

ЧИСЕЛЬНІ МЕТОДИ

М. В. Кутнів

УДК 519.6(075.8)

ББК 22.19 я 73

К 95

Рекомендовано Міністерством освіти і науки України як посібник для студентів вищих навчальних закладів

(Лист № 1.4/18-Г-2813 від 23.12 2008 р.)

Рецензенти:

Макаров В.Л., доктор фізико-математичних наук, професор, академік НАН України, Інституту математики НАН України, м. Київ;

Слоньовський Р.В., доктор фізико-математичних наук, професор Національного університету “Львівська політехніка”;

Хапко Р.С., доктор фізико-математичних наук, професор Львівського національного університету імені Івана Франка.

Кутнів М.В.

К 95 Чисельні методи: Навчальний посібник.— Львів: Видавництво “Растр-7”, 2010. —286 с.

ISBN

У навчальному посібнику розглядаються основи чисельних методів розв’язування задач алгебри, аналізу, звичайних диференціальних рівнянь та рівнянь математичної фізики. Крім традиційних чисельних методів, таких, як методи розв’язування систем лінійних алгебраїчних рівнянь, нелінійних рівнянь та їх систем, апроксимації, чисельного диференціювання та інтегрування, методи розв’язування задач Коші та крайових задач для звичайних диференціальних рівнянь, розглянуто питання апроксимації сплайнами, розв’язування жорстких задач Коші, основні принципи побудови та дослідження різницевих схем для рівнянь математичної фізики, метод скінченних елементів та низку інших питань, які викликають методичний та практичний інтерес. Посібник написано на основі курсу лекцій, які читалися впродовж багатьох років для студентів спеціальності “Прикладна математика” Національного університету “Львівська політехніка”. Посібник розрахований на студентів вищих технічних навчальних закладів, які спеціалізуються з прикладної та обчислювальної математики. Може бути корисним для студентів інших спеціальностей, аспірантів.

УДК 519.6(075.8)

ББК 22.19 я 73

ISBN

©Кутнів М.В., 2010

©“Растр-7”, 2010

ЗМІСТ

Вступ	7
Математичне моделювання та обчислювальний експеримент	7
Класифікація похибок	8
Запис чисел у комп'ютері	8
Абсолютна і відносна похибка	9
Похибка заокруглень	10
 РОЗДІЛ 1. Чисельні методи лінійної алгебри	 12
1.1. Прямі методи розв'язування систем лінійних алгебраїчних рівнянь	12
1.1.1. Метод Гаусса	13
1.1.2. Метод прогонки	22
1.2. Норми та обумовленість матриць систем лінійних алгебраїчних рівнянь	26
1.3. Ітераційні методи	28
1.3.1. Приклади та канонічний вигляд ітераційних методів	28
1.3.2. Матричні нерівності та дії з ними	32
1.3.3. Дослідження збіжності ітераційних методів	36
1.3.4. Оцінка швидкості збіжності стаціонарних ітераційних методів	41
1.3.5. Многочлени Чебишева	44
1.3.6. Ітераційний метод з чебишевським набором параметрів	46
1.4. Обчислення власних значень та власних векторів матриці	50
1.4.1. Алгебраїчна проблема власних значень	50
1.4.2. QR -розклад	54
1.4.3. QR -алгоритм знаходження власних значень та власних векторів	57
Контрольні завдання	61
 РОЗДІЛ 2. Методи розв'язування нелінійних рівнянь та систем	 64
2.1. Чисельне розв'язування нелінійних рівнянь	64
2.1.1. Метод ділення навпіл (метод дихотомії або бісекції)	64
2.1.2. Метод послідовних наближень (простої ітерації)	65
2.1.3. Метод Ньютона (метод дотичних)	66
2.1.4. Метод січних	69
2.2. Розв'язування систем нелінійних рівнянь	72
2.2.1. Метод послідовних наближень (простої ітерації)	72
2.2.2. Метод Ньютона	73
Контрольні завдання	74

РОЗДІЛ 3. Наближення функцій	76
3.1. Постановка задачі наближення функції	76
3.2. Інтерполяційний многочлен Лагранжа	77
3.3. Розділені різниці. Інтерполяційна формула Ньютона	78
3.4. Оцінка залишкового члена інтерполяційного многочлена	84
3.5. Оптимальний вибір вузлів інтерполяції	86
3.6. Розділені різниці та інтерполювання з кратними вузлами	87
3.7. Найкраще наближення в лінійному нормованому просторі	91
3.8. Найкраще наближення в гільбертовому просторі	93
3.9. Метод найменших квадратів	95
3.10. Інтерполяція сплайнами	98
3.11. Чисельне диференціювання	102
Контрольні завдання	103
РОЗДІЛ 4. Чисельне інтегрування	104
4.1. Наближене обчислення інтегралів. Інтерполяційні квадратурні формули	104
4.2. Квадратурні формули Ньютона–Котеса	105
4.3. Квадратурні формули Гаусса	113
4.4. Практична оцінка похибки квадратурних формул	119
4.5. Наближене обчислення невластивих інтегралів	120
Контрольні завдання	122
РОЗДІЛ 5. Чисельні методи розв’язування задачі Коші для звичайних диференціальних рівнянь	123
5.1. Задача Коші для звичайних диференціальних рівнянь	123
5.2. Метод рядів Тейлора	125
5.3. Методи Рунге–Кутта	126
5.4. Практична оцінка похибки та вибір довжини кроку для методів Рунге–Кутта	133
5.5. Лінійні багатокрокові методи	138
5.5.1. Методи Адамса	138
5.5.2. Формули диференціювання назад	141
5.5.3. Порядок апроксимації лінійних багатокрокових методів	144
5.5.4. Стійкість багатокрокових методів	148
5.6. Методи Нордсіка	149
5.7. Чисельне інтегрування жорстких систем звичайних диференціальних рівнянь	152
5.7.1. Поняття жорсткої задачі	152
5.7.2. Абсолютна стійкість чисельних методів	156

5.8. Реалізація лінійних неявних багатокрокових методів	160
Контрольні завдання	164
РОЗДІЛ 6. Чисельне розв'язування крайових задач для звичайних диференціальних рівнянь	
6.1. Крайові задачі для звичайних диференціальних рівнянь	166
6.2. Метод стрільби	167
6.2.1. Метод однократної стрільби	167
6.2.2. Метод многократної стрільби	170
6.3. Метод скінченних різниць	173
6.3.1. Метод заміни похідних скінченними різницями	173
6.3.2. Метод неозначених коефіцієнтів	178
6.3.3. Інтегро-інтерполяційний метод побудови різницевої схеми	180
6.3.4. Збіжність різницевої схеми	187
6.4. Варіаційно-проекційні методи та метод скінченних елементів	190
6.4.1. Метод Рунге	190
6.4.2. Метод Гальоркіна	195
6.4.3. Побудова сіткової схеми методом скінченних елементів	197
6.4.4. Збіжність методу скінченних елементів	201
Контрольні завдання	205
РОЗДІЛ 7. Чисельне розв'язування рівнянь з частинними похідними	
7.1. Крайові задачі для рівнянь з частинними похідними	208
7.2. Основні поняття методу сіток	209
7.3. Сіткові схеми як операторні рівняння	214
7.3.1. Запис сіткових схем у вигляді операторних рівнянь	214
7.3.2. Задача на власні значення для оператора другої різницевої похідної	215
7.3.3. Властивості власних значень та власних функцій	217
7.3.4. Операторні нерівності	219
7.4. Різницева схема для одновимірного рівняння теплопровідності	220
7.4.1. Різницева схема з ваговими коефіцієнтами	220
7.4.2. Порядок апроксимації різницевої схеми з ваговими коефіцієнтами	224
7.4.3. Апроксимація крайових умов третього роду	225
7.5. Різницева схема для рівняння коливання струни	230
7.5.1. Різницева схема з ваговими коефіцієнтами	230
7.5.2. Порядок апроксимації різницевої схеми з ваговими коефіцієнтами	232
7.6. Стійкість двохвирусних та трихвирусних сіткових схем	233

7.6.1. Канонічний вигляд та умови стійкості двоярусних сіткових схем	233
7.6.2. Стійкість різницевої схеми з ваговими коефіцієнтами для рівняння теплопровідності	237
7.6.3. Канонічний вигляд та умови стійкості триярусних сіткових схем	239
7.6.4. Стійкість різницевої схеми з ваговими коефіцієнтами для рівняння коливання струни	244
7.7. Різницева апроксимація задачі Діріхле для рівняння Пуассона .	247
7.8. Принцип максимуму для різницевих схем	249
7.8.1. Принцип максимуму	249
7.8.2. Стійкість та збіжність різницевої задачі Діріхле	254
7.8.3. Монотонні різницеві схеми	256
7.9. Метод скінченних елементів розв'язування задачі Діріхле для рівняння Пуассона	259
7.10. Методи розв'язування сіткових рівнянь	262
7.10.1. Запис сіткових схем розв'язування задачі Діріхле для рівняння Пуассона в операторному вигляді	263
7.10.2. Швидке дискретне перетворення Фур'є	265
7.10.3. Прямі методи. Метод розділення змінних	268
7.10.4. Застосування ітераційних методів для розв'язування задачі Діріхле	271
7.11. Чисельне розв'язування багатовимірних задач теплопровідності	275
7.11.1. Різницеві схеми для багатовимірних рівнянь теплопровідності	275
7.11.2. Метод змінних напрямків побудови економних різницевих схем	276
7.11.3. Локально-одновимірний метод	279
Контрольні завдання	282
Список літератури	285

ВСТУП

Математичне моделювання та обчислювальний експеримент

Математичне моделювання стає все більш поширеним засобом теоретичного дослідження складних науково-технічних задач. Технологією побудови і аналізу за допомогою комп'ютера математичних моделей є обчислювальний експеримент. Обчислювальний експеримент поєднує у собі вибір фізичної та математичної моделі явища, яке вивчається, розробку чисельних методів і алгоритмів розв'язування отриманої задачі, програмування обчислювального алгоритму, проведення обчислень і аналіз результатів. Результати розрахунків порівнюють з наявними даними спостережень натуральних експериментів і у випадку необхідності модифікують математичну модель, чисельні методи та програми, які їх реалізують. На основі моделей, які пройшли таку перевірку виникає можливість прогнозувати поведінку досліджуваного явища.

Предметом курсу “Чисельні методи” є викладення питань, які відображають тільки один з етапів обчислювального експерименту, а саме етап побудови та дослідження обчислювальних методів. Процес побудови чисельного методу для математичної моделі передбачає формулювання дискретної моделі та розробку алгоритму розв'язування дискретної задачі. Одній і тій же математичній задачі можна поставити у відповідність множину різних дискретних задач. Дискретна модель повинна бути адекватною неперервній моделі, тобто правильно передавати фізичні особливості процесів, які вивчаються. Поведінка розв'язку дискретної моделі повинна правильно відображати якісну поведінку розв'язку вихідної задачі. І, нарешті, чисельний метод повинен бути *збіжним*, тобто розв'язок дискретної задачі повинен прямувати до точного розв'язку задачі за умови, що параметри чисельного методу прямують до певних граничних значень. Збіжність чисельного методу тісно зв'язана з його *коректністю*. Припустимо, що математична модель коректна, тобто існує її єдиний розв'язок, який неперервно залежить від вхідних даних. Тоді відповідна дискретна модель повинна бути коректною, тобто однозначно розв'язною і стійкою. Під стійкістю

розуміють неперервну залежність дискретної моделі від вхідних даних. Крім того, дискретна модель повинна допускати можливість реалізації на комп'ютері, тобто можливість отримати розв'язок дискретної задачі за прийнятний час. Реальні обчислювальні алгоритми повинні бути *економними* за числом дій, а також за необхідним обсягом пам'яті.

Класифікація похибок

Процес дослідження вихідного об'єкта методом *математичного моделювання* і *обчислювального експеримента* носить наближений характер. Побудова математичної моделі пов'язана з спрощенням досліджуваного явища, недостатньо точним заданням коефіцієнтів рівняння та інших вхідних даних. По відношенню до чисельних методів, які використовуються для дослідження математичної моделі, вказані похибки є неусувними. При переході від математичної моделі до чисельного методу виникають похибки, які називаються *похибками методу*. Вони зв'язані з тим, що будь-який чисельний метод наближено відтворює відповідну математичну модель. Оскільки реальний комп'ютер не може оперувати при граничних значеннях параметрів за яких дискретна модель збігається, то важливо вміти оцінювати похибку чисельного методу. Найбільш типовими похибками методу є *похибки дискретизації* та *похибки заокруглень*. Різниця між розв'язком дискретної задачі та точним розв'язком вихідної задачі, називається похибкою дискретизації. Оскільки вхідні дані задаються в комп'ютері не точно, а з заокругленням, то в процесі виконання обчислювального алгоритму похибки заокруглень нагромаджуються, і в результаті розв'язок, отриманий на комп'ютері, буде відрізнятися від точного розв'язку дискретної задачі. Результируюча похибка називається похибкою заокруглення (або обчислювальною похибкою). Величина цієї похибки визначається двома факторами: точністю подання дійсних чисел у комп'ютері та чутливістю даного алгоритму до похибок заокруглень (стійкістю).

Запис чисел у комп'ютері

Наближене число можна записати у вигляді

$$x = \pm \sum_{k=1}^t \alpha_k q^{-k} = \pm (\alpha_1, \alpha_2, \dots, \alpha_t), \quad (1)$$

де q — ціла основа системи числення, $\alpha_1, \alpha_2, \dots, \alpha_t$ — цілі, які знаходяться в межах $0 \leq \alpha_k < q$. Такий запис дійсного числа x називається записом з *фіксованою комою*.

Друга форма запису з *плаваючою комою*, найбільш поширена в комп'ютерах, призначених для наукових розрахунків

$$x = \pm q^p \sum_{k=1}^t \alpha_k q^{-k} = \pm q^p (\alpha_1, \alpha_2, \dots, \alpha_t),$$

де p — порядок числа, який задовольняє нерівність $|p| \leq p_0$. Найбільш поширений випадок двійкової системи числення $q = 2$. Випадки $q = 8$ і $q = 10$ використовуються як допоміжні на етапах підготовки і видачі даних.

У сучасних комп'ютерах використовуються записи чисел як з фіксованою комою, так і з плаваючою. В багатьох випадках число розрядів t може задаватися користувачем.

Абсолютна і відносна похибка

Якщо a — точне значення деякої величини, а a^* — відоме наближення до нього, то *абсолютною похибкою* наближеного значення a^* називають величину $\Delta(a^*)$, для якої

$$|a - a^*| \leq \Delta(a^*).$$

Відносною похибкою наближеного значення називають величину $\delta(a^*)$, для якої

$$\left| \frac{a - a^*}{a^*} \right| \leq \delta(a^*).$$

Відносну похибку часто виражають у процентах.

Значущими цифрами числа називають всі цифри в його записі, починаючи з першої не нульової зліва.

Значущу цифру називають *вірною*, якщо абсолютна похибка числа не перевищує одиниці розряду, яка відповідає цій цифрі.

Якщо всі значущі цифри вірні, то кажуть, що число записане з усіма вірними цифрами.

Якщо a^* є наближеним значенням числа a з абсолютною похибкою $\Delta(a^*)$, то точне значення записують у вигляді

$$a = a^* \pm \Delta(a^*),$$

Числа a^* і $\Delta(a^*)$ прийнято записувати з однаковою кількістю знаків після коми. Відповідно, якщо a^* є наближеним значенням числа a з відносною похибкою $\delta(a^*)$, то

$$a = a^*(1 \pm \delta(a^*)).$$

Похибка заокруглень

Обмеження на порядки чисел у комп'ютері $|p| \leq p_0$ іноді приводить до припинення обчислень, в інших випадках відносно невелика розрядність чисел приводить до недопустимого спотворення результатів. Такі алгоритми, коли внаслідок обмеженості p або малості t виникають подібні ефекти, називають *нестійкими*.

Побудова *стійких* алгоритмів, при використанні яких спотворення кінцевого результату обчислювальною похибкою знаходиться в допустимих межах, є важливою задачею теорії обчислювальних методів.

Розглянемо приклад, який показує, що підвищення точності інколи можливо досягнути за рахунок нескладних алгебраїчних перетворень.

Нехай потрібно обчислити корінь рівняння $y^2 - 140y + 1 = 0$. Припустимо, що обчислення на комп'ютері здійснюються в десятковій системі числення, причому в мантисі числа після заокруглення утримується 4 розряди. Тоді будемо мати

$$y = 70 - \sqrt{4899}, \quad \sqrt{4899} = 69,992\dots,$$

Після заокруглення отримаємо

$$\sqrt{4899} \approx 69,99, \quad y \approx 70 - 69,99 = 0,01.$$

Це ж саме значення y можна обчислити ще так. Позбудемося від ірраціональності в чисельнику, тоді отримаємо $y = 1/(70 + \sqrt{4899})$. Послідовно проводячи обчислення, отримаємо $\sqrt{4899} \approx 69,99$, $70 + 69,99 = 139,99$ і після заокруглення

$$70 + 69,99 = 140,0.$$

Нарешті

$$1/140 = 0,00714285\dots,$$

і після заокруглення

$$y \approx 0,007143.$$

Проводячи обчислення з додатковими розрядами, можна перевірити, що в обох випадках всі підкреслені цифри результатів вірні. Однак, в другому випадку точність результату суттєво вища. Зазначимо, що в першому випадку віднімалися близькі великі числа (70 і 69,99). Оскільки ці числа були великі, то вони заокруглені з великою абсолютною похибкою. Отже, при відніманні близьких величин виникає явище зникання значущих цифр. Це явище досить часто приводить до суттєвого спотворення результату при розв'язуванні систем лінійних алгебраїчних рівнянь.

РОЗДІЛ 1

ЧИСЕЛЬНІ МЕТОДИ ЛІНІЙНОЇ АЛГЕБРИ

1.1. Прямі методи розв'язування систем лінійних алгебраїчних рівнянь

У розділах 1.1, 1.3 розглядаються чисельні методи розв'язування систем лінійних алгебраїчних рівнянь (СЛАР)

$$A\mathbf{x} = \mathbf{f}, \quad (1.1)$$

де $A = \{a_{ij}\}_{i,j=1}^n$ — матриця порядку n , $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ — невідомий вектор, $\mathbf{f} = (f_1, f_2, \dots, f_n)^T$ — заданий вектор. Будемо припускати, що визначник матриці A відмінний від нуля. Отже, розв'язок системи (1.1) існує і єдиний.

Методи чисельного розв'язування систем (1.1) ділять на два типи: *прямі та ітераційні*. За допомогою прямих (або точних) методів розв'язок СЛАР знаходять за скінченне число арифметичних дій. Зазначимо, що внаслідок похибок заокруглень при розв'язуванні задач за допомогою комп'ютерів прямі методи не дають точного розв'язку і називати їх точними можна тільки нехтуючи похибками заокруглень. Порівняння різних прямих методів проводиться за кількістю арифметичних дій за великих n , необхідних для одержання розв'язку. Перевага надається за інших рівних умов методу з меншим числом дій. З курсу алгебри відомі два основні методи розв'язування СЛАР: формули Крамера та метод виключення (метод Гаусса). За великих n перший спосіб, який ґрунтується на обчисленні визначників вимагає порядку $n!$ дій, тоді як метод Гаусса тільки $O(n^3)$ дій. Тому метод Гаусса в різних варіантах широко використовують при розв'язуванні задач лінійної алгебри за допомогою комп'ютерів.

Ітераційні методи полягають в тому, що розв'язок \mathbf{x} системи (1.1) знаходять як границю при $k \rightarrow \infty$ послідовності наближень \mathbf{x}_k , де k

додавання і віднімання, обмежимося обчисленням множень і ділень. Розглянемо спочатку прямий хід. На k -му кроці для обчислення l_{ik} , $i = \overline{k+1, n}$, $k = \overline{1, n-1}$, необхідно

$$\sum_{k=1}^{n-1} \sum_{i=k+1}^n 1 = \sum_{k=1}^{n-1} (n-k) = \sum_{k=1}^{n-1} k = n(n-1)/2$$

ділень, а для обчислення $a_{ij}^{(k)}$, $k = \overline{1, n-1}$ за формулами (1.5)

$$\begin{aligned} \sum_{k=1}^{n-1} \sum_{i=k+1}^n \sum_{j=k+1}^n 1 &= \sum_{k=1}^{n-1} (n-k)^2 = \sum_{k=1}^{n-1} k^2 = \sum_{k=1}^{n-1} (2C_k^2 + C_k^1) = \\ &= 2C_n^3 + C_n^2 = (n-1)n(2n-1)/6 \end{aligned}$$

множень. Отже, обчислення елементів $a_{ij}^{(k)}$, $k = \overline{1, n-1}$ матриці системи (1.6) потребує

$$(n-1)n(2n-1)/6 + n(n-1)/2 = (n^2-1)n/3$$

операцій множення та ділення. Обчислення правих частин $f_i^{(k)}$ за формулами (1.5) вимагає $\sum_{k=1}^n (n-k) = n(n-1)/2$ множень. Для прямого ходу методу Гаусса необхідно виконати

$$(n^2-1)n/3 + n(n-1)/2 = (2n+5)(n-1)n/6$$

операцій множення та ділення. За великих n це число дій дорівнює приблизно $n^3/3$. Для здійснення оберненого ходу методу Гаусса за формулами (1.7) потрібно $\sum_{k=1}^n k = n(n+1)/2 \approx n^2/2$ операцій множення та ділення. Отже, для великих n число дій множення і ділення у методі Гаусса приблизно дорівнює $n^3/3$.

Метод Гаусса компактно записують в матричній формі. Покажемо, як при цьому цей метод зв'язаний з LU -розкладом (факторизацією) матриці A .

Введемо нижню трикутну матрицю:

$$L_k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & 0 \\ & & 1 & & \\ & & -l_{k+1,k} & \ddots & \\ 0 & & \vdots & & \ddots \\ & & -l_{n,k} & & & 1 \end{pmatrix},$$

яку називають також *матрицею Фробеніуса*. Тоді k -й крок виключення еквівалентний множенню системи (1.4) на L_k зліва. Зауважимо, що

$$L_k^{-1} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & 0 \\ & & 1 & & \\ & & l_{k+1,k} & \ddots & \\ 0 & & \vdots & & \ddots \\ & l_{n,k} & & & 1 \end{pmatrix},$$

$$L = L_1^{-1} L_2^{-1} \dots L_{n-1}^{-1} = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & & 0 \\ l_{31} & l_{32} & \ddots & & \\ \dots & \dots & \dots & 1 & \\ l_{n1} & l_{n2} & \dots & l_{n,n-1} & 1 \end{pmatrix}.$$

Справджується також і більш загальне твердження

$$L_{n-1} L_{n-2} \dots L_1 A = L^{-1} A = U,$$

де U — верхня трикутна матриця системи (1.6).

Метод виключення Гауса еквівалентний такому процесу:

- 1) виконати LU -розклад матриці A ;
- 2) розв'язати систему рівнянь $L\mathbf{y} = \mathbf{f}$;
- 3) розв'язати систему $U\mathbf{x} = \mathbf{y}$.

Така матрична форма алгоритму Гауса зручна при розв'язанні послідовності систем лінійних алгебраїчних рівнянь з однією і тією ж матрицею і різними правими частинами, оскільки в цьому випадку достатньо лише один раз здійснити LU -розклад матриці A і розв'язати послідовність систем з трикутними матрицями. Процес LU -розкладу матриці A потребує $n^3/3$ операцій множення, тоді як розв'язування системи рівнянь із трикутною матрицею порядку $n^2/2$ операцій.

За допомогою методу виключення можна обчислити також визначник матриці A . Справді, оскільки визначник добутку двох матриць є добутком їх визначників, унаслідок LU -розкладу матриці A маємо

$$\det A = \det LU = \det L \cdot \det U = a_{11} \cdot a_{22}^{(1)} \dots a_{nn}^{(n-1)}.$$

Отже, визначник є добутком діагональних елементів приведеної до трикутного вигляду матриці і для обчислення потрібно тільки $n - 1$ додаткових множень.

Процес виключення Гаусса є також найкращим методом обертання матриці A . Нехай \mathbf{e}_i — вектор, i -й елемент якого дорівнює 1, а решта нульові. Тоді вектор \mathbf{e}_i буде i -м стовпцем одиничної матриці I і з співвідношення $AA^{-1} = I$ випливає, що i -й стовпець матриці A^{-1} є розв'язком системи лінійних рівнянь $A\mathbf{x} = \mathbf{e}_i$. Отже, розв'язуючи n систем рівнянь

$$A\mathbf{x}^{(i)} = \mathbf{e}_i, \quad i = \overline{1, n} \quad (1.8)$$

і, використовуючи знайдені вектори $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ як стовпці, одержимо обернену матрицю A^{-1} . Для ефективного розв'язування задачі (1.8) з однією і тією ж матрицею можна використати метод виключення Гаусса у вигляді LU -розкладу.

Неважко показати, що метод Гаусса можна застосувати лише в тому разі, коли всі головні мінори відмінні від нуля. Запобігти вказаному обмеженню дає змогу метод Гаусса з вибором головного елемента у стовпці. Основна ідея полягає в тому, що на кожному кроці вибираємо головним елементом найбільший за модулем у стовпці. Цього можна досягнути переставлянням рядків системи. Тоді, якщо $\det A \neq 0$, то у процесі обчислень не буде ділення на нуль. k -й крок алгоритму Гаусса з вибором головного елемента у стовпчику має вигляд:

- 1) Знайти $m \geq k$ таке, що $|a_{mk}^{(k-1)}| = \max \left\{ |a_{ik}^{(k-1)}|, i \geq k \right\}$. Якщо $a_{mk}^{(k-1)} = 0$, то A — вироджена і алгоритм закінчуємо, інакше поміняти місцями $a_{kj}^{(k-1)}$ і $a_{mj}^{(k-1)}$, $j = \overline{k, n}$, а також f_k і f_m .
- 2) Обчислити $l_{ik} = a_{ik}^{(k-1)} / a_{kk}^{(k-1)}$, $i = \overline{k+1, n}$, $k = \overline{1, n-1}$ і $a_{ij}^{(k)}$, $i, j = \overline{k+1, n}$, $f_i^{(k)}$, $i = \overline{k+1, n}$, $k = \overline{1, n-1}$ за формулами (1.5).

Зазначимо, що в алгоритмі з вибором головного елемента $|l_{ik}| \leq 1$.

Крім вибору головного елемента в стовпчику, можна вибрати головний елемент у рядку, а також у всій матриці. Алгоритм вибору головного елемента у двовимірному масиві (матриці) використовують рідко через велику кількість операцій.

Опишемо тепер метод Гаусса у вигляді LU -розкладу з вибором головного елемента у стовпчику.

При переставлянні рядків матриці алгоритм виключення Гаусса не буде еквівалентний розкладу матриці A на добуток нижньої і верхньої трикутних матриць. Однак його можна модифікувати.

Введемо матрицю *переставлень* P розміру $n \times n$, у кожному рядку і кожному стовпці якої є лише один елемент, відмінний від нуля і

рівний одиниці. Переставлення рядків матриці A може бути здійснено за допомогою множення зліва на відповідну матрицю переставлень P . Нехай P_k , $k = \overline{1, n-1}$ матриця переставлень, отримана з одиничної матриці тим же переставленням рядків, що застосовувалося на k -му кроці виключення. Тоді послідовно домножимо матрицю A зліва на P_1, P_2, \dots, P_{n-1} , а далі здійснимо трикутну факторизацію:

$$P_{n-1}P_{n-2} \dots P_2P_1A = PA = LU,$$

де $P = P_{n-1}P_{n-2} \dots P_2P_1$ — матриця переставлень. Як відомо, обернена матриця до матриці переставлень є матрицею переставлень. Отже, $A = (P^{-1}L)U$ і перший множник у розкладі A є переставленням нижньої трикутної матриці, а другий як і раніше верхньої трикутної матриці. Якщо на k -му кроці ніяких переставлень не здійснюється, то матриця P_k є одиничною матрицею.

Використання методу Гаусса з вибором головного елемента потребує додаткових затрат для переставлення рядків, тому для систем, стосовно яких відомо, що переставлення рядків не потрібне, ефективнішим є використання звичайного методу Гаусса. Серед таких систем найбільш вживані — системи з діагонально переважаючими та додатно визначеними матрицями.

Унаслідок впливу похибок заокруглень метод Гаусса дозволяє одержати лише наближений розв'язок \mathbf{x}_1 системи (1.1). Якщо обчислення проведені з великою похибкою, то можна використати ітераційне уточнення розв'язку, кожен крок якого описується формулами:

$$\begin{aligned} \mathbf{f}_m &= \mathbf{f} - A\mathbf{x}_m, \\ A\mathbf{r}_m &= \mathbf{f}_m, \\ \mathbf{x}_{m+1} &= \mathbf{x}_m + \mathbf{r}_m, \quad m = 1, 2, \dots \end{aligned} \tag{1.9}$$

Розв'язування системи (1.9) зводиться до розв'язування трикутних систем

$$L\mathbf{y} = \mathbf{f}_m, \quad U\mathbf{r}_m = \mathbf{y},$$

що вимагає порядку n^2 операцій.

Приклад 1.1. Методом Гаусса обчисліть визначник матриці

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 4 \end{pmatrix}.$$

▷ Оскільки $\det(A) = a_{11}a_{22}^{(1)}a_{33}^{(2)}$, то проведемо обчислення за схемою

$$\begin{aligned} l_{21} &= \frac{a_{21}}{a_{11}} = 2, & l_{31} &= \frac{a_{31}}{a_{11}} = 3, & a_{22}^{(1)} &= a_{22} - l_{21}a_{12} = -1, \\ a_{23}^{(1)} &= a_{23} - l_{21}a_{13} = -2, & a_{32}^{(1)} &= a_{32} - l_{31}a_{12} = -2, \\ a_{33}^{(1)} &= a_{33} - l_{31}a_{13} = -5, & l_{32} &= \frac{a_{32}^{(1)}}{a_{22}^{(1)}} = 2, \\ a_{33}^{(2)} &= a_{33}^{(1)} - l_{32}a_{23}^{(1)} = -1. \end{aligned}$$

Тоді $\det(A) = 1 \cdot (-1) \cdot (-1) = 1$. ◀

Приклад 1.2. Методом виключення Гаусса з вибором головного елемента у стовпчику розв'яжіть СЛАР

$$\begin{cases} x_1 + 3x_2 + 2x_3 = -8, \\ 2x_1 + 2x_2 + 3x_3 = 1, \\ 2x_1 + x_2 + 2x_3 = 3. \end{cases}$$

▷ Оскільки головний елемент у першому стовпці матриці системи $a_{21} = 2$, то переставимо перше і друге рівняння. Отримаємо систему

$$\begin{cases} 2x_1 + 2x_2 + 3x_3 = 1, \\ x_1 + 3x_2 + 2x_3 = -8, \\ 2x_1 + x_2 + 2x_3 = 3. \end{cases}$$

Знайдемо

$$\begin{aligned} l_{21} &= \frac{a_{21}}{a_{11}} = \frac{1}{2}, & l_{31} &= \frac{a_{31}}{a_{11}} = 1, & a_{22}^{(1)} &= a_{22} - l_{21}a_{12} = 2, \\ a_{23}^{(1)} &= a_{22} - l_{21}a_{13} = \frac{1}{2}, & a_{32}^{(1)} &= a_{32} - l_{31}a_{12} = -1, \\ a_{33}^{(1)} &= a_{33} - l_{31}a_{13} = -1, & f_2^{(1)} &= f_2 - l_{21}f_1 = -\frac{17}{2}, \\ f_3^{(1)} &= f_3 - l_{31}f_1 = 2. \end{aligned}$$

і одержимо систему

$$\begin{cases} 2x_1 + 2x_2 + 3x_3 = 1, \\ 2x_2 + \frac{1}{2}x_3 = -\frac{17}{2}, \\ -x_2 - x_3 = 3. \end{cases}$$

На другому кроці головний елемент $a_{22}^{(1)} = 2$, а тому переставляти рядки не треба. Знайдемо

$$l_{32} = \frac{a_{32}^{(1)}}{a_{22}^{(1)}} = -\frac{1}{2}, \quad a_{33}^{(3)} = a_{33}^{(1)} - l_{32}a_{23}^{(1)} = -\frac{3}{4},$$

$$f_3^{(3)} = f_3^{(1)} - l_{32}f_2^{(1)} = -\frac{9}{4}.$$

Одержимо систему

$$\begin{cases} 2x_1 + 2x_2 + 3x_3 = 1, \\ 2x_2 + \frac{1}{2}x_3 = -\frac{17}{2}, \\ -\frac{3}{4}x_3 = -\frac{9}{4}. \end{cases}$$

Використовуючи обернений хід методу Гаусса, знайдемо

$$x_3 = 3, \quad x_2 = -5, \quad x_1 = 1. \quad \blacktriangleleft$$

Приклад 1.3. Методом LU — розкладу з вибором головного елемента розв'яжіть СЛАР

$$A\mathbf{x} = \mathbf{f},$$

де

$$A = \begin{pmatrix} 0 & 3 & 1 \\ 7 & -13 & -2 \\ 1 & 2 & 4 \end{pmatrix}, \quad \mathbf{f} = (1, 2, 4)^T, \quad \mathbf{x} = (x_1, x_2, x_3)^T.$$

▷ Оскільки $\max\{|a_{11}|, |a_{21}|, |a_{31}|\} = |a_{21}|$, то головний елемент у першому стовпці $a_{21} = 7$, а тому переставимо перший і другий рядки матриці A . Тоді одержимо

$$P_1 A = \begin{pmatrix} 7 & -13 & -2 \\ 0 & 3 & 1 \\ 1 & 2 & 4 \end{pmatrix},$$

де P_1 — матриця переставлень отримана з одиничної матриці I переставлянням першого та другого рядків, тобто

$$P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Обчислимо

$$l_{21} = \frac{a_{21}}{a_{11}} = 0, \quad l_{31} = \frac{a_{31}}{a_{11}} = \frac{1}{7}, \quad a_{22}^{(1)} = a_{22} - l_{21}a_{12} = 3,$$

$$a_{23}^{(1)} = a_{23} - l_{21}a_{13} = 1, \quad a_{32}^{(1)} = a_{32} - l_{31}a_{12} = \frac{27}{7},$$

$$a_{33}^{(1)} = a_{33} - l_{31}a_{13} = \frac{30}{7},$$

тоді

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -l_{21} & 1 & 0 \\ -l_{31} & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{1}{7} & 0 & 1 \end{pmatrix},$$

$$L_1 P_1 A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{1}{7} & 0 & 1 \end{pmatrix} \begin{pmatrix} 7 & -13 & -2 \\ 0 & 3 & 1 \\ 1 & 2 & 4 \end{pmatrix} = \begin{pmatrix} 7 & -13 & -2 \\ 0 & 3 & 1 \\ 0 & \frac{27}{7} & \frac{30}{7} \end{pmatrix}.$$

На другому кроці головний елемент $a_{32}^{(1)} = 27/7$, а тому переставимо другий і третій рядки матриці $L_1 P_1 A$. Тоді одержимо

$$P_2 L_1 P_1 A = \begin{pmatrix} 7 & -13 & -2 \\ 0 & \frac{27}{7} & \frac{30}{7} \\ 0 & 3 & 1 \end{pmatrix},$$

де

$$P_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Обчислимо

$$l_{32} = \frac{a_{32}^{(1)}}{a_{22}^{(1)}} = \frac{7}{9}, \quad a_{33}^{(3)} = a_{33}^{(1)} - l_{32} a_{23}^{(1)} = -\frac{7}{3},$$

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -l_{32} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{7}{9} & 1 \end{pmatrix}.$$

Тоді

$$U = L_2 P_2 L_1 P_1 A = \begin{pmatrix} 7 & -13 & -2 \\ 0 & \frac{27}{7} & \frac{30}{7} \\ 0 & 0 & -\frac{7}{3} \end{pmatrix}, \quad P = P_2 P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$

$$\begin{aligned}
L &= P_2 L_1^{-1} P_2^{-1} L_2^{-1} = \\
&= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{7} & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{7}{9} & 1 \end{pmatrix} = \\
&= \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{7} & 1 & 0 \\ 0 & \frac{7}{9} & 1 \end{pmatrix}.
\end{aligned}$$

Оскільки $A = P^{-1}LU$, то вихідна система еквівалентна системі

$$LU\mathbf{x} = P\mathbf{f}.$$

Розв'яжемо спочатку СЛАР з нижньою трикутною матрицею L

$$L\mathbf{y} = P\mathbf{f}, \quad \mathbf{y} = (y_1, y_2, y_3)^T$$

або

$$\begin{cases} y_1 &= 2, \\ \frac{1}{7}y_1 + y_2 &= 4, \\ \frac{7}{9}y_2 + y_3 &= 1. \end{cases}$$

Звідси $y_1 = 2, y_2 = 26/7, y_3 = -17/9$. Оскільки $\mathbf{y} = U\mathbf{x}$, то знайдемо розв'язок системи з правою трикутною матрицею

$$\begin{cases} 7x_1 - 13x_2 - 2x_3 = 2, \\ \frac{27}{7}x_2 + \frac{30}{7}x_3 = \frac{26}{7}, \\ -\frac{7}{3}x_3 = -\frac{17}{9}. \end{cases}$$

Звідси

$$x_3 = 17/21, \quad x_2 = 4/63, \quad x_1 = 40/63. \quad \blacktriangleleft$$

1.1.2. Метод прогонки

Найвідоміший різновид методу Гаусса — *метод прогонки*, який застосовують до систем з тридіагональними матрицями. Такі системи часто зустрічаються при розв'язанні крайових задач для звичайних диференціальних рівнянь другого порядку і рівнянь із частинними похідними методом сіток. Вони можуть бути записані у вигляді:

$$\begin{aligned}
-c_0x_0 + b_0x_1 &= -f_0, \\
a_ix_{i-1} - c_ix_i + b_ix_{i+1} &= -f_i, \quad i = \overline{1, N-1}, \\
a_Nx_{N-1} - c_Nx_N &= -f_N,
\end{aligned} \tag{1.10}$$

або

$$A\mathbf{x} = \mathbf{f},$$

де $\mathbf{x} = (x_0, x_1, \dots, x_N)^T$ — вектор невідомих, $\mathbf{f} = (f_0, f_1, \dots, f_N)^T$ — вектор правих частин, а A — тридіагональна матриця розміру $(N+1) \times (N+1)$:

$$A = \begin{pmatrix} c_0 & -b_0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ -a_1 & c_1 & -b_1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -a_i & c_i & -b_i & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & -a_{N-1} & c_{N-1} & -b_{N-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & -a_N & c_N \end{pmatrix}.$$

Унаслідок прямого ходу в методі Гаусса одержимо систему з дво-діагональною верхньою трикутною матрицею. Тому формули зворотнього ходу мають вигляд:

$$x_i = \alpha_{i+1}x_{i+1} + \beta_{i+1}, \quad i = N-1, N-2, \dots, 0, \quad (1.11)$$

де $\alpha_{i+1}, \beta_{i+1}$ — неозначені коефіцієнти. Для визначення $\alpha_{i+1}, \beta_{i+1}$ підставимо $x_{i-1} = \alpha_i x_i + \beta_i$ у i -те рівняння системи (1.10):

$$(a_i \alpha_i - c_i)x_i + b_i x_{i+1} = -(f_i + a_i \beta_i), \quad i = \overline{1, N-1}. \quad (1.12)$$

Рівняння (1.12) розв'яжемо відносно x_i , тоді

$$x_i = \frac{b_i}{c_i - a_i \alpha_i} x_{i+1} + \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}. \quad (1.13)$$

Порівнюючи рівність (1.11) і (1.13), одержимо

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i \alpha_i}, \quad \beta_{i+1} = \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}, \quad i = \overline{1, N-1}. \quad (1.14)$$

З першого рівняння системи (1.10)

$$x_0 = \frac{b_0}{c_0} x_1 + \frac{f_0}{c_0}.$$

Отже,

$$\alpha_1 = \frac{b_0}{c_0}, \quad \beta_1 = \frac{f_0}{c_0}.$$

Визначивши α_1, β_1 , послідовно обчислимо $\alpha_{i+1}, \beta_{i+1}$, $i = \overline{1, N-1}$ за формулами (1.14).

