Київський національний університет імені Тараса Шевченка

Методи обчислень

Навчальний посібник
для студентів фізичного факультету

Київ
Виробнично-поліграфічний центр
"Київський університет"
2001

Єжов С.М. Методи обчислень: Навчальний посібник.
К.: ВПЦ "Київський університет", 2001, - 140 с.

Рецензент Шевчук І.О., д-р. фіз.-мат. наук, професор

Затверджено Радою
фізичного факультету
9 жовтня 2000 року

Навчальне видання

Навчальний посібник
з курсу "Методи обчислень"

для студентів фізичного факультету

Автор ЄЖОВ Станіслав Миколайович

Редактор Т. Мельник

Глава 1

Оцінка похибки чисельного розв'язку

При вивченні багатьох процесів, що характеризуються набором визначених параметрів, можна побудувати фізичну чи якусь іншу модель, основану на відомих чи запропонованих законах. Математичний опис такої моделі приводить до деяких кількісних співвідношень між параметрами задачі. Ці співвідношення можна представити у вигляді

$$\vec{u} = A\vec{v}. \tag{1.1}$$

Тут $\vec{v}$ - сукупність вхідних параметрів, знання яких дозволяє за відомими правилами, що визначаються оператором A, знаходити сукупність шуканих параметрів $\vec{u}$. Оператор A визначає послідовність дій над параметрами $\vec{v}$ і є математичним виразом відповідного закону чи моделі.

Математичний опис явища не завжди адекватний до дійсної моделі. Тому отриманий розв'язок може мати деяку похибку моделювання, яка тут обговорюватись не буде. Нехай рівняння (1.1) достатньо точно описує реальний процес. Похибка результату чисельного розв'язку обумовлена трьома основними причинами.

1.1. Похибка методу

Метод, що використовується для розв'язку рівняння (1.1), як правило, є наближеним. Це дозволяє іноді значно скоротити час обчислень але за рахунок точності розв'язку. Такі зміни ефективної математичної моделі можуть приводити до того, що результат розв'язку рівняння

$$\vec{u}_1 = A_1\vec{v}. \tag{1.2}$$

буде значно відрізнятися від шуканого розв'язку $\vec{u}$ рівняння (1.1). У рівнянні (1.2) оператор A_1 є сукупністю *скінченної* кількості математичних (у більшості випадків арифметичних) дій, що наближено описують початковий оператор A. Тому бажано завжди мати такий оператор A_1, який дозволяє обчислити $\vec{u}_1$ в достатньо близькому околі $\vec{u}$. Знаходження таких операторів і оцінка похибки, що вводиться при цьому, є основним предметом дослідження теорії чисельних методів.

1.2. Заокруглення при обчисленні

При використанні комп'ютера на всіх етапах обчислень проводяться заокруглення, які обумовлені скінченністю розрядної сітки комп'ютера. Ця проблема має цифрове походження. Вона випливає із скінченної точності представлення чисел в пам'яті ЕОМ. Існує певна, іноді дуже значна, різниця між дійсними математичними і комп'ютерними числами з плаваючою комою. Це можна проілюструвати за допомогою такого прикладу. При обчисленні на комп'ютері розбіжний гармонічний ряд

$$S_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$$

збігається при досягненні n деякого значення, яке залежить від комп'ютера, операційної системи, мови програмування тощо. Після цього значення n усі доданки $\dfrac{1}{n+k}$ будуть давати нульовий внесок у суму внаслідок відкидання молодших розрядів.

Як другий приклад впливу скінченної точності представлення чисел можна привести результат множення та додавання трьох чисел в гіпотетичному комп'ютері з трьома десятинними розрядами:

$$\begin{aligned} (0.364 + 0.423) * 0.125 &= 0.098 \neq \\ \neq 0.364 * 0.125 + 0.423 * 0.125 &= 0.099, \end{aligned}$$

тобто в комп'ютерній арифметиці не виконуються асоціативний та дистрибутивний, а при обчисленні в деяких комп'ютерах і комутативний закони.

Тому, при використанні конкретних чисельних методів завжди потрібний аналіз результатів проміжних операцій, який є наближеним внаслідок великих

труднощів оцінки помилок, що з'являються за рахунок особливого представлення алгебраїчних операцій в комп'ютері. Повний аналіз задачі ускладнюється тим, що помилки, яких припускаються за час розрахунку, взаємозв'язані.

Вплив різниці дійсної і комп'ютерної арифметик у значній мірі зменшується при використанні чисел із подвійною точністю. Існують комп'ютери, які допускають представлення чисел із більш високою точністю, але повністю розбіжність усунути не вдається.

Якщо результати роботи програми, що виконана з подвійною точністю, не співпадають з результатами, що отримані із звичайною точністю, то в цьому винні помилки заокруглення. Слід , однак, пам'ятати, що така перевірка потребує значного збільшення оперативної пам'яті та часу обчислень.

Наявність помилок заокруглення накладає вимоги на обмеження кількості операцій, що виконуються в обраному алгоритмі. Так, наприклад, розв'язок системи лінійних алгебраїчних рівнянь

$$A\vec{x} = \vec{b} \tag{1.3}$$

можна шукати або за правилом Крамера, що потребує $\sim n^2 \cdot n!$ операцій, де n - порядок системи (1.3), або методом послідовних виключень Гаусса, в якому необхідно виконати $\sim n^3/3$ операцій. Для наочності приведемо кількість операцій при деяких значеннях n.

Метод	n=3	5	10	20
Крамера	51	2885	$3.6 \cdot 10^8$	$5 \cdot 10^{20}$
Гаусса	17	65	430	4770

Використання методу Гаусса не тільки значно скорочує час обчислень, але також приводить до меншого впливу помилок заокруглення на результати розв'язку.

Аналогічно можна розглянути два різних рекурентних співвідношення для обчислення числа π, які в точній арифметиці сбігаються:

$$\pi = \lim_{n\to\infty}(n \cdot a_{2n}) \quad ,$$

$$a_{2n} = \sqrt{2} \cdot \sqrt{1 - \sqrt{1 - \left(\frac{a_n}{2}\right)^2}}, \quad a_6 = 1\,; \tag{1.4}$$

$$a_{2n} = \frac{a_n}{\sqrt{2} \cdot \sqrt{1 + \sqrt{1 - \left(\frac{a_n}{2}\right)^2}}}, \quad a_6 = 1\,. \tag{1.5}$$

Перше з них при деякому $n = N$, яке визначається довжиною розрядної сітки, дає граничний нульовий результат, у той час як (1.5) приводить до наближено правильного результату.

Використання комп'ютерної арифметики замість дійсної можна умовно позначати за допомогою введення деякого оператора A_2 замість A_1 у рівнянні (1.2).

1.3. Неусувна похибка

Вихідні дані задачі $\vec{v}$ з різних причин можуть бути задані неточно. Як наслідок отримаємо такий вираз для розв'язку

$$\vec{u}_3 = A_2\vec{v}_1\,. \tag{1.6}$$

Різниця $\vec{v} - \vec{v}_1$ є неусувною похибкою вихідних даних, яка в результаті виконання деякої кількості дій в операторі A (чи A_1, A_2 тощо) може призвести до неусувної похибки результату. Як приклад розглянемо обчислення членів ряду за рекурентним співвідношенням: $x_{n+1} = 10x_n - 9x_{n-1}$. Якщо початкові значення x_0 та x_1 задані з похибкою α, то похибка членів ряду зростає з великою швидкістю: 19α, 199α, 2061α і т.д., і обчисленні значення дуже швидко стануть далекими від істинних.

Таким чином, при оцінюванні точності обчисленого на комп'ютері результату необхідно оцінювати похибки, які вносяться усіма переліченими причинами. Часто це приводить до такого аналізу похибок, складність якого значно перевищує складність початкової задачі. В деяких наведених нижче чисельних методах іноді вказуються ці оцінки та приклади.

1.4. Розподіл похибок вимірів

Нехай ми вимірюємо деяку невідому величину, точне значення якої дорівнює x_0, і в результаті вимірів отримуємо послідовність значень $x_1, x_2, \ldots x_n$. Спираючись на ці значення оцінимо величину x_0 . Для найбільш ймовірного її значення введемо позначення $\bar{x}$. Оскільки воно залежить від результатів вимірювання, відповідно до (1.1) запишемо таку рівність

$$\bar{x} = f\left(x_1, x_2, \ldots, x_n\right)\,.$$

Тут f - шукана функція, явний вигляд якої ми і хочемо визначити. Встановимо деякі властивості цієї функції.

Зрозуміло, що результат аналізу не повинен залежати від обраної шкали вимірів усіх x_i. Це приводить до двох рівностей:

$$f\left(cx_1, cx_2, \ldots, cx_n\right) = c \cdot f\left(x_1, x_2, \ldots, x_n\right) \tag{1.7}$$

- *незалежність від одиниць вимірів*;

$$f\left(x_1 + c, x_2 + c, \ldots, x_n + c\right) = f\left(x_1, x_2, \ldots, x_n\right) + c \tag{1.8}$$

- *незалежність від початкової точка відліку.*

Якщо всі виміри проводились на одній експериментальній установці при однакових зовнішніх умовах, то результати вимірів однаково правдоподібні і, як наслідок, функція $f\left(\{x_i\}\right) = f\left(x_1, \ldots, x_n\right)$ повинна бути симетричною відносно перестановок своїх аргументів. Окрім того, припустимо, що $f(x_1, \ldots, x_n)$ - двічі диференційована по всіх своїх аргументах.

Введемо позначення $q_k = \left.\frac{\partial f}{\partial x_k}\right|_{\{x_i\}=0}$ і розкладемо $f(\{x_i\})$ в ряд Тейлора з точністю до членів першого порядку включно

$$\begin{aligned} &f(\{x_i\}) = f\left(x_1, \ldots, x_n\right) = \\ &= f\left(0, \ldots, 0\right) + q_1 x_1 + \ldots + q_n x_n + O\left(\{x_i^2\}\right) . \end{aligned} \tag{1.9}$$

Зрозуміло, що перший доданок у правій частині (1.9) дорівнює нулеві. Тоді використаємо (1.7) і спрямуємо c до нуля. З точністю до членів першого порядку по c маємо:

$$cf(\{x_i\}) = f(\{cx_i\}) = cq_1 x_1 + \ldots + cq_n x_n + c^2 O(1)\,,$$

тобто $f(x_1, \ldots, x_n) = q_1 x_1 + \ldots + q_n x_n$. Використаємо (1.8):

$$\begin{aligned} &f\left(x_1 + c, \ldots, x_n + c\right) = f\left(x_1, \ldots, x_n\right) + c = \\ &= q_1(x_1 + c) + \ldots + q_n(x_n + c) = \\ &= q_1 x_1 + \ldots + q_n x_n + \left(q_1 + \ldots + q_n\right) \cdot c\,. \end{aligned}$$

У результаті $q_1 + q_2 + \ldots + q_n = 1$. Враховуючи симетрію $f(\{x_i\})$ відносно своїх аргументів, маємо $q_k = \frac{1}{n}$.

Таким чином, *найбільш ймовірним значенням* вимірів величини x_0 буде *середнє арифметичне*

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}\,, \tag{1.10}$$

яке мінімізує *середню квадратичну похибку*

$$m = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_0 - x_i)^2}\,. \tag{1.11}$$

Якщо виміри проводилися на різних експериментальних установках, або змінювалися зовнішні умови, то замість (1.10) використовується *середнє зважене значення*

$$\bar{x} = \sum_{i=1}^{n} q_i x_i\ , \quad \left(\sum_{i=1}^{n} q_i = 1\ , q_i \geq 0\right)\ . \tag{1.12}$$

яке мінімізує *середню зважену похибку*

$$m = \sqrt{\sum_{i=1}^{n} q_i (x_0 - x_i)^2}\,. \tag{1.13}$$

Нехай α - *абсолютна похибка* вимірів величини x_0, тобто вимірювані величини $x_1, \ldots, x_n$ знаходяться в інтервалі $[x_0-\alpha, x_0+\alpha]$. Припустимо, що мінімальна ціна шкали вимірювального приладу дорівнює η і вона кратну кількість разів міститься в α, а також те, що кожний вимір може привести до будь-якого доступного значення з однаковою ймовірністю. Тоді *істинна похибка* числа x_0 буде приймати одне з рівноможливих значень (q - ціле число, $\eta = \alpha/q$):

$$-q\eta, -(q-1)\eta, \ldots, -\eta, 0, \eta, 2\eta, \ldots, (q-1)\eta, q\eta\,.$$

Середня квадратична похибка цього ряду значень дорівнює

$$m = \sqrt{\frac{1\cdot 0^2 + 2\cdot\eta^2 + \ldots + 2q^2\eta^2}{2q+1}} = \sqrt{\frac{2\eta^2\left(1^2 + \ldots + q^2\right)}{2q+1}} = \\ = \sqrt{\frac{2\eta^2 q(q+1)(2q+1)}{6(2q+1)}} = \alpha\sqrt{\frac{q+1}{3q}}\,,$$

і при $q \to \infty$ маємо $\alpha = m\sqrt{3}$.

Таким чином, якщо розподіл величин x_i відносно значення x_0 невідомий, то дисперсію цього розподілу m^2 можна зв'язати з абсолютною похибкою вимірів α. Внаслідок центральної граничної теореми теорії ймовірності ми можемо

стверджувати наступне: якщо всі виміри x_i незалежні або слабозалежні, то при достатньо великій кількості вимірів середнє арифметичне буде розподілене за *нормальним* (або *гауссовим*) законом:

$$p(\bar{x}) = \frac{1}{\sqrt{2\pi}\alpha\sqrt{\frac{1}{3n}}} \exp\left[-\frac{(\bar{x}-x_0)^2}{2\frac{\alpha^2}{3n}}\right]. \tag{1.14}$$

В теорії ймовірностей показується, що при дуже великій кількості вимірів x_i в кожній тисячі випадкових похибок приблизно 680 лежить між $-m$ та m; 950 похибок лежить між $-2m$ та $2m$ і 997 похибок між $-3m$ та $3m$. Тобто в кожній тисячі тільки три виміри мають похибки більші за $3m$.

Число $3m$ звичайно обирається за границю випадкових похибок (так звана *гранична* чи *максимальна* похибка):

$$\sigma = 3m = \alpha\sqrt{3}\,. \tag{1.15}$$

Якщо на практиці при вимірюванні значення x_i перевищує σ, воно не враховується при обробці результатів експерименту.

Розглянемо важливий окремий випадок розрахунку похибки суми багатьох доданків. Якщо помилки вимірів кожного доданку носять випадковий характер і не перевищують α, тоді при $z = x_1 + x_2 + \ldots + x_n$ неважко знайти, що

$$\sigma_z = 3m_z = \sqrt{m_{x_1}^2 + \ldots + m_{x_n}^2} = \alpha\sqrt{3n}\,. \tag{1.16}$$

Доведення (1.16) рекомендується провести самостійно.

Глава 2

Інтерполяція

Досліднику часто доводиться мати справу з даними, поданими у вигляді таблиць. Це може бути пов'язане із скінченним числом експериментальних даних або з тим, що об'єм таблиць повинен бути обмеженим і в таблицях можна привести лише частину даних. Як приклад згадаємо психрометричні таблиці або відомі таблиці Брадіса. Задача інтерполяції полягає в тому, щоб відшукати значення функції в деякій проміжній точці. Найпростішим випадком інтерполяції є лінійна інтерполяція, коли крива на площині між точками з координатами (x_1, y_1) та (x_2, y_2) апроксимується прямою, що проходить через ці точки (рис. 2.1). Рівняння прямої має вигляд

$$y = \frac{x - x_2}{x_1 - x_2} y_1 + \frac{x - x_1}{x_2 - x_1} y_2 \,. \tag{2.1}$$

У результаті на інтервалі $[x_1, x_2]$ (а в загальному випадку і поза цим інтервалом - *екстраполяція*) ми можемо за формулою (2.1) знайти значення y при

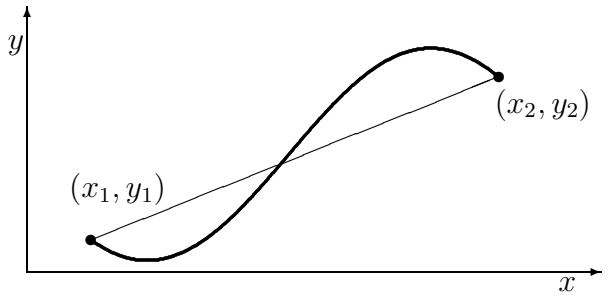

Рис. 2.1. Лінійна інтерполяція

довільному значенні x. Якщо використати не дві, а більшу кількість експериментальних точок, то однозначно можна побудувати інтерполяційний поліном більш високого порядку. Для кривих другого порядку така інтерполяція носить назву *параболічної*.

2.1. Інтерполяційний поліном Лагранжа

Оскільки у загальному випадку через n точок на площині можна провести криву, яка описується поліномом порядку $n-1$, ми будемо вимагати, щоб шуканий поліном набував задані значення у всіх n точках таблиці, тобто

$$P_{n-1}(x_i) = y_i, \quad i = \overline{1,n}\,. \tag{2.2}$$

Поліном $P_{n-1}(x)$ будемо шукати у вигляді

$$P_{n-1}(x) = a_1(x)y_1 + \ldots + a_n(x)y_n\,,$$

де $a_i(x)$ - поліноми степеня $n-1$. Тоді умова (2.2) запишеться у вигляді системи рівнянь:

$$\begin{array}{ccccccc} a_1(x_1)y_1 & + & \ldots & + & a_n(x_1)y_n & = & y_1\,, \\ a_1(x_2)y_1 & + & \ldots & + & a_n(x_2)y_n & = & y_2\,, \\ \vdots & & \ddots & & \vdots & & \vdots \\ a_1(x_n)y_1 & + & \ldots & + & a_n(x_n)y_n & = & y_n\,. \end{array} \tag{2.3}$$

Рівняння (2.3) задовільнятимуться автоматично, якщо обрати

$$a_i(x_j) = \delta_{ij}\,, \tag{2.4}$$

де δ_{ij} - символ Кронекера. Це означає, що довільний поліном $a_i(x)$ дорівнює нулеві при кожному x_j, окрім випадку $i = j$, тобто можна записати

$$a_i(x) = C_i(x - x_1)\cdots(x - x_{i-1})(x - x_{i+1})\cdots(x - x_n)\,.$$

Коефіцієнти C_i легко знаходяться із умови (2.4):

$$C_i = [(x_i - x_1)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)]^{-1}\,.$$

У результаті маємо

$$P_{n-1}(x) = \sum_{i=1}^{n} y_i \left(\prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)} \right)\,. \tag{2.5}$$

Це і є шуканий *інтерполяційний поліном Лагранжа*. Неважко бачити, що він задовольняє умову (2.2).

Для безпосереднього використання полінома (2.5), а також для прикладних цілей, зручно переписати його в іншому вигляді. Визначимо поліном степеня n

$$\Pi(x) = \prod_{k=1}^{n}(x - x_k)\,. \tag{2.6}$$

Продиференціюємо його: $\frac{d\Pi}{dx} = \Pi'(x) = \Pi(x)\sum_{i=1}^{n}\frac{1}{(x-x_i)}$. Якщо тепер спрямувати значення x до значення x_j, то

$$\Pi'(x_j) = \prod_{\substack{k=1\\k\neq j}}^{n}(x_j - x_k)\,. \tag{2.7}$$

Використовуючи (2.5), (2.6), (2.7), можна записати

$$P_{n-1}(x) = \Pi(x)\sum_{i=1}^{n}\frac{y_i}{(x - x_i)\Pi'(x_i)}\,. \tag{2.8}$$

Задача 1. *Знайти значення інтерполяційного полінома Лагранжа в точці* $x = 1.5$ *для функції, що задана у вигляді таблиці:*

x	*0.*	*1.*	*2.*	*3.*
y	*0.0000*	*0.1736*	*0.3420*	*0.5000*

2.2. Залишковий член полінома Лагранжа

Для оцінки абсолютної похибки полінома Лагранжа (тобто його залишкового члена) припустимо, що функція $y = f(x)$, яка задана у вигляді таблиці, має похідну порядку n. Покладемо

$$\varphi(z) = f(z) - P_{n-1}(z) - A\cdot\Pi(z)\,. \tag{2.9}$$

Оберемо коефіцієнт A з умови, що $\varphi(z)$ дорівнює нулеві в $n+1$ точках $x_1, \ldots, x_n, x$. На підставі теореми Ролля її похідна $\varphi'(z)$ дорівнює нулеві щонайменше в n точках. Застосовуючи теорему Ролля до $\varphi'(z)$, дістанемо, що похідна $\varphi''(z)$ дорівнює нулеві щонайменше в $n-1$ точках; продовжуючи ці міркування далі, дістанемо, що n-а похідна $\varphi^{(n)}(z)$ дорівнює нулеві щонайменше в одній точці ξ,

причому ця точка належить відкритому інтервалу, обмеженому максимальним та мінімальним значеннями із сукупності $x_1, \ldots, x_n, x$. На підставі (2.9) маємо $\varphi^{(n)}(z) = f^{(n)}(z) - A \cdot n!$, і, оскільки $\varphi^{(n)}(\xi) = 0$, то $A = \frac{f^{(n)}(\xi)}{n!}$.

Для оцінки залишкового в точці x покладемо $\varphi(x) = 0$, тоді

$$f(x) = P_{n-1}(x) + \frac{f^{(n)}(\xi)}{n!}\Pi(x)\,.$$

Як наслідок, залишковий член або абсолютна похибка алгебраїчного інтерполяційного полінома фіксованого степеня

$$R_n = \frac{1}{n!}\max_{\xi\in[a,b]}|f^{(n)}(\xi)| \cdot \max_{x\in[a,b]}|\Pi(x)| \tag{2.10}$$

буде визначатися не тільки аналітичним виглядом функції, що інтерполюється, і який, як правило, невідомий, але й поведінкою полінома $\Pi(x)$ степеня n, визначеного в (2.6).

Поведінка цього полінома із старшим коефіцієнтом (біля x^n), що дорівнює одиниці, зручно розглядати на відрізку $[-1, 1]$, до якого завжди можна перейти від довільного інтервала $[a, b]$ за допомогою зміни масштабу:

$$t = \frac{1}{(b-a)}\left[2x - (b+a)\right]\,. \tag{2.11}$$

В подальшому будемо писати замість аргументу не t, а x, маючи на увазі, що x масштабований на відрізку $[-1, 1]$ лінійним перетворенням вигляду (2.11).

Виникає питання: як на інтервалі $[-1, 1]$ розташувати вузли (корені полінома) $\{x_i\}$, щоб $\Pi(x)$ із старшим коефіцієнтом 1 найменше відрізнявся від 0? Таку властивість мають *поліноми Чебишева*

$$\overline{T}_n(x) = \frac{1}{2^{n-1}}\cos(n\arccos x)\,. \tag{2.12}$$

За допомогою формули Муавра покажемо, що це дійсно поліном степені n із старшим коефіцієнтом, що дорівнює 1. Якщо позначити $\theta = \arccos x$, то

$$\begin{aligned}
\overline{T}_n(x) &= \frac{1}{2^{n-1}}\cos(n\theta) = \frac{1}{2^n}\left[e^{in\theta} + e^{-in\theta}\right] = \\
&= \frac{1}{2^n}\left[(\cos n\theta + i\sin n\theta) + (\cos n\theta - i\sin n\theta)\right] = \\
&= \frac{1}{2^n}\left[(\cos\theta + i\sin\theta)^n + (\cos\theta - i\sin\theta)^n\right] = \\
&= \frac{1}{2^n}\left[\left(1 + \sqrt{1-\tfrac{1}{x^2}}\right)^n + \left(1 - \sqrt{1-\tfrac{1}{x^2}}\right)^n\right]\,.
\end{aligned}$$

Видно, що всі непарні степені радикалів взаємоскорочуються. Старший коефіцієнт a_n визначимо таким чином:

$$a_n = \lim_{x\to\infty} \frac{\overline{T}_n(x)}{x^n} = \frac{1}{2^n} \lim_{x\to\infty} \frac{(x+\sqrt{x^2-1})^n+(x-\sqrt{x^2-1})^n}{x^n} =$$
$$= \frac{1}{2^n} \lim_{x\to\infty} \left[\left(1+\sqrt{1-\frac{1}{x^2}}\right)^n + \left(1-\sqrt{1-\frac{1}{x^2}}\right)^n \right] = 1\,.$$

Корені цього полінома x_m, які можуть визначати положення нулів інтерполяційного полінома, знаходяться з рівняння

$$\overline{T}_n(x_m) = \frac{1}{2^{n-1}} \cos(n \arccos x_m) = 0\,,$$

звідки $x_m = \cos\left(\frac{\pi}{2}\frac{2m-1}{n}\right)$, $m = \overline{1,n}$.

Точки максимального відхилення $\tilde{x}_m$ знаходимо з умови екстремуму $\overline{T}_n(x)$, тобто

$$\begin{aligned} \overline{T}'_n(\tilde{x}_m) &= \frac{1}{2^{n-1}} \sin(n \arccos \tilde{x}_m) = 0\,, \\ \tilde{x}_m &= \cos\left(\tfrac{\pi m}{n}\right), \quad m = \overline{0,n}\,. \end{aligned}$$

У цих точках

$$\overline{T}_n(\tilde{x}_m) = \frac{1}{2^{n-1}} \cdot (-1)^m\,. \tag{2.13}$$

Доведення того, що поліном (2.12) дійсно найменше ухиляється від нуля на відрізку $[-1,1]$, проведемо методом від протилежного.

Припустимо, що існує інший поліном $S_n(x)$ із старшим коефіцієнтом, рівним 1 при x^n, максимальне по модулю відхилення якого на інтервалі $[-1,1]$ менше, ніж модуль $\overline{T}_n(\tilde{x}_m)$ з виразу (2.13). Тоді поліном $\overline{T}_n(x) - S_n(x)$ має степінь $n-1$, а його знак в точках $\tilde{x}_m$ визначається виразом

$$\operatorname{sign}\left((-1)^m \frac{1}{2^{n-1}} - S_n(\tilde{x}_m)\right) = (-1)^m\,.$$

Таким чином, між кожними двома точками $\tilde{x}_m$ та $\tilde{x}_{m+1}$, поліном $\overline{T}_n(x) - S_n(x)$ змінює знак, і на інтервалі $[-1,1]$ це відбувається n разів, що відповідає наявності у полінома $(n-1)$-ої степені n різних коренів. Знайдене протиріччя і доводить наше твердження.

Результати, що отримані у цьому розділі , дозволяють правильно визначати на інтервалі $[a,b]$ значення точок x_i, в яких потрібно проводити виміри шуканої величини, щоб при наступній інтерполяції похибка була мінімальною. Положення цих точок визначає перетворення, обернене до (2.11):

$$x = \frac{b+a}{2} + \frac{b-a}{2}\cos\left(\frac{\pi}{2}\frac{(2i-1)}{n}\right), \quad i = \overline{1,n}\,.$$

Абсолютна похибка при цьому буде визначатися виразом

$$R_n = \max|f^{(n)}| \cdot \frac{(b-a)^n}{n!2^{2n-1}}\,. \tag{2.14}$$

2.3. Скінченні та розділені різниці

Нехай вузли таблиці x_i розташовані на рівних відстанях: $x_i = x_1 + (i-1)h$; величина h називається кроком таблиці. Для функції, що задана у вигляді таблиці $y_i = f(x_i)$, можна ввести операції, що є дискретним аналогом операцій диференціювання та інтегрування.

Аналогом першої похідної є *різниця першого порядку*

$$\begin{array}{lcll} \Delta y_i & = & y_{i+1} - y_i & - \text{ права,} \\ \bigtriangledown y_i & = & y_i - y_{i-1} & - \text{ ліва,} \\ \delta_i & = & \frac{1}{2}\left(\Delta y_i + \bigtriangledown y_i\right) = \frac{1}{2}\left(y_{i+1} - y_{i-1}\right) & - \text{ центральна.} \end{array}$$

При цьому легко бачити, що $\Delta y_i = \bigtriangledown y_{i+1}$.

Різниці вищого порядку утворюються за допомогою рекурентних співвідношень

$$\begin{array}{lcll} \Delta^m y_i & = & \Delta\left(\Delta^{m-1} y_i\right) & = \Delta^{m-1} y_{i+1} - \Delta^{m-1} y_i\,, \\ \bigtriangledown^m y_i & = & \bigtriangledown\left(\bigtriangledown^{m-1} y_i\right) & = \bigtriangledown^{m-1} y_i - \bigtriangledown^{m-1} y_{i-1}\,, \\ \delta^m y_i & = & \delta\left(\delta^{m-1} y_i\right) & = \frac{1}{2}\left(\delta^{m-1} y_{i+1} - \delta^{m-1} y_{i-1}\right)\,. \end{array}$$

Зокрема, *різниці другого порядку* мають вигляд

$$\begin{array}{lcll} \Delta^2 y_i & = & \Delta\left(\Delta y_i\right) & = y_{i+2} - 2y_{i+1} + y_i\,, \\ \bigtriangledown^2 y_i & = & \bigtriangledown\left(\bigtriangledown y_i\right) & = y_i - 2y_{i-1} + y_{i-2}\,, \\ \delta^2 y_i & = & \delta\left(\delta y_i\right) & = \frac{1}{4}\left(y_{i+2} - 2y_i + y_{i-2}\right)\,. \end{array}$$

Важливу роль в подальшому буде відігравати вираз

$$\Delta \bigtriangledown y_i = \bigtriangledown \Delta y_i = y_{i+1} - 2y_i + y_{i-1}\,.$$

Різницевим аналогом формули диференціювання добутку є формули

$$\begin{aligned}\Delta(u_iv_i) &= u_i\Delta v_i + v_{i+1}\Delta u_i &= u_{i+1}\Delta v_i + v_i\Delta u_i\,,\\ \bigtriangledown(u_iv_i) &= u_{i-1}\bigtriangledown v_i + v_i\bigtriangledown u_i &= u_i\bigtriangledown v_i + v_{i-1}\bigtriangledown u_i\,.\end{aligned}$$

які перевіряються безпосередньо.

Аналогом формул інтегрування по частинах є формули додавання по частинах

$$\begin{aligned}\sum_{i=1}^{N-1} u_i\Delta v_i &= -\sum_{i=2}^{N} v_i\bigtriangledown u_i + (uv)_N - (uv)_1\,,\\ \sum_{i=1}^{N-1} u_i\Delta v_i &= -\sum_{i=2}^{N-1} v_i\bigtriangledown u_i + u_{N-1}v_N - (uv)_1\,.\end{aligned}$$

Якщо вузли не є рівновіддаленими, то аналогом похідної будуть *розділені різниці*. Розділені різниці *нульового порядку* $f(x_i)$ збігаються із значеннями функції $f(x_i)$. Різниці *першого порядку* визначаються рівністю

$$f(x_i;x_j) = \frac{f(x_j)-f(x_i)}{x_j-x_i}\,,$$

різниці *другуго порядку* - рівністю

$$f(x_i;x_j;x_k) = \frac{f(x_j;x_k)-f(x_i;x_j)}{x_k-x_i}\,,$$

і т.д.; різниці *k-го порядку* визначаються рівністю

$$f(x_1;\ldots;x_{k+1}) = \frac{f(x_2;\ldots;x_{k+1})-f(x_1;\ldots;x_k)}{x_{k+1}-x_1}\,.$$

У випадку рівновіддалених вузлів скінченні та розділені різниці зв'язані співвідношенням

$$f(x_1;\ldots;x_{k+1}) = \frac{\Delta^k y_1}{k!h^k} = \frac{\bigtriangledown^k y_{k+1}}{k!h^k}\,. \tag{2.15}$$

2.4. Інтерполяційний поліном Ньютона

Будемо шукати поліном $P_{n-1}(x)$ степеня $n-1$, що проходить через n точок $x_1,\ldots,x_n$, у вигляді

$$\begin{aligned}P_{n-1}(x) = a_1 + a_2(x-x_1) + a_3(x-x_1)(x-x_2) + \ldots +\\ +a_n(x-x_1)\cdots(x-x_{n-1})\,.\end{aligned} \tag{2.16}$$

Коефіцієнти a_i визначаються рівністю (2.2). Тоді $P_{n-1}(x_1) = a_1 = y_1$. Далі, $P_{n-1}(x_2) = a_1 + a_2(x_2 - x_1) = y_2$; звідки

$$a_2 = \frac{y_2 - y_1}{x_2 - x_1} = f(x_1; x_2)\,,$$

тобто a_2 - розділена різниця першого порядку. Аналогічно

$$P_{n-1}(x_3) = a_1 + a_2(x_3 - x_1) + a_3(x_3 - x_1)(x_3 - x_2) = y_3\,,$$

звідки

$$\begin{aligned} a_3 &= \frac{y_3 - y_2 + y_2 - y_1 - a_2(x_3 - x_1)}{(x_3 - x_1)(x_3 - x_2)} = \\ &= \frac{f(x_2; x_3) - f(x_1; x_2)}{x_3 - x_1} = f(x_1; x_2; x_3)\,. \end{aligned}$$

У загальному випадку коефіцієнти в (2.16) дорівнюють розділеним різницям відповідного порядку: $a_k = f(x_1; \ldots; x_k)$. У результаті дістанемо *поліном Ньютона для інтерполяції вперед*:

$$\begin{aligned} P_{n-1}(x) = f(x_1) &+ f(x_1; x_2)(x - x_1) + \ldots + \\ &+ f(x_1; \ldots; x_n)(x - x_1) \cdots (x - x_{n-1})\,. \end{aligned}$$

Цей поліном особливо зручно використовувати у випадку рівновіддалених вузлів ($x_{i+1} - x_i = h$). Зробимо заміну змінних $x = x_1 + h \cdot t$, тоді

$$\begin{aligned} &P_{n-1}(x) \Rightarrow P_{n-1}(t) = y_1 + \Delta y_1 \cdot t + \Delta^2 y_1 \cdot + \\ &+ \frac{t(t-1)}{2!} + \ldots + \frac{\Delta^{n-1} y_1}{(n-1)!} t(t-1) \cdots (t-n+2)\,, \end{aligned} \tag{2.17}$$

що символічно можна переписати у вигляді

$$P_{n-1}(t) = (1 + \Delta)^t y_1\,. \tag{2.18}$$

Аналогічно можна побудувати *поліном Н'Ньютона для інтерполяції назад*,

$$\begin{aligned} P_{n-1}(x) &= f(x_n) + f(x_{n-1}; x_n)(x - x_n) + \ldots + \\ &+ f(x_1; \ldots; x_n)(x - x_n) \cdots (x - x_2)\,. \end{aligned} \tag{2.19}$$

який для рівновіддалених вузлів зручно записати у такому вигляді ($x = x_n - h \cdot t$):

$$\begin{aligned} &P_{n-1}(x) \Rightarrow P_{n-1}(t) = \\ &= y_n - \bigtriangledown y_n \cdot t + \bigtriangledown^2 y_n \frac{t(t-1)}{2!} - \bigtriangledown^3 y_n \frac{t(t-1)(t-2)}{3!} + \ldots = \\ &= (1 - \bigtriangledown)^t y_n\,. \end{aligned}$$

Задача 2. *Знайти скінченні різниці всіх порядків для таблиці із задачі 1. За їх допомогою побудувати інтерполяційний поліном Ньютона третього порядку та знайти його значення в точці* $x = 1.5$.

2.5. Інтерполяція сплайнами

Згідно з теоремою Вейєрштрасса, будь-яка неперервна функція $f(x)$ на інтервалі $[a, b]$ може бути як завгодно точно апроксімована поліномами. Але практичні можливості використання багаточленів Лагранжа (та еквівалентних їм багаточленів Ньютона) обмежені. Передусім ми повинні бути впевнені, що при достатньо великій кількості вузлів інтерполяції буде отримано добре наближення функції, яку ми інтерполюємо.

На жаль, у загальному випадку при довільному розподілі вузлів інтерполяції має місце результат:

$$\lim_{n\to\infty} \max_{x\in[a,b]} |f(x) - P_n(x)| = \infty\,. \tag{2.20}$$

Цей результат (Рунге, 1901; Бернстейн, 1916) вказує, що механічне збільшення кількості вузлів інтерполяції не завжди приводить до бажаного поліпшення наближення.

Як приклад розглянемо функцію Лорентца (розподіл Коші), яку часто використовують при аналізі спектрів в оптиці, атомній та ядерній фізиці:

$$f(x) \sim \frac{1}{a^2 + c^2x^2}, \quad x \in [-b, b]\,.$$

На рис. 2.2 зображені інтерполяційні поліноми Лагранжа для цієї функції ($a = 1, b = 1, c = 5$), що побудовані на рівномірно розподілених 5 та 11 вузлах на відрізку $[-1, 1]$ і позначені P_5 та P_{11} відповідно. Внаслідок симетрії відносно початку координат поведінка функцій зображена тільки для додатних аргументів. При збільшенні степеня інтерполяційного полінома поліпшується опис шуканої функції в області нуля, але поблизу кінців інтервалу $[-1, 1]$ розходження збільшується.

Іноді складності такого типу вдається розв'язати за допомогою спеціального вибору вузлів інтерполяції або за рахунок використання узагальнених поліномів. Але такий шлях дуже ускладнює обрахунки і до того ж не позбавляє нас

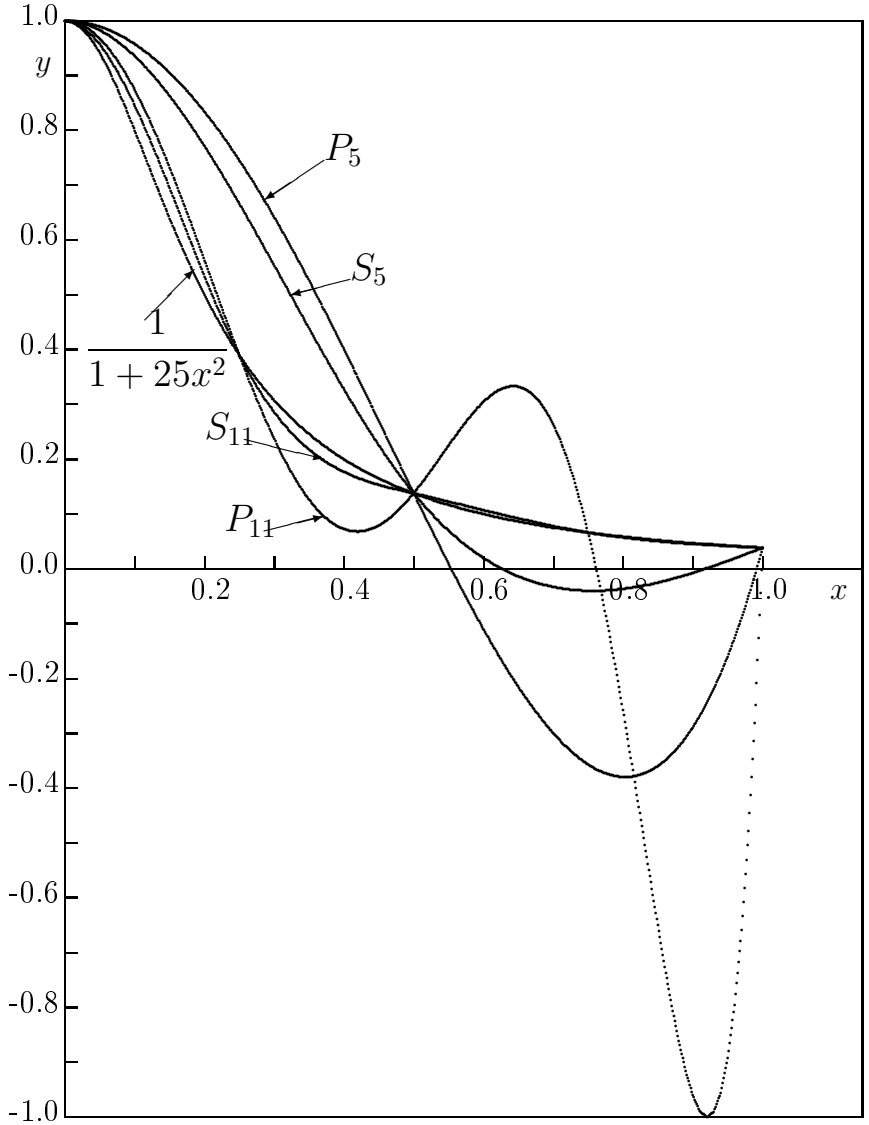

Рис. 2.2. Функція $1/(1+25x^2)$ та її наближення

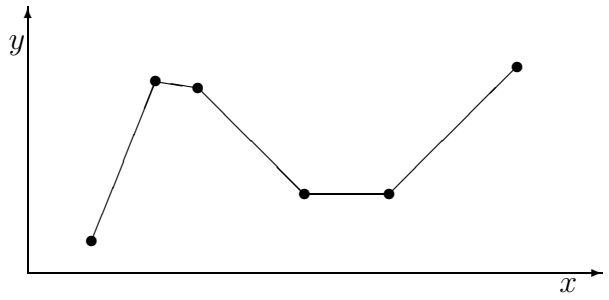

Рис. 2.3. Лінійний сплайн

іншої проблеми – швидкого накопичення похибок заокруглення із зростанням степеня полінома. Тому більш бажано замість побудови інтерполяційного полінома високої степені використовувати інтерполяцію кусковими сплайнами.

