

Міністерство освіти і науки України
Дніпровський національний університет
імені Олеся Гончара

Кафедра математичного аналізу і оптимізації

ПОСІБНИК З ЧИСЕЛЬНИХ МЕТОДІВ:
ОСНОВИ.

Дніпро

2026

УДК 519.6(075.8)

П 61

У посібнику викладено базові поняття та інструменти чисельних методів, необхідні для наближеного розв'язування задач математичного аналізу. Розглянуто ключові теореми аналізу, джерела та оцінювання похибок обчислень, а також машинне представлення цілих і дійсних чисел (числа з рухомою комою). Наведено огляд найпростіших чисельних алгоритмів: схеми Горнера, різницевого наближення похідної, методу Ейлера для задачі Коші, лінійної інтерполяції, методу трапецій, розв'язання тридіагональних систем та крайових задач. Призначено для студентів університету, які вивчають чисельні методи та обчислювальну математику, для аудиторної та самостійної роботи.

Рекомендовано до друку вченою радою механіко-математичного факультету (протокол №7 від 17 лютого 2026 року).

Рецензенти:

О. О. Пипка, завідувач кафедри геометрії та алгебри Дніпровського національного університету імені Олеся Гончара, доктор фізико-математичних наук, доцент.

С. В. Конарева, вчитель математики ДОЛІФМП, кандидат фізико-математичних наук.

П 61 Посібник з чисельних методів: основи. [Електронний ресурс] : метод. посіб. / уклад. О. В. Коваленко. — Електрон. дані. — Дніпро : ДНУ, 2026. — 33 с.

Укладач: доктор фіз.-мат. наук, доцент, О. В. Коваленко

©Коваленко О. В., 2026

Зміст

Вступ	2
1 Основи чисельних методів	3
1.1 Основні теореми аналізу	3
1.2 Похибка обчислень	6
1.3 Машинне представлення чисел	7
1.3.1 Представлення цілих чисел	7
1.3.2 Представлення дійсних чисел. Числа з рухомою комою	8
1.3.3 Наближене обчислення функції $\ln x$	12
2 Огляд найпростіших методів і інструментів наближе- ного обчислення	14
2.1 Схема Горнера	14
2.2 Різницеве наближення похідної	15
2.3 Метод Ейлера для розв'язання задачі Коші	21
2.4 Лінійна інтерполяція	22
2.5 Метод трапеції	24
2.6 Тридіагональні системи лінійних рівнянь	27
2.7 Розв'язання крайової задачі	31
Бібліографія	33

Вступ

Матеріал посібника в основному слідує книзі [1] і розділено на дві логічні частини.

Розділ 1 містить теоретичний фундамент будь-яких обчислень. Ми почнемо з основ математичного аналізу, де ми формулюємо без доведень основні теореми аналізу. Їх доведення можна знайти у будь-якому підручнику з математичного аналізу, напр. [2, 4, 3]. Ми також зосереджуємо увагу на питаннях, які специфічні для роботи з ЕОМ. Ви дізнаєтесь, як виникають похибки, чому комп'ютер не може точно представити деякі числа (стандарт IEEE 754) та як ці обмеження впливають на результат обчислення функцій.

У розділі 2 ми розглянемо класичні інструменти чисельного аналізу. Це набір інструментів, що дозволяють: ефективно обчислювати значення поліномів, знаходити наближені значення похідних та інтегралів, а також розв'язувати диференціальні рівняння та системи лінійних рівнянь спеціального вигляду.

Розділ 1

Основи чисельних методів

1.1 Основні теореми аналізу

У даному підрозділі наведено класичні теореми математичного аналізу, які, не дивлячись на свою простоту, мають широке застосування у чисельних методах.

Теорема 1 (Тейлор). *Нехай функція $f: [a, b] \rightarrow \mathbb{R}$ має $n + 1$ неперервну похідну на відрізку $[a, b]$, $n \geq 0$ і $x_0 \in [a, b]$ — деяка фіксована точка. Тоді функцію f можна подати у вигляді*

$$f(x) = p_n(x_0; x) + R_n(x_0; x), x \in [a, b],$$

де

$$p_n(x_0; x) = \sum_{k=0}^n \frac{(x - x_0)^k}{k!} f^{(k)}(x_0),$$

а

$$R_n(x_0; x) = \frac{1}{n!} \int_{x_0}^x (x - t)^n f^{(n+1)}(t) dt.$$

Функція p_n називається многочленом Тейлора функції f , а R_n — залишковим членом (у інтегральній формі).

Відмітимо, що існує і інші представлення для залишкового члена, зокрема (залишковий член у формі Лагранжа) для деякого ξ між точками x_0 і x справедлива рівність

$$R_n(x_0; x) = \frac{(x - x_0)^{n+1}}{(n + 1)!} f^{(n+1)}(\xi). \quad (1.1)$$

Відмітимо, що многочлен Тейлора однозначно визначається умовами

$$p_n^{(k)}(x_0) = f^{(k)}(x_0), k = 0, 1, \dots, n.$$

Приклад 1. Обчислити значення $\sin x$ з точністю $\varepsilon = 10^{-2}$, $x \in \mathbb{R}$.

Спочатку знайдемо формулу, яка дозволить обчислити значення $\sin x$ з заданою точністю на відрізку $[0, 1]$.

Для цього виберемо у формулі Тейлора число n так, щоб нерівність $|R_n(0; x)| < \varepsilon$ виконувалась для всіх $x \in [0, 1]$. Використовуючи формулу для залишкового члена у формі Лагранжа для функції $f(x) = \sin x$ отримаємо, враховуючи, що $|f^{(k)}(x)| \leq 1$ для всіх $k \in \mathbb{N}$ і $x \in \mathbb{R}$,

$$|R_{2n}(x)| = \left| \frac{x^{2n+1}}{(2n + 1)!} f^{(2n+1)}(\xi) \right| \leq \frac{1}{(2n + 1)!}.$$

Легко бачити, що при $n = 2$ виконується нерівність $\frac{1}{(2n+1)!} < \varepsilon$, а отже для многочлена Тейлора $p_6(x) = x - \frac{x^3}{6} + \frac{x^5}{120}$ виконується нерівність

$$|\sin x - p_6(x)| < \varepsilon, x \in [0, 1].$$

Відмітимо, що на відміну від функції $\sin x$, для обчислення значення $p_6(x)$ у довільній точці x достатньо виконувати дії додавання та множення чисел.

Зазначимо, що $[0, 1] \supset [0, \frac{\pi}{4}]$, а отже многочлен p_6 дає ефективний спосіб обчислення функції $\sin x$ на відрізку $[0, \frac{\pi}{4}]$. Нехай $x \in [\frac{\pi}{4}, \frac{\pi}{2}]$. Враховуючи, що $\sin x = \cos(\frac{\pi}{2} - x)$, $x \in \mathbb{R}$, і той факт, що $x \in [\frac{\pi}{4}, \frac{\pi}{2}] \implies \frac{\pi}{2} - x \in [0, \frac{\pi}{4}]$, отримаємо, що для обчислення функції $\sin x$ на відрізку $[\frac{\pi}{4}, \frac{\pi}{2}]$ з заданою точністю, достатньо вміти обчислювати значення функції $\cos y$ на відрізку $[0, \frac{\pi}{4}]$.