Для обчислення x_i за формулою (1.11) необхідно задати x_N . Підставимо

$$x_{N-1} = \alpha_N x_N + \beta_N$$

в останнє рівняння системи (1.10), тоді

$$(a_N \alpha_N - c_N) x_N = -f_N - a_N \beta_N,$$

$$x_N = \frac{f_N + a_N \beta_N}{c_N - a_N \alpha_N} = \beta_{N+1}.$$

Отже, алгоритм методу прогонки має вигляд:

1. Обчислимо коефіцієнти

$$\begin{aligned} \alpha_1 &= \frac{b_0}{c_0}, \quad \alpha_{i+1} = \frac{b_i}{c_i - a_i \alpha_i}, \quad i = \overline{1, N-1}, \\ \beta_1 &= \frac{f_0}{c_0}, \quad \beta_{i+1} = \frac{f_i + a_i \beta_i}{c_i - a_i \alpha_i}, \quad i = \overline{1, N}. \end{aligned} \quad (1.15)$$

2. Знаходимо розв'язок

$$x_N = \beta_{N+1}, \quad x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1}, \quad i = N-1, N-2, \dots, 0. \quad (1.16)$$

Для реалізації алгоритму методу прогонки необхідно $3N$ множень, $2N+1$ ділень і $3N$ додавань.

Метод прогонки (1.15), (1.16), за допомогою якого x_i обчислюють справа наліво, називають *правою прогонкою*.

Зауважимо, що коефіцієнти α_i не залежать від правої частини системи (1.10), а визначаються лише величинами a_i, b_i, c_i . Тому, якщо необхідно розв'язувати послідовність задач з різними правими частинами, але з однією і тією ж матрицею, то коефіцієнти α_i обчислюються лише при розв'язуванні першої задачі з послідовності. Для кожної подальшої задачі визначаються тільки коефіцієнти β_i і розв'язок x_i .

Формули прогонки можна застосувати, якщо знаменники в (1.15) не перетворюються в нуль. Крім того, при обчисленні x_i за рекурентною формулою (1.16) похибки заокруглень можуть нагромаджуватися. Дійсно, нехай при обчисленні x_{i+1} допущена похибка Δx_{i+1} , тобто знайдено $\tilde{x}_{i+1} = x_{i+1} + \Delta x_{i+1}$. Тоді $\Delta x_i = \tilde{x}_i - x_i = \alpha_{i+1} \Delta x_{i+1}$. Звідси випливає, що за умови $|\alpha_i| > 1$ похибка може сильно зростати. Якщо виконується

умова $|\alpha_i| \leq 1$, то $|\Delta x_i| = |\alpha_{i+1}| \cdot |\Delta x_{i+1}| \leq |\Delta x_{i+1}|$ і кажуть, що метод прогонки *стійкий*.

Доведемо, що для виконання нерівностей $c_i - a_i \alpha_i \neq 0$, $|\alpha_i| \leq 1$, $i = \overline{1, N}$, достатньо, щоб коефіцієнти системи (1.10) задовольняли умови переваги діагональних елементів

$$|c_i| \geq |a_i| + |b_i|, \quad i = \overline{1, N-1}, \quad (1.17)$$

$$|c_0| \geq |b_0|, \quad |c_N| \geq |a_N|, \quad (1.18)$$

причому хоча б в одній з нерівностей (1.17) або (1.18) виконувалася строга нерівність. З умови (1.18) випливає, що $|\alpha_1| \leq 1$. Припустимо, що $|\alpha_i| \leq 1$, тоді на підставі (1.17)

$$|c_i - a_i \alpha_i| \geq |c_i| - |a_i| \cdot |\alpha_i| \geq |b_i| + |a_i| \cdot (1 - |\alpha_i|) \geq |b_i| > 0, \quad (1.19)$$

тобто $c_i - a_i \alpha_i \neq 0$ і

$$|\alpha_{i+1}| = \frac{|b_i|}{|c_i - a_i \alpha_i|} \leq \frac{|b_i|}{|b_i|} = 1, \quad i = \overline{1, N-1}.$$

Покажемо, що $c_N - a_N \alpha_N \neq 0$. Для цього використаємо припущення, що хоча б в одній з нерівностей (1.17) або (1.18) виконується строга нерівність. Якщо $|c_N| > |a_N|$, то $|\alpha_N| \leq 1$ і $|c_N - a_N \alpha_N| \geq |c_N| - |a_N| \cdot |\alpha_N| \geq |c_N| - |a_N| > 0$. Якщо строга нерівність досягається в (1.17) для деякого $i = i_0$, то з (1.19) одержимо, що $|c_{i_0} - a_{i_0} \alpha_{i_0}| > |b_{i_0}| > 0$ і $|\alpha_{i_0+1}| < 1$. За індукцією $|\alpha_i| < 1$ для $i \geq i_0 + 1$. Отже, в цьому випадку будемо мати $|\alpha_N| < 1$, і тому $c_N - a_N \alpha_N \neq 0$. Якщо $|c_0| > |b_0|$, то нерівність $|\alpha_i| < 1$ виконується, починаючи з $i = 1$. Тому знову дістанемо $|\alpha_N| < 1$ і $c_N - a_N \alpha_N \neq 0$.

Приклад 1.4. Розв'яжіть СЛАР

$$3x_0 + x_1 = 0,$$

$$x_{i-1} + 4x_i + x_{i+1} = 0, \quad i = 1, 2, 3,$$

$$x_3 + 3x_4 = -\frac{1}{3}.$$

▷ Якщо покласти $a_i = b_i = 1$, $c_i = 4$, $f_i = 0$, $i = 1, 2, 3$, $N = 4$, $c_0 = -3$, $b_0 = 1$, $f_0 = 0$, $a_4 = 1$, $c_4 = -3$, $f_4 = 1/3$, то систему можна записати у вигляді (1.10). Оскільки виконуються умови (1.17), (1.18), то для розв'язування цієї системи можна застосувати метод прогонки. Здійснимо обчислення за формулами (1.15), (1.16), заповнюючи таблицю 1.1. ◀

Табл. 1.1. Результати розв'язування прикладу 1.4

i	0	1	2	3	4	5
$c_i - a_i \alpha_i$	–	$-11/3$	$-41/11$	$-153/41$	$-418/153$	–
α_i	–	$-1/3$	$-3/11$	$-11/41$	$-41/153$	–
β_i	–	0	0	0	0	$-51/418$
x_i	$-1/1254$	$1/418$	$-11/1254$	$41/1254$	$-51/418$	–

1.2. Норми та обумовленість матриць систем лінійних алгебраїчних рівнянь

Коефіцієнти матриці і правої частини СЛАР не завжди відомі точно. Навіть, якщо ці величини відомі точно, виникають похибки заокруглень. Можна показати, що похибки заокруглень у методі Гаусса мають такий самий вплив на результат, що і похибки у вихідних коефіцієнтах. Тому фактично маємо розв'язок $\tilde{\mathbf{x}}$ деякої іншої системи

$$\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{f}}. \quad (1.20)$$

На практиці важливо знати відносну похибку $\delta\mathbf{x} = \|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$ для будь-якої векторної норми. Найчастіше використовують норми вектора

$$\|\mathbf{x}\|_p = \left(\sum_{j=1}^n |x_j|^p \right)^{1/p}, \quad 1 \leq p < \infty, \quad \|\mathbf{x}\|_\infty = \max_{j=1,n} |x_j|$$

і узгоджені з ними норми матриць (тобто для яких $\|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|$)

$$\|A\|_p = \sup_{\|\mathbf{x}\|_p \neq 0} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \sup_{\|\mathbf{x}\|_p = 1} \|A\mathbf{x}\|_p,$$

$$\|A\|_1 = \max_{k=1,n} \sum_{j=1}^n |a_{jk}|, \quad \|A\|_2 = \sqrt{\lambda_{\max}(AA^T)},$$

$$\|A\|_\infty = \max_{j=1,n} \sum_{k=1}^n |a_{jk}|,$$

де $\lambda_{\max}(AA^T)$ — максимальне власне значення матриці AA^T , а $\sqrt{\lambda_{\max}(AA^T)}$ називається *максимальним сингулярним числом* матриці A . Якщо замість (1.20), брати модель обчислень

$$A\tilde{\mathbf{x}} = \tilde{\mathbf{f}}, \quad (1.21)$$

тобто вважати, що матриця A задана в комп'ютері точно, то з ланцюжка співвідношень

$$\tilde{\mathbf{x}} - \mathbf{x} = A^{-1}(\tilde{\mathbf{f}} - \mathbf{f}), \quad \|\tilde{\mathbf{x}} - \mathbf{x}\| \leq \|A^{-1}\| \cdot \|\tilde{\mathbf{f}} - \mathbf{f}\|,$$

$$\mathbf{f} = A\mathbf{x}, \quad \|\mathbf{f}\| \leq \|A\| \|\mathbf{x}\|$$

випливає оцінка

$$\delta\mathbf{x} = \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\|A^{-1}\| \|\mathbf{f} - \tilde{\mathbf{f}}\| \|\mathbf{f}\|}{\|\mathbf{x}\| \|\mathbf{f}\|} \leq \|A^{-1}\| \|A\| \delta\mathbf{f}, \quad (1.22)$$

де $\text{cond}(A) = \|A^{-1}\| \|A\|$ (відповідно $\text{cond}_p(A) = \|A^{-1}\|_p \|A\|_p$) називається *числом обумовленості* матриці A і, як показує оцінка (1.22), це число є мірою невизначеності розв'язку системи (1.20) при неточних вхідних даних.

Якщо брати модель обчислень

$$\tilde{A}\tilde{\mathbf{x}} = \mathbf{f},$$

в якій збурені елементи лише матриці A , а праві частини задані точно, то використовуючи співвідношення $C^{-1} - B^{-1} = B^{-1}(B - C)C^{-1}$, дістанемо

$$\begin{aligned} \tilde{\mathbf{x}} - \mathbf{x} &= (\tilde{A}^{-1} - A^{-1}) \mathbf{f} = -A^{-1} (\tilde{A} - A) \tilde{A}^{-1} \mathbf{f} \\ &= -A^{-1} (\tilde{A} - A) \tilde{\mathbf{x}}, \end{aligned}$$

$$\|\tilde{\mathbf{x}} - \mathbf{x}\| \leq \|A^{-1}\| \|\tilde{A} - A\| \|\tilde{\mathbf{x}}\|,$$

$$\delta\mathbf{x} = \frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\tilde{\mathbf{x}}\|} \leq \text{cond}(A) \frac{\|\tilde{A} - A\|}{\|A\|}, \quad (1.23)$$

тобто і в цьому випадку число $\text{cond}(A)$ є мірою невизначеності розв'язку при неточних вхідних даних і інтервал цієї невизначеності тим ширший, чим більша величина $\text{cond}(A)$. Можна довести, що таку саму роль відіграє число обумовленості і у випадку моделі обчислень (1.20).

Кажуть, що СЛАР *погано обумовлена*, якщо малі зміни елементів матриці або правих частин викликають великі зміни у розв'язку. У цьому випадку ніякий чисельний метод не дає точного розв'язку, а у багатьох випадках навіть не варто шукати розв'язок.

Матриці погано обумовлених СЛАР мають відносно велике число обумовленості. Такі матриці називають погано обумовленими. Це означення залежить від норми і від комп'ютера, на яких здійснюється обчислення: одна і та ж сама система на різних комп'ютерах може бути добре чи погано обумовленою. Якщо число обумовленості велике, то (див. (1.22), (1.23)) малі зміни даних можуть привести до великих змін розв'язку.

Приклад 1.5. Обчисліть числа обумовленості $\text{cond}_1(A)$, $\text{cond}_\infty(A)$ матриці

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

▷ Обчислимо матрицю

$$A^{-1} = \begin{pmatrix} 1 & 0,8 & -0,6 \\ 0 & 0,2 & 0,1 \\ 0 & -0,4 & 0,3 \end{pmatrix}.$$

Тоді

$$\begin{aligned} \|A\|_1 &= \max \{1 + 0 + 0; 0 + 3 + 4; 2 + 1 + 2\} = 7; \\ \|A^{-1}\|_1 &= \max \{1; 1,4; 1\} = 1,4; \quad \text{cond}_1(A) = \|A\|_1 \|A^{-1}\|_1 = 9,8, \end{aligned}$$

а

$$\begin{aligned} \|A\|_\infty &= \max \{1 + 0 + 2; 0 + 3 + 1; 0 + 4 + 2\} = 6; \\ \|A^{-1}\|_\infty &= \max \{2,4; 0,3; 0,7\} = 2,4; \quad \text{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = 14,4. \end{aligned}$$



1.3. Ітераційні методи

1.3.1. Приклади та канонічний вигляд ітераційних методів

Розглянемо СЛАР:

$$A\mathbf{x} = \mathbf{f}, \tag{1.24}$$

де $A = \{a_{ij}\}_{i,j=1}^N$ — матриця розміру N , $\mathbf{x} = (x_1, x_2, \dots, x_N)^T$ — невідомий вектор, $\mathbf{f} = (f_1, f_2, \dots, f_N)^T$ — заданий вектор.

Для побудови ітераційного методу розв'язування системи (1.24) попередньо перетворимо i -те рівняння системи до вигляду

$$x_i = x_i - \sum_{j=1}^N a_{ij}x_j + f_i, \quad i = \overline{1, N}.$$

Цю систему будемо розв'язувати методом *послідовних наближень*

$$x_i^{n+1} = x_i^n - \sum_{j=1}^N a_{ij}x_j^n + f_i, \quad i = \overline{1, N}, \quad n = \overline{0, n_0}, \quad (1.25)$$

де x_i^n — n -та ітерація i -ї компоненти вектора \mathbf{x} . Початкові наближення x_i^0 , $i = \overline{1, N}$ задаються довільно. Ввівши позначення

$$\mathbf{x}_n = (x_1^n, x_2^n, \dots, x_N^n)^T,$$

ітераційний метод (1.25) запишемо у вигляді

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mathbf{A}\mathbf{x}_n + \mathbf{f}. \quad (1.26)$$

Припустимо, що всі a_{ii} відмінні від нуля. Систему (1.24) перетворимо до вигляду

$$x_i = - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j - \sum_{j=i+1}^N \frac{a_{ij}}{a_{ii}} x_j + \frac{f_i}{a_{ii}}, \quad i = \overline{1, N}. \quad (1.27)$$

Будемо вважати, що значення суми дорівнює нулю, якщо верхня границя сумування менша за нижню. Отже, запишемо рівняння (1.27) при $i = 1$

$$x_1 = - \sum_{j=2}^N \frac{a_{1j}}{a_{11}} x_j + \frac{f_1}{a_{11}}.$$

В *методі Якобі* виходять з запису системи у вигляді (1.27), причому ітерації визначають так:

$$x_i^{n+1} = - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^n - \sum_{j=i+1}^N \frac{a_{ij}}{a_{ii}} x_j^n + \frac{f_i}{a_{ii}}, \quad i = \overline{1, N}, \quad n = \overline{0, n_0}. \quad (1.28)$$

Ітераційний *метод Зейделя* має вигляд

$$x_i^{n+1} = - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{n+1} - \sum_{j=i+1}^N \frac{a_{ij}}{a_{ii}} x_j^n + \frac{f_i}{a_{ii}}, \quad i = \overline{1, N}, \quad n = \overline{0, n_0}. \quad (1.29)$$

Щоб зрозуміти, як знаходять звідси значення x_i^{n+1} , $i = \overline{1, N}$, запишемо детальніше перші два рівняння системи (1.29):

$$x_1^{n+1} = - \sum_{j=2}^N \frac{a_{1j}}{a_{11}} x_j^n + \frac{f_1}{a_{11}}, \quad (1.30)$$

$$x_2^{n+1} = - \frac{a_{21}}{a_{22}} x_1^{n+1} - \sum_{j=3}^N \frac{a_{2j}}{a_{22}} x_j^n + \frac{f_2}{a_{22}}. \quad (1.31)$$

Перша компонента x_1^{n+1} вектора \mathbf{x}_{n+1} знаходиться з рівняння (1.30) явно, для її обчислення потрібно знати вектор \mathbf{x}_n і значення f_1 . При знаходженні x_2^{n+1} з рівняння (1.31) використовуються тільки що знайдені значення x_1^{n+1} і відомі значення x_j^n , $j = 3, 4, \dots, N$, з попередньої ітерації. Таким чином, компоненти вектора \mathbf{x}_{n+1} знаходяться з рівняння (1.29) послідовно, починаючи з $i = 1$.

Щоб записати ітераційні методи Якобі та Зейделя в матричному вигляді, матрицю A системи (1.24) подамо у вигляді суми трьох матриць

$$A = A_1 + D + A_2, \quad (1.32)$$

де

$$D = \begin{pmatrix} a_{11} & & & 0 \\ & a_{22} & & \\ & & \dots & \\ 0 & & & a_{NN} \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & a_{12} & \dots & a_{1N} \\ & 0 & \dots & a_{2N} \\ & & \dots & \dots \\ & & 0 & a_{N-1,N} \\ & & & 0 \end{pmatrix},$$

$$A_1 = \begin{pmatrix} 0 & & & \\ a_{21} & 0 & & \\ \dots & \dots & \dots & \\ a_{N-1,1} & \dots & 0 & \\ a_{N1} & \dots & a_{N,N-1} & 0 \end{pmatrix}.$$

Тоді ітерації Якобі (1.28) можна записати у вигляді

$$\mathbf{x}_{n+1} = -D^{-1}(A_1\mathbf{x}_n + A_2\mathbf{x}_n - \mathbf{f})$$

або

$$D\mathbf{x}_{n+1} + (A_1 + A_2)\mathbf{x}_n = \mathbf{f}. \quad (1.33)$$

Метод Зейделя (1.29) записується у вигляді

$$\mathbf{x}_{n+1} = -D^{-1}(A_1\mathbf{x}_{n+1} + A_2\mathbf{x}_n - \mathbf{f})$$

або

$$(D + A_1)\mathbf{x}_{n+1} + A_2\mathbf{x}_n = \mathbf{f}. \quad (1.34)$$

Враховуючи (1.32), методи (1.33), (1.34) можна переписати відповідно у вигляді

$$D(\mathbf{x}_{n+1} - \mathbf{x}_n) + A\mathbf{x}_n = \mathbf{f}, \quad (1.35)$$

$$(D + A_1)(\mathbf{x}_{n+1} - \mathbf{x}_n) + A\mathbf{x}_n = \mathbf{f}. \quad (1.36)$$

Для прискорення збіжності в ітераційних методах вводять числові параметри, які, взагалі кажучи, залежать від номера ітерації. Наприклад, в методах (1.26), (1.35) можна ввести ітераційні параметри так:

$$\frac{\mathbf{x}_{n+1} - \mathbf{x}_n}{\tau_{n+1}} + A\mathbf{x}_n = \mathbf{f},$$

$$D \frac{\mathbf{x}_{n+1} - \mathbf{x}_n}{\tau_{n+1}} + A\mathbf{x}_n = \mathbf{f}.$$

Узагальненням методу Зейделя (1.36) є метод *верхньої релаксації*

$$(D + \omega A_1) \frac{\mathbf{x}_{n+1} - \mathbf{x}_n}{\omega} + A\mathbf{x}_n = \mathbf{f}, \quad (1.37)$$

де $\omega > 0$ — заданий числовий параметр. Для отримання розрахункових формул перепишемо (1.37) у вигляді

$$(I + \omega D^{-1}A_1)\mathbf{x}_{n+1} = ((1 - \omega)I - \omega D^{-1}A_2)\mathbf{x}_n + \omega D^{-1}\mathbf{f}$$

або у покомпонентному записі

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

В теорії ітераційних методів існує дві основні проблеми:

а) при яких значеннях параметрів метод збіжний;

б) при яких значеннях параметрів збіжність буде найбільш швидкою (відповідні параметри називаються оптимальними).

Надалі ми детальніше зупинимось на цих питаннях.

Наведені вище методи послідовних наближень, Якобі та Зейделя відносяться до *однокрокових* або *двоярусних* ітераційних методів, оскільки для знаходження \mathbf{x}_{n+1} використовують лише одну попередню ітерацію \mathbf{x}_n . Якщо \mathbf{x}_{n+1} виражається через \mathbf{x}_n і \mathbf{x}_{n-1} , то метод називають *двокроковим* або *триярусним*. Надалі ми будемо розглядати лише двоярусні ітераційні методи.

Один і той самий ітераційний метод можна записати багатьма різними способами. Тому доцільно ввести деякий стандартний вигляд запису ітераційних методів.

Канонічним виглядом двоярусного ітераційного методу називається ітераційна схема

$$B_n \frac{\mathbf{x}_{n+1} - \mathbf{x}_n}{\tau_{n+1}} + A\mathbf{x}_n = \mathbf{f}, \quad n = 0, 1, \dots \quad (1.38)$$

де B_n — матриця, для якої існує обернена B_n^{-1} . Числа τ_n називають *ітераційними параметрами*. Якщо $B_n \equiv B$, $\tau_n \equiv \tau$, тобто не залежать від n , то метод (1.38) називають *стаціонарним*, у протилежному випадку — *нестационарним*. При $B_n \equiv I$ ітераційну схему називають *явною*, тому що тоді \mathbf{x}_{n+1} знаходять за явною формулою

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \tau_{n+1} (A\mathbf{x}_n - \mathbf{f}), \quad n = 0, 1, \dots$$

При $B_n \neq I$ метод (1.38) називають *неявним*, бо для знаходження \mathbf{x}_{n+1} необхідно розв'язати систему рівнянь

$$B_n \mathbf{x}_{n+1} = B_n \mathbf{x}_n - \tau_{n+1} (A\mathbf{x}_n - \mathbf{f}), \quad n = 0, 1, \dots \quad (1.39)$$

У неявних методах матрицю B_n вибирають так, щоб рівняння (1.39) можна було розв'язати простіше, ніж (1.24).

1.3.2. Матричні нерівності та дії з ними

При формулюванні умов збіжності ітераційних методів будуть використовуватися матричні нерівності. Нехай H — N -вимірний евклідовий простір зі скалярним добутком

$$(\mathbf{u}, \mathbf{v}) = \sum_{j=1}^N u_j v_j \quad \forall \mathbf{u}, \mathbf{v} \in H.$$

Для дійсної матриці A нерівність $A > 0$ означає, що $(A\mathbf{x}, \mathbf{x}) > 0 \forall \mathbf{x} \in H, \mathbf{x} \neq \mathbf{0}$, а $A \geq 0 \iff (A\mathbf{x}, \mathbf{x}) \geq 0 \forall \mathbf{x} \in H$. З нерівності $A > 0$ випливає, що $(A\mathbf{x}, \mathbf{x}) \geq \delta \|\mathbf{x}\|^2$. Дійсно, якщо $A > 0$ — несиметрична матриця, то для $\forall \mathbf{x} \in H, \mathbf{x} \neq \mathbf{0}$ маємо

$$(A\mathbf{x}, \mathbf{x}) = \frac{1}{2} [(A\mathbf{x}, \mathbf{x}) + (\mathbf{x}, A^T \mathbf{x})] > 0,$$

де A^T — матриця, транспонована до A . Тому за δ можна взяти мінімальне власне значення матриці $A_0 = 0,5(A + A^T)$.

Наведемо потрібні надалі відомості з лінійної алгебри.

1) Якщо A — дійсна симетрична матриця, то існує *ортогональна* матриця Q (тобто $Q^T = Q^{-1}$) така, що $A = Q^T \Lambda Q$, де Λ — *діагональна* матриця. На головній діагоналі матриці Λ знаходяться власні значення матриці A .

2) Для симетричної матриці A нерівність $A \geq 0$ ($A > 0$) еквівалентна невід'ємності (додатності) всіх її власних значень.

Доведення. Використовуючи властивість 1, одержимо $\forall \mathbf{x} \in H$

$$(A\mathbf{x}, \mathbf{x}) = (Q^T \Lambda Q \mathbf{x}, \mathbf{x}) = (\Lambda Q \mathbf{x}, Q \mathbf{x}) = \sum_{i=1}^N \lambda_i y_i^2,$$

де λ_i — власні числа матриці A , y_i — i -та компонента вектора $\mathbf{y} = Q\mathbf{x}$. Звідси випливає, що якщо $\lambda_i \geq 0, i = \overline{1, N}$, ($\lambda_i > 0, i = \overline{1, N}$), то $(A\mathbf{x}, \mathbf{x}) \geq 0 \forall \mathbf{x} \in H$ ($(A\mathbf{x}, \mathbf{x}) > 0 \forall \mathbf{x} \neq \mathbf{0}$). Навпаки, нехай λ_j — власне значення матриці A . Задамо вектор \mathbf{y} , у якого всі компоненти, крім j -ої дорівнюють нулю, а $y_j = 1$. Оскільки матриця $Q^{-1} = Q^T$ існує, для заданого вектора \mathbf{y} знайдеться вектор $\mathbf{x} \in H$ такий, що $Q\mathbf{x} = \mathbf{y}$. Але тоді

$$0 \leq (A\mathbf{x}, \mathbf{x}) = (\Lambda \mathbf{y}, \mathbf{y}) = \lambda_j,$$

тобто $\lambda_j \geq 0$. ■

3) Якщо $A^T = A > 0$, то $\exists A^{-1}$.

Доведення. Згідно з властивістю 2 всі власні числа матриці A додатні, отже, $\det A \neq 0$ і $\exists A^{-1}$. ■

- 4) Для симетричної матриці S і для $\forall \rho > 0$ еквівалентні такі нерівності

$$-\rho I \leq S \leq \rho I, \quad (1.40)$$

$$S^2 \leq \rho^2 I. \quad (1.41)$$

Доведення. Згідно з властивістю 2 умова (1.40) еквівалентна нерівностям

$$|s_k| \leq \rho, \quad k = \overline{1, N},$$

де s_k — власні числа матриці S . Звідси $s_k^2 \leq \rho^2$, $k = \overline{1, N}$, що в свою чергу еквівалентне (1.41). ■

- 5) Якщо $A^T = A$ і $A \geq 0$ ($A > 0$), то існує матриця B , яка володіє такою властивістю

$$B^2 = A, \quad B^T = B, \quad B \geq 0 \quad (B > 0). \quad (1.42)$$

Ця матриця називається *квадратним коренем* з матриці A і позначається $A^{1/2}$.

Доведення. Нехай λ_i , $i = \overline{1, N}$ — власні числа матриці A . Згідно з властивістю 1 існує ортогональна матриця Q така, що

$$Q A Q^T = \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_N).$$

Оскільки всі λ_i невід'ємні, можна визначити матрицю $\Lambda^{1/2}$ як

$$\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_N}).$$

Тоді матриця $B = Q^T \Lambda^{1/2} Q$ володіє властивістю (1.42). ■

- 6) Нехай $A^T = A$ і L — невироджена матриця. Тоді еквівалентні нерівності

$$A \geq 0, \quad L^T A L \geq 0.$$

Аналогічно, еквівалентні строгі нерівності

$$A > 0, \quad L^T A L > 0.$$

Доведення. Для $\forall \mathbf{x} \in H$ маємо $(L^T A L \mathbf{x}, \mathbf{x}) = (A L \mathbf{x}, L \mathbf{x})$. Отже, $L^T A L \geq 0$, якщо $A \geq 0$. Доведемо обернене. Оскільки L^{-1} існує, то довільний вектор $\mathbf{x} \in H$ можна подати у вигляді $\mathbf{x} = L \mathbf{y}$, де $\mathbf{y} = L^{-1} \mathbf{x}$. Тоді

$$(A \mathbf{x}, \mathbf{x}) = (L^T A L \mathbf{y}, \mathbf{y}) \geq 0,$$

тобто $A \geq 0$. ■

- 7) Якщо A і B — симетричні і L — невироджена матриця, то еквівалентні нерівності

$$A \geq B, \quad L^T A L \geq L^T B L.$$

Доведення. Твердження випливає з попередньої властивості. ■

- 8) Нехай $C^T = C > 0$ і α, β — довільні дійсні числа. Тоді еквівалентні нерівності

$$\alpha C \geq \beta I, \quad \alpha I \geq \beta C^{-1}.$$

Доведення. Згідно з властивістю 5 існує матриця $C^{1/2} = (C^{1/2})^T > 0$. Використовуючи властивість 7, перейдемо від першої нерівності до другої за допомогою таких нерівностей:

$$\alpha(C^{-1/2})C(C^{-1/2}) \geq \beta(C^{-1/2})(C^{-1/2}),$$

$$\alpha(C^{-1/2}C^{1/2})(C^{1/2}C^{-1/2}) \geq \beta C^{-1}, \quad \alpha I \geq \beta C^{-1}.$$

■

- 9) Нехай $A^T = A > 0$, $B^T = B > 0$, α і β — довільні дійсні числа. Тоді еквівалентні нерівності

$$\alpha A \geq \beta B, \quad \alpha B^{-1} \geq \beta A^{-1}.$$

Доведення. Помножимо першу нерівність зліва і справа на $B^{-1/2}$, тоді одержимо

$$\alpha C \geq \beta I, \quad C = B^{-1/2} A B^{-1/2}.$$

Згідно з властивістю 8 остання нерівність еквівалентна нерівності $\alpha I \geq \beta C^{-1}$, тобто

$$\alpha I \geq \beta B^{1/2} A^{-1} B^{1/2},$$

помноживши котру зліва і справа на $B^{-1/2}$, одержимо $\alpha B^{-1} \geq \beta A^{-1}$. ■

1.3.3. Дослідження збіжності ітераційних методів

Розглянемо стаціонарний ітераційний метод

$$B \frac{\mathbf{x}_{n+1} - \mathbf{x}_n}{\tau} + A\mathbf{x}_n = \mathbf{f}, \quad n = 0, 1, \dots \quad (1.43)$$

Похибка ітераційного методу характеризується вектором $\mathbf{z}_n = \mathbf{x}_n - \mathbf{x}$, який згідно з (1.24), (1.43) задовольняє однорідне рівняння

$$B \frac{\mathbf{z}_{n+1} - \mathbf{z}_n}{\tau} + A\mathbf{z}_n = 0, \quad n = 0, 1, \dots, \quad \mathbf{z}_0 = \mathbf{x}_0 - \mathbf{x}. \quad (1.44)$$

Кажуть, що ітераційний метод (1.43) *збігається*, якщо $\|\mathbf{z}_n\| \rightarrow 0$ при $n \rightarrow \infty$.

Якщо існує B^{-1} , то з рівняння (1.44) знайдемо

$$\mathbf{z}_{n+1} = S\mathbf{z}_n, \quad (1.45)$$

де $S = I - \tau B^{-1}A$. Тоді

$$\mathbf{z}_{n+1} = S^{n+1}\mathbf{z}_0. \quad (1.46)$$

Отже, за умови що $\|S\| \leq q < 1$ маємо $\|\mathbf{z}_{n+1}\| \leq q^{n+1} \|\mathbf{z}_0\| \rightarrow 0$ при $n \rightarrow \infty$, тобто ітераційний метод збігається зі *швидкістю геометричної прогресії* зі знаменником $q \in (0, 1)$.

Зокрема, ітераційний метод простої ітерації (1.26) збіжний за умови, що $\|I - A\| < 1$.

Розв'язок системи (1.24) будемо надалі розглядати як елемент N — вимірного простору H зі скалярним добутком $(\mathbf{u}, \mathbf{v}) = \sum_{j=1}^N u_j v_j$ та нормою $\|\mathbf{u}\| = (\mathbf{u}, \mathbf{u})^{1/2}$.

► **ТЕОРЕМА 1.1.** Нехай A — симетрична додатно визначена матриця, $\tau > 0$ і виконується умова

$$B > \frac{\tau}{2}A. \quad (1.47)$$

Тоді ітераційний метод (1.43) збігається.

Доведення. Покажемо спочатку, що за умови (1.47) числова послідовність $J_n = (A\mathbf{z}_n, \mathbf{z}_n)$ незростаюча. З рівняння (1.45) знаходимо

$$\mathbf{z}_{n+1} = (I - \tau B^{-1}A)\mathbf{z}_n, \quad A\mathbf{z}_{n+1} = (A - \tau AB^{-1}A)\mathbf{z}_n.$$

Тоді

$$(A\mathbf{z}_{n+1}, \mathbf{z}_{n+1}) = (A\mathbf{z}_n, \mathbf{z}_n) - \tau(AB^{-1}A\mathbf{z}_n, \mathbf{z}_n) - \\ - \tau(A\mathbf{z}_n, B^{-1}A\mathbf{z}_n) + \tau^2(AB^{-1}A\mathbf{z}_n, B^{-1}A\mathbf{z}_n).$$

Оскільки матриця A симетрична, то

$$(AB^{-1}A\mathbf{z}_n, \mathbf{z}_n) = (A\mathbf{z}_n, B^{-1}A\mathbf{z}_n),$$

а тому

$$(A\mathbf{z}_{n+1}, \mathbf{z}_{n+1}) = (A\mathbf{z}_n, \mathbf{z}_n) - \\ - 2\tau((B - 0,5\tau A)B^{-1}A\mathbf{z}_n, B^{-1}A\mathbf{z}_n). \quad (1.48)$$

Звідси, враховуючи умову (1.47), одержимо нерівність

$$(A\mathbf{z}_{n+1}, \mathbf{z}_{n+1}) \leq (A\mathbf{z}_n, \mathbf{z}_n).$$

Отже,

$$0 \leq J_{n+1} \leq J_n \leq \dots \leq J_0,$$

тобто послідовність $J_n = (A\mathbf{z}_n, \mathbf{z}_n)$ — незростаюча і обмежена знизу нулем. Тому за теоремою Вейєрштраса існує границя

$$\lim_{n \rightarrow \infty} J_n = J. \quad (1.49)$$

Доведемо, що $\lim_{n \rightarrow \infty} \|\mathbf{z}_n\| = 0$. З додатності матриці $B - 0,5\tau A$ випливає її додатна визначеність, тобто існує константа $\delta > 0$ така, що

$$((B - 0,5\tau A)B^{-1}A\mathbf{z}_n, B^{-1}A\mathbf{z}_n) \geq \delta \|B^{-1}A\mathbf{z}_n\|^2.$$

Звідси і з (1.48) випливає нерівність

$$J_{n+1} - J_n + 2\delta\tau \|B^{-1}A\mathbf{z}_n\|^2 \leq 0.$$

Враховуючи (1.49) перейдемо в цій нерівності до границі при $n \rightarrow \infty$, тоді одержимо

$$\lim_{n \rightarrow \infty} \|\mathbf{w}_n\| = 0,$$

де $\mathbf{w}_n = B^{-1}A\mathbf{z}_n$. На підставі того, що A — додатно визначена, а тому існує обернена A^{-1} , будемо мати

$$\mathbf{z}_n = A^{-1}B\mathbf{w}_n, \quad \|\mathbf{z}_n\| \leq \|A^{-1}B\| \cdot \|\mathbf{w}_n\|.$$

Звідси випливає, що

$$\lim_{n \rightarrow \infty} \|\mathbf{z}_n\| = 0.$$

Застосуємо цю теорему до конкретних ітераційних методів. ■

► **Наслідок 1.1.** Нехай A — симетрична додатно визначена матриця з діагональною перевагою, тобто

$$a_{ii} > \sum_{\substack{j=1 \\ j \neq i}}^N |a_{ij}|, \quad i = \overline{1, N}. \quad (1.50)$$

Тоді метод Якобі (1.35) збігається.

Доведення. Оскільки для методу Якобі $B = D$, а $\tau = 1$, то умова (1.47) має вигляд $A < 2D$. Покажемо, що ця матрична нерівність випливає з (1.50). Оцінимо квадратичну форму

$$\begin{aligned} (A\mathbf{x}, \mathbf{x}) &= \sum_{i=1}^N \sum_{j=1}^N a_{ij} x_i x_j \leq \\ &\leq \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N |a_{ij}| x_i^2 + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N |a_{ij}| x_j^2 = \\ &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N |a_{ij}| x_i^2 + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N |a_{ji}| x_i^2. \end{aligned}$$

З умов симетричності та додатної визначеності матриці A маємо $a_{ji} = a_{ij}$, $a_{ii} > 0$, $i, j = \overline{1, N}$, а тому з попередньої оцінки випливає нерівність

$$(A\mathbf{x}, \mathbf{x}) \leq \sum_{i=1}^N \sum_{j=1}^N |a_{ij}| x_i^2 = \sum_{i=1}^N \left(\sum_{\substack{j=1 \\ j \neq i}}^N |a_{ij}| + a_{ii} \right) x_i^2. \quad (1.51)$$

Перепишемо умову (1.50) у вигляді

$$a_{ii} + \sum_{\substack{j=1 \\ j \neq i}}^N |a_{ij}| < 2a_{ii}, \quad i = \overline{1, N}.$$

Тоді з нерівності (1.51) одержимо

$$(A\mathbf{x}, \mathbf{x}) < 2 \sum_{i=1}^N a_{ii} x_i^2 = 2(D\mathbf{x}, \mathbf{x}).$$

■

► **Наслідок 1.2.** Нехай A — симетрична додатно визначена матриця. Тоді метод верхньої релаксації (1.37) збігається за умови $0 < \omega < 2$. Зокрема, метод Зейделя ($\omega = 1$) збіжний.

Доведення. Метод верхньої релаксації (1.37) має канонічний вигляд (1.43) з $B = D + \omega A_1$, $\tau = \omega$. Оскільки матрицю A системи (1.24) можна записати у вигляді (1.32), то для симетричної матриці A матриця A_2 є транспонованою до A_1 , а тому

$$(A\mathbf{x}, \mathbf{x}) = (D\mathbf{x}, \mathbf{x}) + (A_1\mathbf{x}, \mathbf{x}) + (A_2\mathbf{x}, \mathbf{x}) = (D\mathbf{x}, \mathbf{x}) + 2(A_1\mathbf{x}, \mathbf{x}).$$

Умова збіжності (1.47) має вигляд

$$\begin{aligned} (B\mathbf{x}, \mathbf{x}) - 0,5\omega(A\mathbf{x}, \mathbf{x}) &= ((D + \omega A_1)\mathbf{x}, \mathbf{x}) - \\ &\quad - 0,5\omega((D\mathbf{x}, \mathbf{x}) + 2(A_1\mathbf{x}, \mathbf{x})) \\ &= (1 - 0,5\omega)(D\mathbf{x}, \mathbf{x}) > 0 \end{aligned}$$

і при $0 < \omega < 2$ виконується.

Розглянемо збіжність методу простої ітерації

$$\frac{\mathbf{x}_{n+1} - \mathbf{x}_n}{\tau} + A\mathbf{x}_n = \mathbf{f} \quad (1.52)$$

з симетричною додатно визначеною матрицею A . Згідно з (1.47) метод збігається за умови

$$I - 0,5\tau A > 0. \quad (1.53)$$

Нехай λ_i , $i = \overline{1, N}$ — власні значення матриці A , які розташовані в порядку зростання. Умова (1.53) еквівалентна тому, що всі власні значення матриці $I - 0,5\tau A$ додатні. Достатньо вимагати додатності мінімального власного числа цієї матриці, рівного $1 - 0,5\tau\lambda_N$. Таким чином, ітераційний метод (1.52) збігається, якщо

$$\tau < 2/\lambda_{\max}, \quad (1.54)$$

де λ_{\max} — максимальне власне число матриці A .

Умова (1.54) є крім того необхідною для збіжності (1.52), тобто, якщо (1.54) не виконується, то знайдеться таке початкове наближення \mathbf{x}_0 , при якому $\lim_{n \rightarrow \infty} \|\mathbf{x}_n - \mathbf{x}\| \neq 0$.

