

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

# **ОБЧИСЛЮВАЛЬНА МАТЕМАТИКА**

*(з виправленнями та доповненнями)*



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

# **ОБЧИСЛЮВАЛЬНА МАТЕМАТИКА**

*(з виправленнями та доповненнями)*

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

Київ  
«Політехніка»  
2003

УДК 519.6 (075.8)  
ББК 22.19я73  
П78

*Гриф надано Методичною  
радою НТУУ «КПІ»  
(Протокол № 3 від 21. 11. 2002 р.)*

Рецензенти: *С. І. Сіренко*, канд. техн. наук, проф.  
*Б. П. Бездітний*, канд. фіз.-мат. наук, доц.

Відповідальний редактор *Ю. С. Синькоп*, канд. техн. наук, проф.

**Прокопенко Ю. В., Татарчук Д. Д., Казміренко В. А.**  
П78 Обчислювальна математика: Навч. посіб. – К.: ІВЦ “Видавництво «Політехніка»”,  
2003. – 120 с.

Викладено навчальний матеріал про числове розв’язання рівнянь і систем, знаходження власних чисел і векторів, інтерполяції та апроксимації табличних залежностей, числового 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 напрямку «Електроніка».

**УДК 519.6 (075.8)**  
**ББК 22.19я73**

© Ю. В. Прокопенко,  
Д. Д. Татарчук,  
В. А. Казміренко, 2003

# Зміст

Вступ.....	5
1. Похибки обчислень .....	6
1.1. Поняття обчислювального експерименту .....	6
1.2. Похибки обчислень.....	7
1.3. Обчислювальна похибка визначення значення функції.....	8
2. Числові методи лінійної алгебри .....	12
2.1. Метод LU-факторизації.....	12
2.2. Метод QR-факторизації.....	15
2.3. Метод Холеського.....	19
2.4. Похибка розв'язання СЛАР .....	23
2.5. Власні числа і власні вектори .....	25
2.6. Степеневий метод .....	26
2.7. Метод скалярного добутку .....	30
2.8. Метод Віландта .....	33
3. Числові методи розв'язання нелінійних рівнянь .....	36
3.1. Метод бісекції (поділу навпіл) .....	37
3.2. Метод Ньютона розв'язання рівнянь .....	38
з однією змінною .....	38
3.3. Квазіньютонівські методи розв'язання рівнянь .....	40
з однією змінною .....	40
3.4. Метод Ньютона розв'язання .....	42
систем нелінійних рівнянь .....	42
3.5. Квазіньютонівські методи.....	44
4. Інтерполяція функцій.....	48
4.1. Інтерполяційна формула Лагранжа.....	48
4.2. Інтерполяційна формула Ньютона.....	49
4.3. Похибка поліноміальної інтерполяції.....	50
4.4. Інтерполяція сплайнами .....	51
5. Числове інтегрування функцій .....	56
5.1. Загальна похибка числового інтегрування.....	57
5.2. Формули Ньютона – Котеса .....	58
5.3. Формули Чебишова .....	61
5.4. Формули Гаусса .....	63
5.5. Апостеріорна оцінка похибки інтегрування .....	64
6. Числове інтегрування звичайних диференціальних рівнянь.....	67
6.1. Метод Ейлера .....	70
6.2. Методи Рунге – Кутта .....	73
6.3. Багатоточкові методи .....	78
6.4. Апостеріорна оцінка похибки розв'язання задачі Коші. Автоматичний вибір кроку інтегрування.....	83

6.5. Жорсткі рівняння .....	87
6.6. Крайові задачі .....	89
6.6.1. Зведення крайових задач до задач Коші .....	90
6.6.2. Метод скінченних різниць .....	92
6.6.3. Проекційні методи розв'язання крайових задач .....	94
6.6.4. Метод скінченних елементів розв'язання крайових задач... ..	98
7. Числове розв'язання задач оптимізації .....	108
7.1. Методи одновимірного пошуку .....	110
7.2. Методи багатовимірного пошуку .....	113
7.3. Градієнтні методи .....	114
7.4. Метод спряжених градієнтів .....	116
7.5. Метод Ньютона .....	118
7.6. Методи змінної метрики .....	120
(квазіньютонівські методи) .....	120
7.7. Методи розв'язання задач умовної оптимізації .....	120
8. Апроксимація функцій .....	128
8.1. Лінійна задача .....	130
про середньоквадратичне наближення .....	130
Список використаної та рекомендованої літератури .....	137

## 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 числовими методами розв'язання задач аналізу і проектування в електроніці, отримання навиків реалізації числових 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 алгебричних рівнянь (СЛАР) – погано зумовленими. Тому основну увагу в курсі приділяють тим методам обчислювальної математики, які пристосовані до задач електроніки.

Жоден з існуючих методів обчислювальної математики не задовольняє висунуті до них вимоги, тому реальні обчислювальні алгоритми, як правило, будують комбінуванням різних методів.

# 1. Похибки обчислень

## 1.1. Поняття обчислювального експерименту

У зв'язку зі швидким розвитком обчислювальної техніки широкого розповсюдження під час проведення наукових досліджень та інженерного проектування набув обчислювальний експеримент. Він ґрунтується на побудові та аналізі за допомогою комп'ютера математичних моделей досліджуваного об'єкта.

Розглянемо схему обчислювального експерименту (рис. 1.1).

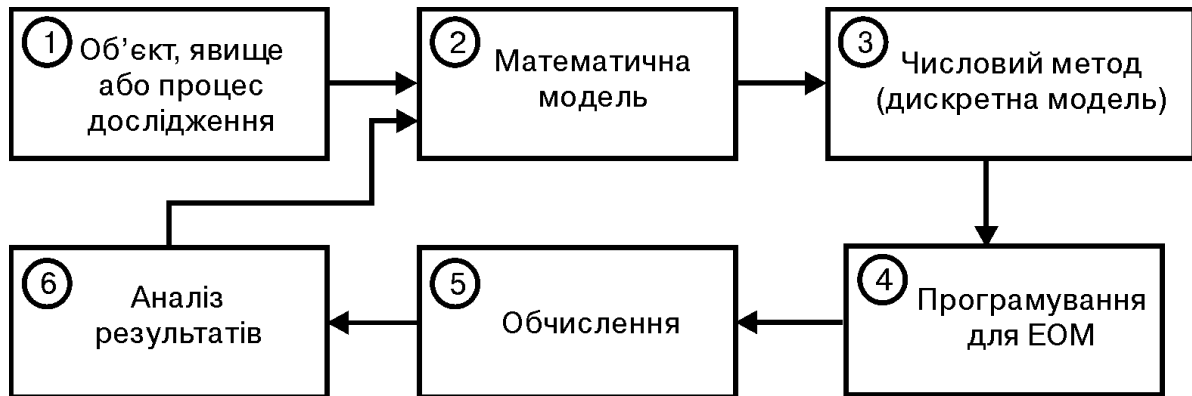


Рис. 1.1. Схема обчислювального експерименту

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 являє 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 вдається 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 настає 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 потрібною точністю.

Обчислювальна математика забезпечує лише один з етапів обчислювального експерименту, а саме етап вибору (побудови) обчислювального методу, від правильного вибору 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 точкою. Така форма відповідає позиційній системі з основою  $r$  і будь-який запис

$$a = \pm a_n a_{n-1} \dots a_0 \cdot a_{-1} a_{-2} \dots a_{-m}$$

означає, що

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

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

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

$$a = \pm M r^p, \quad (1.1)$$

(як правило,  $r = 2$ ).

Число  $M$  записують у формі числа з фіксованою точкою і називають мантиєю числа  $a$ , причому

$$r^{-1} \leq |M| < 1.$$

Таке подання чисел з плаваючою точкою 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. Через скінченність розрядної сітки порядку діапазон чисел, з якими оперує обчислювальна система, обмежений

$$|a| \in [M_0, M_\infty],$$

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

Число  $a$ , яке не можна записати в обчислювальній системі точно, округлюється, тобто замінюється близьким йому числом  $\tilde{a}$ , яке подається точно в обчислювальній системі. Точність подання чисел із плаваючою точкою в обчислювальній системі характеризується відносною похибкою



$$\varepsilon_{\text{маш}} = \left| \frac{\tilde{a} - a}{\tilde{a}} \right|.$$

За найпростішого способу округлення – відкидання всіх розрядів мантиси, які виходять за межі розрядної сітки, відносну похибку округлення можна оцінити як [8]:

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

Оскільки при використанні нормалізованої форми запису мантиси старший біт завжди дорівнює одиниці, то цей біт не запам'ятовують, хоча його присутність завжди мають на увазі при обчисленнях. Це звільняє місце для збереження додаткового біту. Тому в більшості обчислювальних систем мантиса обмежується як

$$1 \leq |M| < r,$$

але старший біт не зберігається. Оскільки при цьому кількість дійсно збережених бітів збільшується на одиницю, то реальна відносна машина похибка оцінюється як:

$$\varepsilon_{\text{маш}} \leq 2^{-t}.$$

### 1.3. Обчислювальна похибка визначення значення функції

Нехай необхідно обчислити значення функції  $y = f(x_1, x_2, \dots, x_n)$  від  $n$  змінних. Кожен з аргументів функції  $x_i, i = 1, \dots, n$  подано в обчислювальній системі як число  $\tilde{x}_i$  з абсолютною похибкою округлення  $\Delta x_i \geq (\tilde{x}_i - x_i), i = 1, \dots, n$ . Нехай  $\tilde{y} = f(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$  – обчислене значення функції. Визначимо, наскільки обчислене значення  $\tilde{y}$  відрізняється від точного  $y$ .

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

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

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

Граничною відносною похибкою називають значення

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

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

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

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

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

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

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

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 аргумента.

### 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	$f(x_1, x_2)$
1	$e^{x_1^2} x_2^3$
2	$10^{x_1} \operatorname{tg}(x_2^5)$
3	$x_1^{x_2} \cos(x_2^5)$
4	$x_1^{\cos(x_2)} x_2^5$
5	$x_1^{\log_2(5x_2^2)} \sin(x_2^3)$
6	$e^{x_1 x_2} \cos(x_1 x_2)$
7	$\operatorname{tg}(x_1 x_2) x_2^{3x_1}$

8	$e^{\cos^2(x_1)} x_2^5$
9	$x_1^2 (1 + \ln x_2)$
10	$e^{x_1+x_2} \sqrt{x_2^5}$
11	$e^{\sin(x_1)} \operatorname{tg}(x_2^2)$
12	$\frac{e^{x_1}}{\sin(x_2)}$
13	$\frac{x_1^5}{1 + \sin(x_2)}$
14	$e^{\frac{x_1}{1+\cos(x_2)}}$
15	$\ln(1 + \cos(x_1)) x_2^3$

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	$f(x_1, x_2)$
16	$x_1^2 (1 + \ln x_2)$
17	$x_1^2 (1 + \sqrt{x_2})$
18	$\sin(x_1)(1 + \cos(x_2))$
19	$\operatorname{tg}(x_1) e^{x_2^2}$
20	$2^{x_1} \cos(x_2)$
21	$x_1^3 e^{\sqrt{x_2}}$
22	$\sqrt{x_1^3 + 3 \cos(e^{x_2})}$
23	$\ln(\cos(x_1)) x_2^3$
24	$x_1^{x_2^2} \ln(x_2^3)$
25	$\cos(\sqrt{x_1^3}) e^{x_2}$
26	$\operatorname{tg}(x_1^3) e^{2x_2}$
27	$\frac{\operatorname{tg}(x_1^3)}{1 + \cos(x_2)}$

28	$e^{x_1} \sqrt{5 + x_2^5}$
29	$\ln(\cos(x_1))(5 + x_2^5)$
30	$x_1^{\cos(x_2)} x_2^3$

## 2. Числові методи лінійної алгебри

Успішний розв'язок багатьох наукових задач значною мірою залежить від уміння швидко і точно розв'язувати системи лінійних рівнянь.

Система лінійних алгебричних рівнянь (СЛАР) у загальному випадку має вигляд:

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

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

Уводячи матрицю коефіцієнтів системи

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

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

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

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

На сьогодні розроблено багато методів розв'язування СЛАР. Розглянемо лише три з них, а саме: метод **LU**-факторизації, метод **QR**-факторизації та метод Холеського.

Такий вибір зумовлено тим, що ці методи добре адаптовані для використання обчислювальною технікою, що робить їх дуже ефективними та актуальними в сучасних умовах [2].

Усі названі методи полягають у зведенні матриці системи  $\mathbf{A}$  до добутку більш простих матриць спеціального виду (трикутних, діагональних, блоково-діагональних). Розглянемо кожен з цих методів окремо.

### 2.1. Метод LU-факторизації

У цьому методі матриця коефіцієнтів  $\mathbf{A}$  подається у вигляді добутку матриць  $\mathbf{L}$  та  $\mathbf{U}$ :

$$\mathbf{A} = \mathbf{LU}, \tag{2.2}$$

де  $\mathbf{L}$  – нижня трикутна матриця; а  $\mathbf{U}$  – верхня трикутна матриця, усі діагональні елементи якої дорівнюють 1:

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ & & \dots & \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 1 & u_{12} & \dots & u_{1n} \\ 0 & 1 & \dots & u_{2n} \\ & & \dots & \\ 0 & 0 & \dots & 1 \end{bmatrix}. \quad (2.3)$$

Метод  $\mathbf{LU}$ -факторизації складається з двох етапів:

- 1) етапу факторизації матриці  $\mathbf{A}$ ;
- 2) етапу отримання розв'язку  $\mathbf{X}$ .

Вектор  $\mathbf{V}$  у процесі факторизації не змінюється.

Із виразів (2.2) та (2.3) випливає, що для  $i > j$  елементи матриці  $\mathbf{A}$  можна виразити через елементи матриць  $\mathbf{L}$  та  $\mathbf{U}$ :

$$a_{ij} = \left( \sum_{k=1}^{j-1} l_{ik} u_{kj} \right) + l_{ij}; \quad (2.4)$$

$$a_{ji} = \left( \sum_{k=1}^{j-1} l_{jk} u_{ki} \right) + l_{jj} u_{ji}.$$

Тоді із рівнянь (2.4) випливає, що факторизація матриці  $\mathbf{A}$  виконується за  $n$  стадій. На кожній  $j$ -й стадії послідовно обчислюють елементи  $l_{ij}$  чергового  $j$ -го стовпця матриці  $\mathbf{L}$  за формулою

$$l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj}, \quad i = \overline{j, n}, \quad (2.5)$$

та елементи  $u_{ji}$  чергового  $j$ -го рядка матриці  $\mathbf{U}$  за формулою

$$u_{ji} = \frac{a_{ji} - \sum_{k=1}^{j-1} l_{jk} u_{ki}}{l_{jj}}, \quad i = \overline{j+1, n}. \quad (2.6)$$

Із формул (2.5), (2.6) випливає, що на першій стадії елементи  $l_{i1} = a_{i1}, i = \overline{1, n}$ , а  $u_{1j} = \frac{a_{1j}}{l_{11}}, j = \overline{2, n}$ .

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

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

Після факторизації матриці  $\mathbf{A}$  система буде мати вигляд

$$\mathbf{LUX} = \mathbf{B}. \quad (2.7)$$

Позначимо  $\mathbf{Y} = \mathbf{UX}$ , тоді із (2.7) маємо рівняння

$$\mathbf{LY} = \mathbf{B}. \quad (2.8)$$

Оскільки  $\mathbf{L}$  – нижня трикутна матриця, то отримуємо розв’язок системи (2.8):

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

Після цього залишається розв’язати ще одну систему з верхньою трикутною матрицею  $\mathbf{U}$ , всі діагональні елементи якої дорівнюють 1:

$$\mathbf{UX} = \mathbf{Y}. \quad (2.9)$$

Розв’язавши систему (2.9), остаточно отримуємо шуканий вектор  $\mathbf{X}$ :

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

**Приклад:**

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

$$j=1 \begin{bmatrix} 9 & -\frac{2}{9} & -\frac{6}{9} \\ -2 & 16 & -12 \\ -6 & -12 & 18 \end{bmatrix}, j=2 \begin{bmatrix} 9 & -\frac{2}{9} & -\frac{2}{3} \\ -2 & \frac{140}{9} & -\frac{6}{7} \\ -6 & -\frac{40}{3} & 18 \end{bmatrix}, j=3 \begin{bmatrix} 9 & -\frac{2}{9} & -\frac{2}{3} \\ -2 & \frac{140}{9} & -\frac{6}{7} \\ -6 & -\frac{40}{3} & \frac{18}{7} \end{bmatrix};$$

$$\mathbf{A}_{\text{LU}} = \begin{bmatrix} 9 & -\frac{2}{9} & -\frac{2}{3} \\ -2 & \frac{140}{9} & -\frac{6}{7} \\ -6 & -\frac{40}{3} & \frac{18}{7} \end{bmatrix} \Rightarrow \mathbf{L} = \begin{bmatrix} 9 & 0 & 0 \\ -2 & \frac{140}{9} & 0 \\ -6 & -\frac{40}{3} & \frac{18}{7} \end{bmatrix}; \quad \mathbf{U} = \begin{bmatrix} 1 & -\frac{2}{9} & -\frac{2}{3} \\ 0 & 1 & -\frac{6}{7} \\ 0 & 0 & 1 \end{bmatrix};$$

$$\mathbf{LY} = \mathbf{B} \Rightarrow \begin{bmatrix} 9 & 0 & 0 \\ -2 & \frac{140}{9} & 0 \\ -6 & -\frac{40}{3} & \frac{18}{7} \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -13 \\ -6 \\ 24 \end{bmatrix},$$

$$y_1 = \frac{-13}{9}; \quad y_2 = \left( -6 - (-2) \times \frac{-13}{9} \right) \times \frac{9}{140} = -\frac{4}{7};$$

$$y_3 = \left( 24 - (-6) \times \frac{-13}{9} - \frac{-40 \cdot (-4)}{3 \cdot 7} \right) \times \frac{7}{18} = \frac{54}{7} \times \frac{7}{18} = 3;$$

$$\begin{bmatrix} 1 & -\frac{2}{9} & -\frac{2}{3} \\ 0 & 1 & -\frac{6}{7} \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -\frac{13}{9} \\ -\frac{4}{7} \\ 3 \end{bmatrix};$$

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

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

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 стовпців) найбільшого за модулем ведучого елемента  $l_{jj}$ . При цьому матрицю  $\mathbf{A}$  можна подати у вигляді  $\mathbf{A} = \mathbf{PLU}$ , де  $\mathbf{P}$  – матриця перестановок.

## 2.2. Метод QR-факторизації

З курсу лінійної алгебри відомо, що будь-яку невинроджену матрицю можна подати у вигляді добутку ортогональної матриці  $\mathbf{Q}$  та верхньої трикутної матриці  $\mathbf{R}$ :

$$\mathbf{A} = \mathbf{QR}. \quad (2.10)$$



Як матрицю  $\mathbf{Q}$  найчастіше використовують матрицю відбиття або матрицю обертання. У першому випадку метод **QR**-факторизації називають методом відбиття, а в другому методом обертання.

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

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

Кожна матриця  $\mathbf{Q}_j$  при множенні обнуляє елементи  $j$ -го стовпця, що знаходяться нижче головної діагоналі, залишаючи без змін перші  $j-1$  стовпців.

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

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

у вектор, паралельний вектору  $\mathbf{Y} = [z_1, z_2, \dots, z_{j-1}, \alpha_j, 0 \dots 0]^T$ , де

$$\alpha_j = \pm \sqrt{\sum_{i=j}^n (a_{ij}^{(j-1)})^2}, \text{ а знак } \alpha_j \text{ вибирається з умови } \alpha_j z_j < 0.$$

Уведемо вектор

$$\mathbf{S}_j = [0, 0, \dots, z_j - \alpha_j, z_{j+1}, \dots, z_n]^T. \quad (2.11)$$

Тоді, враховуючи (2.11), маємо

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

де  $k_j = (a_j^2 + |\alpha_j z_j|)^{-1}$ .

Матрицю  $\mathbf{Q}$  у виразі (2.10) можна знайти як транспонований добуток елементарних матриць  $\mathbf{Q}_j$ . Однак на практиці матрицю  $\mathbf{Q}$  явно не обчислюють, а зводять систему до вигляду

$$\mathbf{A}_{n-1} \mathbf{X} = \mathbf{B}_{n-1}, \quad (2.12)$$

де

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

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

Розв'язок системи (2.12) знаходять з виразу

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

**Приклад:**

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

$$\mathbf{A}_0 = \mathbf{A} = \begin{bmatrix} 9 & -2 & -6 \\ -2 & 16 & -12 \\ -6 & -12 & 18 \end{bmatrix},$$

$$j = 1; \alpha_1^2 = 81 + 4 + 36 = 121; z_1 = 9 > 0 \Rightarrow \alpha_1 = -\sqrt{121} = -11;$$

$$k_1 = \frac{1}{121 + 99} = \frac{1}{220}; \mathbf{S}_1^T = [20; -2; -6].$$

$$\mathbf{S}_1^T \mathbf{A} = [20; -2; -6] \begin{bmatrix} 9 & -2 & -6 \\ -2 & 16 & -12 \\ -6 & -12 & 18 \end{bmatrix} = [220; 0; -204].$$

$$\mathbf{A}_1 = \begin{bmatrix} 9 & -2 & -6 \\ -2 & 16 & -12 \\ -6 & -12 & 18 \end{bmatrix} - \frac{1}{220} \begin{bmatrix} 20 \\ -2 \\ -6 \end{bmatrix} [220; 0; -204] = \begin{bmatrix} -11 & -2 & \frac{138}{11} \\ 0 & 16 & -\frac{762}{55} \\ 0 & -12 & -\frac{684}{55} \end{bmatrix},$$

$$j = 2 \quad \alpha_2^2 = 256 + 144 = 400;$$

$$z_2 = 16 > 0 \Rightarrow \alpha_2 = -\sqrt{400} = -20;$$

$$k_2 = \frac{1}{400 + 320} = \frac{1}{720};$$

$$\mathbf{S}_2^T = [0; 36; -12].$$

$$\mathbf{S}_2^T \mathbf{A}_1 = [0; 36; -12] \begin{bmatrix} -11 & -2 & \frac{138}{11} \\ 0 & -16 & -\frac{762}{55} \\ 0 & -12 & \frac{684}{55} \end{bmatrix} = [0; 720; -648].$$

$$\mathbf{A}_2 = \begin{bmatrix} -11 & -2 & \frac{138}{11} \\ 0 & -16 & -\frac{762}{55} \\ 0 & -12 & \frac{684}{55} \end{bmatrix} - \frac{1}{720} \begin{bmatrix} 0 \\ 36 \\ -12 \end{bmatrix} [0; 720; -648] = \begin{bmatrix} -11 & -2 & \frac{138}{11} \\ 0 & -20 & \frac{204}{11} \\ 0 & 0 & \frac{18}{11} \end{bmatrix},$$

$$\mathbf{S}_1^T \mathbf{B} = [20; -2; -6] \begin{bmatrix} -13 \\ -6 \\ 24 \end{bmatrix} = -392, \quad \mathbf{B}_1 = \begin{bmatrix} -13 \\ -6 \\ 24 \end{bmatrix} + \frac{392}{220} \begin{bmatrix} 20 \\ -2 \\ -6 \end{bmatrix} = \begin{bmatrix} \frac{249}{11} \\ -\frac{526}{55} \\ \frac{732}{55} \end{bmatrix};$$

$$\mathbf{S}_2^T \mathbf{B}_1 = [0; 36; -12] \begin{bmatrix} \frac{249}{11} \\ -\frac{526}{55} \\ \frac{732}{55} \\ \frac{54}{11} \end{bmatrix} = -504, \quad \mathbf{B}_2 = \begin{bmatrix} \frac{249}{11} \\ -\frac{526}{55} \\ \frac{732}{55} \\ \frac{54}{11} \end{bmatrix} + \frac{504}{720} \begin{bmatrix} 0 \\ 36 \\ -12 \end{bmatrix} = \begin{bmatrix} \frac{249}{11} \\ \frac{172}{11} \\ \frac{54}{11} \\ \frac{54}{11} \end{bmatrix};$$

$$\begin{bmatrix} -11 & -2 & \frac{138}{11} \\ 0 & -20 & \frac{204}{11} \\ 0 & 0 & \frac{18}{11} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{249}{11} \\ \frac{172}{11} \\ \frac{54}{11} \end{bmatrix} \Rightarrow x_3 = \frac{54/\cancel{11}}{18/\cancel{11}} = 3; \quad x_2 = \frac{\frac{172}{11} - \frac{204}{11} \cdot 2}{-20} = 2;$$

$$x_1 = \frac{\frac{249}{11} + 2 \cdot 2 - \frac{138}{11} \cdot 3}{-11} = 1 \Rightarrow \mathbf{X} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

### 2.3. Метод Холеського

Метод використовують для розв'язування СЛАР із симетричною матрицею  $\mathbf{A}$  [3]. Він ґрунтується на розкладанні матриці  $\mathbf{A}$  у добуток

$$\mathbf{A} = \mathbf{LDL}^T, \quad (2.13)$$

де  $\mathbf{L}$  – нижня трикутна матриця;  $\mathbf{D}$  – діагональна матриця.

Із виразу (2.13) маємо

$$a_{ij} = \sum_{k=1}^i l_{ik} d_{kk} l_{jk}, \quad i \geq j.$$

Для  $i = j$  одержуємо

$$d_{jj} l_{jj}^2 = a_{jj} - \sum_{k=1}^{j-1} (l_{jk}^2 d_{kk}). \quad (2.14)$$

Нехай  $l_{ii} = 1$ ,  $i = \overline{1, n}$ , тоді з виразу (2.14)

$$d_{jj} = a_{jj} - \sum_{k=1}^{j-1} (l_{jk}^2 d_{kk}), \quad j = \overline{1, n}. \quad (2.15)$$

Якщо  $i > j$ , то

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} d_{kk} l_{jk}}{d_{jj}}, \quad j = \overline{1, n}, i = \overline{j+1, n}. \quad (2.16)$$

Таким чином,  $\mathbf{LDL}^T$  – факторизація матриці, яка виконується за  $n$  стадій. На кожній  $j$ -й стадії вважаємо, що  $l_{jj} = 1$ . За формулою (2.15) спочатку обчислюємо елемент  $d_{jj}$ , а потім елементи  $l_{ij}$  чергового  $j$ -го стовпця за виразом (2.16).

Після факторизації матриці  $\mathbf{A}$  розв'язуємо СЛАР  $\mathbf{LY} = \mathbf{B}$  за допомогою прямої підстановки

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

Потім розв'язуємо СЛАР  $\mathbf{DZ} = \mathbf{Y}$ :

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

Використовуючи вираз, знаходимо  $\mathbf{X}$  із системи  $\mathbf{L}^T \mathbf{X} = \mathbf{Z}$ .

$$x_i = z_i - \sum_{j=i+1}^n (l_{ji} x_j), \quad i = \overline{n, 1}.$$

**Приклад:**

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

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

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

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

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

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

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

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

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

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

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

Як уже відзначалося, цей метод може використовуватися для СЛАР із симетричною матрицею  $\mathbf{A}$ . Якщо ж матриця  $\mathbf{A}$  ще й додатно-визначена, то схему розв'язку можна спростити, використовуючи  $\mathbf{LL}^T$ -факторизацію

$$\mathbf{A} = \mathbf{LL}^T . \quad (2.17)$$

Тоді з виразу (2.17) знаходимо

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

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

Після факторизації матриці  $\mathbf{A}$  розв'язок задачі зводимо до розв'язку двох простих СЛАР з трикутними матрицями  $\mathbf{LY} = \mathbf{B}$ ,  $\mathbf{L}^T \mathbf{X} = \mathbf{Y}$  за формулами:

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

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

Розв'яжемо цим методом  $\mathbf{LL}^T$  СЛАР із попереднього прикладу:

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

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

$$j=2:l_{22} = \sqrt{16 - \frac{4}{9}} = \sqrt{\frac{140}{9}} = \frac{2\sqrt{35}}{3}; \quad l_{32} = \frac{-12 - \frac{4}{3}}{2\sqrt{35}} = -\frac{20}{\sqrt{35}};$$

$$j=3:l_{33} = \sqrt{18 - 4 - \frac{1600}{140}} = \sqrt{\frac{98}{7} - \frac{80}{7}} = 3\sqrt{\frac{2}{7}}.$$

$$\mathbf{L} = \begin{bmatrix} 3 & 0 & 0 \\ -\frac{2}{3} & \frac{2\sqrt{35}}{3} & 0 \\ -2 & -\frac{20}{\sqrt{35}} & 3\sqrt{\frac{2}{7}} \end{bmatrix}.$$

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

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

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

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

## 2.4. Похибка розв'язання СЛАР

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

Можна показати, 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  $\mathbf{A}$ ;  $\|\mathbf{A}^{-1}\|$  – норма матриці, оберненої до  $\mathbf{A}$ .