Абсолютно аналогічно можна отримати, що для многочлена Тейлора $q_5(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24}$ функції $\cos x$ виконується нерівність

$$|\cos x - q_5(x)| < \varepsilon, x \in [0, 1] \supset [0, \frac{\pi}{4}],$$

а тому на відрізку $[0, \frac{\pi}{2}]$ з точністю ε справедлива формула

$$\sin x \approx \begin{cases} p_6(x), & x \in [0, \pi/4] \\ q_5(\pi/2 - x), & x \in (\pi/4, \pi/2]. \end{cases} \quad (1.2)$$

Використовуючи симетрії графіка функції $\sin x$, обчислення значення $\sin x$ можна звести до обчислення значення $\sin y$, $y \in [0, \frac{\pi}{2}]$. Тому значення $\sin x$, $x \in \mathbb{R}$ можна обчислити з заданою похибкою за допомогою формули (1.2).

Теорема 2 (Ролль). *Нехай функція $f: [a, b] \rightarrow \mathbb{R}$ є неперервною на відрізку $[a, b]$, диференційовною на інтервалі (a, b) і така, що $f(a) = f(b)$. Тоді існує $\xi \in (a, b)$, що $f'(\xi) = 0$.*

Теорема 3 (Лагранж). *Нехай функція $f: [a, b] \rightarrow \mathbb{R}$ є неперервною на відрізку $[a, b]$, диференційовною на інтервалі (a, b) . Тоді*

$$\exists \xi \in (a, b): \frac{f(b) - f(a)}{(b - a)} = f'(\xi).$$

Теорема 4 (Больцано, Коші). *Нехай $f: [a, b] \rightarrow \mathbb{R}$ є неперервною на відрізку $[a, b]$ і $f(a) \cdot f(b) < 0$. Тоді існує точка $\xi \in (a, b)$ така, що $f(\xi) = 0$.*

1.2 Похибка обчислень

Метою більшості матеріалу, який вивчається у даному курсі, є наближене обчислення невідомих величин. Тому має сенс більш точно означити поняття похибки. Припустимо, що ми хочемо обчислити невідому величину A , і B — її наближене значення. Тоді числа $A - B$, $|A - B|$ та $\frac{|A-B|}{|A|}$ називають *похибкою*, *абсолютною похибкою* та *відносною похибкою* відповідно (остання — припускаючи, що $A \neq 0$).

Приклад 2. Знайти абсолютну і відносну похибки для

1. $A = 0.1125351747 \times 10^{-6} \approx e^{-16}$ і $B = 0$

2. $A = 0.8886110521 \cdot 10^7 \approx e^{16}$ і $B = 0.888611052 \cdot 10^7$.

У першому випадку абсолютна похибка $|A - B| = A < 1.2 \cdot 10^{-7}$ є достатньо малою, в той же час відносна похибка $\frac{|A-B|}{|A|} = 1$ є великою.

У другому випадку, абсолютна похибка $|A - B| = 10^{-3}$ є достатньо великою, хоча числа A і B відрізняються тільки у останній значущій цифрі запису, а відносна похибка $\frac{|A-B|}{|A|} \approx 4.5 \cdot 10^{-10}$ є малою.

Таким чином, відносна похибка є мірою кількості ”правильних цифр” у наближенні. Враховуючи що, як правило, наближені алгоритми виконуються використовуючи машинне представлення дійсних чисел у вигляді чисел з рухомою комою, які мають фіксовану кількість значущих цифр (про це трохи більш детально ми будемо говорити далі), то відносна похибка може бути мірою якості наближення невідомої величини при комп’ютерних обчисленнях.

1.3 Машинне представлення чисел

1.3.1 Представлення цілих чисел

Як відомо, при записі чисел ми використовуємо позиційний запис у десятковій системі числення. Це означає, що кожне натуральне число представлено у вигляді

$$\overline{a_n a_{n-1} \dots a_1 a_0} = \sum_{k=0}^n a_k \cdot 10^k,$$

де $n \in \mathbb{N}$, $a_0, a_1, \dots, a_n \in \{0, 1, \dots, 9\}$ — це цифри десяткового запису. Число 10, очевидно, пов'язане з кількістю пальців у людей, але не має принципового значення для запису чисел. Якщо $p \in \mathbb{N}$, $p \geq 2$ — це довільне число, то будь-яке натуральне число може бути представлено єдиним чином у вигляді

$$(\overline{b_n b_{n-1} \dots b_1 b_0})_p = \sum_{k=0}^n b_k \cdot p^k,$$

де $n \in \mathbb{N}$, і тепер $b_0, \dots, b_n \in \{0, 1, \dots, p-1\}$ — "цифри" у p -системі (цифр має бути p штук).

З технічних причин комп'ютеру простіше працювати з двома цифрами, відповідно природною для комп'ютера системою числення є двійкова (бінарна).

Приклад 3. Представити число 26 у двійковій системі числення.

Розглянемо два способи переходу від десяткової до двійкової системи числення.

Перший спосіб полягає у представленні числа 26 у вигляді суми степенів двійки, кожен з яких потім легко записати у двійковій системі числення:

$$26 = 16 + 8 + 2 = 2^4 + 2^3 + 2^1 = 10000_2 + 1000_2 + 10_2 = 11010_2.$$

Другий спосіб полягає у послідовному діленні числа на 2 з залишком, як продемонстровано нижче. Двійкове зображення числа отримуємо послідовним записом залишків від ділення починаючи знизу вгору.

26	
13	0
6	1
3	0
1	1
0	1

Відмітимо, що другий спосіб за своєю суттю повторює перший; доведення його коректності залишимо у якості вправи для читача.

Для виконання арифметичних дій над цілими числами на комп'ютері, зазвичай, на представлення кожного числа виділяється певна кількість пам'яті. Наприклад, за допомогою 1 байта ($= 8$ біт) можна представити $2^8 = 256$ різних значень; зазвичай це числа від 0 до 255 (беззнакові числа) або від -128 до 127 (знакові; іноді від -127 до 128).

Відмітимо, що, наприклад, при виконанні додавання $255 + 1$ у випадку 1-байтних беззнакових чисел, відбувається "переповнення" типу:

$$255 + 1 = 11111111_2 + 1_2 = 100000000_2,$$

але оскільки число представляється тільки 1 байтом, то на дев'яту одиницю не вистачає місця, і результатом додавання буде число $0 = 00000000_2$.

1.3.2 Представлення дійсних чисел. Числа з рухомою комою

На відміну від цілих чисел, яких існує скінченна кількість на будь-якому відрізку числової прямої, множина дійсних чисел на будь-

якому (невиродженому) відрізку є нескінченною. Тому представлення дійсних чисел на комп'ютері має принципову відмінність: не кожне дійсне число може бути збереженим у пам'яті. Таким чином при виконанні алгоритмів наближених обчислень на комп'ютері виникає додаткове джерело похибки. Окрім похибки власне наближеного алгоритму, виникає похибка, пов'язана з наближеним представленням чисел у пам'яті комп'ютера. Ця похибка наближеного представлення чисел може акумулюватись і, як результат, стати суттєво більшою за похибку наближеного алгоритму. Для того, щоб мати можливість цього уникнути, варто мати розуміння того, як дійсні числа представляються у пам'яті комп'ютера і які наслідки має цей аспект.