Доведемо останнє твердження. Візьмемо за початкове наближення вектор $\mathbf{x}_0 = \mathbf{x} + \mu$, де \mathbf{x} — точний розв'язок системи (1.24), а μ —

власний вектор матриці A , який відповідає власному числу $\lambda_{\max} = \lambda_N$, тобто $A\mu = \lambda_N\mu$. При такому виборі початкового наближення

$$\mathbf{z}_0 = \mathbf{x}_0 - \mathbf{x} = \mu.$$

З рівняння (1.46)

$$\mathbf{z}_n = (I - \tau A)^n \mu,$$

а тому $\mathbf{z}_n = (1 - \tau\lambda_N)^n \mu$, $\|\mathbf{z}_n\| = |1 - \tau\lambda_N|^n \|\mu\|$.

Якщо $\tau = 2\lambda_N^{-1}$, то $\lim_{n \rightarrow \infty} \|\mathbf{z}_n\| = \|\mu\| \neq 0$. Якщо ж $\tau > 2\lambda_N^{-1}$, то $|1 - \tau\lambda_N| > 1$ і $\lim_{n \rightarrow \infty} \|\mathbf{z}_n\| = \infty$. Отже, умова (1.54) необхідна і достатня для збіжності методу простої ітерації (1.52). ■

► **ТЕОРЕМА 1.2. (НЕОБХІДНА І ДОСТАТНЯ УМОВА ЗБІЖНОСТІ ДВОЯРУСНОГО ІТЕРАЦІЙНОГО МЕТОДУ)** Ітераційний метод (1.43) збігається до розв'язку системи (1.24) для будь-якого початкового наближення тоді і тільки тоді, коли всі власні значення матриці $S = I - \tau B^{-1}A$ за модулем менші від 1.

Доведення. Доведемо спочатку необхідність. Нехай λ_j — власні значення, такі що $|\lambda_j| \geq 1$ і μ_j — відповідний власний вектор матриці S . Тоді при початковому наближенні $\mathbf{x}_0 = \mathbf{x} + r\mu_j$, $r = \text{const} \neq 0$, маємо $S\mathbf{z}_0 = \lambda_j\mathbf{z}$; $\mathbf{z}_0 = r\mu_j$ і з (1.46) $\lim_{n \rightarrow \infty} \|\mathbf{z}_n\| = \lim_{n \rightarrow \infty} \|S^n \mathbf{z}_0\| = \lim_{n \rightarrow \infty} |\lambda_j|^n |r| \|\mu_j\| \neq 0$ при $n \rightarrow \infty$.

Доведення достатності наведемо тільки для випадку, коли матриця S має N лінійно незалежних власних векторів. Нехай λ_j , $j = \overline{1, N}$ — власні числа матриці S і μ_j , $j = \overline{1, N}$ — відповідні лінійно незалежні власні вектори. Розкладемо величину \mathbf{z}_0 за власними векторами μ_j :

$$\mathbf{z}_0 = \sum_{j=1}^N c_j \mu_j.$$

Тоді

$$\mathbf{z}_n = S^n \mathbf{z}_0 = \sum_{j=1}^N c_j \lambda_j^n \mu_j$$

і

$$\|\mathbf{z}_n\| \leq \rho^n \sum_{j=1}^N |c_j| \|\mu_j\|, \quad (1.55)$$

де $\rho = \max_{1 \leq j \leq N} |\lambda_j|$ — спектральний радіус матриці S . З оцінки (1.55) на підставі припущення теореми 1.2 $\rho < 1$, а тому метод збіжний. В загальному випадку, коли система власних векторів матриці S неповна, то доведення достатності умов теореми проводиться за допомогою зведення S до жорданової форми. ■

Приклад 1.6. Доведіть, що для СЛАР $A\mathbf{x} = \mathbf{f}$, де

$$A = \begin{pmatrix} 1,5 & 5 & 0 \\ 0 & 0,5 & 0 \\ 0 & -1 & 0,5 \end{pmatrix}$$

метод послідовних наближень $\mathbf{x}_{n+1} = \mathbf{x}_n - A\mathbf{x}_n + \mathbf{f}$ збіжний для $\forall \mathbf{x}_0$.

▷ Оскільки

$$\mathbf{x}_{n+1} = S\mathbf{x}_n + \mathbf{f},$$

де

$$S = I - A = \begin{pmatrix} -0,5 & -5 & 0 \\ 0 & 0,5 & 0 \\ 0 & 1 & 0,5 \end{pmatrix}.$$

Застосуємо достатню ознаку збіжності методу послідовних наближень: $\|S\|_1 = 6,5 > 1$, $\|S\|_\infty = 5,5 > 1$. Умова не виконується. Застосуємо необхідну і достатню ознаку. Знайдемо власні числа матриці S

$$\begin{aligned} \det(S - \lambda I) &= \begin{vmatrix} -0,5 - \lambda & -5 & 0 \\ 0 & 0,5 - \lambda & 0 \\ 0 & 1 & 0,5 - \lambda \end{vmatrix} \\ &= -(0,5 + \lambda)(0,5 - \lambda)^2 = 0, \end{aligned}$$

$$\lambda_1 = -0,5; \quad \lambda_2 = \lambda_3 = 0,5.$$

Отже, $|\lambda_{1,2,3}| < 1$ і метод послідовних наближень збіжний. ◀

1.3.4. Оцінка швидкості збіжності стаціонарних ітераційних методів

У розділі 1.3.3 доведено, що ітераційний метод (1.43) збігається зі швидкістю геометричної прогресії, тобто, що виконується оцінка

$$\|\mathbf{x}_n - \mathbf{x}\| \leq q^n \|\mathbf{x}_0 - \mathbf{x}\|, \quad n = 0, 1, \dots, q \in (0, 1). \quad (1.56)$$

Використовуючи цю оцінку, можна визначити кількість ітерацій, достатніх для того, щоб початкова похибка зменшилася в задану кількість раз. Дійсно, задамо $\forall \varepsilon > 0$ і будемо вимагати, щоб $q^n < \varepsilon$, тоді

$$n \geq n_0(\varepsilon) = \frac{\ln(1/\varepsilon)}{\ln(1/q)}.$$

З (1.56) одержимо, що

$$\|\mathbf{x}_n - \mathbf{x}\| \leq \varepsilon \|\mathbf{x}_0 - \mathbf{x}\|,$$

тобто після проведення $n_0(\varepsilon)$ ітерацій початкова похибка $\|\mathbf{x}_0 - \mathbf{x}\|$ зменшиться в ε^{-1} раз. Ціла частина числа $n_0(\varepsilon)$ називається *мінімальним числом* ітерацій необхідних для одержання заданої точності ε .

Вираз $\ln(1/q)$, який знаходиться в знаменнику $n_0(\varepsilon)$, називається *швидкістю збіжності* ітераційного методу. Швидкість збіжності визначається властивостями матриці переходу S і не залежить ні від номера ітерації n , ні від вибору початкового наближення \mathbf{x}_0 , ні від заданої точності ε . Чим вища швидкість збіжності, тим кращий метод.

При практичному використанні ітераційних методів велике значення має швидкість збіжності. Відповідь на це питання дає таке твердження.

► **ТЕОРЕМА 1.3.** Нехай A і B — симетричні додатно визначені матриці, для яких справджуються нерівності

$$\gamma_1 B \leq A \leq \gamma_2 B, \quad (1.57)$$

де γ_1, γ_2 — додатні сталі $\gamma_2 \geq \gamma_1 > 0$. При

$$\tau = \frac{2}{\gamma_1 + \gamma_2}$$

ітераційний метод (1.43) збігається і для похибки справджуються оцінки

$$\|\mathbf{x}_n - \mathbf{x}\|_A \leq \rho^n \|\mathbf{x}_0 - \mathbf{x}\|_A, \quad n = 0, 1, \dots, \quad (1.58)$$

$$\|\mathbf{x}_n - \mathbf{x}\|_B \leq \rho^n \|\mathbf{x}_0 - \mathbf{x}\|_B, \quad n = 0, 1, \dots, \quad (1.59)$$

де $\|\mathbf{v}\|_A = (A\mathbf{v}, \mathbf{v})^{1/2}$, $\|\mathbf{v}\|_B = (B\mathbf{v}, \mathbf{v})^{1/2}$ і

$$\rho = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

Доведення. Для дослідження збіжності ітераційної схеми (1.43) перейдемо до задачі для еквівалентної похибки $\mathbf{w}_n = A^{1/2}\mathbf{z}_n$. Тоді з (1.45) одержимо

$$\mathbf{w}_{n+1} = A^{1/2}S\mathbf{z}_n = A^{1/2}SA^{-1/2}\mathbf{w}_n = \tilde{S}\mathbf{w}_n, \quad (1.60)$$

де $\tilde{S} = I - \tau C$, $C = A^{1/2}B^{-1}A^{1/2}$. Оскільки матриця \tilde{S} симетрична, то

$$\|\mathbf{w}_{n+1}\|^2 = (\tilde{S}\mathbf{w}_n, \tilde{S}\mathbf{w}_n) = (\tilde{S}^2\mathbf{w}_n, \mathbf{w}_n). \quad (1.61)$$

Оскільки

$$\frac{1-\rho}{\tau} = \frac{2\xi}{(1+\xi)\tau} = \gamma_1, \quad \frac{1+\rho}{\tau} = \frac{2}{(1+\xi)\tau} = \gamma_2,$$

то матричні нерівності (1.57) можна записати у вигляді

$$\frac{1-\rho}{\tau}B \leq A \leq \frac{1+\rho}{\tau}B$$

або (згідно з властивістю 9)

$$\frac{1-\rho}{\tau}A^{-1} \leq B^{-1} \leq \frac{1+\rho}{\tau}A^{-1}.$$

Помножимо останні нерівності зліва і справа на $A^{1/2}$, тоді

$$\frac{1-\rho}{\tau}I \leq C \leq \frac{1+\rho}{\tau}I.$$

Звідси

$$-\rho I \leq \tilde{S} \leq \rho I. \quad (1.62)$$

На підставі властивості 4 нерівності (1.62) еквівалентні нерівності

$$\tilde{S}^2 \leq \rho^2 I.$$

Отже, з рівності (1.61) випливає оцінка

$$\|\mathbf{w}_{n+1}\| \leq \rho \|\mathbf{w}_n\|,$$

з якої випливає оцінка (1.58). Дійсно,

$$\|\mathbf{w}_n\| = \|A^{1/2}\mathbf{z}_n\| = \|\mathbf{z}_n\|_A \leq \rho^n \|\mathbf{z}_0\|_A.$$

Оцінка (1.59) доводиться аналогічно, якщо за \mathbf{w}_n взяти вектор $B^{1/2}\mathbf{z}_n$, а за C — матрицю $B^{-1/2}AB^{-1/2}$. ■

1.3.5. Многочлени Чебишева

Многочлени Чебишева $T_n(x)$ на відрізку $[-1, 1]$ визначаються співвідношенням

$$T_n(x) = \cos(n \arccos x), \quad n \geq 0 \quad \forall x \in [-1, 1].$$

Розглянемо деякі властивості цих многочленів:

1. Якщо здійснити перетворення при $\forall \theta$ маємо

$$\cos((n+1)\theta) = 2 \cos \theta \cos n\theta - \cos((n-1)\theta).$$

В цій рівності покладемо $\theta = \arccos x$, тоді одержимо

$$\cos((n+1) \arccos x) = 2x \cos(n \arccos x) - \cos((n-1) \arccos x).$$

Отже, справджується рекурентне співвідношення:

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

$$T_0(x) = \cos(0 \cdot \arccos(x)) = 1, \quad T_1(x) = \cos(1 \cdot \arccos(x)) = x.$$

Старший член многочлена $T_{n+1}(x)$ отримується із старшого члена $T_n(x)$ множенням на $2x$, а тому старший член $T_n(x)$ при $n > 0$ є $2^{n-1}x^n$.

2. При $|x| \leq 1$, $|T_n(x)| \leq 1$, отже точки екстремуму $T_n(x)$ на відрізку будуть точки, де $|T_n(x)| = 1$. Це точки, які визначаються формулою

$$n \arccos x = \pi m, \quad x_m = \cos \frac{\pi m}{n}, \quad m = \overline{0, n},$$

причому

$$T_n(x_m) = \cos m\pi = (-1)^m.$$

3. Введемо многочлени $T_n^*(x) = 2^{1-n}T_n(x)$, причому $T_n^*(x_m) = (-1)^m 2^{1-n}$, $m = \overline{0, n}$. Отже, $\max_{x \in [-1, 1]} |T_n^*(x)| = 2^{1-n}$.
4. З рівняння $T_n(x) = \cos(n \arccos x) = 0$ отримаємо, що $x_k = \cos \frac{\pi(2k+1)}{2n}$, $k = \overline{0, n-1}$ — корені многочлена Чебишева.

Зауваження 1.1. Якщо $|x| > 1$, то многочлен $T_n(x)$ визначається формулою

$$T_n(x) = \frac{(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n}{2}.$$

⇒ **Лема 1.1.** Нехай $P_n(x)$ — довільний многочлен степеня n зі старшим коефіцієнтом 1 на відрізку $[-1, 1]$. Серед всіх многочленів $P_n(x)$ многочленом з найменшою верхньою межею $|P_n(x)|$ (многочленом, який найменше відхиляється від нуля) є многочлен $T_n^*(x) = 2^{1-n}T_n(x)$, тобто

$$\max_{x \in [-1, 1]} |P_n(x)| \geq \max_{x \in [-1, 1]} |T_n^*(x)| = 2^{1-n}.$$

Доведення. Припустимо протилежне, тобто

$$\max_{x \in [-1, 1]} |P_n(x)| < \max_{x \in [-1, 1]} |T_n^*(x)| = 2^{1-n}$$

і розглянемо функцію $Q(x) = T_n^*(x) - P_n(x)$, яка є многочленом степеня $n - 1$ і відмінна тотожно від нуля. Розглянемо значення $Q(x_m) = (-1)^m 2^{1-n} - P_n(x_m)$. Згідно з припущенням $|P_n(x_m)| < 2^{1-n}$, а тому

$$\text{sign } Q(x_m) = (-1)^m, \quad m = \overline{0, n}.$$

Отже, многочлен $Q(x)$ на відрізку $[-1, 1]$ міняє знак n разів, тобто має n коренів. Але це неможливо, тому що $Q(x)$ — многочлен степеня $n - 1$, відмінний від тотожного нуля. А це протиріччя, яке доводить лему. Тому многочлени $T_n^*(x)$ називають многочленами, які *найменше відхиляються від нуля*.

Розглянемо тепер многочлени Чебишева на відрізку $[a, b]$. За допомогою заміни змінних $x = \frac{1}{2}[(b - a)t + (b + a)]$, яка відрізок $a \leq x \leq b$ переводить у відрізок $-1 \leq t \leq 1$, многочлен Чебишева $T_n^*(t)$ перетворюємо до вигляду

$$F_n(x) = 2^{1-n} \cos \left(n \arccos \frac{2x - (b + a)}{b - a} \right).$$

Старший коефіцієнт цього многочлена дорівнює $2^n (b - a)^{-n}$. Якщо корені многочлена обчислити за формулами

$$x_k = \frac{b - a}{2} \cos \frac{(2k + 1)\pi}{2n} + \frac{a + b}{2}, \quad k = \overline{0, n - 1},$$

то сам многочлен, який найменше відхиляється від нуля на відріжку $[a, b]$ серед всіх многочленів степеня n зі старшим коефіцієнтом 1, є многочлен:

$$T_n^*(x) = \frac{(b-a)^n}{2^{2n-1}} \cos \left(n \arccos \frac{2x - (b+a)}{b-a} \right).$$

■

1.3.6. Ітераційний метод з чебишевським набором параметрів

Розглянемо явну схему

$$\frac{\mathbf{x}_{n+1} - \mathbf{x}_n}{\tau_{n+1}} + A\mathbf{x}_n = \mathbf{f}, \quad n = 0, 1, \dots, \quad (1.63)$$

при $A = A^* > 0$, $\gamma_1 I \leq A \leq \gamma_2 I$, $\gamma_1 > 0$.

Похибка $\mathbf{z}_n = \mathbf{x}_n - \mathbf{x}$ задовольняє рівняння

$$\frac{\mathbf{z}_{n+1} - \mathbf{z}_n}{\tau_{n+1}} + A\mathbf{z}_n = \mathbf{0}, \quad n = 0, 1, \dots, \quad \mathbf{z}_0 = \mathbf{x}_0 - \mathbf{x}.$$

Звідси

$$\mathbf{z}_n = T_n \mathbf{z}_0, \quad (1.64)$$

а

$$T_n = (I - \tau_n A)(I - \tau_{n-1} A) \dots (I - \tau_1 A).$$

Отже, матриця T_n є поліномом $P_n(A)$ степеня n відносно A , коефіцієнти якого залежать лише від $\tau_1, \tau_2, \dots, \tau_n$. З (1.64) для \mathbf{z}_n одержуємо нерівність

$$\|\mathbf{z}_n\| \leq \|P_n(A)\| \cdot \|\mathbf{z}_0\|.$$

Параметри $\tau_1, \tau_2, \dots, \tau_n$ будемо вибирати так, щоб $\|P_n(A)\|$ була мінімальною. Матричний поліном

$$P_n(A) = \prod_{k=1}^n (I - \tau_k A) = \sum_{k=0}^n c_k A^k, \quad c_0 = 1, \quad P_n(0) = I$$

є самоспряженою матрицею, оскільки A^k — самоспряжена.

Нехай λ_l, ξ_l , $l = \overline{1, N}$ — відповідно власні значення і власні функції матриці A :

$$A\xi_l = \lambda_l \xi_l, \quad l = \overline{1, N}, \quad (\xi_l, \xi_m) = \delta_{l,m} = \begin{cases} 1, & l = m, \\ 0, & l \neq m. \end{cases}$$

Матриця A^k має ті самі власні функції і власні значення λ_l^k , тому

$$P_n(A) \xi_l = \sum_{k=0}^n c_k A^k \xi_l = \sum_{k=0}^n c_k \lambda_l^k \xi_l = P_n(\lambda_l) \xi_l,$$

тобто $\lambda_l(P_n(A)) = P_n(\lambda_l)$. Оскільки $P_n(A)$ — самоспряжена матриця, то

$$\|P_n(A)\| = \max_{1 \leq l \leq N} |P_n(\lambda_l)|.$$

Власні значення матриці A розміщені на відрізку $[\gamma_1, \gamma_2]$, а тому

$$\max_{1 \leq l \leq N} |P_n(\lambda_l)| \leq \max_{\gamma_1 \leq x \leq \gamma_2} |P_n(x)|.$$

Задача найкращого вибору параметрів $\tau_1, \tau_2, \dots, \tau_n$ звелася до задачі знаходження

$$\min_{\{\tau_k\}} \max_{\gamma_1 \leq x \leq \gamma_2} |P_n(x)|.$$

Відобразимо відрізок $x \in [\gamma_1, \gamma_2]$ на відрізок $t \in [-1, 1]$ за допомогою лінійного перетворення

$$x = \frac{1}{2} [(\gamma_2 - \gamma_1)t + \gamma_2 + \gamma_1].$$

Тоді

$$P_n(x) = \tilde{P}_n(t), \quad t = \frac{2}{\gamma_2 - \gamma_1} x - \frac{\gamma_2 + \gamma_1}{\gamma_2 - \gamma_1}.$$

Умова нормування $P_n(0) = 1$ має вигляд

$$\tilde{P}_n(t_0) = 1, \quad t_0 = -\frac{1}{\rho_0}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \quad (1.65)$$

Отже, необхідно знайти поліном $\tilde{P}_n(t)$, який найменше відхиляється від нуля на відрізку $[-1, 1]$, тобто, щоб $\max |\tilde{P}_n(t)|$ був мінімальним і виконувалася умова нормування (1.65). Таким поліномом є

$$\tilde{P}_n(t) = \frac{T_n(t)}{T_n(t_0)},$$

де $T_n(t)$ — поліном Чебишева першого роду, який має вигляд

$$T_n(t) = \begin{cases} \cos(n \arccos t), & |t| \leq 1, \\ \frac{1}{2} \left[\left(t + \sqrt{t^2 - 1} \right)^n + \left(t - \sqrt{t^2 - 1} \right)^n \right], & |t| > 1. \end{cases} \quad (1.66)$$

Поліном Чебишева $T_n(t)$ має n нулів на відрізку $[-1, 1]$, які визначаються за формулою

$$t_k = \cos \frac{2k-1}{2n} \pi, \quad k = \overline{1, n},$$

а поліном $\overline{P_n}(x) = (1 - \tau_1 x)(1 - \tau_2 x) \cdots (1 - \tau_n x)$ має нулі $x_k = 1/\tau_k$, $k = \overline{1, n}$. Враховуючи зв'язок між x і t , $\tilde{P}_n(t)$ буде мати нулі

$$t_k = \frac{2x_k}{\gamma_2 - \gamma_1} - \frac{\gamma_2 + \gamma_1}{\gamma_2 - \gamma_1} = \frac{2}{\tau_k(\gamma_2 - \gamma_1)} - \frac{\gamma_2 + \gamma_1}{\gamma_2 - \gamma_1},$$

які повинні збігатися з нулями $T_n(t)$. Звідси

$$\tau_k = \frac{2}{\gamma_2 + \gamma_1 + (\gamma_2 - \gamma_1)t_k}.$$

Використовуючи позначення

$$\xi = \frac{\gamma_1}{\gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad (1.67)$$

запишемо параметри τ_k у вигляді

$$\tau_k = \frac{\tau_0}{1 + \rho_0 t_k}, \quad k = \overline{1, n}. \quad (1.68)$$

Знайдемо тепер

$$\begin{aligned} q_n &= \max_{\gamma_1 \leq x \leq \gamma_2} |P_n(x)| = \max_{-1 \leq t \leq 1} |\tilde{P}_n(t)| = \\ &= \max_{-1 \leq t \leq 1} \left| \frac{T_n(t)}{T_n(t_0)} \right| = \frac{1}{|T_n(t_0)|}, \end{aligned}$$

оскільки $\max_{-1 \leq t \leq 1} |T_n(t)| = 1$. Зауважимо, що $|t_0| = 1/\rho_0 > 1$, а тому, користуючись формулою (1.66) для $|T_n(t_0)|$, будемо мати

$$|T_n(t_0)| = \frac{1}{2} \left[\left(t_0 + \sqrt{t_0^2 - 1} \right)^n + \left(t_0 - \sqrt{t_0^2 - 1} \right)^n \right].$$

Перетворимо вираз у дужках

$$\begin{aligned} |t_0| \pm \sqrt{t_0^2 - 1} &= \frac{1}{\rho_0} \pm \sqrt{\frac{1}{\rho_0^2} - 1} = \frac{1}{\rho_0} \left(1 \pm \sqrt{1 - \rho_0^2} \right) = \\ &= \frac{1}{\rho_0} \left(1 \pm \frac{2\sqrt{\xi}}{1 + \xi} \right) = \frac{1}{\rho_0} \frac{(1 \pm \sqrt{\xi})^2}{1 + \xi} = \\ &= \frac{(1 \pm \sqrt{\xi})^2}{1 - \xi} = \frac{1 \pm \sqrt{\xi}}{1 \mp \sqrt{\xi}}, \end{aligned}$$

$$|t_0| + \sqrt{t_0^2 - 1} = \frac{1}{\rho_1}, \quad |t_0| - \sqrt{t_0^2 - 1} = \rho_1.$$

Звідси

$$q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}}.$$

Отже, для схеми (1.63) з набором параметрів (1.68), (1.67), яку називають *чебишевським ітераційним методом* (методом Річардсона), виконується оцінка

$$\|\mathbf{x}_n - \mathbf{x}\| \leq q_n \|\mathbf{x}_0 - \mathbf{x}\|,$$

де

$$q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}.$$

Визначимо $n = n(\varepsilon)$ так, щоб $q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}} \leq \varepsilon$. Тоді $\varepsilon\rho_1^{2n} - 2\rho_1^n + \varepsilon \geq 0$. Для цього достатньо $\rho_1^n \leq \varepsilon/2$ або

$$n(\varepsilon) \geq \frac{\ln(2/\varepsilon)}{\ln(2/\rho_1)}. \quad (1.69)$$

Оскільки

$$\ln \frac{1}{\rho_1} = \ln \frac{1 + \sqrt{\xi}}{1 - \sqrt{\xi}} > 2\sqrt{\xi},$$

то (1.69) можна замінити на

$$n(\varepsilon) > n_0(\varepsilon) = \frac{1}{2\sqrt{\xi}} \ln \frac{2}{\varepsilon}. \quad (1.70)$$

Ця оцінка зручніша, ніж (1.69).

Чебишевський ітераційний процес має одну особливість — ріст проміжних значень, який призводить до автоматичної зупинки комп'ютера та нагромадження похибок заокруглень, тобто для нього характерна обчислювальна нестійкість. Причина обчислювальної нестійкості полягає у тому, що норми $\|S_{n+1}\|$ оператора $S_{n+1} = I - \tau_{n+1}A$ для деяких значень ітераційних параметрів τ_n більші за одиницю. Тому, якщо на багатьох ітераціях поспіль використовують параметри τ_{n+1} , для яких $\|S_{n+1}\| > 1$, то відбувається нагромадження похибок заокруглень, що і призводить до обчислювальної нестійкості.

Щоб зменшити вплив обчислювальної нестійкості, необхідно розмістити параметри τ_k у такому порядку, щоб після параметра, для якого

норма оператора більша за одиницю, розміщувався параметр, для якого вона менша за 1. Такий набір параметрів називають стійким. Існують різні стійкі набори. Наприклад, нехай $n = 2^p$, $p > 0$ — ціле. Параметри τ_n однозначно визначаються нулями многочлена Чебишева t_n , а тому будемо говорити про впорядкування t_n . Стійкий набір t_n має вигляд

$$M_n^* = \left\{ -\cos \beta_i, \quad \beta_i = \frac{\pi}{2n} \theta_i^{(n)}, \quad i = \overline{1, n} \right\}, \quad n = 2^p,$$

де $\theta_i^{(n)}$ — одне з непарних чисел $1, 3, 5, \dots, 2n - 1$. Задача зводиться до впорядкування множини n непарних чисел

$$\theta_n = \left\{ \theta_1^{(n)}, \theta_2^{(n)}, \dots, \theta_n^{(n)} \right\}.$$

Виходячи з множини $\theta_1 = \{1\}$, побудуємо множину $\theta_n = \theta_{2^p}$ за формулами

$$\begin{aligned} \theta_{2i-1}^{(2m)} &= \theta_i^{(m)}, & \theta_{2i}^{(2m)} &= 4m - \theta_i^{(m)}, \\ i &= \overline{1, m}, & m &= 1, 2, \dots, 2^{p-1}, \end{aligned}$$

якщо $\theta_i^{(m)}$ відомі. Якщо, наприклад, $n = 2^4$, то послідовно знаходимо

$$\begin{aligned} \theta_1 &= \{1\}, & \theta_2 &= \{1, 3\}, & \theta_4 &= \{1, 7, 3, 5\}, \\ \theta_8 &= \{1, 15, 7, 9, 3, 13, 5, 11\}, \\ \theta_{16} &= \{1, 31, 15, 17, 7, 25, 9, 23, 3, 29, 13, 19, 5, 27, 11, 21\}. \end{aligned}$$

Одержані результати для явної чебишевської схеми можна перенести на неявну схему (1.39), звівши її до еквівалентної явної схеми.

1.4. Обчислення власних значень та власних векторів матриці

1.4.1. Алгебраїчна проблема власних значень

Алгебраїчна проблема власних значень формулюється так: знайти комплексні числа $\lambda_1, \lambda_2, \dots, \lambda_N$ та відповідні ненульові вектори $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$, які задовольняють рівняння

$$A\mathbf{x} = \lambda\mathbf{x}, \tag{1.71}$$

де A — задана комплексна матриця розміру $N \times N$. Числа $\lambda_1, \lambda_2, \dots, \lambda_N$ називаються *власними числами (значеннями)*, а вектори $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$,

власними векторами матриці A . Власні значення є коренями *характеристичного рівняння*

$$\det(A - \lambda I) = 0,$$

де I — одинична матриця. Ліва частина цього рівняння є поліномом степеня N по λ , а тому характеристичне рівняння має рівно N коренів з урахування кратності. Якщо власне значення λ_i знайдено, то відповідний власний вектор можна знайти як розв'язок однорідної системи рівнянь

$$(A - \lambda_i I) \mathbf{x} = 0. \quad (1.72)$$

Зауважимо, якщо навіть матриця A дійсна, її власні значення, а отже і вектори можуть бути комплексними.

Наведена схема — скласти характеристичне рівняння, знайти його корені і розв'язати однорідну систему (1.72), за виключенням найпростіших випадків непридатна як обчислювальна процедура для розв'язування практичних задач. Основна мета цього параграфу полягає у розгляді інших чисельних методів.

Однією з найбільш важливих операцій в теорії матриць є перетворення подібності. Дві матриці A і B розміру $N \times N$ називаються *подібними*, якщо існує невинроджена матриця P така, що $B = P^{-1}AP$. Перетворення подібності виникає з заміни змінних: розглянемо систему рівнянь $A\mathbf{x} = \mathbf{f}$ і зробимо заміну змінних $\mathbf{y} = P^{-1}\mathbf{x}$, $\boldsymbol{\varphi} = P^{-1}\mathbf{f}$, де P — невинроджена матриця. В нових змінних система рівнянь буде мати вигляд $AP\mathbf{y} = P\boldsymbol{\varphi}$ або після множення на P^{-1} , $P^{-1}AP\mathbf{y} = \boldsymbol{\varphi}$. Отже, матриця коефіцієнтів системи в нових змінних є $P^{-1}AP$, одержана перетворенням подібності з матриці A .

Важливою властивістю перетворення подібності є те, що воно зберігає власні значення, тобто матриці A та $P^{-1}AP$ мають однакові власні значення. Це випливає з характеристичного рівняння і того, що визначник добутку матриць дорівнює добутку визначників. Дійсно,

$$\begin{aligned} \det(A - \lambda I) &= \det(P^{-1}P) \det(A - \lambda I) = \\ &= \det P^{-1} \det(A - \lambda I) \det P = \det(P^{-1}AP - \lambda I). \end{aligned}$$

Звідси випливає, що характеристичні поліноми, а, отже, і власні значення матриць A і $P^{-1}AP$ збігаються. Однак власні вектори при перетворенні подібності змінюються. З рівності

$$P^{-1}AP\mathbf{y} = \lambda\mathbf{y}, \quad \text{або} \quad AP\mathbf{y} = \lambda P\mathbf{y}$$

видно, що власний вектор \mathbf{y} матриці $P^{-1}AP$ зв'язаний з власним вектором \mathbf{x} матриці A співвідношенням $\mathbf{x} = P\mathbf{y}$ або $\mathbf{y} = P^{-1}\mathbf{x}$.

Важливим є питання, до якого простого вигляду можна звести матрицю A за допомогою перетворення подібності.

► **ТЕОРЕМА 1.4.** *Матриця A подібна до діагональної матриці тоді і тільки тоді, якщо вона має N лінійно незалежних власних векторів.*

Доведення. Нехай $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ — N лінійно незалежних власних векторів матриці A , які відповідають власним значенням $\lambda_1, \lambda_2, \dots, \lambda_N$ і нехай P — матриця, стовпці якої $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$. Матриця P не вироджена, оскільки її стовпці лінійно незалежні. Тоді з рівності $A\mathbf{x}_i = \lambda_i\mathbf{x}_i$ випливає

$$AP = A(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) = (\lambda_1\mathbf{x}_1, \lambda_2\mathbf{x}_2, \dots, \lambda_N\mathbf{x}_N) = P\Lambda \quad (1.73)$$

Λ — діагональна матриця:

$$\Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{pmatrix}.$$

Оскільки співвідношення (1.73) еквівалентне $A = P\Lambda P^{-1}$, то матриця A подібна до діагональної матриці, у якої на діагоналі стоять власні значення матриці A . Навпаки, якщо матриця A подібна до діагональної матриці, то з (1.73) випливає, що стовпці матриці подібності P є власними векторами A і в силу не виродженості P вони є лінійно незалежні.

■

Наведемо два часткових випадки теореми.

► **ТЕОРЕМА 1.5.** *Якщо всі власні значення матриці A різні, то вона подібна до діагональної.*

► **ТЕОРЕМА 1.6.** *Якщо A — дійсна симетрична матриця (тобто $A = A^T$), то вона подібна до діагональної матриці, причому матриця подібності може бути взята ортогональною, тобто $(PP^T = P^TP = I)$.*

З теореми 1.5 випливає, що якщо матриця A не має N лінійно незалежних векторів, то вона обов'язково має кратні власні значення.

В загальному випадку за допомогою перетворення подібності матриця може бути зведена до найбільш близької до діагональної форми

$$J = \begin{pmatrix} \lambda_1 & \delta_1 & 0 & \dots & \dots & 0 \\ 0 & \lambda_2 & \delta_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & \lambda_{N-1} & \delta_{N-1} \\ 0 & \dots & \dots & \dots & 0 & \lambda_N \end{pmatrix}.$$

Тут λ_i — власні значення A , а δ_i дорівнюють або 1, або 0; припускається, що якщо $\delta_i \neq 0$, то $\lambda_i = \lambda_{i+1}$. Можна показати, що матриця A має $n - q$ лінійно незалежних власних векторів, де q — кількість ненульових δ_i . Отже, матриця J може бути розбита на клітки

$$J = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_p \end{pmatrix}, \quad (1.74)$$

де p — кількість лінійно незалежних власних векторів, і кожна клітка J_i є матриця вигляду

$$J_i = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{pmatrix}$$

з однаковими власними значеннями і одиницями над головною діагоналлю. Матрицю (1.74) називають *канонічною жордановою формою* матриці A . Зауважимо, якщо A має N лінійно незалежних власних векторів, то $p = N$, кожна клітка має розмір 1×1 і матриця J стає діагональною.

Багато обчислювальних алгоритмів використовують ортогональні або унітарні матриці. Комплексна матриця U називається *унітарною*, якщо вона задовольняє співвідношення $U^*U = I$, де U^* — матриця спряжена до U , тобто матриця, одержана заміною елементів U на комплексно спряжені та наступним транспонуванням, I — одинична матриця. Дійсна унітарна матриця називається *ортогональною*.

► **ТЕОРЕМА 1.7. (ШУРА)** Для будь-якої комплексної матриці A розміру $N \times N$ існує унітарна матриця U така, що

$$U^*AU = \begin{pmatrix} \lambda_1 & * & \dots & * \\ 0 & \lambda_2 & \dots & * \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \lambda_N \end{pmatrix},$$

де λ_i , $i = \overline{1, N}$ — власні значення матриці A (не обов'язково різні), а зірочкою позначені елементи, які можуть бути відмінні від нуля.

Якщо $A = A^*$, тобто A є ермітовою, то $(U^*AU)^* = U^*A^*U^{**} = U^*AU$ є також ермітовою і з теореми 1.7 випливає таке твердження.

► **ТЕОРЕМА 1.8.** Для довільної ермітової матриці A існує унітарна матриця U така, що

$$U^{-1}AU = U^*AU = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{pmatrix}.$$

При цьому власні значення λ_i , $i = \overline{1, N}$ матриці A є дійсними, а i -й стовпчик \mathbf{x}_i матриці U є власним вектором, який відповідає λ_i , тобто A має N лінійно незалежних ортогональних власних векторів.

1.4.2. QR -розклад

Метод Гаусса, який ми розглядали раніше ґрунтується на розкладі матриці A у вигляді

$$A = LR,$$

де L — нижня трикутна матриця з одиницями на головній діагоналі, а R — верхня трикутна матриця. Таке зображення матриці справджується не завжди (тоді, коли головні мінори відмінні від нуля), що є недоліком цього підходу.

Розклад

$$A = QR, \tag{1.75}$$

де Q — ортогональна матриця розміру $N \times N$, тобто така, що $QQ^T = Q^TQ = I$, а R — верхня трикутна матриця, називається QR -розкладом матриці A . Якщо ми маємо QR -розклад, то систему лінійних

алгебраїчних рівнянь $A\mathbf{x} = \mathbf{f}$ можна розв'язати (як і у випадку LR —розкладу Гаусса) в два етапи:

- 1) $Q\mathbf{z} = \mathbf{f}$, звідки $\mathbf{z} = Q^T\mathbf{f}$ (прямий хід);
- 2) $R\mathbf{x} = \mathbf{z}$, звідки \mathbf{x} легко знайти зворотньою підстановкою (зворотній хід).

Матрицею відображень Хаусхольдера називається матриця

$$P = I - 2\mathbf{w}\mathbf{w}^T,$$

де $\mathbf{w} \in \mathbb{R}^N$ задовольняє умову $\mathbf{w}^T\mathbf{w} = \|\mathbf{w}\|^2 = 1$.

Безпосередньою перевіркою переконуємося, що $P = P^T$,

$$\begin{aligned} PP^T &= (I - 2\mathbf{w}\mathbf{w}^T)(I - 2\mathbf{w}\mathbf{w}^T)^T = \\ &= I - 2\mathbf{w}\mathbf{w}^T - 2\mathbf{w}\mathbf{w}^T + 4\mathbf{w}\mathbf{w}^T\mathbf{w}\mathbf{w}^T = I. \end{aligned}$$

Тобто, матриця P — симетрична та ортогональна. Оскільки $P^2 = PP^T = I$, а всі власні значення матриці I дорівнюють 1, то всі власні значення матриці P задовольняють умову $\lambda_p^2 = 1$, тобто дорівнюють $+1$ або -1 .

За допомогою таких відображень Хаусхольдера можна довільну дійсну матрицю A звести до верхнього трикутного вигляду. Нехай вектор \mathbf{w}_1 вибраний так, що

$$\begin{aligned} \mathbf{w}_1^T &= \mu_1(a_{11} - s_1, a_{21}, \dots, a_{N1}), \\ s_1 &= \pm \left[\sum_{j=1}^N a_{j1}^2 \right]^{1/2}, \quad \mu_1 = [2s_1(s_1 - a_{11})]^{-1/2}. \end{aligned}$$

Щоб уникнути віднімання близьких чисел, покладемо

$$s_1 = -\text{sign}(a_{11}) \left[\sum_{j=1}^N a_{j1}^2 \right]^{1/2}.$$

Переконуємося, що так вибраний вектор \mathbf{w}_1 задовольняє умову $\mathbf{w}_1^T\mathbf{w}_1 = 1$. Дійсно,

$$\begin{aligned} \mathbf{w}_1^T\mathbf{w}_1 &= \mu_1^2 \left[(a_{11} - s_1)^2 + \sum_{j=2}^N a_{j1}^2 \right] = \\ &= \mu_1^2(a_{11}^2 - 2a_{11}s_1 + s_1^2 + s_1^2 - a_{11}^2) = 2s_1(s_1 - a_{11})\mu_1^2 = 1. \end{aligned}$$

Через \mathbf{a}_1 позначимо перший стовпець матриці A :

$$\mathbf{a}_1 = (a_{11}, a_{21}, \dots, a_{N1})^T.$$

Оскільки

$$\mathbf{w}_1^T \mathbf{a}_1 = \mu_1 \left((a_{11} - s_1)a_{11} + \sum_{j=2}^N a_{j1}^2 \right) = \mu_1 (s_1^2 - s_1 a_{11}) = \frac{1}{2\mu_1},$$

$$\mathbf{a}_1 - 2\mathbf{w}_1 \mathbf{w}_1^T \mathbf{a}_1 = (s_1, 0, \dots, 0)^T.$$

Тоді

$$A^{(1)} = (I - 2\mathbf{w}_1 \mathbf{w}_1^T)A = \begin{pmatrix} * & * & * & \dots & * \\ 0 & * & * & \dots & * \\ 0 & * & * & \dots & * \\ \dots & \dots & \dots & \dots & \dots \\ 0 & * & * & \dots & * \end{pmatrix}.$$

Далі виберемо вектор \mathbf{w}_2 , так, щоб

$$\mathbf{w}_2^T = \mu_2(0, a_{22}^{(1)} - s_2, a_{32}^{(1)}, \dots, a_{N2}^{(1)}),$$

$$s_2 = -\text{sign}(a_{22}^{(1)}) \left[\sum_{j=2}^N (a_{j2}^{(1)})^2 \right]^{1/2}, \quad \mu_2 = [2s_2(s_2 - a_{22}^{(1)})]^{-1/2},$$

де $a_{ij}^{(1)}$ — елементи матриці $A^{(1)}$. Тоді

$$A^{(3)} = (I - 2\mathbf{w}_2 \mathbf{w}_2^T)A^{(1)} = \begin{pmatrix} * & * & * & \dots & * \\ 0 & * & * & \dots & * \\ 0 & 0 & * & \dots & * \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & * & \dots & * \end{pmatrix}.$$

Продовжуючи цей процес за допомогою векторів \mathbf{w}_i , перші $i - 1$ координат яких нулі, отримаємо

$$(I - 2\mathbf{w}_{N-1} \mathbf{w}_{N-1}^T) \dots (I - 2\mathbf{w}_2 \mathbf{w}_2^T)(I - 2\mathbf{w}_1 \mathbf{w}_1^T)A = R, \quad (1.76)$$

де R — верхня трикутна матриця. Якщо покласти

$$Q = (I - 2\mathbf{w}_1 \mathbf{w}_1^T)(I - 2\mathbf{w}_2 \mathbf{w}_2^T) \dots (I - 2\mathbf{w}_{N-1} \mathbf{w}_{N-1}^T), \quad (1.77)$$

то (1.76) можна записати у вигляді $Q^T A = R$. Оскільки кожна матриця $I - 2\mathbf{w}_i \mathbf{w}_i^T$ ортогональна, то їх добуток Q також буде ортогональною матрицею. Отже, $Q^{-1} = Q^T$ і рівність (1.76) еквівалентна співвідношенню (1.75).

