції $g(h)$ використовуючи один із методів одномірного пошуку, наприклад, метод золотого перерізу. Результатом є $h=\frac{2}{3}$. Тоді

$$\mathbf{X}_1=\begin{bmatrix}x_1^{(1)}\\x_2^{(1)}\end{bmatrix}=\mathbf{X}_0-h\frac{\nabla f(\mathbf{X}_0)}{\left\|\nabla f(\mathbf{X}_0)\right\|_2}=\begin{bmatrix}1\\2\end{bmatrix}-\frac{2}{3}\cdot\frac{1}{8}\begin{bmatrix}-8\\0\end{bmatrix}=\begin{bmatrix}\frac{5}{3}\\2\end{bmatrix}.$$

Виконаємо ще одну ітерацію.

$$\nabla f(\mathbf{X}_1)=\begin{bmatrix}8\left(x_1^{(1)}-2\right)+x_2^{(1)}\left(x_1^{(1)}x_2^{(1)}-2\right)\\x_1^{(1)}\left(x_1^{(1)}x_2^{(1)}-2\right)\end{bmatrix}=\begin{bmatrix}0\\\frac{20}{9}\end{bmatrix};$$

$$\left\|\nabla f(\mathbf{X}_1)\right\|_2=\sqrt{0^2+\left(\frac{20}{9}\right)^2}=\frac{20}{9};$$

$$g(h)=f\left(\mathbf{X}_1-h\frac{\nabla f(\mathbf{X}_1)}{\left\|\nabla f(\mathbf{X}_1)\right\|_2}\right)=f\left(\begin{bmatrix}\frac{5}{3}\\2\end{bmatrix}-\frac{h}{\frac{20}{9}}\begin{bmatrix}0\\\frac{20}{9}\end{bmatrix}\right)=f\left(\frac{5}{3},2-h\right)=$$

$$=4\left(\frac{5}{3}-2\right)^2+\frac{\left(\frac{5}{3}(2-h)-2\right)^2}{2}=\frac{4}{9}+\frac{1}{6}(4-5h)^2;$$

$\min g(h)\ \Rightarrow\ h=0{,}8$.

Таким чином

$$\mathbf{X}_2=\begin{bmatrix}x_1^{(2)}\\x_2^{(2)}\end{bmatrix}=\mathbf{X}_1-h\frac{\nabla f(\mathbf{X}_1)}{\left\|\nabla f(\mathbf{X}_1)\right\|_2}=\begin{bmatrix}\frac{5}{3}\\2\end{bmatrix}-0{,}8\cdot\frac{9}{20}\begin{bmatrix}0\\\frac{20}{9}\end{bmatrix}=\begin{bmatrix}\frac{5}{3}\\1{,}2\end{bmatrix}.$$

Як видно, кожна нова ітерація все більше наближає результат до точного розв’язку.

7.4. Метод спряжених градієнтів

Вектори **A** і **B** називають спряженими відносно матриці **Q** (чи **Q**-спряженими), якщо скалярний добуток векторів **A** і **QB** дорівнює нулю, тобто $\mathbf{A}^{\mathrm{T}}\mathbf{QB} = 0$. Спряженість векторів є узагальненням поняття ортогональності, оскільки **Q**-спряженість векторів означає їх ортогональність у випадку, коли **Q** = **E**.

Нехай на i-му ітераційному кроці пошук починається з точки $\mathbf{X}_i$ в напрямі $\mathbf{S}_i$. Тоді нова точка знаходиться за формулою

$$\mathbf{X}_{i+1} = \mathbf{X}_i + h\mathbf{S}_i, \tag{7.7}$$

де значення кроку h слід визначати з умови мінімуму функції $f(\mathbf{X}_i + h\mathbf{S}_i)$ по h. Для цього розкладемо функцію $f(\mathbf{X})$ у ряд Тейлора в околі точки $\mathbf{X} = \mathbf{X}_i$ і обмежимося розглядом тільки трьох членів, що є точним для квадратичної цільової функції

$$f(\mathbf{X}) \approx f(\mathbf{X}_i) + \nabla f(\mathbf{X}_i)^{\mathrm{T}}(\mathbf{X} - \mathbf{X}_i) + \frac{1}{2}(\mathbf{X} - \mathbf{X}_i)^{\mathrm{T}}\mathbf{H}(\mathbf{X}_i)(\mathbf{X} - \mathbf{X}_i),$$

де $\mathbf{H}(\mathbf{X}_i) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{X}_i)}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{X}_i)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{X}_i)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{X}_i)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\mathbf{X}_i)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{X}_i)}{\partial x_2 \partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f(\mathbf{X}_i)}{\partial x_1 \partial x_n} & \frac{\partial^2 f(\mathbf{X}_i)}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f(\mathbf{X}_i)}{\partial x_n^2} \end{bmatrix}$ – матриця Гессе.

Тоді

$$f(\mathbf{X}_i + h\mathbf{S}_i) = f(\mathbf{X}_i) + \nabla f(\mathbf{X}_i)^{\mathrm{T}} h\mathbf{S}_i + \frac{1}{2} h\mathbf{S}_i^{\mathrm{T}}\mathbf{H}(\mathbf{X}_i) h\mathbf{S}_i.$$

З умови мінімуму останньої функції по h маємо

$$\frac{df\left(\mathbf{X}_i + h\mathbf{S}_i\right)}{dh} = \nabla f\left(\mathbf{X}_i\right)^{\mathrm{T}} \mathbf{S}_i + h\mathbf{S}_i^{\mathrm{T}} \mathbf{H}\left(\mathbf{X}_i\right)\mathbf{S}_i = 0.$$

Звідси

$$h = -\frac{\nabla f\left(\mathbf{X}_i\right)^{\mathrm{T}} \mathbf{S}_i}{\mathbf{S}_i^{\mathrm{T}} \mathbf{H}\left(\mathbf{X}_i\right)\mathbf{S}_i}. \qquad (7.8)$$

Можна показати, що після того як за (7.7) і (7.8) обчислено точку $\mathbf{X}_{i+1}$, для продовження пошуку потрібно обрати новий напрям $\mathbf{S}_{i+1}$ так, щоб

$$\mathbf{S}_i^{\mathrm{T}} \mathbf{H}\left(\mathbf{X}_i\right)\mathbf{S}_{i+1} = 0, \qquad (7.9)$$

тобто новий напрям $\mathbf{S}_{i+1}$ має бути спряженим до старого напряму $\mathbf{S}_i$.

Збіжність такого методу строго обґрунтована для квадратичних функцій вигляду

$$f(\mathbf{X}) = \mathbf{B} + \mathbf{C}^{\mathrm{T}}\mathbf{X} + \frac{1}{2}\mathbf{X}^{\mathrm{T}}\mathbf{H}\mathbf{X},$$

де **B** та **C** – вектори-стовпчики, а **H** – матриця постійних коефіцієнтів.

Мінімізацію такої функції за додатньо-визначеною матрицею **H** можна виконати за *n* (чи менше) кроків. У загальному випадку цільові функції не є квадратичними. Тому з використанням такого методу розв'язок буде знайдено за більшу кількість кроків, ніж *n*. Недолік такого методу – трудомісткий розрахунок матриці Гессе. Цей недолік усунутий у модифікованому варіанті, названому методом спряжених градієнтів.

У методі спряжених градієнтів після використання методу найшвидшого градієнтного спуску будується послідовність напрямів пошуку, спряжених до попередніх. Ці напрями є лінійними комбінаціями антиградієнта і попередніх напрямів пошуку. Так, якщо $\mathbf{S}_i = -\nabla f\left(\mathbf{X}_i\right)$, то потрібно знайти новий напрям

$$\mathbf{S}_{i+1} = -\nabla f\left(\mathbf{X}_{i+1}\right) + \beta_{i+1}\mathbf{S}_i, \qquad (7.10)$$

підібравши коефіцієнт β_{i+1} таким чином, щоб він задовольняв умові (7.9) спряженості векторів $\mathbf{S}_{i+1}$ і $\mathbf{S}_i$.

Розкладемо функцію $\nabla f\left(\mathbf{X}\right)$ у ряд Тейлора в околі точки $\mathbf{X} = \mathbf{X}_i$ і обмежимося розглядом тільки двох членів, що є точним для квадратичної цільової функції:

$$\nabla f\left(\mathbf{X}\right) \approx \nabla f\left(\mathbf{X}_i\right) + \mathbf{H}\left(\mathbf{X}_i\right)\left(\mathbf{X} - \mathbf{X}_i\right). \qquad (7.11)$$

Підставляючи у (7.11) $\mathbf{X} = \mathbf{X}_{i+1}$ отримуємо:

$$\nabla f\left(\mathbf{X}_{i+1}\right) - \nabla f\left(\mathbf{X}_i\right) = \mathbf{H}\left(\mathbf{X}_i\right)\left(\mathbf{X}_{i+1} - \mathbf{X}_i\right) = h\mathbf{H}\left(\mathbf{X}_i\right)\mathbf{S}_i. \qquad (7.12)$$

Транспонуючи (7.12) та домножуючи праву та ліву частини цієї рівності на $\mathbf{H}^{-1}\left(\mathbf{X}_i\right)$ маємо:

$$\left(\nabla f\left(\mathbf{X}_{i+1}\right) - \nabla f\left(\mathbf{X}_i\right)\right)^{\mathrm{T}} \mathbf{H}^{-1}\left(\mathbf{X}_i\right) = h\mathbf{S}_i^{\mathrm{T}}. \qquad (7.13)$$

Виразимо з (7.13) $\mathbf{S}_i^{\mathrm{T}}$ та підставимо це значення та вираз (7.10) в (7.9)

$$\left(\nabla f\left(\mathbf{X}_{i+1}\right) - \nabla f\left(\mathbf{X}_i\right)\right)^{\mathrm{T}} \left(-\nabla f\left(\mathbf{X}_{i+1}\right) - \beta_{i+1}\nabla f\left(\mathbf{X}_i\right)\right) = 0. \qquad (7.14)$$

Враховуючи (7.5), отримуємо:

$$\beta_{i+1} = \frac{\nabla f\left(\mathbf{X}_{i+1}\right)^{\mathrm{T}} \nabla f\left(\mathbf{X}_{i+1}\right)}{\nabla f\left(\mathbf{X}_i\right)^{\mathrm{T}} \nabla f\left(\mathbf{X}_i\right)}.$$

Алгоритм методу спряжених градієнтів складається з таких кроків:

1. Обчислити вектор $\mathbf{S}_i = -\nabla f\left(\mathbf{X}_i\right)$.

2. Знайти мінімум $f(\mathbf{X})$ одним з методів одновимірного пошуку в напрямі $\mathbf{S}_i$. Звідси знаходимо $\mathbf{X}_{i+1}$, $\nabla f\left(\mathbf{X}_{i+1}\right)$.