У більшості мов програмування дійсні числа представляються у формі

$$x = \sigma \cdot f \cdot 2^t, \quad (1.3)$$

де $\sigma \in \{1, -1\}$ — це знак числа (на нього відводиться 1 біт), f — це ціле число, яке називається мантиса, t — експонента. Зазвичай на дійсне число відводять 4 (тип float) або 8 (тип double) байтів. Наприклад, для типу float, згідно з стандартом IEEE 754, на мантису відводять 23 біти, а на експоненту — 8 (і ще один біт на знак). Відмітимо, що

$$f \cdot 2^t = (2f) \cdot 2^{t-1}.$$

Тому для забезпечення єдиності представлення числа у вигляді (1.3) біти чисел f та t зазвичай інтерпретують таким чином. Нехай

$$\overline{f_1 \dots f_{23}} \text{ та } \overline{t_1 \dots t_8}$$

— це біти чисел f і t відповідно. Тоді

$$f = (1.f_1 f_2 \dots f_{23})_2 = 1 + f_1 \cdot 2^{-1} + \dots + f_{23} \cdot 2^{-23}, \quad (1.4)$$

так що $f \in [1, 2 - \frac{1}{2^{24}}]$, а t інтерпретується як знакове число з 8 біт, $-127 \leq t \leq 128$. Числа, які можна представити у такому вигляді,

називаються нормалізованими. Легко бачити, що 0 у такому вигляді представити неможливо. Тому виникає необхідність ускладнення представлення деяких чисел. Зокрема, при найменшому значенні експоненти $t = -127$, приходимо до денормалізованого представлення чисел — мантиса інтерпретується як $(0.f_1f_2 \dots f_{23})_2$ замість нормалізованого виду (1.4).

З написаного вище видно, що на відміну від множини дійсних чисел, існує найменше додатне число, яке можна представити у вигляді числа з рухомою комою. Наприклад, для 4-байтних чисел з рухомою комою найменшим додатним числом буде число $2^{-149} = 2^{-23} \cdot 2^{-126} \approx 1.4 \cdot 10^{-45}$: мантиса $0.00 \dots 01_2 = 2^{-23}$ і експонента -127 , яка вказує на денормалізоване число і інтерпретується як 2^{-126} .

Приклад 4. Обчислити суму $a + b$, де $a = 1$ і $b = 2^{-4}$, якщо кожне з чисел представлено у форматі з рухомою комою.

Для числа a мантиса $f_a = 1$, та експонента $t_a = 0$; для числа b маємо $f_b = 1$, $t_b = -4$. Для того, щоб додати ці числа приведемо їх до спільного знаменника: $f_b = 0.0001_2$, $t_b = 0$. Тоді

$$a + b = f_a \cdot 2^{t_a} + f_b \cdot 2^{t_b} = (1 + 0.0001)_2 \cdot 2^0 = 1.0001_2 \cdot 2^0.$$

Приклад 5. Обчислити суми $(-2^{30} + 2^{30}) + 1$ та $-2^{30} + (2^{30} + 1)$, вважаючи, що на мантису відводиться 23 біти.

Перша сума, очевидно, буде дорівнювати 1. Розглянемо другу суму.

Для додавання 2^{30} і 1 треба друге число привести до денормалізованої форми з експонентою 30. Але це означає, що мантиса має бути 2^{-30} , тобто мати вигляд $0.00 \dots 01$, де після коми іде 29 нулів. Оскільки мантиса має лише 23 біти, то отримаємо число $0.00 \dots 0_2 = 0$, і отже $2^{30} + 1 = 2^{30}$. Тому $-2^{30} + (2^{30} + 1) = 0$.

Зокрема бачимо, що додавання чисел з рухомою комою є неасоціативним.

Приклад 6. Представити число 0.1_{10} у двійковій системі з рухомою комою з 8 бітами для мантиси. Оцінити абсолютну похибку представлення.

Легко бачити, що (рівності у десятковій системі числення)

$$\begin{aligned} 0.1 &= 2^{-4} \cdot 1.6 = 2^{-4} + 2^{-4} \cdot 0.6 = 2^{-4} + 2^{-5} + 2^{-5} \cdot 0.2 \\ &= 2^{-4} + 2^{-5} + 2^{-8} + 2^{-8} \cdot 0.6. \end{aligned}$$

Тому двійковий запис числа 0.1_{10} є періодичним двійковим дробом $0.0001(1001)_2$. Нормалізуючи його і залишаючи тільки 8 біт мантиси, отримаємо представлення

$$0.1_{10} \mapsto 1.10011001_2 \cdot 2^{-4}.$$

Неважко перевірити, що

$$1.10011001_2 \cdot 2^{-4} = 1.59765625_{10} \cdot 2^{-4} = 0.099853515625_{10}$$

і абсолютна похибка представлення дорівнює 0.000146484375 .

Приклад 7. Обчислити різницю $a - b$, де

$$a = e^{-1/(100)^2}, b = e^{-1/(1000)^2},$$

використовуючи десяткові наближення з

1. 8 значущими цифрами;
2. 16 значущими цифрами.

Скільки точних цифр мають обчислені різниці?

Для наближень з 8 значущими цифрами маємо $a = 0.99990001$, $b = 0.99999900$, а тому $a - b = -0.00009899$.

Для наближень з 16-ма значущими цифрами маємо

$$a = 0.9999000049998333, b = 0.9999990000004999,$$

а тому $a - b = -0.0000989950006666$.

Відмітимо, що перше наближення має лише 4 правильні цифри різниці, а друге — 12.

Даний приклад ілюструє феномен втрати значущості при відніманні приблизно однакових чисел; аналогічну ситуацію ми бачили у прикладі 5, коли додавали числа, які відрізняються у багато разів.

1.3.3 Наближене обчислення функції $\ln x$

В даному підрозділі ми покажемо, як, враховуючи представлення дійсних чисел (1.3), можна обчислити значення функції $f(x) = \ln x$ з заданою (відносною) точністю для всіх $x > 0$. Враховуючи (1.3), маємо

$$x = f \cdot 2^\beta, f \in [1, 2), \beta \in \mathbb{Z}$$

$$\ln x = \ln(f \cdot 2^\beta) = \ln f + \ln 2^\beta = \ln f + \beta \cdot \ln 2,$$

а тому, вважаючи число $\ln 2$ відомим з необхідною точністю, нам залишається порахувати число $\ln f$, де $f \in [1, 2)$. Враховуючи те, що $\ln f = -\ln f^{-1}$, можемо вважати, що $f \in (1/2, 1]$.