Зазначимо, що розклад (1.75) може бути здійснений завжди, без будь-яких обмежень на матрицю A . Алгоритм QR -розкладу потребує порядку $\frac{2}{3}N^3$ операцій множень і є одним з найбільш стійких до обчислювальної похибки.

У випадку комплексних A та \mathbf{f} матриця відображень $P = I - 2\mathbf{w}\mathbf{w}^*$, $\mathbf{w}^* = (\bar{w}_1, \bar{w}_2, \dots, \bar{w}_N)^T$ (\bar{z} — комплексно спряжене до z) буде унітарною з власними значеннями $\lambda_P = e^{i\varphi}$ (i — уявна одиниця).

1.4.3. QR -алгоритм знаходження власних значень та власних векторів

Розкладемо матрицю $A_0 = A$ у вигляді $Q_0 R_0$. Введемо нову матрицю A_1 , помінявши порядок співмножників Q_0 і R_0 , тобто

$$A_1 = R_0 Q_0.$$

Тоді

$$A_1 = R_0 Q_0 = Q_0^{-1} (Q_0 R_0) Q_0 = Q_0^{-1} A_0 Q_0.$$

Звідси випливає, що A_0 і A_1 подібні.

Припустимо, що матрицю A_1 можна розкласти

$$A_1 = Q_1 R_1,$$

де Q_1 — ортогональна, R_1 — трикутна матриця. Введемо нову матрицю

$$A_2 = R_1 Q_1.$$

І знову матриця A_2 подібна до A_1 . Якщо продовжити цей процес факторизації і переставляння співмножників, тоді отримаємо послідовність матриць

$$A_k = Q_k R_k, \quad A_{k+1} = R_k Q_k, \quad k = 0, 1, \dots \quad (1.78)$$

Всі матриці A_k подібні, а тому мають ті ж власні значення, що і матриця A .

► **ТЕОРЕМА 1.9.** Нехай в розкладі (1.75) матриці A всі діагональні мінори матриці Q не вироджені. Тоді послідовність матриць A_k (1.78) при $k \rightarrow \infty$ збігається по формі до кліткового правого трикутного вигляду.

QR -алгоритм в такому вигляді, як він описаний вище не достатньо ефективний. Оскільки розклад (1.76) вимагає порядку $O(N^3)$ операцій, то кожен крок процесу виконується занадто повільно. Для того, щоб обійти цю проблему, попередньо перетворимо матрицю A до вигляду

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1,N-1} & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2,N-1} & a_{2N} \\ 0 & a_{32} & \dots & a_{3,N-1} & a_{3N} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{N,N-1} & a_{NN} \end{pmatrix}. \quad (1.79)$$

Такі матриці, в яких нижче головної діагоналі є тільки одна ненульова діагональ, яка безпосередньо прилягає до головної, називають *матрицями Хессенберга*.

Матриця A за допомогою перетворень Хаусхольдера може бути ефективно зведена до матриці Хессенберга. Дійсно, якщо вектори \mathbf{w}_i , $i = 1, 2, \dots, N-2$ визначити за формулами

$$\mathbf{w}_1^T = \mu_1(0, a_{21} - s_1, a_{31}, \dots, a_{N1}),$$

$$s_1 = -\text{sign}(a_{21}) \left[\sum_{j=2}^N a_{j1}^2 \right]^{1/2}, \quad \mu_1 = [2s_1(s_1 - a_{21})]^{-1/2},$$

$$\mathbf{w}_2^T = \mu_2(0, 0, a_{32}^{(1)} - s_2, a_{42}^{(1)}, \dots, a_{N2}^{(1)}),$$

$$s_2 = -\text{sign}(a_{32}^{(1)}) \left[\sum_{j=3}^N \left(a_{j2}^{(1)} \right)^2 \right]^{1/2}, \quad \mu_2 = [2s_2(s_2 - a_{32}^{(1)})]^{-1/2},$$

.....

і виконати перетворення подібності $P_{N-2} \dots P_2 P_1 A$ з ортогональними матрицями $P_i = I - 2\mathbf{w}_i \mathbf{w}_i^T$, $i = 1, 2, \dots, N-2$, то в результаті отримаємо матрицю вигляду (1.79).

Отже, будемо вважати, що матриця A зведена до форми Хессенберга, і розглянемо питання про застосування до неї QR -розкладу. Насамперед зазначимо, що в цьому випадку розклад (1.75) здійснюється особливо просто. Це можна було б зробити, як і раніше, за допомогою перетворення Хаусхольдера, але краще використати операції, які

відомі як *перетворення Гівенса*. Ці операції задаються ортогональними матрицями вигляду

$$P_i = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & \sin \theta & \cos \theta & \\ & & & -\cos \theta & \sin \theta & \\ & & & & & 1 \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{pmatrix},$$

де перший синус стоїть в позиції індексу (i, i) . Помножимо тепер матрицю Гівенса P_1 зліва на матрицю Хессенберга

$$\begin{pmatrix} \sin \theta & \cos \theta & & & \\ -\cos \theta & \sin \theta & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix} \cdot \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1,N-1} & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2,N-1} & a_{2N} \\ 0 & a_{32} & \dots & a_{3,N-1} & a_{3N} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{N,N-1} & a_{NN} \end{pmatrix} =$$

$$= \begin{pmatrix} a_{11} \sin \theta + a_{21} \cos \theta & \dots & \dots & \dots & a_{1N} \sin \theta + a_{2N} \cos \theta \\ -a_{21} \cos \theta + a_{21} \sin \theta & \dots & \dots & \dots & -a_{1N} \cos \theta + a_{2N} \sin \theta \\ 0 & a_{32} & \dots & \dots & a_{3N} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{N,N-1} & a_{NN} \end{pmatrix}.$$

Якщо вибрати кут з умови

$$-a_{11} \cos \theta + a_{21} \sin \theta = 0 \quad \text{або} \quad \theta = \arctg \left(\frac{a_{11}}{a_{21}} \right),$$

то елемент $(2, 1)$ добутку перетвориться в нуль. Таким чином, в результаті множення на послідовність матриць Гівенса P_i з відповідними значеннями кутів всі піддіагональні елементи матриці Хессенберга послідовно перетворюються в нуль. Отже, ми прийдемо до розкладу $A = QR$, де $Q^T = P_{N-1} \dots P_1$, а тому $Q = P_1^T \dots P_{N-1}^T$. Цей розклад може бути одержаний за $O(N^2)$ операцій.

Зведення матриці до вигляду Хессенберга не мало б змісту, якби цю процедуру доводилося виконувати після кожного кроку QR -алгоритму.

На щастя це не так. Дійсно, відмінні від нуля не діагональні елементи матриці Гівенса P_i знаходяться тільки в позиціях $(i+1, i)$ і $(i, i+1)$. Звідси випливає, що матриця Q є матрицею Хессенберга. Оскільки R — верхня трикутна матриця, то добуток RQ також є матрицею Хессенберга. Отже, якщо матриця A зведена до вигляду Хессенберга, то і всі матриці A_k , які генеруються QR -алгоритмом, будуть автоматично зберігати цей вигляд.

Для прискорення збіжності застосовують варіант QR -алгоритму *зі зсувом*. А саме, будується послідовність ортогональних матриць Q_k і правих трикутних матриць R_k за рекурентними формулами

$$A_k - \nu_k I = Q_k R_k, \quad A_{k+1} = R_k Q_k + \nu_k I, \quad k = 0, 1, \dots,$$

де ν_k — число, яке близьке до власного значення. Матриці A_k подібні:

$$Q_k^{-1} A_k Q_k = Q_k^{-1} (Q_k R_k + \nu_k I) Q_k = A_{k+1}.$$

На практиці виявляється, що елемент $a_{NN}^{(k)}$ в позиції (N, N) матриці A_k є першим наближенням до власного значення. Тому цей елемент використовується для вибору ν_k . Критерієм збіжності служить достатня малість елемента $a_{N,N-1}^{(k)}$ матриці A_k . Коли ця малість буде досягнута, то $\lambda_N = a_{NN}^{(k)}$ і можна відкинути останній рядок і останній стовпець матриці і перейти до визначення власного значення λ_{N-1} , виходячи з отриманої підматриці розміру $(N-1) \times (N-1)$.

Попередні міркування ґрунтувалися на припущенні, що власне значення λ_N — дійсне. Нехай λ_N — комплексне. Тоді власні значення 2×2 — підматриці, розміщеної в правому нижньому кутку матриці A_k , згенерованої QR -алгоритмом без зсувів, збігаються до пари власних значень λ_N і $\lambda_{N-1} = \bar{\lambda}_N$, де $\bar{\lambda}_N$ — комплексно спряжене власне число. Тому природно за параметри зсувів вибрати власні значення цих 2×2 — підматриць. Нехай ν_1 і $\nu_2 = \bar{\nu}_1$ — власні значення 2×2 — підматриці, яка знаходиться в правому нижньому кутку матриці A_1 , тоді

$$A_1 - \nu_1 I = Q_1 R_1, \quad A_2 = R_1 Q_1 + \nu_1 I,$$

$$A_2 - \nu_2 I = Q_2 R_2, \quad A_3 = R_2 Q_2 + \nu_2 I.$$

Оскільки ν_1 і ν_2 — комплексні, то матриці A_1 , A_2 , Q_1 , Q_2 , R_1 , R_2 будуть взагалі кажучи комплексними.

Контрольні завдання

- ✎ 1.1. Методом Гаусса з вибором головного елемента у стовпчику розв'яжіть СЛАР

$$\begin{cases} x_1 + 4x_2 + 2x_3 = -2, \\ 2x_1 - 8x_2 + 3x_3 = 32, \\ x_2 + x_3 = 1. \end{cases}$$

- ✎ 1.2. Методом LU -розкладу розв'яжіть СЛАР

$$\begin{cases} x_1 - x_2 = -2, \\ -x_1 + 2x_2 - x_3 = -3, \\ -x_2 + 2x_3 = 4. \end{cases}$$

- ✎ 1.3. Методом Гаусса обчисліть визначник матриці

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 \\ 1 & 1 & 3 & 1 \\ 1 & 4 & 1 & 1 \end{pmatrix}.$$

- ✎ 1.4. Доведіть, що при множенні матриці розміру 4×4 зліва на матрицю переставлень

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

другий і четвертий рядок поміняються місцями, а перший і третій залишаються без змін.

- ✎ 1.5. Випишіть матрицю переставлень розміру 4×4 , яка переставляє перший і третій рядки, залишаючи другий і четвертий без змін.


- ✎ 1.6. Доведіть, що добуток двох матриць переставлень розміру $n \times n$ є матрицею переставлень. Доведіть, що матриця, обернена до матриці переставлень, є матрицею переставлень.

- ✎ 1.7. Розв'яжіть систему з завдання 1.2 методом LU -розкладу з вибором головного елемента.


 **1.8.** Методом прогонки розв'яжіть СЛАР

$$\begin{cases} 2x_1 + 2x_2 &= 1, \\ -x_1 + x_2 - 0,5x_3 &= 0, \\ x_2 - 3x_3 - x_4 &= 2, \\ x_3 + 2x_4 &= 2. \end{cases}$$

 **1.9.** Використовуючи властивості матричних норм, доведіть, що $\text{cond}(A) \geq 1$.

 **1.10.** Обчисліть $\text{cond}_1(A)$, $\text{cond}_\infty(A)$ для матриці

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 4 \end{pmatrix}.$$


 **1.11.** Використовуючи необхідну і достатню умову збіжності дослідіть збіжність ітераційних методів Якобі та Зейделя для СЛАР $A\mathbf{x} = \mathbf{f}$, де

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

 **1.12.** Задано СЛАР $A\mathbf{x} = \mathbf{f}$, де

$$A = \begin{pmatrix} 2 & -0,2 & 0,3 & 0,4 \\ 0,3 & -3 & 1 & -1,4 \\ 0,4 & 0,8 & 4 & 2,4 \\ -0,5 & 1,2 & -2,5 & -5 \end{pmatrix}.$$

Дослідіть збіжність методу Якобі розв'язування цієї системи.

 **1.13.** Доведіть, що для СЛАР $A\mathbf{x} = \mathbf{f}$, де

$$A = \begin{pmatrix} 2 & 0,3 & 0,5 \\ 0,1 & 3 & 0,4 \\ 0,1 & 0,1 & 4,8 \end{pmatrix},$$

метод послідовних наближень (1.52) буде збігатися за умови $0 < \tau < 0,4$.

✎ 1.14. Нехай A і B — матриці розміру $N \times N$. Доведіть, що матриці AB і BA подібні.

✎ 1.15. Доведіть, що якщо A і B — дійсні ортогональні матриці розміру $N \times N$, то матриця AB також є ортогональною.

✎ 1.16. Використовуючи результат завдання 1.15, покажіть, що якщо $\mathbf{w}_i^T \mathbf{w}_i = 1$, то матриця (1.77) ортогональна.

✎ 1.17. Виконайте QR -розклад матриці

$$A = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 3 & 1 \\ 1 & -1 & 2 \end{pmatrix}.$$

✎ 1.18. Проробивши детальні викладки, переконайтеся, що ортогональні перетворення подібності $P_i = I - 2\mathbf{w}_i \mathbf{w}_i^T$, $i = 1, 2, \dots, N-2$, де вектори визначені як у розділі 1.4.3, приводять несиметричну матрицю A до вигляду Хессенберга (1.79).

✎ 1.19. Доведіть, що якщо єдині відмінні від нуля не діагональні елементи матриці P_i знаходяться в позиціях $(i+1, i)$ і $(i, i+1)$, то добуток $P_1 P_2 \dots P_{N-1}$ є матрицею Хессенберга.

✎ 1.20. Доведіть, що якщо Q — матриця Хессенберга, а R — верхня трикутна, то добуток RQ є матриця Хессенберга.

РОЗДІЛ 2

МЕТОДИ РОЗВ'ЯЗУВАННЯ НЕЛІНІЙНИХ РІВНЯНЬ ТА СИСТЕМ

2.1. Чисельне розв'язування нелінійних рівнянь

Нехай задано рівняння:

$$f(x) = 0, \quad (2.1)$$

де $f(x)$ — неперервна функція.

Чисельне розв'язування рівняння (2.1) складається з двох етапів:

- 1) відокремлення коренів, тобто пошук проміжків, на яких є тільки один корінь рівняння;
- 2) обчислення коренів з наперед заданою точністю.

Для відокремлення коренів корисне відоме з аналізу твердження:

► **ТЕОРЕМА 2.1.** *Якщо неперервна функція $f(x)$ набуває різних знаків на кінцях відрізка $[a, b]$, тобто $f(a)f(b) < 0$, то в цьому проміжку є принаймні один корінь рівняння.*

Якщо, крім того, похідна існує і зберігає постійний знак у проміжку (a, b) , то корінь єдиний.

Універсальним методом відокремлення коренів є побудова графіка функції $y = f(x)$ за допомогою комп'ютера, тобто графічне відокремлення.

Наближені значення коренів уточнюють різними ітераційними методами. Розглянемо найефективніші з них.

2.1.1. Метод ділення навпіл (метод дихотомії або бісекції)

Нехай ми знайшли такі точки x_0, x_1 , що $f(x_0)f(x_1) < 0$ і на відріжку $[x_0, x_1]$ лежить лише один корінь рівняння (2.1). Обчислення будемо виконувати за такою схемою: покладемо $x_2 = (x_0 + x_1)/2$ і за x_3 виберемо

те із значень x_0 чи x_1 , для якого $f(x_2)f(x_3) < 0$, далі обчислюємо $f(x_4)$, $x_4 = (x_2 + x_3)/2$, і т.д. Цей процес продовжується доти, доки довжина відрізка, який містить корінь не стане меншою за ε . Середина останнього відрізка дає значення кореня з заданою точністю ε . Такий ітераційний процес, очевидно збігається зі швидкістю геометричної прогресії із знаменником $1/2$, тобто

$$|x_{n+1} - x_n| \leq \left(\frac{1}{2}\right)^n |x_1 - x_0|.$$

Основний недолік цього методу — повільна збіжність.

2.1.2. Метод послідовних наближень (простої ітерації)

Нехай на відрізку $[a, b]$ рівняння (2.1) має корінь x^* . Запишемо (2.1) у вигляді

$$x = \varphi(x), \quad (2.2)$$

де $\varphi(x) = x + \rho(x)f(x)$, $\rho(x)$ — довільна функція, яка не має коренів на $[a, b]$. Зокрема, $\rho(x) \equiv 1$.

Метод простої ітерації визначається формулою

$$x_{n+1} = \varphi(x_n), \quad n = 0, 1, 2, \dots, \quad (2.3)$$

де n — номер ітерації, x_0 — початкове наближення.

Справджується твердження.

► **ТЕОРЕМА 2.2.** Нехай функція $\varphi(x)$ у деякому околі $\Delta = \{x : |x - x_0| \leq \delta\}$ задовольняє умову Ліпшиця

$$|\varphi(x'') - \varphi(x')| \leq q |x'' - x'| \quad \forall x', x'' \in \Delta \quad (2.4)$$

із сталою Ліпшиця $q \in (0, 1)$, причому

$$|x_0 - \varphi(x_0)| \leq (1 - q)\delta.$$

Тоді рівняння (2.2) має в околі Δ єдиний корінь x^* , який є границею послідовності $\{x_n\}$, що визначається формулою (2.3).

Доведення. Покажемо, що $\varphi(x)$ відображає в банаховому просторі \mathbb{R}^1 замкнену кулю Δ в себе. Дійсно, якщо $x \in \Delta$, тобто $|x - x_0| \leq \delta$, то

$$\begin{aligned} |\varphi(x) - x_0| &= |\varphi(x) - \varphi(x_0) + \varphi(x_0) - x_0| \\ &\leq |\varphi(x) - \varphi(x_0)| + |\varphi(x_0) - x_0| \\ &\leq q |x - x_0| + (1 - q)\delta \leq q\delta + (1 - q)\delta = \delta. \end{aligned}$$

Крім того, $\varphi(x)$ на Δ — стискаюче відображення в силу умови Ліпшиця (2.4).

Отже, на підставі принципу стискаючих відображень в кулі Δ існує єдиний розв'язок рівняння (2.2). ■

Для похибки $z_{n+1} = x_{n+1} - x^*$ маємо оцінку

$$\begin{aligned} |z_{n+1}| &= |x_{n+1} - x^*| = |\varphi(x_n) - \varphi(x^*)| \leq q |x_n - x^*| = \\ &= q |z_n| \leq \dots \leq q^{n+1} |z_0|, \end{aligned}$$

тому кажуть, що метод послідовних наближень збігається із швидкістю геометричної прогресії зі знаменником q .

Якщо функція $\varphi(x)$ має похідну на Δ , то умова Ліпшиця виконується, коли $|\varphi'(x)| \leq q \ \forall x \in \Delta$, бо тоді згідно з формулою скінченних приростів $|\varphi(x'') - \varphi(x')| = |\varphi'(\xi)| |x'' - x'| \leq q |x'' - x'|$, де $\xi = x' + \theta(x'' - x')$, $0 < \theta < 1$. Більшу швидкість збіжності має метод Ньютона.

2.1.3. Метод Ньютона (метод дотичних)

Використовуючи формулу Тейлора з залишковим членом в формі Лагранжа, запишемо рівність

$$0 = f(x^*) = f(x_n) + (x^* - x_n)f'(x_n) + \frac{1}{2}(x^* - x_n)^2 f''(\xi),$$

$$\xi = x_n + \theta(x^* - x_n), \quad 0 < \theta < 1,$$

де x^* — точне значення кореня. Якщо у цьому розкладі відкинути останній член (залишковий член) і замінити x^* на x_{n+1}

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n)$$

або

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad f'(x_n) \neq 0, \quad n = 0, 1, 2, \dots, \quad (2.5)$$

то отримаємо *метод Ньютона*. Метод Ньютона називають також *методом дотичних*, оскільки нове наближення x_{n+1} є абсцисою точки перетину дотичної до графіка функції $y = f(x)$, проведеної в точці $(x_n, f(x_n))$, з віссю Ox (рис. 2.1). Записавши рівняння (2.5) у вигляді (2.3), де $\varphi(x) = x - f(x)/f'(x)$, помічаємо, що метод Ньютона є методом простої ітерації для (2.2). Припустимо, що відрізок $[a, b]$ містить єдиний

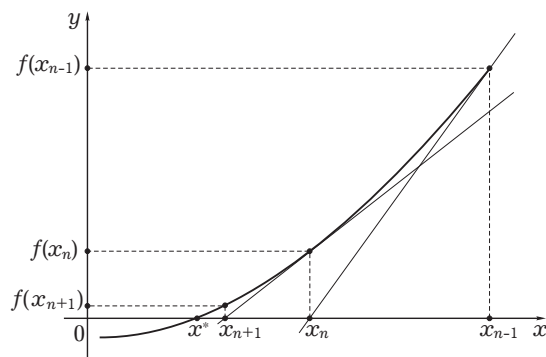


Рис. 2.1.

корінь x^* рівняння $f(x) = 0$ і функція має неперервні похідні першого і другого порядків, які не перетворюються в нуль на $[a, b]$. Тоді

$$\varphi'(x) = 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2},$$

причому $\varphi'(x^*) = 0$. Це означає, що існує окіл точки x^* , в якому $|\varphi'(x)| < 1$, і якщо початкове наближення x_0 взято з цього околу, то за теоремою 2.2 послідовність $\{x_n\}$, знайдена за методом Ньютона, буде збігатися до x^* .

Розглянемо теорему, яка конкретно вказує на вибір початкового наближення для одного класу функцій $f(x)$.

► **ТЕОРЕМА 2.3.** Нехай $f(a)f(b) < 0$, функції $f'(x), f''(x)$ неперервні і відмінні від нуля на $[a, b]$ або, що те саме, зберігають знак на $[a, b]$. Тоді, якщо початкове $x_0 \in [a, b]$ задовольняє умову $f(x_0)f''(x_0) > 0$, то послідовність $\{x_n\}$ методу Ньютона збігається до кореня $x^* \in [a, b]$.

Доведення. За умов теореми рівняння $f(x) = 0$ має лише один корінь x^* на $[a, b]$. Розглянемо випадок $f(a) < 0, f(b) > 0, f'(x) > 0, f''(x) > 0, x \in [a, b]$. Тоді точка $x_0 \in [a, b]$, яка задовольняє умову $f(x_0)f''(x_0) > 0$ міститься, очевидно, справа від x^* , тобто $x_0 > x^*, f(x_0) > 0$. Розглянемо $x_1 = x_0 - f(x_0)/f'(x_0)$. Згідно умов теореми маємо $x_1 < x_0$. Застосовуючи формулу Тейлора, одержимо

$$0 \equiv f(x^*) = f(x_0) + f'(x_0)(x^* - x_0) + \frac{1}{2}f''(\xi)(x^* - x_0)^2, \quad \xi \in (x^*, x_0),$$

$$f(x_0) = -f'(x_0)(x^* - x_0) - \frac{1}{2}f''(\xi)(x^* - x_0)^2,$$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = x_0 - x_0 + x^* + \frac{1}{2} \frac{f''(\xi)}{f'(x_0)} (x^* - x_0)^2 > x^*,$$

тобто $x_1 \in [x^*, x_0] \subset [a, b]$. Застосуємо метод математичної індукції. Припустимо, що $x_k \in [x^*, x_{k-1}] \subset [a, b]$ і доведемо, що $x_{k+1} \in [x^*, x_k]$. Дійсно, за формулою Тейлора

$$0 \equiv f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{1}{2} f''(\xi)(x^* - x_k)^2, \quad \xi \in (x^*, x_k),$$

і звідси

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - x_k + x^* + \frac{1}{2} \frac{f''(\xi)}{f'(x_k)} (x^* - x_k)^2 > x^*.$$

Оскільки за припущенням $x_k \in [x^*, x_{k-1}] \subset [a, b]$, то $f(x_k) > 0$, і тому

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

Отже, $x_{k+1} \in [x^*, x_k]$ що і треба було довести. А це означає, що послідовність $\{x_n\}$ монотонно спадає і обмежена знизу, тобто існує границя $\lim_{k \rightarrow \infty} x_k = \tilde{x}$. Перейшовши до границі в (2.5), переконуємося, що $\tilde{x} = x^*$. Для повного доведення теореми досить аналогічно розглянути інші можливі випадки розміщення знаків $f(a)$, $f(b)$, $f'(x)$, $f''(x)$. ■

Для оцінки похибки припустимо, що

$$\max_{x \in [a, b]} |f''(x)| = M_2, \quad \min_{x \in [a, b]} |f'(x)| = m.$$

Тоді за формулою Лагранжа

$$f(x_n) = f(x^*) + (x_n - x^*)f'(\xi)$$

або

$$x_n - x^* = \frac{f(x_n)}{f'(\xi)}.$$

За формулою Тейлора

$$f(x_n) = f(x_{n-1}) + (x_n - x_{n-1})f'(x_{n-1}) + \frac{1}{2} f''(\eta)(x_n - x_{n-1})^2,$$

$$\eta = x_{n-1} + \theta(x_n - x_{n-1}), \quad 0 < \theta < 1.$$

Оскільки $f(x_{n-1}) + (x_n - x_{n-1})f'(x_{n-1}) = 0$, то

$$|f(x_n)| \leq \frac{1}{2}M_2(x_n - x_{n-1})^2,$$

а тому

$$|x_n - x^*| \leq \frac{M_2}{2m}(x_n - x_{n-1})^2.$$

Ця оцінка є апостеріорною, а тому зручною для практичного застосування і свідчить про високу швидкість збіжності методу Ньютона. Недоліками методу є те, що на кожній ітерації потрібно обчислювати значення функції та її похідної, а також складність вибору початкового наближення.

2.1.4. Метод січних

Якщо в методі Ньютона похідну замінити різницею

$$\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}},$$

то одержимо ітераційний метод

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1})f(x_n)}{f(x_n) - f(x_{n-1})}. \quad (2.6)$$

Така заміна цілком природна, бо

$$\lim_{x_{n-1} \rightarrow x_n} \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} = f'(x_n).$$

Метод січних (2.6), на відміну від попередніх методів, є двокровним, тобто нове наближення x_{n+1} визначається через дві попередні ітерації x_n і x_{n-1} , а тому необхідно задавати два початкових наближення x_0 і x_1 .

Геометрична інтерпретація методу січних полягає у наступному. Через точки $(x_{n-1}, f(x_{n-1}))$, $(x_n, f(x_n))$ проводимо пряму, і абсцисса точки перетину цієї прямої з віссю Ox і є новим наближенням x_{n+1} (рис. 2.2).

Зауваження 2.1. При застосуванні ітераційних методів (послідовних наближень, Ньютона, січних) виникає питання, коли припинити ітераційний процес, щоб одержати розв'язок з заданою точністю ε . Як правило використовують найпростіші умови

$|f(x_{n+1})| < \varepsilon$ або $|x_{n+1} - x_n| < \varepsilon$. Перша умова може дати недостовірний результат, якщо функція $f(x)$ поблизу кореня є дуже пологою, що є можливим у випадку кратного кореня. Друга умова може привести до невірної результату в різних випадках у залежності від конкретного ітераційного методу. Наприклад, у випадку методу Ньютона це може відбутися, якщо на деякій ітерації похідна виявляється дуже великою. Інколи перевіряють обидві умови.

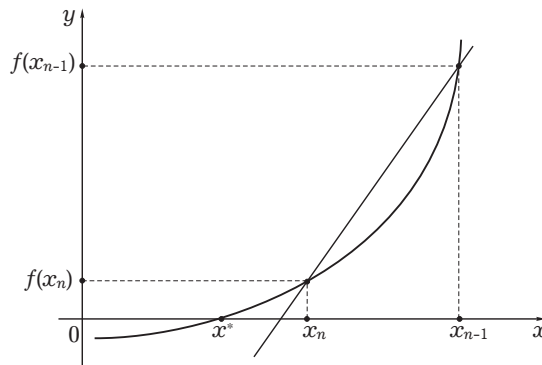


Рис. 2.2.

Приклад 2.1. Доведіть, що для рівняння $x - \frac{1}{1+x^2} = 0$ послідовні наближення $x_{n+1} = \frac{1}{1+x_n^2}$ збігаються до єдиного кореня при $\forall x_0$. Оцініть n , при якому для похибки $z_n = x_n - x^*$ (x^* — точний розв'язок) виконується нерівність $|z_n| \leq \varepsilon |z_0|$, $\varepsilon = 10^{-6}$.

▷ Оскільки функція $\varphi(x) = \frac{1}{1+x^2}$ диференційовна, то умова збіжності методу послідовних наближень $g(x) = |\varphi'(x)| = \frac{2|x|}{(1+x^2)^2} < 1$. Знайдемо найбільше значення функції $g(x)$. З рівностей

$$g'(x) = \frac{2(1-3x^2)}{(1+x^2)^3} = 0, \quad x \geq 0, \quad g'(x) = -\frac{2(1-3x^2)}{(1+x^2)^3} = 0, \quad x < 0$$

випливає, що критичні точки цієї функції $x = \pm\sqrt{3}/3$. Отже, найбільше значення функції $g(x)$ дорівнює $q = g(\pm\sqrt{3}/3) = 3\sqrt{3}/8 < 1$. А тому, згідно принципу стискаючих відображень, рівняння $x - \frac{1}{1+x^2} = 0$ має єдиний розв'язок і послідовні наближення $x_{n+1} = \frac{1}{1+x_n^2}$ збігаються до цього розв'язку при $\forall x_0$.

Для похибки ітераційного методу справджується нерівність $|z_n| \leq q^n |z_0| \leq \varepsilon |z_0|$, з якої випливає

$$\left(\frac{1}{q}\right)^n \geq \frac{1}{\varepsilon}$$

або

$$n(\varepsilon) \geq \frac{\ln(1/\varepsilon)}{\ln(1/q)} \approx 19,93.$$

Отже, $n = 20$. ◀

Приклад 2.2. Виділіть графічно корені рівняння $x - 2 \sin x = 0$. Доведіть, що ітерації $x_{n+1} = 2 \sin x_n$, $n = 0, 1, \dots$, збігаються до кореня з інтервалу $(\pi/2, 2)$ для $\forall x_0 \in (\pi/2, 2)$. Застосуйте метод Ньютона до цього рівняння і встановіть за яких початкових наближень x_0 ітерації будуть збіжні до кореня з інтервалу $(\pi/2)$.

► Побудуємо графіки функцій $y = x/2$, $y = \sin x$. З рис. 2.3 видно, що рівняння $x - 2 \sin x = 0$ має три корені: один $x = 0$, другий на інтервалі $(\pi/2, 2)$, третій на інтервалі $(-2, -\pi/2)$. Дослідимо на збіжність ітераційний метод послідовних наближень $x_{n+1} = 2 \sin x_n$, $n = 0, 1, \dots$, на інтервалі $(\pi/2, 2)$. Оскільки при $x \in (\pi/2, \pi)$ функція

$$|\varphi'(x)| = 2 |\cos x| = -2 \cos x$$

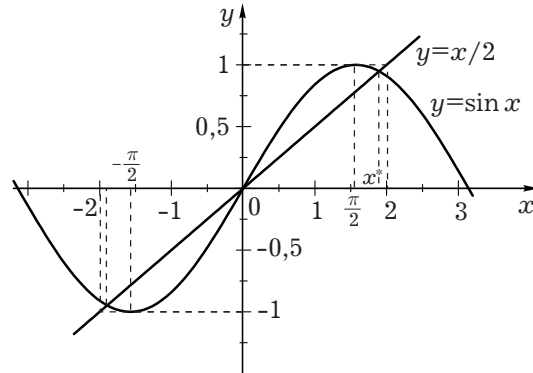
зростає, то $|\varphi'(x)| < -2 \cos \frac{2\pi}{3} < 1 \quad \forall x \in (\pi/2, 2)$. Отже, метод послідовних наближень збіжний $\forall x_0 \in (\pi/2, 2)$. Застосуємо метод Ньютона (2.5), де $f(x) = x - 2 \sin x$, $f'(x) = 1 - 2 \cos x$. Тоді одержимо ітераційний процес

$$x_{n+1} = x_n - \frac{x_n - 2 \sin x_n}{1 - 2 \cos x_n}, \quad n = 0, 1, \dots \quad (2.7)$$

Для знаходження початкового наближення x_0 , за якого цей ітераційний метод буде збіжний на інтервалі $(\pi/2, 2)$, використаємо теорему 2.3. Маємо

$$f(x_0)f''(x_0) = 2(x_0 - 2 \sin x_0) \sin x_0 > 0.$$

Якщо $x_0 \in (\pi/2, 2)$, то $\sin x_0 > 0$. Звідси випливає, що повинна виконуватися нерівність $x_0 - 2 \sin x_0 > 0$. Таким чином (див. рис. 2.3), за умови $x_0 \in (x^*, 2)$ (x^* —точний розв'язок вихідного рівняння) ітераційний метод (2.7) буде збігатися до кореня з інтервалу $(\pi/2, 2)$. ◀

Рис. 2.3. Графіки функцій $y = x/2$ і $y = \sin x$

2.2. Розв'язування систем нелінійних рівнянь

Розглянемо систему нелінійних алгебраїчних рівнянь

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}, \quad (2.8)$$

де $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_N(\mathbf{x}))^T$, $\mathbf{x} = (x_1, x_2, \dots, x_N)^T$ або у координатному вигляді

$$\begin{aligned} f_1(x_1, x_2, \dots, x_N) &= 0, \\ f_2(x_1, x_2, \dots, x_N) &= 0, \\ &\dots\dots\dots \\ f_N(x_1, x_2, \dots, x_N) &= 0. \end{aligned}$$

2.2.1. Метод послідовних наближень (простої ітерації)

Систему (2.8) замінимо еквівалентною системою

$$\mathbf{x} = \boldsymbol{\varphi}(\mathbf{x}), \quad (2.9)$$

або

$$x_i = \varphi_i(x_1, x_2, \dots, x_N), \quad i = \overline{1, N},$$

а ітерації будемо проводити за формулою

$$\mathbf{x}_{n+1} = \boldsymbol{\varphi}(\mathbf{x}_n), \quad n = 0, 1, \dots, \quad \mathbf{x}_n = (x_1^n, x_2^n, \dots, x_N^n)^T \quad (2.10)$$

або

$$x_i^{n+1} = \varphi_i(x_1^n, x_2^n, \dots, x_N^n), \quad n = 0, 1, \dots, \quad i = \overline{1, N}.$$

► **ТЕОРЕМА 2.4.** Нехай вектор-функція $\varphi(\mathbf{x})$ у деякому околі $\Delta = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\| \leq \delta\}$ задовольняє умову Ліпшиця

$$\|\varphi(\mathbf{x}'') - \varphi(\mathbf{x}')\| \leq q \|\mathbf{x}'' - \mathbf{x}'\|, \quad \forall \mathbf{x}', \mathbf{x}'' \in \Delta.$$

із сталою $q \in (0, 1)$, причому

$$\|\mathbf{x}_0 - \varphi(\mathbf{x}_0)\| \leq (1 - q)\delta.$$

Тоді система рівнянь (2.9) має в околі Δ єдиний корінь \mathbf{x}^* , який є границею послідовності $\{\mathbf{x}_n\}$, що визначається формулою (2.10).

Доведення. Доведення аналогічне до доведення теореми 2.2. ■

Зазначимо, що для диференційовної в Δ вектор-функції $\varphi(\mathbf{x})$ умова Ліпшиця виконується, коли $\left\| \frac{\partial \varphi(\mathbf{x})}{\partial \mathbf{x}} \right\| \leq q, \quad \forall \mathbf{x} \in \Delta$. Тут $\frac{\partial \varphi(\mathbf{x})}{\partial \mathbf{x}}$ — матриця Якобі вектор-функції φ , обчислена в точці \mathbf{x} .

2.2.2. Метод Ньютона

Для чисельного розв'язування задачі (2.8) розглянемо ітераційний метод Ньютона

$$\mathbf{x}_{n+1} = \mathbf{x}_n - [J(\mathbf{x}_n)]^{-1} \mathbf{f}(\mathbf{x}_n), \quad n = 0, 1, \dots, \quad (2.11)$$

де $\mathbf{x}_n = (x_1^n, x_2^n, \dots, x_N^n)^T$, $J(\mathbf{x}) = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}$ — матриця Якобі вектор-функції $\mathbf{f}(\mathbf{x})$ в точці \mathbf{x}

$$J(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f_1(\mathbf{x})}{\partial x_N} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f_2(\mathbf{x})}{\partial x_N} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f_N(\mathbf{x})}{\partial x_1} & \frac{\partial f_N(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f_N(\mathbf{x})}{\partial x_N} \end{pmatrix}.$$

Обчислювальну схему (2.11) запишемо у вигляді

$$J(\mathbf{x}_n) \mathbf{z}_n = -\mathbf{f}(\mathbf{x}_n), \quad (2.12)$$

де $\mathbf{z}_n = \mathbf{x}_{n+1} - \mathbf{x}_n$. Отже, кожен крок ітераційного методу Ньютона зводиться до розв'язування системи лінійних алгебраїчних рівнянь (2.12) та знаходження наступної ітерації за формулою


$$\mathbf{x}_{n+1} = \mathbf{x}_n + \mathbf{z}_n, \quad n = 0, 1, \dots$$

Переваги цього ітераційного процесу — в його швидкості збіжності. Недоліки — в складності вибору початкового наближення та у великому обсязі обчислювальної роботи у зв'язку з необхідністю на кожній ітерації обчислювати $J(\mathbf{x}_n)$. Тому часто застосовують *модифікований* метод Ньютона

$$J(\mathbf{x}_0)\mathbf{z}_n = -\mathbf{f}(\mathbf{x}_n), \quad \mathbf{x}_{n+1} = \mathbf{x}_n + \mathbf{z}_n, \quad n = 0, 1, \dots \quad (2.13)$$


У методі (2.13) достатньо один раз обчислити матрицю Якобі та у випадку, коли $n = \overline{0, M-1}$, розв'язати M систем лінійних алгебраїчних рівнянь з однією і тією ж матрицею $J(\mathbf{x}_0)$, що за великих N вимагає $\frac{1}{3}N^3 + N^2M$ операцій множення та ділення, тоді як реалізація алгоритму (2.12) вимагає $\frac{1}{3}MN^3$ операцій множення та ділення.