Як найпростіший приклад на Рис. 2.3 показана сплайн-інтерполяція поліномами першого степеня. Термін *сплайн* походить від англійського *spline*, що в перекладі означає пристосування, яким користувалися кресляри для проведення гладеньких кривих через задані точки площини $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Якщо гнучку пружню лінійку поставити на ребро і помістити між опорами, які розташовані в заданих точках, то за рахунок пружності профіль лінійки набуває форму, для якої потенціальна енергія деформації мінімальна.

Позначимо $S(x)$ функцію, яка описує цей профіль. Виявляється, що ця функція між кожною парою сусідніх точок є кубічним багаточленом; крім цього вона є двічи неперервно диференційованою на всьому інтервалі $[x_1, x_n]$. Функція $S(x)$ відноситься до так званих інтерполяційних кубічних сплайнів, властивості яких забезпечили їм успіх у чисельних застосуваннях.

Головною властивістю сплайна довільного степеня є те, що модуль неперервності (2.20) при нескінченному збільшенні кількості вузлів прямує до нуля, тобто похибка інтерполяції теж прямує до нуля.

Продовжимо розгляд кубічного сплайна. Нехай

$$S(x) = \sum_{i=1}^{n-1} \Theta(x - x_i)\Theta(x_{i+1} - x)S_i(x)\,,$$

де $\Theta(x)$ - функція Хевісайда, $\Theta(x) = \begin{cases} 1, & x > 0\,, \\ 0, & x < 0\,. \end{cases}$, а $S_i(x)$ – поліном третього степеня, визначений на інтервалі $[x_i, x_{i+1}]$, для побудови якого необхідні чотири параметри. Таким чином повний опис сплайна вимагає $4(n-1)$ параметрів,

для знаходження яких необхідні $4n-4$ рівняння. Оскільки сплайн повинен співпадати з таблицею в вузлах, то маємо n умов:

$$S_i(x_i) = y_i, \quad y = \overline{1,n}\,. \tag{2.21}$$

Неперервність сплайна та його перших двох похідних у внутрішніх вузлах таблиці приводить до $3(n-2)$ умов:

$$S_i(x_{i+1}) = S_{i+1}(x_{i+1})\,, \tag{2.22}$$

$$S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})\,, \tag{2.23}$$

$$S_i''(x_{i+1}) = S_{i+1}''(x_{i+1}), \quad i = \overline{2,n-1}\,. \tag{2.24}$$

Таким чином, маємо $4n-6$ умов.

Дві умови, яких бракує, можна обрати різним чином – наприклад використати так званий *природний сплайн*, коли $S_1''(x_1) = S_{n-1}''(x_n) = 0$. Це відповідає випадку, коли кінці пружньої лінійки поза інтервалом $[x_1, x_n]$ вільні. Як інший спосіб задання додаткових умов можна запропонувати рівність похідних сплайна першого, другого чи третього порядку на кінцях інтервалу $[x_1, x_n]$, і відповідних похідних від інтерполяційного полінома Лагранжа, побудованого на декількох перших та останніх вузлах (ці похідні будуть співпадати з відповідними розділеними різницями).

У принципі, для знаходження $4n-4$ коефіцієнтів поліномів третьої степені загального вигляду можна використовувати умови (2.21-2.24). Але зручніше шукати ці поліноми в спеціальному представленні:

$$S_i(x) = \omega_i y_{i+1} + \overline{\omega}_i y_i + h_i^2 \left[(\omega_i^3 - \omega_i)\sigma_{i+1} + (\overline{\omega}_i^3 - \overline{\omega}_i)\sigma_i\right]\,. \tag{2.25}$$

Тут $h_i = x_{i+1} - x_i$, $\omega_i = \omega_i(x) = (x - x_i)/h_i$, $\overline{\omega}_i = 1 - \omega_i$. Оскільки $\omega_i(x_i) = 0$, $\omega_i(x_{i+1}) = 1$, то $S_i(x_i) = y_i$ і $S_i(x_{i+1}) = y_{i+1}$, тобто представлення (2.25) автоматично задовольняє умовам (2.21)-(2.22).

Для знаходження невідомих коефіцієнтів σ_i продиференціюємо двічі вираз (2.25). У вузлових точках $x = x_i$ знайдемо

$$S_i'(x_i) = \frac{y_{i+1} - y_i}{h_i} + h_i \left[(3\omega_i^2 - 1)\sigma_{i+1} - (3\overline{\omega}_i^2 - 1)\sigma_i\right]\,, \tag{2.26}$$

$$S_i''(x_i) = 6\omega_i\sigma_{i+1} + 6\overline{\omega}_i\sigma_i\,. \tag{2.27}$$

Рівняння (2.27) забезпечує виконання умови (2.24). При цьому видно, що $\sigma_i = \frac{1}{6}S_i''(x_i)$, тобто шукані коефіцієнти σ_i визначаються кривизною полінома у вузлах інтерполяції.

Рівняння (2.23) та (2.26) приводять до системи лінійних алгебраїчних рівнянь відносно σ_i:

$$h_i\sigma_i + 2(h_i + h_{i+1})\sigma_{i+1} + h_{i+1}\sigma_{i+2} = \frac{\Delta y_{i+1}}{h_{i+1}} - \frac{\Delta y_i}{h_i}\,. \tag{2.28}$$

Дві додаткові умови можна записати у вигляді:

$$\sigma_1 = \sigma_n = 0\,, \tag{2.29}$$

або

$$\begin{aligned} 6\frac{\sigma_2 - \sigma_1}{h_1} &= f(x_1; x_2; x_3; x_4)\,, \\ 6\frac{\sigma_n - \sigma_{n-1}}{h_{n-1}} &= f(x_{n-3}; x_{n-2}; x_{n-1}; x_n)\,. \end{aligned} \tag{2.30}$$

Система рівнянь (2.28) (плюс (2.29) або (2.30)) має так звану *тридіагональну матрицю*, яка дозволяє використовувати швидкі і ефективні методи розв'язку, розглянуті нижче (*метод прогонки*); у результаті отримаємо розв'язок у вигляді сукупності значень σ_i, за допомогою яких на підставі (2.25) можна знайти значення інтерполяційного сплайна в довільній точці. Окрім задач власне інтерполяції функцій, сплайни знайшли широке використання також при підрахунку визначених інтегралів чисельними методами, що позбавило від деяких ускладнень при розв'язку інтегральних рівнянь.

Для ілюстрації на рис. 2.2 наведені сплайни, що побудовані по 5 та по 11 точках і позначені, відповідно, S_5 та S_{11}. Видно, що якість інтерполяції сплайнами для функції типу розподілу Коші набагато вища, ніж інтерполяції поліномами Лагранжа.

2.6. Інтерполяція методом найменших квадратів

До цього часу ми розглядали побудову інтерполяційних поліномів $P(x)$, значення яких збігаються із значеннями функції $y(x)$ на деякій множині вузлів:

$$P(x_i) = y(x_i) = y_i\,. \tag{2.31}$$

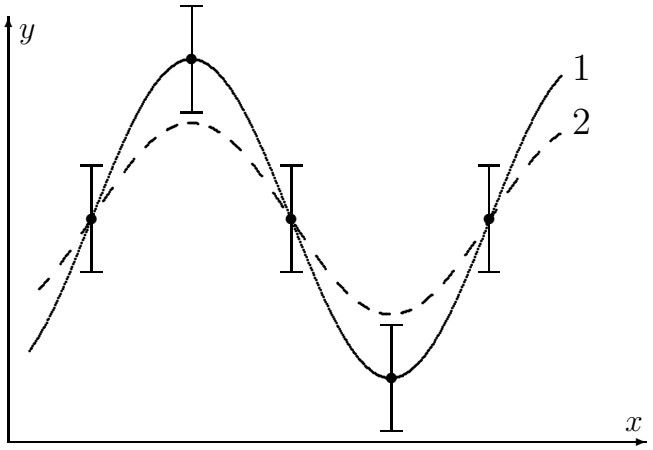

Рис. 2.4. Можливі типи інтерполяції

Функція $P(x)$ наближено описує $y(x)$ на інтервалі сітки, хоча розходження поза вузлами може бути значним (див. (2.20)). У зв'язку з цим постає питання, чи завжди необхідно вимагати виконання умови (2.31)? Справедливість цього питання особливо очевидна у випадку, коли множина $\{y_i\}$ є результатом експериментальних вимірювань і коли данні мають вихідну невиправну похибку.

У цьому випадку присутні на кривій 1 (рис. 2.4) особливості (піки, провали і т.п.) можуть відзеркалювати не реально існуючі особливості явища, а випадкові флюктуації у вимірювальних процесах. При аналізі таких результатів природно побудувати таку апроксимуючу функцію, яка описувала б данні в цілому, без відновлення деяких тонких деталей у поведінці випадкових флюктуацій (крива 2 на рис. 2.4). При такому підході можна відкинути деякі особливості функції, які реально існують, але істотно зменшиться ймовірність появи хибних ефектів
.

Ступінь "загрублення" при використанні середньоквадратичної апроксимації можна регулювати, і вона визначається в кожному конкретному випадку.

Розглянемо спочатку інтерполяцію функції, заданої у вигляді таблиці $y_i = y(x_i), i = 1, 2, \ldots$. Будемо вважати значення $\{y_i\}$ координатами вектора $\vec{y}$ в n-вимірному лінійному просторі із скалярним добутком $(\vec{y} \cdot \vec{z}) = \sum_{i=1}^{n} y_i z_i$ і нормою $\|\vec{y}\| = \sqrt{(\vec{y} \cdot \vec{y})}$.

Інтерполюючу функцію $P(x)$ будемо шукати, замінюючи вимогу (2.31) умовою мінімуму норми $\|\vec{P} - \vec{y}\|$. Тут вектор $\vec{P}$ визначається сукупністю значень $\{P(x_i)\}$ в точках $\{x_i\}$. Пошук $\inf \|\vec{P} - \vec{y}\|$ є задачею про знаходження *найкращого середньоквадратичного наближення* функції, для розв'язування якої слугує

метод найменших квадратів.

Нехай шукана функція $P(x)$ має $m \leq n$ невідомих параметрів C_k, $k = \overline{1,m}$. Явний вигляд цієї функції звичайно вважається відомим наперед (наприклад, виходячи з загальних фізичних міркуваннь і т.п.), і предметом досліджень залишається пошук C_k. Для цього необхідно розв'язати сукупність рівнянь

$$\frac{\partial}{\partial C_k}\|\vec{P} - \vec{y}\|^2 = 0, \quad k = \overline{1,m}\,, \tag{2.32}$$

які в загальному випадку є системою алгебраїчних, або трансцендентних рівнянь з m невідомими C_k. Про способи розв'язку таких систем мова піде далі, а зараз розглянемо окремий випадок, який є важливим у різноманітних застосуваннях, а саме, нехай $P(x)$ є лінійною комбінацією скінченного числа відомих функцій

$$P(x) = C_1\varphi_1(x) + \ldots + C_n\varphi_n(x)\,. \tag{2.33}$$

У цьому випадку рівняння (2.32) набудуть вигляду

$$\sum_{i=1}^{n} \left[C_1\varphi_1(x_i) + \ldots + C_n\varphi_n(x_i) - y_i\right]\varphi_k(x_i) = 0\,,$$

або

$$\begin{array}{ccccccccc} a_{11}C_1 & + & a_{12}C_2 & + & \ldots & + & a_{1n}C_n & = & d_1\,, \\ a_{21}C_1 & + & a_{22}C_2 & + & \ldots & + & a_{2n}C_n & = & d_2\,, \\ \vdots & & & & \ddots & & & & \vdots \\ a_{n1}C_1 & + & a_{n2}C_2 & + & \ldots & + & a_{nn}C_n & = & d_n\,, \end{array} \tag{2.34}$$

де

$$\begin{aligned} a_{kl} = a_{lk} &= \sum_{i=1}^{n} \varphi_k(x_i)\varphi_l(x_i)\,, \\ d_k &= \sum_{i=1}^{n} y_i\varphi_k(x_i)\,. \end{aligned}$$

Якщо вихідна сукупність функцій $\{\varphi_k\}$ є лінійно незалежною, то визначник системи (2.34) відмінний від нуля і розв'язок єдиний.

Як приклад розглянемо випадок, коли $P(x)$ є алгебраїчним поліномом степені $m-1$, тобто коли

$$P_{m-1}(x) = C_1 + C_2x + \ldots + C_mx^{m-1}\,.$$

Коефіцієнти цього полінома знаходяться при розв'язку системи (2.34) з коефіцієнтами

$$a_{kl} = \sum_{i=1}^{n} x_i^{k+l-2}, \quad d_k = \sum_{i=1}^{n} y_i x_i^{k-1},$$

звідки безпосередньо випливає єдиність розв'язку (2.34). При $m = n$ отримаємо інтерполяційний поліном Лагранжа (або Ньютона).

Задача 3. *По таблиці, наведеній у задачі 1., побудувати інтерполяційний поліном другого степеня ($m = 3$) і знайти його значення у вузлах таблиці.*

Часто задача про середньоквадратичну інтерполяцію виникає у тому випадку, коли функція $f(x)$ відома і задана неперервно на інтервалі $[a, b]$, а ми хочемо представити її в іншому вигляді, що може бути зручним, наприклад, при деяких перетвореннях.

Нехай $L_2[a, b]$ – простір дійсних функцій із скалярним добутком

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

та нормою $\|f\|_{L_2} = \sqrt{(f, f)}$.

Так саме, як і у випадку функції, що задана у вигляді таблиці, ми можемо шукати інтерполюючу функцію за допомогою умови мінімуму норми. Якщо $P(x)$ є деякою модельною функцією і належить простору L_2 та якщо для визначення $P(x)$ потрібно знання деякої кількості параметрів C_k, то по аналогії з (2.32) можна записати умову мінімуму норми

$$\frac{\partial}{\partial C_k}\|P - f\|^2 = 0, \quad k = \overline{1, m}, \tag{2.35}$$

що дозволить визначити ці параметри.

У загальному випадку рівняння (2.35) значно складніше за систему алгебраїчних або трансцендентних рівнянь (2.32), оскільки в (2.35) будуть входити інтеграли від функцій, які в свою чергу залежать від шуканих невідомих коефіцієнтів C_k. Очевидно, що така структура рівнянь робить їх мало придатними для практичного використання.

Ситуація значно спрощується, якщо для $P(x)$ використати представлення (2.33), де $\varphi_k \in L_2$. Тоді умова мінімуму норми (2.35) еквівалентна системі рівнянь

$$\begin{array}{rcl} C_1(\varphi_1,\varphi_1) + C_2(\varphi_1,\varphi_2) + \ldots + C_n(\varphi_1,\varphi_n) & = & (\varphi_1, f)\,, \\ C_1(\varphi_2,\varphi_1) + C_2(\varphi_2,\varphi_2) + \ldots + C_n(\varphi_2,\varphi_n) & = & (\varphi_2, f)\,, \\ \vdots & & \vdots \\ C_1(\varphi_n,\varphi_1) + C_2(\varphi_n,\varphi_2) + \ldots + C_n(\varphi_n,\varphi_n) & = & (\varphi_n, f)\,. \end{array} \tag{2.36}$$

Природно, що сукупність *базисних функцій* $\{\varphi_k\}$ повинна бути лінійно незалежною. Розв'язок системи рівнянь (2.36) має особливо простий вигляд у випадку ортогонального базису, тобто якщо $(\varphi_k, \varphi_l) = \delta_{kl}(\varphi_k, \varphi_k)$. Тоді

$$C_k = \frac{(\varphi_k, f)}{(\varphi_k, \varphi_k)}\,.$$

Задача 4. *Функцію $f(x) = x^2$ розкласти на інтервалі $[0, 5]$ по функціях $\varphi_1(x) = 1$ та $\varphi_2(x) = x$. Знайти значення отриманого узагальненого інтерполяційного полінома в точках $x_i = 0,\ 1,\ 2,\ 3,\ 4,\ 5$.*

Глава 3

Алгебраїчні та трансцендентні рівняння

У цьому розділі спочатку подані методи розв'язку рівнянь з одним невідомим, що дозволить значно спростити розуміння запропонованих методів. Потім буде розглянуто необхідні зміни при переході до систем рівнянь з багатьма невідомими. Методи розв'язку найбільш простих з них, систем лінійних алгебраїчних рівнянь, розглядаються у наступному розділі.

Нехай є рівняння

$$f(x) = 0\,, \tag{3.1}$$

де f - відома функція, що визначена на інтервалі $[a, b]$. Необхідно знайти значення $x = \alpha_i\,(i = 1, 2, \ldots)$, при яких значення функції $f(x)$ перетворюється в нуль.

Усі чисельні методи, що використовуються, передбачають наближене знання області, в якій знаходяться корені α_i. В той час, як самі методи досить легко алгоритмізуються і реалізуються на ЕОМ, знаходження цієї області поки що залишається за дослідником, і кожна задача потребує окремого підходу. Якщо $f(x)$ – алгебраїчний поліном, то область, де розташовані усі його корені, легко визначається за допомогою коефіцієнтів полінома. В інших випадках кількість і місце розташування коренів часто залишаються невизначеними. Тому завжди необхідна попередня докомп'ютерна робота з такою задачею. Зокрема, значно полегшує пошук цієї області аналіз фізичних (або інших) основ задачі.

Коли область розташування коренів визначена, постає питання вибору відповідного алгоритму для розв'язку даної задачі. Цінність обраного чисельного методу буде значною мірою визначатися швидкістю отримання розв'язку. Загальним для розглянутих нижче методів є те, що вони ітераційні, тобто процедура розв'язку зводиться до багатократного застосування деякого алгоритму.

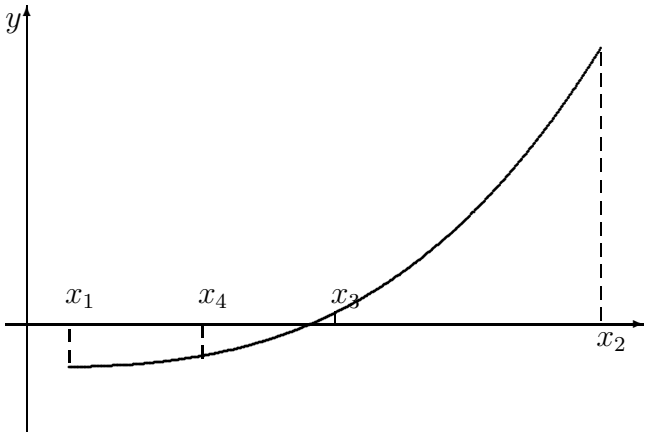

Рис. 3.1. Метод половинного ділення

Отриманий при цьому розв'язок майже завжди є наближеним, хоча і може бути скільки завгодно близьким до точного. Ці методи дуже зручні при реалізації на ЕОМ і тому вони отримали найбільше розповсюдження.

3.1. Метод половинного ділення

Метод складається з таких операцій. Попередньо визначена область розташування коренів поділяється на області меншого розміру $[x_1, x_2]$, в кожній з яких припускається наявність не більше одного кореня (тобто корені повинні бути *відокремлені*). На різних кінцях такого малого інтервалу функція повинна мати протилежні знаки, і якщо функція неперервна, наявність зміни знака вказує на наявність кореня. Корінь може бути і відсутнім в такій малій області, в цьому випадку на її двох кінцях значення функції мають однакові знаки.

На наступному кроці при наявності зміни знака функції, підраховується середнє значення: $x_3 = \dfrac{x_1 + x_2}{2}$ і значення функції в цій точці $f(x_3)$. Якщо знаки $f(x_3)$ і $f(x_1)$ збігаються, то в подальшому замість x_1 та $f(x_1)$ використовуються x_3 та $f(x_3)$. Якщо знак $f(x_3)$ є протилежним знаку $f(x_1)$, то x_3 та $f(x_3)$ використовуються надалі замість x_2 та $f(x_2)$. У результаті, інтервал, який містить корінь, звужується в два рази.

Цей процес повторюється до того часу, доки довжина інтервалу $[x_n, x_{n+1}]$ не стане меншою заданої точності. Тоді як розв'язок можна використовувати значення будь-якого з кінців інтервалу. На рис. 3.1 схематично зображений метод половинного ділення.

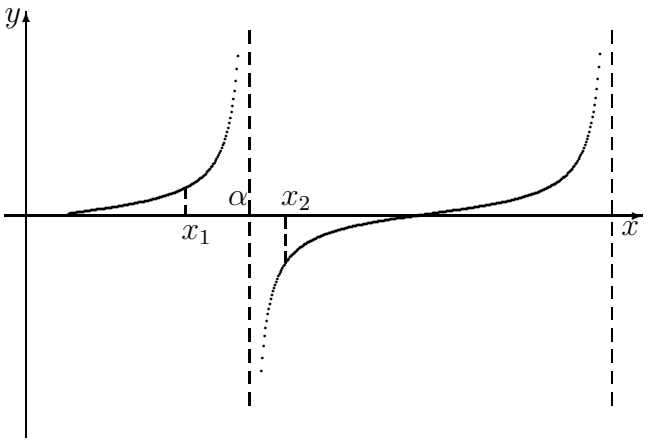

Рис. 3.2. Хибні нулі

Необхідно зробити зауваження в тому випадку, коли наперед невідомо, чи є функція $f(x)$ неперервною в області розташування коренів. Приклад розривної функції типу $\operatorname{tg} x$ показаний на рис. 3.2. В цьому випадку в околі точки розриву α (в даному випадку $\alpha = \pi/2$) знак функції також змінюється ("хибний корінь"), але при цьому є характерним швидке зростання абсолютного значення функції. Тому в алгоритмі методу поряд із вимірами на кожному кроці довжини інтервалу $[x_n, x_{n+1}]$ необхідно передбачити порівняння значення функції $f(x)$, яке обчислене в середній точці цього інтервалу, з будь-яким наперед визначеним малим числом. Процес розрахунків можна вважати закінченим, якщо одночасно виконуються умови малості як довжини інтервалу $[x_n, x_{n+1}]$, так і самої функції.

Задача 5. *Знайти корені рівняння*

$$x^2 - 4x + 3 = 0 \tag{3.2}$$

на інтервалі $[2, 5]$.

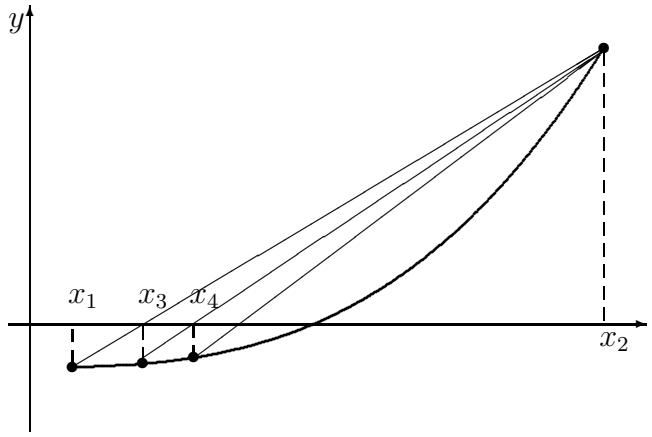

Рис. 3.3. Метод хорд

◁ Тут $f(x) = x^2 - 4x + 3$, $x_1 = 2$, $x_2 = 5$.

$$
\begin{array}{rclrcl}
x_1 &=& 2., & f(x_1) &=& -1. < 0\,, \\
x_2 &=& 5., & f(x_1) &=& 8. > 0, \\
x_3 &=& 3.5, & f(x_3) &=& 1.25 > 0\,, \\
x_4 &=& 2.75, & f(x_4) &=& -0.4375 < 0\,, \\
x_5 &=& 3.125, & f(x_5) &=& 0.2656 > 0\,, \\
x_6 &=& 2.9375, & f(x_6) &=& -0.1211 < 0\,, \\
x_7 &=& 3.0313, & f(x_7) &=& 0.06348 > 0\,, \\
x_8 &=& 2.9844, & f(x_8) &=& -0.03096 < 0\,, \\
x_9 &=& 3.0079, & f(x_9) &=& 0.01586 > 0\,, \\
x_{10} &=& 2.9962, & f(x_{10}) &=& -0.00768 < 0\,.
\end{array}
$$

З точністю до 0.01 корінь дорівнює 3.00. ▷

3.2. Метод хорд

В основі цього методу лежить лінійна інтерполяція функції по двох її значеннях, що мають протилежні знаки (рис. 3.3). Пряма, що проведена через точки з координатами $[x_1, f(x_1)]$ та $[x_2, f(x_2)]$, перетне вісь x в точці x_3:

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

На наступному кроці обчислимо $f(x_3)$ і знайдемо новий інтервал ($[x_1, x_3]$ або $[x_3, x_2]$), в якому розташований корінь. Повторюючи цю процедуру до досяг-

нення заданої точності, знайдемо шуканий корінь. При цьому, як і в методі половинного ділення, необхідно передбачити відкидання "хибних коренів".

Метод хорд, як і попередній метод, є ітераційним у тому розумінні, що наступне значення x_3 обчислюється на підставі двох попередніх (x_1 та x_2). Алгоритм цих методів можна отримати при заміні початкового рівняння (3.1) на рівняння вигляду

$$x = \varphi(x)\,, \tag{3.3}$$

причому вигляд $\varphi(x)$ залежить як від $f(x)$, так і від обраного методу. Зокрема, $\varphi(x)$ завжди можна обрати у вигляді

$$x = \varphi(x) = x - \psi(x) \cdot f(x)\,. \tag{3.4}$$

Дійсно, якщо $x = \alpha_i$ (корінь (3.1)), то (3.4) перетворюється в тотожність, якщо тільки $\psi(x)$ не має особливостей в точках α_i. Для методу хорд маємо

$$\psi(x) = \frac{x - x_1}{f(x) - f(x_1)}\,.$$

Задача 6. *Розв'язати рівняння (3.2) на інтервалі* $[2, 5]$ *за допомогою методу хорд.*

$\lhd$

$$\begin{array}{rclrcl}
x_1 &=& 2., & f(x_1) &=& -1 < 0\,; \\
x_2 &=& 5., & f(x_2) &=& 8 > 0\,; \\
x_3 &=& 2.3333, & f(x_3) &=& -0.8891 < 0\,; \\
x_4 &=& 2.5998, & f(x_4) &=& -0.6403 < 0\,; \\
x_5 &=& 2.7777, & f(x_5) &=& -0.3950 < 0\,; \\
x_6 &=& 2.8823, & f(x_6) &=& -0.2216 < 0\,; \\
x_7 &=& 2.9393, & f(x_7) &=& -0.1176 < 0\,; \\
x_8 &=& 2.9692, & f(x_8) &=& -0.0607 < 0\,; \\
x_9 &=& 2.9845, & f(x_9) &=& -0.0308 < 0\,; \\
x_{10} &=& 2.9922, & f(x_{10}) &=& -0.0155 < 0\,.
\end{array}$$

З точністю до 0.01 корінь дорівнює 2.99. $\rhd$

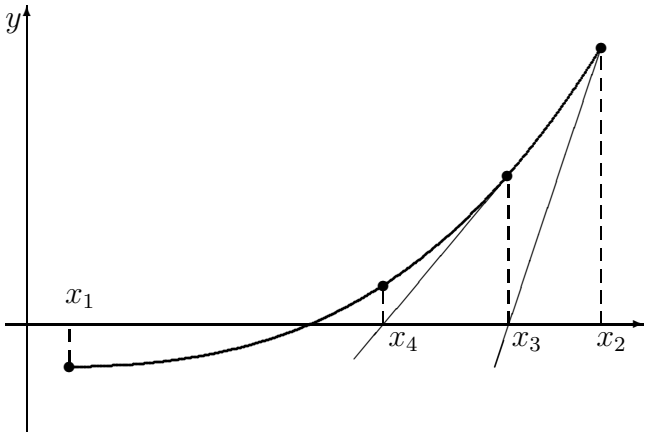

Рис. 3.4. Метод дотичних

3.3. Метод дотичних (метод Ньютона)

Геометрична інтерпретація цього методу зображена на рис. 3.4, де через точку з координатами $[x_2, f(x_2)]$ до кривої $f(x)$ проведена дотична до її перетину з віссю x в точці x_3. На наступному кроці дотичну до кривої проведемо через точку $[x_3, f(x_3)]$ і знайдемо значення x_4. Цей процес будемо повторювати до досягнення необхідної точності.

Рівняння дотичної до кривої $f(x)$ в деякій точці з координатами $[a, f(a)]$ має вигляд:

$$y = f(a) + f'(a)(x - a)\,.$$

Покладаючи $a = x_n$, $x = x_{n+1}$, знайдемо точку перетину дотичної з віссю x, яка буде наступним наближенням до шуканого кореня:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\,, \tag{3.5}$$

що в позначеннях рівняння (3.4) можна записати так $\psi(x) = \frac{1}{f'(x)}$.

Необхідно зазначити, що як початкове наближення тут необхідно використовувати ту координату x кінців інтервалу, в якій знак функції f та її другої похідної f'' збігаються, тобто

$$f(x) \cdot f''(x) > 0\,. \tag{3.6}$$

Це наочно продемонстровано на рис. 3.4, де для точки x_2 умова (3.6) виконується, і процес знаходження кореня збігається. Для точки x_1 ця умова не

виконується, і в результаті дотична перетинає вісь x за межами початкового інтервалу $[x_1, x_2]$.

Теорема 1. (Про збіжність методу дотичних.) *Якщо на проміжку* $[a, b]$, *якому належать корінь* α *і всі наближення* x_i, *функція* $f(x)$ *є така, що* $0 < |f'(x)| \le m$ *та* $|f''| < M$, *і якщо виконується умова* $C = M\dfrac{b-a}{2m} < 1$, *тоді метод Ньютона збігається.*

$\lhd$ Нехай $\alpha = x_i + h_i$, де h_i - похибка i - го кроку ітерації. Тоді

$$f(\alpha) = f(x_i + h_i) = f(x_i) + f'(x_i)\cdot h_i + \frac{1}{2!}f''(x_i + \Theta_i h_i)\cdot h_i^2 = 0\,,$$

де $0 \le \Theta \le 1$. Звідси випливає рівність

$$h_i + \frac{f(x_i)}{f'(x_i)} = -\frac{h_i^2}{2}\frac{f''((x_i + \Theta_i h_i)}{f'(x_i)}\,,$$

ліва частина якої може бути перетворена за допомогою (3.5). У результаті має місце оцінка

$$\begin{aligned} |h_{i+1}| &= \frac{h_i^2}{2}\frac{|f''((x_i + \Theta_i h_i)|}{|f'(x_i)|} \le \frac{h_i^2}{2}\frac{M}{m} \le \frac{M}{2m}\left(\frac{M}{m}\right)^2 h_{i-1}^4 \le \\ &\le \ldots \le \frac{2m}{M}\left(\frac{M}{2m}\right)^{2^{i+1}} (b-a)^{2^{i+1}} = \frac{2m}{M}C^{2^{i+1}} \overset{i\to\infty}{\longrightarrow} 0\,. \end{aligned}$$

$\rhd$

При практичному розв'язку рівнянь рекомендується комбінувати використання методів хорд та дотичних, коли декілька перших ітерацій виконуються за допомогою метода хорд, а потім для прискорення процесу збіжності використовується метод дотичних.

Задача 7. *Розв'язати рівняння (3.2) на інтервалі* $[2, 5]$ *за допомогою методу дотичних.*

$\lhd$

$$f(x)=x^2-4x+3,\, f'(x)=2x-4,\, f''(x)=2,\; f(2)=-1<0,\; f(5)=8>2.$$

З урахуванням умови (3.6) як початкове наближення необхідно обрати значення $x_0 = b = 5$. Далі

$$\begin{aligned} x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} &= 3.6667, f(x_1) = 1.7778, f'(x_1) = 3.3333, \\ x_2 &= 3.1333, f(x_2) = 0.0284, f'(x_1) = 2.2667, \\ x_3 &= 3.0078, f(x_3) = 0.0157. \end{aligned}$$

Вже на третьому кроці розв'язок знайдено з точністю 0.01. $\triangleright$

3.4. Метод послідовних наближень

Як вже зазначалося, всі описані до цього методи є ітераційними. При цьому процедура пошуку практично не залежить від явного вигляду функції $f(x)$, що робить ці методи дуже привабливими з погляду їх використання на ЕОМ.

В той же час існує цілий клас задач, для розв'язку яких потрібно лише мати кишеньковий калькулятор або логарифмічну лінійку. Такі прості методи можна застосовувати тоді, коли вихідна функція $f(x)$ має вигляд, щ о дозволяє легко отримати рівняння типу (3.3).

Метод послідовних наближень (ітерацій) полягає в послідовному знаходженні величин $x_{i+1} = \varphi(x_i)$.

Теорема 2. (Про збіжність послідовних наближень) *Якщо на проміжку $[a, b]$, якому належить корінь α та всі наближення x_i, виконується умова*

$$|\varphi'(x)| \le C < 1\,, \tag{3.7}$$

то метод послідовних наближень збігається до точного розв'язку із швидкістю геометричної прогресії.

$\triangleleft$ Якщо $\alpha = x_i + h_i$, то внаслідок теореми про середнє мають місце рівності

$$\begin{aligned} \alpha - x_1 &= \varphi'(\xi_0)(\alpha - x_0)\,, \\ \alpha - x_2 &= \varphi'(\xi_1)(\alpha - x_1)\,, \\ &\vdots \\ \alpha - x_{i+1} &= \varphi'(\xi_i)(\alpha - x_i)\,, \end{aligned} \tag{3.8}$$

де $\xi_i \in [\alpha, x_i]$. Якщо перемножити ліві та праві частини (3.8), то з урахуванням (3.7) маємо оцінку

$$|\alpha - x_{i+1}| = |h_{i+1}| \le C^{i+1}|h_0| \overset{i\to\infty}{\longrightarrow} 0\,.$$

▷

Задача 8. *Методом послідовних наближень розв'язати рівняння*

$$f(x) = x^5 - x - 1 = 0\,.$$

◁ Попередній аналіз ($f(1) = -1$, $f(2) = 29$) показує, що один із коренів лежить на інтервалі $[1, 2]$. Виходячи з даної функції $f(x)$, необхідно знайти таку $\varphi(x)$, щоб на цьому інтервалі виконувалась умова теореми. Нехай

$$x = \varphi(x) = x^5 - 1\,. \tag{3.9}$$

Тоді на цьому інтервалі $\varphi'(x) = 5x^4 - 1$ може бути більше 1 і представлення (3.9) не підходить. Дійсно, якщо $x_0 = 1$, то $x_1 = 0$, $x_2 = -1$, $x_3 = -2$, $x_4 = -33, \ldots$ Оберемо $\varphi(x)$ в іншому вигляді

$$x = \varphi(x) = (x + 1)^{\frac{1}{5}}\,, \tag{3.10}$$

що дає оцінку для похідної $\varphi'(x) = \frac{1}{5}(1 + x)^{-\frac{4}{5}} < 1$, $x \in [1, 2]$. Тоді:

$$\begin{aligned}
x_0 &= 1.000000\,,\\
x_1 = (1 + 1)^{0.2} &= 1.148698\,,\\
x_2 = (1 + 1.148698)^{0.2} &= 1.165293\,,\\
x_3 &= 1.167087\,,\\
x_4 &= 1.167281\,,\\
x_5 &= 1.167301\,,\\
x_6 &= 1.167304\,,\\
x_7 &= 1.167304\,.
\end{aligned}$$

Таким чином, для цього прикладу була отримана швидка збіжність до точного розв'язку $x = 1.167304\ldots$ ▷

3.5. Розв'язок систем рівнянь

Багато задач на одному із своїх етапів приводять до необхідності розв'язку системи рівнянь вигляду

$$\begin{aligned}
&f_1(x_1, \ldots, x_n) = 0\,,\\
&\vdots\\
&f_n(x_1, \ldots, x_n) = 0\,,
\end{aligned} \tag{3.11}$$

або

$$\vec{f}(\vec{x}) = 0\,. \tag{3.12}$$

Такі системи виникають, наприклад, при згладжуванні функцій методом найменших квадратів або при розв'язку нелінійних рівнянь математичної фізики. Тут ми розглянемо розв'язок систем (3.11) загального вигляду. Важливий, але окремий випадок системи лінійних алгебраїчних рівнянь буде розглянуто у наступному розділі.

Розглянемо лише два основні методи, вважаючи, що їх знання полегшить вивчення літератури та вибір найбільш зручних варіантів того чи іншого методу, найкращим чином пристосованого для розв'язку конкретного класу задач.

Як і у випадку одного рівняння, викладені нижче методи є ітераційними і для їх збіжності велике значення має вдалий вибір початкового наближення.

3.6. Метод Ньютона

Нехай вектор $\vec{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ є точним коренем рівнянь (3.11), (3.12) та нехай відоме достатньо добре наближення для цього розв'язку $\vec{x}^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})^T$, тобто $\vec{\alpha} = \vec{x}^{(k)} + \vec{r}^{(k)}$, де $\vec{r}^{(k)}$ - похибка k-го наближення. Якщо функція $f_i(\vec{x})$ неперервно диференційована в околі точки $\vec{x}^{(k)}$, то

$$\begin{aligned} &f_i(\vec{\alpha}) = f_i(\vec{x}^{(k)} + \vec{r}^{(k)}) = \\ &= f_i(\vec{x}^{(k)}) + \sum_{j=1}^{n} \frac{\partial f_i(x_1^{(k)}, \ldots, x_n^{(k)})}{\partial x_j} r_j^{(k)} + O\left((\vec{r}^{(k)})^2\right) = 0. \end{aligned} \tag{3.13}$$

Якщо ввести матрицю похідних

$$\hat{A}(\vec{x}) = \begin{pmatrix} \dfrac{\partial f_1(\vec{x})}{\partial x_1} & \cdots & \dfrac{\partial f_1(\vec{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_n(\vec{x})}{\partial x_1} & \cdots & \dfrac{\partial f_n(\vec{x})}{\partial x_n} \end{pmatrix}, \tag{3.14}$$

то нехтуючи членами другого порядку малості по $\vec{r}^{(k)}$ та вважаючи, що матрицю $\hat{A}(\vec{x}^{(k)})$ можна обернути, сукупність співвідношень (3.13) можна переписати у вигляді (далі введене позначення $\hat{A}_k = \hat{A}(\vec{x}^{(k)})$)

$$\vec{f}(\vec{x}^{(k)}) + \hat{A}_k \cdot \vec{r}^{(k)} = 0\,, \tag{3.15}$$

що дає значення наступного наближення

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \hat{A}_k^{-1} \cdot \vec{f}(\vec{x}^{(k)}) = \vec{\varphi}(\vec{x}^{(k)}) \,.$$

Такий ітераційний процес називається *методом Ньютона* (очевидно, що при $n = 1$, цей метод збігається з методом дотичних). При цьому на кожному кроці ітерацій необхідно розв'язувати систему лінійних рівнянь (3.15).

Для оцінки похибки $\vec{r}^{(k)}$ використаємо теорему про середнє

$$r_i^{(k+1)} = \alpha_i - x_i^{(k+1)} = \varphi_i(\vec{\alpha}) - \varphi_i(\vec{x}^{(k)}) = \sum_{j=1}^{n} \frac{\partial \varphi_i(\vec{\xi}^{(k)})}{\partial x_j} (\alpha_j - x_j^{(k)}) \,,$$

де $\vec{\xi}^{(k)}$ - координати точки на прямій, що з'єднує $\vec{\alpha}$ та $\vec{x}^{(k)}$.

Позначимо через $\hat{M}_k$ матрицю

$$\hat{M}_k = \begin{pmatrix} \dfrac{\partial \varphi_1(\vec{\xi}^{(k)})}{\partial x_1} & \cdots & \dfrac{\partial \varphi_1(\vec{\xi}^{(k)})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial \varphi_n(\vec{\xi}^{(k)})}{\partial x_1} & \cdots & \dfrac{\partial \varphi_n(\vec{\xi}^{(k)})}{\partial x_n} \end{pmatrix} .$$

Тоді

$$\vec{r}^{(k+1)} = \hat{M}_k \cdot \vec{r}^{(k)} = \hat{M}_k \cdot \hat{M}_{k-1} \cdots \hat{M}_1 \cdot \hat{M}_0 \cdot \vec{r}^{(0)} \,.$$

Таким чином, для того, щоб $\vec{x}^{(k)} \to \vec{\alpha}$ при $k \to \infty$ достатньо виконати умову

$$\hat{M}_k \cdot \hat{M}_{k-1} \cdots \hat{M}_1 \cdot \hat{M}_0 \longrightarrow 0, \quad k \to \infty \,. \tag{3.16}$$

Визначимо матрицю $\hat{M}$ з елементами $\hat{M}_{ij} = \max\limits_{\vec{\xi}} \left| \dfrac{\partial \varphi_i(\vec{\xi})}{\partial x_j} \right|$, де $\vec{\xi}$ знаходиться в області, якій належать корінь $\vec{\alpha}$ та всі наближення $\vec{x}^{(k)}$. Умова (3.16) буде виконуватись, якщо $\hat{M}^k \to 0$ при $k \to \infty$, для чого необхідно і достатньо, щоб усі власні числа матриці $\hat{M}$ були за модулем менші за одиницю. Це виконується, наприклад, тоді, коли будь-яка норма матриці буде меншою одиниці.

Задача 9. *Знайти розв'язок системи рівнянь*

$$\begin{aligned} f_1(x,y) &= 4x^2 + y^2 + 2xy - y - 2 = 0 \,, \\ f_2(x,y) &= 2x^2 + y^2 + 3xy - 3 = 0 \,, \end{aligned}$$

методом Ньютона з початковим наближенням $x^{(0)} = 0.4$, $y^{(0)} = 0.9$.