3. Визначити новий спряжений напрям $\mathbf{S}_{i+1}$ із співвідношення

$$\mathbf{S}_{i+1} = -\nabla f(\mathbf{X}_{i+1}) + \mathbf{S}_i \frac{\nabla f(\mathbf{X}_{i+1})^{\mathrm{T}} \nabla f(\mathbf{X}_{i+1})}{\nabla f(\mathbf{X}_i)^{\mathrm{T}} \nabla f(\mathbf{X}_i)} \qquad (7.15)$$

та знайти мінімум $f(\mathbf{X})$ одним з методів одновимірного пошуку в напрямі $\mathbf{S}_{i+1}$. Звідси знаходимо $\mathbf{X}_{i+2}$ та $\nabla f\left(\mathbf{X}_{i+2}\right)$. Повторюємо обчислення, аналогічні (7.15) для $\mathbf{S}_{i+2}\ldots\ \mathbf{S}_{i+\mathrm{n}}$ та знаходимо $\mathbf{X}_{i+3}\ldots\ \mathbf{X}_{i+n+1}$.

4. Якщо умова припинення пошуку не виконується, то перейти до кроку 1.

Геометричний зміст методу спряжених градієнтів та порівняння траєкторії пошуку цим методом з методом найшвидшого градієнтного спуску показано на рис.7.9.

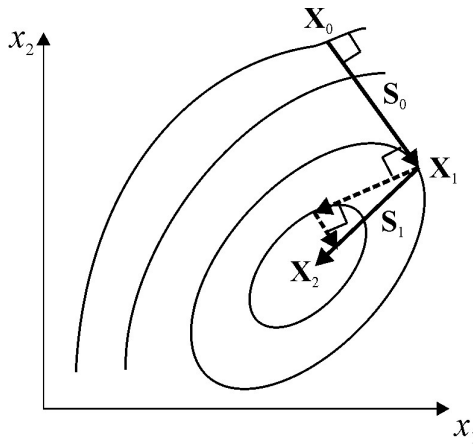


Рис. 7.9. Метод спряжених градієнтів. Для порівняння пунктиром показано траєкторію методу найшвидшого градієнтного спуску

Розглянутий метод називають також методом Флетчера-Ривса. Існують і інші модифікації методу спряжених градієнтів. Так, в методі Полака-Райбера коефіцієнт β_{i+1} розраховують за формулою:

$$\beta_{i+1} = \frac{\nabla f\left(\mathbf{X}_{i+1}\right)^{\mathrm{T}} \left(\nabla f\left(\mathbf{X}_{i+1}\right) - \nabla f\left(\mathbf{X}_i\right)\right)}{\nabla f\left(\mathbf{X}_i\right)^{\mathrm{T}} \nabla f\left(\mathbf{X}_i\right)}.$$

Якщо $\beta_{i+1} < 0$, то приймають це значення за нуль, тобто використовують метод найшвидшого градієнтного спуску.

В методі Хестенса-Штифеля коефіцієнт β_{i+1} розраховують за формулою:

$$\beta_{i+1} = \frac{\nabla f\left(\mathbf{X}_{i+1}\right)^{\mathrm{T}}\left(\nabla f\left(\mathbf{X}_{i+1}\right) - \nabla f\left(\mathbf{X}_{i}\right)\right)}{\nabla f\left(\mathbf{X}_{i}\right)^{\mathrm{T}}\left(\nabla f\left(\mathbf{X}_{i+1}\right) - \nabla f\left(\mathbf{X}_{i}\right)\right)}.$$

Методи спряжених градієнтів мають *n*-крокову квадратичну швидкість збіжності:

$$\left\|\mathbf{X}_{i+n} - \mathbf{X}^{*}\right\| \le c\left\|\mathbf{X}_{i} - \mathbf{X}^{*}\right\|^{2}.$$

Приклад.

Знайти мінімум цільової функції (7.2) методом спряжених градієнтів.

Градієнт цільової функції (7.2) визначається виразом (7.6).

Оберемо початкове наближення $\mathbf{X}_0 = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

Тоді

$$\mathbf{S}_0 = -\nabla f\left(\mathbf{X}_0\right) = -\begin{bmatrix} 8\left(x_1^{(0)} - 2\right) + x_2^{(0)}\left(x_1^{(0)}x_2^{(0)} - 2\right) \\ x_1^{(0)}\left(x_1^{(0)}x_2^{(0)} - 2\right) \end{bmatrix} = \begin{bmatrix} 8 \\ 0 \end{bmatrix}.$$

Знайдемо мінімум $f(\mathbf{X})$ в напрямі $\mathbf{S}_0$, виходячи з обраного початкового наближення. Тоді цільова функція буде мати вигляд

$$g(h) = f\left(\mathbf{X}_0 + h\mathbf{S}_0\right) = f\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix} + h\begin{bmatrix} 8 \\ 0 \end{bmatrix}\right) = f\left(1 + 8h, 2\right) =$$

$$= 4\left(1 + 8h - 2\right)^2 + \frac{\left(\left(1 + 8h\right)2 - 2\right)^2}{2} = 4\left(8h - 1\right)^2 + 128h^2.$$

Знайдемо мінімум функції $g(h)$ використовуючи один із методів одномірного пошуку, наприклад, метод золотого перерізу. Результатом є $h = \frac{1}{12}$. Тоді

$$\mathbf{X}_1 = \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \end{bmatrix} = \mathbf{X}_0 + h\mathbf{S}_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \frac{1}{12}\begin{bmatrix} 8 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{5}{3} \\ 2 \end{bmatrix}.$$

Обчислимо $\nabla f(\mathbf{X}_1)$.

$$\nabla f(\mathbf{X}_1) = \begin{bmatrix} 8\left(x_1^{(1)} - 2\right) + x_2^{(1)}\left(x_1^{(1)}x_2^{(1)} - 2\right) \\ x_1^{(1)}\left(x_1^{(1)}x_2^{(1)} - 2\right) \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{20}{9} \end{bmatrix}.$$

Знайдемо перший спряжений напрям $\mathbf{S}_1$:

$$\mathbf{S}_1 = -\nabla f(\mathbf{X}_1) + \mathbf{S}_0 \frac{\nabla f(\mathbf{X}_1)^{\mathrm{T}} \nabla f(\mathbf{X}_1)}{\nabla f(\mathbf{X}_0)^{\mathrm{T}} \nabla f(\mathbf{X}_0)} = -\begin{bmatrix} 0 \\ \frac{20}{9} \end{bmatrix} + \begin{bmatrix} 8 \\ 0 \end{bmatrix} \frac{0^2 + \left(\frac{20}{9}\right)^2}{(-8)^2 + 0^2} = \begin{bmatrix} \frac{400}{648} \\ -\frac{20}{9} \end{bmatrix}.$$

Знайдемо мінімум $f(\mathbf{X})$ в напрямі $\mathbf{S}_1$, виходячи з точки $\mathbf{X}_1$. Цільова функція для такої задачі буде мати вигляд

$$g(h) = f(\mathbf{X}_1 + h\mathbf{S}_1) = f\left(\begin{bmatrix} \frac{5}{3} \\ 2 \end{bmatrix} + h\begin{bmatrix} \frac{400}{648} \\ -\frac{20}{9} \end{bmatrix}\right) = f\left(\frac{5}{3} + \frac{400}{648}h, 2 - \frac{20}{9}h\right) =$$

$$= 4\left(\frac{5}{3} + \frac{400}{648}h - 2\right)^2 + \frac{\left(\left(\frac{5}{3} + \frac{400}{648}h\right)\left(2 - \frac{20}{9}h\right) - 2\right)^2}{2}.$$

Мінімумом цієї цільової функції є $h \simeq 0{,}454297$.

Таким чином,

$$\mathbf{X}_2 = \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \end{bmatrix} = \mathbf{X}_1 + h\mathbf{S}_1 \simeq \begin{bmatrix} \frac{5}{3} \\ 2 \end{bmatrix} + 0{,}454297 \begin{bmatrix} \frac{400}{648} \\ -\frac{20}{9} \end{bmatrix} \simeq \begin{bmatrix} 1{,}94795 \\ 0{,}991003 \end{bmatrix}.$$

Обчислимо $\nabla f\left(\mathbf{X}_2\right)$.

$$\nabla f\left(\mathbf{X}_2\right) = \begin{bmatrix} 8\left(x_1^{(2)} - 2\right) + x_2^{(2)}\left(x_1^{(2)} x_2^{(2)} - 2\right) \\ x_1^{(2)}\left(x_1^{(2)} x_2^{(2)} - 2\right) \end{bmatrix} \simeq \begin{bmatrix} -0{,}485348 \\ -0{,}135528 \end{bmatrix}.$$

Знайдемо другий спряжений напрям $\mathbf{S}_2$:

$$\mathbf{S}_2 = -\nabla f\left(\mathbf{X}_2\right) + \mathbf{S}_1 \frac{\nabla f\left(\mathbf{X}_2\right)^{\mathrm{T}} \nabla f\left(\mathbf{X}_2\right)}{\nabla f\left(\mathbf{X}_1\right)^{\mathrm{T}} \nabla f\left(\mathbf{X}_1\right)} \simeq$$

$$\simeq -\begin{bmatrix} -0{,}485348 \\ -0{,}135528 \end{bmatrix} + \begin{bmatrix} \frac{400}{648} \\ -\frac{20}{9} \end{bmatrix} \frac{\left(-0{,}485348\right)^2 + \left(-0{,}135528\right)^2}{0^2 + \left(\frac{20}{9}\right)^2} \simeq \begin{bmatrix} 0{,}514946 \\ 0{,}029456 \end{bmatrix}.$$

Цільова функція для пошуку мінімуму $f(\mathbf{X})$ в напрямі $\mathbf{S}_2$, виходячи з точки $\mathbf{X}_2$ буде мати вигляд

$$g(h) = f\left(\mathbf{X}_2 + h\mathbf{S}_2\right) \simeq f\left(\begin{bmatrix} 1{,}94795 \\ 0{,}991003 \end{bmatrix} + h\begin{bmatrix} 0{,}514946 \\ 0{,}029456 \end{bmatrix}\right) \simeq$$

$$f\left(1{,}94795 + 0{,}514946h,\ \ 0{,}991003 + 0{,}029456h\right) =$$

$$= 4\left(1{,}94795 + 0{,}514946h - 2\right)^2 +$$

$$+ \frac{\left(\left(1{,}94795 + 0{,}514946h\right)\left(0{,}991003 + 0{,}029456h\right) - 2\right)^2}{2}.$$

Мінімумом цієї цільової функції є $h \simeq 0{,}103906$.

Таким чином,

$$\mathbf{X}_3 = \begin{bmatrix} x_1^{(3)} \\ x_2^{(3)} \end{bmatrix} = \mathbf{X}_2 + h\mathbf{S}_2 \simeq \begin{bmatrix} 1{,}94795 \\ 0{,}991003 \end{bmatrix} + 0{,}103906 \begin{bmatrix} 0{,}514946 \\ 0{,}029456 \end{bmatrix} \simeq \begin{bmatrix} 2{,}00146 \\ 0{,}994064 \end{bmatrix}.$$

На наступній ітерації послідовно шукають мінімуми в новому напрямку пошуку $\mathbf{S}_3 = -\nabla f(\mathbf{X}_3)$ та спряжених напрямках $\mathbf{S}_4$ та $\mathbf{S}_5$, які знаходять аналогічно напрямкам $\mathbf{S}_1$ та $\mathbf{S}_2$.