Добуток  $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$  називають числом обумовленості матриці  $\mathbf{A}$ :

$$\text{cond}(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|. \quad (2.18)$$

Використовуючи в (2.18) евклідову норму, маємо  $\text{cond}$  – число обумовленості

$$\text{cond}(\mathbf{A}) = \sqrt{\frac{\max_i |\mu_i|}{\min_i |\mu_i|}},$$

де  $\mu_i$  – власні числа матриці  $\mathbf{A}^T \mathbf{A}$ .

Якщо матриця  $\mathbf{A}$  – симетрична, то

$$\text{cond}(\mathbf{A}) = \frac{\max_i |\lambda_i|}{\min_i |\lambda_i|},$$

де  $\lambda_i$  – власне число матриці  $\mathbf{A}$ .

Число обумовленості є мірою виродженості матриці. Якщо воно велике, то похибки округлення можуть суттєво вплинути на результат розв'язку СЛАР. Матриця з великим числом обумовленості називається погано обумовленою. На практиці вважають, що коли  $\text{cond}(\mathbf{A}) \geq \frac{1}{\sqrt{\varepsilon}}$  ( $\varepsilon$  – відносна

похибка округлення), то обчислений розв'язок є ненадійним.

У такому випадку необхідно провести регуляризацію системи, наприклад, методом Тихонова

$$(\mathbf{A}^T \mathbf{A} + \alpha \mathbf{E})\mathbf{X} = \mathbf{A}^T \mathbf{B}, \quad (2.19)$$

де  $\mathbf{E}$  – одинична матриця;  $\alpha$  – параметр регуляризації, деяке досить мале значення. Число  $\alpha$  повинно бути, з одного боку, досить великим, щоб суттєво зменшити число обумовленості матриці СЛАР, а з другого боку, досить малим, щоб СЛАР (2.19) мало відрізнялася від початкової системи.



## Контрольне завдання 2.1

1. Вибрати СЛАР відповідно до свого варіанта.
2. Оцінити обумовленість матриці системи.
3. Розв'язати вибрану СЛАР методом LU-факторизації.
4. Розв'язати вибрану СЛАР методом QR-факторизації.
5. Розв'язати вибрану СЛАР методом Холеського.
6. Порівняти результати розв'язування різними методами. Оцінити відхилення розв'язку.

### Варіанти завдань

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$26. \begin{cases} x_1 + 4x_2 + 8x_3 = 123; \\ 4x_1 + 9x_2 + 4x_3 = 149; \\ 8x_1 + 4x_2 + x_3 = 102. \end{cases}$$

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

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

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

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

## 2.5. Власні числа і власні вектори

У багатьох наукових задачах досить часто виникає потреба у знаходженні власних значень і відповідних їм власних векторів. Наприклад, аналізуючи динамічні системи, власні значення визначають частоти коливань, а власні вектори характеризують їх форму.

Розглянемо квадратну матрицю  $n$ -го порядку:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ & & \dots & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}. \quad (2.20)$$

Власні числа  $\lambda_i, (i = \overline{1, n})$  квадратної матриці  $\mathbf{A}$  є дійсними або комплексними числами, що задовольняють умову

$$(\mathbf{A} - \lambda_i \mathbf{E}) \Psi_i = 0, \quad (2.21)$$

де  $\mathbf{E}$  – одинична матриця;  $\Psi_i$  – власний вектор матриці  $\mathbf{A}$ , що відповідає деякому власному числу  $\lambda_i$ .

Із рівнянь (2.20), (2.21) випливає, що власні числа можна визначити з характеристичного рівняння:

$$\det \begin{pmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ & & \dots & \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{pmatrix} = 0.$$

Визначивши власні числа, з умови (2.21) можна знайти власні вектори. Однак у разі великих розмірностей матриці  $\mathbf{A}$  складність розрахунку значно зростає. Тому розроблено ряд спеціальних методів, що полегшують розв'язання задачі на власні числа і власні вектори [3].

## 2.6. Степеновий метод

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

Розглянемо типову задачу відшукування максимального за модулем власного значення. Нехай  $\{\Psi_i\}$  – система власних векторів, а  $\{\lambda_i\}$  – система власних значень матриці  $\mathbf{A}$ :

$$\mathbf{A} \Psi_i = \lambda_i \Psi_i. \quad (2.22)$$

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

$$|\lambda_1| \leq |\lambda_2| \leq |\lambda_3| \leq \dots \leq |\lambda_{n-1}| \leq |\lambda_n|.$$

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

$$\mathbf{X}_m = \mathbf{A} \mathbf{X}_{m-1}, m = 1, 2, \dots \quad (2.23)$$

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

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

де  $c_i$  – коефіцієнти розкладання. Тоді з (2.22), (2.23) маємо

$$\mathbf{X}_m = \sum_{i=1}^n c_i \mathbf{A}^m \Psi_i = \sum_{i=1}^n c_i \lambda_i^m \Psi_i. \quad (2.24)$$

З огляду на те, що  $|\lambda_n| > |\lambda_{n-1}|$ , для досить великих  $m$  з (2.24) маємо

$$\mathbf{X}_m = \lambda_n^m \left( c_n \Psi_n + o\left(\frac{\lambda_{n-1}}{\lambda_n}\right)^m \right). \quad (2.25)$$

Із виразу (2.25) випливає, що якщо  $c_n \neq 0$ , то при досить великому  $m$

$$|\lambda_n| \approx \frac{\|\mathbf{X}_m\|}{\|\mathbf{X}_{m-1}\|}.$$

Отже, максимальне за модулем власне значення  $\lambda_n$  можна знайти з ітераційного процесу (2.23). При цьому  $m$ -те наближення  $\lambda_n^{(m)}$  до власного значення  $\lambda_n$  оцінюють як

$$|\lambda_n^{(m)}| = \frac{\|\mathbf{X}_m\|}{\|\mathbf{X}_{m-1}\|},$$

а  $n$ -те наближення до власного вектора  $\Psi_n^{(m)}$ :

$$\Psi_n^{(m)} = \frac{\mathbf{X}_m}{\|\mathbf{X}_m\|}. \quad (2.26)$$

Критерієм зупинки ітераційного процесу (2.23) є виконання умови

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

де  $\varepsilon$  – задане значення відносної похибки.

Якщо треба знати знак  $\lambda_n$ , то його знаходять за формулою

$$\text{sign} \lambda_n = \text{sign} \left( x_i^{(m)} \cdot x_i^{(m-1)} \right), \quad \forall i : x_i^{(m)} \neq 0; \quad x_i^{(m-1)} \neq 0.$$

Із виразів (2.21) та (2.24) випливає, що швидкість збіжності ітераційного процесу лінійна і визначається відношенням  $\left| \frac{\lambda_n}{\lambda_{n-1}} \right|$ , оскільки

$$\lambda_n^{(m)} = \lambda_n + 0 \left( \frac{\lambda_{n-1}}{\lambda_n} \right)^m.$$

Для стійкості числового процесу необхідно час від часу нормувати вектор  $\mathbf{X}_m$ , щоб  $\|\mathbf{X}_m\| = 1$ . Інакше, якщо  $|\lambda_n| > 1$ , то, як випливає з виразу (2.25),  $\|\mathbf{X}_m\| \rightarrow \infty$  якщо  $m \rightarrow \infty$ . Тому за досить великого  $m$  може відбутися переповнення розрядної сітки обчислювальної системи. Якщо  $|\lambda_n| < 1$ , то  $\|\mathbf{X}_m\| \rightarrow 0$  і внаслідок скінченності розрядної сітки обчислювальної системи може статися, що починаючи з деякого  $m$   $\mathbf{X}_m = 0$ .

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

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

$$\mathbf{A}\mathbf{X}_m = \mathbf{X}_{m-1}.$$

При цьому  $m$ -те наближення  $\lambda_{\min}^{(m)}$  до власного значення  $\lambda_{\min}$  можна оцінити так:

$$|\lambda_{\min}^{(m)}| = 1 / \lambda_{\max \text{ inv}}^{(m)} = \frac{\|\mathbf{X}_{m-1}\|}{\|\mathbf{X}_m\|}.$$

**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  $\Psi_1 = [1 \ 1]^T$ ,  $\Psi_2 = [1 \ -1]^T$ . Спочатку знаходимо максимальне за модулем власне число та відповідний вектор.

Нехай  $\mathbf{X}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ . Тоді згідно з виразом (2.23) маємо:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\mathbf{A}\mathbf{X}_3 = \mathbf{X}_2 \Rightarrow \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix} \mathbf{X}_3 = \begin{bmatrix} 101/800 \\ 99/800 \end{bmatrix} \Rightarrow \mathbf{X}_3 = \begin{bmatrix} 1001/16000 \\ 999/16000 \end{bmatrix},$$

$$\lambda_{\min}^{(3)} = \frac{\|\mathbf{X}_2\|_2}{\|\mathbf{X}_3\|_2} = \frac{\sqrt{(101/800)^2 + (99/800)^2}}{\sqrt{(1001/16000)^2 + (999/16000)^2}} \approx 2,0001,$$

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

$$\Psi_{\min} \approx \frac{\mathbf{X}_3}{\|\mathbf{X}_3\|_{\infty}} = \begin{bmatrix} 1001/16000 \\ 999/16000 \\ 1001/16000 \end{bmatrix} \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

## 2.7. Метод скалярного добутку

Для симетричної матриці  $\mathbf{A}$  найбільше за модулем власне число доцільно шукати з ітераційного процесу:

$$\mathbf{X}_m = \mathbf{A}\mathbf{X}_{m-1}, \quad \lambda_{\max}^{(m)} = \frac{\mathbf{X}_m^T \mathbf{X}_m}{\mathbf{X}_{m-1}^T \mathbf{X}_{m-1}}, \quad \Psi_{\max}^{(m)} = \frac{\mathbf{X}_m}{\|\mathbf{X}_m\|}, \quad (2.27)$$

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

Найменше за модулем власне число симетричної матриці  $\mathbf{A}$  шукають з ітераційного процесу:

$$\mathbf{A}\mathbf{X}_m = \mathbf{X}_{m-1}, \quad \lambda_{\min}^{(m)} = \frac{\mathbf{X}_{m-1}^T \mathbf{X}_m}{\mathbf{X}_m^T \mathbf{X}_m}, \quad \Psi_{\min}^{(m)} = \frac{\mathbf{X}_m}{\|\mathbf{X}_m\|},$$

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

Можна показати, що метод скалярного добутку має квадратичну збіжність. 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 вектор.

Нехай  $\mathbf{X}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ . Тоді згідно з (2.27) маємо:

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

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

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

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

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

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



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

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

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

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

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

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

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

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

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

$$\lambda_{\min}^{(3)} = \frac{\mathbf{X}_2^T \mathbf{X}_3}{\mathbf{X}_3^T \mathbf{X}_3} = \frac{\begin{bmatrix} 101/800 & 99/800 \end{bmatrix} \times \begin{bmatrix} 1001/16000 \\ 999/16000 \end{bmatrix}}{\begin{bmatrix} 1001/16000 & 999/16000 \end{bmatrix} \times \begin{bmatrix} 1001/16000 \\ 999/16000 \end{bmatrix}} =$$

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

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

$$\Psi_{\min} \approx \frac{\mathbf{X}_3}{\|\mathbf{X}_3\|_{\infty}} = \frac{\begin{bmatrix} 1001/16000 \\ 999/16000 \end{bmatrix}}{1001/16000} \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

## 2.8. Метод Віландта

Часто в практичних задачах виникає необхідність знайти таке власне число, яке є найближчим до деякого заданого числа  $\mu$ . У такому випадку можна розв'язати задачу за допомогою методу Віландта. Суть методу полягає в тому, що спочатку знаходять деяку додаткову матрицю  $\tilde{\mathbf{A}} = (\mathbf{A} - \mu \mathbf{E})^{-1}$ , де  $\mathbf{E}$  – одинична матриця. Після цього визначають максимальне за модулем власне число матриці  $\nu$ . Це можна зробити будь-яким відомим методом (наприклад, методом скалярних добутків). Остаточного шукане власне число  $\lambda$  обчислюють за формулою

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

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

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

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

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

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

$$v^{(2)} = \frac{\mathbf{X}_2^T \mathbf{X}_2}{\mathbf{X}_1^T \mathbf{X}_2} = \frac{\begin{bmatrix} \frac{65}{512} & \frac{63}{512} \end{bmatrix} \times \begin{bmatrix} \frac{65}{512} \\ \frac{63}{512} \end{bmatrix}}{\begin{bmatrix} -\frac{7}{32} & -\frac{9}{32} \end{bmatrix} \times \begin{bmatrix} \frac{65}{512} \\ \frac{63}{512} \end{bmatrix}} = \frac{4097/131072}{-511/8192} \approx -0,5011.$$

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

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

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

### Контрольне завдання 2.2

1. Вибрати матрицю  $\mathbf{A}$  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  $\mu$  з похибкою не більше 1 %. Значення числа  $\mu$  вибрати з табл. 2.1 відповідно до варіанта.

4. Порівняти результати розв'язку різними методами.

Таблиця 2.1. Значення числа  $\mu$

Варіант	$\mu$	Варіант	$\mu$	Варіант	$\mu$	Варіант	$\mu$	Варіант	$\mu$
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. Числові 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  $f(x)$ :

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

Рівняння (3.1), як правило, не допускає аналітичного розв'язання, і його розв'язують числовими методами [4]. 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. Найпростіший спосіб розв'язання задачі на цьому етапі – обчислення таблиці значень функції  $f(x)$  у заданих точках  $x_k \in [a, b]$ ,  $k = 0, 1, \dots, n$ . Якщо виявиться, що за деякого  $k$  числа  $f(x_k), f(x_{k+1})$  мають різні знаки і функція  $f(x)$  неперервна на відрізку  $[x_k, x_{k+1}]$ , то це означає, що на інтервалі  $[x_k, x_{k+1}]$  рівняння (3.1) має, принаймні, один дійсний корінь. При цьому, якщо похідна монотонна, то на вказаному проміжку існує лише один корінь. Потім інтервал  $[x_k, x_{k+1}]$  розбивають на дрібніші інтервали 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  $\{x_n\}$ :  $x_0, x_1, \dots, x_{n-1}$ , причому кожне нове значення аргументу обчислюють на підставі попередніх, тобто в загальному випадку

$$x_{n+1} = \Psi_n(x_{n-m}, x_{n-m+1}, \dots, x_n), \quad n \geq m, \quad (3.2)$$

де  $\Psi_n$  – деяка функція, що залежить від методу розв'язання. Ітераційний процес (3.2) називають  $m$ -кроковим.

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

$$\begin{aligned} f_1(x_1, x_2, \dots, x_n) &= 0; \\ f_2(x_1, x_2, \dots, x_n) &= 0; \\ &\dots \\ f_n(x_1, x_2, \dots, x_n) &= 0. \end{aligned} \quad (3.3)$$

Позначимо через  $\mathbf{X}$  вектор-стовпець  $\mathbf{X} = [x_1, x_2, \dots, x_n]^T$ . Тоді систему рівнянь (3.3) можна записати у вигляді:

$$\begin{aligned}
 f_1(\mathbf{X}) &= 0; \\
 f_2(\mathbf{X}) &= 0; \\
 &\dots \\
 f_n(\mathbf{X}) &= 0,
 \end{aligned}$$

де, уводячи для вектора-стовпця функцій  $[f_1, f_2, \dots, f_n]^T$  позначення  $\mathbf{F}$ , записують всю систему (3.3) у векторній формі

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

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

$$\mathbf{X}_n = \boldsymbol{\psi}_n(\mathbf{X}_{n-m}, \mathbf{X}_{n-m+1}, \dots, \mathbf{X}_{n-1}), \quad n > 0,$$

де  $\boldsymbol{\psi}_n$  – вектор-стовпець функцій, що залежить від методу розв'язання;  $\mathbf{X}_i$  –  $i$ -те наближення до розв'язку.

### 3.1. Метод бісекції (поділу навпіл)

Нехай  $f(x)$  – функція однієї дійсної змінної  $x$ . Щоб почати пошук нуля  $f(x)$ , припустимо, що можна знайти інтервал  $[a, b]$ , 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 не нуль, а малий інтервал  $[\alpha, \beta]$ , на якому  $f(x)$  змінює знак. Такий інтервал можна знайти і звужити настільки, наскільки дозволяє система чисел з плаваючою точкою, тобто так, щоб кінцями цього інтервалу були два сусідні числа цієї системи.

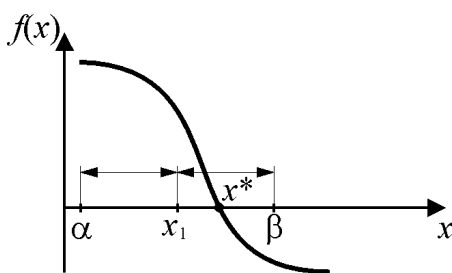


Рис. 3.1. Пошук кореня методом бісекції

У методі бісекції точка  $x_i$  розміщується у центрі інтервалу  $[\alpha, \beta]$  (рис.3.1) і як новий інтервал вибирають той з інтервалів  $[\alpha, x_i]$  і  $[x_i, \beta]$ , який містить корінь. Поділ відрізка пополам продовжують доти, доки довжина відрізка не буде меншою від заданої похибки обчислення. За корінь вибирають середину інтервалу, що залишився. Таким чином, якщо відомі такі  $a$  і  $b$ , що  $f(a)f(b) < 0$ , то за заданої відносної похибки

$\varepsilon$  для визначення кореня  $f(x)$  методу бісекції використовують алгоритм:

- 1) покласти  $\alpha = a, \beta = b$ ;
- 2) покласти  $x_i = \frac{\alpha + \beta}{2}$ ;
- 3) якщо  $f(\alpha)f(x_i) < 0$ , то  $\beta = x_i$  інакше  $\alpha = x_i$ ;

4) якщо  $\left| \frac{\beta - \alpha}{\alpha} \right| > \varepsilon$ , перейти до кроку 2, інакше  $x^* \approx \alpha$ ; 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  $f(\alpha)f(x_i) < 0$  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  $\frac{f(\alpha)}{|f(\alpha)|} f(x_i) < 0$ . Крім того, щоб уникнути переповнень, умову на кроці 4 подають у вигляді  $|\beta - \alpha| > \varepsilon|\alpha|$ .

Метод бісекції працює теоретично завжди. Тому його відносять до класу методів, які збігаються глобально. Дійсно, якщо тільки виконана умова  $f(\alpha)f(x_i) < 0$  і  $f(x)$  неперервна, то незалежно від поведінки  $f(x)$  метод бісекції теоретично завжди дозволить визначити значення кореня із заданою похибкою.

### 3.2. Метод Ньютона розв'язання рівнянь з однією змінною

Нехай  $x_n$  –  $n$ -те наближення розв'язання задачі (3.1). Розкладемо функцію  $f(x)$  у ряд Тейлора в околі точки  $x = x_n$  і обмежимося розглядом тільки двох членів ряду. Тоді

$$f(x) \approx f(x_n) + (x - x_n)f'(x_n). \quad (3.5)$$

Прирівнюючи відповідно до виразу (3.1) вираз (3.5) до нуля, знаходимо нове наближення до розв'язання

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \quad (3.6)$$

Метод Ньютона будується за ітераційною формулою (3.6).

Для установлення геометричного змісту методу Ньютона відзначимо, що вираз (3.5) є рівнянням дотичної до графіка функції  $y = f(x)$  у точці  $x_n$ . Отже, відповідно до формули (3.6) нове наближення  $x_{n+1}$  шукається як точка перетину дотичної до  $y = f(x)$  у точці  $x_n$  з віссю  $Ox$  (рис. 3.2). Тому метод Ньютона іноді називають *методом дотичних*.

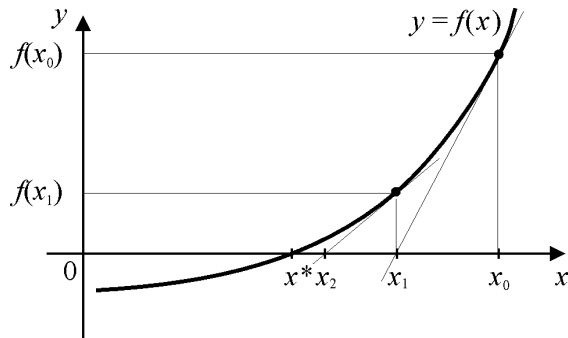


Рис. 3.2. Пошук кореня методом Ньютона

Можна довести, що метод Ньютона збігається, якщо

$$|x_0 - x^*|c \leq 1, \quad (3.7)$$

де  $c = \frac{\sup_x |f''(x)|}{2 \inf_x |f'(x)|}$  характеризує ступінь відносної нелінійності функції  $f(x)$ .

Як випливає з виразу (3.7), 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 передують який-небудь алгоритм типу бісекції, який сходиться глобально. Потім можна переключатися на ітерації методу Ньютона, які швидко збігаються.

Як приклад такого підходу розглянемо метод, що дає наочне уявлення про глобальні методи у  $n$ -мірному просторі. Припустимо, що метод Ньютона генерує не тільки крок  $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ , але й і напрям, на який цей крок указує (передбачається, що  $f'(x_n) \neq 0$ ). Хоча ньютонівський крок може привести до збільшення абсолютного значення функції, його напрям такий, що вздовж нього абсолютне значення функції зменшується (рис. 3.3).

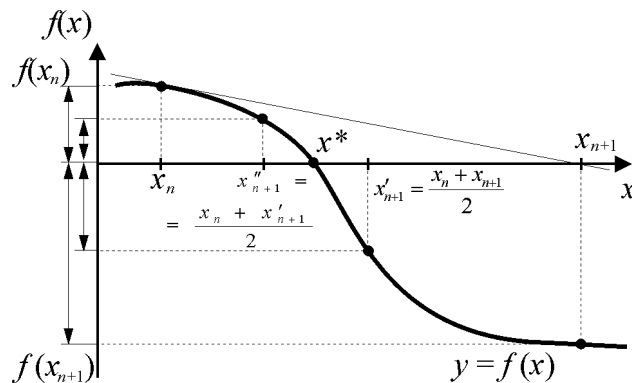


Рис. 3.3. Лінійний пошук



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

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)};$$

while  $|f(x_{n+1})| \geq |f(x_n)|$  do:

$$x_{n+1} = \frac{x_n + x_{n+1}}{2}.$$

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 зменшувалася 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*.

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

$$f'(x) \approx a_n = \frac{f(x_n + h_n) - f(x_n)}{h_n}.$$

Тоді квазіньютонівський ітераційний крок має вигляд

$$x_{n+1} = x_n - \frac{f(x_n)}{a_n}. \quad (3.8)$$

Як видно, для збільшення швидкості збіжності скінченнорізницевого методу Ньютона необхідно зменшувати  $h_n$ . Однак на практиці через наявність арифметики скінченної точності значення  $h_n$  обмежено знизу. Розумний компроміс полягає в тому, щоб збалансувати похибку нелінійності, пов'язану з вибором занадто великих  $h_n$  з похибками арифметики скін-

ченної точності. Тому на практиці значення  $h_n$  вибирають таким, щоб внести зміни приблизно в половину розрядів мантиси  $x_n$ :

$$|h_n| = \sqrt{\varepsilon_{\text{маш}}} \max(|x_{\text{тип}}|, |x_n|),$$

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  $h_n$  вважається рівним  $x_n - x_{n-1}$ , тому

$$a_n = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}.$$

Тоді з виразу (3.8)

$$x_{n+1} = x_n - (x_n - x_{n-1}) \frac{f(x_n)}{f(x_n) - f(x_{n-1})}. \quad (3.9)$$

Таким чином, у ітераційному процесі (3.9) на кожному кроці  $x_{n+1}$  одержують з  $x_n$  і  $x_{n-1}$  як єдиний нуль лінійної функції, що набуває значення  $f(x_n)$  у  $x_n$  і  $f(x_{n-1})$  у  $x_{n-1}$ . Ця лінійна функція є січною до кривої  $y = f(x)$ , що проходить через її точки з абсцисою  $x_n$  і  $x_{n-1}$ . Тому цей метод називають *методом січних* (рис. 3.4).

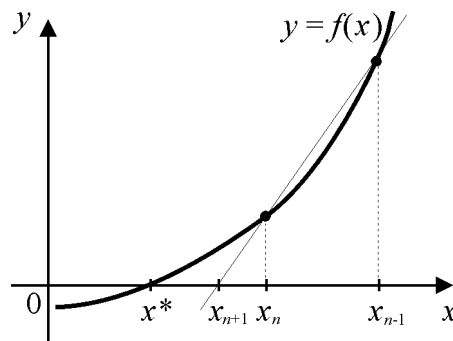


Рис. 3.4. Метод січних

Метод січних належить до двокрокових ітераційних методів, оскільки нове наближення до кореня залежить від двох попередніх наближень (формула (3.9)). Оскільки на початковій ітерації методу січних також потрібні знання двох початкових наближень і значень функції в них, то, як правило, методу січних передуює одна ітерація однокрокового методу, наприклад скінченнорізницевого методу Ньютона. Збіжність методу січних трохи повільніша, ніж методу Ньютона і скінченнорізницевого методу.

### 3.4. Метод Ньютона розв'язання систем нелінійних рівнянь

Розглянемо систему нелінійних рівнянь

$$\begin{cases} f_1(x_1, x_2, \dots, x_n) = 0; \\ f_2(x_1, x_2, \dots, x_n) = 0; \\ \dots \\ f_n(x_1, x_2, \dots, x_n) = 0. \end{cases} \quad (3.10)$$

Нехай  $\mathbf{X}_k = [x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)}]^T$  –  $k$ -й наближений розв'язок зада-

чі (3.10). Розкладемо кожену функцію  $f_i(x_1, x_2, \dots, x_n)$ ,  $i = \overline{1, n}$  у ряд Тейлора в околі точки  $\mathbf{X} = \mathbf{X}_k$  і обмежимося лінійними членами рядів. Тоді замість системи (3.10) маємо наближену систему рівнянь

$$f_i(\mathbf{X}_k) + \sum_{j=1}^n (x_j - x_j^{(k)}) \frac{\partial f_j(\mathbf{X}_k)}{\partial x_i} = 0, \quad i = \overline{1, n}, \quad (3.11)$$

лінійну відносно  $(x_j - x_j^{(k)})$ . Розв'язок системи (3.11) вважатимемо за наступне наближення до розв'язку  $\mathbf{X}_{k+1} = [x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_n^{(k+1)}]$ . Тоді

$$f_i(\mathbf{X}_k) + \sum_{j=1}^n (x_j^{(k+1)} - x_j^{(k)}) \frac{\partial f_j(\mathbf{X}_k)}{\partial x_i} = 0, \quad i = \overline{1, n}. \quad (3.12)$$

Систему (3.12) 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

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

Тоді процес розв'язання нелінійної системи подають у вигляді ітеративної процедури:

$$\mathbf{J}_k(\mathbf{X}_k)\Delta\mathbf{X}_k = -\mathbf{F}(\mathbf{X}_k);$$

$$\mathbf{X}_{k+1} = \mathbf{X}_k + \Delta\mathbf{X}_k.$$

Критерієм зупинки ітераційного процесу є збіжність за аргументом:

$$\frac{\|\Delta\mathbf{X}_k\|}{\|\mathbf{X}_{k+1}\|} \leq \varepsilon, \quad (3.13)$$

де  $\varepsilon$  – задана відносна похибка або збіжність за функцією

$$\|\mathbf{F}(\mathbf{X}_{k+1})\| \leq \Delta, \quad (3.14)$$

де  $\Delta$  – задана абсолютна похибка розв'язку. На практиці корисно поєднувати перевірки за обома критеріями.