Контрольні завдання


 **2.1.** Нехай k — додатне ціле і α — додатне число. Доведіть, що застосування методу Ньютона до рівняння $x^k - \alpha = 0$ приводить до послідовності ітерацій


$$x_{n+1} = \frac{1}{k} \left[(k-1)x_n + \frac{\alpha}{x_n^{k-1}} \right], \quad n = 0, 1, \dots,$$


яка збігається $\forall x_0 > 0$.

 **2.2.** Для рівняння $f(x) = x - x^3 = 0$, яке має корені 0 і ± 1 доведіть, що метод Ньютона локально збігається до кожного з цих коренів. Визначте інтервал, для якого ітерації Ньютона будуть збіжні до коренів рівняння за будь-якого початкового наближення x_0 з цього інтервалу.

 **2.3.** Виділіть графічно корені рівняння $f(x) = x + 2 - \exp(x) = 0$ і вкажіть інтервали розташування коренів. Визначте на якому інтервалі ітерації $x_{n+1} = \exp(x_n) - 2$, $n = 0, 1, \dots$ збігаються до кореня для $\forall x_0$, що належить цьому інтервалу. Застосуйте метод Ньютона. За яких початкових наближень ітерації будуть збігатись до кореня?

 **2.4.** Для рівняння $f(x) = \frac{x}{2} + \operatorname{tg} x - 1 = 0$ виконайте завдання 2.3.

 **2.5.** Для рівняння $f(x) = x - \frac{1}{x^5 + 1} = 0$ виконайте завдання 2.3.

 **2.6.** Покажіть графічно, що система рівнянь

$$\begin{cases} x_1^2 + x_2^2 = 1, \\ x_1^2 - x_1 = 0 \end{cases}$$

має рівно два розв'язки.


 **2.7.** Доведіть, що метод послідовних наближень

$$\begin{cases} x_1^{n+1} = 1,1 - \sin\left(\frac{x_2^n}{3}\right) + \ln\left(1 + \frac{x_1^n + x_2^n}{5}\right), \\ x_2^{n+1} = 0,5 + \cos\left(\frac{x_1^n x_2^n}{6}\right), \quad n = 0, 1, 2, \dots \end{cases}$$

$$x_1^0 = x_2^0 = 1$$


збігається до єдиного розв'язку системи

$$\begin{cases} x_1 = 1,1 - \sin\left(\frac{x_2}{3}\right) + \ln\left(1 + \frac{x_1 + x_2}{5}\right), \\ x_2 = 0,5 + \cos\left(\frac{x_1 x_2}{6}\right). \end{cases}$$

 **2.8.** Обчисліть матрицю Якобі $J(\mathbf{x}) = \partial \mathbf{f}(\mathbf{x}) / \partial \mathbf{x}$ для вектор-функції

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} x_1^2 + x_1 x_2 x_3 + x_3^3 \\ x_1^3 + x_2 x_3^2 \\ x_1 / x_3^2 \end{pmatrix}.$$

 **2.9.** Випишіть ітераційний метод Ньютона для системи рівнянь з вправ 2.6.

 **2.10.** Запишіть ітераційний метод Ньютона (2.11) у вигляді (2.10) та на підставі теореми 2.4 отримайте умови збіжності ітераційного методу Ньютона за умови, що вектор-функція $\mathbf{f}(\mathbf{x})$ неперервно диференційовна в околі розв'язку системи \mathbf{x}^* .

РОЗДІЛ 3

НАБЛИЖЕННЯ ФУНКЦІЙ

3.1. Постановка задачі наближення функції

Найпростіша задача наближення функції полягає у наступному. В дискретні моменти часу x_0, x_1, \dots, x_n спостерігаються (відомі) значення функції $f(x)$; необхідно знайти її значення при інших x .

Інколи з деяких додаткових міркувань відомо, що функцію, яку потрібно наближити, доцільно шукати у вигляді

$$f(x) \approx g(x; a_0, a_1, \dots, a_n).$$

Якщо параметри a_0, a_1, \dots, a_n визначаються з умов

$$f(x_i) = g(x_i; a_0, a_1, \dots, a_n), \quad i = \overline{0, n},$$

де x_i — так звані вузли інтерполяції, то такий спосіб наближення називають *інтерполяцією* або *інтерполюванням*.

Нехай y_1 — найменше з чисел x_i — вузлів інтерполяції, а y_2 — найбільше з них. Якщо точка, в якій обчислюється значення $f(x)$ лежить зовні $[y_1, y_2]$, то разом з терміном інтерполяція використовується термін *екстраполяція*. Якщо вузли інтерполяції вибрано далеко від екстраполяційної точки, то слабо використовується суттєва інформація про поведінку змінної.

Найбільш часто використовується інтерполяція многочленами. Однак, це не єдино можливий тип інтерполяції. Інколи зручно наближати функцію тригонометричними функціями, а також раціональними функціями.

3.2. Інтерполяційний многочлен Лагранжа

Серед способів інтерполювання найбільш поширений випадок лінійного інтерполювання, коли наближення шукають у вигляді:

$$g(x, a_0, \dots, a_n) = \sum_{i=0}^n a_i \varphi_i(x),$$

де $\varphi_i(x)$ — фіксовані функції, значення коефіцієнтів a_i визначаються з умов:

$$f(x_j) = \sum_{i=0}^n a_i \varphi_i(x_j), \quad j = \overline{0, n}. \quad (3.1)$$

Метод розв'язування задачі, при якому коефіцієнти a_i визначаються безпосереднім розв'язування системи (3.1), називається *методом неозначених коефіцієнтів*.

Найчастіше на практиці використовується інтерполяція многочленами:

$$L_n(x) = \sum_{i=0}^n a_i x^i, \quad (3.2)$$

тоді $\varphi_i(x) = x^i$, $i = \overline{0, n}$ і система рівнянь (3.1) має вигляд:

$$\sum_{i=0}^n a_i x_j^i = f(x_j), \quad j = \overline{0, n}. \quad (3.3)$$

Припустимо, що всі x_j різні. Визначник цієї системи є визначником Вандермонда:

$$\Delta = \begin{vmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{vmatrix} \neq 0.$$

Отже, система завжди має єдиний розв'язок. Таким чином, доведено існування та єдиність інтерполяційного многочлена (3.2).

Безпосереднє знаходження коефіцієнтів за допомогою розв'язування цієї системи вже при порівняно невеликих n ($n = 20$) призводить до великої обчислювальної похибки.

Будемо шукати явне представлення інтерполяційного многочлена, не розв'язуючи систему (3.3). Задача інтерполювання буде розв'язана,

якщо побудувати многочлени $\Phi_i(x)$, $i = \overline{0, n}$ степеня не вище n такі, що:

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

Тоді многочлен

$$L_n(x) = \sum_{i=0}^n f(x_i) \Phi_i(x) \quad (3.4)$$

буде шуканим інтерполяційним многочленом. Дійсно,

$$L_n(x_j) = \sum_{i=0}^n f(x_i) \Phi_i(x_j) = f(x_j), \quad j = \overline{0, n}.$$

Крім того, $L_n(x)$ — многочлен степеня n . Многочлени $\Phi_i(x)$ будемо шукати у вигляді:

$$\Phi_i(x) = C_i(x - x_0)(x - x_1) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n),$$

де C_i — неозначені коефіцієнти, які знайдемо з умови $\Phi_i(x_i) = 1$. Тоді

$$C_i = [(x_i - x_0)(x_i - x_1) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)]^{-1}.$$

Отже,

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

Інтерполяційний многочлен, записаний у вигляді:

$$L_n(x) = \sum_{i=0}^n f(x_i) \prod_{j \neq i} \frac{x - x_j}{x_i - x_j} \quad (3.5)$$

називають *інтерполяційним многочленом Лагранжа*.

Існують інші форми запису цього ж інтерполяційного многочлена, наприклад, інтерполяційна формула Ньютона, яку ми будемо розглядати далі.

3.3. Розділені різниці. Інтерполяційна формула Ньютона

За означенням *розділена різниця нульового порядку* $f(x_i)$ від функції $f(x)$ по одному вузлу x_i збігається з значенням функції $f(x_i)$. *Розділені різниці першого порядку* визначаються рівністю:

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

різниці другого 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. Розділені різниці k -го порядку $f(x_0; x_1; \dots; x_k)$ визначаються через різниці $k - 1$ порядку за формулою:

$$f(x_0; x_1; \dots; x_k) = \frac{f(x_1; x_2; \dots; x_k) - f(x_0; x_1; \dots; x_{k-1})}{x_k - x_0}.$$

⇒ **Лема 3.1.** Справджується рівність

$$f(x_0; x_1; \dots; x_k) = \sum_{i=0}^k \frac{f(x_i)}{\prod_{j \neq i} (x_i - x_j)}. \quad (3.6)$$

Доведення. Доведення проведемо методом математичної індукції. При $k = 0$ ця рівність перетворюється в рівність $f(x_0) = f(x_0)$, при $k = 1$

$$f(x_0; x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0}.$$

Нехай (3.6) доведена при $k \leq l$. Тоді

$$\begin{aligned} f(x_0; \dots; x_{l+1}) &= \frac{f(x_1; \dots; x_{l+1}) - f(x_0; \dots; x_l)}{x_{l+1} - x_0} = \\ &= \frac{1}{x_{l+1} - x_0} \left(\sum_{i=1}^{l+1} \frac{f(x_i)}{\prod_{\substack{j \neq i \\ 1 \leq j \leq l+1}} (x_i - x_j)} - \sum_{i=0}^l \frac{f(x_i)}{\prod_{\substack{j \neq i \\ 0 \leq j \leq l}} (x_i - x_j)} \right). \end{aligned}$$

Якщо $i \neq 0, l + 1$, то коефіцієнт при $f(x_i)$ в правій частині є

$$\begin{aligned} &\frac{1}{x_{l+1} - x_0} \left(\frac{1}{\prod_{\substack{j \neq i \\ 1 \leq j \leq l+1}} (x_i - x_j)} - \frac{1}{\prod_{\substack{j \neq i \\ 0 \leq j \leq l}} (x_i - x_j)} \right) = \\ &= \frac{(x_i - x_0) - (x_i - x_{l+1})}{(x_{l+1} - x_0) \prod_{\substack{j \neq i \\ 0 \leq j \leq l+1}} (x_i - x_j)} = \frac{1}{\prod_{\substack{j \neq i \\ 0 \leq j \leq l+1}} (x_i - x_j)}. \end{aligned}$$

Для $i = 0$ значення $f(x_0)$ входить лише в один доданок в правій частині, а коефіцієнт при ньому має вигляд

$$-\frac{1}{(x_{l+1} - x_0) \prod_{\substack{j \neq 0 \\ 0 \leq j \leq l}} (x_0 - x_j)} = \frac{1}{\prod_{\substack{j \neq 0 \\ 0 \leq j \leq l+1}} (x_0 - x_j)},$$

а коефіцієнт при $f(x_{l+1})$ є

$$\frac{1}{(x_{l+1} - x_0) \prod_{\substack{j \neq l+1 \\ 1 \leq j \leq l+1}} (x_{l+1} - x_j)} = \frac{1}{\prod_{\substack{j \neq l+1 \\ 0 \leq j \leq l+1}} (x_{l+1} - x_j)}.$$

Ми показали, що коефіцієнти при $f(x_i)$, $i = 0, 1, \dots, k$ збігаються з коефіцієнтами формули (3.6). ■

Розділена різниця є симетричною функцією своїх аргументів x_0, x_1, \dots, x_k , тобто не змінюється при будь-якому їх переставлянні.

Якщо функція задана в точках x_0, x_1, \dots, x_n , то таблицю

$$\begin{array}{ccccccc} f(x_0) & & & & & & \\ f(x_1) & f(x_0; x_1) & & & & & \\ f(x_2) & f(x_1; x_2) & f(x_0; x_1; x_2) & & & & \\ & \cdots & \cdots & \cdots & \cdots & f(x_0; x_1; \dots; x_n) & \\ \cdots & \cdots & f(x_{n-2}; x_{n-1}; x_n) & \cdots & & & \\ f(x_n) & f(x_{n-1}; x_n) & & & & & \end{array}$$

називають *таблицею розділених різниць*.

Нехай $P_n(x)$ многочлен n -го степеня. Різниця $P_n(x) - P_n(x_0)$ перетворюється в нуль при $x = x_0$, а тому ділиться на $x - x_0$. Отже, перша розділена різниця многочлена n -го степеня

$$P_n(x; x_0) = \frac{P_n(x_0) - P_n(x)}{x_0 - x}$$

є многочленом степеня $n - 1$ відносно x . Розглянемо розділену різницю

$$P_n(x; x_0; x_1) = \frac{P_n(x_0; x_1) - P_n(x; x_0)}{x - x_1}.$$

Чисельник цього дробу перетворюється в нуль при $x = x_1$. Отже, друга розділена різниця є многочленом степеня $n - 2$. Продовжуючи ці

міркування, дійдемо висновку, що $P_n(x, x_0, \dots, x_{n-1})$ є многочленом нульового степеня, тобто константою, а розділені різниці вищого порядку ніж n , дорівнюють нулю.

З означення розділених різниць випливає

$$\begin{aligned} P_n(x) &= P_n(x_0) + (x - x_0)P_n(x; x_0), \\ P_n(x; x_0) &= P_n(x_0; x_1) + (x - x_1)P_n(x; x_0; x_1), \\ P_n(x; x_0; x_1) &= P_n(x_0; x_1; x_2) + (x - x_2)P_n(x; x_0; x_1; x_2), \\ &\dots \end{aligned}$$

$$\begin{aligned} P_n(x; x_0; x_1; \dots; x_{n-1}) &= P_n(x_0; x_1; \dots; x_n) + \\ &\quad + (x - x_n)P_n(x; x_0; x_1; \dots; x_n) = \\ &= P_n(x_0; x_1; \dots; x_n). \end{aligned}$$

Звідси для $P_n(x)$ дістаємо формулу:

$$\begin{aligned} P_n(x) &= P_n(x_0) + (x - x_0)P_n(x_0; x_1) + \\ &\quad + (x - x_0)(x - x_1)P_n(x_0; x_1; x_2) + \dots \\ &\quad + (x - x_0)(x - x_1) \dots (x - x_{n-1})P_n(x_0; x_1; \dots; x_n). \end{aligned}$$

Якщо $L_n(x)$ інтерполяційний многочлен для функції $f(x)$, то його значення у вузлах x_0, x_1, \dots, x_n збігається із значенням функції $f(x)$, а отже, збігаються і розділені різниці, тому інтерполяційний многочлен для функції $f(x)$ можна записати у вигляді:

$$\begin{aligned} L_n(x) &= f(x_0) + (x - x_0)f(x_0; x_1) + \\ &\quad + (x - x_0)(x - x_1)f(x_0; x_1; x_2) + \dots \\ &\quad + (x - x_0)(x - x_1) \dots (x - x_{n-1})f(x_0; x_1; \dots; x_n). \end{aligned} \quad (3.7)$$

Такий запис інтерполяційного многочлена називають інтерполяційним многочленом у формі Ньютона, а формулу (3.7) — *інтерполяційною формулою Ньютона*.

Якщо відомі розділені різниці (таблиця розділених різниць), то многочлен Ньютона зручно обчислювати за схемою Горнера:

$$\begin{aligned} L_n(x) &= f(x_0) + (x - x_0)(f(x_0; x_1) + (x - x_1)(f(x_0; x_1; x_2) + \\ &\quad + \dots + (x - x_{n-1})f(x_0; x_1; \dots; x_n)) \dots). \end{aligned} \quad (3.8)$$

Неважко помітити, що коли покласти

$$b_0 = 0, \quad b_k = (x - x_{n-k+1})b_{k-1} + f(x_0; \dots; x_{n-k+1}), \quad k = \overline{1, n+1},$$

то $L_n(x) = b_{n+1}$. Це рекурентне співвідношення легко програмується. Обчислення $L_n(x)$ для кожного x за схемою (3.8) потребує n множень і $2n$ додавань та віднімань у той час, коли для обчислення значення многочлена Лагранжа, потрібна кількість арифметичних дій порядку $O(n^2)$. Однак, запис інтерполяційного многочлена у формі Лагранжа, як правило, призводить до меншої величини обчислювальної похибки.

Для спрощення обчислень інтерполяційного многочлена зручно використовувати так звану *схему Ейткена*.

Нехай $L_{(k,k+1,\dots,l)}(x)$ — інтерполяційний многочлен степеня $l - k$ з вузлами інтерполяції x_k, x_{k+1}, \dots, x_l , зокрема $L_{(k)}(x) = f(x_k)$. Справджується рівність:

$$L_{(k,k+1,\dots,l+1)}(x) = L_{(k+1,\dots,l+1)}(x) + \frac{(L_{(k+1,\dots,l+1)}(x) - L_{(k,\dots,l)}(x))(x - x_{l+1})}{x_{l+1} - x_k}. \quad (3.9)$$

Дійсно, права частина (3.9) є многочленом степеня $l - k + 1$ і збігається з $f(x)$ в точках x_k, \dots, x_{l+1} . Схема Ейткена обчислення значення $L_{(0,1,\dots,n)}(x)$ полягає в послідовному обчисленні за формулою (3.9) елементів таблиці значень інтерполяційних многочленів

$$\begin{array}{ccccccc} L_{(0)}(x) & & & & & & \\ L_{(1)}(x) & L_{(0,1)}(x) & & & & & \\ L_{(2)}(x) & L_{(1,2)}(x) & L_{(0,1,2)}(x) & & & & \\ & & \dots & \dots & L_{(0,1,\dots,n)}(x) & & \\ \dots & \dots & & & & & \\ L_{(n)}(x) & L_{(n-1,n)}(x) & L_{(n-2,n-1,n)}(x) & \dots & & & \end{array}$$

Нехай вузли x_i розташовані на однакових віддальх: $x_i = x_0 + ih$, де h — величина кроку таблиці. Тоді використовують скінченні різниці, які визначаються так:

1) скінченні різниці першого порядку

$$\Delta f_i = \nabla f_{i+1} = f(x_{i+1}) - f(x_i), \quad i = \overline{0, n-1},$$

Δf_i — різниця вперед, а ∇f_{i+1} — різниця назад;

2) скінченні різниці k -го порядку

$$\Delta^k f_i = \Delta^{k-1} f_{i+1} - \Delta^{k-1} f_i, \quad \nabla^k f_i = \nabla^{k-1} f_i - \nabla^{k-1} f_{i-1}.$$

Очевидно, що

$$\Delta^k f_i = \nabla^k f_{i+k} = k! h^k f(x_i; \dots; x_{i+k}).$$

У формулі (3.7) зробимо заміну $x = x_0 + sh$ і перейдемо від розділених різниць до скінченних, тоді одержимо:

$$\begin{aligned} L_n(x_0 + sh) = f(x_0) + \frac{s}{1!} \Delta f_0 + \frac{s(s-1)}{2!} \Delta^2 f_0 + \\ + \dots + \frac{s(s-1) \cdots (s-n+1)}{n!} \Delta^n f_0. \end{aligned} \quad (3.10)$$

Формулу (3.10) називають інтерполяційною формулою Ньютона для *інтерполювання вперед*. Її зручно використовувати при інтерполюванні на початку таблиці.

Якщо у формулі (3.7) інтерполяційні вузли перенумерувати у порядку x_n, x_{n-1}, \dots, x_0 і зробити заміну $x = x_n + sh$, то одержимо інтерполяційний многочлен

$$\begin{aligned} L_n(x_n + sh) = f(x_n) + \frac{s}{1!} \nabla f_n + \frac{s(s+1)}{2!} \nabla^2 f_n + \\ + \dots + \frac{s(s+1) \cdots (s+n-1)}{n!} \nabla^n f_n, \end{aligned} \quad (3.11)$$

який називається інтерполяційним многочленом Ньютона для *інтерполювання назад* і використовується при інтерполюванні в кінці таблиці.

Приклад 3.1. Побудуйте інтерполяційний многочлен Ньютона для функції заданої таблицею

Табл. 3.1.

x	0	1	2	3	5
$f(x)$	1	0	2	1	4

▷ Побудуємо таблицю розділених різниць

$$\begin{array}{cccccc} 0 & 1 & & & & \\ 1 & 0 & -1 & & & \\ 2 & 2 & 2 & 3/2 & & \\ 3 & 1 & -1 & -3/2 & -1 & \\ 5 & 4 & 3/2 & 5/6 & 7/12 & 19/60 \end{array} .$$

Отже, інтерполяційний многочлен Ньютона буде мати вигляд

$$L_4(x) = 1 - x + \frac{3}{2}x(x-1) - \\ - x(x-1)(x-2) + \frac{19}{60}x(x-1)(x-2)(x-3).$$

◀

3.4. Оцінка залишкового члена інтерполяційного многочлена

Проведемо дослідження похибки, яка виникає при заміні функції інтерполяційним многочленом. Нехай функція $f(x)$ визначена в $n+1$ вузлі інтерполяції $x_i \in [a, b]$, $i = \overline{0, n}$, а $L_n(x)$ — інтерполяційний многочлен. Залишковий член (похибка) інтерполяційного многочлена має вигляд:

$$R_n(x) = f(x) - L_n(x).$$

Очевидно, що у вузлах інтерполяції цей залишковий член дорівнює нулю. Припустимо, що функція $f(x)$ має $n+1$ неперервну похідну на відрізку $[a, b]$, тобто $f(x) \in C^{(n+1)}[a, b]$. Введемо допоміжну функцію

$$\varphi(t) = f(t) - L_n(t) - K\omega_{n+1}(t),$$

де $\omega_{n+1}(t) = (t - x_0) \dots (t - x_n)$, K — константа. Зауважимо, що $\varphi(t) \in C^{(n+1)}[a, b]$, $\varphi(x_i) = 0$, $i = \overline{0, n}$. Виберемо сталу K з умови $\varphi(x) = 0$, де x — точка, в якій оцінюється похибка. Для цього достатньо покласти

$$K = \frac{f(x) - L_n(x)}{\omega_{n+1}(x)}.$$

При такому виборі K функція $\varphi(t)$ перетворюється в нуль в $n+2$ -х точках x_0, \dots, x_n, x . На основі теореми Ролля її похідна має не менше $n+1$ коренів. Послідовно застосовуючи цю теорему до похідних вищого порядку функції $\varphi(t)$ одержимо, що $\varphi''(t)$ має не менше n коренів і.т.д., а функція $\varphi^{(n+1)}(t)$ має принаймні один корінь, тобто $\varphi^{(n+1)}(\xi) = 0$, де $\xi \in [a, b]$. Оскільки

$$\varphi^{(n+1)}(t) = f^{(n+1)}(t) - K(n+1)!,$$

то з умови $\varphi^{(n+1)}(\xi) = 0$ будемо мати:

$$K = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Отже, із того, що $\varphi(x) = 0$ випливає

$$R_n(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)\omega_{n+1}(x)}{(n+1)!}, \quad \xi \in [a, b]. \quad (3.12)$$

Покладаючи

$$M_{n+1} = \sup_{x \in [a, b]} |f^{(n+1)}(x)|,$$

отримаємо оцінку залишкового члена

$$|R_n| \leq \frac{M_{n+1}}{(n+1)!} |\omega_{n+1}(x)|.$$

Зауваження 3.1. Оскільки інтерполяційні многочлени Лагранжа і Ньютона відрізняються тільки формою запису, то формула (3.12) справджується як для інтерполяційного многочлена у формі Лагранжа так і у формі Ньютона.

Приклад 3.2. За таблицею значень функції $f(x) = 1/x$ в точках 2,70, 2,72, 2,74, користуючись лінійною інтерполяцією, знайдіть наближене значення $f(2,718)$ та оцініть залишковий член.

▷ Побудуємо таблицю розділених різниць.

x_i	$f(x_i)$	$f(x_i; x_j)$
2,70	0,3704	-0,14
2,72	0,3676	-0,13
2,74	0,3650	

Інтерполяційний многочлен Ньютона 1-го степеня (лінійна інтерполяція) має вигляд: $L_1(x) = f(x_0) + (x - x_0)f(x_0; x_1) = 0,3704 - 0,14(x - 2,7)$. Тоді $f(2,718) \approx 0,3679$. Для залишкового члена лінійної інтерполяції справджується оцінка

$$R_1(x) = \frac{M_2}{2!} |(x - x_0)(x - x_1)|, \quad M_2 = \max_{x_0 \leq x \leq x_1} |f''(x)|.$$

Враховуючи, що $f''(x) = \frac{2}{x^3}$, $\max_{2,70 \leq x \leq 2,72} |f''(x)| = \frac{2}{(2,7)^3}$ будемо мати

$$|R_1(2,718)| \leq \frac{|2,718 - 2,70| \cdot |2,718 - 2,72|}{(2,7)^3} < 0,2 \cdot 10^{-4}.$$



Приклад 3.3. З якою точністю можна обчислити $\sin 20^\circ$ за відомими значеннями $\{\sin 0^\circ; \sin 30^\circ; \sin 45^\circ\}$, використовуючи інтерполяцію: а) лінійну; б) квадратичну.

▷ Оскільки

$$f(x) = \sin x, \quad f'(x) = \cos x, \quad f''(x) = -\sin x,$$

$$\max_{x \in [0^\circ; 30^\circ]} |f''(x)| = \frac{1}{2}, \quad x_0 = 0, \quad x_1 = \frac{\pi}{6},$$

то для залишкового члена лінійної інтерполяції справджується оцінка

$$\left| R_1 \left(\frac{\pi}{9} \right) \right| \leq \frac{1}{2 \cdot 2} \left| \left(\frac{\pi}{9} - 0 \right) \left(\frac{\pi}{9} - \frac{\pi}{6} \right) \right| \approx 0,015.$$

У випадку квадратичної інтерполяції (інтерполяції многочленом 2-го степеня) з урахуванням $f'''(x) = -\cos x$, $\max_{x \in [0^\circ; 45^\circ]} |\cos x| = 1$, одержимо

$$\left| R_2 \left(\frac{\pi}{9} \right) \right| \leq \frac{1}{6} \left| \left(\frac{\pi}{9} - 0 \right) \left(\frac{\pi}{9} - \frac{\pi}{6} \right) \left(\frac{\pi}{9} - \frac{\pi}{4} \right) \right| \approx 0,004.$$

◀

3.5. Оптимальний вибір вузлів інтерполяції

Мінімізувати залишковий член інтерполяції можна лише за допомогою мінімізації величини $|\omega_{n+1}(x)|$, що можна досягти лише вибором вузлів інтерполяції. Ця задача розв'язується за допомогою многочленів Чебишева (див. розділ 1.3.5).

Нехай функція $f(x)$ наближається на $[a, b]$ за допомогою інтерполяційного многочлена степеня n з вузлами інтерполяції $x_0, x_1, \dots, x_n \in [a, b]$. Згідно з (3.12) маємо

$$R_n(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi) \omega_{n+1}(x)}{(n+1)!},$$

де $\xi \in [a, b]$, якщо $x \in [a, b]$. Звідси випливає оцінка похибки інтерполяції

$$\|R_n\| \leq \frac{\|f^{(n+1)}\| \cdot \|\omega_{n+1}\|}{(n+1)!},$$

де $\|g\| = \max_{x \in [a, b]} |g(x)|$. Многочлен $\omega_{n+1}(x) = \prod_{i=0}^n (x - x_i)$ має старший коефіцієнт 1 і тому за лемою 1.1

$$\|\omega_{n+1}\| = \max_{x \in [a, b]} |\omega_{n+1}(x)| \geq \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

Треба так вибрати корені цього многочлена на відрізок $[a, b]$, щоб мінімізувати величину $\max_{x \in [a, b]} |\omega_{n+1}(x)|$. Для цього достатньо, щоб многочлен $\omega_{n+1}(x)$ збігався із многочленом Чебишева на відрізок $[a, b]$

$$T_{n+1}^*(x) = \frac{(b-a)^{n+1}}{2^{2n+1}} \cos \left((n+1) \arccos \frac{2x - (b+a)}{b-a} \right).$$

Якщо за вузли інтерполяції взяти

$$x_k = \frac{b+a}{2} + \frac{b-a}{2} \cos \frac{\pi(2k+1)}{2(n+1)}, \quad k = \overline{0, n},$$

TO

$$\max_{x \in [a, b]} |\omega_{n+1}(x)| = \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

При такому виборі вузлів справджується найкраща оцінка

$$\|R_n\| \leq \frac{M_{n+1} (b-a)^{n+1}}{2^{2n+1} (n+1)!}.$$

3.6. Розділені різниці та інтерполювання з кратними вузлами

Нехай потрібно побудувати многочлен $P_{s-1}(x)$ степеня $s - 1$, який задовольняє умови:

[illegible]

де всі x_i , $i = \overline{0, n}$ — різні, $s = m_0 + \dots + m_n$. Такий многочлен називають *інтерполяційним многочленом з кратними вузлами*, або *многочленом Ерміта*, а числа m_0, \dots, m_n — кратностями вузлів x_0, \dots, x_n відповідно.

► **ТЕОРЕМА 3.1.** *Інтерполяційний многочлен, який задовольняє умови (3.13), єдиний.*

Доведення. Припустимо від супротивного, що є два многочлени степеня $s-1$, які задовольняють (3.13). Тоді різниця їх $Q_{s-1}(x)$ має властивості:

$$Q_{s-1}(x_0) = \dots = Q_{s-1}^{(m_0-1)}(x_0) = 0,$$

.....,

$$Q_{s-1}(x_n) = \dots = Q_{s-1}^{(m_n-1)}(x_n) = 0,$$

тобто точки x_0, \dots, x_n є нулями многочлена $Q_{s-1}(x)$ кратності m_0, \dots, m_n відповідно. Це означає, що многочлен $Q_{s-1}(x)$ степеня $s-1$ має s нулів, отже, $Q_{s-1}(x) \equiv 0$. ■

Далі припускатимемо, що $f(x)$ неперервно диференційована s разів. Побудуємо явно многочлен $P_{s-1}(x)$, доводячи тим самим його існування.

Розглянемо послідовність точок $x_{ij}^\varepsilon, i = \overline{0, n}, j = \overline{1, m_i}$, причому всі точки x_{ij}^ε різні, $x_i \leq x_{i1}^\varepsilon < x_{i2}^\varepsilon < \dots < x_{i, m_i}^\varepsilon < x_{i+1}, x_{ij}^\varepsilon \rightarrow x_i$ при $\varepsilon \rightarrow 0$. Зокрема, можна взяти $x_{ij}^\varepsilon = x_i + (j-1)\varepsilon$. Побудуємо інтерполяційний многочлен $P_{s-1}^\varepsilon(x)$ степеня $s-1$, який збігається з $f(x)$ у точках x_{ij}^ε , використовуючи при цьому таблицю розділених різниць

$$\begin{array}{ccccccc} f(x_{01}^\varepsilon) & & & & & & \\ f(x_{02}^\varepsilon) & f(x_{01}^\varepsilon; x_{02}^\varepsilon) & & & & & \\ f(x_{03}^\varepsilon) & f(x_{02}^\varepsilon; x_{03}^\varepsilon) & f(x_{01}^\varepsilon; x_{02}^\varepsilon; x_{03}^\varepsilon) & \dots & & & \\ \vdots & \vdots & & & & & \\ f(x_{0m_0}^\varepsilon) & \vdots & & & \dots & \ddots & f(x_{01}^\varepsilon; x_{02}^\varepsilon; \dots; x_{nm_n}^\varepsilon). \\ f(x_{11}^\varepsilon) & f(x_{0m_0}^\varepsilon; x_{11}^\varepsilon) & & \dots & & \ddots & \\ \vdots & & & \dots & & & \\ f(x_{nm_n}^\varepsilon) & \dots & & & & & \end{array} \quad (3.14)$$

Запишемо інтерполяційну формулу Ньютона з розділеними різницями

$$P_{s-1}^\varepsilon(x) = A_0^\varepsilon + A_1^\varepsilon(x - x_{01}^\varepsilon) + A_2^\varepsilon(x - x_{01}^\varepsilon)(x - x_{02}^\varepsilon) + \dots + A_{s-1}^\varepsilon(x - x_{01}^\varepsilon) \dots (x - x_{n, m_n-1}^\varepsilon),$$

де

$$A_0^\varepsilon = f(x_{01}^\varepsilon), \quad A_1^\varepsilon = f(x_{01}^\varepsilon; x_{02}^\varepsilon), \quad A_2^\varepsilon = f(x_{01}^\varepsilon; x_{02}^\varepsilon; x_{03}^\varepsilon), \dots,$$

$$A_{s-1}^\varepsilon = f(x_{01}^\varepsilon; x_{02}^\varepsilon; \dots; x_{n, m_n}^\varepsilon).$$

Виразимо розділені різниці через похідні від функції f . Згідно з формулою (3.6), справджується рівність

$$\begin{aligned} R_n(x) &= f(x) - L_n(x) = f(x) - \sum_{i=0}^n f(x_i) \prod_{j \neq i} \frac{x - x_j}{x_i - x_j} = \\ &= \left(\frac{f(x)}{\prod_{i=0}^n (x - x_i)} + \sum_{i=0}^n \frac{f(x_i)}{(x_i - x) \prod_{j \neq i} (x_i - x_j)} \right) \prod_{i=0}^n (x - x_i) \\ &= f(x; x_0; \dots; x_n) \omega_{n+1}(x). \end{aligned}$$

З іншого боку

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega_{n+1}(x), \quad a \leq \xi \leq b.$$

Отже,

$$f(x; x_0; \dots; x_n) = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Тоді

$$f(x_{il}^\varepsilon; \dots; x_{im}^\varepsilon) = \frac{f^{(m-l)}(x_{ilm}^\varepsilon)}{(m-l)!}, \quad (3.15)$$

де x_{ilm}^ε міститься в найменшому проміжку, що містить всі точки $x_{il}^\varepsilon, \dots, x_{im}^\varepsilon$. Перейшовши до границі при $\varepsilon \rightarrow 0$, одержимо

$$\lim_{\varepsilon \rightarrow 0} f(x_{il}^\varepsilon; \dots; x_{im}^\varepsilon) = \frac{f^{(m-l)}(x_i)}{(m-l)!}.$$

Це означає, що всі розділені різниці $f(x_{il}^\varepsilon; \dots; x_{im}^\varepsilon)$ при $\varepsilon \rightarrow 0$ мають границі, які природно позначити $f(\underbrace{x_i; \dots; x_i}_{m-l+1})$, причому з (3.15) випливає:

$$f(\underbrace{x_i; \dots; x_i}_{p+1}) = \frac{f^{(p)}(x_i)}{p!} \quad \forall p.$$

За індукцією по порядку різниці можна показати, що всі розділені різниці, які входять в таблицю розділених різниць, мають скінченні

границі. Після переходу до границі таблиця (3.14) переходить у таку:

$$\begin{array}{ccccccc}
 f(x_0) & & & & & & \\
 f(x_0) & f(x_0; x_0) & & & & & \\
 f(x_0) & f(x_0; x_0) & f(x_0; x_0; x_0) & \ddots & & & \\
 \vdots & \vdots & & \ddots & & & \\
 f(x_0) & \vdots & \ddots & \ddots & & & \\
 f(x_1) & f(x_0; x_1) & \ddots & & & & \\
 \vdots & & & & & & \\
 f(x_n) & \ddots & & & & &
 \end{array}
 \quad f(x_0; \dots; x_0; \dots; x_n; \dots; x_n) \quad (3.16)$$

А тому на будь-якому відрізку многочлени $P_{s-1}^\varepsilon(x)$ при $\varepsilon \rightarrow 0$ прямують до деякого многочлена.

$$\begin{aligned}
 P_{s-1}(x) = & A_0 + A_1(x - x_0) + A_2(x - x_0)^2 + \dots \\
 & + A_{s-1}(x - x_0)^{m_0} \dots (x - x_{n-1})^{m_{n-1}}(x - x_n)^{m_n-1} = f(x_0) + \\
 & + f(x_0; x_0)(x - x_0) + f(x_0; x_0; x_0)(x - x_0)^2 + \dots, \quad (3.17)
 \end{aligned}$$

де $A_i = \lim_{\varepsilon \rightarrow 0} A_i^\varepsilon$. Многочлен (3.17) можна записати у вигляді:

$$\begin{aligned}
 P_{s-1}(x) = & \sum_{i=0}^{m_0-1} \frac{f^{(i)}(x_0)}{i!} (x - x_0)^i + \\
 & + (x - x_0)^{m_0} F(x - x_1, x - x_2, \dots, x - x_n),
 \end{aligned}$$

де $F(t_1, \dots, t_n)$ — деякий многочлен від t_1, \dots, t_n . Звідси випливає, що $P_{s-1}(x)$ задовольняє умови, які задані в точці x_0 . В силу єдиності інтерполяційного многочлена він не змінюється при перепозначенні x_0 на x_j і x_j на x_0 для довільного j . Тому граничний многочлен буде задовольняти задані умови у будь-якій точці x_j і, таким чином, він є шуканим многочленом.

Приклад 3.4. Побудуйте інтерполяційний многочлен, який задовольняє умови:

x	$f(x)$	$f'(x)$	$f''(x)$
0	3	1	—
1	0	—	—
2	1	2	1
3	5	—	—

▷ Як впливає з таблиці вузол $x_0 = 0$ має кратність 2, $x_1 = 1$ — кратність 1, $x_2 = 2$ — кратність 3, $x_3 = 3$ — кратність 1. Отже, степінь шуканого многочлена буде 6. Складаємо таблицю розділених різниць вигляду (3.16):

0	3						
0	3	1					
1	0	-3	-4	3			
2	1	1	2	-1/2	-7/4	7/8	
2	1	2	1	-1/2	0	1/3	-13/72
2	1	2	1/2	3/2	1		
2	1	2	2				
3	5	4					

За таблицею записуємо інтерполяційний многочлен:

$$P_6(x) = 3 + 1 \cdot x - 4x^2 + 3x^2(x-1) - \frac{7}{4}x^2(x-1)(x-2) + \frac{7}{8}x^2(x-1)(x-2)^2 - \frac{13}{72}x^2(x-1)(x-2)^3.$$

◀

3.7. Найкраще наближення в лінійному нормованому просторі

Нехай задано елемент f лінійного нормованого простору H . Задача побудови *найкращого наближення* для f лінійною комбінацією

$$\sum_{i=1}^n a_i \varphi_i$$

заданих лінійно незалежних елементів $\varphi_1, \varphi_2, \dots, \varphi_n \in H$, полягає в тому, щоб знайти елемент

$$\sum_{i=1}^n a_i^0 \varphi_i$$

такий, що

$$\left\| f - \sum_{i=1}^n a_i^0 \varphi_i \right\| = \inf_{a_1, a_2, \dots, a_n} \left\| f - \sum_{i=1}^n a_i \varphi_i \right\|.$$

Якщо такий елемент існує, то він називається *елементом найкращого наближення*.