7.5. Метод Ньютона

Нехай $\mathbf{X}_i$ – i-те наближення до точки, що відповідає мінімуму функції $f(\mathbf{X})$. Розкладемо в околі цієї точки функцію $f(\mathbf{X})$ у ряд Тейлора й обмежимося розглядом тільки трьох членів. Тоді

$$f(\mathbf{X}) = f(\mathbf{X}_i) + \nabla f(\mathbf{X}_i)^{\mathrm{T}} (\mathbf{X} - \mathbf{X}_i) + \frac{1}{2}(\mathbf{X} - \mathbf{X}_i)^{\mathrm{T}} \mathbf{H}(\mathbf{X}_i)(\mathbf{X} - \mathbf{X}_i), \quad (7.16)$$

де $\mathbf{H}(\mathbf{X}_i)$ – матриця Гессе в точці $\mathbf{X} = \mathbf{X}_i$.

Нове наближення $\mathbf{X}_{i+1}$ знаходимо з необхідної умови існування мінімуму функції (7.16):

$$\nabla f(\mathbf{X}_{i+1}) = \nabla f(\mathbf{X}_i) + \mathbf{H}(\mathbf{X}_i)(\mathbf{X}_{i+1} - \mathbf{X}_i) = 0,$$

звідки

$$\mathbf{X}_{i+1} = \mathbf{X}_i - \mathbf{H}^{-1}(\mathbf{X}_i)\nabla f(\mathbf{X}_i). \quad (7.17)$$

Формулу (7.17) можна також отримати застосуванням методу Ньютона до розв'язання системи нелінійних рівнянь $\nabla f(\mathbf{X}) = 0$. Тому такий метод називають методом Ньютона розв'язання задачі безумовної мінімізації.

У практичному використанні формули (7.17) не вдаються до безпосереднього обертання матриці Гессе, а на кожному ітераційному кроці задачу розв'язують у два етапи. Спочатку розв'язують СЛАР відносно $\Delta\mathbf{X}_i$:

$$\mathbf{H}(\mathbf{X}_i)\Delta\mathbf{X}_i = -\nabla f(\mathbf{X}_i), \quad (7.18)$$

потім знаходять нове наближення

$$\mathbf{X}_{i+1} = \mathbf{X}_i + \Delta\mathbf{X}_i. \quad (7.19)$$

Якщо початкове наближення $\mathbf{X}_0$ задано досить близько до точки $\mathbf{X}^*$, то на основі виразів (7.18) і (7.19) можна отримати квадратичну швидкість збіжності, тобто

$$\left\|\mathbf{X}_{i+1} - \mathbf{X}^*\right\| \le c\left\|\mathbf{X}_i - \mathbf{X}^*\right\|^2,$$

де c – постійна величина.

Розв'язуючи СЛАР (7.18), слід мати на увазі дві важливі властивості матриці Гессе. По-перше, матриця Гессе є симетричною. По-друге, з (7.16) випливає, що якщо $\mathbf{X}_i$ є точкою мінімуму, то $\nabla f\left(\mathbf{X}_i\right) = 0$ й отже, для того, щоб функція $f(\mathbf{X})$ зростала у разі відхилення від точки $\mathbf{X}_i$, необхідно, щоб $\left(\mathbf{X} - \mathbf{X}_i\right)^{\mathrm{T}} \mathbf{H}\left(\mathbf{X}_i\right)\left(\mathbf{X} - \mathbf{X}_i\right) > 0.$ Отже, поблизу мінімуму матриця Гессе додатньо-визначена. Тому для розв'язання СЛАР (7.18) застосовують метод Холеського.

Метод Ньютона безумовної мінімізації має ті ж недоліки, що і метод Ньютона розв'язання нелінійних рівнянь: метод не має глобальної збіжності і потребує аналітично заданої першої і другої похідних цільової функції. Однак метод Ньютона безумовної оптимізації має додатковий недолік. Навіть, будучи локальним, він не завжди приводить до точки мінімуму, оскільки в ньому немає нічого, щоб утримувало б його від просування вбік максимуму чи сідлової точки функції $f(\mathbf{X})$, де $\nabla f(\mathbf{X})$ теж дорівнює нулю. Останнє можливо, якщо матриця Гессе перестане бути додатньо-визначеною. Тому у використовуваних на практиці методах Ньютона, якщо матриця Гессе не є додатньо-визначеною (з'являються від'ємні власні числа), її замінюють матрицею

$$\tilde{\mathbf{H}} = \mathbf{H}\left(\mathbf{X}_i\right) + \mu\mathbf{E},$$

у якій μ вибирають так, щоб матриця $\tilde{\mathbf{H}}$ була додатньо-визначеною і добре обумовленою. Для цього найменше з можливих μ є дещо більшим, ніж модуль найменшого від'ємного власного числа матриці Гессе $\mathbf{H}(\mathbf{X}_i)$.

З метою збільшення області збіжності часто застосовують модифікований метод Ньютона, у якому нове наближення $\mathbf{X}_{i+1}$ визначають за формулою

$$\mathbf{X}_{i+1} = \mathbf{X}_i - \alpha \tilde{\mathbf{H}}^{-1} \nabla f\left(\mathbf{X}_i\right),$$

де $0 < \alpha \le 1$. Величину α вибирають таку, щоб зменшити цільову функцію, тобто виконати умову

$$f(\mathbf{X}_{i+1}) < f(\mathbf{X}_i).$$

Якщо аналітичні вирази для першої і другої похідних цільової функції не можуть бути задані, то для їх визначення використовують скінченно-різницеву апроксимацію.

Приклад. Знайти мінімум цільової функції (7.2) методом Ньютона.

Градієнт цільової функції (7.2) визначається виразом (7.6). Знайдемо матрицю Гессе для заданої цільової функції.

$$\mathbf{H}\left(\mathbf{X}\right) = \begin{bmatrix} 8 + x_2^2 & 2x_1x_2 - 2 \\ 2x_1x_2 - 2 & x_1^2 \end{bmatrix}.$$

Оберемо початкове наближення $\mathbf{X}_0 = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

Тоді

$$\nabla f\left(\mathbf{X}_0\right) = \begin{bmatrix} 8\left(x_1^{(0)} - 2\right) + x_2^{(0)}\left(x_1^{(0)}x_2^{(0)} - 2\right) \\ x_1^{(0)}\left(x_1^{(0)}x_2^{(0)} - 2\right) \end{bmatrix} = \begin{bmatrix} -8 \\ 0 \end{bmatrix};$$

$$\mathbf{H}(\mathbf{X}_0)=\begin{bmatrix} 8+\left(x_2^{(0)}\right)^2 & 2x_1^{(0)}x_2^{(0)}-2 \\ 2x_1^{(0)}x_2^{(0)}-2 & \left(x_1^{(0)}\right)^2 \end{bmatrix}=\begin{bmatrix} 12 & 2 \\ 2 & 1 \end{bmatrix}.$$

Таким чином СЛАР (7.18) на першому ітераційному кроці прийме вигляд

$$\begin{bmatrix} 12 & 2 \\ 2 & 1 \end{bmatrix}\Delta\mathbf{X}_0=-\begin{bmatrix} -8 \\ 0 \end{bmatrix}.$$

Розв'язавши отриману СЛАР маємо $\Delta\mathbf{X}_0=\begin{bmatrix} 1 \\ -2 \end{bmatrix}$. Згідно з (7.19) маємо

$$\mathbf{X}_1=\begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \end{bmatrix}=\mathbf{X}_0+\Delta\mathbf{X}_0=\begin{bmatrix} 1 \\ 2 \end{bmatrix}+\begin{bmatrix} 1 \\ -2 \end{bmatrix}=\begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

Виконаємо ще одну ітерацію.

$$\nabla f\left(\mathbf{X}_1\right)=\begin{bmatrix} 8\left(x_1^{(1)}-2\right)+x_2^{(1)}\left(x_1^{(1)}x_2^{(1)}-2\right) \\ x_1^{(1)}\left(x_1^{(1)}x_2^{(1)}-2\right) \end{bmatrix}=\begin{bmatrix} 0 \\ -4 \end{bmatrix};$$

$$\mathbf{H}(\mathbf{X}_1)=\begin{bmatrix} 8+\left(x_2^{(1)}\right)^2 & 2x_1^{(1)}x_2^{(1)}-2 \\ 2x_1^{(1)}x_2^{(1)}-2 & \left(x_1^{(1)}\right)^2 \end{bmatrix}=\begin{bmatrix} 8 & -2 \\ -2 & 4 \end{bmatrix}.$$

$$\mathbf{H}(\mathbf{X}_1)\Delta\mathbf{X}_1=-\nabla f\left(\mathbf{X}_1\right)\ \Rightarrow\ \begin{bmatrix} 8 & -2 \\ -2 & 4 \end{bmatrix}\Delta\mathbf{X}_1=-\begin{bmatrix} 0 \\ -4 \end{bmatrix}\ \Rightarrow\ \Delta\mathbf{X}_1=\begin{bmatrix} \frac{2}{7} \\ \frac{8}{7} \end{bmatrix}.$$

Таким чином

$$\mathbf{X}_2=\mathbf{X}_1+\Delta\mathbf{X}_1=\begin{bmatrix} 2 \\ 0 \end{bmatrix}+\begin{bmatrix} \frac{2}{7} \\ \frac{8}{7} \end{bmatrix}=\begin{bmatrix} \frac{16}{7} \\ \frac{8}{7} \end{bmatrix}\simeq\begin{bmatrix} 2{,}28571 \\ 1{,}14286 \end{bmatrix}.$$

Як видно ітерації сходяться до точного розв'язку.

7.6. Методи змінної метрики (квазіньютонівські методи)

Основним недоліком методу Ньютона є громіздкі обчислення матриці Гессе. У методах змінної метрики (матрицю Гессе можна інтерпретувати як метрику в просторі градієнтів) матриця Гессе заміняється матрицею, яка обчислюється простіше. Один із способів побудови методів змінної метрики полягає у застосуванні методу січних розв'язання нелінійних рівнянь до системи

$$\nabla f\left(\mathbf{X}\right) = 0.$$

Матриця Гессе є матрицею Якобі, побудованою для компонентів градієнта цільової функції. Використовуючи метод січних на кожному ітераційному кроці, нову матрицю Гессе $\tilde{\mathbf{H}}_{i+1}$ перераховують на основі попередньої $\tilde{\mathbf{H}}_i$.

Нехай $\mathbf{H}_0$ – обчислена матриця Гессе в початковій точці пошуку $\mathbf{X}_0$ або деяка інша додатньо-визначена симетрична матриця, наприклад, одинична матриця. Тоді вектор напряму $\Delta\mathbf{X}_i$ пошуку на i-му кроці та наступну точку траєкторії можна обчислити так:

$$\tilde{\mathbf{H}}_i \Delta\mathbf{X}_i = -\nabla f\left(\mathbf{X}_i\right);$$

$$\mathbf{X}_{i+1} = \mathbf{X}_i + \alpha\Delta\mathbf{X}_i,$$

де α визначають з умови мінімуму функції $f(\mathbf{X}_i + \alpha\Delta\mathbf{X}_i)$ уздовж $\Delta\mathbf{X}_i$ одним з методів одновимірного пошуку.