Як відомо, для $n \in \mathbb{N}$ і $x > 0$, $f^{(n)}(x) = \frac{(-1)^{n-1}(n-1)!}{x^n}$, а тому для $x_0, x > 0, n \in \mathbb{N}$ за формулою Тейлора з залишком у інтегральній формі маємо

$$\begin{aligned} \ln x = \ln x_0 + \frac{x - x_0}{x_0} + \dots + \frac{(-1)^{n-1}}{n} \left(\frac{x - x_0}{x_0} \right)^n \\ + (-1)^n \int_{x_0}^x (x - t)^n t^{-n-1} dt. \end{aligned}$$

Число $x_0 \in (1/2, 1]$ виберемо пізніше. Оцінимо залишок:

$$|R_n(x)| = \left| \int_{x_0}^x (x-t)^n t^{-n-1} dt \right| \leq |x - x_0| \cdot \max_{t \in [x_0, x]} \frac{|x-t|^n}{t^{n+1}}.$$

Маємо для кожного $t \in [x_0, x] \subset [1/2, 1)$,

$$\frac{|x-t|^n}{t^{n+1}} = \frac{1}{t} \cdot \left| \frac{x-t}{t} \right|^n \leq 2 \left| \frac{x-t}{t} \right|^n. \quad (1.5)$$

Розглянемо функцію

$$g(t) = \frac{x-t}{t} = \frac{x}{t} - 1,$$

і знайдемо її екстремуми на відрізку з кінцями в точках x_0 і x .

$$g'(t) = -\frac{x}{t^2} < 0,$$

тобто функція g спадає. Тому її найбільше і найменше значення досягається на кінцях відрізка. Враховуючи, що $g(x) = 0$ і $g(x_0) = \frac{x-x_0}{x_0}$, отримаємо, що

$$\max_t |g(t)| = |g(x_0)|.$$

Тому, продовжуючи ланцюжок (1.5), і вибираючи $x_0 = \frac{3}{4}$, отримаємо

$$\frac{|x-t|^n}{t^{n+1}} \leq 2 \cdot \left| \frac{x-x_0}{x_0} \right|^n \leq 2 \cdot \left| \frac{\frac{1}{4}}{\frac{3}{4}} \right|^n = \frac{2}{3^n}.$$

Таким чином, $|R_n(x)| \leq \frac{2}{3^n}$. Вибираючи n таким, щоб $\frac{2}{3^n}$ було меншим за необхідну похибку, ми отримаємо спосіб обчислення значення $\ln x$.

Розділ 2

Огляд найпростіших методів і інструментів наближеного обчислення

2.1 Схема Горнера

У попередньому розділі ми бачили, що наближення функції поліномом може бути ефективним способом для чисельних обчислень. В даному підрозділі ми обговоримо ефективний спосіб обчислення значень поліномів.

Нехай задано многочлен

$$P_n(x) = a_0 + a_1x + \dots + a_nx^n, n \in \mathbb{N}.$$

Якщо підходити до обчислення значення $P_n(x_0)$, $x_0 \in \mathbb{R}$ наївно, використовуючи k дій множення для обчислення кожного з мономів $a_kx_0^k = a_k \cdot x_0 \cdot \dots \cdot x_0$, то для обчислення значення $P_n(x_0)$ буде виконано $1 + 2 + \dots + n = O(n^2)$, $n \rightarrow \infty$ дій множення, що неефективно.

Якщо ж подати P_n у вигляді

$$P_n(x) = a_0 + x(a_1 + x(a_2 + \dots + x(a_{n-1} + a_n x) \dots)),$$

то обчислення значення $P_n(x_0)$ потребуватиме $O(n)$ множень і $O(n)$ додавань дійсних чисел, $n \rightarrow \infty$, що є значно ефективніше.

2.2 Різницеве наближення похідної

Важливим у математиці взагалі, і у чисельних методах зокрема, є поняття похідної функції. У випадку, коли функція задана аналітично, цю похідну іноді також можна записати аналітично. Але, як ми бачили раніше, навіть у найпростіших випадках, обчислення заданої аналітично функції в деякій точці може бути нетривіальною задачею. Крім того, часто аналітичний вигляд функції є невідомим, або навіть тільки відомі значення функції у деяких точках: наприклад, як результати вимірювання деякої величини. Тому виникає питання наближеного значення похідної в деякій точці, маючи інформацію про значення функції в околі цієї точки.

Найпростіший спосіб наближеного обчислення похідної впливає з означення:

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \approx \frac{f(x+h) - f(x)}{h}, \quad (2.1)$$

тобто при малих значеннях h частка справа має бути близькою до $f'(x)$. Виникає питання, як розуміти цю неформальну фразу, і чи можна оцінити похибку такого наближення значення $f'(x)$. Ми розглянемо задачу у такій постановці.

Нехай задано деякий клас X диференційовних на відрізьку $[a, b]$ функцій і $x_0 \in (a, b)$. Знайти величину

$$e_1(X; x_0, h) = \sup_{f \in X} \left| f'(x_0) - \frac{f(x_0 + h) - f(x_0)}{h} \right|,$$

де h достатньо мале число (таке, щоб $x_0 + h \in [a, b]$). Для простоти нехай $X = W_\infty^2[a, b]$ — це клас двічі неперервно диференційовних на $[a, b]$ функцій f таких, що $\max_{t \in [a, b]} |f''(t)| \leq 1$.

Записуючи залишковий член для многочлена Тейлора у формі Лагранжа (1.1), отримуємо для будь-якої $f \in W_\infty^2[a, b]$ і деякого $\xi \in [a, b]$

$$\left| f'(x_0) - \frac{f(x_0 + h) - f(x_0)}{h} \right| = \left| f'(x_0) - \frac{hf'(x_0) + \frac{1}{2}h^2f''(\xi)}{h} \right| \leq \frac{|h|}{2},$$

і отже $e_1(W_\infty^2[a, b]; x_0, h) \leq \frac{|h|}{2}$. З іншого боку, як легко бачити, функція $f(x) = \frac{x^2}{2}$, $x \in [a, b]$ належить простору $W_\infty^2[a, b]$ і для неї для будь-якого $x_0 \in [a, b]$ маємо

$$\left| f'(x_0) - \frac{f(x_0 + h) - f(x_0)}{h} \right| = \left| x_0 - \frac{(x_0 + h)^2 - x_0^2}{2h} \right| = \frac{|h|}{2},$$

і отже $e_1(W_\infty^2[a, b]; x_0, h) \geq \frac{h}{2}$. Таким чином, ми довели таку теорему:

Теорема 5. Для будь-якого відрізка $[a, b]$, і точок $x_0 \in (a, b)$, $h > 0$ таких, що $x_0 + h \in [a, b]$ має місце рівність

$$e_1(W_\infty^2[a, b]; x_0, h) = \frac{|h|}{2}.$$

Зокрема, з цієї теореми випливає, що похибка формули (2.1) на класі $W_\infty^2[a, b]$ спадає до нуля лінійно відносно h .

Розглянемо тепер клас $X = W_\infty^3[a, b]$ функцій f , які є три рази неперервно диференційовні і для яких $\max_{t \in [a, b]} |f'''(t)| \leq 1$. Очевидно, що функція $f(x) = \frac{x^3}{6}$ лежить у цьому класі і, як ми бачили, похибка наближення похідної цієї функції за допомогою метода (2.1) буде $O(h)$, $h \rightarrow 0$.