Згідно з (3.14) маємо

$$\hat{A}(x,y) = \begin{pmatrix} 8x+2y & 2y+2x-1 \\ 4x+3y & 3x+2y \end{pmatrix}.$$

Тоді

$$\hat{A}_0 = \begin{pmatrix} 5. & 1.6 \\ 4.3 & 3. \end{pmatrix}, \quad \vec{f}(x^{(0)}, y^{(0)}) = \begin{pmatrix} -0.73 \\ -0.79 \end{pmatrix}.$$

Розв'язуючи систему (3.15) з цими коефіцієнтами, отримаємо

$$\begin{pmatrix} x^{(1)} \\ y^{(1)} \end{pmatrix} = \begin{pmatrix} 0.4 \\ 0.9 \end{pmatrix} + \begin{pmatrix} 0.114 \\ 0.100 \end{pmatrix} = \begin{pmatrix} 0.514 \\ 1.000 \end{pmatrix}.$$

Далі

$$\hat{A}_1 = \begin{pmatrix} 6.112 & 2.028 \\ 5.056 & 3.542 \end{pmatrix}, \quad \vec{f}(x^{(1)}, y^{(1)}) = \begin{pmatrix} 0.084784 \\ 0.070392 \end{pmatrix},$$

$$\begin{pmatrix} x^{(2)} \\ y^{(2)} \end{pmatrix} = \begin{pmatrix} 0.514 \\ 1.000 \end{pmatrix} + \begin{pmatrix} -0.013826 \\ -0.000138 \end{pmatrix} = \begin{pmatrix} 0.500174 \\ 0.999862 \end{pmatrix},$$

$$\hat{A}_2 = \begin{pmatrix} 6.00116 & 2.000072 \\ 5.000282 & 3.500246 \end{pmatrix}, \quad \vec{f}(x^{(2)}, y^{(2)}) = \begin{pmatrix} 0.000768 \\ 0.000387 \end{pmatrix},$$

$$\begin{pmatrix} x^{(3)} \\ y^{(3)} \end{pmatrix} = \begin{pmatrix} 0.500174 \\ 0.999862 \end{pmatrix} + \begin{pmatrix} -0.000174 \\ 0.000138 \end{pmatrix} = \begin{pmatrix} 0.500000 \\ 1.000000 \end{pmatrix}.$$

У результаті ми отримали розв'язок з точністю до шістьох значущих цифр.

3.7. Метод найшвидшого спуску

Очевидно, що розв'язок системи (3.11) можна звести до задачі пошуку мінімумів функції

$$\Phi(\vec{x}) = \sum_{i,j} C_{ij} f_i(\vec{x}) f_j(\vec{x}), \tag{3.17}$$

де C_{ij} - елементи деякої додатно визначеної матриці. Симетрична матриця $\hat{C}$ називається додатно визначеною, якщо для довільного вектора $\vec{x} \neq 0$ є справедливими співвідношення

$$\left(\vec{x}, \hat{C}\vec{x}\right) = \left(\hat{C}\vec{x}, \vec{x}\right) = \sum_{i,j} C_{ij} x_i x_j > 0.$$

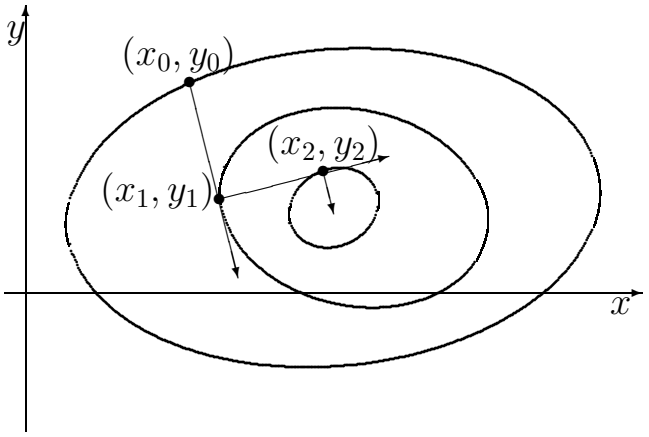

Рис. 3.5. Метод найшвидшого спуску

Оскільки $f_i(\vec{\alpha}) = 0$, то кожний нульовий мінімум функції $\Phi(\vec{x})$ дає розв'язок (3.11). Метод найшвидшого спуску для пошуку нулів $\Phi(\vec{x})$ полягає у подальшому. Нехай відомо $\vec{x}^{(0)}$ - наближене розташування кореня $\vec{\alpha}$. У цій точці функція Φ має деяке значення $\Phi(\vec{x}^{(0)})$. В околі точки $\vec{x}^{(0)}$ знайдемо точку $\vec{x}^{(1)}$, в якій $\Phi(\vec{x}^{(1)}) < \Phi(\vec{x}^{(0)})$. Для цього побудуємо вектор

$$\vec{\nabla}\Phi(\vec{x}^{(0)}) = \operatorname{grad}\Phi(\vec{x}^{(0)}) =$$
$$= \left(\sum_{ij} C_{ij} \left[\frac{\partial f_i}{\partial x_1} f_j + f_i \frac{\partial f_j}{\partial x_1} \right], \ldots, \sum_{ij} C_{ij} \left[\frac{\partial f_i}{\partial x_n} f_j + f_i \frac{\partial f_j}{\partial x_n} \right] \right) .$$

Як відомо, градієнт функції є вектор, що перпендикулярний до поверхні $\Phi(\vec{x}) = \Phi(\vec{x}^{(0)})$ і визначає напрямок найшвидшого зростання функції. У напрямку, що є протилежним до напрямку цього вектора, через точку $\vec{x}^{(0)}$ проведемо пряму

$$\vec{x} = \vec{x}^{(0)} - \lambda\vec{\nabla}\Phi(\vec{x}^{(0)}) .$$

Точка $\vec{x}^{(1)}$ буде знаходитись на цій прямій і її положення буде визначатися із умови мінімуму по λ функції (див. рис. 3.5)

$$\psi_0(\lambda) = \Phi\left(\vec{x}^{(0)} - \lambda\vec{\nabla}\Phi(\vec{x}^{(0)})\right) .$$

Для того, щоб знайти $\min\psi_0$ розв'язуємо рівняння відносно однієї невідомої

$$\frac{d}{d\lambda}\psi_0(\lambda) = \psi_0'(\lambda) = 0 ,$$

розв'язок якого λ_0 дає значення $\vec{x}^{(1)}$:

$$\vec{x}^{(1)} = \vec{x}^{(0)} - \lambda_0 \vec{\nabla}\Phi(\vec{x}^{(0)})\,.$$

Якщо $\Phi(\vec{x}^{(1)}) \neq 0$, то продовжуємо цей процес, виходячи із точки $\vec{x}^{(1)}$ і рухаючись у напрямку, що є протилежним до напрямку $\vec{\nabla}\Phi(\vec{x}^{(1)})$, знаходимо нову точку $\vec{x}^{(2)}$ і т.д. На кожному кроці потрібно буде розв'язувати рівняння

$$\psi_k'(\lambda) = 0\,, \tag{3.18}$$

де

$$\psi_k(\lambda) = \Phi\left(\vec{x}^{(k)} - \lambda\vec{\nabla}\Phi(\vec{x}^{(k)})\right)\,. \tag{3.19}$$

Отже

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \lambda_k\vec{\nabla}\Phi(\vec{x}^{(k)})\,. \tag{3.20}$$

Використовуючи цей метод, на кожному кроці ми рухаємось у напрямку найшвидшого спадання функції $\Phi(\vec{x})$. Тому ця процедура також називається *градієнтним методом.*

Якщо початкове наближення $\vec{x}^{(0)}$ обрано вдало і в точці α немає інших локальних мінімумів $\Phi(\vec{x})$, то цей процес досить швидко дає шуканий розв'язок із заданою точністю. Якщо ж в точці α є інші мінімуми, то при невдалому виборі $\vec{x}^{(0)}$ процес збіжиться, але не приведе до шуканого розв'язку. Тому при практичному використанні методу найшвидшого спуску необхідно, щоб одночасно із збіжністю векторів $\vec{x}^{(k)}$ до деякого граничного значення, мала місце збіжність $\Phi(\vec{x}^{(k)}) \to 0$.

Задача 10. *Розв'язати систему рівнянь*

$$\begin{array}{rcl} f_1(x,y) & = & 2x + y - 5 \quad = \quad 0\,, \\ f_2(x,y) & = & 4x - 3y + 5 \quad = \quad 0\,, \end{array} \tag{3.21}$$

методом найшвидшого спуску із початковим наближенням $x^{(0)} = 2$, $y^{(0)} = 2$.

$\lhd$ Обираємо матрицю $\hat{C}$ у вигляді одиничної матриці, тоді

$$\Phi(x,y) = f_1^2(x,y) + f_2^2(x,y) = 20x^2 + 10y^2 - 20xy + 20x - 40y + 50\,,$$

та

$$\frac{\partial\Phi}{\partial x} = 40x - 20y + 20\,, \quad \frac{\partial\Phi}{\partial y} = -20x + 20y - 40\,.$$

З урахуванням початкового наближення отримаємо

$$\left.\frac{\partial\Phi}{\partial x}\right|_0 = 60\,, \quad \left.\frac{\partial\Phi}{\partial y}\right|_0 = -40\,, \quad x^{(1)} = 2 - 60\lambda\,, \quad y^{(1)} = 2 + 40\lambda\,.$$

Підставимо ці значення в $\Phi(x^{(1)}, y^{(1)})$ і знайдемо, що мінімум цієї функції по λ відповідає значенню $\lambda_0 = 0.019117464$, звідки $x^{(1)} = 0.8529411$, $y^{(1)} = 2.764705$, $\Phi(x^{(1)}, y^{(1)}) = 0.294118$.

Аналогічно отримаємо

$$\left.\frac{\partial\Phi}{\partial x}\right|_1 = -1.176470\,, \quad \left.\frac{\partial\Phi}{\partial y}\right|_1 = -1.764705\,,$$
$$x^{(2)} = 0.8529411 + 1.176470\lambda\,, \quad y^{(2)} = 2.764705 + 1.764705\lambda\,,$$

що дає значення $\lambda_1 = 0.13$ та $x^{(2)} = 1.005802$, $y^{(2)} = 2.9941175$, $\Phi(x^{(2)}, y^{(2)}) = 0.00172$. Ще одна ітерація приводить до значень $\lambda_2 = 0.01911764$ та $x^{(3)} = 0.9991349$, $y^{(3)} = 2.998615$, $\Phi(x^{(3)}, y^{(3)}) = 0.0000102$.

Наведений розв'язок показує, що для даної задачі (3.21) метод найшвидшого спуску за малу кількість ітерацій приводить до правильних значень $x = 1, y = 2$. Зазначимо, що тут рівняння (3.18) є лінійним рівнянням відносно λ, а це є наслідком лінійності вихідної системи рівнянь. $\triangleright$

У загальному випадку рівняння (3.18) є трансцендентним і для його розв'язку необхідно використовувати методи, які було розглянуто вище. Очевидно, що в деяких випадках багатократний розв'язок таких рівнянь може привести до великої обчислювальної роботи. Тому нижче ми розглянемо наближений метод знаходження значення параметру λ. В результаті послідовність нових значень $\vec{x}^{(0)}, \vec{x}^{(1)}, \ldots$ не буде так швидко прямувати до точного розв'язку $\vec{\alpha}$ як при безпосередньому використанні рівняння (3.18), але кожний крок буде обраховуватися набагато швидше. Це означає, що хоча заданої точності буде досягнуто за більшу кількість кроків, загальний час обчислень скоротиться.

Метод полягає в наступному. Підставимо (3.19) в (3.17), розкладемо $\vec{f}(\vec{x})$ в ряд Тейлора поблизу точки $\vec{x}^{(k)}$ та обмежимося членами, що лінійні по λ:

$$\begin{aligned}\Phi(\vec{x}^{(k+1)}) &= \sum_{i,j} C_{ij} f_i\left(\vec{x}^{(k)} - \lambda\vec{\nabla}\Phi\right) f_j\left(\vec{x}^{(k)} - \lambda\vec{\nabla}\Phi\right) \approx \\ &\approx \sum_{i,j} C_{ij}\left[f_i - \lambda\sum_l \frac{\partial f_i}{\partial x_l}\frac{\partial\Phi}{\partial x_l}\right]\left[f_j - \lambda\sum_l \frac{\partial f_j}{\partial x_l}\frac{\partial\Phi}{\partial x_l}\right] = \psi(\lambda)\,.\end{aligned} \tag{3.22}$$

Тут і нижче відсутність аргументів у функціях f_i та Φ означає, що вони обчислюються у точці $\vec{x}^{(k)}$. Використаємо позначення (3.14) та (3.17), з урахуванням симетрії $\hat{C}$ маємо

$$\begin{aligned}\frac{\partial\Phi}{\partial x_l} &= \sum_{ij} C_{ij}\left(\frac{\partial f_i}{\partial x_l}f_j + f_i\frac{\partial f_j}{\partial x_l}\right) = \sum_{ij} C_{ij}\left(A_{il}f_j + A_{jl}f_i\right) = \\ &= 2\sum_{ij} A^T_{li}C_{ij}f_j = 2\left(A^TC\vec{f}\right)_l\,,\end{aligned}$$

або $\vec{\nabla}\Phi = 2A^TC\vec{f}$.

Тепер можна знайти значення λ, що мінімізує (3.22):

$$\frac{\partial\psi(\lambda)}{\partial\lambda} = -4\left(A^TC\vec{f}, A^TC\vec{f}\right) + 8\lambda\left(AA^TC\vec{f}, AA^TC\vec{f}\right) = 0\,,$$

та

$$\lambda = \frac{1}{2}\frac{\left(A^TC\vec{f}, A^TC\vec{f}\right)}{\left(AA^TC\vec{f}, AA^TC\vec{f}\right)}\,. \tag{3.23}$$

Вираз (3.23) дає наближене значення λ, а це означає, що у відповідній точці $\vec{x}^{(k+1)}$ функція Φ не буде мати мінімальне значення вздовж прямої $\vec{x}^{(k)} - \lambda\vec{\nabla}\Phi$. Але оскільки обчислення відповідно до формули (3.23) є дуже простими і вимагають небагато часу, ці втрати точності повністю компенсуються за рахунок збільшення кількості кроків.

Слід звернути увагу на те, що внаслідок додатної визначеності однієї $\hat{C}$ значення λ в (3.23) завжди є додатним. Це є відображенням того факту, що вектор градієнту завжди вказує напрямок максимального зростання функції.

Задача 11. *Розв'язати систему (3.21) модифікованим методом найшвидшого спуску з початковим наближенням* $x^{(0)} = 2$, $y^{(0)} = 2$.

$\lhd$ Оберемо матрицю $\hat{C}$ у вигляді одиничної та використаємо визначення (3.14). Тоді

$$\hat{A}=\begin{pmatrix}2 & 1\\ 4 & 3\end{pmatrix}, \vec{f}(x^{(0)}, y^{(0)})=\begin{pmatrix}1\\ 7\end{pmatrix}, \hat{A}^T\vec{f}=\begin{pmatrix}30\\ 20\end{pmatrix}, \hat{A}\hat{A}^T\vec{f}=\begin{pmatrix}40\\ 180\end{pmatrix}.$$

Відповідно до (3.23) маємо λ_0=0.01911764 і далі $x^{(1)}$=0.8529411, $y^{(1)} = 2.764705$. Далі аналогічно

$$\vec{f}(x^{(1)}, y^{(1)}) = \begin{pmatrix} -0.5294117 \\ 0.1176470 \end{pmatrix}, \hat{A}^T\vec{f} = \begin{pmatrix} -0.5882352 \\ -0.8823529 \end{pmatrix},$$
$$\hat{A}\hat{A}^T\vec{f} = \begin{pmatrix} -2.058823 \\ 0.2941176 \end{pmatrix}.$$

Звідси $\lambda_1 = 0.13$, $x^{(2)} = 1.005882$, $y^{(2)} = 2.994117$.

Видно, що для даної задачі (3.21) точний та наближений методи послідовного обчислення $\vec{x}^{(k)}$ дають збігаючі результати, що є наслідком лінійності вихідної системи рівнянь. $\triangleright$

Глава 4

Методи лінійної алгебри

У цьому розділі розглядаються методи розв'язку систем лінійних алгебраїчних рівнянь

$$A\vec{x} = \vec{b} \tag{4.1}$$

і методи знаходження наближених значень власних чисел і власних векторів

$$A\vec{x} = \lambda\vec{x}\,, \tag{4.2}$$

де A - квадратна матриця з елементами a_{ij}, $\vec{b}$ та $\vec{x}$ - відомий і шуканий вектори з компонентами b_i, x_i, відповідно, $(i, j = 1, \ldots, n)$, λ - власне значення.

Серед методів розв'язку систем (4.1) існують два типи – прямі та ітераційні. Перш за все розглянемо метод послідовного виключення Гаусса для систем загального вигляду та метод матричної прогонки для систем спеціального вигляду (з тридіагональною матрицею). Це - *прямі методи*.

Ефективність ітераційних методів (простої ітерації і Зейделя), що описані далі, залежить від порядку системи та структури матриць. Для розв'язку задачі (4.2) також використовуються ітераційні методи.

4.1. Метод Гаусса

Розв'язок системи (4.1), яку будемо вважати невиродженою, може бути записаний за допомогою формул Крамера

$$x_i = \frac{\Delta_i}{\Delta}\,, \tag{4.3}$$

де $\Delta = \det A \neq 0$ та $\Delta_i = \det A_i$. Матриця A_i отримується із матриці A заміною i-го стовпчика на вектор правих частин $\vec{b}$.

Хоча формули Крамера (4.3) дають розв'язок системи (4.1) в явному вигляді, з обчислювальної точки зору вони не є ефективними. Вони вимагають $\sim n(n+1)!$ операцій та виявляються дуже чутливими до похибок заокруглення. При достатньо великих n обчислення за таким методом буде неможливим і в далекому майбутньому. Для порівняння у розділі 1 приведена таблиця кількостей арифметичних операцій, що необхідні для розв'язку систем лінійних рівнянь методом Крамера та методом Гаусса, який викладається далі.

Основна вимога до методу розв'язку – мінімум кількості арифметичних дій, що є достатніми для пошуку наближеного розв'язку із заданою точністю $\varepsilon > 0$ (економічність чисельного методу).

Особливість більшості сучасних чисельних методів для розв'язку системи (4.1) полягає у відмові від знаходження оберненої матриці A^{-1}, яка б дозволила знайти вектор $\vec{x} = A^{-1}\vec{b}$.

Основна ідея методу виключень Гаусса полягає в тому, що система (4.1) приводиться до еквівалентної їй системі з верхньою трикутною матрицею (*прямий хід виключень*). З отриманої таким чином системи невідомі x_i знаходяться послідовними підстановками (*обернений хід виключень*).

Перейдемо до докладного викладення методу. *Перший крок* методу Гаусса полягає у виключенні з усіх рівнянь, окрім першого, невідомого x_1. Припускаючи, що $a_{11} \neq 0$, перше рівняння системи (4.1) розділимо на a_{11} і отримаємо

$$x_1 + \sum_{j=2}^{n} a_{1j}^{(1)} x_j = b_1^{(1)}\,, \tag{4.4}$$

де $a_{1j}^{(1)} = a_{1j}/a_{11}$ та $b_1^{(1)} = b_1/a_{11}$. Потім із кожного з рівнянь, що залишилися, віднімемо рівняння (4.4), помножене на відповідне a_{i1}. Перше невідоме виявилося виключеним з усіх рівнянь, окрім першого:

$$\begin{aligned} &\sum_{j=2}^{n} a_{ij}^{(1)} x_j = b_i^{(1)}, \quad i = 2, 3, \ldots, n, \\ &a_{ij}^{(1)} = a_{ij} - a_{i1} \cdot a_{1j}^{(1)}, \quad b_i^{(1)} = b_i - a_{i1} \cdot b_1^{(1)}\,. \end{aligned} \tag{4.5}$$

Таким чином, у матриці, яка описує систему рівнянь (4.4) та (4.5), перший стовпчик складається з нулів, за винятком елемента $a_{11}^{(1)}$, що дорівнює одиниці.

Другий крок полягає у виключенні x_2 з системи (4.5). Припускаючи, що $a_{22}^{(1)} \neq 0$, поділимо перше з рівнянь (4.5) на $a_{22}^{(1)}$; помножимо його на $a_{i2}^{(1)}$ і віднімемо

від інших рівнянь з $i = 3, 4, \ldots, n$. У результаті отримаємо систему рівнянь

$$x_2 + \sum_{j=3}^{n} a_{2j}^{(2)} x_j = b_2^{(2)}\,, \quad \left(a_{2j}^{(2)} = \frac{a_{2j}^{(1)}}{a_{22}^{(1)}}, b_2^{(2)} = \frac{b_2^{(1)}}{a_{22}^{(1)}}\right),$$
$$\sum_{j=3}^{n} a_{ij}^{(2)} x_j = b_i^{(2)}\,, \quad i = 3, 4, \ldots, n,$$
$$\left(a_{ij}^{(2)} = a_{ij}^{(1)} - a_{i2} \cdot a_{2j}^{(2)},\, b_i^{(2)} = b_i^{(1)} - a_{i1}^{(1)} \cdot b_2^{(2)}\right).$$

Продовжуючи аналогічні перетворення, після $(n-1)$-го кроку прийдемо до такої трикутної системи:

$$\begin{array}{rcrcrcrcl} x_1 & + & a_{12}^{(1)} x_2 & + & a_{13}^{(1)} x_3 & + & \ldots & + & a_{1n}^{(1)} x_n = b_1^{(1)}\,, \\ & & x_2 & + & a_{23}^{(2)} x_3 & + & \ldots & + & a_{2n}^{(2)} x_n = b_2^{(2)}\,, \\ & & & & x_3 & + & \ldots & + & a_{3n}^{(3)} x_n = b_3^{(3)}\,, \\ & & & & & & \vdots & & \vdots \\ & & & & & & & & x_n = b_n^{(n)}\,. \end{array} \tag{4.6}$$

Обернений хід методу Гаусса полягає в послідовному знаходженні невідомих $x_n, x_{n-1}, \ldots, x_2, x_1$ із системи (4.6). Нескладно показати, що метод Гаусса можна застосувати у випадку, коли всі головні мінори вихідної матриці A відмінні від нуля.

Для зручності використання методу Гаусса при розв'язку системи (4.1) за допомогою ЕОМ покажемо, що перетворення (4.4) - (4.6) є еквівалентними до розкладу матриці A на множники

$$A = F \cdot G\,, \tag{4.7}$$

де F - ліва трикутна матриця, а G - права трикутна матриця з одиницями на головній діагоналі, тобто

$$\begin{array}{l} f_{ik} = 0 \;\text{ при } i < k\,; \quad g_{ik} = 0 \quad \text{при } i > k\,; \\ \qquad\qquad\qquad\quad g_{ii} = 1, \quad i, k = \overline{1, n}\,. \end{array} \tag{4.8}$$

З цією метою відзначимо, що виключення невідомого x_i із рівнянь з номерами $i+1,\, i+2, \ldots, n$ відбувається внаслідок таких операцій:

1) ділення i-го рівняння на $a_{ii}^{(i-1)}$;

2) віднімання i-го рівняння, що помножене на $a_{ki}^{(i-1)}$, від рівнянь з номерами $k = i+1,\, i+2, \ldots, n$.

Перша з цих операцій еквівалентна множенню системи рівнянь зліва на діагональну матрицю

$$C_i = \begin{pmatrix} 1 & 0 & & \dots & & 0 \\ 0 & 1 & & \dots & & 0 \\ & & \ddots & & & \\ 0 & 0 & & \left(a_{ii}^{(i-1)}\right)^{-1} & & 0 \\ & & & & \ddots & \\ 0 & 0 & & \dots & & 1 \end{pmatrix}. \tag{4.9}$$

Друга операція еквівалентна множенню зліва на матрицю

$$\tilde{C}_i = \begin{pmatrix} 1 & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 \\ & & \ddots & & & \\ 0 & 0 & \dots & 1 & \dots & 0 \\ 0 & 0 & \dots & \alpha_{i+1} & \dots & 0 \\ & & & \vdots & \ddots & \\ 0 & 0 & \dots & \alpha_n & \dots & 1 \end{pmatrix}, \tag{4.10}$$

де $\alpha_k = -a_{ki}^{(i-1)}$ при $i < k \leq n$.

Таким чином, система (4.6), яку отримано як результат цих перетворень, можна представити у вигляді

$$CA\vec{x} = C\vec{b}, \tag{4.11}$$

де матриця C дорівнює добутку окремих перетворень (4.9) та (4.10) за виключенням $\tilde{C}_n$:

$$C = C_n\tilde{C}_{n-1}C_{n-1}\cdots\tilde{C}_2C_2\tilde{C}_1C_1\,. \tag{4.12}$$

Оскільки добуток лівих трикутних матриць є лівою трикутною матрицею, то C є також лівою трикутною матрицею.

На підставі формули для елементів оберненої матриці $\left(A^{-1}\right)_{ij} = \dfrac{A_{ji}}{\det A}$, де A_{ji} - алгебраїчне доповнення елемента a_{ji} в $\det A$, випливає, що обернена до лівої трикутної матриці є лівою трикутною матрицею, і це справедливо по відношенню до матриці

$$F = C^{-1}\,. \tag{4.13}$$

Позначимо $CA = G$, де G - права трикутна матриця системи (4.6) з одиницями на ітерації діагоналі. Як результат можна записати шуканий розклад (4.7).

Рівняння (4.7) спільно із умовою $g_{ii} = 1$, $(i = 1, 2, \ldots, n)$ утворюють систему рівнянь відносно елементів трикутних матриць F та G:

$$\sum_{j=1}^{n} f_{ij} g_{jk} = a_{ik} \,. \tag{4.14}$$

За умовою (4.8) рівняння (4.14) можна записати у вигляді

$$\sum_{j=1}^{\min(i,k)} f_{ij} g_{jk} = a_{ik} \,,$$

або

$$\sum_{j=1}^{k} f_{ij} g_{jk} = a_{ik} \, (i < k), \quad \sum_{j=1}^{i} f_{ij} g_{jk} = a_{ik} \, (i \geq k).$$

Скориставшись умовою $g_{ii} = 1$, отримаємо систему рекурентних співвідношень

$$g_{ik} = \frac{a_{ik} - \sum_{j=1}^{i-1} f_{ij} g_{jk}}{f_{ii}} \,, \qquad i < k \,, \tag{4.15}$$

$$f_{ik} = a_{ik} - \sum_{j=1}^{k-1} f_{ij} g_{jk} \,, \qquad i \geq k \,. \tag{4.16}$$

Обчислення проводяться в такій послідовності: обчислюються елементи першого стовпчика матриці F; далі - елементи першого рядка матриці G; обчислюється другий стовпчик F, другий рядок G і т.д. Це вимагає $(n^3 - n)/2$ операцій множення, додавання та ділення.

Позначимо

$$C\vec{b} = \vec{e}. \tag{4.17}$$

Тоді внаслідок (4.11), (4.13) та (4.7) є справедливими такі співвідношення:

$$F\vec{e} = \vec{b}, \tag{4.18}$$

$$G\vec{x} = \vec{e}. \tag{4.19}$$

Послідовне знаходження компонент векторів $\vec{e}$ та $\vec{x}$ за допомогою рівнянь (4.18) та (4.19) називається *оберненим ходом метода Гаусса.* З урахуванням структури матриць F і G отримаємо

$$e_i = \frac{b_i - \sum\limits_{j=1}^{i-1} f_{ij} e_j}{f_{ii}}, \quad i = 1, 2, \ldots, n, \tag{4.20}$$

та

$$x_i = e_i - \sum_{j=i+1}^{n} g_{ij} x_j\,, \quad i = n, n-1, \ldots, 2, 1\,. \tag{4.21}$$

Знаходження e_i та x_i відповідно до (4.20) та (4.21) вимагає $2n^2 - n$ арифметичних операцій, а загалом метод Гаусса в такому варіанті вимагає $\frac{1}{2}(n^3 + 4n^2 - 3n)$ операцій.

Очевидно, що має місце співвідношення

$$\det A = \det F = \prod_{i=1}^{n} f_{ii}\,. \tag{4.22}$$

Зауваження. При практичному застосуванні описаного вище метода Гаусса може статися так, що, хоча вихідна матриця A невироджена ($\det A \neq 0$), на якомусь кроці $a_{ii}^{(i-1)} = 0$ і подальші обчислення стають неможливими. Щоб цього уникнути, знаходять так званий *головний елемент* (максимальний за абсолютною величиною серед усіх елементів субматриці розмірністю $(n-i+1) \times (n-i+1)$, що розміщена у правому нижньому куті). Після цього роблять перестановку рядків та стовпчиків таким чином, щоб головний елемент опинився на місці з індексами (i, i). Далі продовжують описану вище процедуру.

Така модифікація методу Гаусса збільшує необхідну кількість операцій приблизно вдвічі, але при цьому вона не тільки допомагає запобігти випадковому діленню на нуль у процесі обчислення, але й збільшує точність результату за рахунок зменшення відносної похибки заокруглення (при інших рівних умовах відносна похибка частки зменшується із збільшенням величини дільника).

Задача 12. *Використовуючи метод Гаусса без вибору головного елемента розв'язати систему рівнянь*

$$\begin{aligned} x_1 &- 2x_2 + 3x_3 = 6\,, \\ 2x_1 &+ 3x_2 + 4x_3 = 20\,, \\ x_1 &+ x_2 + x_3 = 6\,. \end{aligned} \tag{4.23}$$

$\lhd$ Відповідно до (4.15), (4.16), (4.20) та (4.21) послідовно знаходимо:
прямий хід –

$$\begin{array}{lll} f_{11}{=}a_{11}{=}1, & f_{21}{=}a_{21}{=}2, & f_{31}{=}a_{31}{=}1, \\ g_{11}{=}1, & g_{12}{=}{-}2, & g_{13}{=}3, \\ & f_{22}{=}a_{22}{-}f_{21}g_{12}{=}7, & f_{32}{=}a_{32}{-}f_{31}g_{12}{=}3, \\ & g_{22}{=}1, & g_{23}{=}\dfrac{a_{23}{-}f_{21}g_{13}}{f_{22}}{=}{-}\dfrac{2}{7}, \\ & & f_{33}{=}a_{33}{-}f_{31}g_{13}{-}f_{32}g_{23}{=}{-}\dfrac{8}{7}, \\ & & g_{33}=1; \end{array}$$

обернений хід –

$$\begin{array}{lll} e_1{=}\dfrac{b_1}{f_{11}}{=}6\,, & e_2{=}\dfrac{b_2-f_{21}e_1}{f_{22}}{=}\dfrac{8}{7}\,, & e_3{=}\dfrac{b_3-f_{31}e_1-f_{32}e_2}{f_{33}}=3\,, \\ x_3{=}e_3{=}3\,, & x_2{=}e_2{-}g_{23}x_3{=}2\,, & x_1{=}e_1{-}g_{12}x_2{-}g_{13}x_3{=}1\,. \end{array}$$

Використовуючи співвідношення (4.22), легко отримаємо визначник системи (4.23):

$$\det A = 1\cdot 7\cdot\left(-\frac{8}{7}\right) = -8\,.$$

$\rhd$

4.2. Уточнення розв'язку

У результаті розв'язку системи лінійних рівнянь (4.1) на ЕОМ за допомогою будь-якого методу, в тому числі і методу Гаусса, отримаємо деякий вектор $\vec{x}^{(1)}$, який у загальному випадку внаслідок похибок заокруглення в арифметичних операціях буде відрізнятися від точного розв'язку $\vec{x}$ на деяку величину $\vec{x}_2 = \vec{x} - \vec{x}^{(1)}$. Якщо тепер підставити $\vec{x}^{(1)}$ в ліву частину (4.1), то в правій отримаємо вектор $\vec{b}^{(1)}$, і в результаті для знаходження $\vec{x}_2$ будемо мати рівняння

$$A\vec{x}_2 = A\left(\vec{x} - \vec{x}^{(1)}\right) = \vec{b} - \vec{b}^{(1)}\,. \tag{4.24}$$

Розв'язуючи це рівняння таким же методом як і вихідне (4.1), у загальному випадку замість точної поправки $\vec{x}_2$ отримаємо наближену поправку $\vec{x}^{(2)}$, яка

відрізняється від точної поправки на деяку величину $\vec{x}_3$. Цей вектор задовольняє рівняння

$$A\vec{x}_3 = A\left(\vec{x} - \vec{x}^{(1)} - \vec{x}^{(2)}\right) = \vec{b} - \vec{b}^{(1)} - \vec{b}^{(2)}\,. \tag{4.25}$$

Знаходячи наближено цю та наступні поправки, можна в значній мірі компенсувати вплив похибок заокруглення, а точний розв'язок оцінювати як границю суми $\vec{x} = \vec{x}^{(1)} + \vec{x}^{(2)} + \vec{x}^{(3)} + \ldots$

Таке послідовне уточнення розв'язку називається *методом Гаусса з уточненням*. Характерною особливістю цього методу є те, що розв'язок вихідної системи (4.1) і всіх наступних систем (4.24), (4.25) тощо т.д. вимагає тільки *одного прямого ходу* методу Гаусса, оскільки матриця A у всіх системах одна і та сама. Як наслідок, для знаходження кожного наступного наближення достатньо обчислити відповідній вектор $\vec{b}^{(k)}$ і робити тільки обернений хід, який вимагає $\sim n^2$ операцій, що значно менше кількості операцій, необхідних для прямого ходу.

Як приклад використання методу Гаусса з уточненням розглянемо розв'язок системи рівнянь (4.23) на гіпотетичній ЕОМ, яка проводить обчислення з точністю до двох десяткових розрядів після десяткової крапки на кожному кроці. Отримаємо

$$\begin{aligned}
&f_{11}{=}1.00,\, f_{21}{=}2.00,\, f_{31}{=}1.00,\, g_{11}{=}1.00,\, g_{12}{=}{-}2.00,\, g_{13}{=}3.00,\\
&f_{22}{=}7.00,\, f_{32}{=}3.00,\, g_{22}{=}1.00,\, g_{23}{=}0.29,\, f_{33}{=}{-}1.13,\, g_{33}{=}1.00,\\
&e_1 = 6.00, e_2 = 1.14, e_3 = 3.03,\\
&x_3 = 3.03, x_2 = 2.02, x_1 = 0.95,
\end{aligned}$$

тобто початкове наближення вектора $\vec{x}^{(1)} = \begin{pmatrix} 0.95 \\ 2.02 \\ 3.03 \end{pmatrix}$. Потім знаходимо вектор $\vec{b}^{(1)} = \begin{pmatrix} 6.00 \\ 2.08 \\ 6.00 \end{pmatrix}$ і рівняння для уточнення $\vec{x}_2$:

$$A\vec{x}_2 = \vec{b} - \vec{b}^{(1)} = \begin{pmatrix} 0.00 \\ -0.08 \\ 0.00 \end{pmatrix}\,.$$

Виконуючи тільки обернений хід, отримаємо

$$e_1 = 0.00\,, \qquad e_2 = \frac{-0.08-2.00\cdot 0.}{7.00} = -0.01\,,$$
$$e_3 = \frac{0.00-1.00\cdot 0.-3.00\cdot(-0.01)}{1.13} = -0.03\,,$$
$$x_3^{(2)} = -0.03\,, \quad x_2^{(2)} = -0.02\,, \quad x_1^{(2)} = 0.05\,.$$

У результаті

$$\vec{x} = \vec{x}^{(1)} + \vec{x}^{2)} = \begin{pmatrix} 0.95 \\ 2.02 \\ 3.03 \end{pmatrix} + \begin{pmatrix} 0.05 \\ -0.02 \\ -0.03 \end{pmatrix} = \begin{pmatrix} 1.00 \\ 2.00 \\ 3.00 \end{pmatrix},$$

що співпадає з точним розв'язком.

4.3. Метод прогонки для розв'язку тридіагональних систем

Важливим окремим випадком системи (4.1) є система лінійних рівнянь із тридіагональною матрицею. Незважаючи на свою простоту, тридіагональні матриці мають застосування в різноманітних областях: у теорії ортогональних багаточленів, при інтерполяції сплайнами, у теорії різницевих методів розв'язку задач математичної фізики тощо. Відзначимо, що і в лінійній алгебрі задача на власні значення вимагає попереднього перетворення вихідної матриці до тридіагонального вигляду.

Розглянемо систему рівнянь (4.1) з тридіагональною матрицею

$$\begin{array}{llll} a_{11}x_1 + a_{12}x_2 & & & = b_1\,, \\ a_{21}x_1 + a_{22}x_2 + & a_{23}x_3 & & = b_2\,, \\ & a_{32}x_2 + & a_{33}x_3 + a_{34}x_4 & = b_3\,, \\ & & \ddots & \vdots \\ & a_{n-1,n-2}x_{n-2} + & a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n & = b_{n-1}, \\ & & a_{n,n-1}x_{n-1} + a_{nn}x_n & = b_n\,. \end{array} \tag{4.26}$$

Рівняння цієї системи, за виключенням першого та останнього, зв'язують між собою три невідомі величини. За допомогою еквівалентних перетворень від цієї системи рівнянь можна перейти до системи двочленних рівнянь, які зв'язують дві послідовні невідомі. Покажемо це за допомогою індукції.

Перше рівняння (4.26) вже двочленне. Знаходячи з нього x_1 та підставляючи його в друге рівняння, отримаємо також двочленне рівняння (що буде містити невідомі x_2 та x_3). Подовжуючи цей процес, припустимо, що $(i-1)$-ше рівняння зведено до вигляду

$$x_{i-1} = \alpha_i x_i + \beta_i \,. \tag{4.27}$$

Підставимо це значення в i-те рівняння (4.26):

$$a_{i,i-1}\left(\alpha_i x_i + \beta_i\right) + a_{ii}x_i + a_{i,i+1}x_{i+1} = b_i \,,$$

або

$$x_i = -\frac{a_{i,i+1}}{a_{i,i-1}\alpha_i + a_{ii}}x_{i+1} + \frac{b_i - a_{i,i-1}\beta_i}{a_{i,i-1}\alpha_i + a_{ii}} \,. \tag{4.28}$$

Порівнюючи цей вираз із (4.27), отримаємо рекурентні співвідношення

$$\alpha_{i+1} = -\frac{a_{i,i+1}}{a_{i,i-1}\alpha_i + a_{ii}} \,, \quad \beta_{i+1} = \frac{b_i - a_{i,i-1}\beta_i}{a_{i,i-1}\alpha_i + a_{ii}} \,. \tag{4.29}$$

Оскільки перше рівняння (4.26) вже має двочленний характер, то для знаходження всіх α_i та β_i за допомогою (4.29) можна використати початкові умови

$$\alpha_2 = -\frac{a_{12}}{a_{11}} \,, \quad \beta_2 = \frac{b_1}{a_{11}} \,. \tag{4.30}$$

У результаті ми отримали допоміжні коефіцієнти $(\alpha_i, \beta_i,\ i = 2, \ldots, n)$. Тепер можна використовувати рівняння (4.27) з метою послідовного визначення x_{n-1}, x_{n-2}, …, x_2, x_1, для чого необхідно знати x_n. Зазначимо, що невідомі x_{n-1} та x_n одночасно мають задовольняти останні з рівнянь (4.26) та (4.27), тобто

$$\begin{aligned} a_{n,n-1}x_{n-1} \;+\; a_{nn}x_n \;&=\; b_n \,, \\ x_{n-1} \;-\; \alpha_n x_n \;&=\; \beta_n \,, \end{aligned}$$

звідки

$$x_n = \frac{b_n - a_{n,n-1}\beta_n}{a_{n,n-1}\alpha_n + a_{nn}} \,, \tag{4.31}$$

що збігається з виразом для β_{n+1} (див. (4.29)).

Рекурентні співвідношення (4.29) з початковими умовами (4.30) називаються *прямою прогонкою*, а рекурентні співвідношення (4.28) та (4.31) - *оберненою прогонкою*. Цей метод вимагає усього $\sim 9n$ арифметичних операцій.

Приклад 1. *Розв'язати систему*

$$\begin{array}{rcrcrcrcl} -2x_1 & + & 3x_2 & & & & & = & 4\,, \\ x_1 & + & 2x_2 & + & x_3 & & & = & 8\,, \\ & & x_2 & - & 6x_3 & + & 2x_4 & = & -8\,, \\ & & & & 3x_3 & - & 4x_4 & - & x_5 & = & -12\,, \\ & & & & & & x_4 & + & 4x_5 & = & 24\,. \end{array} \tag{4.32}$$

методом прогонки.