Наступне наближення матриці $\tilde{\mathbf{H}}$ знаходимо за виразом:

$$\tilde{\mathbf{H}}_{i+1} = \tilde{\mathbf{H}}_i + \frac{\alpha\Delta\mathbf{X}_i\Delta\mathbf{X}_i^{\mathrm{T}}}{\Delta\mathbf{X}_i^{\mathrm{T}}\mathbf{P}_i} - \frac{\tilde{\mathbf{H}}_i\mathbf{P}_i\mathbf{P}_i^{\mathrm{T}}\tilde{\mathbf{H}}_i}{\mathbf{P}_i^{\mathrm{T}}\tilde{\mathbf{H}}_i\mathbf{P}_i},$$

де $\mathbf{P}_i = \nabla f\left(\mathbf{X}_{i+1}\right) - \nabla f\left(\mathbf{X}_i\right).$

7.7. Методи розв'язання задач умовної оптимізації

Під задачею умовної оптимізації розуміють задачу пошуку мінімуму функції $f(\mathbf{X})$ у разі обмежень:

$\mathbf{V}(\mathbf{X}) = 0$ (m обмежень-рівностей);

$\mathbf{U}(\mathbf{X}) \geq 0$ (k обмежень-нерівностей).

Цю задачу можна розв'язати за допомогою одного з двох підходів. У першому підході враховується, що більшість розвинутих методів оптимізації орієнтовані на пошук безумовного мінімуму. Тому їх застосування потребує, щоб задача умовної оптимізації була попередньо зведена до задачі безумовної оптимізації. У другому підході використовують методи, спеціально розроблені для розв'язання задач нелінійного програмування з обмеженнями.

У випадках, коли обмеження-нерівності мають простий вигляд, наприклад $a \leq x_i \leq b$, перехід до задачі безумовної оптимізації виконують за допомогою заміни змінних, наприклад:

$$y_i = \operatorname{tg}\left(\pi \frac{x_i - \dfrac{b+a}{2}}{b-a} \right).$$

Якщо всі обмеження-рівності подано в аналітичному вигляді, то перехід до задачі безумовної оптимізації часто виконують методом невизначених множників Лагранжа, у якому нова цільова функція – функція Лагранжа:

$$\psi(\mathbf{X}, \Lambda) = f(\mathbf{X}) + \Lambda^{\mathrm{T}} \mathbf{V}(\mathbf{X}) = f(\mathbf{X}) + \sum_{j=1}^{m} \lambda_j \upsilon_j(\mathbf{X}),$$

де $\mathbf{\Lambda} = \left[\lambda_1, \lambda_2, \ldots, \lambda_m\right]^{\mathrm{T}}$ – вектор невизначених множників Лагранжа, $\upsilon_j(\mathbf{X})$ – j-ті обмеження типу рівності; m – кількість обмежень.

Щоб знайти значення n невідомих $x_1, x_2, \ldots, x_n$ і m множників Лагранжа $\lambda_1, \lambda_2, \ldots, \lambda_m$, розв'язують систему рівнянь, що виражає необхідні умови екстремуму функції Лагранжа:

$$\frac{\partial \psi(\mathbf{X}, \mathbf{\Lambda})}{\partial x_i} = \frac{\partial f(\mathbf{X})}{\partial x_i} + \sum_{j=1}^{m} \lambda_j \frac{\partial \upsilon_j(\mathbf{X})}{\partial x_i} = 0\ , \quad i = \overline{1,n};$$

$$\frac{\partial \psi(\mathbf{X}, \mathbf{\Lambda})}{\partial \lambda_j} = \upsilon_j(\mathbf{X}) = 0\ , \qquad j = \overline{1,m}.$$

Функція Лагранжа $\psi(\mathbf{X}, \mathbf{\Lambda})$ і цільова функція $f(\mathbf{X})$ у допустимій області збігаються, оскільки тут $\mathbf{V}(\mathbf{X}) = 0$. Тому, якщо оптимальне значення функції Лагранжа знайдено, то воно є одночасно і умовним оптимумом цільової функції $f(\mathbf{X})$.

Приклад:

$$f(\mathbf{X}) = x_1^2 + x_2^2,$$
$$\upsilon(\mathbf{X}) = 2x_1 + x_2 - 2 = 0.$$

Знаходимо функцію Лагранжа

$$\psi(\mathbf{X},\ \mathbf{\Lambda}) = x_1^2 + x_2^2 + \lambda(2x_1 + x_2 - 2).$$

Використовуючи (7.1), маємо:

$$\begin{cases} \dfrac{\partial \psi}{\partial x_1} = 2x_1 + 2\lambda = 0 \\ \dfrac{\partial \psi}{\partial x_2} = 2x_2 + \lambda = 0 \\ \dfrac{\partial \psi}{\partial \lambda} = 2x_1 + x_2 - 2 = 0 \end{cases} \Rightarrow \quad \begin{cases} x_1^* = \dfrac{4}{5}, \\ x_2^* = \dfrac{2}{5}, \\ \lambda^* = -\dfrac{4}{5}; \end{cases}$$

$$\min f(\mathbf{X}) = f(\mathbf{X}^*) = \psi(\mathbf{X}^*,\ \mathbf{\Lambda}) = \frac{4}{5}.$$

Слід зазначити, що на практиці задачу розв'язують не шляхом використання умови (7.1), як це зроблено у вище наведеному прикладі, а прямо використовуючи один з методів розв'язання (n+m)-вимірної задачі безумовної оптимізації. Це зауваження стосується і нижче наведених прикладів, де умова (7.1) застосована виключно для того, щоб отримати аналітичний розв'язок задачі. У разі ж чисельного розв'язання задачі слід застосовувати методи безумовної оптимізації.

Основними методами постановки задач безумовної оптимізації за наявності обмежень є методи штрафних функцій:

$$\psi(\mathbf{X}) = f(\mathbf{X}) + S(\mathbf{X}, r),$$

де $S(\mathbf{X}, r)$ – функція штрафу, а r – штрафний параметр. Функцію штрафу задають таким чином, щоб під час розв'язання задачі безумовної оптимізації значення функції $S(\mathbf{X}, r)$ різко зростало у разі виходу точки $\mathbf{X}$ за межі допустимої області (рис. 7.10–7.13). Зазвичай функцію штрафу використовують як бар'єр, тобто

$$\lim_{r\to 0} S(\mathbf{X},\ r) = \begin{cases} 0, & x \in D, \\ \infty, & x \notin D. \end{cases}$$

Загальну функцію штрафу подають у вигляді суми функцій штрафів, накладених кожним з обмежень-нерівностей і обмежень-рівностей:

$$\mathrm{S}(\mathbf{X},\ r) = \sum_{i=1}^{m} S_i\left(\upsilon_i(\mathbf{X}),\ r\right) + \sum_{j=1}^{k} S_j\left(u_j(\mathbf{X}),\ r\right).$$

Найчастіше використовують такі типи функцій штрафу:

– логарифмічну (рис. 7.10):

$$\mathrm{S}(\mathbf{X},\ r) = \begin{cases} -r\ln\left(u(\mathbf{X})\right), & u(\mathbf{X}) > 0, \\ +\infty, & u(\mathbf{X}) \le 0; \end{cases}$$

– задану оберненою функцією (рис. 7.11):

$$S(\mathbf{X},\, r) = \frac{2r}{u(\mathbf{X}) + |u(\mathbf{X})|} = \begin{cases} \dfrac{r}{u(\mathbf{X})}, & u(\mathbf{X}) > 0, \\ +\infty, & u(\mathbf{X}) \le 0; \end{cases}$$

– типу квадрата зрізки (рис. 7.12):

$$S(\mathbf{X},\, r) = \frac{1}{r}\left(\frac{u(\mathbf{X}) - |u(\mathbf{X})|}{2}\right)^2 = \begin{cases} 0, & u(\mathbf{X}) \ge 0, \\ \dfrac{u(\mathbf{X})^2}{r}, & u(\mathbf{X}) < 0; \end{cases}$$

– для обмежень-рівностей використовують квадратичний штраф (рис. 7.13):

$$S(\mathbf{X},\, r) = \frac{1}{r}\upsilon(\mathbf{X})^2.$$

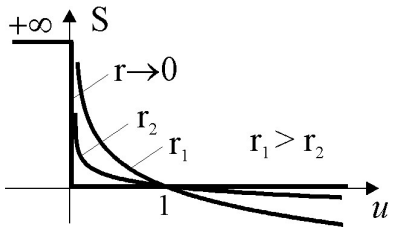


Рис. 7.10. Логарифмічний штраф

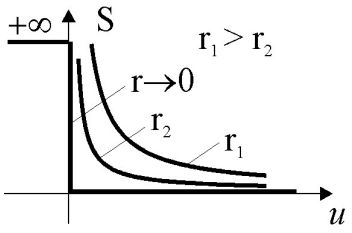


Рис. 7.11. Штраф оберненої функції

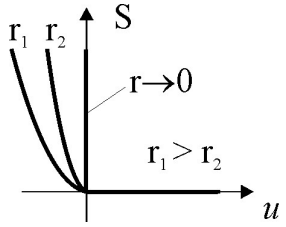


Рис. 7.12. Штраф типу квадрата зрізання

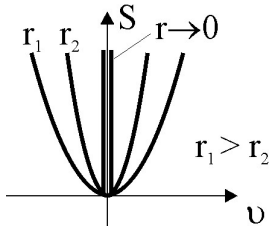


Рис. 7.13. Квадратичний штраф

Після побудови штрафної функції розв'язують задачу безумовної оптимізації для різних r таких, що

$$r_1 > r_2 > r_3 \ldots$$

і знаходять послідовність точок $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$… . Границя цієї послідовності за умови $r \to 0$ і є розв'язком задачі умовної оптимізації. Розв'язання кожної попередньої безумовної задачі розглядають як початкове наближення для розв'язання наступної.

Серед методів штрафних функцій виділяють методи внутрішньої точки (рис. 7.14) і методи зовнішньої точки (рис. 7.15).

Використання методів внутрішньої точки дозволяє утримувати траєкторію пошуку всередині допустимої області. У цьому випадку функція штрафу перешкоджає підходу точок траєкторії пошуку до границі. У методі внутрішньої точки як функції штрафу частіше використовують логарифмічну або обернену функцію.

Приклад:

$$f(\mathbf{X}) = x_1 + x_2,$$
$$u_1(\mathbf{X}) = -x_1^2 + x_2 \geq 0,$$
$$u_2(\mathbf{X}) = x_1 \geq 0.$$

Побудуємо штрафну функцію, використовуючи логарифмічну функцію штрафу:

$$\psi(\mathbf{X}) = x_1 + x_2 - r\ln\left(-x_1^2 + x_2\right) - r\ln x_1.$$

Для аналітичного розв'язання задачі оптимізації скористаємось умовою (7.1):

$$\begin{cases} \dfrac{\partial\psi}{\partial x_1} = 1 + \dfrac{2rx_1}{-x_1^2 + x_2} - \dfrac{r}{x_1} = 0, \\ \dfrac{\partial\psi}{\partial x_2} = 1 - \dfrac{r}{-x_1^2 + x_2} = 0. \end{cases}$$

Розв'язками цієї системи є:

$$x_1(r)=\frac{-1+\sqrt{1+8r}}{4}\ ;\qquad x_2(r)=\frac{\left(-1+\sqrt{1+8r}\right)^2}{16}+r.$$

Розв'язком задачі умовної оптимізації є границя отриманих результатів за умови $r \to 0$:

$$\mathbf{X}^* = \lim_{r\to 0}\mathbf{X}(r) = \left[0,\ 0\right]^{\mathrm{T}}.$$

На практиці ж мінімум шукають методами безумовної оптимізації починаючи з якогось r. Величину r поступово зменшують, використовуючи попередньо отриманий результат в якості першого наближення розв'язку для даного r. Останній розв'язок для r = 0 і є розв'язком задачі умовної оптимізації. Рис. 7.14 ілюструє траєкторію руху розв'язків для вищенаведеного прикладу для різних r. Як видно, всі розв'язки утримуються в дозволеній області. Це і зумовило назву групи методів, як методи внутрішньої точки.