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

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

Виберемо початкове наближення  $\mathbf{X}_0 = [1 \ 5]^T$ . Для оцінювання похибки у (3.13) та (3.14) використаємо евклідові норми  $\|\mathbf{X}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ . Тоді маємо:

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

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

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

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

$$\mathbf{J}(\mathbf{X}_2)\Delta \mathbf{X}_2 = -\mathbf{F}(\mathbf{X}_2) \Rightarrow \begin{bmatrix} 1 & 1 \\ -\frac{25}{136} & \frac{841}{136} \end{bmatrix} \Delta \mathbf{X}_2 = -\begin{bmatrix} 0 \\ \frac{21025}{36992} \end{bmatrix} \Rightarrow \Delta \mathbf{X}_2 = \begin{bmatrix} \frac{21025}{235552} \\ -\frac{21025}{235552} \end{bmatrix} \Rightarrow$$

$$\Rightarrow \mathbf{X}_3 = \begin{bmatrix} -\frac{25}{272} \\ \frac{841}{272} \end{bmatrix} + \begin{bmatrix} \frac{21025}{235552} \\ -\frac{21025}{235552} \end{bmatrix} = \begin{bmatrix} -\frac{625}{235552} \\ \frac{707281}{235552} \end{bmatrix} \approx \begin{bmatrix} -0,0027 \\ 3,0027 \end{bmatrix}, \frac{\|\Delta \mathbf{X}_2\|}{\|\mathbf{X}_3\|} \approx 0,04.$$

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

### 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.

Одним з можливих способів розрахунку якобіана є заміна похідних скінченими різницями

$$\frac{\partial f_j(x_1, x_2, \dots, x_n)}{\partial x_i} \approx \frac{f_j(x_1, x_2, \dots, x_i + \Delta x_i, \dots, x_n) - f_j(x_1, x_2, \dots, x_i, \dots, x_n)}{\Delta x_i},$$

де  $\Delta x_i$  – приріст  $i$ -го аргументу цієї функції. Оптимальне значення кроку скінчених різниць визначають із системи (3.11).

Другим способом модифікації методу Ньютона є використання методу січних. При цьому якобіан системи на кожному  $k + 1$ -му кроці обчислюється за формулою

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

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

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

$\mathbf{G}_k = \mathbf{J}_k^T \Delta \mathbf{F}_k$ ,  $\Delta \mathbf{F}_k = \mathbf{F}(\mathbf{X}_{k+1}) - \mathbf{F}(\mathbf{X}_k)$  – 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 скінченими різницями або аналітичним способом.

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

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

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

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

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

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

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

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

### Контрольні завдання

1. Вибрати з табл. 3.1 нелінійне рівняння відповідно до варіанта.
2. 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'язів 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	$2\ln(-(x + 2)(x + 5)) + x + 3 = 0$
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	$\begin{cases} \sin^2(x) + \cos(y) = 1, \\ \cos(x) - \sin(y) = 1 \end{cases}$
3	$\begin{cases} e^{2x} + \sqrt{y} = 1, \\ e^x - 3\sqrt{y} = 1 \end{cases}$	4	$\begin{cases} \sin(x) + \sqrt{y} = 0, \\ 2\cos(x) + \sqrt{y} = 2 \end{cases}$
5	$\begin{cases} 3\sin(x) + y^2 = 0, \\ \sin(x) - y^2 = 0 \end{cases}$	6	$\begin{cases} \sin^2(x) + \sqrt{y} = 0, \\ \sin^2(x) - 2\sqrt{y} = 0 \end{cases}$

Продовження табл. 3.2

7	$\begin{cases} e^{2x} + y = 1, \\ e^x - 3y = 1 \end{cases}$	8	$\begin{cases} e^x + \cos(y) = 2, \\ \cos(x) + e^y = 2 \end{cases}$
9	$\begin{cases} 2 \cos(x) + \sin(y) = 2, \\ \sin(x) + \cos(y) = 1 \end{cases}$	10	$\begin{cases} \cos^2(x) + \sqrt{y} = 1, \\ \sqrt{x} + \cos(y) = 1 \end{cases}$
11	$\begin{cases} \cos(x) + y = 1, \\ \sqrt{x} - y = 0 \end{cases}$	12	$\begin{cases} e^x + e^y = 2, \\ x + 2y = 0 \end{cases}$
13	$\begin{cases} e^{2x} + e^y = 2, \\ \sin(x) + \sin(y) = 0 \end{cases}$	14	$\begin{cases} e^x + e^y = 2, \\ \cos(2x) + \cos(y) = 2 \end{cases}$
15	$\begin{cases} e^x + e^{2y} = 2, \\ 2\sqrt{x} + \sqrt{y} = 0 \end{cases}$	16	$\begin{cases} e^{3x} - e^y = 0, \\ \sqrt{x} - 2\sqrt{y} = 0 \end{cases}$
17	$\begin{cases} \operatorname{tg}(x) + \sin(y) = 0, \\ \sqrt{x} + \sqrt{y} = 0 \end{cases}$	18	$\begin{cases} \sin(x) + \operatorname{tg}(y) = 0, \\ \sin(x) + \cos(y) = 1 \end{cases}$
19	$\begin{cases} \operatorname{tg}(x) + e^y = 1, \\ e^{2x} + \sqrt{y} = 1 \end{cases}$	20	$\begin{cases} \operatorname{tg}(x) + \operatorname{tg}(y) = 0, \\ e^x + e^{2y} = 2 \end{cases}$
21	$\begin{cases} \sqrt{x} + \operatorname{tg}(y) = 0, \\ x + 2y = 0 \end{cases}$	22	$\begin{cases} 3\operatorname{tg}(2x) + \sqrt{y} = 0, \\ x - 2y = 0 \end{cases}$
23	$\begin{cases} \operatorname{tg}^2(x) + y = 0, \\ 2x + \sqrt{y} = 0 \end{cases}$	24	$\begin{cases} e^x + \operatorname{tg}^2(y) = 0, \\ 2x + e^y = 1 \end{cases}$
25	$\begin{cases} xe^y = 0, \\ \sin(x)y = 0 \end{cases}$	26	$\begin{cases} x\sqrt{y} + y = 0, \\ \cos^2(x) + xy = 0 \end{cases}$
27	$\begin{cases} xe^x + \sin(xy) = 0, \\ \sin(x) + \cos(y) = 1 \end{cases}$	28	$\begin{cases} x\sqrt{y} + \cos(y) = 1, \\ e^x \cos(x) + y = 1 \end{cases}$
29	$\begin{cases} \operatorname{tg}(xy) + e^{xy} = 1, \\ x - y = 0 \end{cases}$	30	$\begin{cases} 3\operatorname{tg}^2(x) + xy = 0, \\ xe^y + y = 0 \end{cases}$



## 4. Інтерполяція функцій

Нехай деяку функцію задано табл. 4.1.

Таблиця 4.1. Деяка таблична функція

$x$	$x_1$	$x_2$	$\dots$	$x_n$
$y(x)$	$y_1$	$y_2$	$\dots$	$y_n$

Під задачею інтерполяції функції розуміють побудову такої функції  $f(x)$ , яка проходила б через усі задані точки  $(x_i, y_i)$ , тобто для кожного  $i$  повинна виконуватися рівність

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

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 [1, 2, 5].

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

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

При цьому вид інтерполяції визначають функцією  $\varphi_j(x)$ . Якщо  $\varphi_j(x)$  є поліномом  $n-1$  степеня, тобто  $\varphi_j(x) = b_{j0} + b_{j1}x + b_{j2}x^2 + \dots + b_{jn-1}x^{n-1}$ , то інтерполяція називається *поліноміальною*. Якщо  $\varphi_j(x) = \cos(b_jx + a_j)$ , 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

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

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

$$\sum_{i=1}^n y_i Q_i(x_j) = y_j, \quad j = \overline{1, n}. \quad (4.2)$$

Співвідношення (4.2) виконуються за умови, що

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

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

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

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

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

Формулу (4.5) називають інтерполяційною формулою Лагранжа.

**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  $P_{n-1}(x)$  через значення  $y_i$  в одному з вузлів та через розділені різниці функції  $f(x)$ , побудованими за іншими вузлами  $x_1, x_2, \dots, x_n$ . Вона є аналогом формули Тейлора.

Розділеними різницями першого порядку називають відношення

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

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

$$\begin{aligned}\Delta(x_1, x_2, x_3) &= \frac{\Delta(x_1, x_2) - \Delta(x_2, x_3)}{x_1 - x_3}, \dots, \Delta(x_{n-2}, x_{n-1}, x_n) = \\ &= \frac{\Delta(x_{n-2}, x_{n-1}) - \Delta(x_{n-1}, x_n)}{x_{n-2} - x_n}, \\ &\dots \\ \Delta(x_j, x_{j+1}, \dots, x_{j+k}, x_{j+k+1}) &= \frac{\Delta(x_j, x_{j+1}, \dots, x_{j+k}) - \Delta(x_{j+1}, \dots, x_{j+k}, x_{j+k+1})}{x_j - x_{j+k+1}}.\end{aligned}$$

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

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

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, запишемо інтерполяційну формулу

$$\begin{aligned}f(x) &\approx y_1 + (x - x_1)\Delta(x_1, x_2) + (x - x_1)(x - x_2)\Delta(x_1, x_2, x_3) + \dots \\ &+ (x - x_1)(x - x_2)\dots(x - x_{n-1})\Delta(x_1, x_2, \dots, x_n) = \\ &= y_1 + \sum_{k=2}^n \left( \Delta(x_1, x_2, \dots, x_k) \prod_{j=1}^{k-1} (x - x_j) \right).\end{aligned}\quad (4.6)$$

Формула (4.6) називається інтерполяційною формулою Ньютона.

Побудуємо інтерполяційний поліном Ньютона для табличної функції, заданої в табл. 4.2:

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

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

### 4.3. Похибка поліноміальної інтерполяції

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

$$R_{n-1} = y(x) - f(x). \quad (4.7)$$

Уважатимемо, що функція  $y(x)$  має всі похідні до  $n$ -го порядку включно. Подамо остаточний член у вигляді полінома степеня  $n$ :

$$R(x) = m \prod_{j=1}^n (x - x_j). \quad (4.8)$$

Ураховуючи (4.7) і (4.8), отримаємо вираз для  $y(x)$ :

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

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

$$f^{(n)}(x) = (n!)m. \quad (4.9)$$

З виразу (4.9) маємо

$$m = \frac{f^{(n)}(x)}{n!}. \quad (4.10)$$

На підставі виразів (4.10) та (4.8) записуємо

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

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

$$|R(x)| \leq \left| \frac{\max_{x \in [x_1; x_n]} (\Delta(x, x_1, x_2, \dots, x_n))}{n!} \prod_{j=1}^n (x - x_j) \right|. \quad (4.11)$$

Якщо відстань між суміжними точками не перевищує деякої величини  $h$ , то наближено вважають, що  $\prod_{j=1}^n (x - x_j) \approx c_n h^n$ , де  $c_n$  – деяка константа, що залежить від способу розбиття відрізка  $[x_1; x_n]$ . Підставимо цей вираз у (4.11). Остаточоно вираз для оцінки похибки має вигляд

$$|R(x)| \leq \left| \frac{\max_{x \in [x_1; x_n]} (\Delta(x, x_1, x_2, \dots, x_n)) c_n h^n}{n!} \right|.$$

#### 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. При цьому точність значення похідної, обчисленої шляхом диференціювання інтерполяційного полінома, може виявитися незадовільною.

Виходом з такої ситуації є використання сплайн-інтерполяції. Сплайн – це функція, що на кожному частковому відрізку є поліномом певного сте-

пеня, а на всьому заданому відрізку – неперервна разом з кількома своїми похідними. При цьому поліном на кожному частковому відрізку має порядок  $n \leq 3$ . Найчастіше використовують сплайни третього степеня.

Нехай функцію задано табл. 4.3 і треба побудувати сплайн-інтерполяцію цієї функції.

**Таблиця 4.3. Деяка задана таблична функція**

$f(x)$	$y_0$	$y_1$	$\dots$	$y_n$
$x$	$x_0$	$x_1$	$\dots$	$x_n$

За допомогою сплайнів функцію можна подати у вигляді

$$f(x) = \begin{cases} S_1(x) = a_{01} + a_{11}(x - x_0) + \dots + a_{m1}(x - x_0)^m, & x \in [x_0; x_1]; \\ S_2(x) = a_{02} + a_{12}(x - x_1) + \dots + a_{m2}(x - x_1)^m, & x \in [x_1; x_2]; \\ \dots \\ S_n(x) = a_{0n} + a_{1n}(x - x_{n-1}) + \dots + a_{mn}(x - x_{n-1})^m, & x \in [x_{n-1}; x_n]. \end{cases}$$

Для побудови сплайн-інтерполяції визначають невідомі коефіцієнти  $a_{01}, a_{02}, \dots, a_{mn}$ . Якщо функція інтерполюється на  $n$  інтервалах за допомогою поліномів порядку  $m$ , то кількість таких невідомих буде  $(m + 1)n$ . Для їх визначення потрібно  $(m + 1)n$  рівнянь. Щоб отримати необхідну кількість рівнянь, можна скористатися умовами рівності функції табличним значенням у вузлах інтерполяції. Ці умови дадуть  $2n$  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  $(n - 1)$  рівняння. Отриману систему треба доповнити ще одним рівнянням. Наприклад, можна задати поведінку функції на одному з кінців інтервалу (наприклад, рівність  $S'_1(x_0) = f'(x_0)$ , причому похідна функції обчислюється чисельним методом на основі табличних значень функції). У разі інтерполяції кубічними сплайнами систему рівнянь необхідно доповнити ще  $n$  рівняннями. Для цього можна поставити вимогу неперервності другої похідної функції, що дасть  $(n - 1)$  рівняння.

Останнє рівняння отримують, задавши поведінку функції на другому (незадіяному) кінці відрізка. На практиці часто задають однакові умови на кінцях інтервалу. Як правило, це умова гладкості функції на кінцях ( $S''_1(x_0) = S''_2(x_n) = 0$ ).

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

На кожному  $i$ -му відрізку:  $S_i(x) = a_i(x - x_{i-1})^3 + b_i(x - x_{i-1})^2 + c_i(x - x_{i-1}) + d_i$ ,  $S'_i(x) = 3a_i(x - x_{i-1})^2 + 2b_i(x - x_{i-1}) + c_i$ ,  $S''_i(x) = 6a_i(x - x_{i-1}) + 2b_i$ . Побудуємо систему рівнянь для визначення невідомих коефіцієнтів.

З умови рівності функції табличним значенням у вузлових точках маємо:

$$\begin{aligned}
 & i = 1; \\
 & x = 0: a_1(0-0)^3 + b_1(0-0)^2 + c_1(0-0) + d_1 = 0; \\
 & x = 1: a_1(1-0)^3 + b_1(1-0)^2 + c_1(1-0) + d_1 = 2; \\
 & i = 2; \\
 & x = 1: a_2(1-1)^3 + b_2(1-1)^2 + c_2(1-1) + d_2 = 2; \\
 & x = 2: a_2(2-1)^3 + b_2(2-1)^2 + c_2(2-1) + d_2 = 10.
 \end{aligned} \tag{4.12}$$

З умови неперервності першої та другої похідних функцій:

$$\begin{aligned}
 & x = 1: \\
 & 3a_1(1-0)^2 + 2b_1(1-0) + c_1 = c_2; \\
 & 6a_1(1-0) + 2b_1 = 6a_2(1-1) + 2b_2.
 \end{aligned} \tag{4.13}$$

З умови гладкості функції:

$$\begin{aligned}
 & x = 0: 6a_1(0-0) + 2b_1 = 0; \\
 & x = 2: 6a_2(2-1) + 2b_2 = 0.
 \end{aligned} \tag{4.14}$$

Розв'язавши рівняння (4.12) – (4.14), отримаємо:

$$\begin{aligned}
 & a_1 = \frac{3}{2}, \quad b_1 = 0, \quad c_1 = \frac{1}{2}, \quad d_1 = 0; \\
 & a_2 = -\frac{3}{2}, \quad b_2 = \frac{9}{2}, \quad c_2 = 5, \quad d_2 = 2.
 \end{aligned}$$

Тоді функція має вигляд

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

### 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 таблиці та порівняти отримані результати.

### Варіанти завдань

1	$x$	-1	0	1	2
---	-----	----	---	---	---

2	$x$	-1	0	1	2
---	-----	----	---	---	---

	$y$	-2	-4	-2	-2
--	-----	----	----	----	----

	$y$	-2	-1	-4	-5
--	-----	----	----	----	----

3	$x$	-1	0	1	2
	$y$	3	4	7	6

4	$x$	-1	0	1	2
	$y$	1	0	-3	-2

5	$x$	-1	0	1	2
	$y$	2	2	6	8

6	$x$	-1	0	1	2
	$y$	7	4	3	10

7	$x$	-1	0	1	2
	$y$	6	2	2	12

8	$x$	-1	0	1	2
	$y$	5	3	1	-7

9	$x$	-1	0	1	2
	$y$	1	-2	-5	-14

10	$x$	-1	0	1	2
	$y$	1	-2	-3	-8

11	$x$	-1	0	1	2
	$y$	4	4	4	10

12	$x$	-1	0	1	2
	$y$	-2	-4	-6	-2

13	$x$	-1	0	1	2
	$y$	-10	-3	0	5

14	$x$	-1	0	1	2
	$y$	7	5	3	-5

15	$x$	-1	0	1	2
	$y$	7	4	3	-2

16	$x$	-1	0	1	2
	$y$	-7	-5	-5	-13

17	$x$	-1	0	1	2
	$y$	3	4	1	-12

18	$x$	-1	0	1	2
	$y$	3	1	1	9

19	$x$	-1	0	1	2
	$y$	-2	-2	-4	-2

20	$x$	-1	0	1	2
	$y$	4	4	2	-8

21	$x$	-1	0	1	2
	$y$	-2	-2	0	10

22	$x$	-1	0	1	2
	$y$	-2	-4	-4	-8

23	$x$	-1	0	1	2
	$y$	2	2	0	-10

24	$x$	-1	0	1	2
	$y$	0	3	4	-3

25	$x$	-1	0	1	2
	$y$	0	-3	-6	-15

26	$x$	-1	0	1	2
	$y$	3	5	3	3

27	$x$	-1	0	1	2
	$y$	-9	-3	1	9

28	$x$	-1	0	1	2
	$y$	4	2	-2	-2

29	$x$	-1	0	1	2
----	-----	----	---	---	---

30	$x$	-1	0	1	2
----	-----	----	---	---	---

	$y$	5	2	1	8
--	-----	---	---	---	---

	$y$	7	5	5	1
--	-----	---	---	---	---



## Числове інтегрування функцій

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

Задача числового 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  $a = x_0 < x_1 < x_2 < \dots < x_n = b$  і шукане значення інтеграла замінюють сумою

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

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

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

де  $f(\xi_{i,j})$  – значення функції  $f(x)$  у вузлах  $x = \xi_{i,j}$ ;  $c_j$  – вагові множники, що залежать тільки від вузлів, а не від вибору  $f(x)$ ,  $h_i = x_{i+1} - x_i$ . Для числового наближення визначених інтегралів часто використовують термін «квадратура», щоб уникнути плутанини з числовим 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 числового інтегрування за допомогою квадратур полягає у відшуванні 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  $R_i = O(h_i^{p+1})$ , 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

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

Для побудови формули числового інтегрування на усьому 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 числового 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  $\Delta I_n$  [1]:

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

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

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

де  $\Delta c_i$  – абсолютна похибка вагових коефіцієнтів;  $\Delta x_i$  – абсолютна похибка вузлів.

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

$$\Delta I_n \leq n \varepsilon_{\text{маш}} (A_1 + A_2) = n A \varepsilon_{\text{маш}},$$

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

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

Похибка квадратурної формули  $R_n(f)$  оцінюється як

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

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

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

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).

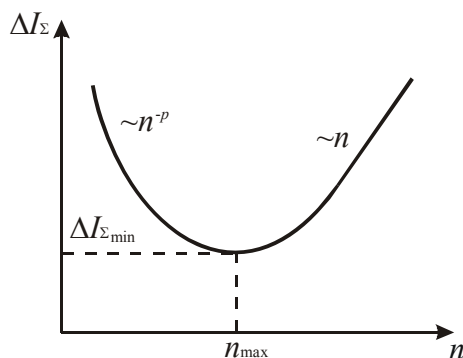


Рис. 5.1. Загальна похибка числового інтегрування

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

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

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

## 5.2. Формули Ньютона – Котеса

Формули Ньютона – Котеса одержують інтегруванням інтерполяційних поліномів, побудованих за рівномірною сіткою [2, 3]. Розрізняють формули відкритого і закритого типів. У формулах закритого типу у розрахунках використовують значення функції, обчислені в обох кінцях еле-

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

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

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

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 розраховують за допомогою табл. 5.1.

Похибка формули Ньютона – Котеса оцінюють як

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

**Таблиця 5.1. Постійні для розрахунку коефіцієнтів Котеса і похибка формули Ньютона – Котеса порядку  $m$**

$M$	$\alpha_0$	$\alpha_1$	$\alpha_2$	$\alpha_3$	$\alpha_4$	$\alpha_5$	$\sum_{j=0}^m \alpha_j$	$R_i$
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	$-\frac{275}{12096 \cdot 5^7}h_i^7 f^{VI}(\xi)$
6	41	216	27	272	27	216	840	$-\frac{9}{1400 \cdot 6^9}h_i^9 f^{VIII}(\xi)$
7	751	3 577	1 323	2 989	2 989	1 323	17 280	$-\frac{8183}{518400 \cdot 7^9}h_i^9 f^{VIII}(\xi)$
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	$\sim h_i^{13} f^{XII}(\xi)$

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.

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 числову нестійкість, що призводить до різкого зростання обчислювальної похибки. Причиною такої нестійкості є те, що коефіцієнти  $c_j$  у формулі (5.6) при великих  $m$  мають різні знаки. Деякі коефіцієнти стають від'ємними при  $m = 8$ . Для  $m = 9$  вони всі додатні, але для  $m \geq 10$  існують як додатні, так і від'ємні значення.

**Приклад.** Обчислимо інтеграл  $I = \int_0^1 \frac{dx}{1+x^2}$ . 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

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

### 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  $m$  [6]. 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:

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

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

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

$$x = \frac{x_{i+1} + x_i}{2} + \frac{h_i}{2} t, \quad (5.10)$$

тоді

$$\xi_{i,j} = \frac{x_{i+1} + x_i}{2} + \frac{h_i}{2} t_j. \quad (5.11)$$

Із рівняння (5.9) одержуємо

$$\int_{-1}^1 f\left(\frac{x_{i+1} + x_i}{2} + \frac{h_i}{2} t\right) dt = \frac{2}{m+1} \sum_{j=0}^m f\left(\frac{x_{i+1} + x_i}{2} + \frac{h_i}{2} t_j\right). \quad (5.12)$$

Уведемо позначення  $f\left(\frac{x_{i+1} + x_i}{2} + \frac{h_i}{2} t\right) = g(t)$ . Тоді з (5.12)

$$\int_{-1}^1 g(t) dt = \frac{2}{m+1} \sum_{j=0}^m g(t_j). \quad (5.13)$$

Для визначення вузлів  $t_j$  використовуємо умову, що формула (5.13) є точною для поліномів  $g(t) = t, t^2, \dots, t^{m+1}$ . Підставляючи ці функції в (5.13), одержуємо систему рівнянь:

$$\begin{aligned}
 t_0 + t_1 + \dots + t_m &= 0; \\
 t_0^2 + t_1^2 + \dots + t_m^2 &= \frac{m+1}{3}; \\
 t_0^3 + t_1^3 + \dots + t_m^3 &= 0; \\
 t_0^4 + t_1^4 + \dots + t_m^4 &= \frac{m+1}{5}; \\
 &\dots \\
 t_0^{m+1} + t_1^{m+1} + \dots + t_m^{m+1} &= \frac{(m+1)(1 - (-1)^{m+1})}{2(m+2)}.
 \end{aligned} \tag{5.14}$$

Із виразів (5.14) можна знайти  $t_j$ ,  $j = \overline{0, m}$ , а потім з (5.11) –  $\xi_{ij}$ . Нижче наведено  $t_j$  для порядків  $m = \overline{1, 4}$ :

$$\begin{aligned}
 m = 1; \quad -t_0 = t_1 &= \frac{1}{\sqrt{3}}; \\
 m = 2; \quad -t_0 = t_2 &= 0,707107; \quad t_1 = 0; \\
 m = 3; \quad -t_0 = t_3 &= 0,794654; \quad -t_1 = t_2 = 0,187592; \\
 m = 4; \quad -t_0 = t_4 &= 0,832498; \quad -t_1 = t_3 = 0,374541; \quad t_2 = 0.
 \end{aligned}$$

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 (5.11) знайдемо розташування вузлів формули Чебишова:

$$\begin{aligned}
 \xi_{0,0} &= \frac{1}{2} + \frac{1}{2} \cdot (-0,794654) = 0,102673; \quad \xi_{0,1} = \frac{1}{2} + \frac{1}{2} \cdot (-0,187592) = 0,406204; \\
 \xi_{0,2} &= \frac{1}{2} + \frac{1}{2} \cdot 0,187592 = 0,593796; \quad \xi_{0,4} = \frac{1}{2} + \frac{1}{2} \cdot 0,794654 = 0,897327.
 \end{aligned}$$

За формулою (5.9) обчислимо наближене значення інтегралу:

$$\begin{aligned}
 I &\approx \frac{1}{4} \left( \frac{1}{1+0,102673^2} + \frac{1}{1+0,406204^2} + \frac{1}{1+0,593796^2} + \frac{1}{1+0,897327^2} \right) = \\
 &\cong 0,785303
 \end{aligned}$$

## 5.4. Формули Гаусса

На відміну від формул Ньютона – Котеса та Чебишова у виведеній з виразу (5.2) формулі Гаусса вузли і вагові коефіцієнти не задаються, а визначаються так, щоб формула була точною для полінома найвищого можливого степеня [6]. Можна показати, що ця вимога виконується, якщо вузли обчислюються за формулою (5.11), причому  $t_j$  є коренями полінома Лежандра  $P_{m+1}(t)$  степеня  $m + 1$ , а вагові коефіцієнти обчислюються за формулою

$$c_j = \frac{[P_{m+1}(1)]^2}{(1-t_j^2)[P'_{m+1}(t_j)]^2}.$$

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

$$P_{m+1}(x) = \frac{2m+1}{m+1} x P_m(x) - \frac{m}{m+1} P_{m-1}(x) = x P_m(x) + \frac{x^2}{m+1} \frac{dP_m(x)}{dx};$$

$$P_0(x) = 1, \quad P_1(x) = x;$$

$$P_2(x) = \frac{1}{2}(3x^2 - 1), \quad P_3(x) = \frac{1}{2}(5x^3 - 3x);$$

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3).$$

Деякі значення параметрів формули Гаусса для методів 1 – 4-го орядків наведено нижче:



$$\begin{array}{ll}
m = 1; & -t_0 = t_1 = \frac{1}{\sqrt{3}}; & c_0 = c_1 = \frac{1}{2}; \\
m = 2; & -t_0 = t_2 = \sqrt{\frac{3}{5}}; & c_0 = c_2 = \frac{5}{18}; \\
& t_1 = 0; & c_1 = \frac{4}{9}; \\
m = 3; & -t_0 = t_3 \cong 0,861136; & c_0 = c_3 \cong 0,173928; \\
& -t_1 = t_2 \cong 0,339981; & c_1 = c_2 \cong 0,326072 \\
m = 4; & -t_0 = t_4 \cong 0,9061798; & c_0 = c_4 \cong 0,1184634; \\
& -t_1 = t_3 = 0,5384693; & c_1 = c_3 \cong 0,2393143; \\
& t_2 = 0; & c_2 \cong 0,2844444.
\end{array}$$

Методи Гаусса мають порядок точності  $2m + 1$ .