◁ Відповідно до (4.30) та (4.29), (4.31) та (4.27) маємо:
1) пряма прогонка –

$$\begin{array}{rclrcl} \alpha_2 & = & \dfrac{3}{2}\,, & \beta_2 & = & -2\,, \\ \alpha_3 & = & -\dfrac{2}{7}\,, & \beta_3 & = & \dfrac{20}{7}\,, \\ \alpha_4 & = & \dfrac{7}{22}\,, & \beta_4 & = & \dfrac{19}{11}\,, \\ \alpha_5 & = & -\dfrac{22}{67}\,, & \beta_5 & = & \dfrac{378}{67}\,, \\ & & & \beta_6 & = & 5\,; \end{array}$$

2) обернена прогонка –

$$x_5 = \beta_6 = 5\,,\ x_4 = 4\,,\ x_3 = 3\,,\ x_2 = 2\,,\ x_1 = 1\,.$$

▷

4.4. Ітераційні методи

Із збільшенням порядку систем лінійних рівнянь, які необхідно розв'язувати, точні методи стають мало придатними через велику кількість арифметичних операцій ($\sim n^3$), а також внаслідок накопичення обчислювальної похибки. В цьому випадку необхідно переходити до ітераційних методів, які достатньо широко використовуються при розв'язку різницевих рівнянь математичної фізики, операторам яких відповідають стрічкові матриці A вищого порядку.

Взагалі видно, що систему (4.1) можна за допомогою невиродженної матриці D та довільної сталої μ записати в еквівалентній формі:

$$\vec{x} = \vec{x} - \mu D\left(A\vec{x} - \vec{b}\right) = \left(E - \mu DA\right)\vec{x} + \mu D\vec{b} = B\vec{x} + \vec{g}\,, \tag{4.33}$$

де E - одинична матриця. Є справедливим і обернений перехід (4.33) $\longrightarrow$ (4.1), якщо тільки B не є одиничною матрицею.

Нехай задано деяке початкове наближення $\vec{x}^{(0)}$ і за допомогою (4.33) можна знайти всі послідовні наближення розв'язку $\vec{x}$ (*метод простої ітерації*, який іноді називають *методом Якобі*):

$$\vec{x}^{(k+1)} = B\vec{x}^{(k)} + \vec{g}\,. \tag{4.34}$$

Методи, що засновані на співвідношенні (4.34), називаються *однокроковими* (або *двошаровими*).

Для оцінки збіжності методу типу (4.34) знадобляться деякі відомості з лінійної алгебри. Нагадаємо означення основних норм векторів та матриць у лінійних просторах.

Якщо в лінійному просторі визначена норма $\|\vec{x}\|$ вектора $\vec{x} = (x_1, x_2, \ldots, x_n)$, то узгоджена з нею норма матриці A визначається як

$$\|A\| = \sup_{\|\vec{x}\neq 0\|} \frac{\|A\vec{x}\|}{\|\vec{x}\|}\,.$$

Найчастіше вживані є такі норми:

$$\|\vec{x}\|_1 = \max_{1\le i\le n} |x_i|\,, \quad \|A\|_1 = \max_{1\le i\le n} \left(\sum_{j=1}^{n} |a_{ij}|\right)\,, \tag{4.35}$$

$$\|\vec{x}\|_2 = \sum_{i=1}^{n} |x_i|\,, \quad \|A\|_2 = \max_{1\le j\le n} \left(\sum_{i=1}^{n} |a_{ij}|\right)\,, \tag{4.36}$$

$$\|\vec{x}\|_3 = \sqrt{\sum_{i=1}^{n} |x_i|^2} = \sqrt{(\vec{x},\vec{x})}\,, \quad \|A\|_3 = \sqrt{\max_{1\le i\le n} \lambda_i}\,, \tag{4.37}$$

де λ_i – i-те власне число симетричної додатно визначеної матриці A^TA.

Теорема 3. (Про збіжність методу простої ітерації.) *Якщо $\|B\| < 1$, то система рівнянь (4.33) має єдиний розв'язок, а ітераційний процес (4.34) прямує до розв'язку із швидкістю геометричної прогресії.*

$\lhd$ Для будь-якого розв'язку рівняння (4.33) має місце оцінка $\|\vec{x}\| \le \|B\| \cdot \|\vec{x}\| + \|\vec{g}\|$, звідки $\|\vec{x}\| \le \dfrac{\|\vec{g}\|}{1-\|B\|}$. Звідси випливає єдиність (та рівність нулеві) розв'язку однорідного рівняння $\vec{x} = B\vec{x}$, та, відповідно, рівняння (4.33).

Нехай $\vec{\alpha}$ - точний розв'язок рівняння (4.1) (або (4.33)). Тоді з (4.33) та (4.34) отримаємо рівняння відносно похибки $\vec{r}^{(k)} = \vec{\alpha} - \vec{x}^{(k)}$:

$$\vec{r}^{(k+1)} = B\vec{r}^{(k)}\,.$$

Очевидно, що $\vec{r}^{(k+1)} = B^{k+1}\vec{r}^{(0)}$, звідки випливає

$$\|\vec{r}^{(k+1)}\| \le \|B\|^{k+1}\|\vec{r}^{(0)}\| \longrightarrow 0\,.$$

$\triangleright$

Наслідок 3.1. *. Для збіжності методу простої ітерації достатньо виконання умови*

$$\sum_{i,j=1}^{n} b_{ij}^2 < 1\,. \tag{4.38}$$

$\triangleleft$ Щоб показати це, скористуємося нормою (4.37); тоді для будь-якого $\vec{x} \neq 0$ є справедливим співвідношення

$$\lambda = \frac{\left(B^TB\vec{x}, \vec{c}\right)}{(\vec{x}, \vec{x})} = \frac{(B\vec{x}, B\vec{c})}{(\vec{x}, \vec{x})} \ge 0\,,$$

тобто всі власні числа матриці B^TB є невід'ємними, до того ж $\lambda \le \lambda_1 + \lambda_2 + \ldots + \lambda_n$, де λ_i - власні числа матриці B^TB.

Очевидно, що

$$\max_i \lambda_i \le \lambda_1 + \lambda_2 + \ldots + \lambda_n = \operatorname{Sp} B^TB = \sum_{i,j=1}^{n} b_{ij}^2\,,$$

звідки випливає виконання умови теореми $\|B\|_3 \le \sqrt{\sum\limits_{i,j=1}^{n} b_{ij}^2}$. $\triangleright$

Приклад 2. *Розв'язати за допомогою методу простої ітерації (Якобі) систему рівнянь*

$$\begin{aligned} 5x &- 2y + z = 4\,,\\ -2x &+ 10y + z = 21\,,\\ x &+ y + 5z = 18\,. \end{aligned} \tag{4.39}$$

$\lhd$ Для того, щоб перейти до рівнянь типу (4.33), знайдемо з першого рівняння (4.39) невідоме x, з другого – y, та з третього – z:

$$\begin{array}{rcrcrcl} x &=& & 0.4y &-& 0.2z &+ \ 0.8\,, \\ y &=& 0.2x & &-& 0.1z &+ \ 2.1\,, \\ z &=& -0.2x &- \ 0.2y & & &+ \ 3.6\,. \end{array} \tag{4.40}$$

У результаті матриця B та вектор $\vec{g}$ набувають вигляду:

$$B = \begin{pmatrix} 0.0 & 0.4 & -0.2 \\ 0.2 & 0.0 & -0.1 \\ -0.2 & -0.2 & 0.0 \end{pmatrix}, \quad \vec{g} = \begin{pmatrix} 0.8 \\ 2.1 \\ 3.6 \end{pmatrix}. \tag{4.41}$$

Оберемо як початкове наближення значення $x^{(0)} = 0$, $y^{(0)} = 0$, $z^{(0)} = 0$ та послідовно використаємо рівняння (4.34):

$$\begin{array}{lcll} x^{(1)} = 0.4 \cdot 0 - 0.2 \cdot 0 & + & 0.8 & = 0.8, \\ y^{(1)} = 0.2 \cdot 0 - 0.1 \cdot 0 & + & 2.1 & = 2.1, \\ z^{(1)} = -0.2 \cdot 0 - 0.2 \cdot 0 & + & 3.6 & = 3.6, \\ x^{(2)} = 0.4 \cdot 2.1 - 0.2 \cdot 3.6 & + & 0.8 & = 0.92, \\ y^{(2)} = 0.2 \cdot 0.8 - 0.1 \cdot 3.6 & + & 2.1 & = 1.90, \\ z^{(2)} = -0.2 \cdot 0.8 - 0.2 \cdot 2.1 & + & 3.6 & = 3.02, \\ x^{(3)} = 0.4 \cdot 1.90 - 0.2 \cdot 3.02 & + & 0.8 & = 0.956, \\ y^{(3)} = 0.2 \cdot 0.92 - 0.1 \cdot 3.02 & + & 2.1 & = 1.982, \\ z^{(3)} = -0.2 \cdot 0.92 - 0.2 \cdot 1.90 & + & 3.6 & = 3.036, \\ x^{(4)} = 0.4 \cdot 1.982 - 0.2 \cdot 3.036 & + & 0.8 & = 0.9856, \\ y^{(4)} = 0.2 \cdot 0.956 - 0.1 \cdot 3.036 & + & 2.1 & = 1.9876, \\ z^{(4)} = -0.2 \cdot 0.956 - 0.2 \cdot 1.982 & + & 3.6 & = 3.0124, \\ x^{(5)} = 0.4 \cdot 1.9876 - 0.2 \cdot 3.0124 & + & 0.8 & = 0.99256, \\ y^{(5)} = 0.2 \cdot 0.9856 - 0.1 \cdot 3.0124 & + & 2.1 & = 1.99588, \\ z^{(5)} = -0.2 \cdot 0.9856 - 0.2 \cdot 1.9876 & + & 3.6 & = 3.00536. \end{array}$$

Видно, що в цій задачі метод простої ітерації досить швидко прямує до точних значень $x = 1$, $y = 2$, $z = 3$. $\rhd$

Запишемо в явному вигляді систему рівнянь (4.34)

$$\begin{array}{rcl} x_1^{(k+1)} & = & b_{11}x_1^{(k)} + b_{12}x_2^{(k)} + \ldots + b_{1n}x_n^{(k)} + g_1\,, \\ x_2^{(k+1)} & = & b_{21}x_1^{(k)} + b_{22}x_2^{(k)} + \ldots + b_{2n}x_n^{(k)} + g_2\,, \\ \vdots & & \vdots \\ x_n^{(k+1)} & = & b_{n1}x_1^{(k)} + b_{n2}x_2^{(k)} + \ldots + b_{nn}x_n^{(k)} + g_n\,. \end{array} \tag{4.42}$$

Обчислення відповідно до (4.42) звичайно проводяться у такій послідовності: спочатку обраховується $x_1^{(k+1)}$, потім $x_2^{(k+1)}$ і т.д. Таким чином, до моменту обчислення $x_m^{(k+1)}$ вже відомі значення компонент $x_1^{(k+1)}, \ldots, x_{m-1}^{(k+1)}$ цієї ж $(k+1)$-ої ітерації, тому при обчисленні $x_m^{(k+1)}$ можна використовувати ці значення замість $x_1^{(k)}, \ldots, x_{m-1}^{(k)}$. Ми отримали *ітераційний метод Зейделя* (чи *Гаусса-Зейделя*):

$$\begin{array}{rcl} x_1^{(k+1)} & = & b_{11}x_1^{(k)} + b_{12}x_2^{(k)} + \ldots + b_{1n}x_n^{(k)} + g_1\,, \\ x_2^{(k+1)} & = & b_{21}x_1^{(k+1)} + b_{22}x_2^{(k)} + \ldots + b_{2n}x_n^{(k)} + g_2\,, \\ \vdots & & \vdots \\ x_n^{(k+1)} & = & b_{n1}x_1^{(k+1)} + b_{n2}x_2^{(k+1)} + \ldots + b_{nn}x_n^{(k)} + g_n\,. \end{array} \tag{4.43}$$

(Цікаво відзначити, що цей метод не був відомий Зейделю, а Гаусс вважав його зовсім некорисним).

Систему рівнянь (4.43) можна записати в матричному вигляді:

$$\vec{x}^{(k+1)} = B_1\vec{x}^{(k+1)} + B_2\vec{x}^{(k)} + \vec{g}\,, \tag{4.44}$$

де

$$B_1 = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ b_{21} & 0 & \ldots & 0 \\ & \ldots & & \\ b_{n1} & b_{n2} & \ldots & 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1n} \\ 0 & b_{22} & \ldots & b_{2n} \\ & \ldots & & \\ 0 & 0 & \ldots & b_{nn} \end{pmatrix}.$$

Такий запис показує, що метод Зейделя (4.44) еквівалентний методу простої ітерації (4.34) з відповідною матрицею. Дійсно, за допомогою (4.44) можна записати:

$$\vec{x}^{(k+1)} = (E - B_1)^{-1}\, B_2\vec{x}^{(k)} + (E - B_1)^{-1}\,\vec{g}\,,$$

що збігається з (4.34) з урахуванням відповідних перетворень.

Теорема 4. (Про збіжність метода Зейделя.) *Якщо* $\|B\|_1 = \nu < 1$, *то метод Зейделя збігається.*

$\lhd$ Якщо вектор $\vec{\alpha}$ є розв'язком вихідного рівняння (4.33), то, враховуючи (4.44), маємо $B = B_1 + B_2$ і

$$\vec{\alpha} - \vec{x}^{(k+1)} = B_1\left(\vec{\alpha} - \vec{x}^{(k+1)}\right) + B_2\left(\vec{\alpha} - \vec{x}^{(k)}\right). \tag{4.45}$$

Введемо позначення: $\beta_i = \sum\limits_{j=1}^{i-1} |b_{ij}|$, $\gamma_i = \sum\limits_{j=i}^{n} |b_{ij}|$. Тоді для i-ої компоненти рівняння (4.45) є справедливим співвідношення

$$\begin{aligned} |\alpha_i - x_i^{(k+1)}| \le & \sum_{j=1}^{i-1} |b_{ij}| \cdot |\alpha_j - x_j^{(k+1)}| + \sum_{j=i}^{n} |b_{ij}| \cdot |\alpha_j - x_j^{(k)}| \le \\ & \le \beta_i \|\vec{\alpha} - \vec{x}^{(k+1)}\|_1 + \gamma_i \|\vec{\alpha} - \vec{x}^{(k)}\|_1 , \end{aligned} \tag{4.46}$$

де використано визначення (4.35).

Умова (4.46) виконується для всіх i, в тому числі і при $i = i_0$, коли ліва частина (4.46) набуває максимального значення, тобто

$$\|\vec{\alpha} - \vec{x}^{(k+1)}\|_1 \le \beta_{i_0} \|\vec{\alpha} - \vec{x}^{(k+1)}\|_1 + \gamma_{i_0} \|\vec{\alpha} - \vec{x}^{(k)}\|_1 ,$$

або

$$\|\vec{\alpha} - \vec{x}^{(k+1)}\|_1 \le \frac{\gamma_{i_0}}{1 - \beta_{i_0}} \|\vec{\alpha} - \vec{x}^{(k)}\|_1 .$$

Введемо позначення $\nu' = \max\limits_i \dfrac{\gamma_i}{1 - \beta_i}$ і покажемо, що $\nu' < \nu$. Оскільки $\sum_{j=1}^{n} |b_{ij}| = \beta_i + \gamma_i \le \nu < 1$, то

$$\beta_i + \gamma_i - \frac{\gamma_i}{1 - \beta_i} = \frac{\beta_i(1 - \beta_i - \gamma_i)}{1 - \beta_i} > 0 ,$$

а отже

$$\nu' = \max_i \frac{\gamma_i}{1 - \beta_i} \le \max_i (\beta_i + \gamma_i) = \nu .$$

Звідси випливає, що $\|\vec{\alpha} - \vec{x}^{(k+1)}\|_1 \le \nu \|\vec{\alpha} - \vec{x}^{(k)}\|_1$. $\rhd$

Приклад 3. *Методом Зейделя розв'язати систему* (4.39), (4.40) *з початковими умовами* $x^{(0)} = 0$, $y^{(0)} = 0$, $z^{(0)} = 0$.

На підставі (4.43) отримаємо

$$\begin{aligned}
x^{(1)} &= 0.4\cdot 0 - 0.2\cdot 0 + 0.8 && = 0.8\,,\\
y^{(1)} &= 0.2\cdot 0.8 - 0.1\cdot 0 + 2.1 && = 2.26\,,\\
z^{(1)} &= -0.2\cdot 0.8 - 0.2\cdot 2.26 + 3.6 && = 2.988\,,\\
x^{(2)} &= 0.4\cdot 2.26 - 0.2\cdot 2.988 + 0.8 && = 1.1064\,,\\
y^{(2)} &= 0.2\cdot 1.1064 - 0.1\cdot 2.988 + 2.1 && = 2.02248\,,\\
z^{(2)} &= -0.2\cdot 1.1064 - 0.2\cdot 2.02248 + 3.6 && = 2.974224\,,\\
x^{(3)} &= 0.4\cdot 2.02248 - 0.2\cdot 2.974224 + 0.8 && = 1.0141472\,,\\
y^{(3)} &= 0.2\cdot 1.0141472 - 0.1\cdot 2.974224 + 2.1 && = 2.005407\,,\\
z^{(3)} &= -0.2\cdot 1.0141472 - 0.2\cdot 2.005407 + 3.6 && = 2.9960892\,,\\
x^{(4)} &= 0.4\cdot 2.005407 - 0.2\cdot 2.9960892 + 0.8 && = 1.002945\,,\\
y^{(4)} &= 0.2\cdot 1.002945 - 0.1\cdot 2.9960892 + 2.1 && = 2.0009801\,,\\
z^{(4)} &= -0.2\cdot 1.002945 - 0.2\cdot 2.0009801 + 3.6 && = 2.999215\,,\\
x^{(5)} &= 0.4\cdot 2.0009801 - 0.2\cdot 2.999215 + 0.8 && = 1.000549\,,\\
y^{(5)} &= 0.2\cdot 1.000549 - 0.1\cdot 2.999215 + 2.1 && = 2.0001883\,,\\
z^{(5)} &= -0.2\cdot 1.000549 - 0.2\cdot 2.0001883 + 3.6 && = 2.9998526\,.
\end{aligned}$$

З цього виходить, що для даного прикладу метод Зейделя збігається значно швидше, ніж метод простої ітерації. Ця обставина пов'язана з тим, що матриця A системи (4.39) є *симетричною додатно визначеною матрицею.*

У загальному ж випадку існують такі системи рівнянь, для яких метод простої ітерації розбігається, а метод Зейделя збігається. Є вірним також обернене твердження. З практичної точки зору метод простої ітерації зручно використовувати, коли вихідна матриця A в (4.1) "майже діагональна", тобто коли

$$\sum_{j=1,j\neq i}^{n} |a_{ij}| < |a_{ii}| \quad \text{або} \sum_{i=1,i\neq j}^{n} |a_{ij}| < |a_{jj}|\,.$$

У той же час використання метода Зейделя припускає, що матриця A є "майже лівою трикутною".

4.5. Обумовленість матриць

Спочатку число обумовленості матриці було введено як засіб *апріорної* оцінки того, наскільки великими можуть бути помилки при розв'язанні системи

$A\vec{x} = \vec{b}$. Але надалі ці числа стали використовуватися для *апостеріорної* оцінки помилок – при цьому спочатку розв'язується задача, а потім оцінюється правильність отриманного результату.

Визначимо *природне число обумовленості*. Для цього розглянемо точний розв'язок рівняння (4.1) і розв'язок $\tilde{\vec{x}}$ збуреної задачі

$$A\tilde{\vec{x}} = \vec{b} + \vec{r}, \tag{4.47}$$

де $\vec{r}$ описує збурення (похибки) як правої частини, так і коефіцієнтів матриці A. Віднімаючи (4.47) з (4.1), отримаємо

$$A\left(\vec{x} - \tilde{\vec{x}}\right) = \vec{r}, \tag{4.48}$$

або $\delta\vec{x} = \vec{x} - \tilde{\vec{x}} = A^{-1}\vec{r}$.

У результаті маємо оцінку

$$\|\delta\vec{x}\| = \|A^{-1}\vec{r}\| \le \|A^{-1}\| \cdot \|\vec{r}\|,$$

і аналогічно для відносної похибки

$$\frac{\|\delta\vec{x}\|}{\|\vec{x}\|} = \frac{\|A^{-1}\| \cdot \|\vec{b}\|}{\|\vec{x}\|} \cdot \frac{\|\vec{r}\|}{\|\vec{b}\|}. \tag{4.49}$$

Тут величина $\dfrac{\|A^{-1}\| \cdot \|\vec{b}\|}{\|\vec{x}\|}$ визначає *природне число обумовленості*.

Але це число не використовується широко в лінійній алгебрі, тому що воно залежить від $\|\vec{x}\|$, а значить не дозволяє отримати апріорну оцінку. Можна виключити залежність від $\|\vec{x}\|$ і отримати *стандартне число обумовленості*. Із (4.1) маємо $\|\vec{b}\| = \|A\vec{x}\| \le \|A\| \cdot \|\vec{x}\|$, звідки

$$\frac{1}{\|\vec{x}\|} \le \frac{\|A\|}{\|\vec{b}\|}. \tag{4.50}$$

Підстановка (4.50) в (4.49) дає оцінку

$$\frac{\|\delta\vec{x}\|}{\|\vec{x}\|} \le \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\vec{r}\|}{\|\vec{b}\|}. \tag{4.51}$$

де $\|A^{-1}\| \cdot \|A\|$ - *стандартне число обумовленості*.

Відзначимо, що обидва ці числа обумовленості є тільки оцінки того, наскільки можуть бути збільшені похибки у розв'язку порівнянно з похибками в $\vec{b}$. Вони завжди дають завищену оцінку.

Розглянемо третій тип числа обумовленості (*Ейрда-Лінча*). Нехай C - матриця, що апроксимує A^{-1} (тобто така, що отримана яким-небудь чисельним методом, один із яких розглянемо нижче). Тоді з (4.48) маємо

$$C\vec{r} = CA\left(\vec{x} - \tilde{\vec{x}}\right), \tag{4.52}$$

і $\|\delta\vec{x}\| \le \|(CA)^{-1}C\vec{r}\| \le \|(CA)^{-1}\| \cdot \|C\vec{r}\|$.

Використаємо матричну рівність для довільної матриці B (якщо $\|B - E\| < 1$)

$$B^{-1} = (E + B - E)^{-1} = E - (B - E) + (B - E)^2 - \dots$$

Тоді

$$\|B^{-1}\| \le \|E\| + \|B - E\| + \|B - E\|^2 + \dots = \frac{1}{\|E\| - \|B - E\|}.$$

Відзначимо, що $\|E\| = 1$; це приводить до оцінки

$$\|\delta\vec{x}\| \le \frac{1}{1 - \|CA - E\|}\|C\vec{r}\|,$$

і для відносної помилки

$$\frac{\|\delta\vec{x}\|}{\|\vec{x}\|} \le \frac{\|C\vec{r}\|}{\|\vec{x}\|(1 - T)}, \tag{4.53}$$

де використано позначення $T = \|CA - E\|$.

Повернемося до (4.52): $\|C\vec{r}\| \le \|CA\| \cdot \|\delta\vec{x}\|$, або $\|\delta\vec{x}\| \ge \dfrac{\|C\vec{r}\|}{\|CA\|}$. Оскільки

$$\|CA\| = \|E + (CA - E)\| \le \|E\| + \|CA - E\| = 1 + T,$$

то

$$\frac{\|\delta\vec{x}\|}{\|\vec{x}\|} \ge \frac{\|C\vec{r}\|}{\|\vec{x}\|(1 + T)}. \tag{4.54}$$

Об'єднаємо (4.53) та (4.54):

$$\frac{\|C\vec{r}\|}{\|\vec{x}\|(1 + T)} \le \frac{\|\delta\vec{x}\|}{\|\vec{x}\|} \le \frac{\|C\vec{r}\|}{\|\vec{x}\|(1 - T)}. \tag{4.55}$$

Таким чином, оцінка похибки результатів, що отримані з умов (4.49), (4.51) та (4.55), буде вимагати $(n^3+2n^2)+(n^3+3n^2)+(2n^3+5n^2)$ додаткових операцій, що приблизно в 4 рази збільшує витрати часу на розв'язок вихідної системи.

4.6. Обернення матриць

Розглянемо рівняння, яке визначає матрицю X, що є оберненою до даної матриці A:

$$A \cdot X = E\,, \tag{4.56}$$

де E - одинична матриця. Нехай $\vec{x}^{(j)}$ та $\vec{y}^{(j)}$ - вектори, компоненти яких є елементами j-х стовпчиків матриць X та E, відповідно, тобто $\left(\vec{x}^{(j)}\right)_i = x_{ij}$, $\left(\vec{y}^{(j)}\right)_i = e_{ij} = \delta_{ij}$. Очевидно, що вони зв'язані співвідношенням

$$A\vec{x}^{(j)} = \vec{y}^{(j)}\,. \tag{4.57}$$

Таким чином, замість розв'язання матричного рівняння (4.56) можна розв'язувати сукупність систем лінійних рівнянь (4.57). Якщо при цьому використовувати метод виключень Гаусса, то необхідно тільки один раз виконати прямий хід і n разів обернений хід для різних правих частин $\vec{y}^{(j)}$. У результаті отримаємо метод обернення матриць *Жордана-Гаусса.*

Приклад 4. *Обернути матрицю в системі (4.23).*

$\lhd$ Для цієї матриці був отриманий розклад (4.7):

$$A = \begin{pmatrix} 1 & -2 & 3 \\ 2 & 3 & 4 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 7 & 0 \\ 1 & 3 & -\frac{8}{7} \end{pmatrix} \cdot \begin{pmatrix} 1 & -2 & 3 \\ 0 & 1 & -\frac{2}{7} \\ 0 & 0 & 1 \end{pmatrix}.$$

Тепер необхідно тричі розв'язати системи рівнянь типу (4.18) та (4.19).

$$\begin{array}{lllllllll} e_1^{(1)}=1, & e_2^{(1)}=-\frac{2}{7}, & e_3^{(1)}=\frac{1}{8}, & x_{31}=\frac{1}{8}, & x_{21}=-\frac{1}{4}, & x_{11}=\frac{1}{8}, \\ e_1^{(2)}=0, & e_2^{(2)}=\frac{1}{7}, & e_3^{(2)}=\frac{3}{8}, & x_{32}=\frac{3}{8}, & x_{22}=\frac{1}{4}, & x_{12}=-\frac{5}{8}, \\ e_1^{(3)}=0, & e_2^{(3)}=0, & e_3^{(3)}=-\frac{7}{8}, & x_{33}=-\frac{7}{8}, & x_{23}=-\frac{1}{4}, & x_{13}=\frac{17}{8}. \end{array}$$

Результат легко перевірити:

$$\begin{pmatrix} 1 & -2 & 3 \\ 2 & 3 & 4 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{8} & -\frac{5}{8} & \frac{17}{8} \\ -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ \frac{1}{8} & \frac{3}{8} & -\frac{7}{8} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Обернення матриць таким способом вимагає приблизно в 5 разів більше часу, ніж розв'язок одиничної системи рівнянь типу (4.1).

4.7. Пошук власних чисел і власних векторів

Обмежимось тут пошуком власних значень і власних векторів симетричних матриць. Розглянемо рівняння вигляду (4.2)

$$A\vec{x}_k = \lambda_k \vec{x}_k$$

де A - симетрична дійсна матриця. Відомо, що в цьому випадку власні числа λ_k - дійсні, а власні вектори $\{\vec{x}_k\}$ можна вважати стовпчиками деякої ортогональної матриці O, яка діагоналізує A:

$$O^T A O = \tilde{A}\,. \tag{4.58}$$

Тут $\tilde{A}$ - діагональна матриця.

Пошук власних чисел еквівалентний розв'язку алгебраїчного рівняння

$$\det(A - \lambda E) = 0\,, \tag{4.59}$$

при знаходженні коренів якого будуть суттєвими два моменти: власне розв'язок рівняння (4.59) і, найголовніше, знаходження коефіцієнтів цього рівняння, оскільки визначник містить $n!$ доданків. Тому ця задача завжди розв'язується ітераційними методами.

В усіх пакетах наукових програм, які реалізують методи лінійної алгебри, рівняння (4.2) та (4.58) розв'язані повністю, тому тут розглянемо тільки один із методів наближеної оцінки власних чисел та власних векторів, який широко використовується в квантово-механічній теорії збурень.

Нехай $A = R + T_0$, де $R_{kl} = \delta_{kl} A_{kk}$, $(T_0)_{kl} = (1 - \delta_{kl}) A_{kl}$, тобто вихідна матриця A представлена у вигляді суми діагональної матриці R і деякої матриці T_0, яку будемо вважати "збуренням". (Зазначимо, що на головній діагоналі матриці T_0 стоять нулі).

Визначимо нову матрицю $T = \dfrac{1}{\varepsilon} T_0$, де ε - "малий"параметр. Тоді рівняння (4.2) буде мати вигляд

$$(R + \varepsilon T)\vec{x}_k = \lambda_k \vec{x}_k \,, \quad k = \overline{1, n} \,. \tag{4.60}$$

Тут $\{\lambda_k\}$ - власні значення, що одночасно є діагональними елементами матриці $\tilde{A}$ (див. (4.58)).

Нехай $\vec{x}_k^{(0)}$ - власний вектор матриці R, що відповідає власному значенню μ_k незбуреної задачі, тобто

$$R\vec{x}_k^{(0)} = \mu_k \vec{x}_k^{(0)} = A_{kk}\vec{x}_k^{(0)} \,.$$

Будемо вважати вектори $\vec{x}_k^{(0)}$ ортонормованими

$$\left(\vec{x}_k^{(0)}, \vec{x}_l^{(0)}\right) = \delta_{kl} \,, \tag{4.61}$$

і такими, що мають вигляд

$$\left(\vec{x}_k^{(0)}\right)^T = (0, \ldots, 0, 1, 0, \ldots, 0) \,,$$

де тільки k-а компонента дорівнює одиниці.

Перехід від головних осей $\{\vec{x}_k^{(0)}\}$, в яких матриця R діагональна, до інших головних осей $\{\vec{x}_k\}$, в яких буде діагональною матриця A, здійснюється за допомогою деякого ортогонального перетворення O:

$$\vec{x}_k = \sum_l O_{kl}\vec{x}_l^{(0)} \,, \tag{4.62}$$

що робить систему векторів $\{\vec{x}_k\}$ також ортонормованою.

Використаємо метод теорії збурень, тобто розкладемо власні числа λ_k та матричні елементи O_{kl} в ряд по степенях параметра ε:

$$\begin{aligned} \lambda_k &= \lambda_k^{(0)} + \varepsilon\lambda_k^{(1)} + \varepsilon^2\lambda_k^{(2)} + \ldots, \\ O_{kl} &= O_{kl}^{(0)} + \varepsilon O_{kl}^{(1)} + \varepsilon^2 O_{kl}^{(2)} + \ldots. \end{aligned} \tag{4.63}$$

Підставимо (4.62) в (4.60)

$$(R+\varepsilon T)\sum_l O_{kl}\vec{x}_l^{(0)} = \lambda_k \sum_l O_{kl}\vec{x}_l^{(0)}$$

та скалярно помножимо це рівняння зліва на $\vec{x}_m^{(0)}$; тоді, враховуючи (4.61) будемо мати

$$R_{mm}O_{km} + \varepsilon\sum_l O_{kl}T_{ml} = \lambda_k O_{km}\,.$$

Тепер тут можна використати розклад (4.63)

$$\begin{aligned}\left(R_{mm}-\lambda_k^{(0)}-\varepsilon\lambda_k^{(1)}-\varepsilon^2\lambda_k^{(2)}-..\right)\left(O_{km}^{(0)}+\varepsilon O_{km}^{(1)}+\varepsilon^2 O_{km}^{(2)}+..\right)+\\ +\varepsilon\sum_l\left(O_{kl}^{(0)}+\varepsilon O_{kl}^{(1)}+\varepsilon^2 O_{kl}^{(2)}+\ldots\right)T_{ml}=0.\end{aligned} \tag{4.64}$$

Оскільки (4.64) повинно бути справедливим при довільних значеннях ε, то необхідно прирівняти нулеві коефіцієнти біля різних степенів ε.

Знайдемо послідовно найнижчі наближення. Коефіцієнт біля ε в нульовому степені

$$\left(R_{mm}-\lambda_k^{(0)}\right)O_{km}^{(0)} = 0\,.$$

Очевидно, що при $\varepsilon\to 0$ матриця O повинна прямувати до одиничної, отже,

$$O_{km}^{(0)} = \delta_{km} \quad \text{та} \quad \lambda_k^{(0)} = R_{kk} = A_{kk}\,. \tag{4.65}$$

Коефіцієнт біля ε в першому степені

$$\begin{aligned}&\left(R_{mm}-\lambda_k^{(0)}\right)O_{km}^{(1)} - \lambda_k^{(1)}O_{km}^{(0)} + \sum_l O_{kl}^{(0)}T_{ml} = 0 \implies\\ &\left(A_{mm}-A_{kk}\right)O_{km}^{(1)} - \lambda_k^{(1)}\delta_{km} + T_{mk} = 0\,,\end{aligned} \tag{4.66}$$

звідки

$$O_{km}^{(1)} = d_{km}T_{mk}\ (k\neq m) \quad \text{та} \quad \lambda_k^{(1)} = T_{kk} = 0\,, \tag{4.67}$$

де введене позначення $d_{km} = (A_{kk}-A_{mm})^{-1}$.

Далі, коефіцієнт біля ε в другому степені

$$\begin{aligned}&\left(R_{mm}-\lambda_k^{(0)}\right)O_{km}^{(2)} - \lambda_k^{(1)}O_{km}^{(1)} - \lambda_k^{(2)}O_{km}^{(0)} + \sum_l O_{kl}^{(1)}T_{ml} = 0\\ &\implies \left(A_{mm}-A_{kk}\right)O_{km}^{(2)} - \lambda_k^{(2)}\delta_{km} + \sum_l O_{kl}^{(1)}T_{ml} = 0\,,\end{aligned}$$

звідки

$$O^{(2)}_{km}=\sum_{l\neq k,m} d_{kl}d_{km}T_{lk}T_{ml}+d_{km}T_{mk}O^{(1)}_{kk}\ (k\neq m)\ \text{та}\ \lambda^{(2)}_k=\sum_{l\neq k} d_{kl}T^2_{kl}.$$

Цей процес можна продовжити і далі. Тут необхідно звернути увагу на два питання. По-перше, поправка першого порядку до власних чисел $\lambda^{(1)}_k$ дорівнює нулеві. Ця властивість широко використовується в задачах квантової механіки. По-друге, залишаються невизначеними діагональні значення поправок $O^{(1)}_{kk}$, $O^{(2)}_{kk}$ тощо.

Для визначення цих діагональних елементів скористуємось ортогональністю матриці O:

$$OO^T = O^TO = E \quad \Longrightarrow \quad \sum_l O_{kl}O_{ml} = \delta_{kl}\,.$$

На підставі розкладу (4.63) маємо

$$\begin{aligned}
&\sum_l \left(O^{(0)}_{kl}+\varepsilon O^{(1)}_{kl}+\varepsilon^2 O^{(2)}_{kl}+\ldots\right)\left(O^{(0)}_{ml}+\varepsilon O^{(1)}_{ml}+\varepsilon^2 O^{(2)}_{ml}+\ldots\right)=\\
&=\sum_l O^{(0)}_{kl}O^{(0)}_{ml}+\varepsilon\sum_l\left(O^{(1)}_{kl}O^{(0)}_{ml}+O^{(0)}_{kl}O^{(1)}_{ml}\right)+\\
&+\varepsilon^2\sum_l\left(O^{(2)}_{kl}O^{(0)}_{ml}+O^{(1)}_{kl}O^{(1)}_{ml}+O^{(0)}_{kl}O^{(2)}_{ml}\right)+\ldots=\delta_{km}\,.
\end{aligned}$$

Умова $\sum\limits_l O^{(0)}_{kl}O^{(0)}_{ml}=\delta_{km}$ виконується автоматично внаслідок (4.65). Далі, прирівнюючи нулеві коефіцієнти біля інших степенів ε, отримаємо

$$\sum_l\left(O^{(1)}_{kl}O^{(0)}_{ml}+O^{(0)}_{kl}O^{(1)}_{ml}\right)=0 \quad \Longrightarrow \quad O^{(1)}_{km}+O^{(1)}_{mk}=0\,.$$

При $k\neq m$ цей результат виходить із (4.67) в силу симетрії вихідної матриці A; при $k=m$ отримаємо оцінку діагонального елементу $O^{(1)}_{kk}=0$.

Далі

$$\sum_l\left(O^{(2)}_{kl}O^{(0)}_{ml}+O^{(1)}_{kl}O^{(1)}_{ml}+O^{(0)}_{kl}O^{(2)}_{ml}\right)+\ldots=O^{(2)}_{mk}+\sum_l O^{(1)}_{kl}O^{(1)}_{ml}+O^{(2)}_{km}=0,$$

що для $k=m$ дає $O^{(2)}_{kk}=-\dfrac{1}{2}\sum\limits_{l\neq k}\left(O^{(1)}_{kl}\right)^2$.

Таким чином, продовжуючи цю процедуру, можна отримати власні числа та власні вектори з визначеною точністю. Видно, що швидкість збіжності розкладу (4.63) визначається абсолютними значеннями d_{kl}, тобто відношенням недіагональних елементів до різниці між діагональними елементами матриці A. Виявляється, що і в тому випадку, коли деякі діагональні елементи вихідної матриці A співпадають, тобто матриця R є виродженою, можна побудувати аналогічну процедуру послідовних наближень.

Приклад 5. *Знайти власні вектори та власні числа до членів другого порядку включно матриці* $A = \begin{pmatrix} 5 & 1 \\ 1 & 10 \end{pmatrix}$.

$\lhd$ Послідовно знайдемо:

$$
\begin{array}{ll}
\lambda_1^{(0)} = 5\,, & \lambda_2^{(0)} = 10\,, \\
O_{11}^{(0)} = O_{22}^{(0)} = 1\,, & O_{12}^{(0)} = O_{21}^{(0)} = 0\,; \\
\lambda_1^{(1)} = 0\,, & \lambda_2^{(1)} = 0\,, \\
O_{11}^{(1)} = O_{22}^{(1)} = 0\,, & O_{12}^{(1)} = -O_{21}^{(1)} = -\dfrac{1}{\varepsilon}0.2\,; \\
\lambda_1^{(2)} = -\dfrac{1}{\varepsilon^2}0.2\,, & \lambda_2^{(2)} = \dfrac{1}{\varepsilon^2}0.2\,, \\
O_{11}^{(2)} = O_{22}^{(2)} = -\dfrac{1}{\varepsilon^2}0.02\,, & O_{12}^{(2)} = -O_{21}^{(2)} = 0\,; \\
\lambda_1 \approx \lambda_1^{(0)} + \varepsilon\lambda_1^{(1)} + \varepsilon^2\lambda_1^{(2)} = 4.8, & \lambda_2 \approx \lambda_2^{(0)} + \varepsilon\lambda_2^{(1)} + \varepsilon^2\lambda_2^{(2)} = 10.2, \\
O_{11} \approx 0.98\,, \quad O_{12} = \approx -0.2\,, & O_{21} \approx 0.2\,, \quad O_{22} \approx 0.98\,.
\end{array}
$$

Точні значення власних чисел є коренями секулярного (в даному випадку квадратного) рівняння

$$\det(A - \lambda E) = 0 \quad \Longrightarrow \quad (5 - \lambda)(10 - \lambda) - 1 = 0\,.$$

Звідси $\lambda_1 = 4.8075$, $\lambda_2 = 10.1925$.

Таким чином, отримані значення задовільно наближують власні числа, а відповідні власні вектори майже ортонормовані:

$$\begin{pmatrix} 0.98 & -0.20 \\ 0.20 & 0.98 \end{pmatrix} \begin{pmatrix} 0.98 & 0.20 \\ -0.20 & 0.98 \end{pmatrix} = \begin{pmatrix} 1.0004 & 0.0 \\ 0.0 & 1.0004 \end{pmatrix}.$$

$\rhd$

Глава 5

Методи чисельного диференціювання

На практиці часто виникає необхідність диференціювати функцію, що задана на дискретній множині значень аргументу (тобто, у вигляді таблиці). Перед тим, як нижче будуть отримані деякі загальні вирази, зауважимо, що в принципі вони можуть давати зовсім неправильні результати. Справа в тому, що наближеність ординат функції до інтерполяційного полінома на відрізку $[a, b]$ зовсім не гарантує зближення на цьому відрізку їх похідних. Тому краще уникати чисельного диференціювання. В будь-якому випадку відповідні методи потрібно використовувати з обережністю.

Загальні методи чисельного диференціювання можуть бути отримані в такий спосіб. Нехай для функції f, що задана таблично на відрізку $[a, b]$, потрібно знайти наближені значення похідних будь-якого порядку. Для цього функцію f замінюють на інтерполяційний поліном P:

$$y = f(x) = P(x) + R(x)\,, \tag{5.1}$$

де $R(x)$ – похибка. Диференціюючи (5.1) довільну кількість разів, отримаємо

$$\begin{aligned} y' &= f'(x) = P'(x) + R'(x)\,, \\ y'' &= f''(x) = P''(x) + R''(x)\,, \\ &\dots \end{aligned}$$

Проте абсолютну похибку фактично не завжди можна встановити.