Натомість, розглянемо такий метод наближеного обчислення похідної:

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h} \quad (2.2)$$

і оцінимо похибку

$$e_2(W_\infty^3[a, b]; x_0, h) := \sup_{f \in W_\infty^3[a, b]} \left| f'(x) - \frac{f(x+h) - f(x-h)}{2h} \right|.$$

Знову, застосовуючи формулу Тейлора (але тепер беремо многочлен Тейлора степеня 2) з залишковим членом у формі Лагранжа, отримаємо для деяких $\xi, \zeta \in (a, b)$

$$f(x_0 + h) = f(x_0) + f'(x_0) \cdot h + \frac{h^2}{2} f''(x_0) + \frac{h^3}{3!} f'''(\xi),$$

$$f(x_0 - h) = f(x_0) - f'(x_0) \cdot h + \frac{h^2}{2} f''(x_0) - \frac{h^3}{3!} f'''(\zeta),$$

звідки, віднімаючи отримані рівності, отримаємо, що

$$f(x_0 + h) - f(x_0 - h) = 2f'(x_0)h + \frac{h^3}{6}(f'''(\xi) + f'''(\zeta)),$$

а отже

$$\left| f'(x_0) - \frac{f(x_0 + h) - f(x_0 - h)}{2h} \right| = \frac{h^2}{12} |f'''(\xi) + f'''(\zeta)| \leq \frac{h^2}{6}.$$

Звідси випливає, що $e_2(W_\infty^3[a, b]; x_0, h) \leq \frac{h^2}{6}$. Аналогічно до попереднього результату, розглядаючи функцію $x \mapsto \frac{x^3}{6} \in W_\infty^3[a, b]$, отримаємо нерівність $e_2(W_\infty^3[a, b]; x_0, h) \geq \frac{h^2}{6}$. Таким чином ми довели таку теорему.

Теорема 6. Для будь-якого відрізка $[a, b]$, і точок $x_0 \in (a, b)$, $h > 0$ таких, що $x_0 \pm h \in [a, b]$ має місце рівність

$$e_1(W_\infty^3[a, b]; x_0, h) = \frac{|h|}{2} \text{ та } e_2(W_\infty^3[a, b]; x_0, h) = \frac{h^2}{6}. \quad (2.3)$$

Зокрема, ми бачимо що на класі $W_\infty^3[a, b]$ метод (2.2) є кращим за (2.1) і похибка наближення (2.2) спадає пропорційно квадрату h .

Отримані у попередніх двох теоремах оцінки означають, що при прямуванні $h \rightarrow 0$ величини похибок e_1 та e_2 також прямують до нуля.

Приклад 8. Обчислити наближені значення похідної $f'(1)$ в точці 1 для функції $f(x) = e^x$, використовуючи методи (2.1) і (2.2) при $h = 2^{-k}$, для декількох перших $k \in \mathbb{N}$.

Відмітимо, що точне значення $f'(1) = e$. Використаємо для обчислень 4-байтні числа з рухомою комою і занесемо результати обчислень у таблицю: у рядку з номером k , $D_1 = (e^{1+2^{-k}} - e^1)/(2^{-k})$, $E_1 = e - D_1$, $D_2 = (e^{1+2^{-k}} - e^{1-2^{-k}})/(2^{1-k})$, $E_2 = e - D_2$.

k	D_1	E_1	D_2	E_2
1	3.526814	-0.808533	2.832968	-0.114686
2	3.088244	-0.369963	2.746686	-0.028404
3	2.895481	-0.177199	2.725367	-0.007085
4	2.805027	-0.086745	2.720053	-0.001771
5	2.761200	-0.042918	2.718723	-0.000442
6	2.739639	-0.021358	2.718391	-0.000110
7	2.728943	-0.010661	2.718307	-0.000026
8	2.723633	-0.005351	2.718292	-0.000010
9	2.720947	-0.002666	2.718262	0.000020
10	2.719727	-0.001445	2.718262	0.000020

Відмітимо декілька аспектів отриманих результатів. По-перше, при зменшенні h вдвічі (тобто при переході до наступного рядка у таблиці), значення $|E_1|$ зменшується приблизно вдвічі, що узгоджується з оцінкою (2.3). В той же час величина E_2 зменшується приблизно у 4 раз до $k = 7$, і перестає зменшуватись при $k > 7$.

Якщо зробити такі самі обчислення, але використовуючи 8-байтні числа з рухомою комою, то отримаємо такі результати.

k	D_2	E_2
1	2.83296779963794	-0.11468597117889
2	2.74668588169833	-0.02840405323929
3	2.72536621980373	-0.00708439134469
4	2.72005188887067	-0.00177006041163
5	2.71872427874553	-0.00044245028648
6	2.71839243697998	-0.00011060852094
7	2.71830948033613	-0.00002765187709
8	2.71828874141249	-0.00000691295345
9	2.71828355669641	-0.00000172823736
10	2.71828226051821	-0.00000043205916

Бачимо, що у цих результатах зменшення h вдвічі зменшує похибку приблизно в чотири рази, що і мало б бути відповідно до рівностей (2.3).

В чому причина "аномалії" у першій таблиці?

Під час доведення формул (2.3), ми вважали, що обчислення роблять точно, тобто ігнорували похибки округлення, які виникають при машинних обчисленнях. При реальних обчисленнях значень з таблиць, обчислення було зроблено використовуючи числа з рухомою комою; при цих обчисленнях виникали похибки округлення. Розглянемо цю ситуацію більш детально. Позначимо через $\bar{f}(x)$ значення функції f в точці x , отримане на комп'ютері, а через $\varepsilon(x) = f(x) - \bar{f}(x)$ — похибку машинного обчислення цього значе-

ння. Ця похибка буде малою (тим меншою, чим більше пам'яті ми відводимо на збереження дійсного числа у пам'яті комп'ютера), але, взагалі кажучи, не буде нульовою. Наближене значення похідної D_2 , насправді, використовує при обчисленні не значення функції f , а значення функції $\bar{f}$. Позначимо $\bar{D}_2(x) = \frac{\bar{f}(x+h) - \bar{f}(x-h)}{2h}$ і розглянемо величину $\bar{E}_2 = f'(x) - \bar{D}_2(x)$ між величиною $f'(x)$, яку ми хочемо обчислити, і величиною $\bar{D}_2(x)$, яку ми можемо обчислити. Маємо

$$\begin{aligned}\bar{E}_2(x) &= f'(x) - \bar{D}_2(x) \\ &= \left(f'(x) - \frac{f(x+h) - f(x-h)}{2h} \right) + \\ &\quad + \frac{(f(x+h) - \bar{f}(x+h)) - (f(x-h) - \bar{f}(x-h))}{2h} \\ &= \left(f'(x) - \frac{f(x+h) - f(x-h)}{2h} \right) + \frac{\varepsilon(x+h) - \varepsilon(x-h)}{2h}.\end{aligned}$$

Відмітимо, що перший доданок, згідно доведеного раніше, прямує до нуля зі швидкістю $O(h^2)$, $h \rightarrow 0$. В той же час, чисельник другого доданку буде маленьким, але ненульовим для достатньо малих h , а отже другий доданок не буде малим при достатньо малих h . Це і пояснює поведінку чисельних значень величини E_2 у першій таблиці. В той же час, навіть при використанні більшої кількості пам'яті для представлення чисел, ефект накопичення помилки (і зникнення зменшення похибки обчислення значення $f'(x)$) буде відбуватись при достатньо великих k .