**Приклад.** Обчислимо інтеграл  $I = \int_0^1 \frac{dx}{1+x^2}$  методом Гаусса 3-го порядку

за одним елементарним відрізком.

$$\begin{array}{l}
m = 3; \quad -t_0 = t_3 \cong 0,861136; \quad c_0 = c_3 \cong 0,173928; \\
\quad \quad -t_1 = t_2 \cong 0,339981; \quad c_1 = c_2 \cong 0,326072.
\end{array}$$

Згідно з (5.11) знайдемо розташування вузлів формули Гаусса:

$$\begin{array}{l}
\xi_{0,0} = \frac{1}{2} + \frac{1}{2} \cdot (-0,861136) = 0,069432; \quad \xi_{0,1} = \frac{1}{2} + \frac{1}{2} \cdot (-0,339981) = 0,330010; \\
\xi_{0,2} = \frac{1}{2} + \frac{1}{2} \cdot 0,339981 = 0,669991; \quad \xi_{0,4} = \frac{1}{2} + \frac{1}{2} \cdot 0,861136 = 0,930568.
\end{array}$$

Тоді згідно з формулою (5.2) маємо

$$\begin{aligned}
I &\approx 1 \cdot \left( 0,173928 \cdot \frac{1}{1+0,069432^2} + 0,326072 \cdot \frac{1}{1+0,330010^2} + \right. \\
&\quad \left. + 0,326072 \cdot \frac{1}{1+0,669991^2} + 0,173928 \cdot \frac{1}{1+0,930568^2} \right) = \\
&= 0,785403
\end{aligned}$$

## 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.

ти, 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, згідно з яким

$$R = \frac{I_{h/q} - I_h}{q^p - 1}, \quad (5.15)$$

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 інтервал інтегрування на елементарні відрізки довжиною  $h$ , потім на елементарні відрізки довжиною  $h/q$ . Після цього за виразом (5.15) 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:

$$p = \frac{\ln\left(\frac{I_h - I_{h/q}}{I_{h/q} - I_{h/q^2}}\right)}{\ln(q)},$$

### 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 Гаусса 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.**  $3x^2 + \operatorname{tg}(x); x \in [-0,5; 0,5].$       **14.**  $3x^2 + \sin(x); x \in [0; 1].$   
**15.**  $3xe^{\cos(x)}; x \in [0,2; 1,2].$       **16.**  $x^2 \operatorname{tg}\left(\frac{x}{2}\right); x \in [0,5; 1,5].$   
**17.**  $x^2/(1+0,25x); x \in [1,1; 2,1].$       **18.**  $(x^3-0,3x)/\sqrt{(1+2x)}; x \in [2; 3].$   
**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. Числове 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 значення. У рамках цього розділу розглянемо розв'язання задачі Коші.

Для ЗДР першого порядку задача Коші полягає в знаходженні такого розв'язку рівняння

$$\frac{dy}{dt} = f(t, y), \quad (6.1)$$

що задовольняє початкову умову

$$y(0) = y_0. \quad (6.2)$$

Для диференціального рівняння  $m$ -го порядку

$$\frac{d^m y}{dt^m} = f\left(t, y, \frac{dy}{dt}, \dots, \frac{d^{m-1}y}{dt^{m-1}}\right) \quad (6.3)$$

задача Коші полягає в знаходженні функції  $y = y(t)$ , що задовольняє рівняння (6.3) і початкові умови:

$$\begin{aligned} y(0) &= y_0; \\ \frac{dy(0)}{dt} &= y_0^{(1)}; \\ &\dots \\ \frac{d^{m-1}y(0)}{dt^{m-1}} &= y_0^{(m-1)}. \end{aligned}$$

Розв'язуючи деякі задачі, слід знайти  $m$  функцій  $y_1(t), y_2(t), \dots, y_m(t)$ , що задовольняють систему диференціальних рівнянь:

$$\begin{aligned} \frac{dy_1}{dt} &= f_1(t, y_1, y_2, \dots, y_m); \\ \frac{dy_2}{dt} &= f_2(t, y_1, y_2, \dots, y_m); \\ &\dots \\ \frac{dy_m}{dt} &= f_m(t, y_1, y_2, \dots, y_m) \end{aligned} \quad (6.4)$$

і початкові умови

$$\begin{aligned}
y_1(0) &= y_{10}; \\
y_2(0) &= y_{20}; \\
&\dots \\
y_m(0) &= y_{m0}.
\end{aligned}
\tag{6.5}$$

Систему диференціальних рівнянь (6.4) називають нормальною.

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  $\mathbf{Y} = [y_1(t), y_2(t), \dots, y_m(t)]^T$ ,  $F(t, \mathbf{Y}) = [f_1(t, \mathbf{Y}), f_2(t, \mathbf{Y}), \dots, f_m(t, \mathbf{Y})]^T$ ,

$\mathbf{Y}_0 = [y_{10}, y_{20}, \dots, y_{m0}]^T$ .

Систему диференціальних рівнянь

$$\frac{dy_1}{dt} = a_{11}y_1 + a_{12}y_2 + \dots + a_{1m}y_m + b_1;$$

$$\frac{dy_2}{dt} = a_{21}y_1 + a_{22}y_2 + \dots + a_{2m}y_m + b_2;$$

...

$$\frac{dy_m}{dt} = a_{m1}y_1 + a_{m2}y_2 + \dots + a_{mm}y_m + b_m,$$

у якій  $a_{ij}$ ,  $b_j$ ,  $i, j = \overline{1, m}$  – функції від  $t$ , називають лінійною.

Якщо ввести позначення

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ & & \dots & \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_m \end{bmatrix},$$

то систему лінійних диференціальних рівнянь записують у матричній формі:

$$\frac{d\mathbf{Y}}{dt} = \mathbf{A}\mathbf{Y} + \mathbf{B}.$$

Рівняння (6.3) за допомогою заміни змінних

$$\begin{aligned}
y_1(t) &= \frac{dy}{dt}; \\
y_2(t) &= \frac{d^2y}{dt^2} = \frac{dy_1}{dt}; \\
&\dots \\
y_{m-1}(t) &= \frac{d^{m-1}y}{dt^{m-1}} = \frac{dy_{m-2}}{dt}
\end{aligned}$$

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

$$\begin{aligned}
\frac{dy}{dt} &= y_1; \\
\frac{dy_1}{dt} &= f(t, y, y_1).
\end{aligned}$$

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 застосовано до систем диференціальних рівнянь з урахуванням того, що всі скалярні операції замінюються операціями над векторами, матрицями і тензорами.

При чисельному розв'язку задачі Коші по змінній  $t$  вводять сітку  $0 < t_1 < t_2 < \dots < t_n$  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

$$R_i = y(t_{i+1}) - y_{i+1} \quad (6.7)$$

При використанні наближених методів основним є питання про збіжність. Найбільше розповсюдження отримало поняття про збіжність при  $h_i \rightarrow 0$ , де

$$h_i = t_{i+1} - t_i.$$

Говорять, що метод збігається в точці  $t_{i+1}$ , якщо  $\lim_{h_i \rightarrow 0} |R_i| = 0$ . 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  $R_i = O(h_i^{p+1})$ , при  $h_i \rightarrow 0$ .

## 6.1. Метод Ейлера

Для чисельного розв'язання задачі Коші (6.1), (6.2) розіб'ємо інтервал  $[0, t_n]$ , на якому шукається розв'язок, на елементарні відрізки  $[t_i, t_{i+1}]$ ,  $i = 0, n-1$ ,  $t_0=0$ . На кожному елементарному відрізку будемо шукати розв'язок у вигляді поліному першого ступеня. Нехай значення функції на границях елементарного відрізка приймають значення  $y_i, y_{i+1}$ . За Інтерполяційною формулою Лагранжа

$$y(t) = \frac{t-t_{i+1}}{t_i-t_{i+1}} y_{i+1} + \frac{t-t_i}{t_i-t_{i+1}} y_i = \frac{t-t_i}{h_i} y_{i+1} - \frac{t-t_{i+1}}{h_i} y_i.$$

Тоді

$$\frac{dy}{dt} = \frac{y_{i+1} - y_i}{h_i}. \quad (6.8)$$

Підставляючи (6.8) в (6.1) та замінюючи функцію  $f(t,y)$  її значенням на початку елементарного відрізка, отримаємо

$$\frac{y_{i+1} - y_i}{h_i} = f(t_i, y_i). \quad (6.9)$$

В методі Ейлера рівняння (6.1) замінюється різницеvim 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

$$y_{i+1} = y_i + h_i f(t_i, y_i) \quad (6.10)$$

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

$$y(t) = y_i + f(y_i, t_i)(t-t_i) + \frac{y''(t_i)}{2}(t-t_i)^2 + \dots$$

Звідки

$$y(t_{i+1}) = y_i + f(y_i, t_i)h_i + \frac{y''(t_i)}{2}h_i^2 + \dots \quad (6.11)$$

Враховуючи (6.10), вважаючи крок інтегрування  $h_i$  малим та нехтуючи малими величинами вищих порядків із (6.7) та (6.11) отримуємо

$$R_i = \frac{y''(t)}{2} h_i^2.$$

Таким чином, метод Ейлера має перший порядок точності. Тому в методі Ейлера зменшення кроку  $h_i$  в два рази в чотири рази зменшує похибку розв'язку.

Існують інші різновиди методу. Так, підставляючи (6.8) до (6.1) і замінюючи функцію  $f(y,t)$  її значення у кінці елементарного відрізка, отримуємо

$$\frac{y_{i+1} - y_i}{h_i} = f(y_i, t_i).$$

Звідки

$$y_{i+1} = y_i + h_i f(t_{i+1}, y_{i+1}). \quad (6.12)$$

Для довільної функції  $f(y,t)$  значення  $y_{i+1}$  не може бути виражене з (6.12) явно. Тому метод (6.12) називають неявним методом Ейлера. При використанні 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

$$\begin{cases} \frac{dy}{dt} = -(t+1)y^2 \\ y(0) = 1 \end{cases}.$$

Точним розв'язком цієї задачі є функція  $y(t) = \frac{2}{(t+1)^2 + 1}$ .

Необхідно обчислити таблицю значень  $y(t)$  на інтервалі  $t \in [0,1]$  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)

$$y_{i+1} = y_i + h \cdot (-(t_i + 1)y_i^2) = y_i - h(t_i + 1)y_i^2.$$

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

$$y_{i+1} = y_i + h \cdot \left( -(t_i + 1)y_{i+1}^2 \right) = y_i - h(t_{i+1} + 1)y_{i+1}^2,$$

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 уже 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

$$\begin{cases} \frac{dx}{dt} = x + t; \\ \frac{dy}{dt} = x + y; \end{cases} \quad x(0) = x_0 = 1, \quad y(0) = y_0 = 2.$$

Розв'язок будемо шукати на інтервалі  $t \in [0; 0,2]$  з кроком  $h = 0,1$ .

Використовуючи формулу (6.10) для системи ЗДР матимемо:

$$\begin{cases} x_{i+1} = x_i + 0,1 \cdot (x_i + t_i) \\ y_{i+1} = y_i + 0,1 \cdot (x_i + y_i) \end{cases}$$

$$t_1 = 0,1:$$

$$\begin{cases} x_1 = x_0 + 0,1 \cdot (x_0 + t_0) \\ y_1 = y_0 + 0,1 \cdot (x_0 + y_0) \end{cases} \Rightarrow \begin{cases} x_1 = 1 + 0,1 \cdot (1 + 0) \\ y_1 = 2 + 0,1 \cdot (1 + 2) \end{cases} \Rightarrow \begin{cases} x_1 = 1,1; \\ y_1 = 2,3. \end{cases}$$

$$t_2 = 0,2 :$$

$$\begin{cases} x_2 = x_1 + 0,1 \cdot (x_1 + t_1) \\ y_2 = y_1 + 0,1 \cdot (x_1 + y_1) \end{cases} \Rightarrow \begin{cases} x_2 = 1,1 + 0,1 \cdot (1,1 + 0,1) \\ y_2 = 2,3 + 0,1 \cdot (1,1 + 2,3) \end{cases} \Rightarrow \begin{cases} x_2 = 1,22; \\ y_2 = 2,64. \end{cases}$$

## 6.2. Методи Рунге – Кутта

У найзагальнішому випадку метод Рунге-Кутта порядку точності  $p$  будується за рекурентною формулою:

$$y_{i+1} = y_i + h_i \sum_{j=1}^m c_j k_j, \quad (6.13)$$

де

$$k_j = f \left( t_i + a_j h_i, y_i + h_i \sum_{l=1}^{j-1} b_{jl} k_l \right), \quad (6.14)$$

$a_1=0$ ,  $a_j$  и  $c_j$  - 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 зрівнюють з рядом Тейлора для функції  $y(t_{i+1})$ . Оскільки локальна похибка визначена згідно з (6.7), то ставиться умова, щоб коефіцієнти при всіх  $h_i^l, l = \overline{0, p}$  в розкладах для  $y(t_{i+1})$  и  $y_{i+1}$  були рівні. Ця вимога приводить до системи рівнянь відносно коефіцієнтів  $b_{jl}$ ,  $a_j$  и  $c_j$ .

Як приклад розглянемо побудову методів Рунге-Кутта другого порядку точності ( $p=2$ ). В цьому випадку можна отримати рішення для коефіцієнтів  $b_{jl}$ ,  $a_j$  и  $c_j$  при  $m=2$ . Тоді формула (6.13) перетвориться:

$$y_{i+1} = y_i + h_i (c_1 k_1 + c_2 k_2), \quad (6.15)$$

де  $k_1 = f(t_i, y_i)$ ,  $k_2 = f(t_i + ah_i, y_i + hb_k k_1)$ .

Розкладемо  $k_1$  і  $k_2$  в ряд Тейлора в околі точки  $(t_i, y_i)$ . Для цього відмітимо, що

$$f(t, y) = f(t_i, y_i) + (y - y_i) \frac{\partial f(t_i, y_i)}{\partial y} + (t - t_i) \frac{\partial f(t_i, y_i)}{\partial t} + \dots$$

Тоді

$$k_1 = f(t_i, y_i), \quad (6.16)$$

$$\begin{aligned} k_2 &= f(t_i, y_i) + (y_i + h_i b k_1 - y_i) \frac{\partial f(t_i, y_i)}{\partial y} + (t_i + a h_i - t_i) \frac{\partial f(t_i, y_i)}{\partial t} + \dots = \\ &= f(t_i, y_i) + h_i b f(t_i, y_i) \frac{\partial f(t_i, y_i)}{\partial y} + a h_i \frac{\partial f(t_i, y_i)}{\partial t} + \dots \end{aligned} \quad (6.17)$$

Підставляючи (6.16) та (6.17) в (6.15), отримаємо

$$y_{i+1} = y_i + h_i f(t_i, y_i)(c_1 + c_2) + h_i^2 c_2 \left( b f(t_i, y_i) \frac{\partial f(t_i, y_i)}{\partial y} + a \frac{\partial f(t_i, y_i)}{\partial t} \right) + \dots$$

Розкладемо функцію  $y(t)$  в ряд Тейлора в околі точки  $(y_i, t_i)$  та знайдемо  $y(t_{i+1})$ . Тоді

$$y(t_{i+1}) = y(t_i) + h_i \frac{dy(t_i)}{dt} + \frac{h_i^2}{2} \frac{d^2 y(t_i)}{dt^2} + \dots \quad (6.18)$$

Врахуємо, що

$$\frac{dy(t_i)}{dt} = f(t_i, y_i), \quad (6.19)$$

$$\begin{aligned} \frac{d^2 y(t_i)}{dt^2} &= \frac{d}{dt} \left( \frac{dy(t_i)}{dt} \right) = \frac{df(t_i, y_i)}{dt} = \frac{\partial f(t_i, y_i)}{\partial y} \frac{dy(t_i)}{dt} + \frac{\partial f(t_i, y_i)}{\partial t} = \\ &= \frac{\partial f(t_i, y_i)}{\partial y} f(t_i, y_i) + \frac{\partial f(t_i, y_i)}{\partial t} \end{aligned} \quad (6.20)$$

Підставляючи (6.19) та (6.20) в (6.18), маємо

$$y(t_{i+1}) = y(t_i) + h_i f(t_i, y_i) + \frac{h_i^2}{2} \left( f(t_i, y_i) \frac{\partial f(t_i, y_i)}{\partial y} + \frac{\partial f(t_i, y_i)}{\partial t} \right) + \dots \quad (6.21)$$

Для того, щоб метод мав другий порядок точності ( $R_i = O(h_i^3)$ ), вимагатимемо, щоб в різниці рівнянь (6.18) та (6.21) були відсутні доданки, які містять  $h_i^0$ ,  $h_i^1$  і  $h_i^2$ . Виходячи з цієї умови коефіцієнти  $a$ ,  $b$ ,  $c_1$  і  $c_2$  повинні задовольняти системі рівнянь:

$$\begin{cases} c_1 + c_2 = 1 \\ c_2 b = \frac{1}{2} \\ c_2 a = \frac{1}{2} \end{cases} \quad (6.22)$$

Рівняння (6.22) складають систему з 3 рівнянь відносно 4 невідомих. Виразимо  $b$ ,  $c_1$  і  $c_2$  через  $a \neq 0$ . Отримаємо:

$$b = a, \quad c_2 = \frac{1}{2a}, \quad c_1 = 1 - \frac{1}{2a} \quad (6.23)$$

Підставляючи (6.23) в (6.15) отримаємо такий узагальнений метод Рунге-Кутта другого порядку точності

$$y_{i+1} = y_i + h_i \left( \left( 1 - \frac{1}{2a} \right) k_1 + \frac{1}{2a} k_2 \right), \quad (6.24)$$

де  $k_1 = f(t_i, y_i)$ ,  $k_2 = f(t_i + ah_i, y_i + h_i a k_1)$ .

Широковідомим методом 2 порядку є окремий випадок методу

(6.24) при  $a = \frac{1}{2}$ :

$$y_{i+1} = y_i + h_i k_2,$$

де  $k_1 = f(t_i, y_i)$ ,  $k_2 = f\left(t_i + \frac{h_i}{2}, y_i + \frac{h_i}{2} k_1\right)$ .

Аналогічно розв'язується задача для інших порядків точності. При цьому 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 для  $m \geq 6$ . Таким чином, для метода Рунге-Кутта 5 порядку точності потрібно щонайменше 6 викликів функції на кожному кроці. Аналогічно для  $p=6$   $m \geq 7$ . Це, зокрема, пояснює популярність класичного методу четвертого порядку, оскільки для того, щоб отримати додатковий порядок точності додатково потрібні 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 [18], або більш стійкі формули методів 4-5 порядку для  $m=7$  [13]. 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  $c_j$ :

$$\begin{aligned}
 y_{i+1} &= y_i + \frac{h_i}{6}(k_1 + 2k_2 + 2k_3 + k_4); \\
 k_1 &= f(t_i, y_i); \quad k_2 = f\left(t_i + \frac{h_i}{2}, y_i + \frac{hk_1}{2}\right); \\
 k_3 &= f\left(t_i + \frac{h_i}{2}, y_i + \frac{hk_2}{2}\right); \\
 k_4 &= f(t_i + h_i, y_i + hk_3).
 \end{aligned} \tag{6.25}$$

Проте, слід зазначити, що використання формул (6.25) не дає змогу побудувати досить ефективні алгоритми розв'язку задачі Коші і його слід використовувати тільки для напів-аналітичних розрахунків або на початкових етапах розробки програмного забезпечення для отримання первинних оцінок.

**Приклад:**

$$\begin{cases} \frac{dy}{dt} = -(t+1)y^2; \\ y(0) = 1 \end{cases};$$

$$t \in [0, 1], \quad h = 0,5.$$

$$\begin{aligned}
 t_0 &= 0; \\
 y_0 &= y(0) = 1; \\
 t_1 &= 0,5;
 \end{aligned}$$

$$k_1 = f(t_0, y_0) = -(t_0 + 1)y_0^2 = -(0 + 1) \cdot 1^2 = -1;$$

$$\begin{aligned} k_2 &= f\left(t_0 + \frac{h}{2}, y_0 + \frac{hk_1}{2}\right) = -\left(t_0 + \frac{h}{2} + 1\right)\left(y_0 + \frac{hk_1}{2}\right)^2 = \\ &= -\left(0 + \frac{0,5}{2} + 1\right)\left(1 - \frac{0,5 \cdot 1}{2}\right)^2 = -0,703125; \end{aligned}$$

$$\begin{aligned} k_3 &= f\left(t_0 + \frac{h}{2}, y_0 + \frac{hk_2}{2}\right) = -\left(t_0 + \frac{h}{2} + 1\right)\left(y_0 + \frac{hk_2}{2}\right)^2 = \\ &= -\left(0 + \frac{0,5}{2} + 1\right)\left(1 - \frac{0,5 \cdot 0,703125}{2}\right)^2 \cong -0,849171; \end{aligned}$$

$$\begin{aligned} k_4 &= f(t_0 + h, y_0 + hk_3) = -(t_0 + h + 1)(y_0 + hk_3)^2 = \\ &= -(0 + 0,5 + 1)(1 - 0,5 \cdot 0,849171)^2 \cong -0,496653; \end{aligned}$$

$$\begin{aligned} y_1 &= y_0 + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + 2k_4) = \\ &= 1 + \frac{0,5}{6}(-1 - 2 \cdot 0,703125 - 2 \cdot 0,849171 - 0,496653) \cong 0,616563; \end{aligned}$$

$$t_2 = 1;$$

$$k_1 = f(t_1, y_1) = -(t_1 + 1)y_1^2 = -(0,5 + 1) \cdot 0,616563^2 = -0,570225;$$

$$\begin{aligned} k_2 &= f\left(t_1 + \frac{h}{2}, y_1 + \frac{hk_1}{2}\right) = -\left(t_1 + \frac{h}{2} + 1\right)\left(y_1 + \frac{hk_1}{2}\right)^2 = \\ &= -\left(0 + \frac{0,5}{2} + 1\right)\left(0,616563 - \frac{0,5 \cdot 570225}{2}\right)^2 = -0,393194; \end{aligned}$$

$$\begin{aligned} k_3 &= f\left(t_1 + \frac{h}{2}, y_1 + \frac{hk_2}{2}\right) = -\left(t_1 + \frac{h}{2} + 1\right)\left(y_1 + \frac{hk_2}{2}\right)^2 = \\ &= -\left(0 + \frac{0,5}{2} + 1\right)\left(0,616563 - \frac{0,5 \cdot 393194}{2}\right)^2 \cong -0,470047; \end{aligned}$$

$$\begin{aligned} k_4 &= f(t_1 + h, y_1 + hk_3) = -(t_1 + h + 1)(y_1 + hk_3)^2 = \\ &= -(0 + 0,5 + 1)(0,616563 - 0,5 \cdot 0,470047)^2 \cong -0,291145; \end{aligned}$$

$$\begin{aligned} y_2 &= y_1 + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + 2k_4) = \\ &= 0,616563 + \frac{0,5}{6}(-0,570225 - 2 \cdot 0,393194 - 2 \cdot 0,470047 - 0,291145) \cong 0,400909; \end{aligned}$$

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 точці.

### 6.3. Багатоточкові методи

В одноточкових методах значення  $y_{i+1}$  обчислюють на основі значення в одній попередній точці  $y_i$ . При цьому логічно припустити, що можна підвищити точність, якщо в розрахунок значення в точці  $y_{i+1}$  на кожному кроці бере участь не одна, а кілька попередніх точок –  $y_i, y_{i-1}, \dots$ .

В загальному випадку лінійним  $m$ -кроковим різницевим методом з постійним кроком називається система різницевих рівнянь

$$\frac{a_0 y_{i+1} + a_1 y_i + \dots + a_m y_{i-m+1}}{h} = b_0 f_{i+1} + b_1 f_i + \dots + b_m f_{i-m+1}; \quad i = \overline{m-1, n-1}; \quad (6.26)$$

де  $a_k, b_k$  — числові 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 знайдені  $y_i, y_{i-1}, y_{i-2}, \dots, y_{i-m+1}$ . Обчислення починається з  $i=m-1$ , тобто з рівняння

$$\frac{a_0 y_m + a_1 y_{m-1} + \dots + a_m y_0}{h} = b_0 f_m + b_1 f_{m-1} + \dots + b_m f_0.$$

Звідси видно, що для початку обчислень слід задати  $m$  початкових значень  $y_0, y_1, \dots, y_{m-1}$ . Значення  $y_0$  задається умовою (6.2). Величини  $y_1, y_2, \dots, y_{m-1}$  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  $y_i, y_{i-1}, \dots, y_{i-m+1}$

$$y_{i+1} = \frac{1}{a_0} \sum_{k=1}^m [-a_k y_{i-k+1} + h b_k f_{i-k+1}] \quad (6.27)$$

Якщо  $b_0 \neq 0$ , то метод називається неявним. Тоді для знаходження  $y_{i+1}$  доводиться розв'язувати нелінійне рівняння

$$y_{i+1} a_0 - b_0 h f(y_{i+1}, t_{i+1}) = \sum_{k=1}^m [-a_k y_{i-k+1} + h b_k f_{i-k+1}] \quad (6.28)$$

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

$$\sum_{k=0}^m b_k = 1 \quad (6.29)$$

Можна показати, що для того, аби різницьвий метод (6.26) мав порядок точності  $p$ , необхідно щоб були виконані умови [8]:

$$\sum_{k=0}^m a_k = 0 \quad (6.30)$$

$$\sum_{k=0}^m k^{l-1} (k a_k + l b_k) = 0, \quad l = \overline{1, p} \quad (6.31)$$

Разом з умовою нормування (6.29) рівняння (6.30) і (6.31) становлять систему з  $p+2$  лінійних алгебраїчних рівнянь відносно  $2m+2$  невідомих  $a_0, a_1, \dots, a_m, b_0, b_1, \dots, b_m$ .

Для того, щоб існував розв'язок цієї системи, необхідно, щоб кількість рівнянь була не більшою від кількості невідомих, тому  $p \leq 2m$ . Ця вимога означає, що лінійні  $m$ -крокові різницьві методи мають порядок точності не більше, ніж  $2m$ . Тому порядок точності неявних  $m$ -крокових методів дорі-



вноє  $2m$ , коли  $a_i \neq 0$ ,  $b_i \neq 0$ ,  $i = \overline{0, m}$ , а явних —  $2m-1$ , 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  $y'$  апроксимується тільки по двох точках  $t_{i+1}$ ,  $t_i$ , тобто

$$a_0 = -a_1 = 1, \quad a_k = 0, \quad k = \overline{2, m}.$$

Таким чином, методи Адамса мають вигляд

$$\frac{y_{i+1} - y_i}{h} = \sum_{k=0}^m b_k f_{i-k+1}. \quad (6.32)$$

У випадку  $b_0=0$  методи Адамса називають явними, а при  $b_0 \neq 0$  — неявними.