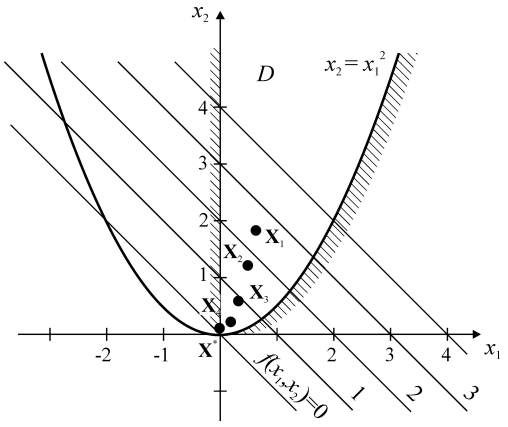


Рис. 7.14. Метод внутрішньої точки

i	r_i	x_1	x_2
1	1,5	0,65	1,92
2	1	0,5	1,25
3	0,57	0,31	0,59
4	0,25	0,18	0,28
5	0,1	0,08	0,11
6	0	0	0

У використанні методів внутрішньої точки важливим є вибір початкових значень вектора **X** і параметра r. Необхідність вибору початкової точки з припустимої області є непростою задачею, і це головне для розглянутого методу. Значення параметра r на першому етапі можна

вважати $r = 1$, або його вибирають з урахуванням перших частинних похідних цільової функції і функції штрафу.

У методі зовнішньої точки функція штрафу підштовхує точку траєкторії пошуку ззовні допустимої області до її границі. Для цього, як функцію штрафу, часто використовують функцію штрафу типу квадрата зрізання.

Приклад:

$$f(\mathbf{X}) = x_1 + x_2,$$
$$u_1(\mathbf{X}) = -x_1^2 + x_2 > 0,$$
$$u_2(\mathbf{X}) = x_1 \geq 0.$$

Побудуємо штрафну функцію, використовуючи функцію штрафу типу квадрата зрізання:

$$\psi(\mathbf{X}) = x_1 + x_2 + \frac{\left(-x_1^2 + x_2\right)^2}{r} + \frac{x_1^2}{r}.$$

З умови (7.1) маємо:

$$\begin{cases} \dfrac{\partial \psi}{\partial x_1} = 1 - \dfrac{4x_1\left(-x_1^2 + x_2\right)}{r} + \dfrac{2x_1}{r} = 0, \\ \dfrac{\partial \psi}{\partial x_2} = 1 + \dfrac{2\left(-x_1^2 + x_2\right)}{r} = 0. \end{cases}$$

Тоді

$$x_1(r) = -\frac{r}{2(1+r)},$$
$$x_2(r) = -\frac{r^2}{4(1+r)^2} - \frac{r}{2};$$
$$\mathbf{X}^* = \lim_{r \to 0} \mathbf{X}(r) = [0,\ 0]^{\mathrm{T}}.$$

Як видно з рис. 7.15, траєкторія руху розв'язків для різних r завжди утримується зовні дозволеної області. Це і зумовлює назву цих методів як методи зовнішньої точки.

i	r_i	x_1	x_2
1	1,5	– 0,3	– 0,58
2	1	– 0,25	– 0,44
3	0,5	– 0,17	– 0,22
4	0,25	– 0,1	– 0,12
5	0,1	– 0,05	– 0,05
6	0	0	0

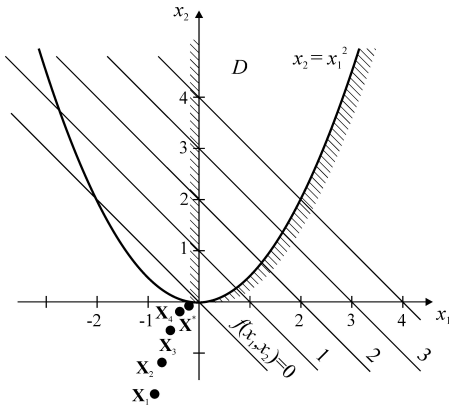


Рис. 7.15. Метод зовнішньої точки

Як приклад методів, орієнтованих на безпосереднє розв'язання задач умовної оптимізації, розглянемо метод проекції градієнта.

Цей метод здебільшого застосовують для розв'язання задач умовної оптимізації з обмеженнями типу рівностей. Умовний мінімум цільової функції знаходиться на n-мірній поверхні обмежень, тому що тільки тут задовольняється умова $\mathbf{V}(\mathbf{X}) = 0$. Тому алгоритмом методу проекції градієнта формують два основні етапи:

Eman 1. Повернення на поверхню обмежень з поточної точки пошуку $\mathbf{X}_i$, якщо ця точка вийшла за припустимі межі порушення обмеження $\Delta\upsilon$:

$$\left\|\mathbf{V}\left(\mathbf{X}_i\right)\right\| \geq \Delta\upsilon.$$

Такий рух відбувається по нормалі до поверхні обмежень (наприклад з точки $\mathbf{X}_0$ у точку $\mathbf{X}_1$, рис. 7.16). Крок переміщення залежить від значення відхилення і визначають за формулою

$$\Delta\mathbf{X}_i = -\mathbf{D}_i\left(\mathbf{D}_i\mathbf{D}_i^{\mathrm{T}}\right)^{-1}\mathbf{V}\left(\mathbf{X}_i\right),$$

де $\mathbf{D}_i$ – матриця розміром $m \times n$, рядками якої є градієнти функцій-обмежень $\mathbf{V}(\mathbf{X})$, обчислені в точці $\mathbf{X}_i$:

$$\mathbf{D}_i = \begin{bmatrix} \dfrac{\partial \upsilon_1}{\partial x_1} & \dfrac{\partial \upsilon_1}{\partial x_2} & \cdots & \dfrac{\partial \upsilon_1}{\partial x_n} \\ \dfrac{\partial \upsilon_2}{\partial x_1} & \dfrac{\partial \upsilon_2}{\partial x_2} & \cdots & \dfrac{\partial \upsilon_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial \upsilon_m}{\partial x_1} & \dfrac{\partial \upsilon_m}{\partial x_2} & \cdots & \dfrac{\partial \upsilon_m}{\partial x_n} \end{bmatrix}.$$

У такий спосіб на етапі 1

$$\mathbf{X}_{i+1} = \mathbf{X}_i + \Delta\mathbf{X}_i.$$

Після потрапляння в малий окіл поверхні обмежень виконують другий етап алгоритму.

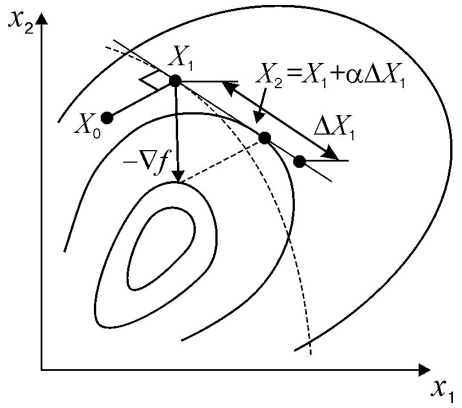


Рис. 7.16. Метод проекції градієнта

Етап 2. Переміщення у бік умовного екстремуму вздовж площини, дотичної в точці $\mathbf{X}_{i+1}$ поверхні обмежень. Напрям пошуку мінімуму визначається за напрямом проекції антиградієнта цільової функції на цю площину:

$$\mathbf{S}_{i+1} = \mathbf{H}_{i+1}\nabla f\left(\mathbf{X}_{i+1}\right),$$

де $\mathbf{H}_{i+1} = \mathbf{E} - \mathbf{D}_{i+1}^{\mathrm{T}}\left(\mathbf{D}_{i+1}\mathbf{D}_{i+1}^{\mathrm{T}}\right)^{-1}\mathbf{D}_{i+1}$ – проектувальна матриця.

Якщо значення кроку h_{i+1} визначено, то нове положення точки траєкторії можна знайти за рекурентним виразом

$$\mathbf{X}_{i+2} = \mathbf{X}_{i+1} + h_{i+1}\mathbf{S}_{i+1}.$$

На рис. 7.16 ця формула зображає перехід з точки $\mathbf{X}_1$ до точки $\mathbf{X}_2$.

Якщо поверхня обмежень нелінійна, то переміщення уздовж дотичної площини може призвести до порушення умови $\left\|\mathbf{V}\left(\mathbf{X}_{i+2}\right)\right\| < \Delta\upsilon$. Тоді наступним етапом буде етап 1 і процес пошуку повториться.

Контрольні завдання

1. Вибрати одномірну цільову функцію відповідно до свого варіанта.
2. Знайти відрізки унімодальності цільової функції.
3. В межах відрізку унімодальності знайти мінімум функції методом золотого перерізу з похибкою, не більшою 1%.

Варіанти завдань

1. $x^4 - 4x^2$.

2. $x^4 - x$.

3. $x^3 - x$.

4. $-x^3 + x$.

5. $x^3 - 3x^2$.

6. $-x^3 + 3x^2$.

7. $-x^3 + 3x$.

8. $x^3 - 3x$.

9. $x^2 - x$.

10. $x^4 + 3x$.

11. $x^4 - 3x$.

12. $\sin\left(x^2\right)$.

13. $\cos\left(x^2\right)$.

14. $\sin(x)/x$.

15. $-\sin^2(x)$.

16. $-\sin^2\left(x\right)e^x$.

17. $\left(\cos\left(x\right)-1\right)/x$.

18. $5\,\mathrm{ch}(x)$.

19. $\mathrm{sh}\left(x\right) - x^2$.

20. $\mathrm{sh}(x-1)/\left(x-1\right)$.

21. $\cosh(x) - x^2$.

22. $x^2 - \mathrm{sh}(x)$.

23. $\mathrm{ch}(x) - 2x$.

24. $\mathrm{Re}(x^x)$.

25. $x^2 - \ln(x)$.

26. $-e^{(-x^2)}$.

27. $5e^{(-x)}\sin(x)$.

28. $2x/(x^2+1)$.

29. $-0,2x^5 + x + 4$.

30. $x^4 - 2x^3 - 1$.

4. Вибрати двомірну цільову функцію відповідно до свого варіанта.
5. Знайти мінімум функції методом покоординатного спуску з похибкою, не більшою 1%.
6. Знайти мінімум функції методом градієнтного спуску з похибкою, не більшою 1%.
7. Знайти мінімум функції методом спряжених градієнтів з похибкою, не більшою 1%.
8. Знайти мінімум функції методом Ньютона з похибкою, не більшою 1%.
9. Порівняти результати пп. 5–8.

Варіанти завдань

1. $(2 - 2x_1 - x_2^2)^2 + (x_1 + x_2 - 1)^2$.

2. $(x_1 - 2x_2 + 2)^2 + (4x_1^2 - x_2 + 1)^2$.

3. $(x_1 - x_2 - 2)^4 + (2x_1^2 + x_2 - 1)^2$.