■► **ТЕОРЕМА 3.2.** *Елемент найкращого наближення існує.*

Доведення. Внаслідок співвідношення (нерівності трикутника для різниці)

$$\begin{aligned} \left| \left\| f - \sum_{i=1}^n a_i^1 \varphi_i \right\| - \left\| f - \sum_{i=1}^n a_i^2 \varphi_i \right\| \right| &\leq \left\| \sum_{i=1}^n (a_i^1 - a_i^2) \varphi_i \right\| \leq \\ &\leq \sum_{i=1}^n |a_i^1 - a_i^2| \|\varphi_i\| \end{aligned}$$

функція

$$F_f(a_1, a_2, \dots, a_n) = \left\| f - \sum_{i=1}^n a_i \varphi_i \right\|$$

є неперервною функцією аргументів a_i для $\forall f \in H$. Нехай $|a|$ — евклідова норма вектора $a = (a_1, a_2, \dots, a_n)$. Функція $F_0(a_1, a_2, \dots, a_n) = \|a_1 \varphi_1 + a_2 \varphi_2 + \dots + a_n \varphi_n\|$ неперервна на одиничній сфері $|a| = 1$, а тому в деякій її точці $(\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_n)$ досягає своєї нижньої грані \tilde{F} на сфері, причому $\tilde{F} \neq 0$, оскільки $\tilde{F} = \|\tilde{a}_1 \varphi_1 + \tilde{a}_2 \varphi_2 + \dots + \tilde{a}_n \varphi_n\| = 0$ суперечить лінійній незалежності елементів $\varphi_1, \varphi_2, \dots, \varphi_n$. Для будь-якого $a = (a_1, a_2, \dots, a_n) \neq (0, 0, \dots, 0)$ справджується оцінка

$$\begin{aligned} \|a_1 \varphi_1 + a_2 \varphi_2 + \dots + a_n \varphi_n\| &= F_0(a_1, a_2, \dots, a_n) = \\ &= |a| F_0\left(\frac{a_1}{|a|}, \frac{a_2}{|a|}, \dots, \frac{a_n}{|a|}\right) \geq |a| \tilde{F}. \end{aligned}$$

Нехай $\gamma > 2\|f\|/\tilde{F}$. Функція $F_f(a_1, a_2, \dots, a_n)$ неперервна в кулі $|a| \leq \gamma$. Отже, в деякій точці кулі $(a_1^0, a_2^0, \dots, a_n^0)$ вона досягає своєї нижньої грані F^* на кулі. Тоді $F^* \leq F_f(0, 0, \dots, 0) = \|f\|$. Зовні кулі виконуються співвідношення

$$\begin{aligned} F_f(a_1, a_2, \dots, a_n) &\geq \|a_1 \varphi_1 + a_2 \varphi_2 + \dots + a_n \varphi_n\| - \|f\| \geq \\ &\geq |a| \tilde{F} - \|f\| \geq \gamma \tilde{F} - \|f\| > \\ &> 2\tilde{F} \|f\| / \tilde{F} - \|f\| = \|f\| > F^*. \end{aligned}$$

Таким чином,

$$F_f(a_1, a_2, \dots, a_n) \geq F^* = F_f(a_1^0, a_2^0, \dots, a_n^0) \quad \forall a_1, a_2, \dots, a_n. \quad \blacksquare$$

Елементів найкращого наближення, взагалі кажучи, може бути декілька.

3.8. Найкраще наближення в гільбертовому просторі

Нехай H — гільбертовий простір зі скалярним добутком (u, v) та нормою $\|u\| = (u, u)^{1/2}$. Для гільбертового простору елемент найкращого наближення єдиний.

Розглянемо функцію

$$\begin{aligned}\Phi(a_1, a_2, \dots, a_n) &= \left\| f - \sum_{i=1}^n a_i \varphi_i \right\|^2 = \left(f - \sum_{i=1}^n a_i \varphi_i, f - \sum_{j=1}^n a_j \varphi_j \right) = \\ &= (f, f) - 2 \sum_{i=1}^n a_i (f, \varphi_i) + \sum_{i=1}^n \sum_{j=1}^n a_i a_j (\varphi_i, \varphi_j).\end{aligned}$$

Коефіцієнти a_i елемента найкращого наближення знаходяться з умови

$$\inf_{a_1, a_2, \dots, a_n} \Phi(a_1, a_2, \dots, a_n).$$

В точці мінімуму повинні виконуватись умови $\partial\Phi/\partial a_i = 0$. Маємо

$$\frac{\partial\Phi}{\partial a_j} = -2(f, \varphi_j) + 2 \sum_{i=1}^n (\varphi_i, \varphi_j) a_i = 0, \quad j = \overline{1, n}.$$

Отже, для знаходження коефіцієнтів a_i отримаємо систему лінійних алгебраїчних рівнянь

$$\sum_{i=1}^n \alpha_{ij} a_i = (f, \varphi_j), \quad j = \overline{1, n}, \quad (3.18)$$

де

$$\alpha_{ij} = (\varphi_i, \varphi_j).$$

Матриця $A = (\alpha_{ij})_{i,j=1}^n$ називається матрицею Грама системи елементів $\varphi_1, \varphi_2, \dots, \varphi_n$. Оскільки $\alpha_{ij} = \alpha_{ji}$, то матриця A симетрична.

⇒ **Лема 3.2.** Якщо елементи $\varphi_1, \varphi_2, \dots, \varphi_n$ лінійно незалежні, то матриця A додатно визначена.

Доведення. Нехай $c = (c_1, c_2, \dots, c_n)$ — довільний вектор з дійсними компонентами. Справджується рівність

$$\begin{aligned}\left\| \sum_{i=1}^n c_i \varphi_i \right\|^2 &= \left(\sum_{i=1}^n c_i \varphi_i, \sum_{j=1}^n c_j \varphi_j \right) = \\ &= \sum_{i=1}^n \sum_{j=1}^n c_i c_j (\varphi_i, \varphi_j) = (Ac, c).\end{aligned} \quad (3.19)$$

Звідси

$$(Ac, c) = \left\| \sum_{i=1}^n c_i \varphi_i \right\|^2 \geq 0.$$

Якщо елементи φ_i лінійно незалежні, то $\left\| \sum_{i=1}^n c_i \varphi_i \right\| = 0$ тільки в тому випадку, коли всі $c_i = 0$. Отже, $(Ac, c) > 0$, якщо $c \neq 0$, тобто матриця A додатно визначена. Визначник додатно визначеної матриці відмінний від нуля, а тому система (3.18) має єдиний розв'язок.

Систему функцій $\{\varphi_i\}$ доцільно вибирати ортонормованою системою або близькою до неї. Якщо $\{\varphi_i\}$ утворюють ортонормовану систему $(\varphi_i, \varphi_j) = \delta_{ij}$ (δ_{ij} — символ Кронекера), то система рівнянь (3.18) буде мати вигляд

$$a_i = (f, \varphi_i).$$

Тоді найкраще наближення записується в вигляді

$$\varphi = \sum_{i=1}^n (f, \varphi_i) \varphi_i$$

і для величини $\|f - \varphi\|^2$ справджується рівність:

$$\begin{aligned} \|f - \varphi\|^2 &= \left(f - \sum_{i=1}^n a_i \varphi_i, f - \sum_{i=1}^n a_i \varphi_i \right) = \\ &= (f, f) - 2 \sum_{i=1}^n a_i (f, \varphi_i) + \sum_{i=1}^n |a_i|^2 = (f, f) - \sum_{i=1}^n |(f, \varphi_i)|^2. \end{aligned}$$

Оскільки $\|f - \varphi\|^2 \geq 0$, то з рівності

$$\|f - \varphi\|^2 = (f, f) - \sum_{i=1}^n |(f, \varphi_i)|^2$$

впливає, зокрема, відома нерівність Бесселя

$$(f, f) \geq \sum_{i=1}^n |(f, \varphi_i)|^2.$$

■

Якщо вихідні елементи не утворюють ортонормованої системи, то, взагалі кажучи, їх можна ортогоналізувати за допомогою процесу ортогоналізації Гільберта–Шмідта [21].

знаходимо точки, в яких може бути екстремум. Вибравши той розв'язок, який належить області зміни параметрів a_1, \dots, a_m і в якому функція $S(a_1, \dots, a_m)$ має абсолютний мінімум, знаходимо незалежні значення a_1, \dots, a_m .

Якщо $f(x, a_1, \dots, a_m)$ лінійно залежить від параметрів a_1, \dots, a_m , тобто

$$f(x, a_1, \dots, a_m) = \sum_{j=1}^m f_j(x) a_j,$$

то система (3.21) набуває вигляду

$$y_i = \sum_{j=1}^m f_j(x_i) a_j, \quad i = \overline{1, n}. \quad (3.22)$$

Метод найменших квадратів розв'язування системи (3.22) полягає у тому, щоб визначити невідомі, які мінімізують суму квадратів нев'язок, тобто суму вигляду

$$S(a_1, \dots, a_m) = \sum_{i=1}^n \left[y_i - \sum_{j=1}^m f_j(x_i) a_j \right]^2.$$

З умови мінімуму величини S як функції від a_1, \dots, a_m отримаємо систему лінійних алгебраїчних рівнянь

$$\frac{\partial S}{\partial a_k} = -2 \sum_{i=1}^n \left[y_i - \sum_{j=1}^m f_j(x_i) a_j \right] f_k(x_i) = 0, \quad k = \overline{1, m}$$

або

$$\sum_{i=1}^n \left[\sum_{j=1}^m f_j(x_i) a_j \right] f_k(x_i) = \sum_{i=1}^n f_k(x_i) y_i, \quad k = \overline{1, m}. \quad (3.23)$$

Розв'язок системи m лінійних алгебраїчних рівнянь (3.23) з m невідомими вважаємо наближеним розв'язком системи (3.22).

Приклад 3.5. Методом найменших квадратів для функції заданої таблицею побудуйте лінійний і квадратичний многочлени.

x_i	0	1/4	1/2	3/4	1
y_i	1	2	1	0	1

▷ Для наближення функції використаємо лінійний многочлен

$$\bar{y}(x) = a_1 + a_2x.$$

Тоді

$$S(a_1, a_2) = \sum_{i=1}^n (y_i - a_1 - a_2x_i)^2.$$

Необхідна умова мінімуму функції S — виконання співвідношень

$$\begin{aligned} \frac{\partial S}{\partial a_1} &= -2 \sum_{i=1}^n (y_i - a_1 - a_2x_i) = 0, \\ \frac{\partial S}{\partial a_2} &= -2 \sum_{i=1}^n x_i(y_i - a_1 - a_2x_i) = 0. \end{aligned}$$

Згрупувавши разом коефіцієнти при a_1 і a_2 , отримаємо систему двох лінійних рівнянь

$$\begin{cases} n \cdot a_1 + \sum_{i=1}^n x_i \cdot a_2 = \sum_{i=1}^n y_i, \\ \sum_{i=1}^n x_i \cdot a_1 + \sum_{i=1}^n x_i^2 \cdot a_2 = \sum_{i=1}^n y_i x_i. \end{cases}$$

Розв'язок цієї системи можна знайти за формулами Крамера

$$\begin{aligned} a_1 &= \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{\frac{15}{8} \cdot 5 - 2 \cdot \frac{5}{2}}{5 \cdot \frac{15}{8} - \frac{25}{4}} = \frac{7}{5}; \\ a_2 &= \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{5 \cdot 2 - 5 \cdot \frac{5}{2}}{5 \cdot \frac{15}{8} - \frac{25}{4}} = -\frac{4}{5}. \end{aligned}$$

У випадку квадратичного многочлена необхідно знайти мінімум функції

$$S(a_1, a_2, a_3) = \sum_{i=1}^n (y_i - a_1 - a_2x_i - a_3x_i^2)^2.$$

Тоді

$$\begin{aligned} \frac{\partial S}{\partial a_1} &= -2 \sum_{i=1}^n (y_i - a_1 - a_2x_i - a_3x_i^2) = 0, \\ \frac{\partial S}{\partial a_2} &= -2 \sum_{i=1}^n x_i(y_i - a_1 - a_2x_i - a_3x_i^2) = 0, \\ \frac{\partial S}{\partial a_3} &= -2 \sum_{i=1}^n x_i^2(y_i - a_1 - a_2x_i - a_3x_i^2) = 0. \end{aligned}$$

Цю систему запишемо у вигляді

$$\begin{cases} n \cdot a_1 + \sum_{i=1}^n x_i \cdot a_2 + \sum_{i=1}^n x_i^2 \cdot a_3 = \sum_{i=1}^n y_i, \\ \sum_{i=1}^n x_i \cdot a_1 + \sum_{i=1}^n x_i^2 \cdot a_2 + \sum_{i=1}^n x_i^3 \cdot a_3 = \sum_{i=1}^n y_i x_i, \\ \sum_{i=1}^n x_i^2 \cdot a_1 + \sum_{i=1}^n x_i^3 \cdot a_2 + \sum_{i=1}^n x_i^4 \cdot a_3 = \sum_{i=1}^n y_i x_i^2. \end{cases}$$

Розв'язавши її за наших даних, знайдемо

$$a_1 = 7/5, \quad a_2 = -4/5, \quad a_3 = 0.$$

Отже, квадратичний многочлен у даному випадку не дає ніякого покращення у порівнянні з лінійною інтерполяцією. ◀

3.10. Інтерполяція сплайнами

Інтерполяція многочленом Лагранжа або Ньютона на всьому відрізку $[a, b]$ з використанням великої кількості вузлів інтерполяції часто призводить до поганого наближення, що пояснюється сильним нагромадженням похибок в процесі обчислення. Крім того, через розбіжність процесу інтерполяції збільшення кількості вузлів не повинно призводити до підвищення точності. Для того, щоб запобігти великим похибкам, весь відрізок $[a, b]$ розбивають на окремі відрізки і на кожному з них наближено заміняють функцію $f(x)$ многочленом невисокого степеня (так звана кусково-поліноміальна інтерполяція).

Одним із способів інтерполювання на всьому відрізку є інтерполяція за допомогою сплайнів. *Сплайном (сплайн-функцією)* називають кусково-поліноміальну функцію, визначену на відрізку $[a, b]$ і таку, що має на цьому відрізку деяку кількість неперервних похідних. Найпоширеніші в інженерних розрахунках сплайни складені з многочленів третього степеня (кубічні сплайни).

Нехай на відрізку $[a, b]$ задано сітку $\omega = \{x_i : x_0 = a < x_1 < \dots < x_n = b\}$ у вузлах якої задано значення $\{f_i\}_{i=0}^n$ функції $f(x)$, визначеної на $[a, b]$. Задача кусково-кубічної інтерполяції ставиться таким чином: знайти функцію $s(x)$, яку називатимемо *кубічним сплайном*, визначену на $[a, b]$ і таку, що

1) функція $s(x)$, а також перша і друга похідні неперервні на $[a, b]$, тобто

$$s(x) \in C^{(2)}[a, b]; \quad (3.24)$$

2) на кожному з відрізків $[x_{i-1}, x_i]$ функція $s(x)$ є кубічним многочленом вигляду

$$s(x) \equiv s_i(x) = \sum_{k=0}^3 a_k^{(i)} (x - x_i)^k, \quad i = \overline{1, n}; \quad (3.25)$$

3) у вузлах сітки ω виконуються рівності

$$s(x_i) = f_i, \quad i = \overline{0, n}; \quad (3.26)$$

4) $s''(x)$ задовольняє граничні умови

$$s''(a) = s''(b) = 0. \quad (3.27)$$

Покажемо, що поставлена задача має єдиний розв'язок і вкажемо алгоритм його обчислення.

Оскільки, згідно з (3.25), $s(x)$ — кубічний многочлен, то $s''(x)$ — лінійна функція. Тому з формули лінійної інтерполяції для $x \in [x_{i-1}, x_i]$, $i = \overline{1, n}$, маємо

$$\begin{aligned} s''(x) &= m_{i-1} + (x - x_{i-1}) \frac{m_i - m_{i-1}}{h_i} = \\ &= m_{i-1} \frac{x_i - x}{h_i} + m_i \frac{x - x_{i-1}}{h_i}, \end{aligned} \quad (3.28)$$

де $h_i = x_i - x_{i-1}$, $m_i = s''(x_i)$. Зауважимо, що $s''(x)$ неперервна на $[a, b]$ функція: $s''(x_i - 0) = s''(x_i + 0) = m_i$. Два рази інтегруючи обидві частини рівності (3.28), одержимо

$$\begin{aligned} s'(x) &= -m_{i-1} \frac{(x_i - x)^2}{2h_i} + m_i \frac{(x - x_{i-1})^2}{2h_i} + C_i, \\ s(x) &= m_{i-1} \frac{(x_i - x)^3}{6h_i} + m_i \frac{(x - x_{i-1})^3}{6h_i} + \\ &+ A_i \frac{x_i - x}{h_i} + B_i \frac{x - x_{i-1}}{h_i}, \end{aligned} \quad (3.29)$$

де C_i, A_i, B_i — деякі сталі інтегрування. Знайдемо A_i, B_i з умов $s(x_{i-1}) = f_{i-1}$, $s(x_i) = f_i$ (див. (3.26)). Підставляючи в (3.29) $x = x_i$ і $x = x_{i-1}$, маємо

$$m_i \frac{h_i^2}{6} + B_i = f_i, \quad m_{i-1} \frac{h_i^2}{6} + A_i = f_{i-1} \quad (3.30)$$

звідки $B_i = f_i - m_i h_i^2/6$, $A_i = f_{i-1} - m_{i-1} h_i^2/6$. Підставивши ці значення в (3.29) знайдемо $s(x)$ для $x \in [x_{i-1}, x_i]$:

$$s(x) = m_{i-1} \frac{(x_i - x)^3}{6h_i} + m_i \frac{(x - x_{i-1})^3}{6h_i} + \left(f_{i-1} - \frac{m_{i-1} h_i^2}{6} \right) \frac{x_i - x}{h_i} + \left(f_i - \frac{m_i h_i^2}{6} \right) \frac{x - x_{i-1}}{h_i}. \quad (3.31)$$

Тоді

$$s'(x) = -m_{i-1} \frac{(x_i - x)^2}{2h_i} + m_i \frac{(x - x_{i-1})^2}{2h_i} + \frac{f_i - f_{i-1}}{h_i} - \frac{m_i - m_{i-1}}{6} h_i. \quad (3.32)$$

Із виразу (3.32) випливає рівність

$$s'(x_i - 0) = \frac{h_i}{6} m_{i-1} + \frac{h_i}{3} m_i + \frac{f_i - f_{i-1}}{h_i}. \quad (3.33)$$

Записавши (3.32) для відрізка $[x_i, x_{i+1}]$, тобто

$$s'(x) = -m_i \frac{(x_{i+1} - x)^2}{2h_{i+1}} + m_{i+1} \frac{(x - x_i)^2}{2h_{i+1}} + \frac{f_{i+1} - f_i}{h_{i+1}} - \frac{m_{i+1} - m_i}{6} h_{i+1},$$

знайдемо

$$s'(x_i + 0) = -\frac{h_{i+1}}{3} m_i - \frac{h_{i+1}}{6} m_{i+1} + \frac{f_{i+1} - f_i}{h_{i+1}}. \quad (3.34)$$

За умовою (3.24) функції $s'(x)$ і $s''(x)$ мають бути неперервними на $[a, b]$. Враховуючи (3.33), (3.34), з умови неперервності $s'(x)$ в точках x_i , $i = \overline{1, n-1}$, одержимо $n-1$ рівняння

$$\begin{aligned} \frac{h_i}{6} m_{i-1} + \frac{h_i + h_{i+1}}{3} m_i + \frac{h_{i+1}}{6} m_{i+1} &= \\ &= \frac{f_{i-1}}{h_i} - \left(\frac{1}{h_i} + \frac{1}{h_{i+1}} \right) f_i + \frac{f_{i+1}}{h_{i+1}}, \quad i = \overline{1, n-1}. \end{aligned} \quad (3.35)$$

Доповнимо (3.35) рівностями

$$m_0 = m_n = 0, \quad (3.36)$$

які впливають з (3.27).

Систему (3.35), (3.36) можна розв'язати методом прогонки (див. розділ 1.1.2), матриця цієї системи є матрицею із строго діагональною перевагою оскільки

$$\frac{h_i + h_{i+1}}{3} > \frac{h_i}{6} + \frac{h_{i+1}}{6},$$

а тому ця система має єдиний розв'язок і метод прогонки стійкий.

Отже, доведено, що існує єдиний кубічний сплайн, визначений умовами (3.24) — (3.27).

Приклад 3.6. Побудуйте інтерполяційний кубічний сплайн для функції, заданої таблицею

x_i	0	1/4	1/2	3/4	1
$f(x_i)$	1	2	1	0	1

▷ За формулою $h_i = x_i - x_{i-1}$, $i = \overline{1, 4}$ знаходимо, $h_1 = h_2 = h_3 = h_4 = 1/4$. Тоді система (3.35), (3.36) буде мати вигляд:

$$\begin{cases} \frac{1}{6}m_1 + \frac{1}{24}m_2 = -8, \\ \frac{1}{24}m_1 + \frac{1}{6}m_2 + \frac{1}{24}m_3 = 0, \\ \frac{1}{24}m_2 + \frac{1}{6}m_3 = 8, \end{cases}$$

$$m_0 = m_4 = 0.$$

Розв'язавши цю систему, отримаємо

$$m_1 = -48, \quad m_2 = 0, \quad m_3 = 48.$$

Тепер згідно з (3.31), запишемо кубічний сплайн

$$s(x) = \begin{cases} -32x^3 - 4\left(x - \frac{1}{4}\right) + 10x, & 0 \leq x \leq \frac{1}{4}, \\ 32\left(x - \frac{1}{2}\right)^3 - 10\left(x - \frac{1}{2}\right) + 4\left(x - \frac{1}{4}\right), & \frac{1}{4} \leq x \leq \frac{1}{2}, \\ 32\left(x - \frac{1}{2}\right)^3 - 4\left(x - \frac{3}{4}\right) - 2\left(x - \frac{1}{2}\right), & \frac{1}{2} \leq x \leq \frac{3}{4}, \\ -32(x-1)^3 + 2(x-1) + 4\left(x - \frac{3}{4}\right), & \frac{3}{4} \leq x \leq 1. \end{cases}$$



3.11. Чисельне диференціювання

Задача чисельного диференціювання формулюється так: за заданими значеннями функції $f(x)$ в точках $x_i, i = \overline{0, n}$ і заданими x та k знайти наближене значення $f^{(k)}(x)$, $k \geq 1$ і оцінити похибку.

Найпростіші формули чисельного диференціювання дістають за допомогою диференціювання інтерполяційних многочленів $L_n(x)$.

Зауважимо, що задача чисельного диференціювання *не є коректною* в $C[a, b]$, бо немає неперервної залежності норми похідної від норми функції.

Зобразимо функцію $f(x)$ через інтерполяційний многочлен Ньютона

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0)f(x_0; x_1) + \dots + \\ &\quad + (x - x_0)(x - x_1) \dots (x - x_{n-1})f(x_0; x_1; \dots; x_n) + \\ &\quad + (x - x_0)(x - x_1) \dots (x - x_n)f(x; x_0; \dots; x_n) = \\ &= L_n(x) + R_n(x). \end{aligned}$$

Звідси одержуємо таке співвідношення для похідної k -го порядку

$$f^{(k)}(x) = L_n^{(k)}(x) + R_n^{(k)}(x),$$

де похідна від залишкового члена за допомогою формули Лейбніца зображується так

$$R_n^{(k)}(x) = \sum_{i=0}^k C_k^i f^{(i)}(x; x_0; \dots; x_n) \omega_{n+1}^{(k-i)}(x), \quad (3.37)$$

$$C_k^i = \frac{k!}{i!(k-i)!}.$$

За означенням розділеної різниці з кратними вузлами

$$f^{(q)}(x; x_0; \dots; x_n) = f(\underbrace{x; \dots; x}_{q+1}; x_0; \dots; x_n) q!.$$


Отже, співвідношення (3.37) можна записати у вигляді


$$R_n^{(k)}(x) = \sum_{i=0}^k C_k^i i! f(x; \dots; x; x_0; \dots; x_n) \omega_{n+1}^{(k-i)}(x). \quad (3.38)$$

Виражаючи розділену різницю через похідну, дістаємо оцінку


$$\begin{aligned} |R_n^{(k)}(x)| &\leq \\ &\leq \sum_{i=0}^k \frac{k!}{(k-i)!(n+i+1)!} \max_{\xi \in [y_1, y_2]} |f^{(n+i+1)}(\xi)| \cdot \left| \omega_{n+1}^{(k-i)}(x) \right|, \\ y_1 &= \min(x, x_0, \dots, x_n), \quad y_2 = \max(x, x_0, \dots, x_n). \end{aligned}$$


Контрольні завдання


 **3.1.** Побудуйте многочлен другого степеня $L_2(x)$, який задовольняє умови $L_2(0) = 0$, $L_2(1) = 1$, $L_2(2) = 0$, використовуючи інтерполяційний многочлен Лагранжа та Ньютона, і переконатися, що ці многочлени однакові.

 **3.2.** Побудуйте інтерполяційний многочлен Ньютона для функції заданої таблицею

x_i	-2	-1	1	2	3
$f(x_i)$	26	5	3	-4	-7

 **3.3.** Нехай $f(x) = \sin(\pi x/2)$, а $L_2(x)$ — інтерполяційний многочлен другого степеня, значення якого в точках $x = 0, 1, 2$ дорівнюють відповідним значенням функції $f(x)$. Дайте оцінку залишковий член інтерполяційного многочлена. Порівняйте цю оцінку з фактичною похибкою в точках $x = 1/4$ і $x = 3/4$.


 **3.4.** З якою точністю можна обчислити $\sqrt{116}$ за відомими значеннями $\{\sqrt{100}; \sqrt{121}; \sqrt{144}\}$, використовуючи квадратичну інтерполяцію?

 **3.5.** Знайдіть найкраще наближення функції $f(x) = 2^x$ у просторі $L_2[-1, 1]$ многочленом третього степеня.

 **3.6.** Методом найменших квадратів для функції заданої таблицею

x_i	1	2	3	4
y_i	2	3	5	3

побудуйте лінійний і квадратичний многочлени.

 **3.7.** Побудуйте інтерполяційний кубічний сплайн для функції заданої таблицею

x_i	-15	-12	-6	0	6	12	15
$f(x_i)$	1	1	1	1	1	1	0

РОЗДІЛ 4

ЧИСЕЛЬНЕ ІНТЕГРУВАННЯ

4.1. Наближене обчислення інтегралів.

Інтерполяційні квадратурні формули

Нехай потрібно обчислити інтеграл

$$I = \int_a^b \rho(x) f(x) dx, \quad (4.1)$$

де $\rho(x) > 0$ — задана інтегровна на (a, b) вагова функція, $f(x)$ — задана достатньо гладка на $[a, b]$ функція. Для наближеного обчислення (4.1) будемо розглядати формули вигляду

$$I \approx \sum_{i=0}^n c_i^{(n)} f(x_i), \quad (4.2)$$

які називаються *квадратурними формулами*. Числа x_i , $i = \overline{0, n}$ називаються *вузлами квадратурної формули*, а числа $c_i^{(n)}$ — *коефіцієнтами*, або *ваговими коефіцієнтами*. Величина

$$R_n = I - \sum_{i=0}^n c_i^{(n)} f(x_i)$$

називається *залишковим членом*, або похибкою квадратурної формули.

Якщо залишковий член квадратурної формули дорівнює нулю для будь-якого многочлена не вище m -го степеня, то кажуть, що квадратурна формула має *алгебраїчну степінь точності m* .

Якщо функцію $f(x)$ на $[a, b]$ замінити інтерполяційним поліномом Лагранжа

$$L_n(x) = \sum_{i=0}^n f(x_i) \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j},$$

то одержимо квадратурну формулу *інтерполяційного типу*. У цьому випадку

$$c_i^{(n)} = \int_a^b \rho(x) \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j} dx. \quad (4.3)$$

Очевидно, що алгебраїчна степінь точності квадратурної формули інтерполяційного типу (4.2) з ваговими коефіцієнтами (4.3) є щонайменше n . Дійсно, якщо $f(x)$ — многочлен степеня n , то його можна записати у вигляді інтерполяційного многочлена Лагранжа, а тому $R_n = 0$.

Дамо оцінку похибки квадратурної формули інтерполяційного типу. Запишемо функцію $f(x)$ у вигляді $f(x) = L_n(x) + r_n(x)$, де $r_n(x)$ — похибка інтерполяції. Тоді

$$\begin{aligned} \int_a^b \rho(x) f(x) dx &= \int_a^b \rho(x) L_n(x) dx + \int_a^b \rho(x) r_n(x) dx \\ &= \sum_{i=0}^n c_i^{(n)} f(x_i) + \int_a^b \rho(x) r_n(x) dx. \end{aligned}$$

Отже, залишковий член R_n квадратурної формули інтерполяційного типу дорівнює

$$R_n = \int_a^b \rho(x) r_n(x) dx = \frac{1}{(n+1)!} \int_a^b \rho(x) \omega_{n+1}(x) f^{(n+1)}(\xi(x)) dx,$$

де $\omega_{n+1}(x) = (x - x_0) \dots (x - x_n)$. Звідси, якщо $|f^{(n+1)}(x)| \leq M_{n+1} \forall x \in [a, b]$, то для залишкового члена квадратурної форми інтерполяційного типу справджується оцінка

$$R_n \leq \frac{M_{n+1}}{(n+1)!} \int_a^b \rho(x) |\omega_{n+1}(x)| dx. \quad (4.4)$$

4.2. Квадратурні формули Ньютона–Котеса

Якщо у квадратурній формулі (4.2) з ваговими коефіцієнтами (4.3) для інтегралів (4.1) з $\rho(x) \equiv 1$ вузли рівновіддалені, тобто $x_{i+1} - x_i = h$, $i = 0, n-1$, $n = 1, 2, \dots$, то така формула називається *квадратурною формулою Ньютона–Котеса*. У формулах Ньютона–Котеса крок

$h = (b - a)/n$. Тоді квадратурна формула (4.2), (4.3) буде мати вигляд

$$I = \int_a^b f(x) dx \approx \sum_{i=0}^n c_i^{(n)} f(a + ih), \quad (4.5)$$

де

$$c_i^{(n)} = \int_a^b \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j} dx.$$

Зробимо заміну змінних $x = a + ht$. Тоді $x - x_j = h(t - j)$ і

$$\begin{aligned} \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j} &= \frac{(x - x_0) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)} = \\ &= \frac{t(t-1) \cdots (t-i+1)(t-i-1) \cdots (t-n)}{(-1)^{n-i} i! (n-i)!}. \end{aligned}$$

Отже,

$$\begin{aligned} c_i^{(n)} &= \frac{(-1)^{n-i} h}{i! (n-i)!} \int_0^n t(t-1) \cdots (t-i+1)(t-i-1) \cdots (t-n) dt, \\ i &= \overline{0, n}. \end{aligned}$$

Позначимо

$$c_i^{(n)} = (b - a) D_i^{(n)}, \quad (4.6)$$

тоді

$$\begin{aligned} D_i^{(n)} &= \frac{(-1)^{n-i}}{n i! (n-i)!} \int_0^n t(t-1) \cdots (t-i+1)(t-i-1) \cdots (t-n) dt, \\ i &= \overline{0, n}. \end{aligned} \quad (4.7)$$

Коефіцієнти $D_i^{(n)}$ не залежать від проміжку інтегрування $[a, b]$ і можуть бути обчислені один раз.

Оцінка залишкового члена квадратурних формул Ньютона–Котеса має вигляд

$$|R_n| \leq \frac{M_{n+1}}{(n+1)!} \int_a^b |(x - x_0)(x - x_1) \cdots (x - x_n)| dx. \quad (4.8)$$

На практиці використовують часткові випадки формул Ньютона–Котеса при невеликих n , оскільки при великих n деякі коефіцієнти $c_i^{(n)}$ стають від’ємними, що призводить до великих похибок заокруглень.

Розглянемо детальніше формули Ньютона–Котеса. Нехай $n = 0$, $x_0 = a$, тобто підінтегральну функцію замінимо інтерполяційним многочленом нульового степеня $f(x_0)$, тоді

$$I \approx I_L = (b - a)f(a).$$

Ця формула називається формулою *лівих прямокутників*. Якщо $x_0 = b$, то одержимо формулу *правих прямокутників*

$$I \approx I_R = (b - a)f(b).$$

А при $x_0 = (a + b)/2$ будемо мати формулу *середніх прямокутників* (див. рис. 4.1)

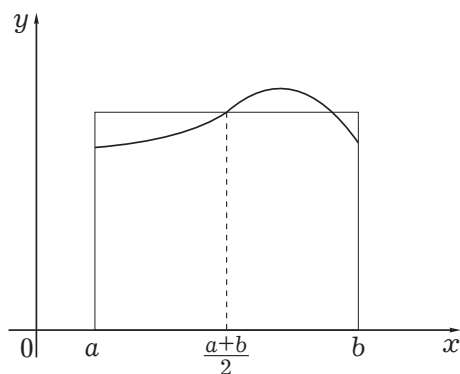


Рис. 4.1.

$$I \approx I_M = (b - a)f\left(\frac{a + b}{2}\right).$$

Оцінка залишкового члена (4.8) при $n = 0$, $x_0 = a$ (тобто для формули лівих прямокутників) буде мати вигляд

$$|R_L| \leq M_1 \int_a^b |x - a| dx = \frac{M_1(x - a)^2}{2} \Big|_a^b = \frac{M_1(b - a)^2}{2}.$$

Ця оцінка не зміниться, якщо $x_0 = b$. Застосуємо (4.8) до формули

середніх прямокутників

$$\begin{aligned}
 |R_M| &\leq M_1 \int_a^b \left| x - \frac{a+b}{2} \right| dx = \\
 &= M_1 \int_a^{\frac{a+b}{2}} \left(\frac{a+b}{2} - x \right) dx + M_1 \int_{\frac{a+b}{2}}^b \left(x - \frac{a+b}{2} \right) dx = \\
 &= -\frac{1}{2} M_1 \left(\frac{a+b}{2} - x \right)^2 \Big|_a^{\frac{a+b}{2}} + \frac{1}{2} M_1 \left(x - \frac{a+b}{2} \right)^2 \Big|_{\frac{a+b}{2}}^b = \\
 &= \frac{1}{2} M_1 \frac{(b-a)^2}{4} + \frac{1}{2} M_1 \frac{(b-a)^2}{4} = \frac{M_1 (b-a)^2}{4}.
 \end{aligned}$$

Якщо від підінтегральної функції $f(x)$ існує неперервна друга похідна $f''(x)$, то для формули середніх прямокутників можна одержати іншу оцінку точності. Для цього розкладемо $f(x)$ в ряд Тейлора в околі точки $(a+b)/2$:

$$\begin{aligned}
 f(x) &= f\left(\frac{a+b}{2}\right) + \left(x - \frac{a+b}{2}\right) f'\left(\frac{a+b}{2}\right) + \\
 &\quad + \frac{1}{2} \left(x - \frac{a+b}{2}\right)^2 f''(\xi),
 \end{aligned}$$

$$\begin{aligned}
 R_M &= I - (b-a) f\left(\frac{a+b}{2}\right) = \\
 &= \int_a^b \left[f\left(\frac{a+b}{2}\right) + \left(x - \frac{a+b}{2}\right) f'\left(\frac{a+b}{2}\right) + \right. \\
 &\quad \left. + \frac{1}{2} \left(x - \frac{a+b}{2}\right)^2 f''(\xi) \right] dx - (b-a) f\left(\frac{a+b}{2}\right) = \\
 &= f\left(\frac{a+b}{2}\right) (b-a) + \frac{1}{2} \left(x - \frac{a+b}{2}\right)^2 f'\left(\frac{a+b}{2}\right) \Big|_a^b + \\
 &\quad + \frac{1}{6} \left(x - \frac{a+b}{2}\right)^3 f''(\xi) \Big|_a^b - f\left(\frac{a+b}{2}\right) (b-a) =
 \end{aligned}$$

$$= \frac{(b-a)^2}{8} f' \left(\frac{a+b}{2} \right) - \frac{(a-b)^2}{8} f' \left(\frac{a+b}{2} \right) + \\ + \frac{(b-a)^3}{48} f''(\xi) - \frac{(a-b)^3}{48} f''(\xi) = \frac{(b-a)^3 f''(\xi)}{24},$$

де $\xi \in [a, b]$. Тоді

$$|R_M| \leq \frac{M_2(b-a)^3}{24}.$$

Покладемо в (4.7) $n = 1$, тоді

$$D_0^{(1)} = - \int_0^1 (t-1) dt = - \frac{(t-1)^2}{2} \Big|_0^1 = \frac{1}{2},$$

$$D_1^{(1)} = \int_0^1 t dt = \frac{t^2}{2} \Big|_0^1 = \frac{1}{2},$$

$$c_0^{(1)} = \frac{b-a}{2}, \quad c_1^{(1)} = \frac{b-a}{2}.$$

Підставимо коефіцієнти $c_0^{(1)}, c_1^{(1)}$ у (4.5), тоді одержимо *формулу трапеції* (див. рис. 4.2)

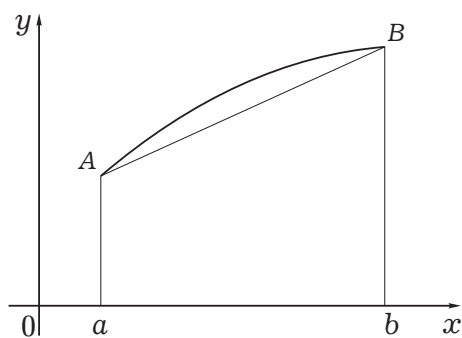


Рис. 4.2.

$$I \approx I_T = \frac{(b-a)}{2} [f(a) + f(b)]$$

з оцінкою залишкового члена

$$\begin{aligned}
 |R_T| &\leq \frac{M_2}{2!} \int_a^b |(x-a)(x-b)| dx = \\
 &= \frac{M_2}{2!} \int_a^b [-x^2 + (a+b)x - ab] dx = \\
 &= \frac{M_2}{2} \left[-\frac{x^3}{3} + (a+b)\frac{x^2}{2} - abx \right]_a^b = \\
 &= \frac{M_2}{2} \left[-\frac{b^3}{3} + (a+b)\frac{b^2}{2} - ab^2 + \frac{a^3}{3} - (a+b)\frac{a^2}{2} + a^2b \right] = \\
 &= \frac{M_2}{12} (b^3 - 3ab^2 + 3a^2b - a^3) = \frac{M_2}{12} (b-a)^3.
 \end{aligned}$$

Нехай $n = 2$, тоді

$$\begin{aligned}
 D_0^{(2)} &= \frac{1}{4} \int_0^2 (t-1)(t-2) dt = \frac{1}{4} \int_0^2 (t^2 - 3t + 2) dt = \\
 &= \frac{1}{4} \left(\frac{t^3}{3} - 3\frac{t^2}{2} + 2t \right) \Big|_0^2 = \frac{1}{6},
 \end{aligned}$$

$$D_1^{(2)} = -\frac{1}{2} \int_0^2 t(t-2) dt = -\frac{1}{2} \left(\frac{t^3}{3} - t^2 \right) \Big|_0^2 = \frac{2}{3},$$

$$D_2^{(2)} = \frac{1}{4} \int_0^2 t(t-1) dt = \frac{1}{4} \left(\frac{t^3}{3} - \frac{t^2}{2} \right) \Big|_0^2 = \frac{1}{6},$$

$$c_0^{(2)} = \frac{b-a}{6}, \quad c_1^{(2)} = \frac{2(b-a)}{3}, \quad c_2^{(2)} = \frac{b-a}{6}.$$

Отже, отримаємо формулу

$$I \approx I_S = \frac{b-a}{6} \left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right],$$

яка називається *формулою Сімпсона (парабол)* (див. рис. 4.3). Якщо припустити, що існує неперервна четверта похідна від підінтегральної

функції $f(x)$, то аналогічно як для формули середніх прямокутників можна показати, що:

$$|R_s| \leq \frac{M_4(b-a)^5}{2880}.$$

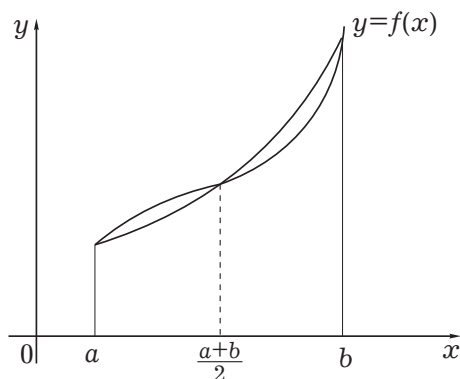


Рис. 4.3.