Для методів Адамса умови (6.29), (6.31) приймають вигляд

$$l \sum_{k=1}^m k^{l-1} b_k = 1; \quad l = \overline{1, p} \quad (6.33)$$

$$b_0 = 1 - \sum_{k=1}^m b_k \quad (6.34)$$

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) при  $p=m$ . При  $m=1$  отримуємо метод Ейлера

$$\frac{y_{i+1} - y_i}{h} = f_i$$

При  $m=2$  (6.33) приймає вигляд

$$\begin{aligned} b_1 + b_2 &= 1 \\ 2(b_1 + 2b_2) &= 1 \end{aligned}$$

Звідки  $b_1 = \frac{3}{2}$ ;  $b_2 = -\frac{1}{2}$

Таким чином, при  $m=2$ , 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

$$y_{i+1} = y_i + h \left( \frac{3}{2} f(t_i, y_i) - \frac{1}{2} f(t_{i-1}, y_{i-1}) \right);$$

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  $m$ :

$$m = 1:$$

$$y_{i+1} = y_i + \frac{h}{2} (f(t_{i+1}, y_{i+1}) + f(t_i, y_i));$$

$$m = 2:$$

$$y_{i+1} = y_i + \frac{h}{12} (5f(t_{i+1}, y_{i+1}) + 8f(t_i, y_i) - f(t_{i-1}, y_{i-1}));$$

$$m = 3:$$

$$y_{i+1} = y_i + \frac{h}{24} (9f(t_{i+1}, y_{i+1}) + 19f(t_i, y_i) - 5f(t_{i-1}, y_{i-1}) + f(t_{i-2}, y_{i-2})).$$

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 прогнозне значення  $y_{i+1}^n$  за допомогою явного методу, що називається предиктором. Знайдене значення використовується як початкове наближення для його уточнення не-

явним методом, що називається коректором. Такі методи називають методами предиктор-коректор. Для  $m=3$  обчислювальна схема матиме вигляд:

$$y_{i+1}^n = y_i + \frac{h_i}{12} (23f(t_i, y_i) - 16f(t_{i-1}, y_{i-1}) + 5f(t_{i-2}, y_{i-2}));$$

$$y_{i+1} = y_i + \frac{h}{24} (9f(t_{i+1}, y_{i+1}^n) + 19f(t_i, y_i) - 5f(t_{i-1}, y_{i-1}) + f(t_{i-2}, y_{i-2})).$$

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

$$\begin{cases} \frac{dy}{dt} = -(t+1)y^2 \\ y(0) = 1 \end{cases}$$

Необхідно обчислити таблицю значень  $y(t)$  на інтервалі  $t \in [0, 1]$ , якщо  $h=0,25$ .

Застосуємо трьохточковий явний метод Адамса. Використовуючи таблицю 6.1 отримаємо загальну формулу для цього методу:

$$y_{i+1} = y_i + \frac{h_i}{12} (23f(t_i, y_i) - 16f(t_{i-1}, y_{i-1}) + 5f(t_{i-2}, y_{i-2}));$$

Спочатку обчислимо значення  $y_1, y_2$  явним методом Ейлера:

$$y_0 = y(0) = 1;$$

$$y_1 = y(0,25) = y_0 - h(t_0 + 1)y_0^2 = 1 - 0,25 \cdot (0 + 1) \cdot 1^2 = 0,75;$$

$$y_2 = y(0,5) = y_1 - h(t_1 + 1)y_1^2 = 0,75 - 0,25 \cdot (0,25 + 1) \cdot 0,75^2 \cong 0,574219;$$

Після цього продовжимо розв'язок заданим методом:

$$t_3 = 0,75;$$

$$\begin{aligned} y_3 &= y_2 + \frac{h}{12} \left( 23(-(t_2 + 1)y_2^2) - 16(-(t_1 + 1)y_1^2) + 5(-(t_0 + 1)y_0^2) \right) = \\ &= 0,574219 + \frac{0,25}{12} \left( -23(0,5 + 1) \cdot 0,574219^2 + 16(0,25 + 1) \cdot 0,75^2 - 5(0 + 1) \cdot 1^2 \right) \cong \\ &\cong 0,467436; \end{aligned}$$

$$t_4 = 1;$$

$$\begin{aligned} y_4 &= y_3 + \frac{h}{12} \left( 23(-(t_3 + 1)y_3^2) - 16(-(t_2 + 1)y_2^2) + 5(-(t_1 + 1)y_1^2) \right) = \\ &= 0,467436 + \frac{0,25}{12} \left( -23(0,75 + 1) \cdot 0,467436^2 + 16(0,5 + 1) \cdot 0,574219^2 - 5(0,25 + 1) \cdot 0,75^2 \right) \cong \\ &\cong 0,375839. \end{aligned}$$

#### 6.4. Апостеріорна оцінка похибки розв'язання задачі Коші. Автоматичний вибір кроку інтегрування.

При чисельному розв'язанні задачі Коші використовують різні методи апостеріорної оцінки похибки в залежності від обраного методу та інших факторів. Найбільш широко застосовується правило Рунге:

$$R_i \approx \frac{y_{i+1, \frac{h}{q}} - y_{i+1, h}}{q^p - 1} \quad (6.35)$$

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 уже були проведені в параграфі 6.1. Результатом є  $y(1)_{\frac{h}{2}} = 0,361752$ . Тоді, враховуючи, що метод Ейлера має перший поряд-

док 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 від-

значити, що апостеріорна оцінка похибки тим ближча до дійсної, чим менший крок  $h$ .

При розв'язанні задачі Коші неявними методами використання правила Рунге є неефективним, оскільки при зменшенні кроку доводиться додатково розв'язувати  $q$  нелінійних рівнянь. Тому при використанні неявних методів застосовують інші методи оцінки похибки. Так, наприклад, при порівнянні формул для похибок явного та неявного методів Ейлера видно, що при  $h \rightarrow 0$  модулі цих похибок є однаковими, але вони мають протилежний знак. Ця особливість дає змогу апостеріорно оцінити похибку неявного методу за формулою:

$$R_i \approx \frac{\tilde{y}_{i+1} - y_{i+1}}{2} \quad (6.36)$$

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 точці  $t=1$  при розв'язанні задачі з кроком  $h=0,25$ . Використовуючи приклад з параграфу 6.1 маємо  $\tilde{y}_4=0,361752$  а  $y_4=0,431175$ . Тоді

$$R_i \approx \frac{0,361752 - 0,431175}{2} \cong -0,034712$$

Як видно ця оцінка близька до дійсної похибки.

При застосуванні методів Рунге-Кутта використовують інший підхід. В ньому оцінка похибки розраховується як різниця результатів отриманих методами з різними порядками точності:

$$R_i \approx y_{i+1} - y_{i+1}^* \quad (6.37)$$

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:

$$R_i \approx h_i \sum_{j=1}^m (c_j - c_j^*) k_j$$

Існують і інші підходи для оцінки похибки розв’язання задачі Коші. Так наприклад можливо оцінити похідну, яка входить в формулу похибки. Так, наприклад, для явного методу Ейлера досить оцінити  $\frac{d^2 y(t_i)}{dt^2}$ . Для цього можна скористатися апроксимацією похідної функції  $f(t, y)$  скінченними різницями:

$$y''(t_i) = \frac{d}{dt} \left( \frac{dy(t_i)}{dt} \right) = \frac{d}{dt} (f(t_i, y_i)) = \frac{\partial f}{\partial y} \frac{dy}{dt} \Big|_{t=t_i} + \frac{\partial f}{\partial t} \Big|_{t=t_i} = f(t_i, y_i) \frac{\partial f}{\partial y} \Big|_{t=t_i} + \frac{\partial f}{\partial t} \Big|_{t=t_i}.$$

$$\text{де } \frac{\partial f}{\partial y} \Big|_{t=t_i} \approx \frac{f(t_i, y_i + \Delta y_i) - f(t_i, y_i)}{\Delta y_i}, \quad \frac{\partial f}{\partial t} \Big|_{t=t_i} \approx \frac{f(t_i + \Delta t_i, y_i) - f(t_i, y_i)}{\Delta t_i},$$

$$\Delta y_i = \sqrt{\varepsilon_{\text{max}}} y_i, \Delta t_i = \sqrt{\varepsilon_{\text{max}}} t_i.$$

Інший спосіб оцінки похідної ґрунтується на диференціюванні інтерполяційних поліномів. Так, наприклад, для оцінки  $\frac{d^2 y(t_i)}{dt^2}$  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:

$$|R_i| \leq \varepsilon |y_{i+1}| \quad (6.38)$$

де  $\varepsilon$  – задана відносна похибка розв’язку.

Якщо умова (6.38) не виконується, то крок зменшують в  $q$  разів або в  $\sqrt[p+1]{\frac{\varepsilon |y_{i+1}|}{|R_i|}}$  раз. Зменшення кроку виконують до тих пір, поки не виконається умова (6.38). Після чого переходять до наступного елементарного відрізка.

Крок на наступному інтервалі вибирають за формулою

$$h_{i+1} = \left( \frac{\varepsilon |y_{i+1}|}{|R_i|} \right)^{\frac{1}{p+1}} h_i,$$

або за правилом зон:

$$h_{i+1} = \begin{cases} h_i, & \frac{\varepsilon}{q^{p+1}} \leq \left| \frac{R_i}{y_{i+1}} \right| \leq \varepsilon; \\ qh_i, & \left| \frac{R_i}{y_{i+1}} \right| < \frac{\varepsilon}{q^{p+1}}. \end{cases}$$

Якщо необхідно обчислити значення функції в наперед заданій вузлах, можна використовувати метод екстраполяції нульового кроку. 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:

$$y_{i+1, h/2^k}^{(j)} = y_{i+1, h/2^k}^{(j-1)} + \frac{y_{i+1, h/2^k}^{(j-1)} - y_{i+1, h/2^{k-1}}^{(j-1)}}{2^{p+j} - 1}, \quad k = \overline{j, n}$$

Крок послідовно зменшують доти, поки нова екстраполяція не буде відрізнятися від попередньої не більше, ніж на задану величину похибки розв’язку. При цьому остання екстраполяція є екстраполяцією нульового кроку. В цьому методі автоматичний вибір кроку не потрібний.

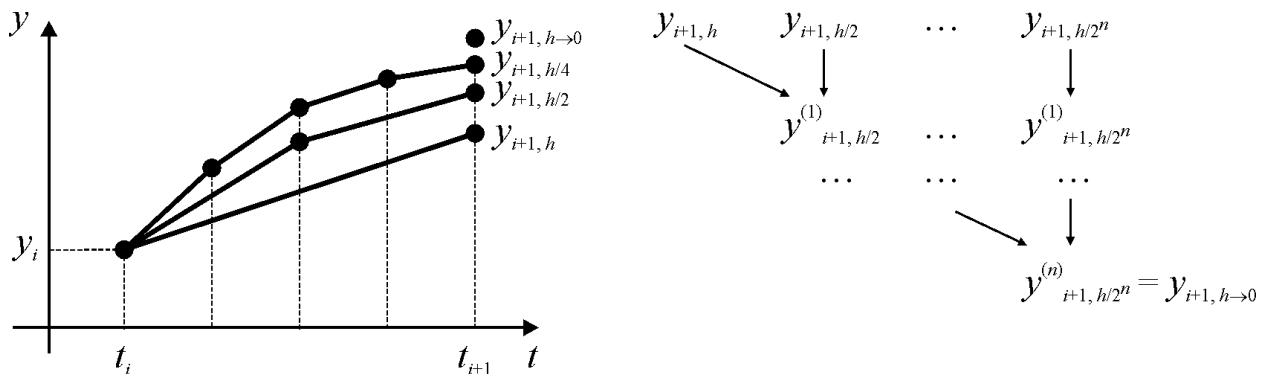


Рис. 6.1. Екстраполяція нульового кроку.

## 6.5. Жорсткі рівняння

Поняття жорсткого рівняння пов'язано з жорсткою умовою стійкості для різницевих схем розв'язання цих рівнянь [1]. Розглянемо задачу Коші:

$$\frac{dy}{dt} = \lambda y, \quad y(0) = y_0, \quad (6.39)$$

де  $\lambda$  – досить велике від'ємне число. Точний розв'язок задачі (6.39) має вигляд

$$y(t) = y_0 e^{\lambda t}. \quad (6.40)$$

Із виразу (6.40) випливає, що  $\forall t_i$  і  $h_i > 0$ :

$$|y(t_{i+1})| = |y(t_i + h_i)| < |y(t_i)|. \quad (6.41)$$

Застосуємо для розв'язання задачі (6.39) явний метод Ейлера:

$$y_{i+1} = (1 + \lambda h_i) y_i = (1 + \lambda h_i)^i y_0. \quad (6.42)$$

Із виразу (6.42) випливає, що умова (6.41) виконується, якщо

$$|1 + \lambda h_i| < 1.$$

Отже, процес (6.42) буде стійким для виконання співвідношення (6.41) тільки за умови, що

$$h_i \leq \frac{2}{|\lambda|}. \quad (6.43)$$

За великих значень  $\lambda$  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 інтервалу інтегрування:

$$S = |\lambda| t_n,$$

де  $t_n$  – права границя інтервалу інтегрування.

Для системи диференціальних рівнянь жорсткість можна оцінити як

$$S = \frac{\max_{i, t \in [0, t_n]} (-\operatorname{Re} \lambda_i(t))}{\min_{i, t \in [0, t_n]} |\operatorname{Re} \lambda_i(t)|},$$



де  $\lambda_i$  — 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

$$\sum_{k=0}^m a_k y_{i-k+1} = hf(y_{i+1}, t_{i+1}), \quad (6.44)$$

яка є окремим випадком багатоточкового методу (6.26) при

$$b_i = 0, i = \overline{1, m}, b_0 = 1. \quad (6.45)$$

Для методу (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

$$\begin{cases} \frac{dy}{dt} = -(t+1)y^2 \\ y(0) = 1 \end{cases}$$

Необхідно обчислити таблицю значень  $y(t)$  на інтервалі  $t \in [0, 1]$  методом Гіра 3 порядку, якщо  $h=0,25$ .

Використовуючи таблицю 6.2 отримаємо з (6.44) загальну формулу для цього методу:

$$\frac{11}{6}y_{i+1} - 3y_i + \frac{3}{2}y_{i-1} - \frac{1}{3}y_{i-2} - hf(t_{i+1}, y_{i+1}) = 0;$$

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:

$$t_3 = 0,75;$$

$$\frac{11}{6}y_3 - 3y_2 + \frac{3}{2}y_1 - \frac{1}{3}y_0 - h(-(t_3 + 1)y_3^2) = 0; \Rightarrow$$

$$\Rightarrow 0,25(0,75 + 1)y_3^2 + \frac{11}{6}y_3 - 3 \cdot 0,644320 + \frac{3}{2} \cdot 0,8 - \frac{1}{3} \cdot 1 = 0; \Rightarrow$$

$$\Rightarrow y_3 \cong 0,517665;$$

$$t_4 = 1;$$

$$\frac{11}{6}y_4 - 3y_3 + \frac{3}{2}y_2 - \frac{1}{3}y_1 - h(-(t_4 + 1)y_4^2) = 0; \Rightarrow$$

$$\Rightarrow 0,25(1 + 1)y_4^2 + \frac{11}{6}y_4 - 3 \cdot 0,517665 + \frac{3}{2} \cdot 0,644320 - \frac{1}{3} \cdot 0,8 = 0; \Rightarrow$$

$$\Rightarrow y_4 \cong 0,417772;$$

## 6.6. Крайові задачі

Часто фізичні задачі зводять до диференціального рівняння  $m$ -го порядку

$$\frac{d^m y}{dx^m} = f\left(x, y, \frac{dy}{dx}, \dots, \frac{d^{m-1}y}{dx^{m-1}}\right), \quad (6.46)$$

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

$$\sum_{k=0}^{m-1} \left[ \alpha_{kj} \frac{d^k y(a)}{dx^k} + \beta_{kj} \frac{d^k y(b)}{dx^k} \right] = c_j, \quad j = \overline{1, m}, \quad (6.47)$$

де  $\alpha_{kj}, \beta_{kj}, c_j$  – 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 та спеціальні методи, призначені для розв'язання власне крайових задач [1, 3].

### 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:

$$Ly = f(x) \quad (6.48)$$

$$\sum_{i=0}^{m-1} \left[ \alpha_{ij} \frac{d^i y(0)}{dx^i} + \beta_{ij} \frac{d^i y(l)}{dx^i} \right] = \gamma_j, \quad j = \overline{0, m-1} \quad (6.49)$$

де  $L$  – диференціальний оператор і  $\alpha_{ij}, \beta_{ij}, \gamma_j$  – сталі.

Задачу (6.48), з умовою (6.49) найпростіше звести до задачі Коші, якщо  $L$  – лінійний диференціальний оператор

$$L = a_m(x) \frac{d^m}{dx^m} + a_{m-1}(x) \frac{d^{m-1}}{dx^{m-1}} + \dots + a_1(x) \frac{d}{dx} + a_0(x).$$

У цьому випадку розв'язок шукають у вигляді

$$y(x) = z_0(x) + \sum_{k=1}^m c_k z_k(x), \quad (6.50)$$

де  $z_k(x)$  знаходять з  $(m+1)$  задачі Коші

$$\begin{aligned} Lz_0 &= f(x), \quad z_0^{(i)}(0) = 0, \quad i = \overline{0, m-1} \\ Lz_k &= 0, \quad z_k^{(i)}(0) = \begin{cases} 0, & i \neq k-1 \\ 1, & i = k-1 \end{cases}, \quad i = \overline{0, m-1}, \quad k = \overline{1, m} \end{aligned} \quad (6.51)$$

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:**

$$\frac{d^2 y}{dx^2} + 4y = x;$$

$$y(0) = 1;$$

$$y(1) = 2.$$

Шукаємо розв'язок у вигляді

$$y(x) = z_0(x) + c_1 z_1(x) + c_2 z_2(x).$$

Із розв'язків задач Коші на відрізку  $[0, 1]$

$$\frac{d^2 z_0}{dx^2} + 4z_0 = x, \quad z_0(0) = 0, \quad z_0'(0) = 0;$$

$$\frac{d^2 z_1}{dx^2} + 4z_1 = 0, \quad z_1(0) = 1, \quad z_1'(0) = 0;$$

$$\frac{d^2 z_2}{dx^2} + 4z_2 = 0, \quad z_2(0) = 0, \quad z_2'(0) = 1.$$

отримуємо  $z_0(x)$ ,  $z_1(x)$ ,  $z_2(x)$ . Далі знаходимо сталі  $c_1$  та  $c_2$ .

З першої крайової умови маємо

$$0 + c_1 \cdot 1 + c_2 \cdot 0 = 1 \Rightarrow c_1 = 1.$$

З другої умови:

$$z_0(1) + c_1 z_1(1) + c_2 z_2(1) = 2 \Rightarrow c_2 = \frac{2 - z_0(1) - z_1(1)}{z_2(1)}.$$

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

$$Lz = f(x); \tag{6.52}$$

$$z^{(i)}(0) = z_i, \quad i = \overline{0, m-1}. \tag{6.53}$$

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

$$\frac{d^2 y}{dx^2} = f(x, y, y')$$

за умов:

$$y(0) = y_0;$$

$$y(l) = y_l.$$

Можна застосувати такий ітераційний метод:

1) вибирати  $\xi$ , 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  $|z(l) - y_l| < \varepsilon$ , де  $\varepsilon$  – задана похибка, то покласти  $y(x) \approx z(x)$ , у противному випадку – змінити  $\xi$  (виконати ітерацію розв'язку нелінійного рівняння) і повернутися до пункту 2).

### 6.6.2. Метод скінченних різниць

Найпростішим і 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  $[0, l]$ . Для цього відрізок  $[0, l]$  ділять на  $n \geq m+1$  рівних відрізків вузлами  $x_0 = 0, x_1 = x_0 + h, x_2 = x_0 + 2h, \dots, x_n = x_0 + nh$ . 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 скінчен-

норізницевої апроксимації похідної  $y_i^{(j)} = \frac{d^j y(x_i)}{dx^j}$  будемо інтерполяційний поліном Лагранжа  $P_k(x), k \geq j$ , який проходить через  $k+1$  точку  $y_{i-\left[\frac{k+1}{2}\right]}, y_{i-\left[\frac{k+1}{2}\right]+1}, \dots, y_i, \dots, y_{i+\left[\frac{k+1}{2}\right]-1}, y_{i+\left[\frac{k+1}{2}\right]}$ .

Тоді

$$y_i^{(j)} = \frac{d^j}{dx^j} P_k(x_i).$$

Похибка такої апроксимації оцінюється як

$$R_i^{(j)} \sim f^{(k)}(\xi) h^{k+1},$$

$$\text{де } \xi \in \left[ x_{i-\left[\frac{k+1}{2}\right]}, x_{i+\left[\frac{k+1}{2}\right]} \right].$$

Як приклад знайдемо скінченнорізницеву 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  $O(h^2)$ .

У методі скінченних різниць диференціальне рівняння (6.48) замінюється скінченнорізницевими рівняннями, записаними для  $n - m + 1$  внутрішнього вузла відрізка  $[0, l]$ . Граничні умови (6.49) замінюються  $m$  скінченнорізницевими рівняннями. У підсумку одержуємо СЛАР 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:**

$$\frac{d^2 y}{dx^2} + y = x;$$

$$y(0) = 0;$$

$$y(1) = \frac{1}{3}.$$

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 скінченнорізницеву апроксимацію  $y''_i = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}$ , запишемо скінченнорізницеве рівняння для вузлів  $x_1, x_2$ :

$$\frac{y_0 - 2y_1 + y_2}{\left(\frac{1}{3}\right)^2} + y_1 = x_1; \quad -17y_1 + 9y_2 = \frac{1}{3};$$

$$\frac{y_1 - 2y_2 + y_3}{\left(\frac{1}{3}\right)^2} + y_2 = x_2; \quad 9y_1 - 17y_2 = -\frac{7}{3}.$$

З крайових умов:

$$y_0 = 0;$$

$$y_3 = 1.$$

Одержимо систему чотирьох рівнянь стосовно чотирьох невідомих  $y_0, y_1, y_2, y_3$ . Розв'язуючи її, одержимо:

$$y_1 = \frac{23}{312} = 0,0737; \quad y\left(\frac{1}{3}\right) = 0,0741;$$

$$y_2 = \frac{55}{312} = 0,1763; \quad y\left(\frac{2}{3}\right) = 0,1768.$$

### 6.6.3. Проекційні методи розв'язання крайових задач

Подамо крайові умови (6.49) у вигляді

$$L_s y = B_s, \quad (6.54)$$

де  $L_s$  – векторний лінійний диференціальний оператор,  $j$ -й елемент якого дорівнює

$$L_{sj} = \sum_{i=0}^{m-1} \left[ \alpha_{ij} \frac{d^i}{dx^i} \Big|_{x=0} + \beta_{ij} \frac{d^i}{dx^i} \Big|_{x=l} \right];$$

$B_s = [\gamma_0, \gamma_1, \dots, \gamma_{m-1}]^T$  – вектор сталих величин.

Для побудови проекційних методів уведемо дві повні системи лінійно-незалежних функцій – базисну (координатну)  $\{\varphi_i(x)\}$  і проекційну (вагову)  $\{\psi_i(x)\}$ . Подамо розв'язок задачі (6.48) з умовою (6.54) у вигляді розкладання за системою базисних функцій:

$$y(x) = \sum_{i=1}^{\infty} c_i \varphi_i(x), \quad (6.55)$$

де  $c_i$  – сталі, які треба визначити.

Підставивши вираз (6.55) у (6.48) і (6.54), одержимо:

$$L \sum_{i=1}^{\infty} c_i \varphi_i(x) - f(x) = R(x);$$

$$L_s \sum_{i=1}^{\infty} c_i \varphi_i(x) - B_s = 0, \quad (6.56)$$

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 числового розв’язання задачі, очевидно, неможливе. Алгоритм проєкційних 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 проєкції відхилу на множину проєкційних функцій  $\{\psi_i(x)\}$ . Якщо цю вимогу виконано, усі скалярні добутки  $(R(x), \psi_j(x))$ ,  $j = \overline{1, \infty}$  дорівнюють нулю. Звідси

$$\left( L \sum_{i=1}^{\infty} c_i \varphi_i(x), \psi_j(x) \right) - (f(x), \psi_j(x)) = 0, \quad (6.57)$$

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 проєкційних методів. Відомі окремі випадки методу моментів, що мають спеціальні назви. Так, у методі Гальоркіна – Бубнова базисна система функцій збігається з проєкційною, 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. При цьому в розкладанні (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 проєкційного відхилу на  $n - m$  перших проєкційних 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



$$\max_{x \in [0, l]} |y_{n1}(x) - y_{n2}(x)| < \xi. \quad (6.58)$$

Якщо для деяких  $n_1, n_2$  умова (6.58) виконується, то кажуть, що проекційний метод збігається. Якщо з ростом  $n$   $|y_{n1}(x) - y_{n2}(x)|$  збільшується, то кажуть, що метод розбігається і є нестійким.

У багатьох задачах задають однорідні крайові умови, тобто такі, що з (6.54) одержані при  $B_s = 0$ . Крім того, багато крайових умов можна звести до однорідних за допомогою заміни змінних. 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 щодо  $c_i, i = \overline{1, \infty}$ :

$$\sum_{i=1}^{\infty} c_i a_{ij} = b_j, \quad j = \overline{1, \infty}, \quad (6.59)$$

де  $a_{ij} = (L \varphi_i(x), \psi_j(x)), \quad b_j = (f(x), \psi_j(x))$ .