2.3 Метод Ейлера для розв'язання задачі Коші

Розглянемо задачу Коші для звичайного диференціального рівняння

$$y'(t) = f(t, y), y(t_0) = y_0. \quad (2.4)$$

Якщо використати наближення (2.1) для значення похідної $y'(t)$, то отримаємо

$$\frac{y(t+h) - y(t)}{h} = f(t, y(t)) + \frac{1}{2}hy''(t_h), t_h \in (t, t+h),$$

звідки

$$y(t+h) = y(t) + hf(t, y(t)) + \frac{1}{2}h^2y''(t_h). \quad (2.5)$$

Отримана формула мотивує до розгляду послідовності $\{y_n\}$, що задається рекурентно

$$y_{n+1} = y_n + hf(t_n, y_n), n \in \mathbb{N}.$$

де $t_n = t_{n-1} + h$, $n \in \mathbb{N}$, $t_0 = x_0$. Числа y_n — це наближені значення невідомого розв'язку y рівняння (2.4).

Оцінку для відхилення між обчисленим наближенням функції y у точках t_k , $k \in \mathbb{N}$ від розв'язку задачі (2.4) дає така теорема.

Теорема 7. *Припустимо, що функція f є ліпшицевою з константою $L > 0$ по другій змінній, тобто виконується нерівність*

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2| \forall t, y_1, y_2,$$

а про розв'язок y задачі (2.4) відомо, що $\max_t |y''(t)| \leq M$, $M \geq 0$. Тоді для всіх n має місце нерівність

$$|y(t_n) - y_n| \leq \frac{M}{2L}[e^{L(t_n - t_0)} - 1]h.$$

Доведення. Нехай $e_n = y(t_n) - y_n$. Тоді використовуючи (2.5) і умови теореми, отримаємо

$$\begin{aligned} |e_{n+1}| &= |y(t_{n+1}) - [y_n + hf(t_n, y_n)]| \\ &= \left| \left[y(t_n) + hf(t_n, y(t_n)) + \frac{1}{2}h^2y''(\xi) \right] - [y_n + hf(t_n, y_n)] \right| \\ &= \left| e_n + h[f(t_n, y(t_n)) - f(t_n, y_n)] + \frac{1}{2}h^2y''(\xi) \right| \leq |e_n| + hL|e_n| + \frac{h^2M}{2}. \end{aligned}$$

Таким чином, ми отримали рекурентну нерівність

$$|e_{n+1}| \leq (1 + hL)|e_n| + \frac{h^2M}{2}, n \geq 0, e_0 = 0,$$

з якої індукцією по n легко вивести нерівність

$$|e_n| \leq \frac{Mh}{2L} ((1 + hL)^n - 1), n \geq 0.$$

Нарешті, використовуючи нерівність $1 + x \leq e^x$, $x \in \mathbb{R}$, і той факт, що $nh = t_n - t_0$, отримаємо

$$|e_{n+1}| \leq \frac{Mh}{2L} (e^{hLn} - 1) = \frac{M}{2L} (e^{L(t_n - t_0)} - 1) h.$$

□

Зазначимо, що метод Ейлера дає оцінки для значень невідомого розв'язку задачі Коші тільки у наборі точок t_k , $k \in \mathbb{N}$, і якщо відрізок, на якому нас цікавить розв'язок задачі фіксований, то похибка наближення має асимптотичну оцінку $O(h)$, $h \rightarrow 0$.

2.4 Лінійна інтерполяція

У попередньому підрозділі ми отримали спосіб чисельного наближення значень розв'язку диференціального рівняння у деякому наборі

точок. Спираючись на нього, можна побудувати неперервну функцію, що є наближенням розв'язку вже на всьому відрізку. В даному підрозділі ми обговоримо лінійну інтерполяцію, яка важлива також і в багатьох інших питаннях.

Нехай задано функцію $f: [a, b] \rightarrow \mathbb{R}$. Легко бачити, що лінійна функція p , що задається формулою

$$p(x) = \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b)$$

інтерполуює функцію f в точках a, b , тобто справедливі рівності.

$$p(a) = f(a), p(b) = f(b).$$

Теорема 8. *Якщо функція f є двічі неперервно диференційовною, то для кожного $x \in [a, b]$*

$$|f(x) - p(x)| \leq \frac{1}{8}(b-a)^2 \max_{t \in [a, b]} |f''(t)|.$$

Доведення. Як впливає з загальної теореми про відхилення функції від її інтерполяційного многочлену (див. напр., [1, Theorem 4.3]), для кожного $t \in [a, b]$ існує $\xi \in (a, b)$ таке, що

$$f(t) - p(t) = \frac{f''(\xi)}{2}(t-a)(t-b).$$

Твердження теореми тепер впливає з того факту, що

$$\max_{t \in [a, b]} |(t-a)(t-b)| = \frac{1}{4}(b-a)^2.$$

□

2.5 Метод трапеції

Однією з найважливіших і найдавніших задач чисельних методів є (наближене) обчислення інтеграла функції, тобто значення функціонала

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

Простою, але важливою ідеєю для отримання наближеного значення функціонала від 'складної' функції f є така: замінити функцію f деякою, більш 'простою' функцією p , для якої значення функціоналу можна порахувати точно, і у якості наближеного значення для функції f взяти обчислене значення функціонала на функції p .

Якщо замість функції f ми візьмемо лінійну функцію p , що інтегрує f у точках a і b , то отримаємо, враховуючи результати попереднього підрозділу,

$$\begin{aligned} I(f) &\approx I(p) = \int_a^b p(x)dx = \int_a^b \left(\frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b) \right) dx \\ &= \frac{1}{2(b-a)} \left(-(b-x)^2 f(a) + (x-a)^2 f(b) \right) \Big|_a^b = \frac{1}{2}(b-a)(f(a) + f(b)). \end{aligned}$$

Крім того, для двічі неперервно диференційовної функції $f: [a, b] \rightarrow \mathbb{R}$, маємо

$$\begin{aligned} |I(f) - I(p)| &= \left| \int_a^b (f(x) - p(x))dx \right| \leq \int_a^b |f(x) - p(x)| dx \\ &= \int_a^b \left| \frac{f''(\xi(t))}{2}(t-a)(t-b) \right| dt \\ &\leq \frac{\max_{t \in [a,b]} |f''(t)|}{2} \int_a^b |(t-a)(t-b)| dt = \frac{1}{12}(b-a)^3 \max_{t \in [a,b]} |f''(t)|. \end{aligned}$$

Відмітимо, що для функції

$$f(x) = c \cdot (x - a)(x - b), c \in \mathbb{R}, \quad (2.6)$$

маємо $p(x) \equiv 0$ і усі нерівності у попередньому ланцюжку перетворюються на рівність. Тому ми довели таку теорему.

Теорема 9. *Для формули трапеції наближеного інтегрування*

$$\int_a^b f(x)dx \approx \frac{1}{2}(b-a)(f(a) + f(b)) \quad (2.7)$$

справедлива така оцінка похибки:

$$\left| \int_a^b f(x)dx - \frac{1}{2}(b-a)(f(a) + f(b)) \right| \leq \frac{(b-a)^3}{12} \max_{t \in [a,b]} |f''(t)|.$$

Ця оцінка є точною на класі двічі неперервно диференційовних функцій, в тому сенсі, що вона перетворюється на рівність на функціях (2.6).

Відмітимо, що права частина у оцінці похибки формули трапеції пропорційна величині $\max_{t \in [a,b]} |f''(t)|$ (яка характеризує швидкість зміни похідної функції f) та кубу довжини відрізка інтегрування. Тому формула трапеції у вигляді (2.7) може давати велику похибку на відрізках великої довжини, і в той же похибка швидко стає малою при зменшенні довжини відрізка.