Використаємо як $P(x)$ інтерполяційний поліном Лагранжа, побудований на n точках $x_1, x_2, \dots, x_n$; тоді $f(x)$=$P_{n-1}(x)$ +$R_n(x)$, де залишковий член визна-

чений в (2.10). У результаті отримаємо:

$$y' = f'(x) \approx P'_{n-1}(x) = \frac{d}{dx}\left[\Pi(x)\sum_{i=1}^{n}\frac{y_i}{(x-x_i)\Pi'(x_i)}\right] =$$
$$= \Pi(x)\sum_{i=1}^{n}\sum_{k=1,k\neq i}^{n}\frac{y_i}{(x-x_i)(x-x_k)\Pi'(x_i)}\,.$$

Спрямовуючи тепер $x \longrightarrow x_l$, отримаємо формулу чисельного диференціювання, яка часто використовується на практиці:

$$y'_l = \Pi'(x_l)\sum_{i=1,i\neq l}^{n}\frac{y_i}{(x_l-x_i)\Pi'(x_i)} + \sum_{i=1,i\neq l}^{n}\frac{y_l}{x_l-x_i} + \frac{f^{(n)}(\xi)}{n!}\Pi'(x_l)\,. \tag{5.2}$$

Якщо вузли $\{x_i\}$ є рівновіддаленими з кроком h, то (5.2) значно спрощується. Випишемо деякі окремі випадки:
n=3 (три вузли) –

$$\begin{aligned} y'_1 &= \frac{1}{2h}(-3y_1+4y_2-y_3) + \frac{h^2}{3}y^{(3)}\,,\\ y'_2 &= \frac{1}{2h}(-y_1+y_3) + \frac{h^2}{6}y^{(3)}\,,\\ y'_3 &= \frac{1}{2h}(y_1-4y_2+3y_3) + \frac{h^2}{3}y^{(3)}\,, \end{aligned} \tag{5.3}$$

$n = 7$ (сім вузлів) –

$$y'_1 = \frac{1}{60h}(-147y_1+360y_2-450y_3+400y_4-225y_5+72y_6-10y_7) + \frac{h^6y^{(7)}}{7}$$

...

$$y'_4 = \frac{1}{60h}(-y_1+9y_2-45y_3+45y_5-9y_6+y_7) + \frac{h^6}{140}y^{(7)}\,,$$

...

Якщо не враховувати похибки, то формулу (5.2) можна переписати у вигляді $\vec{y}' = A\vec{y}$, де $\vec{y}$ - вектор-стовпчик з координатами $y_1, \ldots, y_n$ і аналогічно для вектора $\vec{y}'$, а елементи матриці A розмірності $(n \times n)$ визначені в (5.2). Бачимо, що послідовне використання (5.2) дасть результат

$$\vec{y}^{(m)} = A\vec{y}^{(m-1)} = \ldots = A^m\vec{y}$$

для знаходження похідних вищого порядку. Безпосередньою перевіркою легко встановити, що A^m є нульовою матрицею, якщо $m \geq n$. Це є наслідком того, що інтерполяційний поліном має степінь $n-1$.

Як інтерполюючий поліном оберемо зараз поліном Ньютона P_{n-1} для інтерполяції вперед (2.17)

$$\begin{aligned} P_{n-1}(x) &= Q_{n-1}(t) = y_1 + \Delta y_1 \cdot t + \Delta^2 y_1 \cdot \frac{t(t-1)}{2!} + \\ &+ \ldots + \frac{\Delta^{n-1} y_1}{(n-1)!} t(t-1)\cdots(t-n+2)\,, \end{aligned} \tag{5.4}$$

де $t = (x - x_1)/h$. Розкладемо $Q(t)$ в ряд Тейлора в околі $t = 0$ без зазначення залишкового члена розкладу:

$$Q(t)=Q(0)+Q'(0)t+\frac{1}{2!}Q''(0)t^2+\ldots+\frac{1}{(n-1)!}Q^{(n-1)}(0)t^{n-1}. \tag{5.5}$$

Порівнюючи коефіцієнти в (5.4) і (5.5) при однакових степенях t, отримаємо

$$\begin{aligned} Q'(0) &= \Delta y_1 - \frac{1}{2}\Delta^2 y_1 + \frac{1}{3}\Delta^3 y_1 - \frac{1}{4}\Delta^4 y_1 + \ldots \\ Q''(0) &= \Delta^2 y_1 - \Delta^3 y_1 + \frac{11}{12}\Delta^4 y_1 - \frac{5}{6}\Delta^5 y_1 + \ldots \\ &\ldots \end{aligned}$$

Врахуємо, що $y_1' \approx \dfrac{dP_{n-1}}{dx}|_{x=x_1} = \dfrac{1}{h}Q'(0)$ тощо. В результаті

$$\begin{aligned} y_1' &= \frac{1}{h}\left(\Delta y_1 - \frac{1}{2}\Delta^2 y_1 + \frac{1}{3}\Delta^3 y_1 - \frac{1}{4}\Delta^4 y_1 + \ldots\right), \\ y_1'' &= \frac{1}{h^2}\left(\Delta^2 y_1 - \Delta^3 y_1 + \frac{11}{12}\Delta^4 y_1 - \frac{5}{6}\Delta^5 y_1 + \ldots\right), \\ &\ldots \end{aligned} \tag{5.6}$$

Це *різницеві формули чисельного диференціювання.*

Формули (5.6) можна переписати в компактному вигляді, якщо врахувати, що

$$\frac{d^r}{dx^r} \approx \left[\frac{1}{h}\ln(1+\Delta)\right]^r = \frac{1}{h^r}\left(\Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 - \ldots\right)^r.$$

Приклад 6. *Знайти похідні в точках* x_i $(i = \overline{1,5})$ *за допомогою* (5.3) *від функції* $\exp(x/10.)$, *що задана у вигляді таблиці.*

x_i	0.	1.	2.	3.	4.
y_i	1.	1.105170	1.221402	1.349858	1.491824

◁ Бачимо, що найменшу похибку має друге рівняння з рівнянь (5.3), яке будемо використовувати для точок з номерами $i = 2, 3, 4$. Для знаходження y_1' та y_5' використаємо перше та останнє з рівнянь (5.3) відповідно. Для порівняння в дужках наведені точні значення похідної від цієї функції у відповідних вузлах.

$$
\begin{aligned}
{y_1}' &= \frac{1}{2h}\left(-3y_1 + 4y_2 - y_3\right) &&= 0.099639 \quad (0.100000)\,,\\
{y_2}' &= \frac{1}{2h}\left(-y_1 + y_3\right) &&= 0.110701 \quad (0.110517)\,,\\
{y_3}' &= \frac{1}{2h}\left(-y_2 + y_4\right) &&= 0.122344 \quad (0.122140)\,,\\
{y_4}' &= \frac{1}{2h}\left(-y_3 + y_5\right) &&= 0.135211 \quad (0.134986)\,,\\
{y_5}' &= \frac{1}{2h}\left(y_3 - 4y_4 + 3y_5\right) &&= 0.148721 \quad (0.149182)\,.
\end{aligned}
$$

▷

Глава 6

Методи чисельного інтегрування

Чисельне інтегрування – це обчислення визначеного інтеграла $\int\limits_a^b f(x)dx$ по ряду чисельних значень підінтегральної функції $y_i = f(x_i),\ i = \overline{1,n}$.

Задача чисельного інтегрування виникає в таких випадках:
1) $f(x)$ задана у вигляді таблиці;
2) $f(x)$ не має первісної у класі елементарних функцій;
3) $f(x)$ невідома (приклад - інтегральне рівняння).

Формули наближеного обчислення визначених інтегралів часто називають *квадратурними формулами*. В загальному випадку вони отримуються заміною підінтегральної функції на деякий інтерполяційний поліном.

6.1. Квадратурні формули Ньютона-Котеса

Розглянемо інтеграл

$$I = \int\limits_a^b f(x)dx\,. \tag{6.1}$$

Відрізок $[a,b]$ розіб'ємо на $n-1$ однакових частин довжиною $h = \dfrac{b-a}{n-1}$ кожна. Побудуємо інтерполяційний поліном Лагранжа по точках $x_1 = a, x_2 = a + h, \ldots, x_n = a + (n-1)h = b$ і в (6.1) зробимо заміну $f(x) \approx P_{n-1}(x)$. Тоді

$$I \approx \int\limits_a^b P_{n-1}(x)dx = \int\limits_a^b \sum_{i=1}^{n} f(x_i) \prod_{j=1, j\neq i}^{n} \frac{(x-x_j)}{(x_i-x_j)}dx = \sum_{i=1}^{n} \omega_i y_i. \tag{6.2}$$

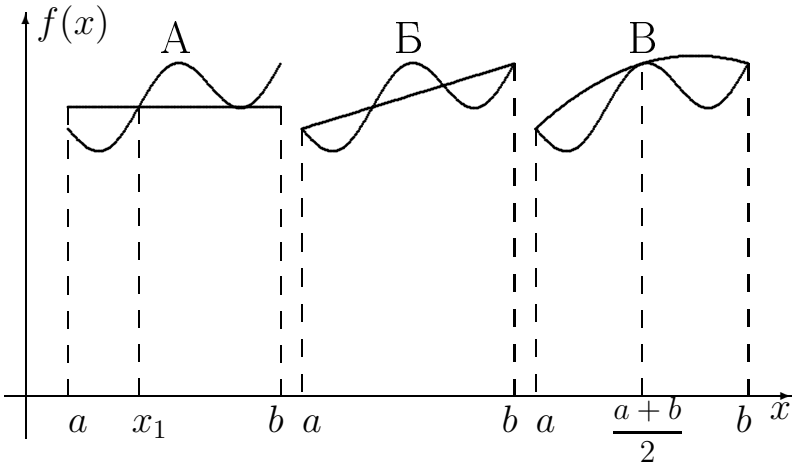

Рис. 6.1. Методи прямокутників (А), трапецій (Б), Сімпсона (В)

Тут $\{x_i\}$ – *вузли* та $\{\omega_i\}$ – *ваги квадратурної формули Ньютона-Котеса*, які визначаються за формулою

$$\omega_i = \int_a^b \prod_{j=1, j\neq i}^{n} \frac{(x - x_j)}{(x_i - x_j)} dx \,. \tag{6.3}$$

Розглянемо деякі окремі випадки:
1) $n = 1,\ \omega_1 = b - a$;

$$\int_a^b f(x)dx \approx (b-a)f(x_1), \quad \text{де} \quad x_1 \in [a, b] \,. \tag{6.4}$$

Нижче буде показано, що, хоча положення точки x_1 довільне, краще за все обирати її посередині інтервалу $[a, b]$. Вираз (6.4) називається *формулою прямокутників*, оскільки відповідає заміні підінтегральної функції $f(x)$ постійною величиною $f(x_1)$ (рис. 6.1А).
2) $n = 2,\ \omega_1 = \omega_2 = (b - a)/2$;

$$\int_a^b f(x)dx \approx \frac{b-a}{2} \left[f(a) + f(b)\right] \,. \tag{6.5}$$

Вираз (6.5) називається *формулою трапецій* і відповідає заміні підінтегральної функції $f(x)$ прямою лінією (рис. 6.1Б).

3) $n = 3,\ \omega_1 = \omega_3 = (b-a)/6,\ \omega_2 = 2(b-a)/3;$

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left[f(a) + 4f(\frac{a+b}{2}) + f(b)\right]. \tag{6.6}$$

Це - *формула Сімпсона* (параболічних трапецій) (рис. 6.1В).

Аналогічно можна отримати формули і більш високого порядку, хоча вони практично не використовуються.

Заміна змінних $x = a + (b-a)z$ приводить до еквівалентного інтегралу

$$\int_a^b f(x)dx = \int_0^1 \Phi(z)dz, \quad \Phi(z) = f\left(a + (b-a)z\right).$$

Вузлам x_i відрізку $[a,b]$ відповідають вузли $z_i = \dfrac{i-1}{n-1}$ на інтервалі $[0,1]$, причому $y_i = f(x_i) = \Phi(z_i)$. Внаслідок (6.2) та (6.3) маємо

$$\int_a^b f(x)dx \approx (b-a)\sum_{i=1}^{n} \overline{\omega}_i y_i\,, \tag{6.7}$$

де вага

$$\overline{\omega}_i = \int_0^1 \prod_{j=1, j\neq i}^{n} \frac{(n-1)z - j}{i-j}dx$$

не залежить від відрізку $[a,b]$.

Якщо вихідний відрізок $[a,b]$ розділити на m інтервалів довжиною $(b-a)/m$ і на кожному з них застосувати формули (6.5) або (6.6), то отримаємо їх узагальнення:

$$\int_a^b f(x)dx \approx \frac{b-a}{2m}\left[y_1 + 2y_2 + 2y_3 + \ldots + 2y_m + y_{m+1}\right] \tag{6.8}$$

– *формула трапецій*;

$$\begin{aligned}\int_a^b f(x)dx \ \approx\ & \frac{b-a}{3m}\left[y_1 + 4y_2 + 2y_3 + 4y_4 + \ldots\right. \\ & \left. + 2y_{m-1} + 4y_m + y_{m+1}\right]\end{aligned} \tag{6.9}$$

– *формула Сімпсона* (тут m повинно бути парним).

6.2. Оцінка похибки

Формули Ньютона-Котеса (6.2), (6.7) базуються на заміні підінтегральної функції інтерполяційним поліномом Лагранжа. Це дає змогу оцінити похибку цих квадратурних формул. Дійсно, оскільки $f(x) = P_{n-1}(x) + R_n(x)$, то похибка при обчисленні інтеграла дорівнює

$$R_n[f] = \int_a^b f(x)dx - \int_a^b P_{n-1}(x)dx = \int_a^b \frac{f^{(n)}(\xi)}{n!}\Pi(x)dx\,.$$

Тоді

$$|R_n[f]| \le \max_{x\in[a,b]} \left|f^{(n)}(x)\right| \cdot \frac{1}{n!} \int_a^b |\Pi(x)|\,dx\,,$$

де $\Pi(x) = (x - x_1)(x - x_2)\cdots(x - x_n)$. Тут зручно перейти до інтервалу $[0, 1]$, тоді

$$|R_n[f]| \le \max_{x\in[a,b]} \left|f^{(n)}(x)\right| \cdot \frac{(b-a)^{n+1}}{n!} \int_0^1 |\omega_n(t)|\,dt\,,$$

де $\omega(t) = (t - 0)(t - \frac{1}{n-1})\cdots(t - \frac{n-1}{n-1})$.

Зокрема, це дає такі оцінки:

формула прямокутників ($n = 1,\ \omega_1(t) = t - \frac{1}{2}$) –

$$|R_1[f]| \le \max_{x\in[a,b]} |f'(x)| \cdot \frac{(b-a)^2}{4}\,; \tag{6.10}$$

формула трапецій ($n = 2,\ \omega_2(t) = t(t - 1)$) –

$$|R_2[f]| \le \max_{x\in[a,b]} |f''(x)| \cdot \frac{(b-a)^3}{12}\,; \tag{6.11}$$

формула Сімпсона ($n = 3,\ \omega_3(t) = t(t - \frac{1}{2})(t - 1)$) –

$$|R_3[f]| \le \max_{x\in[a,b]} \left|f^{(3)}(x)\right| \cdot \frac{(b-a)^4}{192}\,. \tag{6.12}$$

Виявляється, що оцінки (6.10) і (6.12) дещо завищені. Справа в тому, що, наприклад, формула Сімпсона (6.6) справедлива не тільки для довільних поліномів другої степені, але і для довільних поліномів третьої степені, в чому не важко переконатися безпосередньою перевіркою. Це є проявом більш загальної властивості квадратурних формул Ньютона-Котеса – якщо вузли розміщенні симетрично відносно середини інтервалу $[a,b]$, то ваги, що відповідають симетричним вузлам, збігаються між собою. Звідси випливає, що ці квадратурні формули є точними і для поліномів степені n (якщо n непарне). Щоб отримати відповідну оцінку похибки, потрібно підінтегральну функцію замінити інтерполяційним поліномом Лагранжа степені n, в якому точка $x = (a+b)/2$ є двократним вузлом інтерполяції.

Зокрема для *формули прямокутника* отримаємо:

$$\begin{aligned} |R_1[f]| &\le \max_{x\in[a,b]} |f''(x)| \cdot \frac{(b-a)^3}{2!} \int\limits_0^1 \left|(t-\tfrac{1}{2})\right|^2 dt = \\ &= \max_{x\in[a,b]} |f''(x)| \cdot \frac{(b-a)^3}{24}\,, \end{aligned}$$

і аналогічно для *формули Сімпсона*

$$\begin{aligned} |R_3[f]| &\le \max_{x\in[a,b]} \left|f^{(4)}(x)\right| \cdot \frac{(b-a)^5}{4!} \int\limits_0^1 \left|t(t-\tfrac{1}{2})^2(t-1)\right| dt = \\ &= \max_{x\in[a,b]} \left|f^{(4)}(x)\right| \cdot \frac{(b-a)^5}{2880}\,. \end{aligned} \tag{6.13}$$

Порівняння формул (6.8), (6.11) та (6.9), (6.13) пояснює причину популярності формули Сімпсона: практично при тій же кількості операцій точність тут значно вища.

6.3. Загальна постановка задачі про квадратури

В загальному випадку по аналогії з (6.2) можна записати:

$$I = \int\limits_a^b f(x)dx \approx \sum_{i=1}^{n} \omega_i f(x_i)\,, \tag{6.14}$$

де ω_i - ваги (що не обов'язково визначені за допомогою (6.3)) та x_i - вузли квадратурної формули.

Постає питання: як при практично однаковій кількості обчислювальної роботи, тобто при однаковій кількості вузлів n, досягти більшої точності формули типу (6.14)?
1. Цього можна досягти за рахунок зміни ваг ω_i, як в цьому ми переконались вище при переході від формули прямокутників до формули трапецій і далі до формули Сімпсона.
2. Цього ж можна досягти за рахунок вдалого вибору вузлів квадратурної формули. Точки, в яких обчислюються значення підінтегральної функції для формули трапецій, парабол тощо, визначаються тим, що відрізок інтегрування $[a, b]$ поділяється на n рівних частин. Але такий розподіл не завжди є вигідним. На тих ділянках, де функція змінюється повільно, точки ділення можна обирати на більшій відстані, а на ділянках із швидкою зміною – густішими. Далі розглянемо реалізацію цієї ідеї (формули Чебишева), а потім одночасний оптимальний вибір вузлів і ваг (формули Гаусса).

Очевидно, що замість довільного інтервалу $[a, b]$ можна розглядати інтервал $[-1, 1]$. Дійсно, за допомогою заміни $x = \dfrac{a+b}{2} + \dfrac{b-a}{2}t$ завжди має місце співвідношення

$$\int_a^b f(x)dx = \frac{b-a}{2} \int_{-1}^{1} f\left(\frac{a+b}{2} + \frac{b-a}{2}t\right) dt\,. \tag{6.15}$$

Тому надалі ми можемо обмежитися розглядом квадратурних формул на інтервалі $[-1, 1]$.

6.4. Формули Чебишева

Нехай значення $y_i = f(x_i)$, що входять в квадратурну формулу (6.14), знаходяться в результаті вимірів і містять в собі випадкові похибки. Припустимо, що всі y_i одержані з однаковою точністю (наприклад, на одній експериментальній установці). Тоді значення суми (6.14) теж буде містити в собі випадкову похибку. Виберемо ваги ω_i таким чином, щоб величина $\sum_i \omega_i y_i$ мала найменшу квадратичну похибку. При цьому додатково будемо вимагати, щоб формула

(6.14) була точною при $f(x) \equiv 1$, тобто (на інтервалі $[-1, 1]$) маємо умову:

$$\sum_{i=1}^{n} \omega_i = 2\,. \tag{6.16}$$

Розглянемо суму $I = \omega_1 y_1 + \omega_2 y_2 + \ldots + \omega_n y_n$. Якщо величини y_i мають середньоквадратичні похибки m_i, то середньоквадратична похибка цієї суми дорівнює

$$m_I = \sqrt{\omega_1^2 m_1^2 + \omega_2^2 m_2^2 + \ldots + \omega_n^2 m_n^2}\,.$$

Тоді у випадку $m_i \equiv m$ необхідно шукати мінімум величини $\omega_1^2 + \omega_2^2 + \ldots + \omega_n^2$ при додатковій умові (6.16). Для цього використаємо метод невизначених множників Лагранжа.

Будемо шукати мінімум величини

$$\Phi = \omega_1^2 + \omega_2^2 + \ldots + \omega_n^2 + \alpha \left(\sum_{i=1}^{n} \omega_i - 2 \right)\,,$$

що приводить до системи рівнянь

$$\frac{\partial \Phi}{\partial \alpha} = 0\,, \quad \frac{\partial \Phi}{\partial \omega_i} = 0,\ i = \overline{1, n}\,,$$

або $2\omega_i + \alpha = 0$, тобто всі ваги задовольняють одне і те ж рівняння, а отже рівні між собою.

Наведені міркування показують, що вибір однакових ваг мінімізує середню квадратичну похибку суми. В результаті приходимо до *квадратурних формул Чебишева*:

$$\int_{-1}^{1} f(x)dx \approx \omega \sum_{i=1}^{n} f(x_i)\,. \tag{6.17}$$

Формула (6.17) містить $n + 1$ параметрів $\{\omega, x_1, x_2, \ldots, x_n\}$, за допомогою яких можна добитися того, щоб ця рівність була точною для довільних багаточленів степеня $\leq n$. Зокрема, ці багаточлени можна обирати у вигляді x^k, $k = \overline{0, n}$. У результаті отримаємо систему рівнянь відносно ваги ω та вузлів x_i:

$$\int_{-1}^{1} 1 \cdot dx = \omega \sum_{i=1}^{n} 1 = 2\,,$$

звідки $\omega = \dfrac{2}{n}$, та

$$\begin{aligned}
\int_{-1}^{1} x \cdot dx &= \omega \sum_{i=1}^{n} x_i = 0\,, \\
\int_{-1}^{1} x^2 \cdot dx &= \omega \sum_{i=1}^{n} x_i^2 = \frac{2}{3}\,, \\
&\dots \\
\int_{-1}^{1} x^k \cdot dx &= \omega \sum_{i=1}^{n} x_i^k = \begin{cases} 0 & , \ k - \text{непарне}; \\ \dfrac{2}{k+1} & , \ k - \text{парне}; \end{cases} \\
&\dots
\end{aligned} \tag{6.18}$$

Систему рівнянь (6.18) можна привести до одного алгебраїчного рівняння n-го степеня, корені якого є вузлами квадратурної формули Чебишева. Як було показано, система (6.18) має дійсні розв'язки при $n = 1, 2, 3, 4, 5, 6, 7, 9$; при $n = 8$ і $n \geq 10$ серед коренів обов'язково будуть комплексні числа.

Приклад 7. *Розв'яжемо цю задачу для* $n = 2,\ \omega = 1$.

$\lhd$ Система (6.18) набуває вигляду

$$x_1 + x_2 = 0\,,\ x_1^2 + x_2^2 = \frac{2}{3}\,.$$

Звідси $x_1 = -x_2 = -\sqrt{\frac{1}{3}} = -0.577350$. $\rhd$

Наведемо також корисну таблицю для $n = 9,\ \omega = \dfrac{2}{9}$:

$$\begin{aligned}
x_1 &= -x_9 = -0.911589\,, \\
x_2 &= -x_8 = -0.601019\,, \\
x_3 &= -x_7 = -0.528762\,, \\
x_4 &= -x_6 = -0.167906\,, \\
x_5 &= 0\,.
\end{aligned}$$

6.5. Квадратурні формули Гаусса

Квадратурна формула

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} \omega_i f(x_i) \tag{6.19}$$

при фіксованій кількості вузлів n містить $2n$ параметрів ω_i та x_i. Вище було показано, що можна підвищувати точність (6.19), якщо змінювати ваги ω_i при рівновіддалених вузлах x_i (формули Ньютона-Котеса), або змінювати розташування вузлів x_i при однакових значеннях ваг ω_i (формули Чебишева). Гаусс запропонував параметри ω_i та x_i вважати вільними і знаходити їх з умови, що формула (6.19) буде точною для довільного алгебраїчного полінома степеня $\leq 2n-1$.

Таким чином, для довільної функції

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_{2n-1} x^{2n-1} \tag{6.20}$$

повинна виконуватися рівність (без зменшення загальності перейдемо до інтервалу $[-1, 1]$)

$$\int_{-1}^{1} f(x)dx = \sum_{i=1}^{n} \omega_i f(x_i)\,. \tag{6.21}$$

Функцію f вигляду (6.20) завжди можна подати у вигляді

$$f_{2n-1}(x) = P_n(x)Q_{n-1}(x) + R_{n-1}(x)\,,$$

де $P_n(x)$ – деякий поліном степеня n, корені якого будуть слугувати вузлами квадратурної формули Гаусса, тобто

$$P_n(x) = (x - x_1)(x - x_2)\cdots(x - x_n)\,, \tag{6.22}$$

а Q_{n-1} та R_{n-1} – деякі поліноми степеня не вище ніж $n-1$ (частка і залишок).

Очевидно, що

$$f_{2n-1}(x_1) = R_{n-1}(x_1), \ldots, f_{2n-1}(x_n) = R_{n-1}(x_n)\,.$$

Тому внаслідок (6.21) маємо

$$\int_{-1}^{1} f(x)dx = \int_{-1}^{1} P_n(x)Q_{n-1}(x)dx + \int_{-1}^{1} R_{n-1}(x)dx = \sum_{i=1}^{n} \omega_i R_{n-1}(x_i),$$

звідки випливає, що

$$\int_{-1}^{1} P_n(x)Q_{n-1}(x)dx = 0\,. \tag{6.23}$$

Оскільки рівність (6.21) повинна бути справедливою для будь-якого полінома f степеня $2n-1$, то умова (6.23) повинна виконуватися для будь-якого полінома Q степеня $\le n-1$, в тому числі для $Q(x) = 1, x, x^2, \ldots, x^{n-1}$. Це приводить до системи рівнянь для визначення вузлів x_i:

$$\int_{-1}^{1} P_n(x)x^k dx = 0\,, \quad k = \overline{0, n-1}\,. \tag{6.24}$$

Після знаходження вузлів x_i ваги ω_i знаходяться відповідно до (6.3), тобто

$$\omega_i = \int_{-1}^{1} \prod_{j=1, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} dx\,. \tag{6.25}$$

Приклад 8. *Розглянемо випадок з* $n = 2$.

$\lhd$ Для визначення вузлів на підставі (6.24) та (6.25) маємо систему рівнянь:

$$\int_{-1}^{1} (x - x_1)(x - x_2)dx = 0\,, \quad \int_{-1}^{1} x(x - x_1)(x - x_2)dx = 0\,,$$

або

$$x_1 x_2 = -\frac{1}{3}\,, \quad x_1 + x_2 = 0\,.$$

Звідси

$$\begin{aligned} x_1 &= -\sqrt{\frac{1}{3}}\,, & x_2 &= \sqrt{\frac{1}{3}}\,, \\ \omega_1 &= \int_{-1}^{1} \frac{x - \frac{1}{\sqrt{3}}}{-\frac{1}{\sqrt{3}} - \frac{1}{\sqrt{3}}} dx = 1\,, & \omega_2 &= \int_{-1}^{1} \frac{x + \frac{1}{\sqrt{3}}}{\frac{1}{\sqrt{3}} + \frac{1}{\sqrt{3}}} dx = 1\,. \end{aligned} \tag{6.26}$$

$\triangleright$

Проте при збільшенні n безпосереднє використання формул (6.24) та (6.25) для визначення вузлів і ваг стає затрудненим внаслідок того, що (6.24) являє собою систему n нелінійних рівнянь, яка зводиться до одного рівняння степеня n відносно однієї невідомої. І отримання цього рівняння, і його подальший розв'язок приводять до значних труднощів. Тому ми розглянемо тут інший спосіб розв'язку цієї задачі.

Багаточлен n-ого степеня $P_n(x)$ можна розглядати як n-у похідну від деякого багаточлена $F_{2n}(x)$ степеня $2n$. Для однозначного визначення коефіцієнтів цього багаточлена необхідно поставити вимоги щодо виконання n додаткових умов; зокрема, можна вимагати, щоб виконувалися такі умови:

$$F_{2n}(-1) = F'_{2n}(-1) = \ldots = F^{(n-1)}_{2n}(-1) = 0\,. \tag{6.27}$$

Тоді, інтегруючи n разів по частинах вираз (6.23), отримаємо

$$\begin{aligned}
&\int_{-1}^{1} F^{(n)}_{2n}(x) Q_{n-1}(x)dx = \\
&= \left[F^{(n-1)}_{2n}(x)Q_{n-1}(x) - F^{(n-2)}_{2n}(x)Q'_{n-1}(x) + \ldots\right]\Big|^{1}_{-1} \pm \\
&\pm \int_{-1}^{1} F_{2n}(x) Q^{(n)}_{n-1}(x)dx = 0\,.
\end{aligned}$$

Зважаючи на те, що $Q^{(n)}_{n-1} \equiv 0$, і на умову (6.27), а також на довільність полінома $Q_{n-1}(x)$, приходимо до таких додаткових умов:

$$F_{2n}(1) = F'_{2n}(1) = \ldots = F^{(n-1)}_{2n}(1) = 0\,.$$

Таким чином, точки $x = -1$ та $x = 1$ є коренями полінома $F_{2n}(x)$ n-ої кратності:

$$F_{2n}(x) = C(x+1)^n(x-1)^n = c(x^2-1)^n\,,$$

де C - довільна стала. Якщо, наприклад, $C = \dfrac{1}{2^n n!}$, то $P_n(x)$ - поліном Лежандра:

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n\,.$$

Ці поліноми є ортогональними на інтервалі $[-1,1]$:

$$\int_{-1}^{1} P_m(x)P_n(x)dx = \frac{2}{2n+1}\delta_{mn}\,. \tag{6.28}$$

Відповідно до (6.25), (6.22) та (2.8) ваги мають вигляд:

$$\omega_i = \frac{1}{C\cdot\prod\limits_{k\neq i}(x_i - x_k)}\int_{-1}^{1}\frac{C\cdot\prod\limits_{k=1}^{n}(x-x_k)}{x-x_i}dx = \frac{1}{P_n'(x_i)}\int_{-1}^{1}\frac{P_n(x)}{x-x_i}dx. \tag{6.29}$$

Для обчислення інтеграла в правій частині (6.29) використаємо тотожність Кристофеля-Дарбу

$$(x-t)\sum_{k=0}^{n}P_k(x)P_k(t) = \frac{a_n}{a_{n+1}}\left[P_{n+1}(x)P_n(t) - P_n(x)P_{n+1}(t)\right]\,,$$

яка у випадку $t = x_i$ (корінь P_n) набуває вигляду:

$$\sum_{k=0}^{n}P_k(x)P_k(x_i) = -\frac{a_n}{a_{n+1}}\frac{P_n(x)P_{n+1}(x_i)}{x-x_i}\,. \tag{6.30}$$

Тут a_n, a_{n+1} - коефіцієнти біля старших степенів у P_n, P_{n+1} відповідно.

Проінтегруємо (6.30) на інтервалі $[-1,1]$ і врахуємо (6.28):

$$2 = -\frac{a_n}{a_{n+1}}P_{n+1}(x_i)\int_{-1}^{1}\frac{P_n(x)}{x-x_i}dx\,.$$

Звідси

$$\omega_i = -\frac{a_{n+1}}{a_n}\frac{2}{P_n'(x_i)P_{n+1}(x_i)}.$$

Якщо тепер скористатися відомим співвідношенням для поліномів Лежандра $-(n+1)P_{n+1}(x_i) = (1-x_i^2)P_n'(x_i)$, то

$$\omega_i = \frac{2}{(1-x_i^2)\left[P_n'(x_i)\right]^2}\,. \tag{6.31}$$

Приклад 9. *Розглянемо випадок з* $n = 2$.

$\lhd$

$$P_2(x) = \frac{1}{2^2 2!}\frac{d^2}{dx^2}(x^2-1)^2 = \frac{1}{2}(3x^2-1)\,, \qquad P_2'(x) = 3x\,,$$
$$x_1 = -\frac{1}{\sqrt{3}} = -0.5773503\,,\ x_2 = 0.5773503\,,\ \ \omega_1 = \omega_2 = 1\,.$$

Як і було очікувано, результат збігається з (6.26).

Обчислимо згідно формулі Гаусса з $n = 2$ значення такого інтеграла:

$$\int_0^1 e^x dx = e - 1 = 1.718281828\ldots \tag{6.32}$$

Згідно з (6.15), (6.26) маємо:

$$\int_0^1 e^x dx = 0.5\int_{-1}^1 \exp(0.5+0.5t)dt \approx 0.5\,[\exp(0.5-0.5\cdot 0.5773503)+$$
$$+\exp(0.5+0.5\cdot 0.5773503)] = 1.717896\,.$$

Результат досить близький до точного значення (6.32), беручи до уваги, що розрахунки проводилися тільки по двох точках. $\rhd$

Для оцінки похибки квадратурної формули Гаусса необхідно знайти похибку, яка виникає при заміні підінтегральної функції $f(x)$ деяким інтерполяційним поліномом $H_{2n-1}(x)$ степеня $2n-1$, що побудований по n вузлах. Для того, щоб цей поліном побудувати однозначно, необхідно задати $2n$ умов, зокрема, можна вимагати виконання таких рівностей:

$$H_{2n-1}(x_i) = f(x_i)\,, \quad H'_{2n-1}(x_i) = f'(x_i)\,, \quad i = \overline{1,n}\,.$$

Повторюючи міркування розділу 2, легко отримати оцінку похибки інтерполяції

$$R_n(x) = \frac{f^{(2n)}(\xi)}{(2n)!}\Pi^2(x)\,,$$

де ξ - деяка точка, яка лежить на інтервалі, що містить вузли x_i та точку x.

Тоді для оцінки похибки формули Гаусса будемо мати вираз

$$R_n[f] = \frac{f^{(2n)}(\xi)}{(2n)!}\int\limits_{-1}^{1}[(x-x_1)\dots(x-x_n)]^2\,dx = \\ = f^{(2n)}(\xi)\frac{2^{n+1}}{(2n+1)!}\left[\frac{(n!)^2}{(2n)!}\right]^2, \quad \xi\in[-1,1]\,.$$

Або для довільного відрізку $[a,b]$:

$$R_n[f] = f^{(2n)}(\xi)\frac{(n!)^4}{[(2n)!]^3}\frac{(b-a)^{2n+1}}{2n+1}, \quad \xi\in[a,b]\,.$$

Зокрема, при $n=2$

$$R_n[f] = f^{(4)}(\xi)\frac{(b-a)^5}{4320},$$

що краще за оцінку похибки для формули Сімпсона (6.13). Наприклад, для інтеграла (6.32) при n=2 оцінка дорівнює 0.000629.

6.6. Обчислення невласних інтегралів

При проведенні досліджень часто виникає задача про обчислення інтегралів на нескінченному відрізку, наприклад $[0,\infty)$. На кожному скінченному відрізку $0\le x\le b<\infty$ можна побудувати поліном $P(x)$, що рівномірно наближує $f(x)$ з наперед заданою точністю. Але $P(x)$ може не давати рівномірного наближення $f(x)$ на всій напіввісі і різниця $f-P$ може мати великі значення при великих x. У цьому випадку інтеграл $\int\limits_0^\infty f(x)dx$ можна замінити на інтеграл $\int\limits_0^\infty p(x)\cdot\tilde{f}(x)dx$, де $p(x)$ - деякий ваговий множник, а

$$\tilde{f}(x) = \frac{f(x)}{p(x)}\,.$$

Тоді, якщо $p(x)$ спадає достатньо швидко, може статися, що величина інтеграла $\int\limits_0^\infty\left|p(x)\left(\tilde{f}(x)-P(x)\right)\right|dx$ може стати такою малою, що нею можна знехтувати. Це означає, що поліноми $P(x)$ повинні наближати функцію $\tilde{f}(x)$, а не $f(x)$.

Нехай, наприклад, $f(x) = \varphi(x)e^{-x}$. Тоді можна обрати $p(x) = e^{-x}$ та $\tilde{f}(x) = \varphi(x)$ і в результаті шукати квадратурну формулу для інтеграла

$$\int_0^\infty f(x)dx = \int_0^\infty p(x)\varphi(x)dx \approx \int_0^\infty e^{-x}P(x)dx\,, \tag{6.33}$$

де як інтерполяційні поліноми будемо використовувати поліноми Лягерра:

$$\begin{aligned} &P(x) \longrightarrow L_n(x) = (-1)^n e^x \frac{d^n}{dx^n}\left(x^n e^{-x}\right), \\ &L_1(x) = x - 1\,, \\ &L_2(x) = x^2 - 4x + 2\,, \\ &\dots \end{aligned} \tag{6.34}$$

Тоді можна записати

$$\int_0^\infty e^{-x}\varphi(x)dx \approx \sum_{i=1}^n \omega_i \varphi(x_i)\,, \tag{6.35}$$

де x_i (корені $L_n(x)$) і ω_i - вузли і ваги *квадратурної формули Чебишева-Лягерра*:

$$\omega_i = \frac{(n!)^2}{x_i\left[L_n'(x_i)\right]^2}\,. \tag{6.36}$$

Приклад 10. *Обчислити інтеграл*

$$I = \int_0^\infty \frac{x}{\operatorname{sh} x}dx = \frac{\pi^2}{4} \approx 2.4674011\dots$$

за допомогою (6.35) *при* $n = 3$.

◁ Відповідно до (6.34) та (6.36) маємо:

$$\begin{aligned} L_3(x) = x^3 - 9x^2 + 18x - 6\,, &\qquad L_3'(x) = 3x^2 - 18x + 18\,, \\ x_1 = 0.4157746\,, &\qquad \omega_1 = 0.7110930\,, \\ x_2 = 2.2942804\,, &\qquad \omega_2 = 0.2785177\,, \\ x_3 = 6.2899451\,, &\qquad \omega_3 = 0.01038926\,, \\ I = \int_0^\infty \frac{2x}{e^x - e^{-x}}dx &= \int_0^\infty e^{-x}\frac{2x}{1 - e^{-2x}}dx \approx \\ \approx 2\sum_{i=1}^3 \omega_i \frac{x_i}{1 - \exp(-2x_i)} &\approx 2.4690772\,. \end{aligned}$$

▷

Для обчислення інтегралів на всій вісі $(-\infty, \infty)$ як вагову функцію можна використовувати $p(x) = e^{-x^2}$; тоді аналогічно до (6.33) отримаємо

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} e^{-x^2}\tilde{f}(x)dx \approx \int_{0}^{\infty} e^{-x^2}P(x)dx\,, \tag{6.37}$$

де вузли інтерполяційного полінома дорівнюють кореням полінома Ерміта

$$\begin{aligned} &P(x) \longrightarrow H_n(x)(-1)^n e^{x^2}\frac{d^n}{dx^n}\left(e^{-x^2}\right), \\ &H_1(x) = 2x, \\ &H_2(x) = 4x^2 - 2\,, \\ &\dots \end{aligned} \tag{6.38}$$

У результаті

$$\int_{-\infty}^{\infty} e^{-x^2}\varphi(x)dx \approx \sum_{i=1}^{n}\omega_i\varphi(x_i)\,, \tag{6.39}$$

де x_i (корені $H_n(x)$) і ω_i - вузли і ваги *квадратурної формули Чебишева-Ерміта*:

$$\omega_i = \frac{2^{n+1}n!\sqrt{\pi}}{[H_n'(x_i)]^2}\,. \tag{6.40}$$

Приклад 11. *Обчислити інтеграл*

$$I = \int_{-\infty}^{\infty} e^{-x^2}\left(e^{-\frac{1}{x^2}} - 1\right)dx = \sqrt{\pi}\left(\frac{1}{e^2} - 1\right) \approx -1.532578\dots$$

за допомогою (6.39) *при* $n = 3$.

◁ Відповідно до (6.38) та (6.40) маємо:

$$\begin{aligned} &H_3(x) = 8x^3 - 12x\,, && H_3'(x) = 24x^2 - 12\,, \\ &x_1 = -1.2247449\,, && \omega_1 = 0.2954090\,, \\ &x_2 = 0.0\,, && \omega_2 = 1.1816359\,, \\ &x_3 = 1.2247449\,, && \omega_1 = 0.2954090\,, \\ &I \approx \sum_{i=1}^{3}\omega_i\left[\exp(-\frac{1}{x_i^2})-1\right] &&= 0.2954090\left\{\left[\exp(-\frac{1}{x_1^2})-1\right] + \right. \\ &\left. + \left[\exp(-\frac{1}{x_3^2})-1\right]\right\} &&+ 1.1816359\cdot\{-1\} \approx -1.469117\,. \end{aligned}$$

▷

Формули (6.35) та (6.39) мають найвищу алгебраїчну точність.

6.7. Кратне інтегрування

Нехай необхідно обчислити інтеграл вигляду $\int\limits_D f(\vec{x})d\vec{x}$, де D – область в n-вимірному просторі, $\vec{x} = (x_1, \ldots, x_n)$ - точка цього простору, $d\vec{x}$ - елементарний об'єм. Задача наближеного обчислення кратного інтеграла набагато складніша і менш досліджена порівняно із випадком однієї змінної. Такі труднощі викликані не тільки громіздкими обчисленнями, що пов'язані із наявністю великої кількості змінних. Задачу також ускладнює розмаїтість видів областей інтегрування.