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), то ця функція зводить до мінімуму функціонал енергії:

$$F(x) = (Ly, y) - 2(y, f(x)). \quad (6.60)$$

Підставляючи рівняння (6.55) у (6.60), одержимо

$$F(y) = F(c_1, c_2, \dots, c_n, \dots). \quad (6.61)$$

Коефіцієнти  $c_i$  визначають з умови існування мінімуму функціонала (6.61)

$$\frac{\partial F(c_1, c_2, \dots, c_n, \dots)}{\partial c_j} = 0, \quad j = \overline{1, \infty}, \quad (6.62)$$

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:**

$$\frac{d^2 y}{dx^2} + y = x; \quad y(0) = 0; \quad y(1) = \frac{1}{3}.$$

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

$$\frac{d^2 \tilde{y}}{dx^2} + \tilde{y} = \frac{2}{3}x; \quad \tilde{y}(0) = 0; \quad \tilde{y}(1) = 0.$$

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 розкладання для  $\tilde{y}(x)$  у вихідне диференціальне рівняння, одержуємо:

$$\sum_{i=1}^{\infty} \left[ c_i (1 - i^2 \pi^2) \sin(i\pi x) \right] - \frac{2}{3}x = 0.$$

Оскільки  $\int_0^1 \sin(i\pi x) \sin(j\pi x) dx = 0$  при  $i \neq j$ , з умови рівності нулю проекції відхили на систему функцій  $\{\sin(i\pi x)\}$  одержуємо:

$$c_i (1 - i^2 \pi^2) \int_0^1 \sin^2(i\pi x) dx - \int_0^1 \frac{2}{3} x \sin(i\pi x) dx = 0, \quad i = \overline{1, \infty}.$$

Обчислюючи інтеграли, знаходимо:

$$a_{ij} c_i = b_i, \quad i = \overline{1, \infty};$$

$$a_{ij} = \frac{1}{2} (1 - i^2 \pi^2), \quad b_i = \frac{2}{3} \frac{(-1)^{i+1}}{i\pi};$$

$$c_i (1 - i^2 \pi^2) \frac{1}{2} = \frac{2}{3} \frac{(-1)^{i+1}}{i\pi}, \quad i = \overline{1, \infty}.$$

Розв'язуючи СЛАР, маємо

$$c_i = \frac{b_i}{a_{ij}} = \frac{4(-1)^{i+1}}{3i\pi(1 - i^2 \pi^2)}.$$

Тоді:

$$\tilde{y}(x) \approx \sum_{i=1}^n \frac{4(-1)^{i+1}}{3i\pi(1-i^2\pi^2)} \sin(i\pi x);$$

$$y(x) \approx \sum_{i=1}^n \frac{4(-1)^{i+1}}{3i\pi(1-i^2\pi^2)} \sin(i\pi x) + \frac{x}{3}.$$

$n$	$y(1/3)$	$y(2/3)$
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, 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  $\varphi_{ij}(x)$ ,  $1, r$  (рис. 6.1), які дорівнюють нулю поза цим елементом і мають властивість

$$\varphi_{ij}(x_{ik}) = \begin{cases} 0, & j \neq k, \\ 1, & j = k, \end{cases} \quad (6.63)$$

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

$$y(x) = \sum_{i=1}^n \sum_{j=1}^r y_{ij} \varphi_{ij}(x). \quad (6.64)$$

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:

$$y(x) = \sum_{i=1}^n \sum_{j=2}^{r-1} y_{ij} \varphi_{ij}(x) + \frac{1}{2} \sum_{i=1}^n y_{ir} (\varphi_{ir}(x) + \varphi_{i+1,r}(x)) + y_{11} \varphi_{11}(x) + y_{nr} \varphi_{nr}(x). \quad (6.65)$$

Підставляючи формулу (6.65) у (6.54) і враховуючи, що в розкладанні (6.65) залишаються тільки базисні функції, носіями яких є елементи  $D_1$  і  $D_n$ , що лежать на межі, одержимо  $m$  рівнянь:

$$L_s \left( \sum_{j=1}^r (y_{1j} \varphi_{1j}(x) + y_{nj} \varphi_{nj}(x)) \right) = B_s. \quad (6.66)$$

Далі використаємо стандартну процедуру проекційних методів, причому за проекційний базис візьмемо базисні функції, носіями яких є всі  $n(r-2)$  внутрішні вузли всіх скінченних елементів  $D_1, D_3, \dots, D_n$ , а також  $n-1$  базисна функція  $(\varphi_{i1}(x) + \varphi_{i+1,r}(x))$  з носіями у вузлах, що одночасно належать сусіднім елементам. Дві функції  $\varphi_{11}(x)$  і  $\varphi_{nr}(x)$  у проекційний базис не включають, оскільки за їх допомогою вже визначено значення функції на кінцях відрізка (із граничних умов). При цьому враховують, що всі скалярні добутки базисних функцій на носіях, що не збігаються, дорівнюють нулю. У підсумку одержимо систему рівнянь:

$$\begin{aligned} & \left( L \sum_{j=2}^{r-1} y_{ij} \varphi_{ij}(x), \varphi_{is}(x) \right) - (f(x), \varphi_{is}(x)) = 0, \quad i = \overline{1, n}, \quad s \neq 1, \quad s \neq r; \\ & \left( L \sum_{j=1}^r y_{ij} \varphi_{ij}(x), \varphi_{ir}(x) \right) + \left( L \sum_{j=1}^{r-1} y_{i+1,j} \varphi_{i+1,j}(x), \varphi_{i+1,1}(x) \right) - \\ & \quad - (f(x), \varphi_{ir}(x) + \varphi_{i+1,1}(x)) = 0, \quad i = \overline{2, n-1}. \end{aligned} \quad (6.67)$$

Таким чином, одержали систему  $m(r-1) - 1$  нелінійних рівнянь щодо  $m(r-1) - 1$  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:**

$$\begin{cases} \frac{d^2 y}{dx^2} + y = x; \\ y(0) = 0; \\ y(1) = \frac{1}{3}; \end{cases} \Rightarrow \begin{cases} Ly = f(x); \\ L = \frac{d^2}{dx^2} + 1; \\ f(x) = x. \end{cases}$$

Розіб'ємо відрізок  $[0, 1]$  на два скінченні елементи –  $D_1 : x \in \left[0, \frac{1}{2}\right]$ ,  $D_2 : x \in \left[\frac{1}{2}, 1\right]$ . 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:

$$\varphi_{11}(x) = \frac{\left(x - \frac{1}{3}\right)\left(x - \frac{1}{2}\right)}{\left(-\frac{1}{3}\right)\left(-\frac{1}{2}\right)} = 6\left(x - \frac{1}{3}\right)\left(x - \frac{1}{2}\right);$$

$$\varphi_{21}(x) = \frac{\left(x - \frac{2}{3}\right)(x - 1)}{\left(\frac{1}{2} - \frac{2}{3}\right)\left(\frac{1}{2} - 1\right)} = 12\left(x - \frac{2}{3}\right)(x - 1);$$

$$\varphi_{12}(x) = \frac{x\left(x - \frac{1}{2}\right)}{\frac{1}{3}\left(\frac{1}{3} - \frac{1}{2}\right)} = -18x\left(x - \frac{1}{2}\right);$$

$$\varphi_{22}(x) = \frac{\left(x - \frac{1}{2}\right)(x - 1)}{\left(\frac{2}{3} - \frac{1}{2}\right)\left(\frac{2}{3} - 1\right)} = -18\left(x - \frac{1}{2}\right)(x - 1);$$

$$\varphi_{13}(x) = \frac{x\left(x - \frac{1}{3}\right)}{\frac{1}{2}\left(\frac{1}{2} - \frac{1}{3}\right)} = 12x\left(x - \frac{1}{3}\right);$$

$$\varphi_{23}(x) = \frac{\left(x - \frac{1}{2}\right)\left(x - \frac{2}{3}\right)}{\left(1 - \frac{1}{2}\right)\left(1 - \frac{2}{3}\right)} = 6\left(x - \frac{1}{2}\right)\left(x - \frac{2}{3}\right).$$

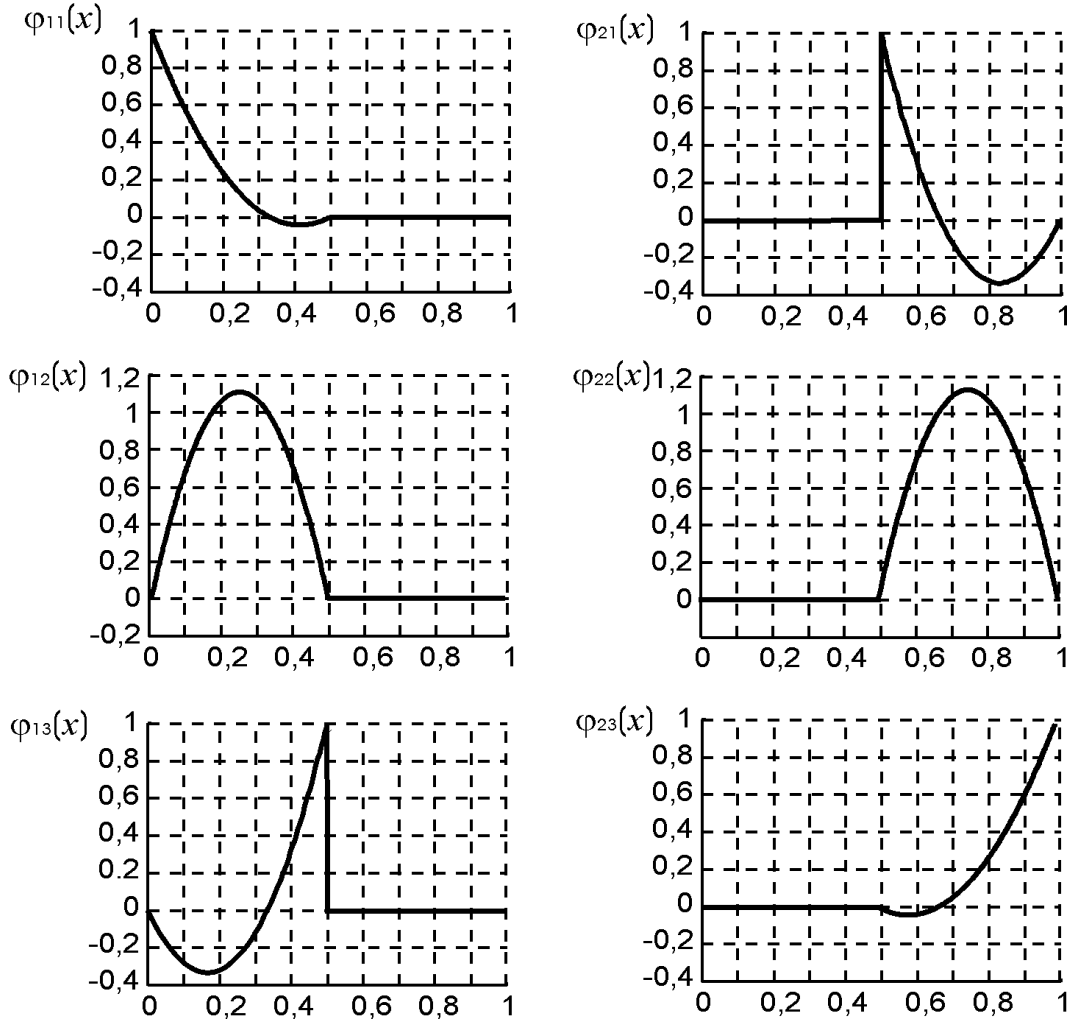


Рис. 6.1. Базис фінітних функцій

$$y(x) = y_{11}\varphi_{11}(x) + y_{12}\varphi_{12}(x) + \frac{1}{2}y_{13}(\varphi_{13}(x) + \varphi_{21}(x)) + y_{22}\varphi_{22}(x) + y_{23}\varphi_{23}(x).$$

Побудуємо проєкційну схему, причому за проєкційний 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  $D_2$ :

$$\begin{aligned}
& y_{11} \underbrace{\int_0^{1/2} L\varphi_{11}(x)\varphi_{12}(x)dx}_{\frac{183}{40}} + y_{12} \underbrace{\int_0^{1/2} L\varphi_{12}(x)\varphi_{12}(x)dx}_{\frac{1053}{80}} + y_{13} \underbrace{\int_0^{1/2} L\varphi_{13}(x)\varphi_{12}(x)dx}_{\frac{717}{80}} = \\
& = \underbrace{\int_0^{1/2} f(x)\varphi_{12}(x)dx}_{\frac{3}{32}}; \\
& y_{21} \underbrace{\int_{1/2}^1 L\varphi_{21}(x)\varphi_{22}(x)dx}_{\frac{717}{80}} + y_{22} \underbrace{\int_{1/2}^1 L\varphi_{22}(x)\varphi_{22}(x)dx}_{\frac{1053}{80}} + y_{23} \underbrace{\int_{1/2}^1 L\varphi_{23}(x)\varphi_{22}(x)dx}_{\frac{183}{40}} = \\
& = \underbrace{\int_{1/2}^1 f(x)\varphi_{22}(x)dx}_{\frac{9}{32}}; \\
& y_{11} \underbrace{\int_0^{1/2} L\varphi_{11}(x)\varphi_{13}(x)dx}_{\frac{7}{240}} + y_{12} \underbrace{\int_0^{1/2} L\varphi_{12}(x)\varphi_{13}(x)dx}_{\frac{3}{80}} + \frac{1}{2}y_{13} \times \\
& \times \left[ \underbrace{\int_0^{1/2} L\varphi_{13}(x)\varphi_{12}(x)dx}_{\frac{1}{30}} + \underbrace{\int_{1/2}^1 L\varphi_{21}(x)\varphi_{21}(x)dx}_{\frac{1}{30}} \right] + \\
& + y_{22} \underbrace{\int_{1/2}^1 L\varphi_{22}(x)\varphi_{21}(x)dx}_{\frac{3}{80}} + y_{23} \underbrace{\int_{1/2}^1 L\varphi_{23}(x)\varphi_{21}(x)dx}_{\frac{7}{240}} = \underbrace{\int_0^{1/2} f(x)\varphi_{13}(x)dx}_{\frac{1}{48}} + \underbrace{\int_{1/2}^1 f(x)\varphi_{21}(x)dx}_{\frac{1}{48}}.
\end{aligned}$$

Ураховуючи, 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:

$$-\frac{1053}{80}y_{12} + \frac{717}{80}y_{13} = \frac{3}{32}, \quad \frac{717}{80}y_{13} - \frac{1053}{80}y_{22} = -\frac{199}{160},$$

$$-\frac{3}{80}y_{12} + \frac{2}{15}y_{13} - \frac{3}{80}y_{22} = \frac{7}{720}.$$

Розв'язуючи систему, одержуємо:

$$y_{12} = 0,0745, \quad 0,0741,$$

$$y_{13} = y_{21} = 0,1209, \quad 0,1202,$$

$$y_{22} = 0,1784, \quad 0,1768.$$

### 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. 1 – 5.
7. Розв'язати систему ЗДР методом Ейлера. Оцінити похибку роз'язку.

### Варіанти завдань

1.  $\frac{dy}{dt} = t^2 + 3y;$   
 $y(0) = 0, \quad t \in [0; 1].$ 

$$\begin{cases} \frac{dx}{dt} = y, \\ \frac{dy}{dt} = x, \end{cases}$$

$$x(0) = 1, \quad y(0) = 1, \quad t \in [0; 1].$$
2.  $\frac{dy}{dt} = y(1-t);$   
 $y(0) = 1, \quad t \in [0; 1].$ 

$$\begin{cases} \frac{dx}{dt} = x - y, \\ \frac{dy}{dt} = xy, \end{cases}$$

$$x(0) = 1, \quad y(0) = 1, \quad t \in [0; 1].$$
3.  $\frac{dy}{dt} = ty - t;$   
 $y(0) = 0, \quad t \in [0; 1].$ 

$$\begin{cases} \frac{dx}{dt} = \sin(t) - y, \\ \frac{dy}{dt} = \cos(t) - x, \end{cases}$$

$$x(0) = 0, \quad y(0) = 0, \quad t \in [0; 1].$$



4.  $\frac{dy}{dt} = 4t^2 + y;$   
 $y(0) = 0, t \in [0; 1].$
5.  $\frac{dy}{dt} = \cos(t + y);$   
 $y(0) = 0, t \in [0; 1].$
6.  $\frac{dy}{dt} = \sin(t - y);$   
 $y(0) = 0, t \in [0; 1].$
7.  $\frac{dy}{dt} = \sin(t + y);$   
 $y(0) = 0, t \in [0; 1].$
8.  $\frac{dy}{dt} = \cos(t) + y;$   
 $y(0) = 0, t \in [0; 1].$
9.  $\frac{dy}{dt} = \sin(t) + y,$   
 $y(0) = 0, t \in [0; 1].$
10.  $\frac{dy}{dt} = \sin(t) - y;$   
 $y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = x - y, \\ \frac{dy}{dt} = x + y, \end{cases}$   
 $x(0) = 1, y(0) = 1, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y, \\ \frac{dy}{dt} = t, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = x + y, \\ \frac{dy}{dt} = t, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y - x, \\ \frac{dy}{dt} = t, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y - x, \\ \frac{dy}{dt} = t^2, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + t, \\ \frac{dy}{dt} = x - t, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + t^2, \\ \frac{dy}{dt} = x - t^2, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$

11.  $\frac{dy}{dt} = \cos(t) - y;$   
 $y(0) = 0, t \in [0; 1].$
12.  $\frac{dy}{dt} = y\sqrt{t};$   
 $y(0) = 1, t \in [0; 1].$
13.  $\frac{dy}{dt} = t\sqrt{y};$   
 $y(0) = 1, t \in [0; 1].$
14.  $\frac{dy}{dt} = ty;$   
 $y(0) = 1, t \in [0; 1].$
15.  $\frac{dy}{dt} = y^2;$   
 $y(0) = 1, t \in [0; 1].$
16.  $\frac{dy}{dt} = ty^2;$   
 $y(0) = 1, t \in [0; 1].$
17.  $\frac{dy}{dt} = t^2 y;$   
 $y(0) = 1, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = t - y, \\ \frac{dy}{dt} = x - t, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + \sin(t), \\ \frac{dy}{dt} = x + t, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + \cos(t), \\ \frac{dy}{dt} = x + t, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y - x^2, \\ \frac{dy}{dt} = x + t, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y - \cos(t), \\ \frac{dy}{dt} = x + t, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y - \cos(t), \\ \frac{dy}{dt} = t - x, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y - \sin(t), \\ \frac{dy}{dt} = t - x, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$

18.  $\frac{dy}{dt} = e^t + y;$   
 $y(0) = 0, t \in [0; 1].$
19.  $\frac{dy}{dt} = e^t - y;$   
 $y(0) = 0, t \in [0; 1].$
20.  $y' = -\sqrt{t}y^2;$   
 $y(0) = 1, t \in [0; 1].$
21.  $\frac{dy}{dt} = t^2\sqrt{y};$   
 $y(0) = 1, t \in [0; 1].$
22.  $\frac{dy}{dt} = \sqrt{ty};$   
 $y(0) = 1, t \in [0; 1].$
23.  $\frac{dy}{dt} = t^2y^2;$   
 $y(0) = 1, t \in [0; 1].$
24.  $\frac{dy}{dt} = y\sin(t);$   
 $y(0) = 1, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y - e^t, \\ \frac{dy}{dt} = t - x, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + e^t, \\ \frac{dy}{dt} = t - x, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + e^t, \\ \frac{dy}{dt} = t + x, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y - e^t, \\ \frac{dy}{dt} = x + t, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + x - t, \\ \frac{dy}{dt} = y + x + t, \end{cases}$   
 $x(0) = 1, y(0) = 1, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y - x - t, \\ \frac{dy}{dt} = y - x + t, \end{cases}$   
 $x(0) = 1, y(0) = 1, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = 2y - t, \\ \frac{dy}{dt} = t - x, \end{cases}$   
 $x(0) = 1, y(0) = 1, t \in [0; 1].$

25.  $\frac{dy}{dt} = y \cos(t);$   
 $y(0) = 1, t \in [0; 1].$
26.  $\frac{dy}{dt} = y^2 \cos(t);$   
 $y(0) = 1, t \in [0; 1].$
27.  $\frac{dy}{dt} = y^2 \sin(t);$   
 $y(0) = 1, t \in [0; 1].$
28.  $\frac{dy}{dt} = y \sin^2(t);$   
 $y(0) = 1, t \in [0; 1].$
29.  $\frac{dy}{dt} = y \cos^2(t);$   
 $y(0) = 1, t \in [0; 1].$
30.  $\frac{dy}{dt} = -yt^3;$   
 $y(0) = 1, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + \cos^2(t), \\ \frac{dy}{dt} = x, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + \sin^2(t), \\ \frac{dy}{dt} = x, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + e^{2t}, \\ \frac{dy}{dt} = x, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + e^{(t/2)}, \\ \frac{dy}{dt} = x, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + \sin(2t), \\ \frac{dy}{dt} = x, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$
- $\begin{cases} \frac{dx}{dt} = y + \cos(2t), \\ \frac{dy}{dt} = x, \end{cases}$   
 $x(0) = 0, y(0) = 0, t \in [0; 1].$

## 7. Числове розв'язання задач оптимізації

Часто наукові та інженерно-технічні обчислення полягають у визначенні максимуму чи мінімуму (і відповідних аргументів) дійсної функції  $f(x_1, x_2, \dots, x_n)$  від  $n$  дійсних змінних у області  $D$   $n$ -мірного простору. Слово «оптимізація» означає або мінімізацію, або максимізацію функції. Іноді  $D$  збігається з усім  $n$ -мірним 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  $\mathbf{X}$   $n$ -мірного простору з координатами  $(x_1, x_2, \dots, x_n)$  належить  $D$  тоді й тільки тоді, коли  $X$  задовольняє нерівності

$$g_i \geq 0, \quad i = \overline{1, n},$$

де  $g_i$  – задані функції від  $\mathbf{X}$ .

Функцію  $f(x_1, x_2, \dots, x_n)$ , що підлягає мінімізації, називають *цільовою функцією*. Методи обчислення мінімумів тривіально переносяться на задачу максимізації (оскільки мінімуми функції  $f$  є максимумами для мінус  $f$ ).

Якщо функція  $f$  і всі обмеження  $g_i$  є лінійними функціями, то кажуть про задачу лінійного програмування. У цьому випадку розв'язання лежить у вершині опуклого багатогранника, описаного обмеженнями в  $n$ -мірному просторі. Звичайний метод розв'язання полягає в пошуку потрібної вершини переміщенням від наступної вершини до суміжної. Якщо функція  $f$  або яке-небудь з обмежень нелінійні, то кажуть про задачу нелінійного програмування.

Нехай функція  $f(x_1, x_2, \dots, x_n)$  неперервно диференційована в області  $D$ . Нехай відомо, що в  $D$  існує точка  $\mathbf{X} = (x_1, x_2, \dots, x_n)$ , тоді в ній виконуються співвідношення:

$$\begin{aligned} \frac{\partial}{\partial x_1} f(x_1, x_2, \dots, x_n) &= 0; \\ \frac{\partial}{\partial x_2} f(x_1, x_2, \dots, x_n) &= 0; \\ &\dots \\ \frac{\partial}{\partial x_n} f(x_1, x_2, \dots, x_n) &= 0. \end{aligned} \tag{7.1}$$

Розв'язання системи (7.1) дозволяє визначити невідомі експериментальні точки. Потім, використовуючи достатні умови мінімуму, варто виділити ті точки області  $D$ , у яких досягається локальний мінімум. Порівнявши значення цільової функції в точках локального мінімуму, треба вибрати точку з мінімальним значенням локального мінімуму, тобто визначити глоба-

льний мінімум  $f(x_1, x_2, \dots, x_n)$  у області  $D$ . Однак надійні й одночасно економічні методи пошуку глобального мінімуму в цей час невідомі.

Необхідна умова екстремуму (7.1) – умова стаціонарності можна записати у вигляді

$$\nabla f(\mathbf{X}) = 0,$$

де  $\nabla f(\mathbf{X}) = \left[ \frac{\partial f(\mathbf{X})}{\partial x_1}, \frac{\partial f(\mathbf{X})}{\partial x_2}, \dots, \frac{\partial f(\mathbf{X})}{\partial x_n} \right]^T$  – градієнт цільової функції в точці

$$\mathbf{X} = [x_1, x_2, \dots, x_n]^T.$$

Слід відзначити, що на відміну від нелінійних рівнянь  $f(x) = 0$ , які можна розв'язати з точністю  $\varepsilon$ , задачу  $\min \{f(x)\}$  – тільки з точністю  $\sqrt{\varepsilon}$ .

Найчисленнішу групу методів розв'язання задач оптимізації складають локальні методи безумовної оптимізації [6]. Деяке уявлення про методи, які широко застосовуються цією групою, ілюструє рис. 7.1.

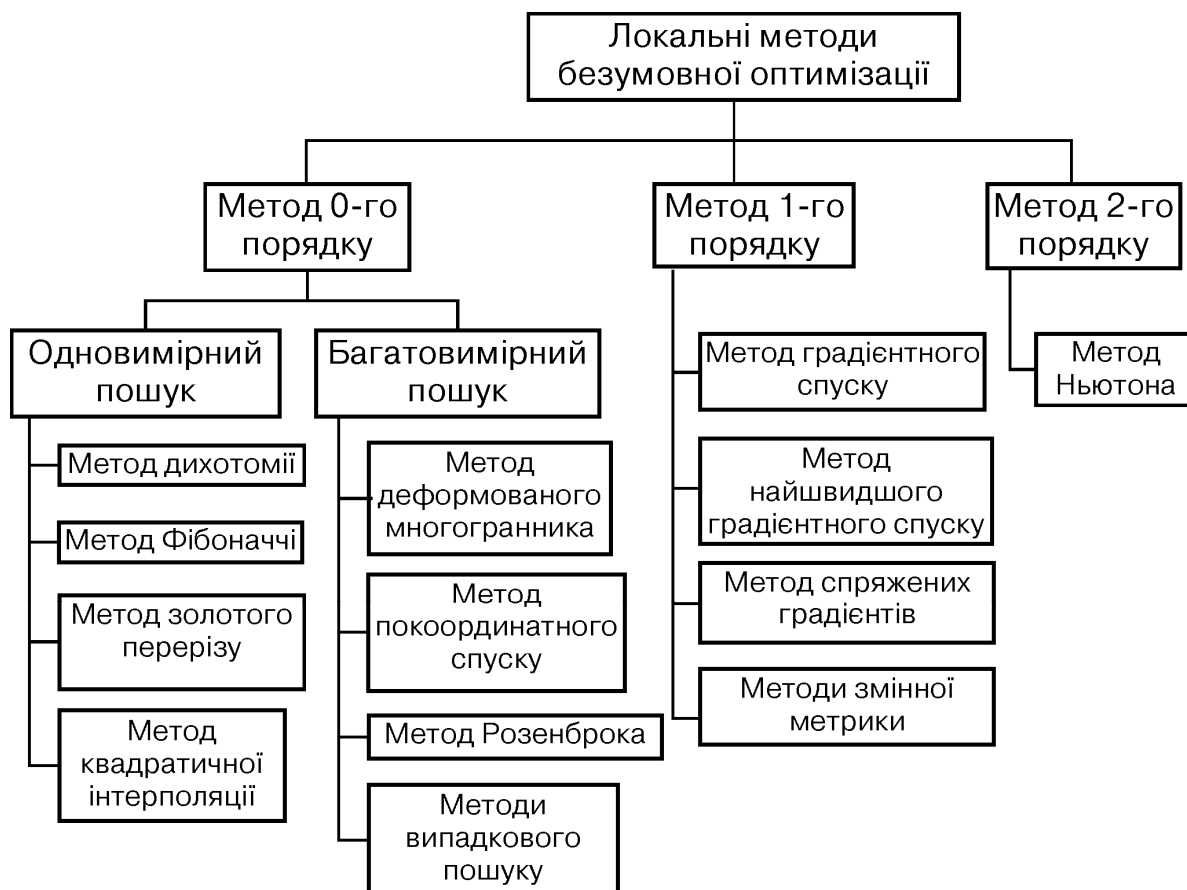


Рис. 7.1. Класифікація методів локальної безумовної оптимізації

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  $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \dots$ , за якої  $f(\mathbf{X}_0) > f(\mathbf{X}_1) > f(\mathbf{X}_2) > \dots$ . У цьому випадку метод

забезпечує збіжність і можна сподіватися, що мінімум функції буде знайдений.

Схему алгоритму пошуку для загального випадку показано на рис. 7.2. Сутність методу оптимізації визначають 2 і 3 етапами алгоритму, на яких вибирається напрям подальшого пошуку й обчислюються координати наступної точки  $X_{i+1}$  на траєкторії пошуку.

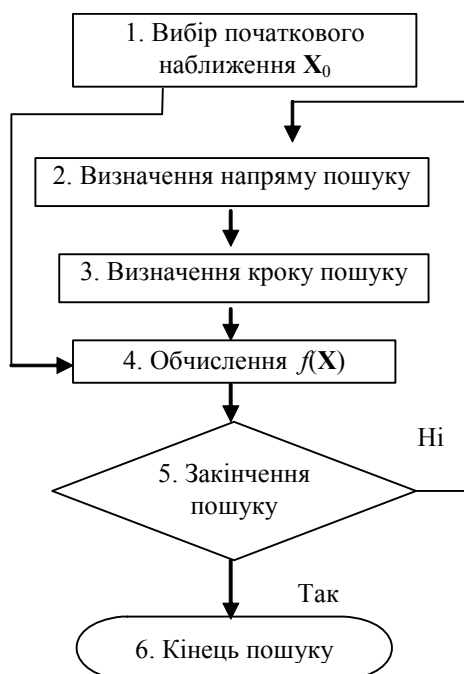


Рис. 7.2. Узагальнена схема оптимізації

## 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  $x < x^*$  і строго зростає для  $x > x^*$ . Така функція називається унімодальною і для її графіка існують три можливі форми, показані на рис. 7.3.

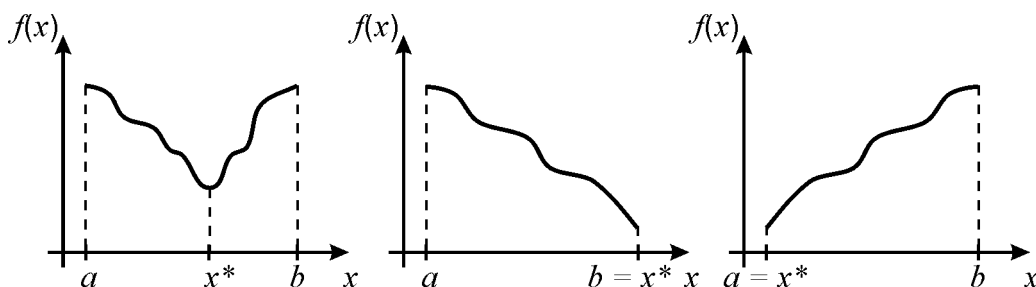


Рис. 7.3. Унімодальні функції

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) [4].