4. $\left(\frac{6x_1}{x_2 + 5} + 1\right)^2 + (2x_2 - x_1 - 3)^2$.

5. $(2x_1 + x_2 + 1)^2 + \left(\frac{8x_2}{x_1 + 4} + 2\right)^2$.

6. $\left(x_1 + \frac{x_2}{4} + 1\right)^2 + (x_1 + x_2 + 1)^4$.

7. $\left(2-x_1x_2\right)^2+\left(x_1+2x_2-5\right)^2$.

8. $\left(2x_1-x_2-3\right)^2+\left(\frac{x_1x_2}{2}-1\right)^2$.

9. $\left(x_1^2-x_2+1\right)^2+\frac{\left(3x_1+x_2+1\right)^2}{2}$.

10. $\left(x_1-4x_2^2+2\right)^2+\left(x_1-x_2-3\right)^2$.

11. $\left(2x_1-x_2-4\right)^4+\left(x_1+x_2-2\right)^2$.

12. $\left(\frac{x_1-4}{x_2+2}+1\right)^2+\left(x_1+2x_2-4\right)^2$.

13. $\left(x_1+x_2+2\right)^2+\left(\frac{x_2+2}{x_1+4}-1\right)^2$.

14. $\left(x_1+\frac{x_2}{2}+1\right)^2+\left(x_1-x_2-2\right)^4$.

15. $\left(x_1x_2+3\right)^2+\left(x_1+x_2-2\right)^2$.

16. $\left(2x_1+3x_2-7\right)^2+\left(x_1x_2-x_1+2\right)^2$.

17. $\left(x_1-x_2^2+8\right)^2+\left(x_1+x_2-4\right)^2$.

18. $\left(x_1-2x_2-1\right)^2+\left(x_1^2-4x_2-5\right)^2$.

19. $\left(2x_1-x_2-6\right)^4+\left(x_1+2x_2+2\right)^2$.

20. $\left(\frac{5x_1}{x_2+3}+2\right)^2+\left(2x_2-x_1-6\right)^2$.

21. $\left(x_1-x_2+3\right)^2+\left(\frac{4x_2}{x_1-2}+1\right)^2$.

22. $\left(x_1+x_2+1\right)^2+\left(x_1-x_2-3\right)^4$.

23. $\left(3+x_1x_2\right)^2+\left(2x_1+x_2+5\right)^2$.

24. $\left(x_1-x_2-4\right)^2+\left(x_1x_2+2x_1+1\right)^2$.

25. $\left(x_1^2-x_2+3\right)^2+\frac{\left(3x_1+x_2-3\right)^2}{2}$.

26. $\left(2x_1-x_2^2-6\right)^2+\left(x_1+2x_2-3\right)^2$.

27. $\left(x_1-x_2+1\right)^4+\left(x_1+x_2-5\right)^2$.

28. $\left(\frac{x_1+1}{x_2+2}-1\right)^2+\left(x_1-2x_2+1\right)^2$.

29. $\left(2x_1+x_2+1\right)^2+\left(\frac{x_2+1}{x_1-2}+1\right)^2$.

30. $\left(x_1+x_2-1\right)^2+\left(x_1-x_2-5\right)^4$.

8. Апроксимація функцій

Нехай деяку функцію $y(x)$ задано табл. 8.1.

Таблиця. 8.1. **Деяка таблична функція**

x	x_1	x_2	…	x_n
$y(x)$	y_1	y_2	…	y_n

За отриманими даними необхідно знайти порівняно просту функцію для обчислення її значення в будь-якій точці x. У четвертому розділі розглянуто застосування інтерполяції для розв'язання цієї задачі. У прикладних задачах таблиця значень функції може задаватись із похибкою. Тоді доцільно шукати функцію, яка б не проходила точно через усі вузли, але разом з тим деякою мірою відтворювала характер вихідної функції [2, 5].

Тип апроксимуючої функції визначається характером вихідних даних. Прикладами апроксимуючих функцій є:

1) $y = ax + b;$ 5) $y = \dfrac{1}{ax + b};$

2) $y = ax^2 + bx + c;$ 6) $y = a\ln(x) + b;$

3) $y = ax^m;$ 7) $y = a\dfrac{1}{x} + b;$

4) $y = ae^{(mx)};$ 8) $y = \dfrac{x}{ax + b}.$

Нехай $f(x)$ – деяка функція, що наближено відтворює $y(x)$. Уведемо поняття відхилу в точці x_i

$$r_i = y_i - f(x_i)$$

і вектора відхилу

$$\mathbf{R} = \left[r_1, \ldots, r_n\right]^T.$$

Під задачею апроксимації розуміють знаходження такої функції $f(x)$, яка мінімізує норму вектора відхилу $\|\mathbf{R}\|$. Задачі апроксимації розрізняють за типом використовуваних норм. Так, мінімізація евклідової норми відхилу

$$\|\mathbf{R}\|_2 = \sqrt{\sum_{i=1}^{n} r_i^2} = \sqrt{\sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2}$$

називається середньоквадратичним наближенням, а норми Чебишова

$$\|\mathbf{R}\|_\infty = \max_{i=1,n} |r_i| = \max_{i=1,n} |y_i - f(x_i)|$$

– рівномірним наближенням.

Якщо $\|\mathbf{R}\| = 0$, то задача апроксимації збігається із задачею інтерполяції (рис. 8.1).

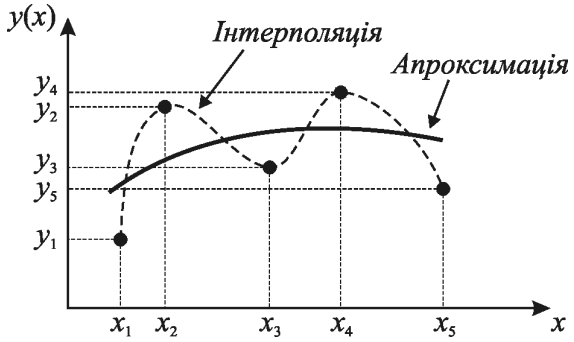


Рис. 8.1. Інтерполяція та апроксимація табличних залежностей

Оскільки мінімум норми Евкліда збігається з мінімумом її квадрата, то задачу середньоквадратичного наближення формулюють так:

$$\min \|\mathbf{R}\|_2^2 = \min \left[\sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 \right]. \qquad (8.1)$$

Апроксимація функцій шляхом розв’язання (8.1) називається методом найменших квадратів (МНК).

Апроксимуючу функцію будують у вигляді

$$f(x) = f(x,\ c_1,\ldots,\ c_m),$$

де c_i – коефіцієнти, які знаходять із необхідної умови існування мінімуму норми відхилу:

$$\frac{\partial \|\mathbf{R}\|}{\partial c_i} = 0, \qquad i = \overline{1,m}\,. \tag{8.2}$$

Вираз (8.2) є системою m рівнянь відносно коефіцієнтів c_i. Якщо ця система лінійна, то кажуть про лінійну задачу апроксимації.

Задачу можна поставити так, що не всі точки таблиці, що апроксимують, є рівноцінними, наприклад, одні з них мають меншу похибку чи більш важливі за певним критерієм. У таких випадках використовують зважену норму, наприклад, зважену норму Евкліда:

$$\min\left[\sum_{i=1}^{n}\rho(x)\left(y_i - f(x_i)\right)^2\right],$$

де $\rho(x)$ – вагова функція, яка, зокрема, може описувати ступінь імовірності чи важливості окремих точок.

Інколи неперервну функцію $y(x)$, задану на відрізку $[a, b]$, необхідно апроксимувати простішою функцією $f(x)$. Як міру близькості $f(x)$ до $y(x)$ використовують умови:

1) $\min \max\limits_{x\in[a,b]} \left|y(x) - f(x)\right|$ – неперервне рівномірне наближення;

2) $\min \int\limits_a^b \rho(x)\left(y(x) - f(x)\right)^2 dx$ – неперервне середньоквадратичне наближення.

8.1. Лінійна задача про середньоквадратичне наближення

Апроксимуючу функцію будують у вигляді узагальненого полінома:

$$f(x)=\sum_{j=1}^{m} c_j \varphi_j(x),$$

де $\varphi_j(x)$ – система відомих лінійно незалежних функцій; c_j – коефіцієнти, які необхідно визначити. Коефіцієнти беруть такими, щоб мінімізувати вираз

$$\sigma=\sum_{i=1}^{n}\rho(x_i)\left(y_i-f(x_i)\right)^2=\sum_{i=1}^{n}\rho(x_i)\left(y_i-\sum_{j=1}^{m}c_j\varphi_j(x_i)\right)^2. \qquad (8.3)$$

З необхідної умови існування мінімуму (8.3) знаходимо:

$$\begin{aligned}
\frac{\partial\sigma}{\partial c_1}&=2\sum_{i=1}^{n}\rho(x_i)\varphi_1(x_i)\left(y_i-\sum_{j=1}^{m}c_j\varphi_j(x_i)\right)=0;\\
\frac{\partial\sigma}{\partial c_2}&=2\sum_{i=1}^{n}\rho(x_i)\varphi_2(x_i)\left(y_i-\sum_{j=1}^{m}c_j\varphi_j(x_i)\right)=0;\\
&\cdots\\
\frac{\partial\sigma}{\partial c_m}&=2\sum_{i=1}^{n}\rho(x_i)\varphi_m(x_i)\left(y_i-\sum_{j=1}^{m}c_j\varphi_j(x_i)\right)=0.
\end{aligned} \qquad (8.4)$$

Введемо позначення:

$$\begin{aligned}
(\varphi_k,\ \varphi_j)&=\sum_{i=1}^{n}\varphi_k(x_i)\varphi_j(x_i)\rho(x_i);\\
(\varphi_k,\ y)&=\sum_{i=1}^{n}\varphi_k(x_i)y_i\rho(x_i).
\end{aligned} \qquad (8.5)$$

З урахуванням позначень (8.5), які називають скалярними добутками, необхідна умова існування мінімуму (8.4) має вигляд:

$$\sum_{j=1}^{m}\left(\varphi_k,\ \varphi_j\right)c_j=\left(\varphi_k,\ y\right),\quad k=\overline{1,m}\,. \tag{8.6}$$

Визначник системи (8.6)

$$\begin{vmatrix}\left(\varphi_1,\ \varphi_1\right) & \left(\varphi_1,\ \varphi_2\right) & \cdots & \left(\varphi_1,\ \varphi_m\right)\\ \left(\varphi_2,\ \varphi_1\right) & \left(\varphi_2,\ \varphi_2\right) & \cdots & \left(\varphi_2,\ \varphi_m\right)\\ & \cdots & & \\ \left(\varphi_m,\ \varphi_1\right) & \cdots & \cdots & \left(\varphi_m,\ \varphi_m\right)\end{vmatrix}$$

називається визначником Грама і дорівнює нулю тільки за умови, якщо система функцій $\varphi_j(x)$ лінійно залежна. Виходячи з припущення про лінійну незалежність системи функцій $\varphi_j(x)$, визначник Грама не дорівнює нулю, тому система (8.6) має єдиний розв'язок.

Матриця коефіцієнтів системи (8.6) симетрична і додатньо-визначена, тому для її розв'язання доцільно використовувати метод $\mathbf{L}^{\dagger}\mathbf{L}$-факторизації.