Як і в одновимірному випадку, найчастіше за наближене значення інтеграла приймається лінійна комбінація

$$\int\limits_D f(\vec{x})d\vec{x} \approx \sum_{i=1}^{N} \omega_i f(\vec{x}^i)\,. \tag{6.41}$$

Сума в правій частині (6.41) називається *кубатурною формулою*, а $\vec{x}^i$ та ω_i - *вузли* та *ваги* кубатурної формули. Якщо рівність (6.41) переписати у вигляді

$$\int\limits_D f(\vec{x})d\vec{x} = \sum_{i=1}^{n} \omega_i f(\vec{x}^i) + R(f)\,, \tag{6.42}$$

то лишок $R(f)$ в правій частині (6.41) буде давати похибку при заміні інтеграла на суму. Однією із головних характеристик формули (6.42) є степінь точності цієї формули.

Подібно до одновимірного випадку можна казати, що (6.42) має алгебраїчну степінь точності m, якщо лишок $R(f)$ дорівнює нулеві при умові, що f – довільний поліном від $\vec{x}$ степеня $\leq m$:

$$f(\vec{x}) \equiv P(\vec{x}) = \sum_{m_1,\ldots,m_n} C_{m_1\ldots m_n} x_1^{m_1} \cdots x_n^{m_n}\ (m_1 + \ldots + m_n \leq m),$$

та існує хоча б один багаточлен степеня $m+1$, для якого лишок відмінний від нуля.

Одним із найпростіших способів побудови кубатурних формул є повторне застосування відомих квадратурних формул. Цей спосіб грунтується на обчисленні інтегралів шляхом повторного інтегрування. Розглянемо його на прикладі двократного інтеграла

$$I = \iint\limits_D f(x,y)dxdy\,,$$

де область інтегрування D є прямокутник $a \le x \le b$, $c \le y \le d$. Подвійний інтеграл приводиться до двох простих інтегралів, до яких у свою чергу можна застосувати які-небудь квадратурні формули (не обов'язково однакові для різних змінних інтегрування):

$$I = \int\limits_a^b dx \int\limits_c^d f(x,y)dy = \sum_{i=1}^{N}\sum_{j=1}^{M}\omega_{ij}f(x_i,y_j) + R\,.$$

Тут вага $\omega_{ij} = \omega_i^{(1)} \cdot \omega_j^{(2)}$, тобто є добутком відповідних ваг квадратурних формул.

Очевидно, що застосування таких формул вимагає набагато більше часу для обчислень порівняно з одновимірним інтегралом, особливо при збільшенні розмірності простору. Тому нижче без доказів будуть наведені деякі приклади кубатурних формул, що побудовані за допомогою багатовимірних інтерполяційних поліномів і які можуть бути корисними.

Для двократного інтегрування можна запропонувати просту формулу, степінь точності якої $m = 5$:

$$\begin{aligned}\int\limits_{-1}^{1}\int\limits_{-1}^{1} f(x,y)dxdy = &\frac{88}{45}f(0,0)+\frac{16}{45}\left[f(-1,0)+f(1,0)+f(0,-1)+\right.\\ &\left.+f(0,1)\right]+\frac{7}{45}\left[f(-1,-1)+f(1,-1)+f(-1,1)+f(1,1)\right]+R.\end{aligned} \tag{6.43}$$

Приклад 12. *Обчислити значення інтеграла*

$$\int\limits_0^{\pi/2}\int\limits_0^{\pi/2} \frac{\sin y\sqrt{1-k^2\sin^2 x\sin^2 y}}{1-k^2\sin^2 y}dxdy = \frac{\pi}{1\sqrt{1-k^2}}$$

при різних значеннях k.

$\lhd$ Для того, щоб можна було користуватися формулами типу (6.43), необхідно перейти до стандартної області інтегрування:

$$\int_a^b\int_c^d f(x,y)dxdy=\frac{b-a}{2}\frac{d-c}{2}\int_{-1}^{1}\int_{-1}^{1} f(\frac{a+b}{2}+\frac{b-a}{2}u,\frac{c+d}{2}+\frac{d-c}{2}v)dudv.$$

Результати використання кубатурної формули (6.43) представлені у вигляді таблиці

Значення k	Наближене значення	Точне значення
0.1	1.581874	1.578710
0.5	1.799189	1.813799
0.9	3.357491	3.603654

Видно, що хоча із зростанням k точність розрахунків погіршується, але враховуючи характер підінтегральної функції результати можна вважати добрими. $\rhd$

Потрійні інтеграли можна обраховувати за формулою, степінь точності якої $m = 5$:

$$\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} f(x,y,z)dxdy = -\frac{496}{45}f(0,0,0)+$$
$$+\frac{128}{45}[f(-0.5,0,0)+f(0.5,0,0)+f(0,-0.5,0)+$$
$$+f(0,0.5,0)+f(0,0,-0.5)+f(0,0,0.5)]+$$
$$+\frac{8}{45}[f(-1,0,0)+f(1,0,0)+f(0,-1,0)+$$
$$+f(0,1,0)+f(0,0,-1)+f(0,0,1)]+$$
$$+\frac{1}{9}[f(-1,-1,-1)+f(-1,-1,1)+f(-1,1,-1)+f(-1,1,1)+$$
$$+f(1,-1,-1)+f(1,-1,1)+f(1,1,-1)+f(1,1,1)]+R\,.$$

Глава 7

Розв'язок звичайних диференціальних рівнянь

Багато прикладних задач, у тому числі й фізичних, приводять до необхідності розв'язку звичайних диференціальних рівнянь (ЗДР). При цьому часто розв'язки отриманих рівнянь не можна виразити в квадратурах (тобто в аналітичному вигляді). Тому виникає необхідність застосовувати ті чи інші методи, які дають наближений розв'язок задачі. Ці методи можна поділити на дві групи. Одні з них дають наближений розв'язок у вигляді аналітичного виразу (*аналітичні наближені методи*), інші у вигляді таблиці значень функції в скінченній кількості точок (*чисельні методи*).

7.1. Метод послідовних наближень

Розглянемо ЗДР першого порядку

$$y' = f(x, y) \tag{7.1}$$

з початковою умовою

$$y(x_0) = y_0\,. \tag{7.2}$$

Якщо функція $f(x, y)$ визначена в області $D = \{|x - x_0| \le a,\ |y - y_0| \le b\}$ і задовольняє в цій області умову Ліпшиця по змінній y

$$|f(x, y_1) - f(x, y_2)| \le K|y_1 - y_2|$$

для всіх $(x, y) \in D$ і K не залежить від x, y_1, y_2, то задача Коші (7.1), (7.2) має єдиний розв'язок. При цьому точний розв'язок можна отримати як границю

послідовності $y_0(x)$, $y_1(x)$,..., $y_n(x)$, ..., де

$$y_{n+1}(x) = y_0 + \int\limits_{x_0}^{x} f\left(t, y_n(t)\right) dt\,. \tag{7.3}$$

Очевидно, що (7.3) випливає із (7.1) та (7.2). Пошук розв'язку за допомогою формули (7.3) називається *методом послідовних наближень Пікара.*

Приклад 13. *Розв'язати задачу Коші*

$$\left.\begin{array}{l} y' = x - y\,, \\ y(0) = 1\,. \end{array}\right\} \tag{7.4}$$

$\lhd$ Послідовно використовуємо формулу (7.3)

$$\begin{aligned}
&y_0(x) = y_0 = 1\,,\\
&y_1(x) = y_0 + \int\limits_0^x (t - y_0)dt = 1 + \int\limits_0^x (t-1)dt = 1 - x + \frac{x^2}{2}\,,\\
&y_2(x) = 1 + \int\limits_0^x (t - 1 + t - \frac{t^2}{2})dt = 1 - x + x^2 - \frac{x^3}{6}\,,\\
&y_3(x) = 1 + \int\limits_0^x (t - 1 + t - t^2 + \frac{t^3}{6})dt = 1 - x + x^2 - \frac{x^3}{3} + \frac{x^4}{24}\,,\\
&\ldots\\
&y_n(x) = 1 - x + x^2 - \frac{x^3}{3} + \ldots (-1)^n 2 \cdot \frac{x^n}{n!} + (-1)^{n+1}\frac{x^n}{(n+1)!} \longrightarrow\\
&\longrightarrow 2e^{-x} - 1 + x\,,
\end{aligned}$$

що збігається з точним розв'язком. $\rhd$

7.2. Метод степеневих рядів

Розглянемо ЗДР n-го порядку

$$y^{(n)}(x) = f\left(x, y, y', \ldots, y^{(n-1)}\right) \tag{7.5}$$

з початковими умовами

$$y(x_0) = y_0\,,\ y'(x_0) = y_0'\,, \ldots, y^{(n-1)}(x_0) = y_0^{(n-1)}\,. \tag{7.6}$$

Розкладемо $y(x)$ в ряд Тейлора поблизу точки x_0:

$$y(x) = y_0 + \sum_{i=1}^{n-1} \frac{y_0^{(i)}}{i!}(x-x_0)^i + \frac{y_0^{(n)}}{n!}(x-x_0)^n + \dots \tag{7.7}$$

Коефіцієнти в перших n доданках правої частини (7.7) визначаються початковими умовами (7.6). Для знаходження подальших коефіцієнтів необхідно потрібну кількість разів продиференціювати (7.5), використавши при цьому (7.6). Дійсно,

$$y^{(n)}(x_0) = f(x_0, y_0, y_0', \dots, y_0^{(n-1)}), \tag{7.8}$$
$$y^{(n+1)}(x) = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}y' + \frac{\partial f}{\partial y'}y'' + \dots + \frac{\partial f}{\partial y^{(n-1)}}y^{(n)},$$

де $y^{(n)}$ визначається за допомогою (7.5). Тоді

$$y^{(n+1)}(x_0) = f_1(x_0, y_0, y_0', \dots, y_0^{(n-1)}). \tag{7.9}$$

Аналогічно

$$\begin{aligned} y^{(n+2)}(x_0) &= \frac{\partial f_1}{\partial x} + \frac{\partial f_1}{\partial y}y' + \frac{\partial f_1}{\partial y'}y'' + \dots + \frac{\partial f_1}{\partial y^{(n-1)}}y^{(n)} = \\ &= f_2(x_0, y_0, y_0', \dots, y_0^{(n-1)}) \end{aligned}$$

тощо. Підставляючи (7.8), (7.9) тощо в (7.7), отримаємо ряд з будь-якою кількістю доданків.

Приклад 14. *Методом степеневих рядів розв'язати задачу Коші*

$$y'' = 3\sin x - 4y,$$
$$y(0) = 1.,\ y'(0) = 1.$$

◁ Послідовно знаходимо

$$\begin{aligned} & & y''(0) &= -4, \\ y'''(x) &= 3\cos x - 4y', & y'''(0) &= -1, \\ y^{(4)}(x) &= -3\sin x - 4y'', & y^{(4)}(0) &= 16, \\ y^{(5)}(x) &= -15\cos x + 16y', & y^{(5)}(0) &= 1, \\ & & y^{(6)}(0) &= -64, \\ & & y^{(7)}(0) &= -1, \\ & & \dots & \\ & & y^{(2n)}(0) &= (-4)^n, \\ & & y^{(2n+1)}(0) &= (-1)^n, \end{aligned}$$

$\dots$

У результаті

$$\begin{aligned} y(x) &= 1 + x - 4 \cdot \frac{x^2}{2!} - \frac{x^3}{3!} + 16\frac{x^4}{4!} + \frac{x^5}{5!} - 64\frac{x^6}{6!} - \frac{x^7}{7!} + \dots \\ &= 1 - \frac{(2x)^2}{2!} + \frac{(2x)^4}{4!} - \frac{(2x)^6}{6!} + \dots \\ &\quad \dots + x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots \\ &= \cos 2x + \sin x \,, \end{aligned}$$

що збігається з точним розв'язком.

7.3. Метод Рунге-Кутта

Розглянемо задачу Коші (7.1), (7.2). При чисельному розв'язку задача ставиться так: у точках $x_1, x_2, \dots, x_n$, треба знайти наближення y_n до значень точного розв'язку $y(x_n)$. Різниця $h_n = x_{n+1} - x_n$ називається *кроком сітки*; будемо вважати крок постійним.

Припустимо, що відомі розв'язки в точках $x_0, ..., x_n$. Тоді можна записати:

$$y(x_{n+1}) = y_n + \int\limits_{x_n}^{x_{n+1}} f\left(t, y(t)\right) dt \,. \tag{7.10}$$

Для обчислення інтегралу тепер можна використати яку-небудь квадратурну формулу. Наприклад, формула прямокутників приводить до виразу:

$$y(x_{n+1}) = y_n + h \cdot f\left(\xi, y(\xi)\right) + O(h^2) \,, \tag{7.11}$$

де $\xi \in [x_n, x_{n+1}]$. Якщо знехтувати похибкою та обрати $\xi = x_n$, то отримаємо *формулу Ейлера*:

$$y_{n+1} = y_n + h \cdot f(x_n, y_n) \,. \tag{7.12}$$

Формула трапецій у правій частині (7.10) приводить до виразу

$$y(x_{n+1}) = y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}\,, y(x_{n+1}))\right] + O(h^3) \,, \tag{7.13}$$

Якщо знехтувати похибкою, то отримаємо трансцендентне рівняння відносно y_{n+1}:

$$y_{n+1} = y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1})\right] \,, \tag{7.14}$$

розв'язок якого може бути пов'язаний із значними труднощами. Але можна дещо змінити процедуру розрахунків із збереженням точності кінцевого виразу.

Визначимо початкове наближення:

$$y^*_{n+1} = y_n + h \cdot f(x_n, y_n)\,. \tag{7.15}$$

Тоді (7.13) з урахуванням (7.15) та (7.11) можна переписати:

$$\begin{aligned} y(x_{n+1}) &= y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}\,, y^*_{n+1} + O(h^2))\right] + O(h^3) = \\ &= y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}, y^*_{n+1}) + O(h^2)\right] + O(h^3) = \\ &= y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}, y^*_{n+1})\right] + O(h^3)\,. \end{aligned}$$

Якщо знехтувати похибкою, отримаємо розрахункову формулу

$$y_{n+1} = y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}, y^*_{n+1})\right]\,, \tag{7.16}$$

і точність цієї формули збігається з точністю (7.14). Послідовне використання (7.15) та (7.16) на кожному кроці називається *методом Ейлера з уточненням.*

Згадаємо, що формула прямокутників для вузла в середині інтервалу має похибку $O(h^3)$, тобто із (7.10) випливає

$$y(x_{n+1}) = y_n + h \cdot f\left(x_n + \frac{h}{2}\,, y(x_n + \frac{h}{2})\right) + O(h^3)\,. \tag{7.17}$$

Побудуємо допоміжну величину

$$y_{n+\frac{1}{2}} = y_n + \frac{h}{2} f(x_n, y_n) \tag{7.18}$$

із похибкою $O(h^2)$ і підставимо її в (7.17). У результаті отримаємо

$$y_{n+1} = y_n + hf\left(x_n + \frac{h}{2}\,, y_{n+\frac{1}{2}}\right) \tag{7.19}$$

з похибкою $O(h^3)$, тобто точність методу (7.18)-(7.19) така сама, як і методу Ейлера з уточненням. Ці методи також мають назву *предиктор-коректор*: спочатку обраховується наближене значення, а потім воно уточнюється. Хоча уточнення потребує деякого збільшення розрахункової роботи, але це повністю компенсується значним збільшенням точності.

Формули (7.15), (7.16) та (7.18), (7.19) належать до родини формул Рунге-Кутта. Загальний вигляд методу Рунге-Кутта такий. Оцінка значення функції в точці x_{n+1} знаходиться у вигляді

$$y_{n+1} = y_n + \sum_{i=1}^{q} p_i k_i(h)\,, \tag{7.20}$$

де

$$\begin{aligned}
&k_1(h) = h \cdot f(x_n, y_n),\\
&k_2(h) = h \cdot f(x_n + \alpha_2 h, y_n + \beta_{21} k_1),\\
&k_3(h) = h \cdot f(x_n + \alpha_3 h, y_n + \beta_{31} k_1 + \beta_{32} k_2),\\
&\ldots\\
&k_q(h) = h \cdot f(x_n + \alpha_q h, y_n + \beta_{q1} k_1 + \ldots + \beta_{q,q-1} k_{q-1}).
\end{aligned} \tag{7.21}$$

Тут $p_i, \alpha_i, \beta_{ij}$ - невідомі константи, вигляд яких визначається деякими додатковими умовами, головним з яких є точність, яка вимагається.

Введемо *похибку методу* $\varphi(h) = y(x+h) - y_{n+1}$. Припустимо, що

$$\varphi(0) = \varphi'(0) = \varphi''(0) = \ldots = \varphi^{(s)}(0) = 0\,.$$

Будемо говорити, що s - *порядок точності*, тоді можна записати

$$\begin{aligned}
\varphi(h) &= \sum_{i=0}^{s} \frac{\varphi^{(i)}(0)}{i!} h^i + \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!} h^{s+1} = \\
&= \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!} h^{s+1}, \quad 0 \le \theta \le 1\,.
\end{aligned}$$

Нехай $q = 1$:

$$\begin{aligned}
&y_{n+1} = y_n + p_1 \cdot k_1(h) = y_n + p_1 h \cdot f(x_n, y_n)\,,\\
&\varphi(h) = y(x_n + h) - y_{n+1} = y(x_n + h) - y_n - p_1 h \cdot f(x_n, y_n)\,.
\end{aligned}$$

Очевидно, що $\varphi(0) = 0$, тоді

$$\varphi'(h) = [y'(x_n + h) - p_1 f(x_n, y_n)]|_{h=0} = (1 - p_1) f(x_n, y_n), \tag{7.22}$$

$$\varphi''(h) = y''(x_n + h)\,.$$

У загальному випадку величини $\varphi''(0)$, $\varphi'''(0)$ тощо не дорівнюють нулеві, а рівність нулеві виразу (7.22) для довільної f буде виконуватися при $p_1 = 1$. У результаті маємо формулу Ейлера (7.12) при $s = 1$.

Нехай $q = 2$:

$$y_{n+1} = y_n + p_1 h f(x_n, y_n) + p_2 h f\left(x_n + \alpha_2 h, y_n + \beta_{21} h f(x_n, y_n)\right) .$$

Для скорочення кількості позначень нижче запис f буде означати, що аргументами є значення x_n, y_n. Тоді

$$\varphi(h) = y(x_n + h) - y_n - p_1 f - p_2 h f(\overline{x}_n, \overline{y}_n),$$

де $\overline{x}_n = x_n + \alpha_2 h, \overline{y}_n = y_n + \beta_{21} h f$. Використаємо далі також позначення $\overline{f} = f(\overline{x}_n, \overline{y}_n)$.

Очевидно, що $\varphi(0) = 0$. Далі

$$\begin{aligned}
\varphi'(h) &= y'(x_n + h) - p_1 f - p_2 \overline{f} - p_2 h \left[\alpha_2 \frac{\partial \overline{f}}{\partial x} + \beta_{21} f \frac{\partial \overline{f}}{\partial y}\right], \\
\varphi''(h) &= y''(x_n + h) - 2p_2 \left[\alpha_2 \frac{\partial \overline{f}}{\partial x} + \beta_{21} f \frac{\partial \overline{f}}{\partial y}\right] - \\
&- p_2 h \left[\alpha_2^2 \frac{\partial^2 \overline{f}}{\partial x^2} + 2\alpha_2 \beta_{21} f \frac{\partial^2 \overline{f}}{\partial x \partial y} + \beta_{21}^2 f^2 \frac{\partial^2 \overline{f}}{\partial y^2}\right], \\
\varphi'''(h) &= y'''(x_n + h) - 3p_2 \left[\alpha_2^2 \frac{\partial^2 \overline{f}}{\partial x^2} + 2\alpha_2 \beta_{21} f \frac{\partial^2 \overline{f}}{\partial x \partial y} + \beta_{21}^2 f^2 \frac{\partial^2 \overline{f}}{\partial y^2}\right] .
\end{aligned}$$

Тоді

$$\begin{aligned}
\varphi'(0) &= (1 - p_1 - p_2) f , \\
\varphi''(0) &= y''(x_n) - 2p_2 \left[\alpha_2 \frac{\partial f}{\partial x} + \beta_{21} f \frac{\partial f}{\partial y}\right] = \\
&= (1 - 2p_2\alpha_2) \frac{\partial f}{\partial x} + (1 - 2p_2\beta_{21}) f \frac{\partial f}{\partial y} , \\
\varphi'''(0) &= y'''(x_n) - 3p_2 \left[\alpha_2^2 \frac{\partial^2 f}{\partial x^2} + 2\alpha_2 \beta_{21} f \frac{\partial^2 f}{\partial x \partial y} + \beta_{21}^2 f^2 \frac{\partial^2 f}{\partial y^2}\right] = \\
&= (1 - 3p_2\alpha_2^2) \frac{\partial^2 f}{\partial x^2} + +2(1 - 3p_2\alpha_2\beta_{21}) f \frac{\partial^2 f}{\partial x \partial y} + \\
&\quad + (1 - 3p_2\beta_{21}^2) f^2 \frac{\partial^2 f}{\partial y^2} + \frac{\partial f}{\partial y} y''(x_n) .
\end{aligned}$$

Очевидно, що при довільному $f(x, y)$ величина $\varphi'''(0) \neq 0$, а рівності $\varphi'(0) = \varphi''(0) = 0$ виконуються, якщо

$$p_1 + p_2 = 1 , \quad 2p_2\alpha_2 = 1 , \quad 2p_2\beta_{21} = 1 ,$$

тобто чотири параметри задовольняють три умови. Ми можемо довільним чином задавати значення одного із параметрів і отримувати різні формули для метода Рунге-Кутта. Наприклад, $p_1 = \dfrac{1}{2}$. Тоді $p_2 = \dfrac{1}{2}$, $\alpha_2 = 1$, $\beta_{21} = 1$ і в результаті

$$y_{n+1} = y_n + \frac{h}{2}\left[f + f(x_n + h, y_n + hf)\right] ,$$

що збігається із методом Ейлера з уточненням (7.15), (7.16) ($s = 2$). Інший можливий розв'язок: $p_1 = 0$, $p_2 = 1$, $\alpha_2 = 1/2$, $\beta_{21} = 1/2$ і

$$y_{n+1} = y_n + hf\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}f\right) ,$$

що збігається із (7.18), (7.19) ($s = 2$).

Одна з найбільш поширених формул Рунге-Кутта отримується в результаті вибору в (7.20) та (7.21) таких параметрів: $q = 4$, $s = 4$. При цьому

$$\begin{aligned}
&p_1 = \frac{1}{6}, \quad p_2 = \frac{1}{3}, \quad p_3 = \frac{1}{3}, \quad p_4 = \frac{1}{6},\\
&\alpha_2 = \frac{1}{2}, \quad \beta_{21} = \frac{1}{2},\\
&\alpha_3 = \frac{1}{2}, \quad \beta_{31} = 0, \quad \beta_{31} = \frac{1}{2},\\
&\alpha_4 = 1, \quad \beta_{41} = 0, \quad \beta_{42} = 0, \quad \beta_{43} = 1 .
\end{aligned}$$

Коли в літературі говорять про формулу Рунге-Кутта, як правило, мають на увазі саме такий вибір, який приводить до наступної розрахункової формули:

$$y_{n+1} = y_n + \frac{1}{6}\left[k_1 + 2k_2 + 2k_3 + k_4\right] , \tag{7.23}$$

де

$$\begin{aligned}
k_1 &= h \cdot f(x_n, y_n) ,\\
k_2 &= h \cdot f\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}\right) ,\\
k_3 &= h \cdot f\left(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}\right) ,\\
k_4 &= h \cdot f(x_n + h, y_n + k_3) .
\end{aligned} \tag{7.24}$$

Приклад 15. *Знайти розв'язок рівняння (7.4) в точці* $x = 0.2$ *з кроком* $h = 0.1$.

$\lhd$ Спочатку розв'яжемо задачу за допомогою формули Ейлера (7.12) ($s = 1$):

$$\begin{aligned} y_1 &= y_0 + hf(x_0, y_0) = 1. + 0.1 \cdot (0. - 1.) = 0.9\,, \\ y_2 &= y_1 + hf(x_1, y_1) = 0.9 + 0.1 \cdot (0.1 - 0.9) = \mathbf{0.82}\,. \end{aligned}$$

Метод Ейлера з уточненням (7.15), (7.16) ($s = 2$):

$$\begin{aligned} y_1^* &= y_0 + hf(x_0, y_0) = 0.9\,, \\ y_1 &= y_0 + \frac{h}{2}\left[f(x_0, y_0) + f(x_1, y_1^*)\right] = \\ &= 1 + 0.05 \cdot [-1. - 0.8] = 0.91\,, \\ y_2^* &= y_1 + hf(x_1, y_1) = 0.91 + 0.1 \cdot (0.1 - 0.91) = 0.829\,, \\ y_2 &= y_1 + \frac{h}{2}\left[f(x_1, y_1) + f(x_2, y_2^*)\right] = \\ &= 0.91 + 0.05 \cdot [(0.1 - 0.91) + (0.2 - 0.829)] = \mathbf{0.83805}\,. \end{aligned}$$

Метод Рунге-Кутта (7.18), (7.19) ($s = 2$):

$$\begin{aligned} y_{\frac{1}{2}}^* &= y_0 + \frac{h}{2} f(x_0, y_0) = 1 + 0.05 \cdot (0. - 1.) = 0.95\,, \\ y_1 &= y_0 + hf\left(x_0 + \frac{h}{2}, y_{\frac{1}{2}}^*\right) = 1 + 0.1 \cdot (.05 - 0.95) = 0.91\,, \\ y_{\frac{3}{2}}^* &= y_1 + \frac{h}{2} f(x_1, y_1) = 0.91 + 0.05 \cdot (0.1 - 0.91) = 0.8695\,, \\ y_2 &= y_1 + hf\left(x_1 + \frac{h}{2}\,, y_{\frac{3}{2}}^*\right) = \\ &= 0.91 + 0.1 \cdot (0.15 - 0.8695) = \mathbf{0.83805}\,. \end{aligned}$$

Метод Рунге-Кутта (7.23), (7.24) ($s = 4$):

$$\begin{aligned} k_1 &= h \cdot f(x_0, y_0) = 0.1 \cdot (0. - 1.) = -0.1\,, \\ k_2 &= h \cdot f\left(x_0 + \frac{h}{2}, y_0 + \frac{k_1}{2}\right) = -0.09\,, \\ k_3 &= h \cdot f\left(x_0 + \frac{h}{2}, y_0 + \frac{k_2}{2}\right) = -0.0905\,, \\ k_4 &= h \cdot f(x_0 + h, y_0 + k_3) = -0.08095\,, \\ y_1 &= y_0 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) = 0.909675\,, \end{aligned}$$

$$
\begin{aligned}
k_1 &= h \cdot f(x_1, y_1) = -0.0809675\,, \\
k_2 &= h \cdot f\left(x_1 + \frac{h}{2}, y_1 + \frac{k_1}{2}\right) = -0.07191913\,, \\
k_3 &= h \cdot f\left(x_1 + \frac{h}{2}, y_1 + \frac{k_2}{2}\right) = -0.07237155\,, \\
k_4 &= h \cdot f(x_1 + h, y_1 + k_3) = -0.06373035\,, \\
y_1 &= y_0 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) = \mathbf{0.8374618}\,.
\end{aligned}
$$

Для порівняння наведемо точне значення

$$y(x) = 2e^{-x} - 1 + x\,,\ y(0.2) = \mathbf{0.8374614}\ldots$$

▷

7.4. Метод Адамса

Розглянуті вище чисельні методи є однокроковими. Іншими словами, для обчислення значення y_{n+1} використовується значення тільки y_n, і для підвищення точності при фіксованому кроці необхідно проводити обчислення все більшої кількості допоміжних величин (типу k_i в (7.21)). Це є причиною того, що для багатьох задач застосування формул Рунге-Кутта є неможливим внаслідок великих втрат комп'ютерного часу. Тому частіше більш раціонально переходити до багатокрокових методів, які дають можливість, використовуючи значення $f(x_i, y_i)$, що обчислені на попередніх кроках, отримати прийнятну точність.

Серед k-крокових методів найбільш часто використовуються методи інтегрування на сітці з постійним кроком, які називаються *кінцево-різницевими схемами*. Розглянемо ЗДР (6.41) і припустимо, що вже відомі розв'язки на множині значень x_i (i=0, 1, . . ., n). Тобто можна записати рівняння (7.10):

$$y(x_{n+1}) = y_n + \int\limits_{x_n}^{x_{n+1}} f\left(t, y(t)\right) dt.$$

При обчисленні інтеграла в правій частині цього виразу підінтегральну функцію замінимо на інтерполяційний поліном Ньютона для інтерполяції назад (2.19) на сітці $x_n, x_{n-1}, x_{n-2}, \ldots$ При цьому

$$f(x,y)\equiv f(x)=a_0+a_1(x-x_n)+a_2(x-x_n)(x-x_{n-1})+ \\ +\ldots+a_m(x-x_n)\cdots(x-x_{n-m+1})+R_m(x)\,, \tag{7.25}$$

де $a_k=\dfrac{\nabla^k f_n}{k!h^k}$ і $R_m(x)$ – похибка інтерполяції, яка і буде визначати похибку отриманих нижче формул. Нагадаємо, що $\nabla^k f_n$ - кінцеві ліві різниці k-го порядку функції $f(x,y)$ в точці x_n. Підставивши (7.25) в праву частину (7.10) і знехтувавши оцінкою похибки, отримаємо

$$\int\limits_{x_n}^{x_{n+1}} f\left(t,y(t)\right)dt\approx\sum_{k=0}^{m}\frac{\nabla^k f_n}{k!h^k}\int\limits_{x_n}^{x_{n+1}}(t-x_n)\ldots(t-x_{n-k+1})dt. \tag{7.26}$$

Обрахуємо декілька перших інтегралів в (7.26):

$$k=0\,,\quad \int\limits_{x_n}^{x_{n+1}} dt=h\,,$$
$$k=1\,,\quad \int\limits_{x_n}^{x_{n+1}}(t-x_n)dt=\frac{h^2}{2}\,,$$
$$k=2\,,\quad \int\limits_{x_n}^{x_{n+1}}(t-x_n)(t-x_{n-1})dt=\frac{5}{6}h^3\,,$$

тощо. У результаті отримаємо *формулу Адамса*

$$y_{n+1}=y_n+h\left[f_n+\frac{1}{2}\nabla f_n+\frac{5}{12}\nabla^2 f_n+\frac{3}{8}\nabla^3 f_n+\frac{251}{720}\nabla^4 f_n+\ldots\right], \tag{7.27}$$

де порядок точності методу збігається з кількістю доданків у квадратних дужках.

На практиці, для користування цією формулою залежно від порядку точності, необхідно знати певну кількість значень f_i (а значить і y_i) у вузлах x_i. Тому для "розгону" звичайно використовують однокроковий метод (наприклад Рунге-Кутта) для знаходження y_i в декількох початкових точках поблизу x_0, а потім переходять до формули Адамса.

Приклад 16. *Знайти розв'язок рівняння (7.4) в точці* $x=0.5$ *з кроком* $h=0.1$ *методом Адамса* 3*-го порядку.*

$\lhd$ Для розгону можна використати значення, обчисленні на стор. 100-101:

$$f_0 = -1\,,\ f_1 = -0.809675\,,\ f_2 = -0.6374618.$$

Окрім того, є зручним такий вираз

$$\begin{aligned} y_{n+1} &= y_n + h\left[f_n + \frac{1}{2}\nabla f_n + \frac{5}{12}\nabla^2 f_n\right] = \\ &= y_n + \frac{h}{12}\left[23f_n - 16f_{n-1} + 5f_{n-2}\right]\,. \end{aligned} \tag{7.28}$$

Тоді

$$\begin{aligned} y_3 &= y_2 + \frac{h}{12}\left[23f_2{-}16f_1{+}5f_0\right] = 0.7815716\,,\ f_3 = -0.4815716\,,\\ y_4 &= y_3 + \frac{h}{12}\left[23f_3{-}16f_2{+}5f_1\right] = 0.7405288\,,\ f_4 = -0.3405288,\\ y_5 &= y_4 + \frac{h}{12}\left[23f_4{-}16f_3{+}5f_2\right] = \mathbf{0.7129094}, (= \mathbf{0.7130614})\,. \end{aligned}$$

Для порівняння у дужках наведено точне значення розв'язку рівняння (7.4) в точці $x = 0.5$. $\rhd$

Формула (7.28) називається *екстраполяційною*, оскільки підінтегральна функція в (7.10) представляється у вигляді інтерполяційного полінома, що побудований на вузлах поза інтервалом інтегрування. Для того щоб зменшити вплив похибки, що пов'язана з екстраполяцією, замість (7.25) побудуємо поліном Ньютона на сітці $x_{n+1},\ x_n,\ x_{n-1},\ \ldots$. Тоді

$$\begin{aligned} f(x,y) \cong\ & a_0 + a_1(x - x_{n+1}) + a_2(x - x_{n+1})(x - x_n) + \\ & + \ldots + a_m(x - x_{n+1})\cdots(x - x_{n-m+2})\,, \end{aligned} \tag{7.29}$$

де $a_k = \dfrac{\nabla^k f_{n+1}}{k!h^k}$. Підставимо (7.29) в (7.10) і одержимо

$$\begin{aligned} y_{n+1} &= y_n + \sum_{k=0}^{m} \frac{\nabla^k f_{n+1}}{k!h^k} \int\limits_{x_n}^{x_{n+1}} (t{-}x_{n+1})\ldots(t - x_{n-k+2})dt = \\ &= y_n + h\left[f_{n+1} - \frac{1}{2}\nabla f_{n+1} - \frac{1}{12}\nabla^2 f_{n+1} - \frac{1}{24}\nabla^3 f_{n+1} - \right. \\ &\qquad \left. - \frac{19}{720}\nabla^4 f_{n+1} + \ldots\right]\,, \end{aligned} \tag{7.30}$$

так звану *інтерполяційну формулу Адамса*, яка може слугувати коректором по відношенню до предиктора (7.27).

Приклад 17. *Розв'язати попередню задачу за допомогою формул (7.27) та (7.30) третього порядку.*

◁ Перетворимо (7.30):

$$\begin{aligned} y_{n+1} &= y_n + h\left[f_{n+1} - \frac{1}{2}\nabla f_{n+1} - \frac{1}{12}\nabla^2 f_{n+1}\right] = \\ &= y_n + \frac{h}{12}\left[5f_{n+1} + 8f_n - f_{n-1}\right] \end{aligned} \tag{7.31}$$

і далі одержимо

$$y_3^* = y_2 + \frac{h}{12}\left[23f_2 - 16f_1 + 5f_0\right] = 0.7815716\,,\ f_3^* = -0.4815716\,,$$
$$y_3 = y_2 + \frac{h}{12}\left[5f_3^* + 8f_2 - f_1\right] = 0.7816461\,,\ f_3 = -0.4816461\,,$$
$$(\mathbf{0.7816366})$$

$$y_4^* = y_3 + \frac{h}{12}\left[23f_3 - 16f_2 + 5f_1\right] = 0.7405890\,,\ f_4^* = -0.3405890\,,$$
$$y_4 = y_3 + \frac{h}{12}\left[5f_4^* + 8f_3 - f_2\right] = 0.7406573\,,\ f_4 = -0.3406573\,,$$
$$(\mathbf{0.7406400})$$
$$y_5^* = y_4 + \frac{h}{12}\left[23f_4 - 16f_3 + 5f_2\right] = 0.7130232\,,\ f_5^* = -0.2130232\,,$$
$$y_5 = y_4 + \frac{h}{12}\left[5f_5^* + 8f_4 - f_3\right] = 0.7130845\,,\ (= \mathbf{0.7130614})\,.$$

Поруч у дужках для порівняння приведені точні значення. ▷

7.5. Метод Мілна

Одним із найбільш простих і практично зручних методів чисельного розв'язку ЗДР є метод Мілна, який також є кінцево-різницевим. Як і в методі Адамса, наведені нижче формули Мілна можна використовувати по схемі предиктор-коректор.

Розглянемо ЗДР (6.41) і припустимо, що відомі значення розв'язку y_i на множині значень x_i, $i = 0, 1, 2, \ldots, n$. Аналогічно до (7.10) можна записати

$$y(x_{n+1}) = y_{n-3} + \int\limits_{x_{n-3}}^{x_{n+1}} f\left(t, y(t)\right) dt\,. \tag{7.32}$$

Замінимо підінтегральну функцію на інтерполяційний поліном Ньютона для інтерполяції вперед (2.17) на сітці $x_k, x_{k+1} \ldots$

$$\begin{aligned} f(x,y) \;\cong\; & a_0 + a_1(x - x_k) + a_2(x - x_k)(x - x_{k+1}) + \\ & + \ldots + a_m(x - x_k)\cdots(x - x_{k+m-1})\,, \end{aligned} \tag{7.33}$$

де $a_i = \dfrac{\nabla^i f_k}{i! h^i}$. Підставимо (7.33) в (7.32):

$$y_{n+1} = y_{n-3} + \sum_{i=0}^{m} \frac{\nabla^i f_k}{i! h^i} \int\limits_{x_{n-3}}^{x_{n+1}} (t - x_k)\ldots(t - x_{k+i-1}) dt.$$

Якщо покласти $k = n - 3$ та $m = 3$, то отримаємо

$$\begin{aligned} y_{n+1} &= y_{n-3} + h\left[4f_{n-3} + 8\nabla f_{n-3} + \frac{20}{3}\nabla^2 f_{n-3} + \frac{8}{3}\nabla^3 f_{n-3}\right] = \\ &= y_{n-3} + \frac{4h}{3}\left[2f_n - f_{n-1} + 2f_{n-2}\right]. \end{aligned}$$

Це – *перша формула Мілна* (предиктор).

По аналогії з (7.32) можна записати

$$y(x_{n+1}) = y_{n-1} + \int\limits_{x_{n-1}}^{x_{n+1}} f\left(t, y(t)\right) dt\,.$$

Підставимо сюди (7.33) при $k = n - 1$, $m = 3$ і одержимо

$$y_{n+1} = y_{n-1} + \frac{h}{3}\left[f_{n+1} + 4f_n + f_{n-1}\right].$$

Це – *друга формула Мілна* (коректор). Відмітимо, що обидві формули Мілна мають четвертий порядок точності .

Приклад 18. *Розв'язати рівняння (7.4) з кроком* $h = 0.1$ *до* $x = 0.5$ *і порівняти із розв'язком, що одержаний методом Адамса.*

$\lhd$ Як значення y_2, f_2, y_3, f_3, використаємо результати, що отримані в попередніх прикладах.

$$y_4^* = y_0 + \frac{4h}{3}\,[2f_3 - f_2 + 2f_1] = 0.7406426,\ f_4^* = -0.3406426,$$
$$y_4 = y_2 + \frac{h}{3}\,[f_4^* + 4f_3 + f_2] = 0.7406388,\ f_4 = -0.3406388,$$
$$(\mathbf{0.7406400})$$
$$y_5^* = y_1 + \frac{4h}{3}\,[2f_4 - f_3 + 2f_2] = 0.7129343\,,\ f_5^* = -0.2129343\,,$$
$$y_5 = y_3 + \frac{h}{3}\,[f_5^* + 4f_4 + f_3] = 0.7130749,\ (= \mathbf{0.7130614})\,.$$

Поруч у дужках для порівняння наведені точні значення. $\rhd$

7.6. Розв'язок диференціальних рівнянь другого порядку

Математичний опис багатьох фізичних задач приводить до необхідності розв'язувати ЗДР другого порядку. Як приклад можна назвати рівняння Ньютона і Лагранжа в класичній механіці, рівняння Шредингера в квантовій механіці. За допомогою застосування додаткових нових невідомих диференціальні рівняння порядку вище першого та їх системи зводяться до системи ЗДР першого порядку. Дійсно, розв'язок ЗДР

$$y^{(n)} = f\left(x, y, y', \ldots, y^{(n-1)}\right)$$

з початковими умовами

$$y(x_0) = y_0\,,$$
$$y'(x_0) = y_0'\,,$$
$$\ldots$$
$$y^{(n-1)}(x_0) = y_0^{(n-1)}$$

еквівалентний розв'язку такої системи ЗДР першого порядку:

$$\begin{cases} y' = u_1\,, \\ u_1' = u_2\ (= y'')\,, \\ \dots \\ u_{n-2}' = u_{n-1}\ (= y^{(n-1)})\,, \\ u_{n-1}' = f(x, y, u_1, \dots, u_{n-1}) \end{cases}$$

з початковими умовами

$$\begin{aligned} &y(x_0) = y_0\,, \\ &u_1(x_0) = y_0'\,, \\ &\dots \\ &u_{(n-1)}(x_0) = y_0^{(n-1)}\,. \end{aligned}$$

Методи розв'язку таких систем загального вигляду будуть розглянуті нижче. Проте в силу поширеності ЗДР другого порядку значно ефективнішим є використовування методів, що спеціально для цього пристосовані. Зокрема, тут розглядаються такі ЗДР другого порядку, які явно не містять першої похідної:

$$y'' = f(x, y)\,.$$

Саме до такого вигляду зводиться більшість задач небесної механіки і рівняння Шредингера для багатьох типів потенціалу.