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 інтервал, у якому знаходиться точка  $x^*$ , яка відповідає мінімуму  $f(x)$ , причому цей інтервал має довжину меншу, ніж вихідний. Нехай  $x_1 < x_2$ . Можливі варіанти, показано на рис. 7.4.

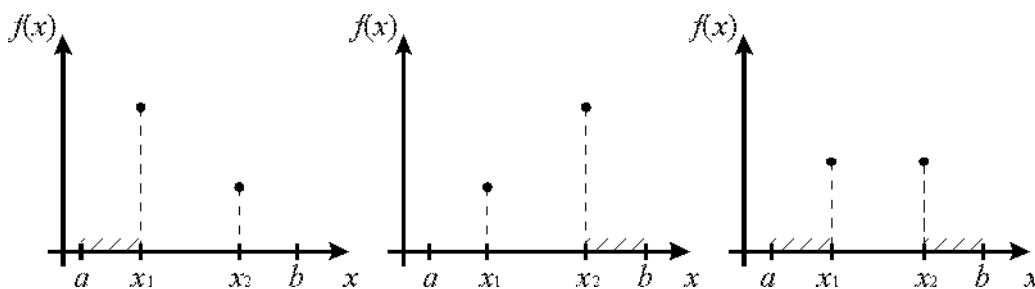


Рис. 7.4. Послідовний пошук мінімуму

У першому випадку слід відкинути інтервал  $[a, x_1]$ , у другому випадку –  $[x_2, b]$ , у третьому –  $[a, x_1], [x_2, b]$ , оскільки в цих інтервалах не може знаходитися  $x^*$ , за інших випадків порушується припущення про унімодальність  $f(x)$ . Завдання полягає в тому, щоб знайти таку множину абсцис  $x_1, x_2, \dots, x_n$ , у якій мінімум функції  $f(x)$  лежить у деякому інтервалі  $x^* \in [x_{i-1}, x_i]$ . Такий інтервал 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 інтервалу:

$$x_1 = \frac{a+b}{2} - \frac{\delta}{2}, \quad x_2 = \frac{a+b}{2} + \frac{\delta}{2}.$$

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 інтервалу  $[a, b]$  точкою  $x_1$  на дві частини таким чином, щоб відношення довжини всього інтервалу 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  $x_2$  – інтервалу  $[x_1, b]$ . Тому на інтервалі, що залишився, потрібно визначити тільки одну точку, що утворює золотий переріз. На кожному кроці довжина першого інтервалу невизначеності дорівнює 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 – квадратну 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  $f(x_{i+1}) \leq f(x_i)$ , 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 точ-

ки  $x$ , 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

$$x^* = x_c + \frac{h(f(x_a) - f(x_b))}{2(f(x_a) - 2f(x_c) + f(x_b))}.$$

## 7.2. Методи багатовимірної пошуку

У методі покоординатного спуску в області  $D$  (рис. 7.5) задають початкове наближення  $\mathbf{X}_0 = [x_1^0, x_2^0, \dots, x_n^0]^T$  і послідовно проводять одновимірні пошуки вздовж різних напрямів, паралельних координатним осям.

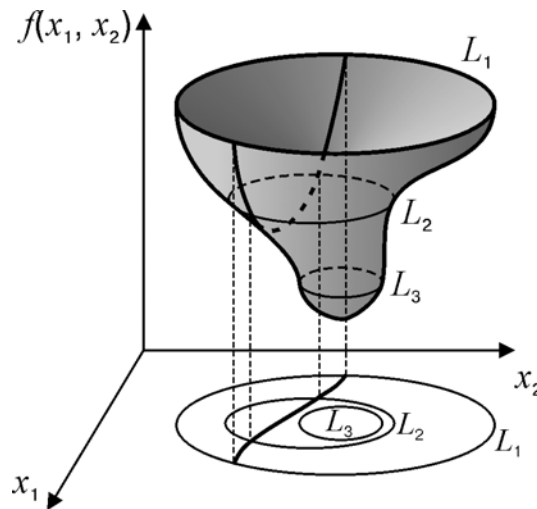


Рис. 7.5. Пошук мінімуму покоординатним спуском

Для цього розглядають функцію однієї змінної  $f(x_1^0, x_2^0, \dots, x_n^0)$  і методом одновимірної пошуку знаходять  $\min_{x_1 \in D} f(x_1, x_2^0, \dots, x_n^0)$ . Значення  $x_1$ , що відповідає мінімуму, позначають через  $x_1^1$ :

$$f(x_1^1, x_2^0, \dots, x_n^0) \leq f(x_1^0, x_2^0, \dots, x_n^0).$$

Далі розглядають функцію  $f(x_1^1, x_2^0, \dots, x_n^0)$  і знаходять  $\min_{x_2 \in D} f(x_1^1, x_2, \dots, x_n^0)$ . Значення  $x_2$ , що відповідає мінімуму функції, позначають через  $x_2^1$ . При цьому

$$f(x_1^1, x_2^1, \dots, x_n^0) \leq f(x_1^1, x_2^0, \dots, x_n^0).$$

Після  $n$  кроків одержують

$$f(x_1^1, x_2^1, \dots, x_n^1) \leq f(x_1^1, x_2^0, \dots, x_n^0).$$

У результаті одного кроку покоординатного спуску відбувається перехід із точки  $\mathbf{X}_0$  у точку  $\mathbf{X}_1 = [x_1^1, x_2^1, \dots, x_n^1]^T$ .

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:

$$\left| \frac{f(\mathbf{X}_{i+1}) - f(\mathbf{X}_i)}{f(\mathbf{X}_{i+1})} \right| < \varepsilon,$$

де  $\varepsilon$  – задана точність.

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. При цьому 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 напрям.

Точка  $\mathbf{X}_{i+1}$  вважається обчисленою, якщо  $f(\mathbf{X}_{i+1}) < f(\mathbf{X}_i)$ . Інакше роблять спробу досягти успіху або за рахунок зміни кроку, або за рахунок зміни напрямку на протилежний. Якщо всі спроби виявляються невдалими, то вибирають новий випадковий напрям пошуку. Випадковий напрям пошуку знаходять за допомогою випадкової точки  $\mathbf{A} = [a_1, a_2, \dots, a_n]$ , компоненти якої є випадковим значенням, рівномірно розподіленим на відрізку  $[-1, 1]$ .

### 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.

Усі градієнтні методи використовують указані 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 напрямку пошуку на  $i$ -му кроці.

Спосіб вибирання кроку, напряму чи того й другого одночасно визначає сутність методу.

У разі пошуку мінімуму цільової функції потрібно рухатися в напрямі, протилежному градієнту функції точки  $f(\mathbf{X})$ . Тому напрям

$$\mathbf{S}_i = -\frac{\nabla f(\mathbf{X}_i)}{\|\nabla f(\mathbf{X}_i)\|_2}$$

дозволяє одержати таку рекурентну формулу для методу градієнтного спуску:

$$\mathbf{X}_{i+1} = \mathbf{X}_i - h \frac{\nabla f(\mathbf{X}_i)}{\|\nabla f(\mathbf{X}_i)\|_2}.$$

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:

$$f(\mathbf{X}_{i+1}) = f\left(\mathbf{X}_i - h \frac{\nabla f(\mathbf{X}_i)}{\|\nabla f(\mathbf{X}_i)\|_2}\right) = g(h).$$

Тому крок  $h$  можна вибрати так, щоб  $f(\mathbf{X}_{i+1})$  максимально зменшила своє значення:

$$f(\mathbf{X}_{i+1}) = \min_h f\left(\mathbf{X}_i - h \frac{\nabla f(\mathbf{X}_i)}{\|\nabla f(\mathbf{X}_i)\|_2}\right).$$

Вибір оптимального кроку зводиться до розв'язання одновимірної оптимізації функції  $g(h)$ .

*Алгоритм методу найшвидшого спуску:*

1. Обчислити всі частинні похідні цільової функції.
2. Знайти один з методів одновимірного пошуку оптимального кроку уздовж антиградієнтного напряму. Крок  $h$  визначають з умови мінімуму функції  $f\left(\mathbf{X}_i - h \frac{\nabla f(\mathbf{X}_i)}{\|\nabla f(\mathbf{X}_i)\|_2}\right)$  по  $h$ .

3. Обчислити координати нової точки  $\mathbf{X}_{i+1}$ .

4. Якщо умова припинення пошуку не виконується, то перейти до кроку 1.

Траєкторію пошуку цим методом показано на рис. 7.6, з якого видно, що рух уздовж одного напряму припиняється, коли лінія напряму пошуку стає

дотичною до якої-небудь лінії рівня. Кожний новий напрям руху до екстремуму ортогональний попередньому, якщо цільова функція квадратична.

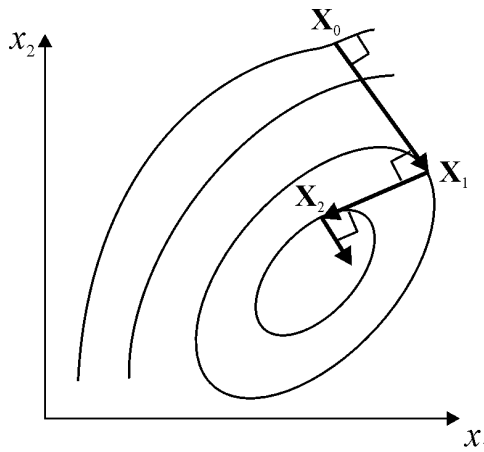


Рис. 7.6. Метод найвищого спуску

Розглянуті методи збігаються до локального мінімуму зі швидкістю геометричної прогресії, тобто лінійно.

#### 7.4. Метод спряжених градієнтів

Вектори  $\mathbf{A}$  і  $\mathbf{B}$  називають спряженими відносно матриці  $\mathbf{Q}$  (чи  $\mathbf{Q}$ -спряженими), якщо скалярний добуток векторів  $\mathbf{A}$  і  $\mathbf{QB}$  дорівнює нулю, тобто  $\mathbf{A}^T \mathbf{QB} = 0$ . Спряженість векторів є узагальненням поняття ортогональності, оскільки, якщо  $\mathbf{Q} = \mathbf{E}$ , то  $\mathbf{Q}$ -спряженість векторів означає їх ортогональність.

Нехай пошук починається в точці  $\mathbf{X}_0$  з початковим напрямом  $\mathbf{S}_0$ , вибраним довільно. Тоді нова точка

$$\mathbf{X}_1 = \mathbf{X}_0 + h\mathbf{S}_0, \quad (7.2)$$

де значення кроку  $h$  слід визначати за умови мінімуму функції  $f(\mathbf{X}_0 + h\mathbf{S}_0)$  по  $h$ . Для цього розкладемо функцію  $f(\mathbf{X})$  у ряд Тейлора в околі точки  $\mathbf{X} = \mathbf{X}_0$  і обмежимося розглядом тільки трьох членів, що є точні, для квадратичної цільової функції

$$f(\mathbf{X}) \approx f(\mathbf{X}_0) + \nabla f(\mathbf{X}_0)^T (\mathbf{X} - \mathbf{X}_0) + \frac{1}{2} (\mathbf{X} - \mathbf{X}_0)^T \mathbf{H}(\mathbf{X}_0) (\mathbf{X} - \mathbf{X}_0),$$

$$\text{де } \mathbf{H}(\mathbf{X}_0) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{X}_0)}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{X}_0)}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f(\mathbf{X}_0)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{X}_0)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\mathbf{X}_0)}{\partial x_2^2} & \dots & \frac{\partial^2 f(\mathbf{X}_0)}{\partial x_2 \partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial^2 f(\mathbf{X}_0)}{\partial x_1 \partial x_n} & \frac{\partial^2 f(\mathbf{X}_0)}{\partial x_2 \partial x_n} & \dots & \frac{\partial^2 f(\mathbf{X}_0)}{\partial x_n^2} \end{bmatrix} \text{ – матриця Гессе.}$$

Тоді

$$f(\mathbf{X}_0 + h\mathbf{S}_0) = f(\mathbf{X}_0) + \nabla f(\mathbf{X}_0)^T h\mathbf{S}_0 + \frac{1}{2} h\mathbf{S}_0^T \mathbf{H}(\mathbf{X}_0) h\mathbf{S}_0.$$

З умови мінімуму останньої функції по  $h$  маємо

$$\frac{df(\mathbf{X}_0 + h\mathbf{S}_0)}{dh} = \nabla f(\mathbf{X}_0)^T \mathbf{S}_0 + h\mathbf{S}_0^T \mathbf{H}(\mathbf{X}_0) \mathbf{S}_0 = 0.$$

Звідси

$$h_0 = -\frac{\nabla f(\mathbf{X}_0)^T \mathbf{S}_0}{\mathbf{S}_0^T \mathbf{H}(\mathbf{X}_0) \mathbf{S}_0}. \quad (7.3)$$

Можна показати, що після того як за (7.2) і (7.3) обчислено точку  $\mathbf{X}_1$ , для продовження пошуку потрібно обрати новий напрям  $\mathbf{S}_1$  так, щоб

$$\mathbf{S}_1^T \mathbf{H}(\mathbf{X}_0) \mathbf{S}_0 = 0,$$

тобто новий напрям  $\mathbf{S}_1$  має бути спряженим до старого напрямку  $\mathbf{s}_0$ . Збіжність такого методу строго обґрунтована для квадратичних функцій вигляду

$$f(\mathbf{X}) = \mathbf{c}^T \mathbf{X} + \frac{1}{2} \mathbf{X}^T \mathbf{H} \mathbf{X},$$

де  $\mathbf{c}$  – вектор-стовпець, а  $\mathbf{H}$  – матриця постійних коефіцієнтів.

Мінімізацію такої функції за додатно визначеною матрицею  $\mathbf{H}$  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 Флетчером і Ривсом варіанті, названо методом спряженого градієнта.

У методі спряжених градієнтів будується послідовність напрямів пошуку  $\mathbf{S}_i$ , що є лінійними комбінаціями антиградієнта  $-\nabla f(\mathbf{X}_i)$  і попередніх напрямів пошуку  $(\mathbf{S}_0, \dots, \mathbf{S}_{i-1})$ . Так, якщо  $\mathbf{S}_0 = -\nabla f(\mathbf{X}_0)$ , то  $\mathbf{X}_1 = h_0 \mathbf{S}_0$ , потрібно знайти новий  $\mathbf{S}_1 = -\nabla f(\mathbf{X}_1) + \beta_1 \mathbf{S}_0$  напрям, підбравши коефіцієнт  $\beta_1$  так, щоб він задовольняв умови спряженості векторів  $\mathbf{S}_0$  і  $\mathbf{S}_1$ , якщо

$$\beta_1 = \frac{\nabla f(\mathbf{X}_1)^T \nabla f(\mathbf{X}_1)}{\nabla f(\mathbf{X}_0)^T \nabla f(\mathbf{X}_0)}.$$

Алгоритм загального випадку методу спряжених градієнтів буде таким.

1. Обчислити вектор  $\mathbf{S}_0 = -\nabla f(\mathbf{X}_0)$ .
2. Знайти мінімум  $f(\mathbf{X})$  одним з методів одновимірного пошуку в напрямі  $\mathbf{S}_i$ . Звідси знаходимо  $\mathbf{X}_{i+1}$ ,  $f(\mathbf{X}_{i+1})$ ,  $\nabla f(\mathbf{X}_{i+1})$ .
3. Визначити новий спряжений напрям  $\mathbf{S}_{i+1}$  із співвідношення

$$\mathbf{S}_{i+1} = -\nabla f(\mathbf{X}_{i+1}) + \mathbf{S}_i \frac{\nabla f(\mathbf{X}_{i+1})^T \nabla f(\mathbf{X}_{i+1})}{\nabla f(\mathbf{X}_i)^T \nabla f(\mathbf{X}_i)}.$$

Після обчислення  $(n + 1)$ -го кроку (при  $i = n$ ) обчислення циклічно повторюють.

4. Якщо  $\|\mathbf{S}_i\| < \varepsilon$ , пошук припинити.

Методи спряжених градієнтів мають  $n$ -крокову квадратичну швидкість збіжності, тобто

$$\|\mathbf{X}_{i+n} - \mathbf{X}^*\| \leq c \|\mathbf{X}_i - \mathbf{X}^*\|^2.$$

## 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

$$f(\mathbf{X}) = f(\mathbf{X}_i) + \nabla f(\mathbf{X}_i)^T (\mathbf{X} - \mathbf{X}_i) + \frac{1}{2} (\mathbf{X} - \mathbf{X}_i)^T \mathbf{H}(\mathbf{X}_i) (\mathbf{X} - \mathbf{X}_i), \quad (7.4)$$

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 (7.4):

$$\nabla f(\mathbf{X}_{i+1}) = \nabla f(\mathbf{X}_i) + \mathbf{H}(\mathbf{X}_i) (\mathbf{X}_{i+1} - \mathbf{X}_i), \quad (7.5)$$

звідки

$$\mathbf{X}_{i+1} = \mathbf{X}_i - \mathbf{H}^{-1}(\mathbf{X}_i) \nabla f(\mathbf{X}_i). \quad (7.6)$$

Формулу (7.5) одержуємо 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 (7.6) не вдаються до безпосереднього обергання матриці Гессе, а на кожному ітераційному кроці задачу розв'язують у два етапи. Спочатку розв'язують СЛАР відносно  $\Delta \mathbf{X}_i$ :

$$\mathbf{H}(\mathbf{X}_i) \Delta \mathbf{X}_i = -\nabla f(\mathbf{X}_i), \quad (7.7)$$

потім знаходять нове наближення

$$\mathbf{X}_{i+1} = \mathbf{X}_i + \Delta \mathbf{X}_i. \quad (7.8)$$

Якщо початкове наближення  $\mathbf{X}_0$  задано досить близько до точки  $\mathbf{X}^*$ , то на основі виразів (7.7) і (7.8) можна побудувати квадратичну швидкість збіжності, тобто

$$\|\mathbf{X}_{i+1} - \mathbf{X}^*\| \leq c \|\mathbf{X}_i - \mathbf{X}^*\|^2,$$

де  $c$  – постійна величина.

Розв'язуючи СЛАР (7.7), слід мати на увазі дві важливі властивості матриці Гессе. По-перше, матриця Гессе є симетричною. По-друге, з (7.4) випливає, що якщо  $\mathbf{X}_i$  є точкою мінімуму, то  $\nabla f(\mathbf{X}_i) = 0$  й отже, для того, щоб функція  $f(\mathbf{X})$  зростала у разі відхилення від точки  $\mathbf{X}_i$ , необхідно, щоб  $(\mathbf{X} - \mathbf{X}_i)^T \mathbf{H}(\mathbf{X}_i)(\mathbf{X} - \mathbf{X}_i) > 0$ . Отже, поблизу мінімуму матриця Гессе додатно визначена. Тому для розв'язання СЛАР (7.7) застосовують метод декомпозиції 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 в бік максимуму чи сідлової точки функції  $f(\mathbf{X})$ , де  $\nabla f(\mathbf{X})$  теж дорівнює нулю. Останнє можливо, якщо матриця Гессе перестане бути позитивно визначеною. Тому у використовуваних на практиці методах Ньютона, якщо матриця Гессе не є позитивно визначеною (з'являються негативні власні значення), її замінюють матрицею

$$\tilde{\mathbf{H}} = \mathbf{H}(\mathbf{X}_i) + \mu \mathbf{E},$$

у якій  $\mu$  вибирають так, щоб матриця  $\tilde{\mathbf{H}}$  була позитивно визначеною і добре обумовленою. При цьому найменше з можливих  $\mu$  є дещо більшим, ніж модуль найменшого від'ємного власного значення 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

$$\mathbf{X}_{i+1} = \mathbf{X}_i - \alpha \tilde{\mathbf{H}}^{-1} \nabla f(\mathbf{X}_i),$$

де  $0 < \alpha \leq 1$ . Величину  $\alpha$  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 скінченорізницеву апроксимацію.

## 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 замінюється матрицею, яка обчислюється простіше. Один із способів побудови методів змінної метрики полягає у застосуванні методу для січних розв'язання нелінійних рівнянь до системи

$$\nabla f(\mathbf{X}) = 0.$$

При цьому 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 напрямку  $\Delta\mathbf{X}_i$  пошуку на  $i$ -му кроці та наступну точку траєкторії можна обчислити так:

$$\tilde{\mathbf{H}}_i \Delta\mathbf{X}_i = -\nabla f(\mathbf{X}_i);$$

$$\mathbf{X}_{i+1} = \mathbf{X}_i + \alpha \Delta\mathbf{X}_i,$$

де  $\alpha$  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:

$$\tilde{\mathbf{H}}_{i+1} = \tilde{\mathbf{H}}_i + \frac{\alpha \Delta\mathbf{X}_i \Delta\mathbf{X}_i^T}{\Delta\mathbf{X}_i^T (\nabla f(\mathbf{X}_{i+1}) - \nabla f(\mathbf{X}_i))} - \frac{\tilde{\mathbf{H}}_i (\nabla f(\mathbf{X}_{i+1}) - \nabla f(\mathbf{X}_i)) (\nabla f(\mathbf{X}_{i+1}) - \nabla f(\mathbf{X}_i))^T \tilde{\mathbf{H}}_i}{(\nabla f(\mathbf{X}_{i+1}) - \nabla f(\mathbf{X}_i))^T \tilde{\mathbf{H}}_i (\nabla f(\mathbf{X}_{i+1}) - \nabla f(\mathbf{X}_i))},$$

$$\tilde{\mathbf{H}}_{i+1} = \tilde{\mathbf{H}}_i + \frac{\alpha \Delta\mathbf{X}_i \Delta\mathbf{X}_i^T}{\Delta\mathbf{X}_i^T \mathbf{P}_i} - \frac{\tilde{\mathbf{H}}_i \mathbf{P}_i \mathbf{P}_i^T \tilde{\mathbf{H}}_i}{\mathbf{P}_i^T \tilde{\mathbf{H}}_i \mathbf{P}_i},$$

де  $\mathbf{P}_i = \nabla f(\mathbf{X}_{i+1}) - \nabla f(\mathbf{X}_i)$ .

## 7.7. Методи розв'язання задач умовної оптимізації

Під задачею умовної оптимізації розуміють задачу пошуку мінімуму функції  $f(\mathbf{X})$  у разі обмежень:

$V(\mathbf{X}) = 0$  ( $m$  обмежень-рівностей);  
 $U(\mathbf{X}) \geq 0$  ( $k$  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 умов 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:

$$y_i = \operatorname{tg} \left( \pi \frac{x_i - \frac{b+a}{2}}{b-a} \right).$$

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:

$$\psi(\mathbf{X}, \Lambda) = f(\mathbf{X}) + \Lambda^T \mathbf{V}(\mathbf{X}) = f(\mathbf{X}) + \sum_{j=1}^m \lambda_j v_j(\mathbf{X}),$$

де  $\Lambda = [\lambda_1, \lambda_2, \dots, \lambda_m]^T$  – вектор невизначених множників Лагранжа,  $v_j(\mathbf{X})$  –  $j$ -і обмеження типу рівності;  $m$  – кількість обмежень.

Щоб знайти значення  $n$  невідомих  $x_1, x_2, \dots, x_n$  і  $m$  множників Лагранжа  $\lambda_1, \lambda_2, \dots, \lambda_m$ , розв'язують систему рівнянь, що виражає необхідні умови екстремуму функції Лагранжа:

$$\frac{\partial \psi(\mathbf{X}, \Lambda)}{\partial x_i} = \frac{\partial f(\mathbf{X})}{\partial x_i} + \sum_{j=1}^m \lambda_j \frac{\partial v_j(\mathbf{X})}{\partial x_i} = 0, \quad i = \overline{1, n};$$

$$\frac{\partial \psi(\mathbf{X}, \Lambda)}{\partial \lambda_j} = v_j(\mathbf{X}) = 0, \quad j = \overline{1, m}.$$

Функція Лагранжа  $\psi(\mathbf{X}, \Lambda)$  і цільова функція  $f(\mathbf{X})$  у допустимій області збігаються, оскільки тут  $V(\mathbf{X}) = 0$ . Тому, якщо оптимальне значення функції Лагранжа знайдено, то воно є одночасним і умовним оптимумом цільової функції  $f(\mathbf{X})$ .

**Приклад:**

$$f(\mathbf{X}) = x_1^2 + x_2^2;$$

$$v(\mathbf{X}) = 2x_1 + x_2 - 2 = 0.$$

Знаходимо функцію Лагранжа

$$\psi(\mathbf{X}, \Lambda) = x_1^2 + x_2^2 + \lambda(2x_1 + x_2 - 2);$$

$$\left. \begin{aligned} \frac{\partial \psi}{\partial x_1} = 2x_1 + 2\lambda = 0 \\ \frac{\partial \psi}{\partial x_2} = 2x_2 + \lambda = 0 \\ \frac{\partial \psi}{\partial \lambda} = 2x_1 + x_2 - 2 = 0 \end{aligned} \right\} \begin{aligned} x_1^* &= \frac{4}{5}, \\ x_2^* &= \frac{2}{5}, \\ \lambda^* &= -\frac{4}{5}; \end{aligned}$$

$$\min f(\mathbf{X}) = f(\mathbf{X}^*) = \psi(\mathbf{X}^*, \Lambda^*) = \frac{4}{5}.$$

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. 7.8–7.10). Зазвичай функцію штрафу використовують як бар’єр, тобто

$$\lim_{r \rightarrow 0} S(\mathbf{X}, r) = \begin{cases} 0, & x \in D, \\ \infty, & x \notin D. \end{cases}$$

Загальну функцію штрафу подають у вигляді суми функцій штрафів, накладених кожним з обмежень-нерівностей і обмежень-рівностей:

$$S(\mathbf{X}, r) = \sum_{i=1}^m S_i(v_i(\mathbf{X}), r) + \sum_{j=1}^k S_j(v_j(\mathbf{X}), r).$$

Найчастіше використовують такі типи функцій штрафу:

– логарифмічну (рис. 7.7):

$$S(\mathbf{X}, r) = \begin{cases} -r \ln(u(\mathbf{X})), & u(\mathbf{X}) > 0, \\ +\infty, & u(\mathbf{X}) \leq 0; \end{cases}$$

– задану оберненою функцією (рис. 7.8):

$$S(\mathbf{X}, r) = \begin{cases} \frac{r}{u(\mathbf{X})}, & u(\mathbf{X}) > 0, \\ +\infty, & u(\mathbf{X}) \leq 0; \end{cases}$$

– типу квадрата зрізки (рис. 7.9):

$$S(\mathbf{X}, r) = \frac{1}{r} \left( \frac{u(\mathbf{X}) - (u(\mathbf{X}))^2}{2} \right)^2 = \begin{cases} 0, & u(\mathbf{X}) \geq 0, \\ \frac{u(\mathbf{X})^2}{r}, & u(\mathbf{X}) < 0; \end{cases}$$

– для обмежень-рівностей використовують квадратичний штраф (рис. 7.10):

$$S(\mathbf{X}, r) = \frac{1}{r} v(\mathbf{X})^2$$

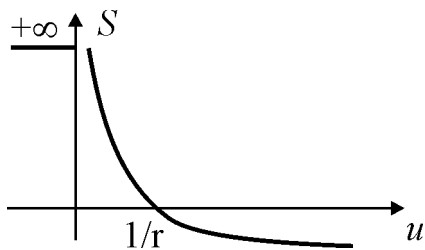


Рис. 7.7. Логарифмічний штраф

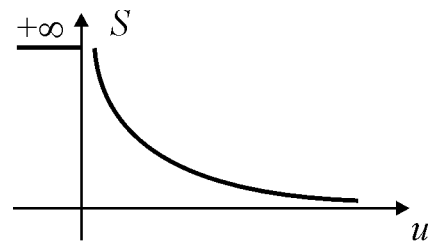


Рис. 7.8. Штраф оберненої функції

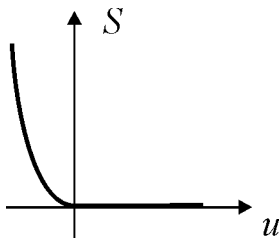


Рис. 7.9. Штраф типу квадрата зрізання

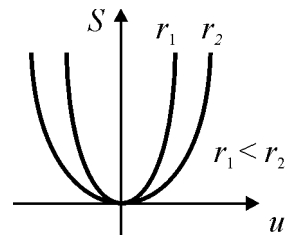


Рис. 7.10. Квадратичний штраф

Після побудови штрафної функції розв'язують задачу безумовної оптимізації при різних  $r$  таких, що

$$r_1 > r_2 > r_3 \dots$$

і знаходять послідовність точок  $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3 \dots$ . Границя цієї послідовності при  $r \rightarrow 0$  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 (7.11) і методи зовнішньої точки (рис. 7.12).