Аналогічно розв'язується задача неперервного середньоквадратичного наближення. Можна показати, що задача зводиться до розв'язання СЛАР типу (8.6) відносно коефіцієнтів c_j, причому відповідні скалярні добутки обчислюються за формулами

$$\left(\varphi_k,\ \varphi_j\right)=\int_a^b\rho(x)\varphi_k(x)\varphi_j(x)\,dx;$$

$$\left(\varphi_k,\ y\right)=\int_a^b\rho(x)\varphi_k(x)y(x)\,dx.$$

Задача спрощується, якщо система функцій $\{\varphi_j(x)\}$ ортогональна. Тоді

$$\left(\varphi_k,\ \varphi_j\right)=0,\ k\neq j$$

і матриця системи (8.6) буде діагональною, а шукані коефіцієнти можна знайти так:

$$c_j = \frac{\left(\varphi_j,\, y\right)}{\left(\varphi_j,\, \varphi_j\right)}.$$

Розглянемо приклади неперервного та дискретного середньоквадратичного наближення.

Приклад. На відрізку $x \in [0,1]$ наблизити функцію $y(x) = x^2 + x$ поліномом першого степеня $f(x) = c_1 + c_2 x$.

В даному випадку системою базисних функцій є $\varphi_1(x) = 1;\ \varphi_2(x) = x$, а вагова функція $\rho(x) \equiv 1$. Тоді:

$$\left(\varphi_1,\, \varphi_1\right) = \int_0^1 1 \cdot 1 dx = 1;$$

$$\left(\varphi_1,\, \varphi_2\right) = \left(\varphi_2,\, \varphi_1\right) = \int_0^1 1 \cdot x dx = \frac{1}{2};$$

$$\left(\varphi_2,\, \varphi_2\right) = \int_0^1 x \cdot x dx = \frac{1}{3};$$

$$\left(\varphi_1,\, y\right) = \int_0^1 1 \cdot \left(x^2 + x\right) dx = \frac{5}{6};$$

$$\left(\varphi_2,\, y\right) = \int_0^1 x \cdot \left(x^2 + x\right) dx = \frac{7}{12}.$$

Тоді система (8.6) має вигляд:

$$\begin{cases} c_1 + \dfrac{1}{2} c_2 = \dfrac{5}{6}; \\ \dfrac{1}{2} c_1 + \dfrac{1}{3} c_2 = \dfrac{7}{12}; \end{cases} \quad \Rightarrow \quad \begin{cases} c_1 = -\dfrac{1}{6}; \\ c_2 = 2. \end{cases}$$

Таким чином, поліном, апроксимуючий функцію $y(x) = x^2 + x$ на відрізку $x \in [0,1]$, має вигляд $f(x) = 2x - \frac{1}{6}$.

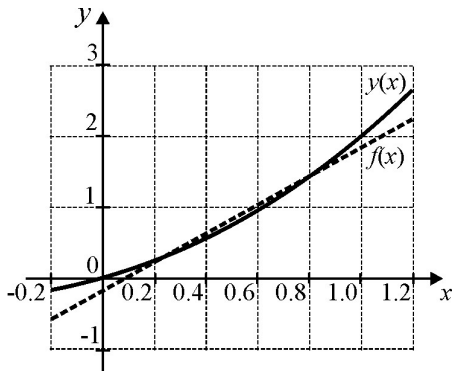


Рис. 8.2. Апроксимація неперервної функції на відрізку

Приклад. Апроксимуємо поліномом першого степеня $f(x) = c_1 + c_2 x$ таблицю:

x	–2	0	1	2	4
y	–1	1	0	3	3

Оскільки системою базисних функцій є $\varphi_1(x) = 1;\ \varphi_2(x) = x$, а вагова функція $\rho(x) \equiv 1$, то:

$$(\varphi_1,\ \varphi_1) = \sum_{i=1}^{5} 1 \cdot 1 = 5;$$

$$(\varphi_1,\ \varphi_2) = (\varphi_2,\ \varphi_1) = \sum_{i=1}^{5} 1 x_i = -2 + 0 + 1 + 2 + 4 = 5;$$

$$(\varphi_2,\ \varphi_2) = \sum_{i=1}^{5} x_i x_i = 4 + 0 + 1 + 4 + 16 = 25;$$

$$(\varphi_1,\ y) = \sum_{i=1}^{5} 1 y_i = -1 + 1 + 0 + 3 + 3 = 6;$$

$$(\varphi_2,\ y) = \sum_{i=1}^{5} x_i y_i = 2 + 0 + 0 + 6 + 12 = 20.$$

Систему (8.6) записуємо у вигляді:

$$\begin{cases} 5c_1 + 5c_2 = 6; \\ 5c_1 + 25c_2 = 20; \end{cases} \Rightarrow \quad \begin{cases} c_1 = 0,5; \\ c_2 = 0,7. \end{cases}$$

Відповідно апроксимуючу функцію записуємо так: $f(x) = 0,7x + 0,5$.

x	–2	0	1	2	4
y	–1	1	0	3	3
$f(x)$	–0,9	0,5	1,2	1,9	3,3
$r(x)$	–0,1	0,5	–1,2	1,1	–0,3

Точки табличної функції і графік одержаної аналітичної залежності показано на рис. 8.3. Норма відхилу $\|\mathbf{R}\|_2 = \sqrt{3}$.

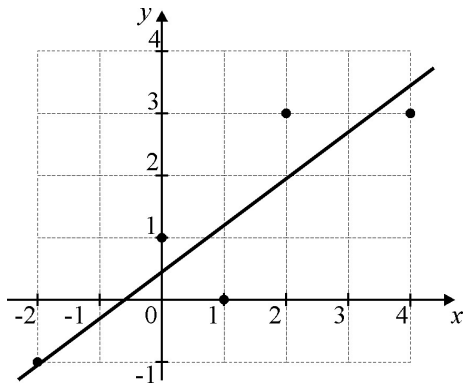


Рис. 8.3. Апроксимація табличної функції

Слід звернути увагу, що задача в розглянутому прикладі є лінійною не тому, що апроксимуюча функція лінійна, а тому, що є лінійною система рівнянь відносно c_0 та c_1.

Не всі задачі апроксимації є лінійними. Якщо система рівнянь, отримана з (8.2) є нелінійною відносно параметрів апроксимації c_i, то використовують метод Ньютона-Гауса [6], який є модифікацією метода Ньютона для задачі безумовної оптимізації на випадок спеціальної структури цільової функції, яка може бути представлена у вигляді $\mathbf{R}^T\mathbf{R}$. Проте, в багатьох випадках нелінійна задача апроксимації може бути

зведена до лінійної за допомогою спеціальних прийомів, зокрема логарифмування. Проілюструємо деякі прийоми на прикладах:

1. $f\left(x,\ c_1,\ c_2\right)=c_1x^{c_2}$.

Прологарифмуємо $f\left(x,\ c_1,\ c_2\right)$. У результаті маємо:

$$\ln\left(f\left(x,\ c_1,\ c_2\right)\right)=\ln\left(c_1\right)+c_2\ln\left(x\right).$$

Уведемо новий параметр $\tilde{c}_1=\ln\left(c_1\right)$. Тоді

$$\ln\left(f\left(x,\ c_1,\ c_2\right)\right)=\tilde{f}\left(x,\ \tilde{c}_1,\ c_2\right)=\tilde{c}_1+c_2\ln\left(x\right).$$

Тобто задача звелась до лінійної. В ній $\varphi_1\left(x\right)=1,\ \varphi_2\left(x\right)=\ln\left(x\right)$.

Практично треба виконати такі кроки:

– прологарифмувати значення функції y_i вихідної таблиці:

x	x_1	x_2	…	x_n
$\tilde{y}$	$\ln(y_1)$	$\ln(y_2)$	…	$\ln(y_n)$

– за новою таблицею знайти параметри $\tilde{c}_1,\ c_2$ з системи лінійних рівнянь:

$$\begin{cases} n\tilde{c}_1+\left(\sum\limits_{i=1}^{n}\ln\left(x_i\right)\right)c_2=\sum\limits_{i=1}^{n}\tilde{y}_i; \\ \left(\sum\limits_{i=1}^{n}\ln\left(x_i\right)\right)\tilde{c}_1+\left(\sum\limits_{i=1}^{n}\ln^2\left(x_i\right)\right)c_2=\sum\limits_{i=1}^{n}\ln\left(x_i\right)\tilde{y}_i; \end{cases}$$

– обчислити $c_1=e^{\tilde{c}_1}$.

2. $f\left(x,\ c_1,\ c_2\right)=c_1e^{c_2x}$.

Прологарифмуємо $f\left(x,\ c_1,\ c_2\right)$. Тоді

$$\ln\left(f\left(x,\ c_1,\ c_2\right)\right)=\ln\left(c_1\right)+c_2x.$$

Позначимо:

$$\ln\left(c_1\right)=\tilde{c}_1.$$

У результаті маємо

$$\ln\left(f\left(x,\ c_1,\ c_2\right)\right)=\tilde{f}\left(x,\ \tilde{c}_1,\ c_2\right)=\tilde{c}_1+c_2x.$$

Задачу зведено до лінійної. В ній $\varphi_1\left(x\right)=1,\ \varphi_2\left(x\right)=x$.

Для практичної реалізації необхідно:

– прологарифмувати значення функції y_i вихідної таблиці;

– за новою таблицею знайти параметри $\tilde{c}_1,\ c_2$;

– обчислити $c_1=e^{\tilde{c}_1}$.

3. $f\left(x,\ c_1,\ c_2\right)=c_1xe^{c_2x}$.

Прологарифмуємо $\dfrac{f\left(x,\ c_1,\ c_2\right)}{x}$. Тоді

$$\ln\left(\frac{f\left(x,\ c_1,\ c_2\right)}{x}\right)=\ln\left(c_1\right)+c_2x.$$

Позначимо:

$$\ln\left(c_1\right)=\tilde{c}_1.$$

У результаті маємо

$$\ln\left(\frac{f\left(x,\ c_1,\ c_2\right)}{x}\right)=\tilde{f}\left(x,\ \tilde{c}_1,\ c_2\right)=\tilde{c}_1+c_2x.$$

Задачу зведено до лінійної. В ній $\varphi_1\left(x\right)=1,\ \varphi_2\left(x\right)=x$.

Для практичної реалізації необхідно:

– прологарифмувати відношення функції до аргументу $\dfrac{y_i}{x_i}$ вихідної таблиці:

x	x_1	x_2	…	x_n
$\tilde{y}$	$\ln\left(\frac{y_1}{x_1}\right)$	$\ln\left(\frac{y_2}{x_2}\right)$	…	$\ln\left(\frac{y_n}{x_n}\right)$

– за новою таблицею знайти параметри $\tilde{c}_1,\ c_2$;

– обчислити $c_1 = e^{\tilde{c}_1}$.

4. $f\left(x,\ c_1,\ c_2\right) = \dfrac{1}{c_1 x + c_2}$.

Введемо нову функцію

$$\tilde{f}\left(x,\ c_1,\ c_2\right) = \frac{1}{f\left(x,\ c_1,\ c_2\right)} = c_1 x + c_2.$$

Задачу зведено до лінійної. Зверніть увагу, що в даному випадку $\varphi_1\left(x\right) = x,\ \varphi_2\left(x\right) = 1$.