Будемо шукати розв'язок у рівновіддалених точках x_n, $n = 1, 2, \dots$ з очевидними позначеннями:

$$x_{n+1} - x_n = h\,,\ y(x_n) = y_n\,,\ f(x_n, y_n) = f_n = y_n''\,.$$

Розкладемо шукану функцію за формулою Тейлора поблизу точки x_n із залишковим членом в інтегральній формі

$$y(x) = y(x_n) + y'(x_n)(x - x_n) + \int\limits_{x_n}^{x} (x - t) y''(t) dt\,,$$

і оберемо значення x рівним x_{n+1} та x_{n-1} відповідно :

$$\begin{aligned} y_{n+1} &= y_n + y_n' h + \int\limits_{x_n}^{x_{n+1}} (x_{n+1} - t) y''(t) dt\,, \\ y_{n-1} &= y_n - y_n' h + \int\limits_{x_{n-1}}^{x_n} (t - x_{n-1}) y''(t) dt\,. \end{aligned}$$

Додамо ці два рівняння і отримаємо

$$y_{n-1} - 2y_n + y_{n+1} = h^2 \int\limits_{x_{n-1}}^{x_{n+1}} G(t)y''(t)dt\,, \tag{7.34}$$

де

$$G(t) = \begin{cases} \dfrac{1}{h^2}(t - x_{n-1}), & t \le x_n, \\ \dfrac{1}{h^2}(x_{n+1} - t), & t > x_n\,. \end{cases}$$

Для другої похідної $y''(t)$ під знаком інтеграла в правій частині (7.34) використаємо поліном Ньютона для інтерполяції назад по трьох точках x_n, x_{n-1}, x_{n-2}:

$$y''(x) = f_n + \frac{\nabla f_n}{h}(x - x_n) + \frac{\nabla^2 f_n}{2h^2}(x - x_n)(x - x_{n-1}) + O(h^3). \tag{7.35}$$

Інтегрування окремих доданків дає

$$\int\limits_{x_{n-1}}^{x_{n+1}} G(t)dt = 1\,,$$

$$\int\limits_{x_{n-1}}^{x_{n+1}} G(t)(t - x_n)dt = 0\,,$$

$$\int\limits_{x_{n-1}}^{x_{n+1}} G(t)(t - x_n)(t - x_{n-1})dt = \frac{1}{6}\,.$$

У результаті одержимо *явну формулу Штермера* (4-го порядку):

$$\begin{aligned} y_{n+1} &= -y_{n-1} + 2y_n + h^2\left[f_n + \frac{1}{12}\nabla^2 f_n\right] + O(h^5) = \\ &= -y_{n-1} + 2y_n + h^2\left[\frac{13}{12}f_n - \frac{1}{6}f_{n-1} + \frac{1}{12}f_{n-2}\right] + O(h^5) = \\ &= -y_{n-1} + 2y_n + \frac{h^2}{12}\left[13f_n - 2f_{n-1} + f_{n-2}\right] + O(h^5)\,. \end{aligned} \tag{7.36}$$

Цю формулу можна розглядати як предиктор, а уточнюючу *неявну формулу Штермера* (коректор) можна одержати аналогічно, якщо замість (7.35)

використати інтерполяційний поліном Ньютона по трьох точках x_{n+1}, x_n, x_{n-1}:

$$\begin{aligned} y_{n+1} &= -y_{n-1}+2y_n+h^2\left[f_{n+1}-\nabla f_n+\frac{1}{12}\nabla^2 f_{n+1}\right]+O(h^5)= \\ &= -y_{n-1}+2y_n+h^2\left[\frac{1}{12}f_{n+1}+\frac{5}{6}f_n+\frac{1}{12}f_{n-1}\right]+O(h^5)= \\ &= -y_{n-1}+2y_n+\frac{h^2}{12}\left[f_{n+1}+10f_n+f_{n-1}\right]+O(h^5)\,. \end{aligned} \tag{7.37}$$

де при обчисленні f_{n+1} необхідно використовувати y_{n+1} із (7.36).

Особливо простим буде розв'язок для однорідного рівняння типу

$$y'' = f(x)y\,,$$

прикладом якого є рівняння Шредингера. У цьому випадку можна відразу використовувати (7.37):

$$y_{n+1}-\frac{h^2}{12}f_{n+1}y_{n+1} = 2y_n+\frac{5}{6}h^2 f_n y_n - y_{n-1}+\frac{h^2}{12}f_{n-1}y_{n-1}$$

або

$$y_{n+1} = \frac{\left(2+\frac{5}{6}h^2 f_n\right)y_n-\left(1-\frac{1}{12}h^2 f_{n-1}\right)y_{n-1}}{1-\frac{1}{12}h^2 f_{n+1}}\,.$$

Цей вираз називається *формулою Ковелла.*

7.7. Розв'язок систем диференціальних рівнянь

Усі розглянуті вище методи практично без змін переносяться на випадок систем ЗДР першого порядку

$$\left.\begin{aligned} y_1' &= f_1(x, y_1, \ldots, y_n)\,, \\ &\cdots \\ y_n' &= f_n(x, y_1, \ldots, y_n)\,, \end{aligned}\right\}$$

або

$$\vec{y'} = \vec{f}\,(x, \vec{y})\ . \tag{7.38}$$

Формальна відмінність полягає лише в тому, що у відповідних співвідношеннях замість скалярних величин приймають участь деякі матриці, вектори або тензори.

Приклад 19. *Розглянемо застосування методу послідовних наближень для розв'язку системи ЗДР*

$$\begin{cases} y' &= -z + x\,, \\ z' &= -y \end{cases} \tag{7.39}$$

з початковими умовами $y(0) = 1$, $z(0) = 0$.

$\lhd$ Відповідно до (6.43) та (7.38) можна записати:

$$\vec{y}_{n+1}(x) = \vec{y}_0 + \int\limits_{x_0}^{x} \vec{f}\,(t, \vec{y}_n(t))\,dt\,.$$

Для нашого прикладу це дає такі практичні формули:

$$\begin{aligned} y_{n+1}(x) &= 1 + \int\limits_0^x [-z_n(t) + t]\,dt\,, \\ z_{n+1}(x) &= -\int\limits_0^x y_n(t)dt. \end{aligned}$$

Отримаємо послідовні наближення

$$\begin{aligned} y_0(x) &= y(0) = 1\,, \\ z_0(x) &= z(0) = 0\,, \\ y_1(x) &= 1 + \int\limits_0^x (-0 + t)dt = 1 + \frac{x^2}{2}\,, \\ z_1(x) &= -\int\limits_0^x 1 \cdot dt = -x\,, \end{aligned}$$

$$
\begin{aligned}
y_2(x) &= 1 + x^2\,,\\
z_2(x) &= -x - \frac{x^3}{6}\,,\\
y_3(x) &= 1 + x^2 + \frac{x^4}{24}\,,\\
z_3(x) &= -x - \frac{x^3}{3}\,,\\
y_4(x) &= 1 + x^2 + \frac{x^4}{12}\,,\\
z_4(x) &= -x - \frac{x^3}{3} - \frac{x^5}{120}\,,\\
&\dots\\
y_7(x) &= 1 + x^2 + \frac{x^4}{12} + \frac{x^6}{360} + \frac{x^8}{8!}\,,\\
z_7(x) &= -x - \frac{x^3}{3} - \frac{x^5}{60} - 2\cdot\frac{x^7}{7!}\,,\\
&\dots\\
y_n(x) &= 1 + 2\frac{x^2}{2!} + 2\frac{x^4}{4!} + 2\frac{x^6}{6!} + \ldots = 2\,\mathrm{ch}\,x - 1\,,\\
z_n(x) &= -x - 2\frac{x^3}{3!} - 2\frac{x^5}{5!} - 2\cdot\frac{x^7}{7!} - \ldots = -2\,\mathrm{sh}\,x + x\,.
\end{aligned}
$$

Легко впевнитися, що отримані розв'язки задовольняють (7.39). $\triangleright$

Приклад 20. *Розглянемо застосування методу Рунге-Кутта для розв'язку системи двох ЗДР*

$$
\begin{cases} y' = z\,,\\ z' = y - x + 1\,, \end{cases} \tag{7.40}
$$

з початковими умовами $y(0) = 1$, $z(0) = -1$.

$\triangleleft$ Розрахункові формули (7.23) та (7.24) будуть мати вигляд

$$
\vec{y}_{n+1} = \vec{y}_n + \frac{1}{6}\left(\vec{k}_1 + 2\vec{k}_2 + 2\vec{k}_3 + \vec{k}_4\right)\,,
$$

де

$$\begin{aligned}
\vec{k}_1 &= h\vec{f}(x, \vec{y}_n)\ , \\
\vec{k}_2 &= h\vec{f}\left(x + \frac{h}{2}, \vec{y}_n + \frac{\vec{k}_1}{2}\right)\ , \\
\vec{k}_3 &= h\vec{f}\left(x + \frac{h}{2}, \vec{y}_n + \frac{\vec{k}_2}{2}\right)\ , \\
\vec{k}_4 &= h\vec{f}\left(x + h, \vec{y}_n + \vec{k}_3\right)\ .
\end{aligned}$$

Тоді з кроком $h = 0.1$ одержимо

$$\begin{aligned}
k_1 &= 0.1 \cdot (-1) = -0.1\,, \\
l_1 &= 0.1 \cdot (1 - 0 + 1) = 0.2\,, \\
k_2 &= 0.1 \cdot (-1 + 0.1) = -0.09\,, \\
l_2 &= 0.1 \cdot \left(1 - \frac{0.1}{2} - \frac{0.1}{2} + 1\right) = 0.19\,, \\
k_3 &= 0.1 \cdot \left(-1 + \frac{0.19}{2}\right) = -0.0905\,, \\
l_3 &= 0.1 \cdot \left(1 - \frac{0.09}{2} - \frac{0.1}{2} + 1\right) = 0.1905\,, \\
k_4 &= 0.1 \cdot (-1 + 0.1905) = -0.08095\,, \\
l_4 &= 0.1 \cdot (1 - 0.0905 - 0.1 + 1) = 0.18095\,, \\
y_1 &= y_0 + \frac{1}{6}\,(k_1 + 2k_2 + 2k_3 + k_4) = 0.909675\,, \\
z_1 &= z_0 + \frac{1}{6}\,(l_1 + 2l_2 + 2l_3 + l_4) = -0.809675\,.
\end{aligned}$$

Це розв'язок у точці $x = 0.1$. Знайдемо ще розв'язок у точці $x = 0.2$.

$$\begin{aligned}
k_1 &= 0.1 \cdot (-0.809675) = -0.0809675\,, \\
l_1 &= 0.1 \cdot (0.909675 - 0.1 + 1) = 0.1809675\,, \\
k_2 &= -0.07191913\,, \qquad l_2 = 0.17191913\,, \\
k_3 &= -0.07237155\,, \qquad l_3 = 0.17237455\,, \\
k_4 &= -0.06373075\,, \qquad l_4 = 0.16373035\,,
\end{aligned}$$

$$\begin{aligned}
y_2 &= y_1 + \frac{1}{6}\,(k_1 + 2k_2 + 2k_3 + k_4) = 0.8374619\,, \\
z_2 &= z_1 + \frac{1}{6}\,(l_1 + 2l_2 + 2l_3 + l_4) = -0.6374618\,.
\end{aligned}$$

▷

Приклад 21. *Зараз можна розглянути приклад, що демонструє метод Штермера з попереднього розділу. Для цього знайдемо розв'язок рівняння*

$$\begin{aligned} &y'' = y - x + 1\,, \\ &y(0) = 1\,, \quad y'(0) = -1\,, \end{aligned} \tag{7.41}$$

в точках $x = 0.1\,,\ 0.2\,,\ 0.3\,,\ 0.4\,,\ 0.5$ *за допомогою методу Штермера (7.36), (7.37) з кроком* $h = 0.1$.

◁ Відповідно до цього методу для знаходження значення y_{n+1} необхідно знати y_n та y_{n-1}, тобто значення функції у двох попередніх точках. З цією метою для розгону використаємо метод Рунге-Кутта, для чого перетворимо (7.41) до системи ЗДР першого порядку

$$\begin{cases} y' &= z\,, \\ z' &= y - x + 1\,, \end{cases}$$

з початковими умовами $y(0) = 1$, $z(0) = -1$, що збігається з системою (7.40). Тобто можна використати розв'язок попереднього прикладу в точках $x = 0.1$ та 0.2 і повернутися до методу Штермера:

$$\begin{aligned} y_3^* &= -y_1 + 2y_2 + \frac{h^2}{12}(13f_2 - 2f_1 + f_0) = 0.7816385\,, \\ y_3 &= -y_1 + 2y_2 + \frac{h^2}{12}(f_3^* + 10f_2 + f_1) = 0.7816370\,, (\mathbf{0.7816366}) \\ y_4^* &= -y_2 + 2y_3 + \frac{h^2}{12}(13f_3 - 2f_2 + f_1) = 0.7406421\,, \\ y_4 &= -y_2 + 2y_3 + \frac{h^2}{12}(f_4^* + 10f_3 + f_2) = 0.7406408\,, (\mathbf{0.7406400}) \\ \\ y_5^* &= -y_3 + 2y_4 + \frac{h^2}{12}(13f_4 - 2f_3 + f_2) = 0.7130633\,, \\ y_5 &= -y_3 + 2y_4 + \frac{h^2}{12}(f_5^* + 10f_4 + f_3) = 0.7130621\,, (\mathbf{0.7130614}). \end{aligned}$$

Для порівняння в дужках наведені точні значення $y(x) = 2e^{-x} + x - 1$. ▷

Глава 8

Розв'язок рівнянь з частинними похідними

Більшість явищ гідродинаміки, електрики і магнетизму, механіки, оптики та переносу тепла можуть бути описані за допомогою рівнянь із частинними похідними (РЧП). Як приклад можна привести рівняння Максвелла, закон теплообміну Ньютона, рівняння Нав'є-Стокса, рівняння механіки у формі Лагранжа і Гамільтона, рівняння Шредингера в квантовій механіці, рівняння гідродинаміки. В усіх цих рівняннях фізичні явища описуються на мові просторових та часових похідних. Є цілий арсенал методів розв'язку РЧП, в яких рівняння з частинними похідними зводяться до звичайних диференціальних рівнянь, що розглянуті у розділі 7. Проте у багатьох випадках це неможливо, що приводить до необхідності наближеного розв'язку рівнянь з частинними похідними за допомогою комп'ютера, тобто використовується заміна РЧП кінцево-різницевим рівнянням.

У даному розділі будуть розглянути чисельні методи розв'язку рівнянь із частинними похідними другого порядку з двома незалежними змінними x, y вигляду

$$F(x, y, u_x, u_y, u_{xx}, u_{yy}, u_{xy}) = f(x, y)\,, \tag{8.1}$$

де u - шукана функція, $u_x, u_y, u_{xx}, u_{yy}, u_{xy}$ - перші та другі частинні похідні функції u по аргументах x та y.

Рівняння (8.1) називається *цілком лінійним*, якщо воно першої степені відносно шуканої функції, усіх її похідних та не містить їх добутків. Таке рівняння задається виразом:

$$au_{xx} + bu_{xy} + cu_{yy} + du_x + eu_y + gu = f(x, y), \tag{8.2}$$

у якому $a,\ b,\ c,\ d,\ e,\ g$ та f у загальному випадку є відомими функціями змінних x, y.

Визначимо дискримінант рівняння (8.2) $D(x,y)=b^2(x,y)-4a(x,y)c(x,y)$. Залежно від знаку функції D лінійне диференціальне рівняння з частинними похідними відноситься у даній області до одного із наступних типів:

$D<0$ – еліптичний тип,

$D=0$ – параболічний тип,

$D>0$ – гіперболічний тип,

D не зберігає постійного знаку – мішаний тип.

Рівняння еліптичного типу описують усталені процеси. Наприклад, температура $u(x,y)$ точки (x,y) пластинки при стаціонарному розподілі (тобто розподілі, що не залежить від часу) та відсутності джерел тепла відповідає рівнянню Лапласа

$$\triangle u = u_{xx} + u_{yy} = 0, \tag{8.3}$$

яке при наявності джерел тепла, що описуються функцією $f(x,y)$, переходить у рівняння Пуассона:

$$\triangle u = u_{xx} + u_{yy} = f(x,y). \tag{8.4}$$

Тут та у (8.3) $a=1,\ b=0,\ c=1$, тобто $D<0$.

Рівняння параболічного типу описують процеси теплопровідності та дифузії. Наприклад, температура $u(x,t)$ точки однорідного тонкого стержня з абсцисою x в кожний момент часу t відповідає одновимірному рівнянню теплопровідності

$$u_t - a^2 u_{xx} = f(x,t)\,,$$

де a - стала, що залежить від фізичних властивостей стержня, $f(x,t)$ - функція, що зв'язана з густиною розподілу джерел тепла. У відсутності джерел тепла $(f(x,t)=0)$ за допомогою заміни $\tau = a^2 t$ рівняння теплопровідності можна записати у вигляді

$$\frac{\partial u}{\partial \tau} = \frac{\partial^2 u}{\partial x^2}\,.$$

У цих двох рівняннях $b=c=0$, тобто $D=0$.

Рівняння гіперболічного типу описують коливальні системи і хвильові рухи. Наприклад, зміщення $u=u(x,t)$ точки однорідної струни з абсцисою x (рис.

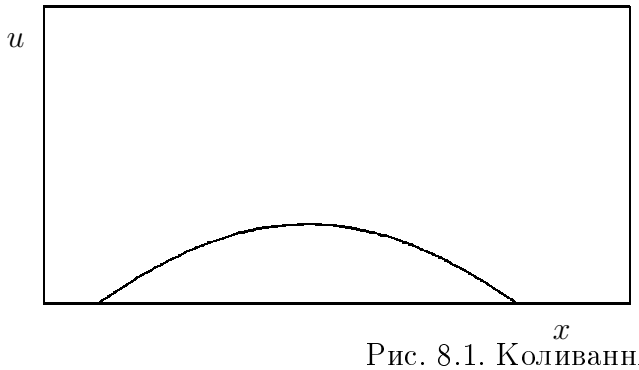

Рис. 8.1. Коливання струни

8.1) при наявності зовнішньої сили в момент часу t задовольняє неоднорідному хвильовому рівнянню

$$u_{tt} - a^2 u_{xx} = f(x,t), \tag{8.5}$$

де a - стала, $f(x,t)$ - функція, що залежить від зовнішньої сили. У відсутності зовнішньої сили ($f(x,t) = 0$, вільні коливання) рівняння коливань струни набуває вигляду:

$$u_{tt} - a^2 u_{xx} = 0. \tag{8.6}$$

Рівняння (8.5) та (8.6) відносяться до гіперболічного типу ($D > 0$).

8.1. Граничні та початкові умови. Задача Коші

Диференціальне РЧП в загальному випадку має нескінченну множину розв'язків, тому для однозначності отриманих розв'язків необхідно до рівняння додати деякі додаткові умови. У найпростішому випадку ці додаткові умови визначаються із початкових і граничних (крайових) умов. Відрізняються ці умови у випадку, коли одна із незалежних змінних РЧП відіграє роль часу, а інша - просторової координати. При цьому умови, що відносяться до початкового моменту часу, називаються *початковими*, а умови, що відносяться до фіксованих значень координат (звичайно це координати границі області, що розглядається), називаються *граничними* або *крайовими*. Як правило, крайові задачі ставляться для рівнянь еліптичного типу, а для рівнянь параболічного і гіперболічного типів розв'язуються задачі з початковими умовами, що називаються *задачами Коші*.

Розглянемо для визначеності постановку крайової задачі для рівнянь еліптичного типу, що є найбільш характерною. Запишемо лінійне диференціальне

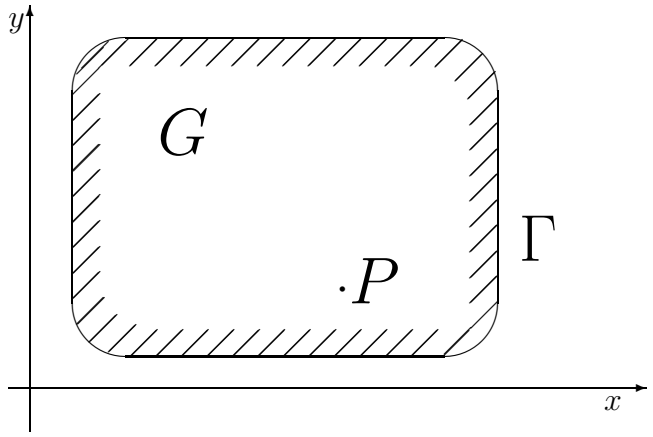

Рис. 8.2. Можлива область існування роз'язку

РЧП еліптичного типу в операторному вигляді:

$$L[u] = \triangle u + du_x + eu_y + gu = f(x, y), \tag{8.7}$$

де $d = d(x, y)$, $e = e(x, y)$, $g = g(x, y)$, $f(x, y)$ - неперервні функції, $L[u]$ - лінійний оператор.

Найчастіше зустрічаються три типи крайових задач.

Перша крайова задача. На контурі Γ, що обмежує область G (рис. 8.2), задана неперервна функція $\varphi(P) = \varphi(x, y)$. Тут $P = \{x, y\}$ - точка площини. Необхідно знайти функцію $u(P) = u(x, y)$, що задовольняє усередині G рівняння (8.7) і приймає на границі Γ значення $\varphi(P)$, тобто повинні виконатися умови

$$L[u(P)] = f(P),\ P \in G;\ u(P) = \varphi(P),\ P \in \Gamma.$$

Визначення значень шуканої функції на границі називається *граничною умовою першого типу.*

Друга крайова задача. Нехай на контурі Γ, що обмежує область G, задана неперервна функція $\varphi_1(P) = \varphi_1(x, y)$. Необхідно знайти функцію $u(P) = u(x, y)$, що задовольняє усередині G рівняння (8.7),і нормальна похідна до якої на границі Γ приймає задані значення $\varphi_1(P)$, тобто повинні виконатися умови

$$L[u(P)] = f(P),\ P \in G;\ \frac{\partial u(P)}{\partial n} = \varphi_1(P),\ P \in \Gamma.$$

Умова, коли на границі задається значення похідної шуканої функції, називається *граничною умовою другого типу.*

Третя крайова задача. Нехай на контурі Γ, що обмежує область G, задана неперервна функція $\psi(P) = \psi(x, y)$. Необхідно знайти функцію $u(P) = u(x, y)$, що

задовольняє усередині G рівняння (8.7),і для якої на границі Γ виконуються такі додаткові умови

$$L[u(P)]=f(P),\ P\in G;\ \ \alpha u(P)+\beta\frac{\partial u(P)}{\partial n}=\psi(P),\ P\in\Gamma. \tag{8.8}$$

де α та β відомі і $|\alpha|+|\beta|\neq 0$. Умова на границі (8.8) називається *граничною умовою третього типу.*

Третю крайову задачу можна розглядати як загальну, оскільки при $\alpha=1$ та $\beta=0$ отримаємо першу крайову задачу, а при $\alpha=0$ та $\beta=1$ - другу крайову задачу. Зазначимо, що для обмеженої області G відповідна крайова задача називається внутрішньою, а для необмеженої - зовнішньою. Для рівняння Лапласа (8.3) перша крайова задача називається задачею Діріхле, друга - задачею Неймана, третя - мішаною крайовою задачею.

Розглянемо загальну постановку задачі з початковими умовами. Нехай дано лінійне диференціальне РЧП

$$L[u(x,y)]=f(x,y), \tag{8.9}$$

де $L[u]$ визначається лівою частиною рівняння (8.2). Необхідно знайти розв'язок $u=u(x,y)$ рівняння (8.9), що задовольняє початкові умови Коші:

$$u(x,y_0)=\varphi(x),\ u_y(x,y_0)=\varphi_1(x). \tag{8.10}$$

Рівняння (8.9) при умовах (8.10) називається *задачею Коші.*

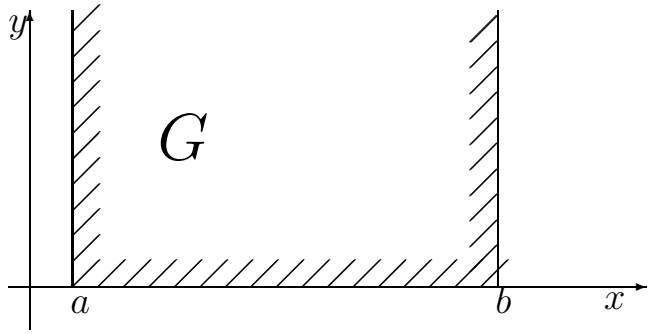

Рис. 8.3. Мішана задача

До мішаної задачі приводить необхідність отримати розв'язок рівняння (8.8) $u(x,y)$ на напівсмузі $G(a\leq x\leq b, 0\leq y\leq\infty)$ (див. рис. 8.3), для якого є заданими початкові умови на відрізку $a\leq x\leq b$. Для однозначності розв'язку необхідно також задати умови на прямих $x=a,\ y=b$. Необхідно знайти розв'язок

рівняння (8.9), що задовольняє початкові та крайові умови:

$$u(x,0)=\varphi(x),\ u_y(x,0)=\varphi_1(x),\ a\le x\le b,\ y=b,$$
$$\alpha_1 u(a,y)+\beta_1 u_x(a,y)=\psi_1(y),\ \ |\alpha_1|+|\beta_1|\neq 0,$$
$$\alpha_2 u(b,y)+\beta_2 u_x(b,y)=\psi_2(y),\ \ |\alpha_2|+|\beta_2|\neq 0.$$

Зазначимо, що мішана задача може бути поставлена для області довільної конфігурації, але розв'язок такої задачі буде мати реальну цінність тільки у тому випадку, коли мішана задача поставлена коректно, тобто невеликим відхиленням початкових і граничних умов будуть відповідати невеликі відхилення відповідного розв'язку. Наприклад, для рівнянь еліптичного типу задача Коші поставлена некоректно і тому звичайно не розглядається.

8.2. Рівняння еліптичного типу

Чисельні методи розв'язку лінійних диференціальних рівнянь РЧП ґрунтуються, як правило, на заміні похідних, що входять у рівняння, на кінцеві різниці. Тому ці методи називаються кінцево-різницевими. До них відноситься і так званий *метод сіток*, що є найчастіше уживаним. Викладемо його на прикладі розв'язку задачі Діріхле для рівняння

$$L[u]=au_{xx}+bu_{xy}+cu_x+du_y+gu=f(x,y), \tag{8.11}$$

де a,b,c,d,g – функції незалежних змінних x та y, що визначені в обмеженій області G з границею Γ (Рис. 8.2). Вважаємо, що функції неперервні в $G+\Gamma$, а g - недодатня в ній і $a(x,y)\cdot b(x,y)>0$.

Необхідно знайти розв'язок рівняння (8.11), що є неперервним в області $G+\Gamma$ і приймає в точках на границі певні значення $u(P)=\varphi(P),\ p\in\Gamma$, де $\varphi(P)$ – неперервна функція.

Для наближеного розв'язку цієї задачі проведемо дві сім'ї паралельних прямих

$$x_i=x_0+ih,\ i=0,\pm1,\pm2,\ldots,\ \ y_j=y_0+jl,\ j=0,\pm1,\pm2,\ldots.$$

Точки їх перетину назвемо вузлами, а h та l - кроками сітки по осях x та y (рис. 8.4).

Розглянемо ті вузли, що належать $G+\Gamma$. Вузли, у яких усі чотири сусідніх вузла належать цій множині, будемо називати внутрішніми, а множину внутрішніх вузлів назвемо сіточною областю G^*. Ті вузли, у яких хоча б один сусідній

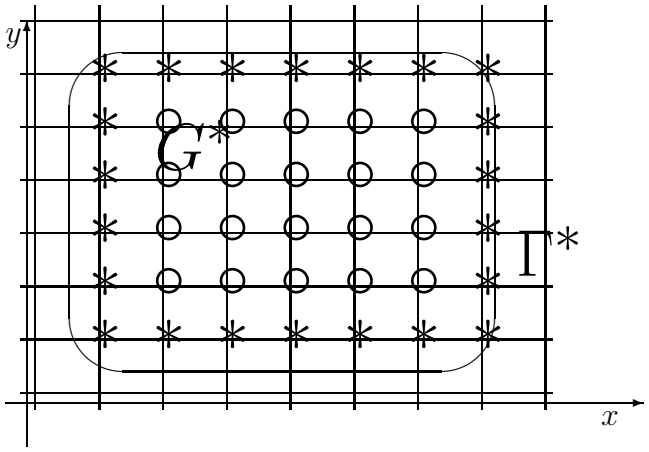

Рис. 8.4. Множина точок (∘) утворює область G^*, точки (∗) утворюють її границю Γ^*

вузол не належить множині G^*, назвемо граничними, а їх сукупність - границею сіточної області, позначивши Γ^*.

Складемо різницеве рівняння для кожного внутрішнього вузла (i,j), замінюючи в точці (x_0+ih, y_0+jl) похідні в рівнянні (8.11) кінцевими різницями,

$$\left(\frac{\partial u}{\partial x}\right)_{ij} \approx \frac{u_{i+1,j}-u_{i-1,j}}{2h}, \left(\frac{\partial^2 u}{\partial x^2}\right)_{ij} \approx \frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{h^2},$$

$$\left(\frac{\partial u}{\partial y}\right)_{ij} \approx \frac{u_{i,j+1}-u_{i,j-1}}{2l}, \left(\frac{\partial^2 u}{\partial y^2}\right)_{ij} \approx \frac{u_{i,j+1}-2u_{ij}+u_{i,j-1}}{l^2}, \tag{8.12}$$

де позначено $u_{ij} = u(x_i, y_j)$. Введемо позначення для коефіцієнтів рівняння (8.11) у вузлі: $a_{ij} = a(x_i, y_j)$ і т.д. Як результат отримаємо кінцево-різницеве рівняння для вузла (i,j):

$$\begin{aligned} L[u_{ij}] = a_{ij}\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{h^2} + b_{ij}\frac{u_{i,j+1}-2u_{ij}+u_{i,j-1}}{l^2} + \\ + c_{ij}\frac{u_{i+1,j}-u_{i-1,j}}{2h} + d_{ij}\frac{u_{i,j+1}-u_{i,j-1}}{2l} + g_{ij}u_{ij} = f_{ij}. \end{aligned} \tag{8.13}$$

Рівняння (8.13) можна записати для кожного внутрішнього вузла. Якщо вузол (i,j) є граничним, то u_{ij} в цьому вузлі визначається значенням функції φ на границі Γ в точці, що є найближчою до цього вузла.

Таким чином, для отримання розв'язку u у внутрішніх вузлах потрібно розв'язати повністю визначену систему лінійних алгебраїчних рівнянь (СЛАР). Розв'язавши таку систему, отримаємо наближене значення шуканого розв'язку на скінченній множині внутрішніх вузлів. Для інших значень x та y розв'язки можна отримати за допомогою інтерполяції.

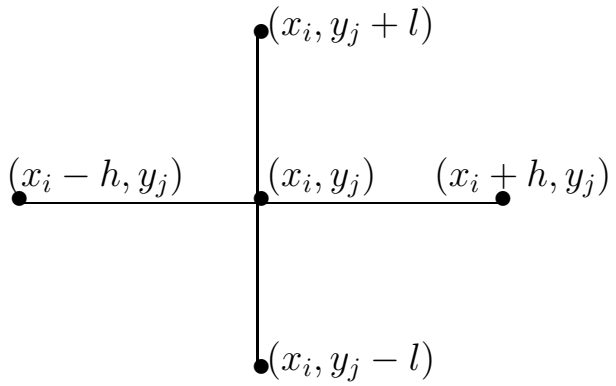

Рис. 8.5.

Як приклад розглянемо розв'язок рівняння Пуассона (8.4)

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y)\,,$$

яке при $f(x, y) = 0$ переходить у рівняння Лапласа (8.3). Візьмемо сітку, що побудована по п'яти вузлах, які, як показано на рис. 8.5, розташовані хрестом. Замінивши диференціал у рівнянні (8.4) на кінцеві різниці (8.12), отримаємо різницеве рівняння

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{ij} = h^2 f_{ij}. \tag{8.14}$$

Для спрощення тут обрані кроки сіток по x та y однаковими.

Зазначимо, що заміна диференціального РЧП на різницеве припускає наявність у розв'язку рівняння похідних необхідного порядку. Це у свою чергу накладає певні вимоги до коефіцієнтів рівняння в області G та до функцій, які входять до граничних умов. Тому не завжди треба застосовувати наближення більш високих порядків, оскільки це може призвести до ускладнення розрахунків, але не покращити точність результату.

Отримана система різницевих рівнянь (8.14) для внутрішніх вузлів сітки та відповідні рівняння на границя Г замінюють РЧП другого порядку еліптичного типу з лінійними граничними умовами і представляють собою СЛАР, кількість яких дорівнює кількості невідомих. Виникає питання про те, чи має розв'язки ця система? Для кожного конкретного РЧП необхідно перевірити коректність задачі та єдиність її розв'язку.

Сформулюємо без доказів низку властивостей функції $u(x, y)$, які визначають єдиність розв'язку задачі Діріхле (для рівняння Лапласа).

Властивість 1 (*принцип максимуму*). Гармонічна в обмеженій області функція, що неперервна в замкненій області $\overline{G} = G + \Gamma$, не може приймати всередині цієї області значень більших, ніж максимум її значень на границі, та менших, ніж мінімум її значень на Γ. Нагадаємо, що $u(x, y)$ називається гармонічною, якщо вона має неперервно частинні похідні другого порядку в області G та задовольняє в середині цієї області рівняння Лапласа.

Як наслідок, для функції $u(x, y)$, що є гармонічною в області G та неперервною в замкненій області $\overline{G}$, справедливі нерівності

$$\underline{u} \le u(x, y) \le \overline{u},$$

де $\underline{u} = \min u(x, y)$ на Γ та $\overline{u} = \max u(x, y)$ на Γ.

Властивість 2 (*єдиність розв'язку задачі Діріхле*). Задача Діріхле для замкненої та обмеженої області може мати єдиний розв'язок, оскільки не існує двох неперервних гармонічних функцій в замкненій та обмеженій області $\overline{G}$, які приймають на границі одні й ті самі значення.

Властивість 3 (*коректність задачі Діріхле*). Розв'язок задачі Діріхле для замкненої та обмеженої області неперервно залежить від граничних умов.

Таким чином, визначивши єдиність та коректність задачі Діріхле, можна використати для розв'язку системи лінійних алгебраїчних рівнянь (8.14) якийнебудь із методів лінійної алгебри. Найчастіше тут використовують метод простої ітерації чи метод Зейделя. Це пов'язано з великим порядком отриманих СЛАР.

8.3. Рівняння гіперболічного типу

Застосуємо метод сіток до розв'язку крайових задач для РЧП гіперболічного типу

$$L[u] = au_{xx} - bu_{yy} + cu_x + du_y + gu = f, \tag{8.15}$$

де a, b, c, d, f – задані функції незалежних змінних x та y, $a \cdot b > 0$.

Як приклад розглянемо розв'язок мішаної задачі $u(x, y)$ для рівняння (8.15) в області $G\{\beta \le x \le \gamma,\ y \ge 0\}$, яке задовольняє початкові умови

$$u|_{y=0} = \varphi(x),\ u_y|_{y=0} = \psi(x),\ \beta \le x \le \gamma, \tag{8.16}$$

якщо задана гранична умова 3-го роду:

$$(\alpha_1 u + \beta_1 u_x)|_{x=\beta} = F_1(y),\ (\alpha_2 u + \beta_2 u_x)|_{x=\gamma} = F_2(y),\ y \ge 0, \tag{8.17}$$

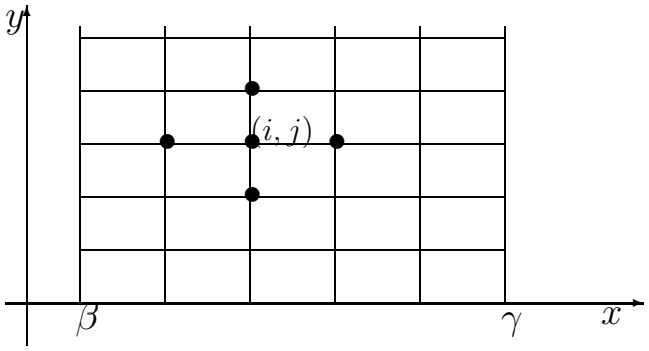

Рис. 8.6.

де φ, ψ, F_1, F_2 - задані функції змінних x та y.

При розв'язку РЧП гіперболічного типу методом сіток, всі схеми заміни диференціального рівняння кінцево-різнецевим, які застосовані тут, поділяються на явні та неявні. В явних різницевих схемах при будь-якому номері вузла розбиття (i, j) в кожне з різницевих рівнянь, які зв'язують значення шуканого розв'язку в горизонтальних рядках з номерами $j+1, j, j-1, \ldots, j-m$ входить лише одна точка рядку $j+1$ (рис. 8.5). Але явні схеми накладають значні обмеження на крок розбиття сітки і, відповідно, збільшують об'єм обчислень.

Для того, щоб обійти це обмеження, використовують різницеві схеми розв'язку РЧП, в яких для визначення значень розв'язку у вузлах $j+1$-го рядка при відомих значеннях у всіх попередніх рядках треба розв'язати систему рівнянь, яка пов'язує значення розв'язків у вузлах $j+1$-го рядка. Переваги кожної із схем будуть розглянуті згодом. Для ілюстрації розглянемо більш детально розв'язок мішаної задачі. Побудуємо на напівплощині $y \geq 0$ сітку прямокутників зі сторонами h та l по осях x та y (рис. 8.6). Вершини прямокутників назвемо внутрішніми вузлами, якщо $\beta < x_i < \gamma$ і $y > 0$, а вузли на прямих $x = \beta, x = \gamma, y = 0$, на яких визначені граничні та початкові значення, граничними вузлами.

Складемо різницеве рівняння для кожного внутрішнього вузла (i, j), замінивши в диференціальному рівнянні (8.15) похідні центральними кінцевими різницями (8.12), де $u_{ij} = u(x_i = ih, y_j = jl)$. Отримаємо різницеве рівняння

$$\begin{aligned} Lu_{ij} = a_{ij}\frac{u_{i+1,j}-2u_{ij}+u_{i-1,j}}{h^2} - b_{ij}\frac{u_{i,j+1}-2u_{ij}+u_{i,j-1}}{l^2}+ \\ +c_{ij}\frac{u_{i+1,j}-u_{i-1,j}}{2h} + d_{ij}\frac{u_{i,j+1}-u_{i,j-1}}{2l} = \\ =A_{ij}u_{i,j+1}+B_{ij}u_{i,j-1}+C_{ij}u_{i+1,j}+D_{ij}u_{i-1,j}+E_{ij}u_{ij}=f_{ij}, \end{aligned} \tag{8.18}$$

де запроваджені очевидні позначення для величин A_{ij}, B_{ij}, C_{ij}, D_{ij}, E_{ij}. З цього рівняння видно, що для відшукання розв'язку у $j+1$-му шарі використовуються

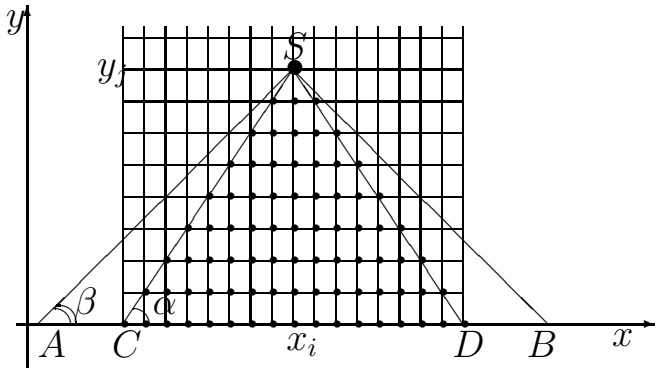

Рис. 8.7.

значення в двох попередніх шарах з номерами j та $j+1$, тобто використовується *явна різницева схема* розв'язку РЧП.

Для розв'язку (8.18) необхідно знати початкові значення шуканої функції в двох початкових шарах з $j = 0$ та $j = 1$. Їх можна знайти, наприклад, з початкових умов (8.16). Замінивши в початкових умовах похідну $u_x|_{y=0}$ різницевим відношенням $\dfrac{u_{i1} - u_{i0}}{l}$, отримаємо вираз для значень функції u_{ij} в двох перших рядках

$$u_{i0} = \varphi(ih) = \varphi_i,\ u_{i1} = \varphi_i + l\psi_i. \tag{8.19}$$

Якщо $x = \beta = i_{min}h$ або $x = \gamma = i_{max}h$, то в системі (8.18) величини $u_{i_{min},j}$ та $u_{i_{max},j}$ будуть визначатися кінцево-різницевими рівняннями, які випливають із (8.17):

$$\begin{aligned} \alpha_1 u_{i_{min},j} + \frac{\beta_1}{h}\left(u_{i_{min}+1,j} - u_{i_{min},j}\right) &= F_1(y_j), \\ \alpha_2 u_{i_{max},j} + \frac{\beta_2}{h}\left(u_{i_{max},j} - u_{i_{max}-1,j}\right) &= F_2(y_j). \end{aligned} \tag{8.20}$$

Таким чином, розв'язок задач (8.15) – (8.17) зводиться до розв'язку спільної системи лінійних алгебраїчних рівнянь (8.18) - (8.20). Як приклад розглянемо задачу Коші, в якій треба відшукати розв'язок рівняння (8.15) в області $G(y > 0)$, який задовольняє тільки початкові умови (8.16), коли $-\infty < x < \infty$. У цьому випадку ми також розв'язуємо СЛАР (8.18), але індекс "i" може приймати будь-яке ціле значення (додатне чи від'ємне). Якщо знати розв'язок у вузлах перших двох рядків (з номерами 0 та 1), то можна послідовно отримати значення розв'язку у 2-ому та 3-ому і т.д. рядках. (рис. 8.7). При цьому розв'язок у вузлі (i, j) буде визначатися даними на відрізку вісі $[C, D]$, якій

визначається прямими, що виходять із цього вузла та утворюють з віссю Ox кути, тангенси яких дорівнюють $\operatorname{tg}\alpha = \dfrac{l}{h}$. Трикутник CSD, що утворений цими прямими, називається *трикутником визначеності різницевої схеми.* Якщо з цього ж вузла провести дві характеристики до перетину з віссю Ox, то ми отримаємо *трикутник визначеності диференціального рівняння* ASB. У ньому $\operatorname{tg}\beta = \sqrt{\dfrac{a}{b}}$, де a та b - коефіцієнти у рівнянні (8.15).