В усі отримані оцінки похибок входять степені довжини відрізка $[a, b]$. Якщо ця довжина не буде малою, то, взагалі кажучи, не будуть малими ці оцінки. Однак на практиці будемо застосовувати квадратурні формули тільки на досить малих відрізках, які одержуються в результаті розбиття даного відрізка $[a, b]$. Розбиваючи $[a, b]$ на N рівних частин довжини $h = (b-a)/N$, матимемо

$$\int_a^b f(x)dx = \sum_{i=1}^N \int_{x_{i-1}}^{x_i} f(x)dx,$$

де $x_i = a + ih$, $i = \overline{0, N-1}$, $x_N = b$. Якщо тепер на кожному з відрізків $[x_{i-1}, x_i]$ застосувати формулу лівих прямокутників, то одержимо складену формулу лівих прямокутників

$$I \approx I_{CL} = h \sum_{i=1}^N f(x_{i-1})$$

з оцінкою залишкового члена

$$|R_{CL}| \leq \frac{M_1}{2} \sum_{i=1}^N h^2 = \frac{M_1(b-a)}{2} h.$$

Застосовуючи до кожної частини відрізка $[a, b]$ відповідну формулу, отримаємо *складені формули середніх прямокутників і трапецій*

$$\begin{aligned}
 I &\approx I_{CM} = h \sum_{i=1}^N f\left(\frac{x_{i-1} + x_i}{2}\right) = h \sum_{i=1}^N f\left(a + \frac{2i-1}{2}h\right), \\
 |R_{CM}| &\leq \frac{M_2}{24}(b-a)h^2, \\
 I &\approx I_{CT} = \frac{h}{2} \sum_{i=1}^N [f(x_{i-1}) + f(x_i)] = \\
 &= h \left[\frac{1}{2}(f(a) + f(b)) + \sum_{i=1}^{N-1} f(a + ih) \right], \\
 |R_{CT}| &\leq \frac{M_2}{12}(b-a)h^2,
 \end{aligned}$$

де $f_i = f(x_i)$, $i = \overline{0, N}$. Поклавши $h = (b-a)/2N$, $x_i = a + 2ih$, маємо *складену формулу Сімпсона (парабол)*

$$\begin{aligned}
 I &\approx I_{CS} = \frac{h}{3} \sum_{i=1}^N \left[f(x_{i-1}) + 4f\left(\frac{x_{i-1} + x_i}{2}\right) + f(x_i) \right] = \\
 &= \frac{h}{3} \left[f(a) + 4 \sum_{i=1}^N f\left(\frac{x_{i-1} + x_i}{2}\right) + 2 \sum_{i=1}^{N-1} f(x_i) + f(b) \right] = \\
 &= \frac{h}{3} \left[f(a) + 4 \sum_{i=1}^N f(a + (2i-1)h) + 2 \sum_{i=1}^{N-1} f(a + 2ih) + f(b) \right], \\
 |R_{CS}| &\leq \frac{M_4}{2880} \sum_{i=1}^N (2h)^5 = \frac{M_4(b-a)h^4 2^4}{2880} = \frac{M_4(b-a)h^4}{180}.
 \end{aligned}$$

Приклад 4.1. Застосуйте формулу трапецій для обчислення інтеграла від функції $f(x) = x^4$ на відрізку $[0, 1]$, порівняйте фактичну похибку з оцінкою залишкового члена.

▷ Обчислимо точне значення інтеграла та наближене за формулою трапецій

$$I = \int_0^1 x^4 dx = 0,2, \quad I_T = \frac{b-a}{2} [f(a) + f(b)] = 0,5.$$

Знайдемо фактичну похибку $|I_T - I| = 0,3$. Оскільки $M_2 = \max_{x \in [0,1]} |f''(x)| = 12$, то оцінка залишкового члена $R_T \leq \frac{M_2}{12}(b-a)^3 = 1$. Отже, $|I_T - I| < 1$, тобто фактична похибка менша за оцінку залишкового члена. ◀

Приклад 4.2. Яким повинен бути крок h у складеній формулі середніх прямокутників при обчисленні інтеграла від функції $f(x) = x^4$ на відрізку $[0, 1]$, щоб похибка не перевищувала 10^{-6} ?

▷ Використовуючи оцінку залишкового члена складеної формули середніх прямокутників будемо мати

$$|R_{CM}| \leq \frac{M_2}{24}(b-a)h^2 = \frac{h^2}{2} \leq 10^{-6}.$$

Звідси $h \leq \sqrt{2} \cdot 10^{-3}$. ◀

4.3. Квадратурні формули Гаусса

Алгебраїчна степінь точності формул Ньютона–Котеса з n вузлами у загальному випадку є $n - 1$. Виявляється, що за рахунок вибору вузлів можна одержати квадратурні формули, які будуть точними для многочленів степеня вище $n - 1$.

Нехай квадратурна формула

$$\int_a^b \rho(x)f(x)dx = \sum_{k=1}^n c_k^{(n)} f(x_k^{(n)}) + R_n, \quad (4.9)$$

де $\rho(x) > 0 \ \forall x \in (a, b)$ — вагова функція така, що

$$\left| \int_a^b \rho(x)x^i dx \right| < \infty, \quad i = 0, 1, \dots,$$

є формулою інтерполяційного типу, тобто

$$c_k^{(n)} = \int_a^b \rho(x) \prod_{\substack{j=1 \\ j \neq k}}^n \frac{x - x_j^{(n)}}{x_k^{(n)} - x_j^{(n)}} dx. \quad (4.10)$$

Покажемо, що існують формули вигляду (4.9), для яких залишковий член R_n дорівнює нулю на будь-яких многочленах степеня не вище $2n - 1$. Такі квадратурні формули називатимемо *формулами найвищого алгебраїчного степеня точності*.

Нехай $\{P_i(x)\}_{i=0}^{\infty}$ — система ортогональних з вагою $\rho(x)$ многочленів на $[a, b]$, тобто

$$\int_a^b \rho(x) P_i(x) P_j(x) dx = 0 \quad \text{при} \quad i \neq j.$$

Зазначимо, що будь-який многочлен $Q_m(x)$ степеня m можна єдиним чином подати у вигляді

$$Q_m(x) = c_m P_m(x) + \dots + c_1 P_1(x) + c_0 P_0(x),$$

де c_i — сталі. Звідси випливає, що $P_n(x)$ ортогональний з вагою $\rho(x)$ до всіх многочленів $Q_m(x)$ степеня $m \leq n - 1$.

Існує єдиний ортогональний з вагою $\rho(x)$ многочлен $a_{0,n}^{-1} P_n(x)$, де $a_{0,n}$ — коефіцієнт при x^n у многочлені $P_n(x)$.

Справджується таке твердження.

► **ТЕОРЕМА 4.1.** Для того, щоб квадратурна формула (4.9) з коефіцієнтами (4.10) була квадратурною формулою найвищого алгебраїчного степеня точності, необхідно і достатньо, щоб $\omega(x) = a_{0,n}^{-1} P_n(x)$, тобто вузли $x_k^{(n)}$, $k = \overline{1, n}$ збігалися з нулями ортогональних з вагою $\rho(x)$ многочленів $P_n(x)$, причому така квадратурна формула єдина.

Доведення. Нехай (4.9) є квадратурною формулою найвищого алгебраїчного степеня точності, тобто точною для функції $f(x) = Q(x)\omega(x)$, де $Q(x)$ — многочлен степеня $m \leq n - 1$, $\omega(x) = \prod_{j=1}^n (x - x_j^{(n)})$. Оскільки $\omega(x_k^{(n)}) = 0$, $k = \overline{1, n}$ і залишковий член квадратурної формули $R_n = 0$, то

$$\int_a^b \rho(x) Q(x) \omega(x) dx = \sum_{k=1}^n c_k^{(n)} Q(x_k^{(n)}) \omega(x_k^{(n)}) + R_n = R_n = 0, \quad (4.11)$$

Отже, многочлен $\omega(x)$ ортогональний до всіх многочленів степеня не вище $n - 1$. Звідси випливає, що $x_k^{(n)}$, $k = \overline{1, n}$ збігаються з нулями єдиного ортогонального многочлена $a_{0,n}^{-1} P_n(x)$.

Нехай вузли $x_k^{(n)}$, $k = \overline{1, n}$ збігаються з нулями многочленів $P_n(x)$, тобто $\omega(x) = a_{0,n}^{-1} P_n(x)$. Покажемо, що $R_n = 0$ для будь-якого многочлена степеня не вище $2n - 1$. Дійсно,

$$f(x) = \omega(x)Q(x) + r(x), \quad (4.12)$$

де $Q(x)$ і $r(x)$ — многочлени степеня не вище $n - 1$. Оскільки $\omega(x) = a_{0,n}^{-1} P_n(x)$ ортогональний до $Q(x)$, то

$$\begin{aligned} \int_a^b \rho(x)f(x)dx &= \int_a^b \rho(x)\omega(x)Q(x)dx + \int_a^b \rho(x)r(x)dx \\ &= \int_a^b \rho(x)r(x)dx. \end{aligned} \quad (4.13)$$

З (4.12) випливає, що

$$f(x_i^{(n)}) = r(x_i^{(n)}), \quad i = \overline{1, n},$$

тобто $r(x)$ — інтерполяційний многочлен функції $f(x)$. Але формула (4.9) є формулою інтерполяційного типу, яка точна для многочленів степеня $n - 1$. Отже,

$$\int_a^b \rho(x)f(x)dx = \int_a^b \rho(x)r(x)dx = \sum_{k=1}^n c_k^{(n)} r(x_k^{(n)}) = \sum_{k=1}^n c_k^{(n)} f(x_k^{(n)}),$$

тобто $R_n = 0$, якщо $f(x)$ — будь-який многочлен степеня не вище $2n - 1$.

Єдиність квадратурної формули найвищого алгебраїчного степеня точності впливає з єдиності з точністю до сталого множника многочлена $P_n(x)$ ортогонального з вагою $\rho(x)$ на $[a, b]$. ■

Квадратурна формула (4.9) з коефіцієнтами (4.10), яка має найвищий алгебраїчний степінь точності, називається ще *формулою Гаусса*. Вагові коефіцієнти $c_k^{(n)}$ в (4.9), (4.10) часто називають *коефіцієнтами Крістоффеля*.

Можна показати, що $2n - 1$ — це найвища степінь точності формули Гаусса, тобто, що існує многочлен степеня $2n$, для якого ця формула не є точною. Дійсно, покладемо в (4.9)

$$f(x) = \prod_{i=1}^n (x - x_i^{(n)})^2.$$

Тоді

$$\int_a^b \rho(x) \prod_{i=1}^n (x - x_i^{(n)})^2 dx = \sum_{k=1}^n c_k^{(n)} \omega^2(x_k^{(n)}) + R_n = R_n.$$

Отже,

$$R_n = \int_a^b \rho(x) \omega^2(x) dx > 0.$$

Доведемо тепер, що при будь-якому n коефіцієнти $c_k^{(n)}$ формули Гаусса додатні. Розглянемо многочлени

$$\varphi_i(x) = \left(\prod_{\substack{j=1 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j} \right)^2, \quad i = \overline{1, n}$$

степеня $2n - 2$, для яких $\varphi_i(x_k) = \delta_{ik}$.

Оскільки для цих многочленів формула Гаусса точна, то справджується рівність

$$\int_a^b \rho(x) \varphi_i(x) dx = \sum_{k=1}^n c_k^{(n)} \varphi_i(x_k) = c_i^{(n)} > 0, \quad i = \overline{1, n}.$$

Властивість додатності коефіцієнтів важлива для стійкості обчислень і дозволяє використовувати формули з великим числом вузлів n . На практиці використовують формули Гаусса з числом вузлів до 100.

Можна показати, що для залишкового члена формули Гаусса справджується зображення

$$R_n = \frac{1}{(2n)!} \int_a^b \rho(x) \omega^2(x) f^{(2n)}(\xi(x)) dx, \quad \xi \in (a, b),$$

$$\omega(x) = \prod_{j=1}^n (x - x_j^{(n)}),$$

а, отже, і оцінка

$$|R_n| \leq \frac{M_{2n}}{(2n)!} \int_a^b \rho(x) \omega^2(x) dx.$$

Таким чином, для побудови квадратурної формули найвищого алгебраїчного степеня точності необхідно побудувати відповідну систему ортогональних многочленів і знайти їх корені. Для вагових функцій $\rho(x)$, пов'язаних з класичними ортогональними многочленами є таблиці вагових коефіцієнтів і абсцис відповідних квадратурних формул Гаусса.

Розглянемо інтеграл

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

Зробимо заміну

$$x = \frac{a+b}{2} + \frac{b-a}{2}t, \quad t \in [0, 1].$$

Тоді

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

Як відомо, ортогональну систему на $[-1, 1]$ з вагою $\rho(t) \equiv 1$ утворюють многочлени Лежандра, які можна знайти, наприклад, з рекурентного співвідношення

$$P_{n+1}(t) = \frac{2n+1}{n+1}tP_n(t) - \frac{n}{n+1}P_{n-1}(t), \quad (4.14)$$

де

$$P_{-1}(t) = 0, \quad P_0(t) = 1.$$

Нехай необхідно обчислити інтеграл

$$\int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx,$$

з ваговою функцією

$$\rho(x) = \frac{1}{\sqrt{1-x^2}}.$$

Ортогональними з такою вагою є многочлени Чебишева 1-го роду

$$T_n(x) = \cos(n \arccos(x)).$$

Тому вузли квадратурної формули знаходять з рівнянь

$$T_n(x) = \cos(n \arccos(x)) = 0.$$

Звідки

$$x_k^{(n)} = \cos \frac{2k-1}{2n} \pi, \quad k = \overline{1, n}, \quad c_k^{(n)} = \frac{\pi}{n}, \quad k = \overline{1, n}.$$

Отже,

$$\int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx \approx \frac{\pi}{n} \sum_{k=1}^n f(x_k).$$

Приклад 4.3. Побудуйте квадратурну формулу Гаусса з двома вузлами для інтегралів вигляду

$$\int_0^1 f(x) dx.$$

▷ Зробимо заміну $x = \frac{1}{2} + \frac{1}{2}t$, тоді

$$I = \frac{1}{2} \int_{-1}^1 f\left(\frac{1}{2} + \frac{1}{2}t\right) dt.$$

Шукана квадратурна формула буде мати вигляд

$$I \approx \frac{1}{2} \left[c_1^{(2)} f\left(\frac{1}{2} + \frac{1}{2}t_1^{(2)}\right) + c_2^{(2)} f\left(\frac{1}{2} + \frac{1}{2}t_2^{(2)}\right) \right],$$

де $t_1^{(2)}, t_2^{(2)}$ — нулі многочлена $P_2(t)$ другого степеня із системи ортогональних на $[-1, 1]$ з вагою $\rho(t) \equiv 1$ многочленів Лежандра $\{P_i(t)\}_{i=0}^\infty$, які можна знайти з рекурентного співвідношення (4.14): $P_1(t) = t$, $P_2(t) = \frac{3}{2}t^2 - \frac{1}{2}$. Таким чином, $t_1^{(2)} = -\frac{\sqrt{3}}{3}$, $t_2^{(2)} = \frac{\sqrt{3}}{3}$ і є вузлами квадратурної формули. Тоді згідно з формулою (4.10)

$$c_1^{(2)} = \int_{-1}^1 \frac{t - t_2^{(2)}}{t_1^{(2)} - t_2^{(2)}} dt = -\frac{\sqrt{3}}{2} \int_{-1}^1 \left(t - \frac{\sqrt{3}}{3}\right) dt = -\frac{\sqrt{3}}{4} \left(t - \frac{\sqrt{3}}{3}\right)^2 \Big|_{-1}^1 = 1,$$

$$c_2^{(2)} = \int_{-1}^1 \frac{t - t_1^{(2)}}{t_2^{(2)} - t_1^{(2)}} dt = \frac{\sqrt{3}}{2} \int_{-1}^1 \left(t + \frac{\sqrt{3}}{3}\right) dt = \frac{\sqrt{3}}{4} \left(t + \frac{\sqrt{3}}{3}\right)^2 \Big|_{-1}^1 = 1.$$

Отже, формула Гаусса буде мати вигляд

$$I \approx \frac{1}{2} \left[f\left(\frac{3-\sqrt{3}}{6}\right) + f\left(\frac{3+\sqrt{3}}{6}\right) \right].$$

Ця формула є квадратурною формулою алгебраїчного степеня точності 3. ◀

4.4. Практична оцінка похибки квадратурних формул

У багатьох практичних задачах похідні високих порядків від підінтегральної функції знайти важко, крім того підінтегральна функція може навіть не задаватися аналітичною формулою, а її значення знаходяться за допомогою довгого ланцюжка обчислень. Тоді оцінки залишкових членів, виведені в попередньому параграфі, практично непридатні. В будь-якій реальній задачі похибка оцінюється на основі числових (а не аналітичних) даних. Розповсюджений метод отримання практичних оцінок похибки полягає в комбінації двох або більшої кількості квадратурних формул. У найпростішому випадку застосовують дві формули і за оцінку похибки менш точної з них використовують модуль різниці двох наближень. Наприклад, якщо застосувати складені формули трапецій і Сімпсона, то різницю отриманих наближень можна використати для оцінки першої. Нижче буде показано, як для оцінки похибки можна використати квадратурні формули з різною кількістю елементарних відрізків.

Нехай необхідно обчислити інтеграл

$$I = \int_a^b f(x) dx \quad (4.15)$$

за допомогою деякої квадратурної формули I_h . Припустимо $f(x) \in C^{(p+1)}[a, b]$, тоді можна показати, що

$$I = I_h + Ch^p + O(h^{p+1}), \quad (4.16)$$

де C не залежить від h , $p = 2$ для формули середніх прямокутників і трапецій і $p = 4$ для формули Сімпсона. Величина Ch^p називається головною частиною похибки квадратурної формули. Проведемо обчислення за квадратурною формулою з кроками h і θh , $\theta > 0$. Тоді справджується формула (4.16) і

$$I = I_{\theta h} + C\theta^p h^p + O(h^{p+1}). \quad (4.17)$$

Віднімемо (4.17) від (4.16), тоді одержимо

$$I_h - I_{\theta h} = Ch^p(\theta^p - 1) + O(h^{p+1}).$$

$$I - I_h \approx Ch^p = \frac{I_h - I_{\theta h}}{\theta^p - 1}. \quad (4.18)$$

Отже, використання квадратурної формули з кроками h і θh дозволяє оцінити головну частину похибки квадратурної формули I_h . Зокрема, якщо вибрати $\theta = 2$, тоді для формули середніх прямокутників і трапецій маємо оцінку похибки

$$I - I_h \approx \frac{I_h - I_{2h}}{3},$$

а для формули Сімпсона

$$I - I_h \approx \frac{I_h - I_{2h}}{15}.$$

Можна виключити знайдену похибку (4.18) з формули (4.16) і одержати результат з вищою точністю

$$I_h^* = I_h + \frac{I_h - I_{\theta h}}{\theta^p - 1},$$

а саме з похибкою

$$I - I_h^* = O(h^{p+1}).$$

4.5. Наближене обчислення невластивих інтегралів

Нехай необхідно обчислити інтеграл

$$\int_a^\infty f(x)dx, \quad a > 0, \quad (4.19)$$

де функція $f(x)$ неперервна при $a \leq x < \infty$.

За допомогою заміни змінних

$$x = a/(1 - t)$$

півпряма $[a, \infty)$ перетворюється у відрізок $[0, 1]$. Якщо після перетворення підінтегральна функція разом з деяким числом похідних залишається обмеженою, то можна застосувати відомі квадратурні формули.

Другий спосіб обчислення інтеграла (4.19) полягає в заміні цього інтеграла інтегралом вигляду

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

де b настільки велике, щоб інтеграл

$$\int_b^\infty f(x)dx$$

був менший за допустиму похибку обчислень.

Ще одним способом обчислення є його приведення до вигляду

$$\int_a^\infty \rho(x)\tilde{f}(x)dx,$$

де $\rho(x) > 0$ та використання формули Гаусса.

Нехай тепер необхідно обчислити інтеграл

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

де $f(x)$ перетворюється у безмежність в деякій точці відрізка $[a, b]$. Будемо припускати, що в околі особливої точки $|f(x)| \leq M|x - \bar{x}|^\alpha$, де \bar{x} —особлива точка, $-1 < \alpha < 0$. Розглянемо методи обчислення інтеграла на окремому відрізку, де особливими точками є тільки одна або обидві границі інтеграла.

Перший спосіб — метод адитивного виділення особливостей. Розіб'ємо підінтегральну функцію на суму

$$f(x) = \varphi(x) + \psi(x),$$

де $\varphi(x)$ обмежена функція, а $\psi(x)$ інтегрується аналітичними методами. Тоді

$$\int_a^b \psi(x)dx$$

обчислюється точно, а

$$\int_a^b \varphi(x) dx$$

за допомогою квадратурних формул.

Другий спосіб — мультиплікативне виділення особливостей. Підінтегральну функцію зобразимо у вигляді

$$f(x) = \varphi(x)\rho(x),$$

де $\varphi(x)$ обмежена, а $\rho(x)$ додатна та інтегровна на відрізку $[a, b]$. Тоді можна розглядати $\rho(x)$, як вагову функцію і застосувати квадратурну формулу Гаусса.

Контрольні завдання

- ✎ 4.1. Покажіть, що формула трапецій точно інтегрує будь-яку лінійну функцію, а формула Сімпсона точно інтегрує будь-який кубічний многочлен.
- ✎ 4.2. Застосуйте формулу середніх прямокутників та Сімпсона для обчислення інтеграла від функції $f(x) = x^4$ на відрізку $[0, 1]$, порівняйте фактичну похибку з оцінкою залишкового члена.
- ✎ 4.3. Яким повинен бути крок h при використанні складених формул трапецій (Сімпсона) при обчисленні інтеграла від функції $f(x) = x^4$ на відрізку $[0, 1]$, щоб похибка не перевищувала 10^{-6} ?
- ✎ 4.4. Виберіть крок чисельного інтегрування так, щоб при обчисленні інтеграла від функції $\int_0^1 \frac{x}{1+x} dx$ за допомогою складеної формули трапецій (Сімпсона) отримати точність 10^{-4} .
- ✎ 4.5. Доведіть, що для коефіцієнтів інтерполяційної квадратурної формули справджується рівність

$$\sum_{i=0}^n c_i^{(n)} = \int_a^b \rho(x) dx.$$

- ✎ 4.6. Виведіть квадратурну формулу Ньютона–Котеса, яка має 4 вузли.

РОЗДІЛ 5

ЧИСЕЛЬНІ МЕТОДИ РОЗВ'ЯЗУВАННЯ ЗАДАЧІ КОШІ ДЛЯ ЗВИЧАЙНИХ ДИФЕРЕНЦІАЛЬНИХ РІВНЯНЬ

5.1. Задача Коші для звичайних диференціальних рівнянь

Нехай на відрізку $[t_0, T]$ необхідно знайти розв'язок задачі:

$$\mathbf{u}' = \mathbf{f}(t, \mathbf{u}), \quad \mathbf{u}(t_0) = \mathbf{u}_0, \quad (5.1)$$

де

$$\mathbf{u} = (u_1, u_2, \dots, u_N)^T, \quad \mathbf{f} = (f_1, f_2, \dots, f_N)^T,$$

або в покомпонентному вигляді:

$$u'_m = f_m(t, u_1, u_2, \dots, u_N), \quad m = \overline{1, N}, \quad (5.2)$$

$$u_m(t_0) = u_{0m}, \quad m = \overline{1, N}. \quad (5.3)$$

Зауважимо, що систему звичайних диференціальних рівнянь (ЗДР) будь-якого порядку можна звести до системи рівнянь першого порядку (5.1).

Добре відомі умови, які гарантують існування та єдиність розв'язку задачі (5.2), (5.3). Нехай функції f_m , $m = \overline{1, N}$ неперервні в замкненій області

$$D = \{ |t - t_0| \leq a, \quad |u_m - u_{0m}| \leq b, \quad m = \overline{1, N} \}.$$

З неперервності функцій f_m випливає їх обмеженість, тобто існування константи $M > 0$ такої, що скрізь у D виконуються нерівності $|f_m| \leq M$, $m = \overline{1, N}$. Крім того, припустимо, що в D функції f_m задовольняють умову Ліпшиця, тобто

$$\begin{aligned} & \left| f_m(t, u_1^{(1)}, u_2^{(1)}, \dots, u_N^{(1)}) - f_m(t, u_1^{(2)}, u_2^{(2)}, \dots, u_N^{(2)}) \right| \leq \\ & \leq L \left\{ |u_1^{(1)} - u_1^{(2)}| + |u_2^{(1)} - u_2^{(2)}| + \dots + |u_N^{(1)} - u_N^{(2)}| \right\} \end{aligned}$$

для будь-яких точок $(t, u_1^{(1)}, u_2^{(1)}, \dots, u_N^{(1)})$, $(t, u_1^{(2)}, u_2^{(2)}, \dots, u_N^{(2)})$ області D . Якщо умови, накладені на f_m , $m = \overline{1, N}$ виконуються, то існує єдиний розв'язок $u_1 = u_1(t)$, $u_2 = u_2(t)$, \dots , $u_N = u_N(t)$ системи (5.2), визначений при $|t - t_0| \leq t^* = \min(a, b/M)$, і такий, що при $t = t_0$ задовольняє задані початкові умови (5.3).

Методи розв'язування звичайних диференціальних рівнянь можна розділити на *точні*, *наближені* і *чисельні*. До точних відносять методи, які дозволяють виразити розв'язок диференціального рівняння через елементарні функції, або зобразити його за допомогою квадратур від елементарних функцій. Однак класи рівнянь, для яких розроблені методи знаходження точних розв'язків, порівняно вузькі і охоплюють дуже малу частину задач, що виникають на практиці. Наближеними називають методи, в яких розв'язок одержується як границя $\mathbf{u}(t)$ деякої послідовності $\{\mathbf{u}_n(t)\}$, причому елементи цієї послідовності $\mathbf{u}_n(t)$ виражаються через елементарні функції або за допомогою квадратур. Обмежуючись скінченним числом n , одержимо наближений вираз для $\mathbf{u}(t)$. Прикладом може бути метод розкладу розв'язку в степеневий ряд або метод послідовних наближень Пікара. Ці методи також можуть бути застосовані лише для порівняно простих задач (таких, як лінійні). Чисельні методи — це алгоритми обчислення наближених значень шуканого розв'язку $\mathbf{u}(t)$ на деякій вибраній сітці значень аргументу. Розв'язок при цьому одержується у вигляді таблиці. Чисельні методи не дозволяють знайти загальний розв'язок; вони можуть дати тільки який-небудь частинний розв'язок, наприклад, розв'язок задачі Коші (5.1). Це основний недолік чисельних методів. Однак, ці методи можна застосувати до дуже широкого класу рівнянь і всіх типів задач для них. Тому з появою комп'ютерів чисельні методи стали основним засобом розв'язування конкретних практичних задач.

Для розв'язування задачі Коші будемо використовувати чисельні методи. Виберемо на відрізку $[t_0, T]$ деяку, взагалі кажучи, *нерівномірну сітку* $\{t_n, 0 \leq n \leq n_0\}$ значень аргументу з кроком $\tau = t_{n+1} - t_n$ так, щоб для вузлів сітки виконувалися співвідношення $t_0 < t_1 < t_2 < \dots < t_{n_0} = T$. Будемо позначати через $\mathbf{u}(t)$ точний розв'язок задачі, а через $\mathbf{y}_n = \mathbf{y}(t_n)$ — наближений розв'язок. Зазначимо, що наближений розв'язок є сітковою функцією, тобто визначений тільки в точках сітки. Чисельні методи розв'язування звичайних диференціальних рівнянь розділяють на два класи: *однокрокові* (методи рядів Тейлора, методи Рунге-Кутта) та *багатокрокові* (методи Адамса, формули диферен-

ціювання назад тощо). Однокрокові методи для обчислення розв'язку \mathbf{y}_{n+1} в точці t_{n+1} на кожному кроці використовують лише значення \mathbf{y}_n , тоді як багатокрокові використовують значення наближеного розв'язку в кількох попередніх точках.

5.2. Метод рядів Тейлора

Для простоти викладення будемо надалі розглядати одне диференціальне рівняння

$$u' = f(t, u), \quad (5.4)$$

розв'язок якого задовольняє початкову умову

$$u(t_0) = u_0. \quad (5.5)$$

Якщо функція $f(t, u)$ аналітична в точці (t_0, u_0) , то розв'язок $u(t)$ задачі (5.4), (5.5) можна розкласти у *ряд Тейлора*. Обмежуючись скінченним числом членів ряду, запишемо наближену рівність:

$$\begin{aligned} u(t) \approx & u(t_0) + \frac{t - t_0}{1!} u'(t_0) + \frac{(t - t_0)^2}{2!} u''(t_0) + \\ & + \dots + \frac{(t - t_0)^p}{p!} u^{(p)}(t_0). \end{aligned} \quad (5.6)$$

Для знаходження похідних, які стоять у правій частині (5.6), послідовно продиференціюємо рівняння (5.4)

$$\begin{aligned} u' &= f(t, u), \quad u'' = f_t(t, u) + f(t, u) f_u(t, u), \\ u''' &= f_{tt}(t, u) + 2f(t, u) f_{tu}(t, u) + f^2(t, u) f_{uu}(t, u) + \\ &+ f_u(t, u) (f_t(t, u) + f(t, u) f_u(t, u)), \dots \end{aligned} \quad (5.7)$$

Похибка наближеної рівності (5.6) за умови, що величина $|t - t_0|$ менша від радіуса збіжності ряду Тейлора, прямує до нуля при $p \rightarrow \infty$.

На практиці застосовують покроковий варіант методу рядів. Виберемо на відрізку $[t_0, T]$ нерівномірну сітку $\{t_n, 0 \leq n \leq n_0\}$ з кроком $\tau = t_{n+1} - t_n$. Будемо вважати, що наближений розв'язок в точці t_n вже знайдено, тобто відоме значення $u(t_n) \approx y_n$. Для побудови однокрокового чисельного методу знаходження розв'язку в наступній точці

$t_{n+1} = t_n + \tau$ достатньо у формулі (5.6) покласти $t_0 = t_n$, а за t взяти $t_n + \tau$. Тоді одержимо

$$y_{n+1} = y_n + \frac{\tau}{1!} y'_n + \frac{\tau^2}{2!} y''_n + \cdots + \frac{\tau^p}{p!} y_n^{(p)}, \quad (5.8)$$

де $y_n^{(i)}$ обчислюються за формулами (5.7). Якщо б значення y_n збігалося зі значенням $u_n = u(t_n)$, то похибка від заміни $u(t_{n+1})$ на y_{n+1} мала б порядок $O(\tau^{p+1})$. Використовуючи пакети програм аналітичного диференціювання функцій, за допомогою формул (5.8), (5.7) можна послідовно при заданому $y_0 = u_0$ обчислити значення y_n , $n = 1, n_0$. Однак застосовувати для розрахунків формулу (5.8) з великим числом членів не вигідно, тому що навіть при порівняно простій правій частині вирази для похідних можуть бути громіздкими.

5.3. Методи Рунге–Кутта

У найпростішому випадку, обмежуючись тільки першими двома членами розкладу (5.6), одержимо *метод Ейлера (сіткову схему Ейлера)*

$$y_{n+1} = y_n + \tau f(t_n, y_n). \quad (5.9)$$

Метод Ейлера дуже простий для реалізації на комп'ютері: на n -ому кроці обчислюється значення $f(t_n, y_n)$, яке потім підставляється в (5.9).

Основне питання при використанні чисельних методів полягає в оцінці точності наближених значень y_n . Величину $z_{n+1} = y_{n+1} - u_{n+1}$ називають *похибкою (глобальною похибкою)* чисельного методу, тоді як величину $l_{n+1} = y_{n+1} - u_{n+1}$, де y_{n+1} — чисельний розв'язок, одержаний при точному значенні $y_n = u_n$ називають *локальною похибкою*. Кажуть, що метод *збіжний* в точці t_{n+1} , якщо $|z_{n+1}| \rightarrow 0$ при $\tau \rightarrow 0$. Метод збіжний на відрізку $(t_0, T]$, якщо він збіжний в кожній точці $t \in (t_0, T]$. Кажуть, що метод має p -й *порядок точності*, якщо існує число $p > 0$ таке, що $|z_{n+1}| = O(\tau^p)$ при $\tau \rightarrow 0$.

Підставимо $y_n = z_n + u_n$ в (5.9) і поділимо одержану рівність на τ , тоді будемо мати

$$\frac{z_{n+1} - z_n}{\tau} = f(t_n, u_n + z_n) - \frac{u_{n+1} - u_n}{\tau}.$$

Оскільки за теоремою про скінченні прирости

$$f(t_n, u_n + z_n) = f(t_n, u_n) + f_u(t_n, u_n + \theta z_n) z_n, \quad 0 < \theta < 1,$$

то

$$\frac{z_{n+1} - z_n}{\tau} - f_u(t_n, u_n + \theta z_n) z_n = f(t_n, u_n) - \frac{u_{n+1} - u_n}{\tau}.$$

Функція $\psi_n = f(t_n, u_n) - (u_{n+1} - u_n)/\tau$ називається *похибкою апроксимації* або *нев'язкою* сіткового рівняння на розв'язку вихідного рівняння (5.4).

Кажуть, що чисельний метод *апроксимує вихідне диференціальне рівняння*, якщо $\psi_n \rightarrow 0$ або $l_{n+1} \rightarrow 0$ при $\tau \rightarrow 0$. Метод має p -й *порядок апроксимації*, якщо $\psi_n = O(\tau^p)$ або $l_{n+1} = O(\tau^{p+1})$.

Встановимо порядок апроксимації схеми Ейлера (5.9). Для цього розкладемо l_{n+1} в ряд Тейлора в околі точки t_n

$$\begin{aligned} l_{n+1} &= y_{n+1} - u_{n+1} = \\ &= u_n + \tau f(t_n, u_n) - u_n - \tau u'_n - \frac{\tau^2}{2} u''_n + O(\tau^3) = \\ &= -\frac{\tau^2}{2} u''_n + O(\tau^3) = O(\tau^2). \end{aligned}$$

Оскільки $l_{n+1} = u_n + \tau f(t_n, u_n) - u_{n+1} = \tau \psi_n$, то $\psi_n = O(\tau)$. Отже, метод Ейлера має перший порядок апроксимації.

Доведемо, що схема Ейлера збіжна при $\tau \rightarrow 0$ і має перший порядок точності. Доведення проведемо, припускаючи, що $|f_u| \leq L$, для всіх $t_0 \leq t \leq T$ і величина кроку τ — стала. Тоді

$$\begin{aligned} |z_{n+1}| &= |z_n + \tau f_u(t_n, u_n + \theta z_n) z_n + \tau \psi_n| \leq \\ &\leq (1 + \tau L) |z_n| + \tau |\psi_n| \leq \\ &\leq (1 + \tau L)^2 |z_{n-1}| + (1 + \tau L) \tau |\psi_{n-1}| + \tau |\psi_n| \leq \\ &\leq \dots \leq (1 + \tau L)^{n+1} |z_0| + (1 + \tau L)^n \tau |\psi_0| + \\ &\quad + \dots + (1 + \tau L) \tau |\psi_{n-1}| + \tau |\psi_n| \leq \\ &\leq (n + 1) \tau (1 + \tau L)^n \max_{0 \leq i \leq n} |\psi_i|. \end{aligned}$$

Звідси з урахуванням нерівності $1 + \tau L \leq e^{\tau L}$, отримаємо

$$\begin{aligned} |z_{n+1}| &\leq (t_{n+1} - t_0) e^{L(t_n - t_0)} \max_{0 \leq i \leq n} |\psi_i| \leq \\ &\leq (T - t_0) e^{L(T - t_0)} \max_{0 \leq i \leq n} |\psi_i|, \end{aligned}$$

тобто $|z_{n+1}| = O(\tau)$ і наближений розв'язок збіжний до точного з першим порядком точності.

Для одержання точніших розрахункових формул обчислимо проміжне значення $y_{n+1/2}$, використовуючи схему Ейлера

$$y_{n+1/2} = y_n + \frac{\tau}{2} f(t_n, y_n), \quad (5.10)$$

а потім обчислимо y_{n+1} за формулою

$$y_{n+1} = y_n + \tau f\left(t_n + \frac{\tau}{2}, y_{n+1/2}\right). \quad (5.11)$$

Формули (5.10), (5.11) називають *методом прогнозу–корекції*, оскільки на першому етапі (5.10) наближене значення розв'язку прогнозується, а на другому етапі (5.11) прогнозоване значення коректується. Цей метод можна реалізувати інакше. А саме, спочатку обчислимо послідовно функції

$$k_1 = f(t_n, y_n), \quad k_2 = f\left(t_n + \frac{\tau}{2}, y_n + \frac{\tau k_1}{2}\right), \quad (5.12)$$

а потім знайдемо

$$y_{n+1} = y_n + \tau k_2. \quad (5.13)$$

Така форма реалізації схем (5.10), (5.11) називається *методом Рунге–Кутта*. Оскільки вимагається обчислити дві проміжні функції k_1, k_2 , то такий метод називають двоступеневим. Далі ми покажемо, що цей метод має другий порядок апроксимації.

У загальному випадку явні s -ступеневі методи Рунге–Кутта мають вигляд:

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + c_2\tau, y_n + \tau a_{21}k_1), \\ k_3 &= f(t_n + c_3\tau, y_n + \tau a_{31}k_1 + \tau a_{32}k_2), \\ &\dots\dots\dots \\ k_s &= f(t_n + c_s\tau, y_n + \tau a_{s1}k_1 + \dots + \tau a_{s,s-1}k_{s-1}), \\ y_{n+1} &= y_n + \tau(b_1k_1 + b_2k_2 + \dots + b_s k_s), \end{aligned}$$

де $c_i, a_{ij}, i = \overline{2, s}, j = \overline{1, s-1}, b_i, i = \overline{1, s}$ — дійсні коефіцієнти. Компа-

кне зображення методу Рунге–Кутта дає таблиця:

$$\begin{array}{cccccc} 0 & & & & & \\ c_2 & a_{21} & & & & \\ c_3 & a_{31} & a_{32} & & & \\ \dots & \dots & \dots & \dots & & \\ c_s & a_{s1} & a_{s2} & \dots & a_{s,s-1} & \\ & b_1 & b_2 & \dots & b_{s-1} & b_s. \end{array}$$

Переважно коефіцієнти c_i задовольняють такі умови

$$c_2 = a_{21}, \quad c_3 = a_{31} + a_{32}, \dots, \quad c_s = a_{s1} + a_{s2} + \dots + a_{s,s-1}, \quad (5.14)$$

зміст яких полягає в тому, що всі точки, в яких обчислюється f , є наближеннями першого порядку до розв'язків $u(t_n + c_i\tau)$, $i = \overline{1, s}$. Ці умови значно спрощують вивід методів Рунге–Кутта високого порядку.

Коефіцієнти c_i , a_{ij} , b_i вибираються з міркувань точності. Локальна похибка s -ступеневого методу Рунге–Кутта дорівнює

$$\begin{aligned} l_{n+1} &= y_{n+1} - u_{n+1} = \\ &= u_n + \tau(b_1 k_1 + b_2 k_2 + \dots + b_s k_s) - u_{n+1}, \end{aligned} \quad (5.15)$$

де

$$\begin{aligned} k_1 &= f(t_n, u_n), \\ k_i &= f(t_n + c_i\tau, u_n + \tau a_{i1}k_1 + \tau a_{i2}k_2 + \dots + \tau a_{i,i-1}k_{i-1}), \\ i &= \overline{2, s}. \end{aligned}$$

Зазначимо, що

$$l_{n+1} = \tau \left(-\frac{u_{n+1} - u_n}{\tau} + b_1 k_1 + b_2 k_2 + \dots + b_s k_s \right) = \tau \psi_n,$$

де ψ_n — похибка апроксимації або нев'язка методу.