Фактично треба лише побудувати нову таблицю, у якій значення аргументу залишаються незмінними, а значення функції будуть обернені до початкових і, використовуючи нову таблицю, знайти параметри $c_1,\ c_2$.

5. $f\left(x,\ c_1,\ c_2\right) = \dfrac{x}{c_1 x + c_2}$.

Розглянемо обернену функцію

$$\tilde{f}\left(x,\ c_1,\ c_2\right) = \frac{1}{f\left(x,\ c_1,\ c_2\right)} = \frac{c_1 x + c_2}{x} = c_1 + \frac{c_2}{x}.$$

Задачу зведено до лінійної. Далі будуємо нову таблицю, у якій значення аргументу будуть обернені до заданих, і для отриманих значень знаходимо параметри $c_1,\ c_2$, враховуючи, що $\varphi_1\left(x\right) = 1,\ \varphi_2\left(x\right) = \dfrac{1}{x}$.

Слід зазначити, що перелік наведених прикладів є далеко не вичерпним, а є лише ілюстрацією деяких можливостей для того, щоб нелінійні задачі апроксимації звести до лінійних.

Контрольні завдання

1. Вибрати із четвертого розділу табличну функцію відповідно до свого варіанта.
2. Методом найменших квадратів апроксимувати таблицю функцією $f(x,\ a,\ b,\ c) = ax^2 + bx + c$.
3. Побудувати графік отриманої залежності та нанести точки таблиці. Порівняти отримані результати з результатами із четвертого розділу.
4. Відповідно до свого варіанта наблизити аналітичну функцію на заданому відрізку поліномом першого степеня.
5. Побудувати графік апроксимуючої функції та отриманого полінома в одних координатних осях.

Варіанти завдань

1. $x^3 - 2x^2 + 3x - 4, \quad x \in [1,2]$.

2. $-x^3 + 3x^2 - 4x + 5, \quad x \in [2,3]$.

3. $\dfrac{(x+2)^2}{x}, \quad x \in [-4,-3]$.

4. $-\dfrac{x}{x+4}, \quad x \in [-7,-6]$.

5. $\dfrac{x-5}{x+6}, \quad x \in [-3,-1]$.

6. $x^2 - \dfrac{2}{x^2} + 3x - 1, \quad x \in [2,5]$.

7. $4x - \dfrac{1}{(x-3)^3} + 2, \quad x \in [4,6]$.

8. $x^3 + \dfrac{1}{x+2} - 3, \quad x \in [2,4]$.

9. $\dfrac{3-x}{5+x}, \quad x \in [-3,-2]$.

10. $2x^2 - \dfrac{1}{(1-x)^2}, \quad x \in [-4,-2]$.

11. $1+2x-3x^2+4x^3, \quad x \in [3,5]$.

12. $-x^4+2x^2-3x-5, \quad x \in [3,4]$.

13. $-\dfrac{(x+1)^3}{x}, \quad x \in [-6,-4]$.

14. $\dfrac{x}{3-x}, \quad x \in [5,6]$.

15. $\dfrac{1-x}{x+1}, \quad x \in [1,3]$.

16. $(x-3)^2 - \dfrac{3}{(3-x)^2}, \quad x \in [-5,-2]$.

17. $\dfrac{3}{x^3} - 2x - 4, \quad x \in [-10,-2]$.

18. $(x-1)^3 + \dfrac{1}{x-1} - 1, \quad x \in [3,6]$.

19. $\dfrac{(1-x)^4}{x}, \quad x \in [-2,-1]$.

20. $-(x+1)^2 + \dfrac{2}{(2+x)^2}, \quad x \in [-1,1]$.

21. $\dfrac{1}{x^3} - \dfrac{x^2}{4} - 2x - 1, \quad x \in [1,6]$.

22. $\dfrac{x^3}{8} + \dfrac{1}{x^2} + x - 3, \quad x \in [-4,-1]$.

23. $\dfrac{(2-x)^3}{x^2}, \quad x \in [-10,-5]$.

24. $\dfrac{(10x-50)^2}{x^3}, \quad x \in [5,10]$.

25. $\dfrac{300-500x}{x^4}, \quad x \in [-7,-5]$.

26. $\dfrac{x^5}{16} - \dfrac{1}{x^3} + \dfrac{x}{10} - 50, \quad x \in [4,5]$.

27. $\dfrac{(x-4)^3}{10} - \dfrac{1}{(x+2)^2}, \quad x \in [-1,2]$.

28. $\dfrac{x^3}{5} - (x-4)^2 + \dfrac{1}{3-x}, \quad x \in [-5,-1]$.

29. $\dfrac{x^4-x^2-10}{x^3}, \quad x \in [2,6]$.

30. $-\dfrac{x^3-2x-20}{x^2}, \quad x \in [-5,-4]$.

Список використаної літератури

1. *Мэтьюз Джон Д., Финк Куртис Д.* Численные методы. Использование Matlab: Пер. с англ. – 3-е изд. – М.: Издат. дом «Вильямс», 2001. – 720 с.
2. *Бахвалов Н. С., Жидков Н.П., Кобельков Г.М.* Численные методы. – М.: Наука, 1987. – 600 с.
3. *Самарский А.А., Гулин А.В*. Численные методы. - М.:Наука,1989. –432 с.
4. *Демидович Б. П., Марон И. А.* Основы вычислительной математики. – М.: Наука, 1970. – 665 с.
5. *Вержбицкий В.М.* Численные методы. Линейная алгебра и нелинейные уравнения. – М.:Высшая школа, 2000.– 266 с. – ISBN 5-06-003654-5.
6. *Денис Дж., Шнабель Р.* Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1988. – 440 с. – ISBN 5-03-001102-1.
7. *Сигорский В.П., Петренко А.И.* Алгоритмы анализа электронных схем. – М.:Сов.радио, 1976. – 608 с.
8. *Фадеев Д.К., Фадеева В.Н.* Вычислительные методы линейной алгебры. – М.:Гос. издат. физ.-мат. литературы, 1960. – 656 с.
9. *Воеводин В.В.* Численные методы алгебры. Теория и алгорифмы. – – М.:Гос. издат. физ.-мат. литературы, 1966.– 248 с.
10. *Гловацкая А. П.* Методы и алгоритмы вычислительной математики. – М.: Радио и связь, 1999. – 408 с. – ISBN 5-256-01458-7.
11. *Амосов А. А., Дубинский Ю. А., Копченова Н. В.* Вычислительные методы для инженеров.– М.: Высшая школа, 1994.– 544 с. – ISBN 5-06-000625-5.
12. *Годунов С.К., Антонов А.Г., Кирилюк О.П., Костин В.И.* Гарантированная точность решения систем линейных уравнений в эвклидовых пространствах. – Новосибирск: Наука, 1988. – 456 с. – ISBN 5-02-028593-5.
13. *Верлань А.Ф., Сизиков В.С.* Интегральные уравнения: методы, алгоритмы, программы. Справочное пособие. –К.: Наукова думка, 1986. –544 с.
14. *Калиткин Н. Н.* Численные методы.: Под редакцией А. А. Самарского – М.: Наука, 1978. – 512 с.
15. *Молчанов И.Н*. Машинные методы решения прикладных задач. Алгебра, приближение функций.- Киев: Наукова думка, 1987.- 288 с.
16. *Боглаев Ю.П.* Вычислительная математика и программирование. - М.: Высшая школа,1990. –540 с. –ISBN 5-06-00623-9.
17. *Golub, Gene H., Van Loan, Charles F*. Matrix computations. – Baltimore: John Hopkins University Press. – 1996. – 694 p. – ISBN:0-8018-5414-8.
18. *William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery* Numerical Recipes in C. The art of scientific Computing. –

Second edition. – Cambrige University Press. –1992. – 997 p. – ISBN 0-521-43108-5.
19. *Икрамов Х. Д.* Несимметричная проблема собственных значений. – М.: Наука, 1991, с. 100.
20. *Березин И.С., Жидков Н.П.* Методы вычислений: В 2 т. – М.: Физматиздат, 1962. – 640 с.
21. *Ортега Дж.*, *Рейнболдт В*. Итерационные методы решения нелинейных систем уравнений со многими неизвестными. – М.: Мир, 1975. – 560 с.
22. *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений. – М.: Мир, 1980. – 279 с.
23. *Григоренко Я.М., Панкратова Н.Д*. Обчислювальні методи в задачах прикладної математики: Навч. посібник. –К.:Либідь, 1995. – 280 с.– ISBN 5-325-00486-7.
24. *На Ц.* Вычислительные методы решения прикладных граничных задач. – М.: Мир, 1982. – 296 с.
25. *Положий Г.Н., Пахарева Н.А., Степаненко И.З., Бондаренко П.С., Великоиваненко И.М.* Математический практикум.: Под редакцией Г.Н. Положего.– М.: Гос. издат. физ.-мат. литературы, 1960. – 512 с.
26. *Хемминг Р.В.* Численные методы для научных работников и инженеров. – М.: Гос. издат. физ.-мат. литературы, 1972. – 400 с.
27. *Вержбицкий В.М.* Численные методы (математический анализ и обыкновенные дифференциальные уравнения). – М.: Высшая школа, 2001. – 382 с. –ISBN 5-06-003982-X.
28. *Васильков Ю. В., Василькова Н. Н.* Компьютерные технологии вычислений в математическом моделировании. – М.: Финансы и статистика, 1999. — 255 с.
29. Справочник по специальным функциям с формулами, графиками и математическими таблицами.: Под редакцией М.Абрамовица и И.Стигана. – М.: Наука, 1979. – 832 с.
30. *Молчанов И.Н*. Машинные методы решения прикладных задач. Дифференциальные уравнения.- Киев: Наукова думка, 1988.- 344 с.
31. *Михлин С. Г., Смолицкий Х. Л.* Приближенные методы решения дифференциальных и интегральных уравнений. Серия: Справочная математическая библиотека.: Под общей редакцией Л.А.Люстерника и А.Р.Янпольского.– М.: Наука, Главная редакция физ.-мат. литературы, 1965. – 384 с.
32. *Положий Г. Н.* Уравнения математической физики.— М.: Высш.шк., 1964. — 560 с.
33. *Хургин Я.И., Яковлев В.П.* Финитные функции в физике и технике. — М.: Наука, 1971. – 408 с.

34. Методи обчислень: Практикум на ЕОМ: навч. посібник / В.Л.Бурківська, С.О.Войцехівський, І.П.Гаврилюк та ін. –К.: Вища шк. , 1995. — 303 с. – ISBN 5-11-004030-3.
35. *Евдокимов А.Г.* Минимизация функций. Харьков: Издательское объединение «Вища школа» , 1977. – 160 с.
36. *Волков А. Е.* Численные методы. – М.: Наука, 1982. – 248 с.
37. *Воробьева Г. Н., Данилова А. Н.* Практикум по вычислительной математике. – М.: Высш. шк., 1990. – 308 с.
38. *Крылов В. И., Бобков В. Б., Монастырный П. И.* Вычислительные методы: В 2 т. – М.: Наука, 1977. – 339 с.
39. *Мак-Кракен Д., Дорн У.* Численные методы и программирование на ФОРТРАНЕ. – М.: Мир, 1977. – 293 с.
40. *Турчак Л. И.* Основы численных методов. – М.: Наука, 1987. – 350 с.
41. *Шуп Т. Е.* Решение инженерных задач на ЭВМ. – М.: Мир, 1990. – 235 с.