Нехай крок l по вісі y відноситься до кроку h по вісі x так, що $\operatorname{tg}\alpha > \operatorname{tg}\beta$. Тоді трикутник визначеності різницевої схеми цілком розміщується всередині трикутника визначеності диференціального рівняння. Легко показати, що при постійному l та при прямуванні h до нуля таким чином, щоб точка (i, j) була весь час вузлом сітки, значення розв'язку, що отримане методом сіток, в цьому вузлі може не збігатися із справжнім значенням у цій точці.

Звідси можна зробити висновок, що для прямування послідовності наближених розв'язків до розв'язку задачі Коші (8.15), (8.16), отриманих за допомогою методу сіток при постійному відношенні $\operatorname{tg}\alpha = \dfrac{l}{h}$, повинна виконуватись умова $\operatorname{tg}\alpha \le \operatorname{tg}\beta$, тобто, трикутник визначеності РЧП повинен розміщуватись всередині трикутника визначеності різницевої схеми, який має ту саму вершину. Ця умова справедлива також і для криволінійних трикутників.

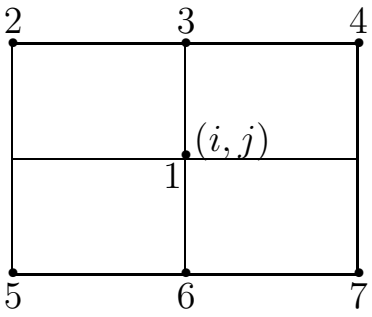

Рис. 8.8.

Обмеження, які накладаються на вибір кроків сітки трикутника визначеності різницевої схеми та РЧП, утворюють інколи значні обчислювальні труднощі, оскільки для порівняльно невеликого кроку сітки по одній координаті треба зробити дуже маленький крок по інший. Щоб уникнути цього, використовують *неявні схеми* розв'язку РЧП. Розглянемо (без доведення збіжності та стійкості) неявну схему для найпростішого РЧП гіперболічного типу:

$$L[u] = u_{xx} - u_{yy} = f(x, y). \tag{8.21}$$

Для цього використаємо квадратну ($l = h$) сітку та відповідну різницеву схему (рис. 8.8) для обчислення других похідних, коли u_{xx} у точці 1 з координатами (i, j) визначаються напівсумою виразів типу (8.12) для $(j + 1)$-го та $(j - 1)$-го рядків (точки з номерами 2, 3, 4 та 5, 6, 7 на рис. 8.8 відповідно), а u_{yy}

обчислюється по точках з номерами 3, 1, 6:

$$u_{xx}=\frac{1}{2}\left(\frac{u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1}}{h^2}+\frac{u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1}}{h^2}\right),$$
$$u_{yy}=\frac{u_{i,j+1}-2u_{i,j}+u_{i,j+1}}{h^2}.$$

У результаті для (8.21) отримаємо систему різницевих рівнянь (з точністю до h^2)

$$u_{i-1,j+1}-4u_{i,j+1}+u_{i+1,j+1}=-4u_{ij}-u_{i-1,j-1}-u_{i+1,j-1}+4u_{i,j-1}+h^2f_{ij} \tag{8.22}$$

яку необхідно розв'язати спільно з початковими умовами (8.19). Для мішаних задач необхідно також врахувати граничні умови (8.20). Відмітимо, що СЛАР (8.22) визначає значення функції $u(x,y)$ відразу на всьому $(j+1)$-му шарі. Такі системи зазвичай розв'язуються методом прогонки (розділ 4).

8.4. Рівняння параболічного типу

Запишемо лінійне РЧП параболічного типу

$$L[u]=u_t-a(x,t)u_{xx}-d(x,t)u_x-g(x,t)u=f(x,t), \tag{8.23}$$

де $a(x,t)>0$. Розглянемо спочатку розв'язок задачи Коші для цього рівняння, а потім розв'язок мішаної задачі.

В задачі Коші необхідно знайти розв'язок $u(x,t)$ рівняння (8.23) у напівплощині $y>0$, що задовольняє початкову умову

$$u(x,0)=\varphi(x),\ \ -\infty<x<\infty, \tag{8.24}$$

де $\varphi(x)$ - задана функція.

Як і у випадку РЧП інших типів, для пошуку розв'язку побудуємо прямокутну сітку, що утворена перетином двох сімейств паралельних прямих. Для кожного вузла (i,j) запишемо різницеве рівняння, що апроксимує РЧП (8.23), замінивши похідну по x кінцевими різницями (8.12), а похідну по t одним із трьох різницевих відношень:

$$\left(\frac{\partial u}{\partial t}\right)_{ij}=\frac{u_{i,j+1}-u_{ij}}{l}\ \text{ або }\ \frac{u_{ij}-u_{i,j-1}}{l}\ \text{ або }\ \frac{u_{i,j+1}-u_{i,j-1}}{2l}\,.$$

Відповідно до цього отримаємо три типи різницевої апроксимації РЧП (8.23):

$$L_1[u_{ij}] = \frac{1}{l}(u_{i,j+1}-u_{ij})-\frac{1}{h^2}a_{ij}(u_{i+1,j}-2u_{ij}+u_{i-1,j})- \\ -\frac{1}{2h}d_{ij}(u_{i+1,j}-u_{i-1,j}) - g_{ij}u_{ij} = f_{ij}\ ; \tag{8.25}$$

$$L_2[u_{ij}] = \frac{1}{l}(u_{ij}-u_{ij-1})-\frac{1}{h^2}a_{ij}(u_{i+1,j}-2u_{ij}+u_{i-1,j})- \\ -\frac{1}{2h}d_{ij}(u_{i+1,j}-u_{i-1,j}) - g_{ij}u_{ij} = f_{ij}\ ; \tag{8.26}$$

$$L_3[u_{ij}] = \frac{1}{2l}(u_{ij+1}-u_{ij-1})-\frac{1}{h^2}a_{ij}(u_{i+1,j}-2u_{ij}+u_{i-1,j})- \\ -\frac{1}{2h}d_{ij}(u_{i+1,j}-u_{i-1,j}) - g_{ij}u_{ij} = f_{ij}\ ; \tag{8.27}$$

Різницеве рівняння (8.25) містить значення розв'язку в чотирьох вузлах (рис. 8.9а) і апроксимує РЧП з точністю $O(l^2+h^2)$. Таку ж точність має рівняння (8.26), якому відповідає рис. 8.9б. Різницеве рівняння (8.27) містить значення в п'яти вузлах (рис. 8.9в) і має таку ж точність. Для вузлів нульового горизонтального рядку $j=0$ із початкової умови (8.24) випливає

$$u_{i0} = \varphi(ih) = \varphi_i,\ i = 0, \pm1, \pm2, \dots$$

Перша (L_1) і третя (L_3) різницеві схеми є явними, друга (L_2) – неявна.

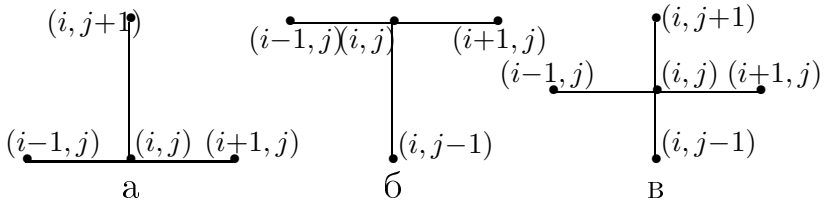

Рис. 8.9.

Як приклад розглянемо різницеве рівняння для рівняння теплопровідності

$$L[u] = \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x,t)\,. \tag{8.28}$$

Позначимо $\alpha = \dfrac{l}{h^2}$, тоді із рівнянь (8.25) - (8.27) отримаємо відповідно

$$u_{i,j+1} = (1-2\alpha)u_{ij} + \alpha(u_{i+1,j}+u_{i-1,j}) + \alpha h^2 f_{ij}, \\ (1+2\alpha)u_{ij} - \alpha(u_{i+1,j}+u_{i-1,j}) = u_{i,j-1} + \alpha h^2 f_{ij}, \\ u_{i,j+1} = 2\alpha(u_{i+1,j}-2u_{ij}+u_{i-1,j}) + +u_{i,j-1} + \alpha h^2 f_{ij}.$$

З погляду простоти розрахункових формул значення α потрібно обирати таким чином, щоб різницеве рівняння було якомога простішим, наприклад $\alpha = \frac{1}{2}$. У результаті отримаємо прості різницеві рівняння:

$$u_{i,j+1} = \frac{1}{2}(u_{i+1,j} + u_{i-1,j}) + \frac{1}{2}h^2 f_{ij}\,, \tag{8.29}$$

$$-u_{i-1,j} + 4u_{ij} - u_{i+1,j} = 2u_{i,j-1} + h^2 f_{ij}\,, \tag{8.30}$$

$$u_{i,j+1} = u_{i+1,j} - 2u_{ij} + u_{i-1,j} + \frac{1}{2}h^2 f_{ij}\,. \tag{8.31}$$

Для практичного використання найпростішою є перша схема, оскільки із початкових умов відомі розв'язки у вузлах початкового (нульового) рядка. При використанні другої схеми потрібно розв'язувати систему зв'язаних рівнянь (8.30), а при використанні третьої схеми потрібно якимось чином знайти значення розв'язків у вузлах першого рядка, після чого обчислення проводяться так само легко, як і за першою схемою.

З погляду апроксимації найбільш точною є третя схема, але вона є нестійкою відносно похибки обчислень. Покажемо це на прикладі так званої ε-схеми. Нехай обчислення проводяться за схемою (8.31) і на j-му кроці при розрахунку u_{ij} було припущено похибки ε. Вважаючи подальші розрахунки абсолютно точними, отримаємо похибку, що буде розповсюджуватись відповідно до таблиці 1.. Звідси видно, що мала похибка швидко зростає при переході до наступного шару, тим більше що похибки можуть утворюватись на кожному кроці і будуть певним чином взаємодіяти.

У таблиці 2. наведено ε-схему при заміні рівняння (8.28) різницевим рівнянням (8.30). Тут похибка не збільшується, а навіть зменшується. Така схема називається *стійкою*.

Таблиця 1.

	$i-6$	$i-5$	$i-4$	$i-3$	$i-2$	$i-1$	i
$j-1$	0	0	0	0	0	0	0
j	0	0	0	0	0	0	ε
$j+1$	0	0	0	0	0	ε	-2ε
$j+2$	0	0	0	0	ε	-4ε	7ε
$j+3$	0	0	0	ε	-6ε	17ε	-24ε
$j+4$	0	0	ε	-8ε	31ε	-68ε	89ε
$j+5$	0	ε	-10ε	49ε	-144ε	273ε	-338ε
$j+6$	ε	-12ε	71ε	-260ε	641ε	-1096ε	1311ε

Таблиця 2.

	$i-6$	$i-5$	$i-4$	$i-3$	$i-2$	$i-1$	i
$j-1$	0	0	0	0	0	0	0
j	0	0	0	0	0	0	ε
$j+1$	0	0	0	0	0	0.5ε	0
$j+2$	0	0	0	0	0.25ε	0	0.5ε
$j+3$	0	0	0	0.125ε	0	0.375ε	0
$j+4$	0	0	0.0625ε	0	0.25ε	0	0.375ε
$j+5$	0	0.0312ε	0	0.1562ε	0	0.3125ε	0
$j+6$	0.0156ε	0	0.0938ε	0	0.2344ε	0	0.3125ε

При переході до мішаної задачі необхідно знайти розв'язок рівняння типу (8.23) у напівполосі (Рис. 8.3) чи у прямокутнику $\{a \le x \le b, 0 \le t \le T\}$, що задовольняє початкову умову

$$u(x,0) = \varphi(x),\ a \le x \le b,$$

і граничні умови

$$[\beta_1 u_x + \gamma_1 u]_{x=a} = \psi_1(t),\ [\beta_2 u_x + \gamma_2 u]_{x=b} = \psi_2(t),$$

де $\beta_1, \beta_2, \gamma_1, \gamma_2, \psi_1, \psi_2$ - відомі функції змінної t. При $\beta_1 = \beta_2 \equiv 0,\ \gamma_1 = \gamma_2 \equiv 1$ маємо першу крайову задачу, а при $\beta_1 = \beta_2 \equiv 1,\ \gamma_1 = \gamma_2 \equiv 0$ – другу крайову задачу.

Розв'язок мішаної задачі методом сіток вимагає зробити не тільки заміну диференціального оператора кінцево-різницевим, але й провести різницеву апроксимацію граничних умов. Заміна оператора здійснюється за схемами (8.25) – (8.27), а граничні умови апроксимують із заданою точністю, що не є нижчою заданої точності наближення диференціальних операторів кінцевими різницями (див. рівняння (8.20)).

При будь-якому способі апроксимації для відшукання розв'язків мішаної задачі одержується стільки рівнянь, скільки є невідомих. Розв'язуючи цю систему лінійних алгебраїчних рівнянь, знайдемо наближені розв'язки в усіх вузлах сітки.

Для явних схем розв'язність системи не викликає сумнівів, для неявних її потрібно досліджувати у кожному конкретному випадку. Явні схеми (8.25) і (8.27) дозволяють досить просто обчислити значення шуканого розв'язку у вузлах j-го горизонтального рядка, якщо відомі розв'язки у попередніх рядках. Але ці схеми мають суттєвий недолік: для їх стійкості потрібно накласти сильні

обмеження на сітку. Наприклад, для схеми (8.25) повинна виконуватися умова $\alpha = \dfrac{l}{h^2} \leq \dfrac{1}{2}$, що приводить до дуже малого кроку l. Зменшення в процесі обчислення кроку по x тягне за собою різке зменшення кроку по t.

Неявна схема (8.26) звільнена від цього недоліку, але при її реалізації доводиться стикатися з іншою проблемою: для відшукання розв'язків у вузлах j-го горизонтального рядка треба розв'язувати систему лінійних алгебраїчних рівнянь із великою кількістю невідомих. Якщо для їх розв'язку застосовувати метод ітерацій, як для РЧП еліптичного типу, то із збільшенням кроку по часу, що припускається у даному випадку, значно зросте число ітерацій, які потрібні для розв'язку СЛАР із заданою точністю. Через це для розв'язку різницевої схеми (8.26) найчастіше використовується метод прогонки, що описаний в розділі 4.

Глава 9

Методи розв'язку інтегральних рівнянь

Інтегральними прийнято називати рівняння, в яких невідома функція входить під знак інтегралу. Як приклад можна назвати рівняння (6.43), до якого приводиться звичайне диференціальне рівняння першого порядку (6.36). Методи інтегральних рівнянь (ІР) широко використовуються в газовій динаміці і електродинаміці, квантовій механіці (рівняння Ліппмана-Швінгера, Фадєєва), в теорії переносу випромінювання і в багатьох інших областях.

Тут розглянемо методи розв'язку найбільш простих ІР, а саме лінійних ІР. Інтегральне рівняння вигляду

$$g(x)y(x) = f(x) + \lambda \int\limits_{\Omega} K(x,t)y(t)dt,\ x \in \Omega, \tag{9.1}$$

називається *лінійним*. Тут $y(x)$ - шукана функція з областю визначення Ω, яка може бути постійною або змінною (залежати від x), $f(x), g(x)$ - відомі функції з областю визначення Ω, $K(x,t)$ - ядро ІР.

Одним із найбільш важливих класів лінійних ІР є *рівняння Фредгольма*. Якщо $g(x) \equiv 0$, то ми маємо *рівняння Фредгольма 1 роду*

$$\int\limits_{\Omega} K(x,t)y(t)dt = f(x)\,,$$

де область Ω стала ($\Omega = [a,b]$). Якщо ж $g(x) \neq 0$, то (9.1) допускає ділення на $g(x)$; в результаті отримаємо *рівняння Фредгольма 2 роду*

$$y(x) = f(x) + \lambda \int\limits_{a}^{b} K(x,t)y(t)dt\,,\ x \in [a,b]\,. \tag{9.2}$$

Зазначимо, що інтервал $[a, b]$ може бути скінченним і нескінченним.

Якщо область Ω змінна, то при $g(x) \equiv 0$ отримаємо *рівняння Вольтерра 1 роду*:

$$\int_a^x K(x,t)y(t)dt = f(x)\,,$$

а при $g(x) \neq 0$ приходимо до *рівняння Вольтерра 2 роду*

$$y(x) = f(x) + \lambda \int_a^x K(x,t)y(t)dt\,. \tag{9.3}$$

Якщо в рівнянні (9.2) або (9.3) функція $f(x)$ дорівнює нулеві, то воно буде називатися *однорідним*.

Тут будуть розглянуті методи наближеного розв'язку інтегральних рівнянь Фредгольма; рівняння Вольтерра за допомогою перевизначення ядра можна розглядати як частковий випадок рівнянь Фредгольма.

9.1. Метод послідовних наближень

Запишемо рівняння (9.2) в операторному вигляді

$$y = f + \lambda K y\,, \tag{9.4}$$

де K - лінійний інтегральний оператор з ядром $K(x,t)$, дія якого на функцію визначається як $Ky = \int_a^b K(x,t)y(t)dt$. Внаслідок лінійності рівняння (9.4) можна переписати як

$$(I - \lambda K)y = f\,,$$

де I - одиничний оператор. Тоді $y = (I - \lambda K)^{-1} f$, що при підстановці в праву частину (9.4) дає

$$y = f + \lambda K(I - \lambda K)^{-1} f = f + R_\lambda f\,. \tag{9.5}$$

Інтегральний оператор $R_\lambda = \lambda K(I - \lambda K)^{-1}$ називається *резольвентою* оператора λK і має ядро, що задовольняє інтегральному рівнянню

$$R_\lambda(x,t) = \lambda K(x,t) + \lambda \int_a^b K(x,s)R_\lambda(s,t)ds\,.$$

Цей оператор визначає розв'язок вихідного ІР

$$y(x) = f(x) + \int_a^b R_\lambda(x,t) f(t) dt \,.$$

Видно, що резольвента не залежить від вільного члена, а визначається лише "внутрішніми" властивостями рівняння, його ядром.

Метод послідовних наближень полягає в представленні розв'язку через нескінченний ряд по степенях параметра λ. Це еквівалентно такому запису

$$R_\lambda = \lambda K(I + \lambda K + (\lambda K)^2 + \ldots) = \sum_{n=1}^{\infty} (\lambda K)^n \,.$$

Використаємо як початкове наближення для функції $y(x)$ значення $f(x)$, тобто $y_0(x) = f(x)$, і підставимо його в праву частину (9.2):

$$y_1(x) = f(x) + \lambda \int_a^b K(x,t) y_0(t) dt = f(x) + \lambda \int_a^b K(x,t) f(t) dt \,.$$

Отримано перше наближення функції $y_1(x)$, яке можна підставити в праву частину (9.2) і отримати друге наближення:

$$y_2(x) = f(x) + \lambda \int_a^b K(x,t) y_1(t) dt =$$

$$= f(x) + \lambda \int_a^b K(x,t) f(t) dt + \lambda^2 \int_a^b \int_a^b K(x,t) K(t,s) f(s) ds \,.$$

Продовжуючи цей процес за допомогою рекурентної формули

$$y_{n+1}(x) = f(x) + \lambda \int_a^b K(x,t) y_n(t) dt \,,$$

при граничному наближенні можна отримати шуканий розв'язок із наперед заданою точністю. При цьому практичною ознакою близькості отриманих розв'язків до шуканої функції є досягнення малої величини різниці двох наближень, що йдуть одне за одним.

Достатньою умовою застосування методу послідовних наближень є нерівність

$$b^2 = \lambda^2 \int_a^b \int_a^b K(x,t)K(t,x)dtdx < 1\,, \tag{9.6}$$

яка і за формою, і за змістом нагадує умову (4.38) для збіжності метода простої ітерації при розв'язку систем лінійних алгебраїчних рівнянь.

Приклад 22. *Методом послідовних наближень розв'язати ІР*

$$y(x) = 1 + \int_0^1 xt^2 y(t)dt\,. \tag{9.7}$$

$\lhd$ Для визначення збіжності методу обчислимо відповідно до (9.6) ($\lambda = 1$)

$$b^2 = \int_0^1 \int_0^1 x^3t^3 dtdx = \frac{1}{16} < 1\,.$$

Далі $y_0(x) = 1$,

$$y_1(x) = 1 + \int_0^1 xt^2 \cdot 1dt = 1 + \frac{x}{3} = 1 + 0.3333333x,$$

$$y_2(x) = 1 + \int_0^1 xt^2 \cdot (1 + \frac{t}{3})dt = 1 + \frac{5}{12}x = 1 + 0.4166667x,$$

$$y_3(x) = 1 + \int_0^1 xt^2 \cdot (1 + \frac{5}{12}t)dt = 1 + \frac{7}{16}x = 1 + 0.4375000x,$$

$$y_4(x) = 1 + \int_0^1 xt^2 \cdot (1 + \frac{7}{16}t)dt = 1 + \frac{85}{192}x = 1 + 0.4427083x,$$

$$y_5(x) = 1 + \int_0^1 xt^2 \cdot (1 + \frac{85}{192}t)dt = 1 + \frac{341}{768}x = 1 + 0.4440104x.$$

Ця послідовність наближень досить швидко прямує до точного розв'язку (див.(9.18))

$$y(x) = 1 + \frac{4}{9}x \approx 1 + 0.4444444x\,. \tag{9.8}$$

$\rhd$

9.2. Метод скінченних сум

Найчастіше на практиці використовують метод, в якому задача розв'язку інтегрального рівняння зводиться до розв'язку апроксимуючої системи лінійних алгебраїчних рівнянь, що отримуються в результаті використання деякої квадратурної формули для обчислення інтегралу в правій частині.

Розглянемо деяку квадратурну формулу

$$\int_a^b \varphi(x)dx = \sum_{i=1}^{n} \omega_i \varphi(x_i) + r(\varphi)\,,$$

де ω_i, x_i ваги та вузли квадратурної формули, φ – залишковий член (похибка) квадратурної формули. Запроваджуючи позначення $y_i = y(x_i)$, $K_{ij} = K(x_i, x_j)$, $f_i = f(x_i)$, отримаємо для рівняння (9.2) представлення

$$y_i = f_i + \lambda \sum_{j=1}^{n} \omega_j K_{ij} y_j + \lambda r_i\,,\ i = \overline{1,n}. \tag{9.9}$$

Якщо відкинути малі похибки λr_i, то система рівнянь (9.9) дає можливість отримати наближені значення функції $y_i \approx Y_i$ у вузлах квадратурної формули

$$\sum_{j=1}^{n} \left[\delta_{ij} - \lambda \omega_j K_{ij}\right] Y_j = f_i\,. \tag{9.10}$$

Після розв'язку цієї системи отримаємо аналітичний вираз для наближеного розв'язку

$$y(x) \approx Y(x) = f(x) + \lambda \sum_{j=1}^{n} \omega_j K(x, x_j) Y_j. \tag{9.11}$$

Приклад 23. *Розв'язати рівняння (9.7) методом скінчених сум за допомогою квадратурної формули Сімпсона.*

Обираємо вузли $x_1 = 0$, $x_2 = \frac{1}{2}$, $x_3 = 1$ та відповідні ваги $\omega_1 = \omega_3 = \frac{1}{6}$, $\omega_2 = \frac{2}{3}$. Тоді система (9.10) буде мати вигляд:

$$\begin{array}{rcrcl} Y_1 & & & = & 1, \\ & \frac{11}{12}Y_2 & - \frac{1}{12}Y_3 & = & 1, \\ & -\frac{1}{6}Y_2 & + \frac{5}{6}Y_3 & = & 1. \end{array}$$

Розв'язок цієї системи $Y_1 = 1$, $Y_2 = \dfrac{11}{9}$, $Y_3 = \dfrac{13}{9}$ відповідно до (9.11) приводить до загального наближеного розв'язку

$$Y(x) = 1 + \frac{2}{3}x \cdot \frac{1}{4} \cdot \frac{11}{9} + \frac{1}{6}x \cdot 1 \cdot \frac{13}{9} = 1 + \frac{4}{9}x\,,$$

що збігається з точним розв'язком (9.8). Цей збіг є наслідком того, що шукана функція є поліномом першої степені, а квадратурна формула Сімпсона є точною для усіх поліномів третього степеня.

9.3. Метод вироджених ядер

Ядро інтегрального рівняння називається *виродженим*, якщо його можна представити у вигляді

$$K(x,t) = \sum_{i=1}^{n} \alpha_i(x)\beta_i(t)\,. \tag{9.12}$$

Наявність вироджених ядер в ІР Фредгольма дозволяє застосувати простий та ефективний метод, що приводить до розв'язку в явному вигляді.

Підставимо (9.12) у (9.2):

$$y(x)=f(x)+\lambda\sum_{i=1}^{n}\alpha_i(x)\int_a^b \beta_i(t)y(t)dt=f(x)+\lambda\sum_{i=1}^{n}C_i\alpha_i(x), \tag{9.13}$$

де використане позначення

$$C_i = \int_a^b \beta_i(t)y(t)dt\,. \tag{9.14}$$

Тепер підставимо в інтеграл (9.14) праву частину (9.13)

$$C_i = \int_a^b \beta_i(t)\left[f(t) + \lambda\sum_{i=1}^{n} C_i\alpha_i(t)\right]dt = \gamma_i + \lambda\sum_{i=1}^{n}\omega_{ij}C_j\,,$$

або

$$\sum_{j=1}^{n}\left[\delta_{ij} - \lambda\omega_{ij}\right]C_j = \gamma_i\,,\ i = \overline{1,n}, \tag{9.15}$$

де використані позначення

$$\gamma_i = \int_a^b \beta_i(t) f(t) dt\,, \quad \omega_{ij} = \int_a^b \beta_i(t)\alpha_j(t) dt\,.$$

Розв'язуючи систему лінійних алгебраїчних рівнянь відносно C_i та підставляючи ці значення в (9.13), отримаємо розв'язок в остаточному вигляді.

Особливо простим цей метод буде у випадку, коли в (9.12) входить тільки один доданок $K(x,t) = \alpha(x)\beta(t)$; такі ядра називаються *сепарабельними*. Тоді із (9.15) одразу випливає

$$C = \frac{\int_a^b \beta(t) f(t) dt}{1 - \lambda \int_a^b \alpha(t)\beta(t) dt}\,, \tag{9.16}$$

та розв'язок дорівнює

$$y(x) = f(x) + \lambda C \alpha(x)\,. \tag{9.17}$$

Якщо ядро вихідного рівняння не є виродженим, то в багатьох випадках його з деякою мірою точності можна замінити на таке. Наприклад,

$$K(x,t) = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{\partial^n}{\partial x^n} K(x_0,t)(x - x_0)^n.$$

Зрозуміло, що ядро можна розкласти і по іншій змінній або по обох змінних одразу, а потім ряд обірвати.

Приклад 24. *Розв'язати рівняння (9.7) методом вироджених ядер.*

◁ Оскільки ядро в (9.7) сепарабельне, $K(x,t) = x \cdot t^2$, то спираючись на (9.16) та (9.17) маємо

$$C = \frac{\int_0^1 t^2 \cdot 1 dt}{1 - \int_0^1 t \cdot t^2 dt} = \frac{4}{9} \quad \text{та} \quad y(x) = 1 + \frac{4}{9}x\,. \tag{9.18}$$

▷

9.4. Метод найменших квадратів

Визначимо нев'язку інтегрального рівняння (9.2)

$$U\left(y(x)\right) = y(x) - f(x) - \lambda \int\limits_a^b K(x,t)y(t)dt\,. \tag{9.19}$$

Якщо замість точного розв'язку $y(x)$ у виразі (9.19) підставити деякий наближений розв'язок

$$y(x) \cong Y(x) = \Phi(x, c_1, \ldots, c_n)\,, \tag{9.20}$$

де вигляд функції Φ відомий, а невідомі значення параметрів $\{c_i\}$, то нев'язка (9.19) не обов'язково дорівнює нулеві, і наша задача буде полягати в тому, щоб знайти такі вільні параметри $\{c_i\}$, для яких нев'язка буде малою у деякому розумінні.

Визначимо інтеграл

$$I = \int\limits_a^b U^2\left(\Phi(x, c_1, \ldots, c_n)\right) dx\,. \tag{9.21}$$

Метод найменших квадратів полягає у знаходженні c_i із умові мінімуму (9.21), тобто з умов $\dfrac{\partial I}{\partial c_i} = 0,\ i = \overline{1,n}$. Використовуючи явні вирази (9.19) та (9.20) отримаємо систему рівнянь

$$\begin{aligned}
&\int\limits_a^b \Phi(x,\{c_j\})\frac{\partial\Phi(x,\{c_j\})}{\partial c_i}dx - \int\limits_a^b f(x)\frac{\partial\Phi(x,\{c_j\})}{\partial c_i}dx + \\
&+\lambda\int\limits_a^b\int\limits_a^b K(x,t)f(x)\frac{\partial\Phi(t,\{c_j\})}{\partial c_i}dxdt - \\
&-\lambda\int\limits_a^b\int\limits_a^b \left[K(x,t)+K(t,x)\right]\frac{\partial\Phi(t,\{c_j\})}{\partial c_i}\Phi(t,\{c_j\})dxdt + \\
&+\lambda^2\int\limits_a^b\int\limits_a^b\int\limits_a^b K(x,t)K(t,s)\Phi(t,\{c_j\})\frac{\partial\Phi(s,\{c_j\})}{\partial c_i}dsdtdx
\end{aligned} \tag{9.22}$$

відносно параметрів c_i.

У загальному випадку система (9.22) є досить складною, тому частіше за все найбільш зручно наближений розв'язок представляти у вигляді функції, що лінійно залежить від параметрів c_i, тобто у вигляді узагальнених поліномів

$$y(x) \cong Y(x) = \sum_{i=1}^{n} c_i \varphi_i(x), \tag{9.23}$$

коли $\varphi_i(x)$ - відомі лінійно незалежні функції, які називаються координатними функціями. У даному випадку система (9.22) набуває вигляду

$$\sum_{j=1}^{n} a_{ij}(\lambda) c_j = b_i(\lambda), \tag{9.24}$$

де

$$\begin{aligned} a_{ij}(\lambda) &= \int_a^b \left[\varphi_i(x) - \lambda \int_a^b K(x,t)\varphi_i(t)dt \right] \cdot \\ &\quad \cdot \left[\varphi_j(x) - \lambda \int_a^b K(x,s)\varphi_j(s)ds \right] dx, \\ b_i(\lambda) &= \int_a^b f(x) \left[\varphi_i(x) - \lambda \int_a^b K(x,t)\varphi_i(t)dt \right] dx. \end{aligned} \tag{9.25}$$

Із (9.25) випливає, що матриця системи (9.24) симетрична та, якщо виконується умова (9.6), ця матриця також додатньо визначена. Тоді, якщо λ не є власним значенням ядра $K(x,t)$, то система (9.24) завжди розв'язна та $Y(x) \to y(x)$ при $n \to \infty$.

Приклад 25. *Розв'язати рівняння (9.7) методом найменших квадратів.*

$\lhd$ Як координатні функції оберемо систему алгебраїчних поліномів $\varphi_1(x) = 1$,

$\varphi_2(x) = x$, $\varphi_3(x) = x^2$. Тоді

$$a_{11} = \int_0^1 \left[1 - \int_0^1 xt^2 \cdot 1dt\right] \left[1 - \int_0^1 xs^2 \cdot 1ds\right] dx = \frac{19}{27};$$

$$a_{12} = a_{21} = \int_0^1 \left[1 - \int_0^1 xt^2 \cdot 1dt\right] \left[x - \int_0^1 xs^2 \cdot sds\right] dx = \frac{7}{24};$$

$$a_{13} = a_{31} = \int_0^1 \left[1 - \int_0^1 xt^2 \cdot 1dt\right] \left[x^2 - \int_0^1 xs^2 \cdot s^2ds\right] dx = \frac{31}{180};$$

$$a_{22} = \int_0^1 \left[x - \int_0^1 xt^2 \cdot tdt\right] \left[x - \int_0^1 xs^2 \cdot sds\right] dx = \frac{3}{16};$$

$$a_{23} = a_{32} = \int_0^1 \left[x - \int_0^1 xt^2 \cdot tdt\right] \left[x^2 - \int_0^1 xs^2 \cdot s^2ds\right] dx = \frac{11}{80};$$

$$a_{33} = \int_0^1 \left[x^2 - \int_0^1 xt^2 \cdot t^2dt\right] \left[x^2 - \int_0^1 xs^2 \cdot s^2ds\right] dx = \frac{17}{150};$$

$$b_1 = \int_0^1 1 \cdot \left[1 - \int_0^1 xt^2 \cdot 1dt\right] dx = \frac{5}{6};$$

$$b_2 = \int_0^1 1 \cdot \left[x - \int_0^1 xt^2 \cdot tdt\right] dx = \frac{3}{8};$$

$$b_3 = \int_0^1 1 \cdot \left[x^2 - \int_0^1 xt^2 \cdot t^2dt\right] dx = \frac{7}{30}.$$

Система (9.24) набуває вигляду

$$\begin{cases} \frac{19}{27}c_1 + \frac{7}{24}c_2 + \frac{31}{180}c_3 = \frac{5}{6}, \\ \frac{7}{24}c_1 + \frac{3}{16}c_2 + \frac{11}{80}c_3 = \frac{3}{8}, \\ \frac{31}{180}c_1 + \frac{11}{80}c_2 + \frac{17}{150}c_3 = \frac{7}{30}, \end{cases}$$

та її розв'язок дорівнює

$$c_1 = 1,\ c_2 = \frac{4}{9},\ c_3 = 0 \ \ \Rightarrow Y(x) = 1 + \frac{4}{9}x,$$

що збігається з точним розв'язком.

9.5. Метод колокації

Згідно методу колокації наближений розв'язок (9.20) знаходиться з умови того, що нев'язка $U(\Phi(x, c_1, \ldots, c_n))$ перетворюється у нуль в заданій системі точок $\{x_i\}$ на відрізку $[a, b]$, тобто покладається

$$U(\Phi(x, c_1, \ldots, c_n)) = 0,\ i = \overline{1, n}.$$

Зокрема, для узагальнених поліномів (9.23) ця вимога приводить до системи лінійних алгебраїчних рівнянь

$$\begin{aligned} &\sum_{j=1}^{n} c_j \varphi_j(x_i) - f(x_i) - \lambda \int_a^b K(x_i, t) \sum_{j=1}^{n} c_j \varphi_j(t) dt = 0 \Rightarrow \\ &\Rightarrow \sum_{j=1}^{n} a_{ij} c_j = f(x_i), \end{aligned} \tag{9.26}$$

де $a_{ij} = \varphi_j(x_i) - \lambda \int_a^b K(x_i, t)\varphi_j(t)dt$. Якщо визначник системи $D(\lambda) = \det a_{ij} \neq 0$, то однозначно отримаємо значення c_i і знаходимо наближений розв'язок відповідно до (9.23).

Приклад 26. *Розв'язати рівняння (9.7) методом колокації.*

$\lhd$ Оберемо систему функцій $\varphi_1(x) = 1$, $\varphi_2(x) = x$, $\varphi_3(x) = x^2$ та систему точок $x_1 = 0$, $x_2 = \frac{1}{2}$ $x_3 = 1$. Тоді

$$a_{11} = 1,\ a_{12} = 0,\ a_{13} = 0,$$

$$a_{21} = 1 - \frac{1}{2}\int_0^1 t^2 \cdot 1dt = \frac{5}{6},\ a_{22} = \frac{1}{2} - \frac{1}{2}\int_0^1 t^2 \cdot tdt = \frac{3}{8},$$

$$a_{23} = \frac{1}{4} - \frac{1}{2}\int_0^1 t^2 \cdot t^2dt = \frac{3}{20},\ a_{31} = 1 - \frac{1}{2}\int_0^1 t^2 \cdot 1dt = \frac{2}{3},$$

$$a_{32} = 1 - \int_0^1 t^2 \cdot tdt = \frac{3}{4},\ a_{33} = 1 - \int_0^1 t^2 \cdot t^2dt = \frac{4}{5},$$

$$f(x_1) = f(x_2) = f(x_3) = 1.$$

Система (9.26) набуває вигляду

$$\begin{cases} c_1 & & & = & 1, \\ \frac{5}{6}c_1 & + & \frac{3}{8}c_2 & + & \frac{3}{20}c_3 & = & 1, \\ \frac{2}{3}c_1 & + & \frac{3}{4}c_2 & + & \frac{4}{5}c_3 & = & 1, \end{cases}$$

та її розв'язок, як і в прикладі 25., дорівнює

$$c_1 = 1\,,\ c_2 = \frac{4}{9}\,,\ c_3 = 0 \ \ \Rightarrow Y(x) = 1 + \frac{4}{9}x\,'$$

що збігається з точним розв'язком. $\rhd$

9.6. Метод моментів

Цей метод, який іноді називають методом Бубнова-Галеркіна, полягає у наступному. Наближений розв'язок (9.23) знаходиться у вигляді

$$y(x) \cong Y(x) = f(x) + \sum_{i=1}^{n} c_i\varphi_i(x)\,,$$

де невідомі коефіцієнти c_i знаходяться із умови ортогональності нев'язки $U(Y(x))$ до кожної із лінійно незалежних функцій $\varphi_i(x),\ i = \overline{1,n}$.

Визначимо, що дві функції $\varphi(x)$ та $\psi(x)$ ортогональні на інтервалі $[a, b]$, якщо їх скалярний добуток дорівнює нулеві, тобто

$$(\varphi, \psi) = \int_a^b \varphi(x)\psi(x)dx = 0\,.$$

Умова ортогональності приводить до системи лінійних алгебраїчних рівнянь:

$$\left(f + \sum_{j=1}^{n} c_j\varphi_j - f - \lambda K(f + \sum_{j=1}^{n} c_j\varphi_j), \varphi_i \right) = 0,$$

або

$$\sum_{j=1}^{n} \left[(\varphi_j, \varphi_i) - \lambda(K\varphi_j, \varphi_i)\right] c_j - \lambda(Kf, \varphi_i) = 0.$$

Це еквівалентно такій системі рівнянь відносно c_i

$$\sum_{j=1}^{n} a_{ij}c_j = b_i, \tag{9.27}$$

де

$$a_{ij} = \int_a^b \varphi_i(x)\varphi_j(x)dx - \lambda \int_a^b \int_a^b K(x,t)\varphi_i(x)\varphi_j(t)dtdx,$$

$$b_i = \lambda \int_a^b \int_a^b K(x,t)\varphi_i(x)f(t)dtdx.$$

Приклад 27. *Розв'язати рівняння (9.7) методом моментів.*

$\lhd$ Виберемо систему функцій у вигляді $\varphi_1(x)=1$, $\varphi_2(x)=x$, $\varphi_3(x)=x^2$. Тоді

$$a_{11}=\int\limits_0^1 1\cdot 1dx-\int\limits_0^1\int\limits_0^1 xt^2dtdx=\frac{5}{6},\ a_{12}=\int\limits_0^1 xdx-\int\limits_0^1\int\limits_0^1 xt^2\cdot tdtdx=\frac{3}{8},$$

$$a_{13}=\int\limits_0^1 x^2dx-\int\limits_0^1\int\limits_0^1 xt^2\cdot t^2dtdx=\frac{7}{30},\ a_{21}=\int\limits_0^1 xdx-\int\limits_0^1\int\limits_0^1 xt^2\cdot xdtdx=\frac{7}{18},$$

$$a_{22}=\frac{1}{4},\ a_{23}=\frac{11}{60},\ a_{31}=\frac{1}{4},\ a_{32}=\frac{3}{16},\ a_{33}=\frac{3}{20},$$

$$b_1=\int\limits_0^1\int\limits_0^1 xt^2dtdx=\frac{1}{6},\ b_2=\int\limits_0^1\int\limits_0^1 xt^2xdtdx=\frac{1}{9},\ b_3=\int\limits_0^1\int\limits_0^1 xt^2x^2dtdx=\frac{1}{12}.$$

Система (9.27) набуває вигляду

$$\begin{cases} \frac{5}{6}c_1 + \frac{3}{8}c_2 + \frac{7}{30}c_3 = \frac{1}{6}, \\ \frac{7}{18}c_1 + \frac{1}{4}c_2 + \frac{11}{60}c_3 = \frac{1}{9}, \\ \frac{1}{4}c_1 + \frac{3}{16}c_2 + \frac{3}{20}c_3 = \frac{1}{12}. \end{cases}$$

ЇЇ розв'язок, як і в прикладах 25. та 26., дорівнює

$$c_1=1,\ c_2=\frac{4}{9},\ c_3=0\ \ \Rightarrow Y(x)=1+\frac{4}{9}x,$$

що збігається з точним розв'язком. $\rhd$

Оглавление