Використання методів внутрішньої точки дозволяє утримувати траєкторію пошуку всередині допустимої області. При цьому функція штрафу перешкоджає підходу точок траєкторії пошуку до границі. У методі внутрішньої точки як функції штрафу частіше використовують логарифмічну або обернену функцію.

**Приклад:**

$$f(\mathbf{X}) = x_1 + x_2,$$

$$v_1(\mathbf{X}) = -x_1^2 + x_2 \geq 0,$$

$$v_2(\mathbf{X}) = x_1 \geq 0.$$

Побудуємо штрафну функцію, використовуючи логарифмічну функцію штрафу:

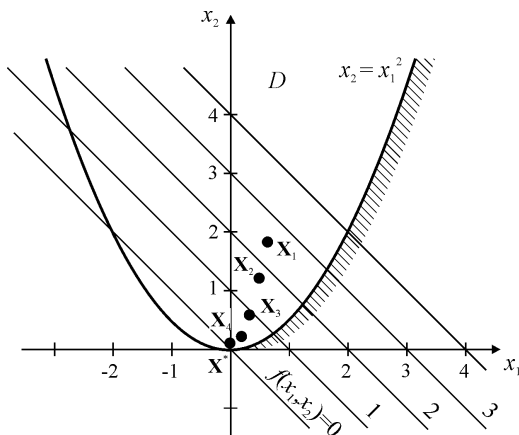
$$\psi(\mathbf{X}) = x_1 + x_2 - r \ln(-x_1^2 + x_2) - r \ln x_1,$$

$$\frac{\partial \psi}{\partial x_1} = 1 + \frac{2rx_1}{-x_1^2 + x_2} - \frac{r}{x_1} = 0,$$

$$\frac{\partial \psi}{\partial x_2} = 1 - \frac{r}{-x_1^2 + x_2} = 0,$$

$$x_1(r) = \frac{-1 + \sqrt{1 + 8r}}{4}; \quad x_2(r) = \frac{(-1 + \sqrt{1 + 8r})^2}{16} + r;$$

$$\mathbf{X}^* = \lim_{r \rightarrow 0} \mathbf{X}(r) = [0, 0]^T.$$



$i$	$r_i$	$x_1$	$x_2$
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

Рис. 7.11. Метод внутрішньої точки

У використанні методу внутрішньої точки важливим є вибір початкових значень вектора  $\mathbf{X}$  і параметра  $r$ . Необхідність задання 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. При цьому як функцію штрафу часто використовують функцію штрафу типу квадрата зрізання.

**Приклад:**

$$f(\mathbf{X}) = x_1 + x_2,$$

$$v_1(\mathbf{X}) = -x_1^2 + x_2 > 0,$$

$$v_2(\mathbf{X}) = x_1 \geq 0;$$

$$\psi(\mathbf{X}) = x_1 + x_2 + \frac{(-x_1^2 + x_2)^2}{r} + \frac{x_1^2}{r},$$

$$\frac{\partial \psi}{\partial x_1} = 1 - \frac{4x_1(-x_1^2 + x_2)}{r} + \frac{2x_1}{r} = 0,$$

$$\frac{\partial \psi}{\partial x_2} = 1 + \frac{2(-x_1^2 + x_2)}{r} = 0;$$

$$x_1(r) = -\frac{r}{2(1+r)},$$

$$x_2(r) = -\frac{r^2}{4(1+r)^2} - \frac{r}{2};$$

$$\mathbf{X}^* = \lim_{r \rightarrow 0} \mathbf{X}(r) = [0, 0]^T.$$

$i$	$r_i$	$x_1$	$x_2$
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

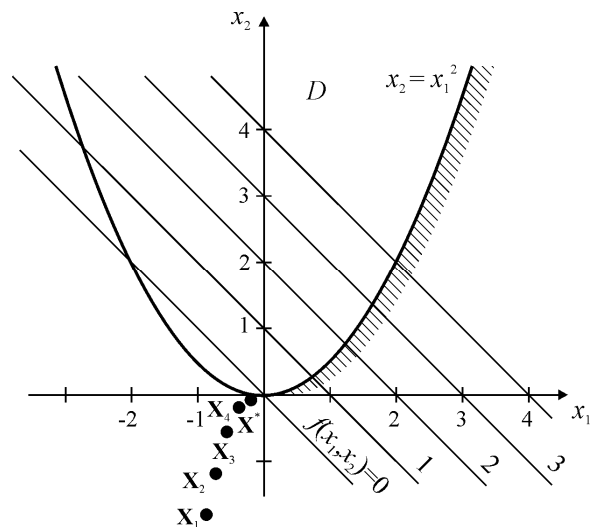


Рис. 7.12. Метод зовнішньої точки

Як приклад методів, орієнтованих на безпосереднє розв'язання задач умовної оптимізації розглянемо метод проєкції градієнта.

Цей метод здебільшого застосовують для розв'язання задач умовної оптимізації з обмеженнями типу рівностей. При цьому умовний мінімум цільової функції знаходиться на  $n$ -мірній поверхні обмежень, тому що тільки тут задовольняється умова  $\mathbf{V}(\mathbf{X}) = 0$ . Тому алгоритмом методу проєкції градієнта формують два основні етапи:

*Етап 1.* Повернення на поверхню обмежень з поточної точки пошуку  $\mathbf{X}_i$ , якщо ця точка вийшла за припустимі межі порушення обмеження  $\Delta v$ :

$$\|\mathbf{V}(\mathbf{X}_i)\| \geq \Delta v.$$

Такий рух відбувається по нормалі до поверхні обмежень (наприклад з точки  $\mathbf{X}_0$  у точку  $\mathbf{X}_1$ , рис. 7.13). Крок переміщення залежить від значення відхилення і визначають за формулою

$$\Delta \mathbf{X}_i = -\mathbf{D}_i (\mathbf{D}_i \mathbf{D}_i^T)^{-1} \mathbf{V}(\mathbf{X}_i),$$

де  $\mathbf{D}_i$  – матриця розміром  $m \times n$ , рядками якої є градієнти функцій-обмежень  $\mathbf{V}(\mathbf{X})$ , обчислені в точці  $\mathbf{X}_i$ :

$$\mathbf{D}_i = \begin{bmatrix} \frac{\partial v_1}{\partial x_1} & \frac{\partial v_1}{\partial x_2} & \dots & \frac{\partial v_1}{\partial x_n} \\ \frac{\partial v_2}{\partial x_1} & \frac{\partial v_2}{\partial x_2} & \dots & \frac{\partial v_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial v_m}{\partial x_1} & \frac{\partial v_m}{\partial x_2} & \dots & \frac{\partial v_m}{\partial x_n} \end{bmatrix}.$$

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.

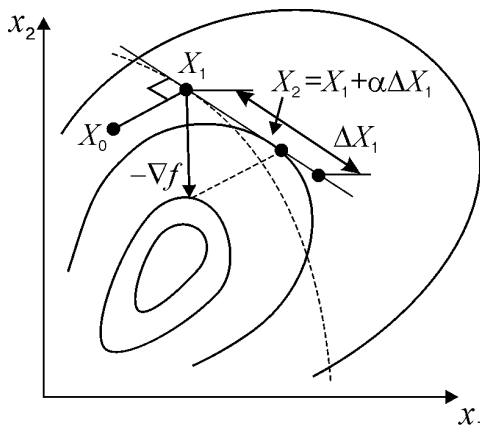


Рис. 7.13. Метод проекції градієнта

*Етап 2.* Переміщення у бік умовного екстремуму вздовж площини, дотичної в точці  $\mathbf{X}_{i+1}$  поверхні обмежень. Напрямок пошуку мінімуму визначається за напрямом проекції антиградієнта цільової функції на цю площину:

$$\mathbf{S}_{i+1} = \mathbf{H}_{i+1} \nabla f(\mathbf{X}_{i+1}),$$

де  $\mathbf{H}_{i+1} = \mathbf{E} - \mathbf{D}_{i+1}^T (\mathbf{D}_{i+1} \mathbf{D}_{i+1}^T)^{-1} \mathbf{D}_{i+1}$  – 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. 7.13 це 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  $\|\mathbf{V}(\mathbf{X}_{i+2})\| < \Delta v$ . Тоді наступним етапом буде етап 1 і процес пошуку повториться.

### Контрольні завдання

1. Вибрати функцію відповідно до свого варіанта.
2. Знайти мінімум функції методом золотого перетину.
3. Знайти мінімум функції методом градієнтного спуску.
4. Знайти мінімум функції методом Ньютона.
5. Порівняти результати пп. 2 – 4.

### Варіанти завдань

- |  |  |
|--|--|
| 1. $x^4, x \in [-1; 1]$ .                | 2. $x^4 - x, x \in [0,5; 1]$ .                     |
| 3. $x^3 - x, x \in [0; 1]$ .             | 4. $-x^3 + x, x \in [-1; 0]$ .                     |
| 5. $x^3 - 3x^2, x \in [1; 3]$ .          | 6. $-x^3 + 3x^2, x \in [-0,5; 0,5]$ .              |
| 7. $-x^3 + 3x, x \in [-2; 0]$ .          | 8. $x^3 - 3x, x \in [0; 2]$ .                      |
| 9. $x^2 - x, x \in [-1; 2]$ .            | 10. $x^4 + 3x, x \in [-2; 0]$ .                    |
| 11. $x^4 - 3x, x \in [0; 2]$ .           | 12. $\sin(x^2), x \in [-1; 1]$ .                   |
| 13. $\cos(x^2), x \in [1; 2]$ .          | 14. $\sin(x)/x, x \in [3; 6]$ .                    |
| 15. $-\sin^2(x), x \in [1; 2]$ .         | 16. $-\sin^2(x)e^x, x \in [1; 3]$ .                |
| 17. $(\cos(x) - 1)/x, x \in [1,5; 3]$ .  | 18. $5 \cosh(x), x \in [-1; 1]$ .                  |
| 19. $\sinh(x) - x^2, x \in [1; 3]$ .     | 20. $\sin(x)/x, x \in [-1; 1]$ .                   |
| 21. $\cosh(x) - x^2, x \in [1; 3]$ .     | 22. $x^2 - \sinh(x), x \in [0; 2]$ .               |
| 23. $\cosh(x) - 2x, x \in [0; 3]$ .      | 24. $\operatorname{Re}(x^x), x \in [-1,5; -0,5]$ . |
| 25. $x^2 - \ln(x), x \in [0,1; 1,6]$ .   | 26. $-e^{(-x^2)}, x \in [-1; 1]$ .                 |
| 27. $5e^{(-x)} \sin(x), x \in [-3; 2]$ . | 28. $2x/(x^2 + 1), x \in [-2; 0]$ .                |
| 29. $-0,2x^5 + x + 4, x \in [1; 2]$ .    | 30. $x^4 - 2x^3 - 1, x \in [1; 2]$ .               |



## 8. Апроксимація функцій

Нехай деяку функцію  $y(x)$  задано табл. 8.1.

**Таблиця. 8.1.** Деяка таблична функція

$x$	$x_1$	$x_2$	$\dots$	$x_n$
$y(x)$	$y_1$	$y_2$	$\dots$	$y_n$

За отриманими даними необхідно знайти порівняно просту функцію для обчислення цього значення. У четвертому розділі розглянуто застосування інтерполяції для розв'язання цієї задачі. Одним із суттєвих обмежень інтерполяційних формул є неможливість їх застосування поза кінцями відрізка, на якому проводилися спостереження. Крім того, 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) $y = \frac{1}{ax + b};$  |
| 2) $y = ax^2 + bx + c;$ | 6) $y = a \ln(x) + b;$      |
| 3) $y = ax^m;$          | 7) $y = a \frac{1}{x} + b;$ |
| 4) $y = ae^{(mx)};$     | 8) $y = \frac{x}{ax + b}.$  |

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

$$\mathbf{R} = [r_1, \dots, r_n]^T.$$

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

$$\|\mathbf{R}\|_2 = \sqrt{\sum_{i=1}^n r_i^2} = \sqrt{\sum_{i=1}^n (y_i - f(x_i))^2}$$

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).

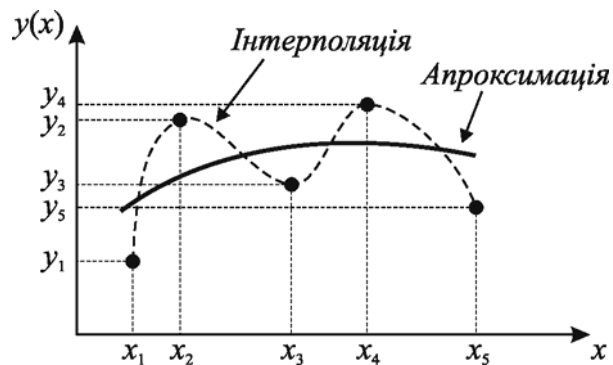


Рис. 8.1. Інтерполяція та апроксимація табличних залежностей

Оскільки мінімум норми Евкліда збігається з мінімумом її квадрата, то задачу середньоквадратичного наближення формулюють так:

$$\min \|\mathbf{R}\|_2^2 = \min \left[ \sum_{i=1}^n (y_i - f(x_i))^2 \right]. \quad (8.1)$$

Апроксимація функцій шляхом розв'язання (8.1) називається методом найменших квадратів (МНК).

Апроксимуючу функцію будують у вигляді

$$f(x) = f(x, c_1, \dots, c_m),$$

де  $c_i$  — коефіцієнти, які знаходять із необхідної умови існування мінімуму норми відхилу:

$$\frac{\partial \|\mathbf{R}\|}{\partial c_i} = 0, \quad i = \overline{1, m}. \quad (8.2)$$

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:

$$\min \left[ \sum_{i=1}^n \rho(x) (y_i - f(x_i))^2 \right],$$

де  $\rho(x)$  — вагова функція, яка, зокрема, може описувати ступінь імовірності чи важливості окремих точок.

Інколи неперервну функцію  $y(x)$ , задану на інтервалі  $[a, b]$ , необхідно апроксимувати простішою функцією  $f(x)$ . Як міру близькості  $f(x)$  до  $y(x)$  використовують умови:

$$1) \min \max_{x \in [a, b]} |y(x) - f(x)| - \text{неперервне рівномірне наближення};$$

$$2) \min \int_a^b \rho(x)(y(x) - f(x))^2 dx - \text{неперервне середньоквадратичне наближення}.$$

## 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

$$\sigma = \sum_{i=1}^n \rho(x_i)(y_i - f(x_i))^2 = \sum_{i=1}^n \rho(x_i) \left( y_i - \sum_{j=1}^m c_j \varphi_j(x_i) \right)^2. \quad (8.3)$$

З необхідної умови існування мінімуму (8.3) знаходимо:

$$\begin{aligned} \frac{\partial \sigma}{\partial c_1} &= 2 \sum_{i=1}^n \rho(x_i) \varphi_1(x_i) \left( y_i - \sum_{j=1}^m c_j \varphi_j(x_i) \right) = 0; \\ \frac{\partial \sigma}{\partial c_2} &= 2 \sum_{i=1}^n \rho(x_i) \varphi_2(x_i) \left( y_i - \sum_{j=1}^m c_j \varphi_j(x_i) \right) = 0; \\ &\dots \\ \frac{\partial \sigma}{\partial c_m} &= 2 \sum_{i=1}^n \rho(x_i) \varphi_m(x_i) \left( y_i - \sum_{j=1}^m c_j \varphi_j(x_i) \right) = 0. \end{aligned} \quad (8.4)$$

Уведемо позначення:

$$\begin{aligned} (\varphi_k, \varphi_j) &= \sum_{i=1}^n \varphi_k(x_i) \varphi_j(x_i) \rho(x_i); \\ (\varphi_k, y) &= \sum_{i=1}^n \varphi_k(x_i) y_i \rho(x_i). \end{aligned} \quad (8.5)$$

З урахуванням позначень (8.5), які називають скалярними добутками, необхідна умова існування мінімуму (8.4) має вигляд

$$\sum_{j=1}^m (\varphi_k, \varphi_j) c_j = (\varphi_k, y), \quad k = \overline{1, m}. \quad (8.6)$$

Визначник системи (8.6)

$$\begin{vmatrix} (\varphi_1, \varphi_1) & (\varphi_1, \varphi_2) & \cdots & (\varphi_1, \varphi_m) \\ (\varphi_2, \varphi_1) & (\varphi_2, \varphi_2) & \cdots & (\varphi_2, \varphi_m) \\ & & \cdots & \\ (\varphi_m, \varphi_1) & \cdots & \cdots & (\varphi_m, \varphi_m) \end{vmatrix}$$

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 додатно визначена, тому для її розв'язання доцільно використовувати метод  $\mathbf{LL}^T$ -декомпозиції.

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 вираховуються за формулами

$$\begin{aligned} (\varphi_k, \varphi_j) &= \int_a^b \rho(x) \varphi_k(x) \varphi_j(x) dx; \\ (\varphi_k, y) &= \int_a^b \rho(x) \varphi_k(x) y(x) dx. \end{aligned}$$

Задача спрощується, якщо система функцій  $\{\varphi_j(x)\}$  ортогональна. Тоді

$$(\varphi_k, \varphi_j) = 0, \quad k \neq j$$

і матриця системи (8.6) буде діагональною, а шукані коефіцієнти можна знайти так:

$$c_j = \frac{(\varphi_j, y)}{(\varphi_j, \varphi_j)}.$$

**Приклад.** Апроксимуємо поліномом першого степеня  $f(x) = c_1 + c_2x$  таблицю:

$x$	-2	0	1	2	4
$y$	-1	1	0	3	3

В даному випадку системою базисних функцій є  $\varphi_1(x) = 1$ ;  $\varphi_2(x) = x$ , а вагова функція  $\rho(x) \equiv 1$ . Тоді:

$$(\varphi_1, \varphi_1) = \sum_{i=1}^5 1 \cdot 1 = 5;$$

$$(\varphi_1, \varphi_2) = (\varphi_2, \varphi_1) = \sum_{i=1}^5 1x_i = -2 + 0 + 1 + 2 + 4 = 5;$$

$$(\varphi_2, \varphi_2) = \sum_{i=1}^5 x_i x_i = 4 + 0 + 1 + 4 + 16 = 25;$$

$$(\varphi_1, y) = \sum_{i=1}^5 1y_i = -1 + 1 + 0 + 3 + 3 = 6;$$

$$(\varphi_2, y) = \sum_{i=1}^5 x_i y_i = 2 + 0 + 0 + 6 + 12 = 20.$$

Систему (8.6) записуємо у вигляді

$$\begin{cases} 5c_1 + 5c_2 = 6; \\ 5c_1 + 25c_2 = 20 \end{cases} \Rightarrow \begin{cases} c_1 = 0,5; \\ c_2 = 0,7. \end{cases}$$

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. 8.2. Норма відхилу  $\|\mathbf{R}\|_2 = \sqrt{3}$ .

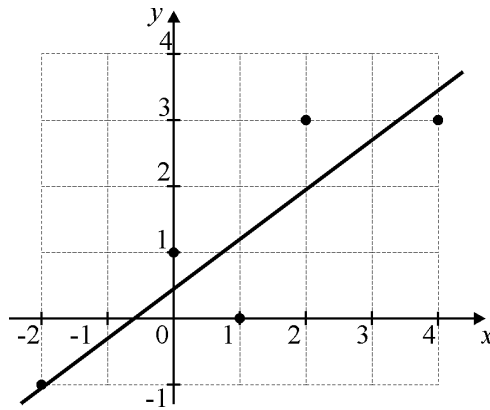


Рис. 8.2. Апроксимуюча функція

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 [4], 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 апроксимація може бути зведена до лінійної за допомогою спеціальних прийомів, зокрема логарифмування. Проілюструємо деякі прийоми на прикладах [2]:

1.  $f(x, c_1, c_2) = c_1 x^{c_2}$ .

Прологарифмуємо  $f(x, c_1, c_2)$ . У результаті маємо

$$\ln(f(x, c_1, c_2)) = \ln(c_1) + c_2 \ln(x).$$

Уведемо новий параметр  $\tilde{c}_0 = \ln(c_0)$ . Тоді

$$\ln(f(x, c_0, c_1)) = \tilde{f}(\tilde{x}, c_0, \tilde{c}_1) = \tilde{c}_0 + c_1 \ln(x)$$

Тобто задача звелась до лінійної. В ній  $\varphi_1(x) = 1$ ,  $\varphi_2(x) = \ln(x)$ .

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:

$$\left\{ \begin{array}{l} n\tilde{c}_1 + \left( \sum_{i=1}^n \ln(x_i) \right) c_2 = \sum_{i=1}^n \tilde{y}_i; \\ \left( \sum_{i=1}^n \ln(x_i) \right) \tilde{c}_1 + \left( \sum_{i=1}^n \ln^2(x_i) \right) c_2 = \sum_{i=1}^n \ln(x_i) \tilde{y}_i; \end{array} \right.$$

– обчислити  $c_1 = e^{\tilde{c}_1}$ .

2.  $f(x, c_1, c_2) = c_1 e^{c_2 x}$ .

Прологарифмуємо  $f(x, c_1, c_2)$ . Тоді

$$\ln(f(x, c_1, c_2)) = \ln(c_1) + c_2 x.$$

Позначимо:

$$\ln(c_1) = \tilde{c}_1.$$

У результаті маємо

$$\ln(f(x, c_1, c_2)) = \tilde{f}(x, \tilde{c}_1, c_2) = \tilde{c}_1 + c_2 x$$

Задачу зведено до лінійної. В ній  $\varphi_1(x) = 1$ ,  $\varphi_2(x) = x$ .

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.  $f(x, c_1, c_2) = c_1 x e^{c_2 x}$ .

Прологарифмуємо  $\frac{f(x, c_1, c_2)}{x}$ . Тоді

$$\ln\left(\frac{f(x, c_1, c_2)}{x}\right) = \ln(c_1) + c_2 x.$$

Позначимо:

$$\ln(c_1) = \tilde{c}_1.$$

У результаті маємо

$$\ln\left(\frac{f(x, c_1, c_2)}{x}\right) = \tilde{f}(x, \tilde{c}_1, c_2) = \tilde{c}_1 + c_2 x$$

Задачу зведено до лінійної. В ній  $\varphi_1(x) = 1$ ,  $\varphi_2(x) = x$ .

Для практичної реалізації необхідно:

– прологарифмувати відношення функції до аргументу  $\frac{y_i}{x_i}$  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  $c_1 = e^{\tilde{c}_1}$ .

$$4. f(x, c_1, c_2) = \frac{1}{c_1 x + c_2}.$$

Уведемо нову функцію

$$\tilde{f}(x, c_1, c_2) = \frac{1}{f(x, c_1, c_2)} = c_1 x + c_2.$$

Задачу зведено до лінійної. Зазначте, що в даному випадку  $\varphi_1(x) = x$ ,  $\varphi_2(x) = 1$ .

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  $c_1, c_2$ .

$$5. f(x, c_1, c_2) = \frac{x}{c_1 x + c_2}.$$

Розглянемо обернену функцію

$$\tilde{f}(x, c_1, c_2) = \frac{1}{f(x, c_1, c_2)} = \frac{c_1 x + c_2}{x} = c_1 + \frac{c_2}{x}.$$

Задачу зведено до лінійної. Далі будемо нову таблицю, у якій значення аргументу будуть обернені до заданих, і для отриманих значень знаходимо параметри  $c_1, c_2$ , враховуючи, що  $\varphi_1(x) = 1$ ,  $\varphi_2(x) = \frac{1}{x}$ .

Слід зазначити, що перелік наведених прикладів є далеко не вичерпним, а є лиш 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. Побудувати емпіричну залежність у вигляді  $f(x, a, b, c) = ax^2 + bx + c$ .
3. Побудувати графік отриманої залежності, та нанести точки таблиці. Порівняти отримані результати з результатами із четвертого пункту.

## Список використаної та рекомендованої 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. *Заварыкин В. М., Житомирский В. Г., Лапчик М. П.* Численные методы. – М.: Просвещение, 1991. – 176 с.
3. *Гловацкая А. П.* Методы и алгоритмы вычислительной математики. – М.: Радио и связь, 1999. – 406 с.
4. *Денис Дж., Шнабель Р.* Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1988. – 249 с.
5. *Демидович Б. П., Марон И. А.* Основы вычислительной математики. – М.: Наука, 1970. – 665 с.
6. *Васильков Ю. В., Василькова Н. Н.* Компьютерные технологии вычислений в математическом моделировании. – М.: Финансы и статистика, 1999. – 255 с.
7. *Бахвалов Н. С., Жидков Н. П., Кобельков Г. М.* Численные методы. – М.: Наука, 1987. – 600 с.
8. *Самарский А.А., Гулин А.В.* Численные методы. - М.:Наука,1989. – XXX с.
9. *Березин И. С., Жидков Н. П.* Методы вычислений: В 2 т. – М.: Физматиздат, 1962. – 640 с.
10. *Волков А. Е.* Численные методы. – М.: Наука, 1982. – 248 с.
11. *Воробьева Г. Н., Данилова А. Н.* Практикум по вычислительной математике. – М.: Высш. шк., 1990. – 308 с.
12. *Молчанов И.Н.* Машинные методы решения прикладных задач. Алгебра, приближение функций.- Киев: Наукова думка, 1987.- 288 с.
13. *Молчанов И.Н.* Машинные методы решения прикладных задач. Дифференциальные уравнения.- Киев: Наукова думка, 1988.- 344 с.
14. *Калиткин Н. Н.* Численные методы. – М.: Наука, 1978. – 500 с.
15. *Крылов В. И., Бобков В. Б., Монастырский П. И.* Вычислительные методы: В 2 т. – М.: Наука, 1977. – 339 с.
16. *Мак-Кракен Д., Дорн У.* Численные методы и программирование на ФОРТРАНЕ. – М.: Мир, 1977. – 293 с.
17. *Турчак Л. И.* Основы численных методов. – М.: Наука, 1987. – 350 с.
18. *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений. – М.: Мир, 1980. – 279 с.
19. *Шуп Т. Е.* Решение инженерных задач на ЭВМ. – М.: Мир, 1990. – 235 с.
20. *Боглаев Ю.П.* Вычислительная математика и программирование. - М.: Высшая школа,1990. –XXX с.

Навчальне видання

**Прокопенко** Юрій Васильович, **Татарчук** Дмитро Дмитрович,  
**Казміренко** Віктор Анатолійович

ОБЧИСЛЮВАЛЬНА МАТЕМАТИКА

Навчальний посібник

Редактор	<i>Г. І. Лоза</i>
Коректор	<i>Л. А. Мельник</i>
Комп'ютерна верстка	<i>В. Б. Лісничий</i>

Темплан 2002 р., поз. I / 59  
Темплан 2003 р., поз. I-2 / 009

Підп. до друку 16.04.2003. Формат 60×84<sup>1</sup>/<sub>16</sub>. Папір офс. Спосіб друку – ризографія.  
Ум. друк. арк. 6,97. Обл.-вид. арк. 11,6. Зам. № 3-43. Наклад 300 пр.

---

Інформаційно-видавничий центр “Видавництво «Політехніка»” НТУУ «КПІ»  
Свідоцтво про держреєстрацію ДК № 211 від 09.10.2000  
03056, Київ-56, вул. Політехнічна, 14, корп. 15