Припустимо тепер, що задано деяке розбиття $a = x_0 < x_1 < \dots < x_n = b$ відрізка $[a, b]$, яке ми будемо для простоти вважати рівномірним, тобто $x_i - x_{i-1} = h = \frac{b-a}{n}$, $i = 1, \dots, n$ і ми знаємо значення функції f у точках розбиття. Позначимо через p_n кусково лінійну функцію, яка інтерполює значення f у точках розбиття. Застосовуючи на кожному відрізку $[x_{i-1}, x_i]$, $i = 1, \dots, n$ формулу

трапеції (2.7), прийдемо до такої формули наближеного обчислення інтеграла $I(f)$, яку також іноді називають формулою трапеції:

$$\int_a^b f(x)dx \approx \frac{h}{2}(f(x_0) + 2f(x_1) + \dots + 2f(x_{n-1}) + f(x_n)).$$

Теорема 10. *На класі двічі неперервно диференційовних функцій справедлива така оцінка похибки обчислення інтеграла:*

$$\left| \int_a^b f(x)dx - \frac{h}{2}(f(x_0) + 2f(x_1) + \dots + 2f(x_{n-1}) + f(x_n)) \right| \leq \frac{h^2(b-a)}{12} \cdot \max_{t \in [a,b]} |f''(t)|.$$

Ця оцінка є точною в тому сенсі, що множник $\frac{h^2(b-a)}{12}$ справа не може бути зменшеним.

Доведення.

$$\begin{aligned} |I(f) - I(p_n)| &= \left| \sum_{i=1}^n \left(\int_{x_{i-1}}^{x_i} f(x)dx - \frac{1}{2}(x_i - x_{i-1})(f(x_i) + f(x_{i-1})) \right) \right| \\ &\leq \sum_{i=1}^n \left| \int_{x_{i-1}}^{x_i} f(x)dx - \frac{1}{2}(x_i - x_{i-1})(f(x_i) + f(x_{i-1})) \right| \\ &\leq n \cdot \frac{h^3}{12} \max_{t \in [a,b]} |f''(t)| = \frac{(b-a)h^2}{12} \max_{t \in [a,b]} |f''(t)| \end{aligned}$$

і оцінку зверху доведено. Відмітимо, що ця оцінка стає рівністю для функції

$$f^*(t) = c \cdot (t - x_{i-1})(x_i - t), t \in [x_{i-1}, x_i], i = 1, \dots, n, c \in \mathbb{R}.$$

Ця функція (при $c \neq 0$) є неперервною, але не диференційовною. Але для кожного $\varepsilon > 0$ стандартними методами можна побудувати

двічі (або навіть нескінченно) диференційовну функцію f_ε^* , для якої $f_\varepsilon^*(x_i) = 0$, $i = 0, \dots, n$ і $\max_{t \in [a, b]} |f_\varepsilon^*(t) - f^*(t)| < \varepsilon$. З існування f_ε^* буде випливати точність наведеної у теоремі оцінки. $\square$

2.6 Тридіагональні системи лінійних рівнянь

Розв'язання системи лінійних рівнянь є однією з найважливіших задач прикладної математики взагалі, і чисельних методів зокрема. У цьому підрозділі ми зупинимось на одному важливому частковому випадку систем лінійних рівнянь.

Нехай задано набори чисел a_{ij} , $i, j = 1, \dots, n$ і f_i , $i = 1, \dots, n$. Систему рівнянь

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = f_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = f_2 \\ \dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = f_n \end{cases} \quad (2.8)$$

відносно невідомих $x_1, \dots, x_n$ можна також записати у матричному вигляді $Ax = f$, де

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}, \quad f = \begin{pmatrix} f_1 \\ f_2 \\ \dots \\ f_n \end{pmatrix}$$

Нас буде цікавити випадок, коли елементи матриці A є нулями окрім, можливо, елементів, що стоять на головній діагоналі і двох сусідніх

діагоналях матриці:

$$A = \begin{pmatrix} a_{11} & a_{12} & 0 & 0 & \dots & 0 & 0 \\ a_{21} & a_{22} & a_{21} & 0 & \dots & 0 & 0 \\ 0 & a_{32} & a_{33} & a_{34} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & a_{n-1n} & a_{nn} \end{pmatrix}.$$

Наприклад, матриця

$$A = \begin{pmatrix} 4 & 2 & 0 & 0 & 0 \\ 1 & 5 & 6 & 0 & 0 \\ 0 & 3 & 9 & 5 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 2 & 4 \end{pmatrix}$$

є тридіагональною.

Для таких матриць зручно ввести такі позначення елементів матриці: $l_i = a_{ii-1}$, $2 \leq i \leq n$, $d_i = a_{ii}$, $i = 1, \dots, n$, $u_i = a_{ii+1}$, $1 \leq i \leq n-1$.

Приклад 9. *Розв'язати систему рівнянь*

$$\begin{cases} 4x_1 + x_2 = 6 \\ x_1 + 4x_2 + x_3 = 12 \\ x_2 + 4x_3 + x_4 = 18 \\ x_3 + 4x_4 = 19 \end{cases}$$

Нагадаємо, що один зі способів розв'язання систем алгебраїчних рівнянь є метод Гаусса. Він полягає у приведенні матриці системи до верхньотрикутної матриці (прямий хід метода Гаусса), а потім до діагональної матриці (зворотній хід метода Гаусса) за допомогою елементарних операцій над рядками матриці (заміна місцями двох

рядків, множення одного рядка на ненульове число та заміна одного рядка на його суму з деяким іншим рядком).

$$\begin{aligned}
& \begin{pmatrix} 4 & 1 & 0 & 0 & | & 6 \\ 1 & 4 & 1 & 0 & | & 12 \\ 0 & 1 & 4 & 1 & | & 18 \\ 0 & 0 & 1 & 4 & | & 19 \end{pmatrix} \iff \begin{pmatrix} 4 & 1 & 0 & 0 & | & 6 \\ 0 & \frac{15}{4} & 1 & 0 & | & \frac{21}{2} \\ 0 & 1 & 4 & 1 & | & 18 \\ 0 & 0 & 1 & 4 & | & 19 \end{pmatrix} \\
& \iff \begin{pmatrix} 4 & 1 & 0 & 0 & | & 6 \\ 0 & \frac{15}{4} & 1 & 0 & | & \frac{21}{2} \\ 0 & 0 & \frac{56}{15} & 1 & | & \frac{76}{5} \\ 0 & 0 & 0 & \frac{209}{56} & | & \frac{209}{14} \end{pmatrix} \iff \begin{pmatrix} 4 & 1 & 0 & 0 & | & 6 \\ 0 & \frac{15}{4} & 1 & 0 & | & \frac{21}{2} \\ 0 & 0 & \frac{56}{15} & 1 & | & \frac{76}{5} \\ 0 & 0 & 0 & \frac{209}{56} & | & \frac{209}{14} \end{pmatrix} \\
& \iff \begin{pmatrix} 4 & 1 & 0 & 0 & | & 6 \\ 0 & \frac{15}{4} & 1 & 0 & | & \frac{21}{2} \\ 0 & 0 & \frac{56}{15} & 1 & | & \frac{76}{5} \\ 0 & 0 & 0 & 1 & | & 4 \end{pmatrix} \iff \begin{pmatrix} 4 & 1 & 0 & 0 & | & 6 \\ 0 & \frac{15}{4} & 1 & 0 & | & \frac{21}{2} \\ 0 & 0 & 1 & 0 & | & 3 \\ 0 & 0 & 0 & 1 & | & 4 \end{pmatrix} \\
& \iff \begin{pmatrix} 4 & 1 & 0 & 0 & | & 6 \\ 0 & 1 & 0 & 0 & | & 2 \\ 0 & 0 & 1 & 0 & | & 3 \\ 0 & 0 & 0 & 1 & | & 4 \end{pmatrix} \iff \begin{pmatrix} 1 & 0 & 0 & 0 & | & 1 \\ 0 & 1 & 0 & 0 & | & 2 \\ 0 & 0 & 1 & 0 & | & 3 \\ 0 & 0 & 0 & 1 & | & 4 \end{pmatrix}.
\end{aligned}$$

Таким чином, єдиним розв'язком системи рівняння є набір $(1, 2, 3, 4)$.

Застосуємо метод Гаусса до тридіагональної системи.

$$[A|f] = \left(\begin{array}{ccccc|c} d_1 & u_1 & 0 & \dots & 0 & f_1 \\ l_2 & d_2 & u_2 & \dots & 0 & f_2 \\ 0 & l_3 & d_3 & \dots & 0 & f_3 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & d_n & f_n \end{array} \right)$$

$$\Leftrightarrow \left(\begin{array}{ccccc|c} d_1 & u_1 & 0 & \dots & 0 & f_1 \\ 0 & d_2 - u_1 \cdot \frac{l_2}{d_1} & u_2 & \dots & 0 & f_2 - f_1 \cdot \frac{l_2}{d_1} \\ 0 & l_3 & d_3 & \dots & 0 & f_3 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & d_n & f_n \end{array} \right)$$

...

$$\Leftrightarrow \left(\begin{array}{ccccc|c} \delta_1 & u_1 & 0 & \dots & 0 & g_1 \\ 0 & \delta_2 & u_2 & \dots & 0 & g_2 \\ 0 & 0 & \delta_3 & \dots & 0 & g_3 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \delta_n & g_n \end{array} \right),$$

де $\delta_1 = d_1$,

$$\delta_k = d_k - u_{k-1} \cdot \frac{l_k}{\delta_{k-1}}, 2 \leq k \leq n,$$

$g_1 = f_1$,

$$g_k = f_k - g_{k-1} \cdot \frac{l_k}{\delta_{k-1}}, 2 \leq k \leq n.$$

Відмітимо, що прямий хід метода Гаусса є можливим тоді і тільки тоді, коли

$$\delta_k \neq 0, k = 1, \dots, n. \quad (2.9)$$

Зворотній хід метода Гаусса тепер також можливий при тій самій умові (2.9). При цьому, складність розв'язання тридіагональної системи рівнянь є лінійною (на відміну від складності розв'язання системи рівнянь у загальному випадку).

Ми будемо казати, що тридіагональна матриця A має *діагональне домінування*, якщо

$$|d_i| > |l_i| + |u_i|, i = 1, \dots, n, \quad (2.10)$$

де для зручності покладемо $l_1 = u_n = 0$.

Теорема 11. *Якщо тридіагональна матриця A має діагональне домінування, то система (2.8) має єдиний розв'язок, який можна знайти за допомогою метода Гаусса описаного вище.*

Доведення. Зазначимо, що $d_1 = \delta_1$, і за умовою діагонального домінування $|d_1| > |u_1| \geq 0$.

$$|\delta_2| = \left| d_2 - u_1 \cdot \frac{l_2}{d_1} \right| \geq |d_2| - \left| u_1 \cdot \frac{l_2}{d_1} \right| > |l_2| + |u_2| - |l_2| = |u_2| \geq 0,$$

де ми використали нерівність $\frac{|u_1|}{|d_1|} < 1$. Повторюючи аналогічні міркування отримаємо справедливість умови (2.9), а отже і твердження теореми. $\square$

2.7 Розв'язання крайової задачі

Для ілюстрації ситуації, в якій при наближеному розв'язанні деякої задачі виникає необхідність розв'язувати тридіагональну систему рівнянь, розглянемо крайову задачу

$$-u'' + u = f(x), x \in [0, 1], \quad (2.11)$$

$$u(0) = u(1) = 0. \quad (2.12)$$

Відмітимо, що однорідність крайових умов не зменшує загальність, бо заміною $v = u(x) + Ax + B$ з відповідним чином підібраними коефіцієнтами A і B неоднорідні крайові умови можна звести до однорідних.

Розглянемо рівномірне розбиття $0 = x_0 < x_1 < \dots < x_n = 1$ відрізка $[0, 1]$, $x_i - x_{i-1} = h = \frac{1}{n}$, $i = 1, \dots, n$. Аналогічно до того, як було отримано різницеве наближення похідної у підрозділі 2.2, неважко отримати, що для деякого $\xi \in (x, x + h)$ справедлива рівність

$$u''(x) - \frac{u(x - h) - 2u(x) + u(x + h))}{h^2} = \frac{1}{12}h^2 u^{(4)}(\xi). \quad (2.13)$$

Підставляючи цей вираз у рівняння (2.11) і нехтуючи виразом у правій частині (2.13), отримаємо

$$\frac{-u(x-h) + 2u(x) - u(x+h)}{h^2} + u(x) \approx f(x), x \in [0, 1]$$

Позначимо $u_k = u(x_k)$, $f_k = f(x_k)$, $k = 0 \dots, n$ і розглянемо систему рівнянь

$$\frac{-u_{k-1} + 2u_k - u_{k+1}}{h^2} + u_k = f_k, 1 \leq k \leq n-1,$$

відносно наближених значень u_k невідомої функції u , $k = 1, \dots, n-1$; за умовою (2.12), маємо $u_0 = u_n = 0$. Перепишучи цю систему у вигляді

$$-u_{k-1} + (2 + h^2)u_k - u_{k+1} = h^2 f_k, 1 \leq k \leq n-1,$$

приходимо до тридіагональної системи рівнянь з матрицею

$$\begin{pmatrix} 2 + h^2 & -1 & 0 & \dots & 0 \\ -1 & 2 + h^2 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 2 + h^2 \end{pmatrix},$$

в якій $l_i = -1$, $i = 2, \dots, n$, $u_1 = -1$, $i = 1, \dots, n-1$, і $d_i = 2 + h^2$, $i = 1, \dots, n$. Бачимо, що виконується умова (2.10) діагонального домінування, і отже цю систему рівнянь можна за лінійний час розв'язати за допомогою метода Гаусса.

На оцінці похибки наближеного розв'язання ми тут не зупиняємось.

Бібліографія

- [1] J.F. Epperson. *An Introduction to Numerical Methods and Analysis, Solutions Manual*. John Wiley & Sons, 2014.
- [2] А. Я. Дороговцев. *Математичний аналіз: Підручник. У 2 ч.* Київ: Либідь, 1993.
- [3] М. В. Заболоцький, О. Г. Сторож, С. І. Тарасюк. *Математичний аналіз: Підручник*. Київ: Знання, 2008.
- [4] І. І. Ляшко, В. Ф. Ємельянов, О. К. Боярчук. *Математичний аналіз*. Київ: Вища школа, 1993.