Зупинимось детальніше на окремих методах. При $s = 1$ одержимо схему Ейлера, а при $s = 2$ — множину методів:

$$\begin{aligned} k_1 &= f(t_n, y_n), \quad k_2 = f(t_n + c_2\tau, y_n + a_{21}\tau k_1), \\ y_{n+1} &= y_n + \tau(b_1 k_1 + b_2 k_2). \end{aligned}$$

Дослідимо локальну похибку двоступеневого методу залежно від вибору параметрів, припускаючи достатню гладкість розв'язку $u(t)$ і

функції $f(t, u)$. Для цього розкладемо всі величини, які входять до виразу локальної похибки (5.15), за формулою Тейлора в околі точки t_n . Запишемо спочатку розклад за степенями τ значення u_{n+1}

$$u_{n+1} = u_n + \tau u'_n + \frac{\tau^2}{2} u''_n + \frac{\tau^3}{6} u'''_n + O(\tau^4).$$

Використовуючи співвідношення (5.7), одержимо

$$\begin{aligned} u_{n+1} = & u_n + \tau f + \frac{\tau^2}{2} (f_t + f \cdot f_u) \\ & + \frac{\tau^3}{6} [f_{tt} + 2f \cdot f_{tu} + f^2 f_{uu} + f_u (f_t + f \cdot f_u)] + \\ & + O(\tau^4). \end{aligned} \quad (5.16)$$

Тут значення функції $f(t, u)$ та її частинних похідних беруться при $t = t_n, u = u_n$.

Розкладемо тепер k_2 як функцію τ за формулою Тейлора

$$k_2(\tau) = k_2(0) + \tau \frac{dk_2(0)}{d\tau} + \frac{\tau^2}{2} \frac{d^2 k_2(0)}{d\tau^2} + O(\tau^3).$$

Знайдемо похідні

$$\begin{aligned} \frac{dk_2(\tau)}{d\tau} &= \frac{\partial f(t, u)}{\partial t} c_2 + \frac{\partial f(t, u)}{\partial u} k_1 a_{21}, \\ \frac{d^2 k_2(\tau)}{d\tau^2} &= \frac{\partial^2 f(t, u)}{\partial t^2} c_2^2 + 2 \frac{\partial^2 f(t, u)}{\partial t \partial u} c_2 a_{21} k_1 + \frac{\partial^2 f(t, u)}{\partial u^2} a_{21}^2 k_1^2, \\ t &= t_n + c_2 \tau, \quad u = u_n + a_{21} \tau k_1, \end{aligned}$$

Враховуючи, що $k_1 = f(t_n, u_n)$, $a_{21} = c_2$, одержимо

$$\frac{dk_2(0)}{d\tau} = c_2 (f_t + f f_u), \quad \frac{d^2 k_2(0)}{d\tau^2} = c_2^2 (f_{tt} + 2f f_{tu} + f^2 f_{uu}).$$

Тоді

$$\begin{aligned} k_2 = & f + \tau c_2 (f_t + f \cdot f_u) + \\ & + \frac{\tau^2}{2} c_2^2 (f_{tt} + 2f \cdot f_{tu} + f^2 f_{uu}) + O(\tau^3) \end{aligned} \quad (5.17)$$

i

$$\begin{aligned}
l_{n+1} = & \tau (b_1 + b_2 - 1) f + \tau^2 \left(b_2 c_2 - \frac{1}{2} \right) (f_t + f \cdot f_u) + \\
& + \tau^3 \left[\frac{1}{2} \left(b_2 c_2^2 - \frac{1}{3} \right) (f_{tt} + 2f \cdot f_{tu} + f^2 f_{uu}) - \right. \\
& \left. - \frac{1}{6} f_u (f_t + f \cdot f_u) \right] + O(\tau^4).
\end{aligned}$$

Прирівнюючи до нуля коефіцієнти при τ і τ^2 , одержимо систему рівнянь, яку повинні задовольняти коефіцієнти двоступеневого методу для того, щоб цей метод мав другий порядок апроксимації

$$b_1 + b_2 = 1, \quad b_2 c_2 = \frac{1}{2}, \quad a_{21} = c_2.$$

Зокрема, при $b_1 = 0$, $b_2 = 1$, $c_2 = a_{21} = 1/2$ будемо мати метод другого порядку (5.12), (5.13). Оскільки головний член (перший ненульовий член) тейлорівського розкладу локальної похибки

$$\begin{aligned}
l_{n+1} = & \tau^3 \left[\frac{1}{2} \left(b_2 c_2^2 - \frac{1}{3} \right) (f_{tt} + 2f \cdot f_{tu} + f^2 f_{uu}) - \right. \\
& \left. - \frac{1}{6} f_u (f_t + f \cdot f_u) \right] + O(\tau^4),
\end{aligned} \tag{5.18}$$

то звідси випливає, що двоступеневого методу третього порядку апроксимації не існує.

Щоб побудувати триступеневу схему третього порядку апроксимації, необхідно розкласти функції, які входять до (5.15) за формулою Тейлора до величин порядку τ^3 включно. Аналогічно до k_2 , розкладемо k_3 за степенями τ :

$$k_3(\tau) = k_3(0) + \tau \frac{dk_3(0)}{d\tau} + \frac{\tau^2}{2} \frac{d^2 k_3(0)}{d\tau^2} + O(\tau^3).$$

Обчислимо

$$\frac{dk_3(\tau)}{d\tau} = \frac{\partial f(t, u)}{\partial t} c_3 + \frac{\partial f(t, u)}{\partial u} \left(a_{31} k_1 + a_{32} k_2 + \tau a_{32} \frac{dk_2}{d\tau} \right),$$

$$\begin{aligned}
\frac{d^2 k_3(\tau)}{d\tau^2} &= \frac{\partial^2 f(t, u)}{\partial t^2} c_3^2 + \\
&+ 2 \frac{\partial^2 f(t, u)}{\partial t \partial u} c_3 \left(a_{31} k_1 + a_{32} k_2 + \tau a_{32} \frac{dk_2}{d\tau} \right) + \\
&+ \frac{\partial^2 f(t, u)}{\partial u^2} \left(a_{31} k_1 + a_{32} k_2 + \tau \frac{dk_2}{d\tau} \right)^2 + \\
&+ \frac{\partial f(t, u)}{\partial u} \left(2a_{32} \frac{dk_2(\tau)}{d\tau} + a_{32} \tau \frac{d^2 k_2}{d\tau^2} \right), \\
t &= t_n + c_3 \tau, \quad u = u_n + a_{31} \tau k_1 + a_{32} \tau k_2.
\end{aligned}$$

Звідси на підставі (5.14) будемо мати

$$\begin{aligned}
\frac{dk_3(0)}{d\tau} &= c_3(f_t + f f_u), \\
\frac{d^2 k_3(0)}{d\tau^2} &= c_3^2(f_{tt} + 2f f_{tu} + f^2 f_{uu}) + 2a_{32} c_2 f_u(f_t + f f_u).
\end{aligned}$$

Тоді

$$\begin{aligned}
k_3 &= f + \tau c_3(f_t + f \cdot f_u) + \\
&+ \frac{\tau^2}{2} [c_3^2(f_{tt} + 2f \cdot f_{tu} + f^2 f_{uu}) + 2a_{32} c_2 f_u(f_t + f \cdot f_u)] + \\
&+ O(\tau^3).
\end{aligned} \tag{5.19}$$

Враховуючи (5.16), (5.17), (5.19) локальна похибка триступеневого методу буде мати вигляд

$$\begin{aligned}
l_{n+1} &= \tau(b_1 + b_2 + b_3 - 1)f + \tau^2 \left(b_2 c_2 + b_3 c_3 - \frac{1}{2} \right) (f_t + f \cdot f_u) + \\
&+ \tau^3 \left[\frac{1}{2} \left(b_2 c_2^2 + b_3 c_3^2 - \frac{1}{3} \right) (f_{tt} + 2f \cdot f_{tu} + f^2 f_{uu}) + \right. \\
&\left. + \left(b_3 a_{32} c_2 - \frac{1}{6} \right) f_u(f_t + f \cdot f_u) \right] + O(\tau^4).
\end{aligned} \tag{5.20}$$

Прирівняємо до нуля коефіцієнти при τ, τ^2, τ^3 , тоді одержимо умови

третього порядку апроксимації

$$\begin{aligned} b_1 + b_2 + b_3 &= 1, & b_2 c_2 + b_3 c_3 &= \frac{1}{2}, & b_2 c_2^2 + b_3 c_3^2 &= \frac{1}{3}, \\ b_3 a_{32} c_2 &= \frac{1}{6}, & c_2 &= a_{21}, & c_3 &= a_{31} + a_{32}. \end{aligned}$$

Один з методів, який задовольняє ці умови, має вигляд

$$\begin{array}{c} 0 \\ \frac{1}{2} \quad \frac{1}{2} \\ 1 \quad -1 \quad 2 \\ \frac{1}{6} \quad \frac{4}{6} \quad \frac{1}{6} \end{array}$$

У випадку $s = 3$ не існує формул четвертого порядку апроксимації.

При $s = 4, 5$ не можна побудувати формул п'ятого порядку апроксимації. Наведемо таблицю коефіцієнтів найбільш вживаного чотири-ступеневого методу четвертого порядку

$$\begin{array}{c} 0 \\ \frac{1}{2} \quad \frac{1}{2} \\ \frac{1}{2} \quad 0 \quad \frac{1}{2} \\ 1 \quad 0 \quad 0 \quad 1 \\ \frac{1}{6} \quad \frac{2}{6} \quad \frac{2}{6} \quad \frac{1}{6} \end{array} \quad (5.21)$$

При певних припущеннях щодо гладкості правої частини рівняння (5.4), подібно до того як це було зроблено для схеми Ейлера, можна довести що, якщо метод Рунге–Кутта має p -й порядок апроксимації, то він збіжний з p -м порядком точності (див., напр., [18]).

5.4. Практична оцінка похибки та вибір довжини кроку для методів Рунге–Кутта

Найбільш традиційним способом оцінки локальної похибки є *правило Рунге* (подвійний перерахунок, стратегія $\tau - \tau/2$), який полягає в

повторенні обчислень із зменшеною вдвоє довжиною кроку і в порівнянні результатів.

Виходячи з точки (t_n, y_n) , методом Рунге–Кутта p -го порядку точності знайдемо чисельний розв'язок y_{n+1}^τ в точці t_{n+1} . Тоді локальна похибка розв'язку y_{n+1}^τ буде мати вигляд

$$l_{n+1} = y_{n+1}^\tau - u_{n+1} = C\tau^{p+1} + O(\tau^{p+2}), \quad (5.22)$$

де C виражається через коефіцієнти методу і частинні похідні правої частини (див. (5.18)), обчислені в точці $(t_n, y_n = u_n)$. Припустимо, що в результаті двох кроків, проведених, цим самим методом, виходячи з цієї ж точки (t_n, y_n) , обчислено $y_{n+1/2}^{\tau/2}$ і $y_{n+1}^{\tau/2}$. Похибка розв'язку $y_{n+1}^{\tau/2}$ складається з двох частин: з локальної похибки першого кроку, яка має вигляд

$$C(\tau/2)^{p+1} + O(\tau^{p+2}),$$

і локальної похибки другого кроку, яка також виражається формулою (5.22), але з частинними похідними, що входять в C , обчисленими в точці $t_{n+1/2} = t_n + \tau/2, y_{n+1/2} = y_n + O(\tau)$. Отже,

$$\begin{aligned} y_{n+1}^{\tau/2} - u_{n+1} &= C(\tau/2)^{p+1} + (C + O(\tau))(\tau/2)^{p+1} + O(\tau^{p+2}) = \\ &= 2C(\tau/2)^{p+1} + O(\tau^{p+2}) \end{aligned} \quad (5.23)$$

Якщо від (5.23) відняти (5.22), то отримаємо

$$\begin{aligned} y_{n+1}^{\tau/2} - y_{n+1}^\tau &= 2C(\tau/2)^{p+1} - C\tau^{p+1} + O(\tau^{p+2}) = \\ &= 2C(\tau/2)^{p+1}(1 - 2^p) + O(\tau^{p+2}). \end{aligned}$$

Звідси

$$2C(\tau/2)^{p+1} = \frac{y_{n+1}^{\tau/2} - y_{n+1}^\tau}{1 - 2^p} + O(\tau^{p+2}).$$

Тоді похибка (5.23) може бути обчислена за формулою

$$y_{n+1}^{\tau/2} - u_{n+1} = \frac{y_{n+1}^{\tau/2} - y_{n+1}^\tau}{1 - 2^p} + O(\tau^{p+2}), \quad (5.24)$$

а вираз

$$\hat{y}_{n+1} = y_{n+1}^{\tau/2} + \frac{y_{n+1}^{\tau/2} - y_{n+1}^\tau}{2^p - 1} \quad (5.25)$$

апроксимує величину u_{n+1} з порядком $p + 1$.

Розглянемо алгоритм автоматичного вибору довжини кроку τ так, щоб локальна похибка не перевищувала допустимої точності ε . На основі формули (5.24) обчислюється похибка

$$E = \frac{1}{2^p - 1} \cdot \frac{|y_{n+1}^{\tau/2} - y_{n+1}^\tau|}{d},$$

де d — масштабний множник. Для обчислення абсолютної похибки кладуть $d = 1$, а для відносної $d = |\hat{y}_{n+1}|$. Можна використовувати змішане масштабування типу

$$d = \max(|\hat{y}_{n+1}|, |y_n|, 1)$$

або

$$d = \max(|\hat{y}_{n+1}|, |y_n|, 10^{-6}).$$

Потім величина E порівнюється з заданою величиною допустимої похибки ε . Довжина нового кроку $\tau_H = \theta\tau$ вибирається з умови

$$\frac{1}{d} |2C(\tau_H/2)^{p+1}| = \frac{1}{d} |2C(\theta\tau/2)^{p+1}| \approx \theta^{p+1} E = \varepsilon.$$

Звідси

$$\theta = \left(\frac{\varepsilon}{E}\right)^{\frac{1}{p+1}},$$

і

$$\tau_H = 0,9 \tau \left(\frac{\varepsilon}{E}\right)^{\frac{1}{p+1}}. \quad (5.26)$$

Коефіцієнт 0,9 дозволяє врахувати вплив величини $O(\tau^{p+2})$. Тоді, якщо $E \leq \varepsilon$, то обчислений крок вважається успішним і розв'язування продовжується, виходячи з \hat{y}_{n+1} або $y_{n+1}^{\tau/2}$, причому довжина нового кроку вибирається рівною τ_H . У протилежному випадку крок відкидається і обчислення повторюються з новою довжиною кроку τ_H . Щоб запобігти занадто великому збільшенню довжини кроку, максимальний коефіцієнт збільшення кроку θ , як правило, вибирають між 1,5 і 5. Після неуспішного кроку рекомендується не збільшувати розмір кроку.

Другий підхід до оцінки локальної похибки полягає в тому, щоб побудувати такі формули Рунге–Кутта, які б дозволяли обчислити крім розв'язку y_{n+1} , і розв'язок більш високого порядку точності \hat{y}_{n+1} . Тобто,

необхідно знайти таку таблицю коефіцієнтів

$$\begin{array}{cccccc}
 0 & & & & & \\
 c_2 & a_{21} & & & & \\
 c_3 & a_{31} & a_{32} & & & \\
 \dots & \dots & \dots & \dots & & \\
 c_s & a_{s1} & a_{s2} & \dots & a_{s,s-1} & \\
 & b_1 & b_2 & \dots & b_{s-1} & b_s \\
 & \hat{b}_1 & \hat{b}_2 & \dots & \hat{b}_{s-1} & \hat{b}_s,
 \end{array}$$

щоб величина

$$y_{n+1} = y_n + \tau (b_1 k_1 + b_2 k_2 + \dots + b_s k_s)$$

мала порядок точності p , а

$$\hat{y}_{n+1} = y_n + \tau (\hat{b}_1 k_1 + \hat{b}_2 k_2 + \dots + \hat{b}_s k_s)$$

— порядок точності $p+1$. Такі методи називають *вкладеними*. Локальні похибки методів p -го та $p+1$ -го порядків точності мають вигляд:

$$y_{n+1} - u_{n+1} = C\tau^{p+1} + O(\tau^{p+2}), \quad (5.27)$$

$$\hat{y}_{n+1} - u_{n+1} = \hat{C}\tau^{p+2} + O(\tau^{p+3}). \quad (5.28)$$

Якщо від формули (5.27) відняти (5.28), то одержимо

$$y_{n+1} - \hat{y}_{n+1} = C\tau^{p+1} + O(\tau^{p+2})$$

Отже, для вкладених методів

$$E = \frac{|\hat{y}_{n+1} - y_{n+1}|}{d}.$$

Довжина нового кроку вибирається з умови

$$\frac{1}{d} |C\tau_H^{p+1}| = \frac{1}{d} |C(\theta\tau)^{p+1}| \approx \theta^{p+1} E = \varepsilon, \quad \theta = \left(\frac{\varepsilon}{E}\right)^{\frac{1}{p+1}},$$

тобто за формулою (5.26).

Виведемо вкладені формули другого і третього порядків апроксимації, таблиця яких має вигляд

$$\begin{array}{ccc} 0 & & \\ c_2 & a_{21} & \\ c_3 & a_{31} & a_{32} \\ & b_1 & b_2 & b_3 \\ & \hat{b}_1 & \hat{b}_2 & \hat{b}_3 \end{array}$$

З (5.20) та аналогічного розкладу для $\hat{l}_{n+1} = \hat{y}_{n+1} - u_{n+1}$ випливає, що для збереження відповідних порядків апроксимації, повинні задовольнятися рівняння

$$\begin{aligned} b_1 + b_2 + b_3 &= 1, & \hat{b}_1 + \hat{b}_2 + \hat{b}_3 &= 1, \\ b_2 c_2 + b_3 c_3 &= \frac{1}{2}, & \hat{b}_2 c_2 + \hat{b}_3 c_3 &= \frac{1}{2}, \\ c_2 &= a_{21}, & \hat{b}_2 c_2^2 + \hat{b}_3 c_3^2 &= \frac{1}{3}, \\ c_3 &= a_{31} + a_{32}, & \hat{b}_3 a_{32} c_2 &= \frac{1}{6}. \end{aligned}$$

Вибравши $c_2 = 1$ і $b_3 = 0$, з перших двох рівнянь одержимо $b_1 = b_2 = 1/2$. Якщо покласти $c_3 = 1/2$, $a_{32} = 1/4$, то $\hat{b}_1 = 1/6$, $\hat{b}_2 = 1/6$, $\hat{b}_3 = 4/6$, $a_{31} = 1/4$. Одержаний метод наведений в таблиці

$$\begin{array}{ccc} 0 & & \\ 1 & 1 & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ & \frac{1}{2} & \frac{1}{2} & 0 \\ & \frac{1}{6} & \frac{1}{6} & \frac{4}{6}. \end{array}$$

В основу ефективної програми розв’язування задачі Коші для звичайних диференціальних рівнянь лягли вкладені методи четвертого і п’ятого порядків Дорманта і Прінса (див. [23]), які визначаються коефіціє-

єнтами:

$$\begin{array}{cccccccc}
 0 & & & & & & & \\
 1 & 1 & & & & & & \\
 \hline 5 & \hline 5 & & & & & & \\
 3 & 3 & 9 & & & & & \\
 \hline 10 & \hline 40 & \hline 40 & & & & & \\
 4 & 44 & 56 & 32 & & & & \\
 \hline 5 & \hline 55 & \hline 15 & \hline 9 & & & & \\
 8 & 19372 & 25360 & 64448 & 212 & & & \\
 \hline 9 & \hline 6561 & \hline 2187 & \hline 6561 & \hline 729 & & & \\
 1 & 9017 & 355 & 46732 & 49 & 5103 & & \\
 \hline 1 & \hline 3168 & \hline 33 & \hline 5247 & \hline 176 & \hline 18656 & & \\
 & 35 & & 500 & 125 & 2187 & 11 & \\
 1 & & 0 & & & & & \\
 & \hline 384 & & \hline 1113 & \hline 192 & \hline 6784 & \hline 84 & & \\
 & 35 & & 500 & 125 & 2187 & 11 & \\
 & & 0 & & & & & \\
 & \hline 384 & & \hline 1113 & \hline 192 & \hline 6784 & \hline 84 & & 0 \\
 & 5179 & & 7571 & 393 & 92097 & 187 & 1 \\
 \hline & \hline 57600 & & \hline 16695 & \hline 640 & \hline 339200 & \hline 2100 & \hline 40
 \end{array}$$

Зазначимо, що застосування цього методу вимагає обчислення лише шести правих частин диференціального рівняння, оскільки k_7 збігається з k_1 для наступного кроку.

5.5. Лінійні багатокрокові методи

5.5.1. Методи Адамса

Введемо на інтервалі $[t_0, T]$ рівномірну сітку $\bar{\omega}_\tau = \{t_n = t_0 + n\tau, n = \overline{0, n_0}\}$ з кроком $\tau = (T - t_0)/n_0$. Якщо рівняння (5.4) проінтегрувати на відрізьку $[t_n, t_{n+1}]$, то одержимо

$$u_{n+1} = u_n + \int_{t_n}^{t_{n+1}} f(t, u(t)) dt. \quad (5.29)$$

Припустимо, що нам відомі наближені значення

$$y_{n-k+1}, y_{n-k+2}, \dots, y_n$$

точного розв'язку $u_{n-k+1}, u_{n-k+2}, \dots, u_n$ задачі (5.4), (5.5), тоді можна вважати також, що ми маємо і величини $f_j = f(t_j, y_j)$, $j = n - k + 1, n$.

Замінімо функцію $f(t, u)$ в (5.29) інтерполяційним многочленом Ньютона, який проходить через точки $\{(t_j, f_j), j = \overline{n-k+1}, n\}$. Його можна виразити через різниці назад:

$$\begin{aligned} P(t) &= P(t_n + s\tau) = \\ &= \nabla^0 f_n + \frac{s}{1!} \nabla f_n + \frac{s(s+1)}{2!} \nabla^2 f_n + \dots \\ &\quad + \frac{s(s+1) \dots (s+k-2)}{(k-1)!} \nabla^{k-1} f_n = \\ &= \nabla^0 f_n + \sum_{j=1}^{k-1} \frac{s(s+1) \dots (s+j-1)}{j!} \nabla^j f_n. \end{aligned} \quad (5.30)$$

Тоді чисельний аналог (5.29) буде задаватися формулою

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} P(t) dt.$$

Після заміни змінної $s = (t - t_n)/\tau$ в останньому інтегралі та підставлення виразу (5.30) будемо мати

$$y_{n+1} = y_n + \tau \int_0^1 P(t_n + s\tau) ds = y_n + \tau \sum_{j=0}^{k-1} \gamma_j \nabla^j f_n, \quad (5.31)$$

де коефіцієнти γ_j обчислюються за формулами

$$\gamma_0 = 1, \quad \gamma_j = \frac{1}{j!} \int_0^1 s(s+1) \dots (s+j-1) ds, \quad j = \overline{1, k-1}.$$

Формула (5.31) дозволяє визначити y_{n+1} явно, тому її називають *явним методом Адамса*.

Розглянемо частинні випадки (5.31). Якщо для $k = \overline{1, 4}$, обчислити $\gamma_j, j = \overline{0, 3}$ ($\gamma_0 = 1, \gamma_1 = 1/2, \gamma_2 = 5/12, \gamma_3 = 3/8$) та виразити різниці

назад через f_{n-j} , то одержимо такі формули

$$\begin{aligned} y_{n+1} &= y_n + \tau f_n, \quad k = 1, \\ y_{n+1} &= y_n + \tau \left(\frac{3}{2} f_n - \frac{1}{2} f_{n-1} \right), \quad k = 2, \\ y_{n+1} &= y_n + \tau \left(\frac{23}{12} f_n - \frac{16}{12} f_{n-1} + \frac{5}{12} f_{n-2} \right), \quad k = 3, \\ y_{n+1} &= y_n + \tau \left(\frac{55}{24} f_n - \frac{59}{24} f_{n-1} + \frac{37}{24} f_{n-2} - \frac{9}{24} f_{n-3} \right), \quad k = 4. \end{aligned}$$

Зауважимо, що при $k = 1$ ми маємо явний метод Ейлера.

Формули (5.31) одержані при інтегруванні інтерполяційного многочлена від t_n до t_{n+1} , тобто зовні інтервалу інтерполяції (t_{n-k+1}, t_n) . Добре відомо, що зовні цього інтервалу інтерполяційний многочлен дає досить погане наближення. Тому дослідимо також методи, що ґрунтуються на інтерполяційному многочлені, який додатково використовує точку (t_{n+1}, f_{n+1}) , тобто

$$\begin{aligned} P^*(t) &= P^*(t_n + s\tau) = \\ &= \nabla^0 f_{n+1} + \frac{s-1}{1!} \nabla f_{n+1} + \frac{(s-1)s}{2!} \nabla^2 f_{n+1} + \dots \\ &\quad + \frac{(s-1)s \dots (s+k-2)}{k!} \nabla^k f_{n+1} = \\ &= \nabla^0 f_{n+1} + \sum_{j=1}^k \frac{(s-1)s \dots (s+j-2)}{j!} \nabla^j f_{n+1}. \end{aligned} \quad (5.32)$$

Підставляючи цей многочлен у формулу (5.29) замість $f(t, u)$ та зробивши заміну $s = (t - t_n)/\tau$, одержимо *неявний метод Адамса*:

$$y_{n+1} = y_n + \tau \sum_{j=0}^k \gamma_j^* \nabla^j f_{n+1}, \quad (5.33)$$

де коефіцієнти γ_j^* визначаються за формулами

$$\gamma_0^* = 1, \quad \gamma_j^* = \frac{1}{j!} \int_0^1 (s-1)s \dots (s+j-2) ds, \quad j = \overline{1, k}.$$

Наведемо приклади формул (5.33). При $k = 0$, $\gamma_0^* = 1$ будемо мати неявний метод Ейлера

$$y_{n+1} = y_n + \tau f_{n+1},$$

при $k = 1$, $\gamma_1^* = -1/2$ правило трапецій

$$y_{n+1} = y_n + \tau \left(\frac{1}{2} f_{n+1} + \frac{1}{2} f_n \right).$$

Насправді ці два методи — однокрокові. При $k = 2, 3$, $\gamma_2^* = -1/12$, $\gamma_3^* = -1/24$ одержимо відповідно такі методи

$$\begin{aligned} y_{n+1} &= y_n + \tau \left(\frac{5}{12} f_{n+1} + \frac{8}{12} f_n - \frac{1}{12} f_{n-1} \right), \\ y_{n+1} &= y_n + \tau \left(\frac{9}{24} f_{n+1} + \frac{19}{24} f_n - \frac{5}{24} f_{n-1} + \frac{1}{24} f_{n-2} \right). \end{aligned}$$

Формули (5.33) визначають y_{n+1} неявно (на кожному кроці для обчислення y_{n+1} необхідно розв'язати нелінійне рівняння), а тому вони називаються неявними методами Адамса.

Неявні формули Адамса мають загальний вигляд:

$$y_{n+1} = y_n + \tau \sum_{j=0}^k \beta_j f_{n-j+1}. \quad (5.34)$$

5.5.2. Формули диференціювання назад

Багатокрокові формули Адамса ґрунтуються на чисельному інтегруванні, тобто інтеграл в (5.29) апроксимується деякою квадратурною формулою. Тепер розглянемо багатокрокові методи, які ґрунтуються на чисельному диференціюванні.

Припустимо, що відомі значення $y_{n-k+1}, y_{n-k+2}, \dots, y_n$ розв'язку диференціального рівняння (5.4). Щоб вивести формулу для y_{n+1} , використаємо інтерполяційний многочлен $Q(t)$, який проходить через точки $\{(x_j, y_j) \mid j = n - k + 1, n + 1\}$. Як і многочлен (5.32), його можна ви-

разити через різниці назад, а саме

$$\begin{aligned}
 Q(t) &= Q(t_n + s\tau) = \\
 &= \nabla^0 y_{n+1} + \frac{s-1}{1!} \nabla y_{n+1} + \frac{(s-1)s}{2!} \nabla^2 y_{n+1} + \dots \\
 &\quad + \frac{(s-1)s \dots (s+k-2)}{k!} \nabla^k y_{n+1} = \\
 &= \nabla^0 y_{n+1} + \sum_{j=1}^k \frac{(s-1)s \dots (s+j-2)}{j!} \nabla^j y_{n+1}. \tag{5.35}
 \end{aligned}$$

Визначимо тепер невідоме значення y_{n+1} так, щоб многочлен $Q(t)$ задовольняв диференціальне рівняння хоча б в одному вузлі сітки, тобто

$$Q'(t_{n+1-r}) = f(t_{n+1-r}, y_{n+1-r}). \tag{5.36}$$

Враховуючи, що $s = (t - t_n)/\tau$, продиференціюємо (5.35) по змінній t

$$Q'(t) = \frac{1}{\tau} \sum_{j=1}^k \frac{d}{ds} \left(\frac{(s-1)s \dots (s+j-2)}{j!} \right) \nabla^j y_{n+1}.$$

Для $r = 1$ одержимо явні формули

$$\sum_{j=1}^k \delta_j \nabla^j y_{n+1} = \tau f_n,$$

де

$$\delta_1 = 1, \quad \delta_j = \frac{d}{ds} \left[\frac{(s-1)s \dots (s+j-2)}{j!} \right]_{s=0} = -\frac{1}{j(j-1)}, \quad j \geq 2.$$

При $k = 1$ будемо мати явний метод Ейлера, а при $k = 2$ правило середньої точки

$$\frac{1}{2} y_{n+1} - \frac{1}{2} y_{n-1} = \tau f_n.$$

У випадку $k = 3$ формула має вигляд

$$\frac{1}{3} y_{n+1} + \frac{1}{2} y_n - y_{n-1} + \frac{1}{6} y_{n-2} = \tau f_n.$$

Однак вона нестійка, як і всі решта формул при $k > 3$ (див. розділ 5.5.4), а тому непридатна для розрахунків.

Кращі властивості мають формули, які одержуються з (5.36) при $r = 0$. Це неявні формули

$$\sum_{j=1}^k \delta_j^* \nabla^j y_{n+1} = \tau f_{n+1} \quad (5.37)$$

з коефіцієнтами

$$\delta_j^* = \frac{d}{ds} \left[\frac{(s-1)s \dots (s+j-2)}{j!} \right]_{s=1},$$

які після диференціювання набувають вигляду

$$\delta_j^* = \frac{1}{j} \quad \text{при} \quad j \geq 1.$$

Тому (5.37) зводиться до формули

$$\sum_{j=1}^k \frac{1}{j} \nabla^j y_{n+1} = \tau f_{n+1}.$$

Такі багатокрокові методи називають *формулами диференціювання назад*. Вони вперше були виведені Кертісом і Хіршфельдером.

Наведемо приклади цих формул, виразивши різниці назад через y_{n-i}

$$\begin{aligned} y_{n+1} - y_n &= \tau f_{n+1}, & k = 1, \\ \frac{3}{2}y_{n+1} - 2y_n + \frac{1}{2}y_{n-1} &= \tau f_{n+1}, & k = 2, \\ \frac{11}{6}y_{n+1} - 3y_n + \frac{3}{2}y_{n-1} - \frac{1}{3}y_{n-2} &= \tau f_{n+1}, & k = 3, \\ \frac{25}{12}y_{n+1} - 4y_n + 3y_{n-1} - \frac{4}{3}y_{n-2} + \frac{1}{4}y_{n-3} &= \tau f_{n+1}, & k = 4, \\ \frac{137}{60}y_{n+1} - 5y_n + 5y_{n-1} - \frac{10}{3}y_{n-2} + \frac{5}{4}y_{n-3} - \frac{1}{5}y_{n-4} &= \tau f_{n+1}, & k = 5. \end{aligned}$$

Формули диференціювання назад мають вигляд:

$$\sum_{j=0}^k \alpha_j y_{n-j+1} = \tau f_{n+1}.$$

5.5.3. Порядок апроксимації лінійних багатокрокових методів

Методи Адамса та формули диференціювання назад є частинними випадками *лінійних багатокрокових (різницевих)* методів, які описуються загальною формулою

$$\sum_{j=0}^k \alpha_j y_{n-j+1} = \tau \sum_{j=0}^k \beta_j f_{n-j+1}, \quad (5.38)$$

де $f_n = f(t_n, y_n)$, α_j, β_j — дійсні числа. Будемо вважати, що виконані такі умови :

$$\alpha_0 \neq 0, \quad |\alpha_k| + |\beta_k| > 0.$$

Рівняння (5.38) треба розглядати як рекурентне співвідношення, яке виражає нове значення y_{n+1} через знайдені раніше значення $y_n, y_{n-1}, \dots, y_{n-k+1}$. Для початку розрахунку необхідно задати k значень y_0, y_1, \dots, y_{k-1} . Значення y_0 визначається вихідною задачею, а саме $y_0 = u_0$. Величини y_1, y_2, \dots, y_{k-1} можна обчислити, наприклад, за допомогою методів Рунге–Кутта або багатокрокових методів нижчого порядку точності (з меншим значенням k). Надалі будемо вважати, що початкові значення задані.

Якщо $\beta_0 = 0$, то формула (5.38) задає явний метод; якщо $\beta_0 \neq 0$, то — неявний.

Зауважимо, коефіцієнти в (5.38) визначені з точністю до множника. Щоб ліквідувати цю довільність, будемо вважати, що виконується умова

$$\sum_{j=0}^k \beta_j = 1, \quad (5.39)$$

яка означає, що права частина різницевого рівняння (5.38) апроксимує праву частину рівняння (5.4).

Локальною похибкою багатокрокового методу (5.38) називається величина $l_{n+1} = y_{n+1} - u_{n+1}$ при точних значеннях $y_j = u_j$, $j = \overline{n-k+1, n}$. При $k = 1$ це означення збігається з означенням локальної похибки однокрокових методів.

Похибкою апроксимації або *нев'язкою* різницевого методу (5.38) називається функція

$$\psi_n = \frac{1}{\tau} \sum_{j=0}^k [-\alpha_j u_{n-j+1} + \tau \beta_j f(t_{n-j+1}, u_{n-j+1})], \quad (5.40)$$

яка одержується в результаті підставлення точного розв'язку $u(t)$ диференціальної задачі (5.4), (5.5) в різницеве рівняння (5.38) та ділення на τ .

⇒ **Лема 5.1.** Розглянемо диференціальне рівняння (5.4) з неперервно диференційовною функцією $f(t, u)$ і розв'язком $u(t)$. Тоді для локальної похибки лінійного багатокрокового методу виконується рівність

$$l_{n+1} = \tau(\alpha_0 - \tau\beta_0 f_u(t_{n+1}, u_{n+1} + \theta y_{n+1}))^{-1} \psi_n, \text{ де } 0 < \theta < 1.$$

Доведення. З означення локальної похибки

$$\begin{aligned} l_{n+1} &= y_{n+1} - u_{n+1} = \\ &= \frac{1}{\alpha_0} \sum_{j=1}^k [-\alpha_j u_{n-j+1} + \tau\beta_j f(t_{n-j+1}, u_{n-j+1})] + \\ &\quad + \frac{\tau\beta_0}{\alpha_0} f(t_{n+1}, y_{n+1}) - u_{n+1} = \\ &= \frac{\tau}{\alpha_0} \psi_n + \frac{\tau\beta_0}{\alpha_0} [f(t_{n+1}, y_{n+1}) - f(t_{n+1}, u_{n+1})]. \end{aligned}$$

За теоремою про скінченні прирости будемо мати

$$l_{n+1} = \frac{\tau}{\alpha_0} \psi_n + \frac{\tau\beta_0}{\alpha_0} f_u(t_{n+1}, u_{n+1} + \theta y_{n+1}) l_{n+1}.$$

Звідси випливає твердження леми. ■

Кажуть, що багатокроковий метод має p -й порядок апроксимації, якщо виконується одна з умов: $\psi_n = O(\tau^p)$ або $l_{n+1} = O(\tau^{p+1})$. Зауважимо, що згідно з лемою ці умови еквівалентні.

Дослідимо, як впливає вибір коефіцієнтів на похибку апроксимації багатокрокового методу (5.38). Розкладаючи функції $u_{n-j+1} = u(t_{n+1} - j\tau)$ в точці $t = t_{n+1}$ за формулою Тейлора, одержимо

$$u_{n-j+1} = u(t_{n+1} - j\tau) = \sum_{s=0}^p \frac{(-j\tau)^s u^{(s)}(t_{n+1})}{s!} + O(\tau^{p+1}),$$

$$\begin{aligned} f(t_{n-j+1}, u_{n-j+1}) &= u'(t_{n+1} - j\tau) = \\ &= \sum_{m=0}^{p-1} \frac{(-j\tau)^m u^{(m+1)}(t_{n+1})}{m!} + O(\tau^p) \\ &= \sum_{s=1}^p \frac{(-j\tau)^{s-1} u^{(s)}(t_{n+1})}{(s-1)!} + O(\tau^p), \quad j = \overline{0, k}. \end{aligned}$$

Підставляючи ці розклади у вираз для похибки апроксимації (5.40), будемо мати

$$\begin{aligned}\psi_n &= \frac{1}{\tau} \sum_{j=0}^k \left[-\alpha_j \sum_{s=0}^p \frac{(-j\tau)^s u^{(s)}(t_{n+1})}{s!} + \beta_j \tau \sum_{s=1}^p \frac{(-j\tau)^{s-1} u^{(s)}(t_{n+1})}{(s-1)!} \right] + \\ &+ O(\tau^p) = \\ &= -\frac{1}{\tau} \sum_{j=0}^k \alpha_j u(t_{n+1}) + \\ &+ \sum_{s=1}^p \frac{(-\tau)^{s-1} u^{(s)}(t_{n+1})}{s!} \left[\sum_{j=0}^k \alpha_j j^s + s \sum_{j=0}^k \beta_j j^{s-1} \right] + O(\tau^p).\end{aligned}$$

Звідси випливає, що похибка апроксимації має p -й порядок, якщо виконуються умови

$$\sum_{j=0}^k \alpha_j = 0, \quad \sum_{j=0}^k \alpha_j j^s = -s \sum_{j=0}^k \beta_j j^{s-1}, \quad s = \overline{1, p}. \quad (5.41)$$

Разом з умовою нормування (5.39) ми отримали систему $p + 2$ лінійних алгебраїчних рівнянь відносно $2(k+1)$ невідомих $\alpha_0, \alpha_1, \dots, \alpha_k, \beta_0, \beta_1, \dots, \beta_k$.

Для того, щоб ця система не була перевизначена, необхідно вимагати, щоб $p+2 \leq 2(k+1)$. Ця вимога означає, що порядок апроксимації лінійних k -крокових методів не може перевищувати $2k$. Отже, *найвищий досяжний порядок апроксимації* неявних k -крокових методів дорівнює $2k$, а явних — $2k-1$.

Явні методи Адамса дають точний розв'язок тоді, коли $f(t, u) \equiv f(t)$ є многочленом, степінь якого менший k . Тоді за допомогою явних методів Адамса одержуємо точні розв'язки диференціальних рівнянь

$$u' = -s(t_{n+1} - t)^{s-1}, \quad u(t_{n+1}) = 0, \quad s = \overline{0, k},$$

які мають вигляд $u(t) = (t_{n+1} - t)^s$, $s = \overline{0, k}$. Це означає, що похибка апроксимації дорівнює нулю, а отже враховуючи, що $u_{n-j+1} = (j\tau)^s$, $u'_{n-j+1} = -s(j\tau)^{s-1}$, з (5.40) отримаємо

$$\psi_n = -\tau^{s-1} \left(\sum_{j=0}^k \alpha_j j^s + s \sum_{j=0}^k \beta_j j^{s-1} \right) = 0, \quad s = \overline{0, k}.$$

Оскільки ця рівність збігається з рівністю (5.41) при $p = k$, то явні методи Адамса мають порядок не нижче k .

Неявні методи Адамса дають точний розв'язок, якщо $f(t, u) \equiv f(t)$ є многочленом на одиницю вищого степеня, ніж для явних, тому вони мають порядок не нижче $k + 1$.

Аналогічно можна показати, що формули диференціювання назад мають порядок не нижче k .

Приклад 5.1. Знайдіть головний член похибки апроксимації двокрокового методу

$$y_{n+1} + 4y_n - 5y_{n-1} = \tau(4f_n + 2f_{n-1}).$$

Доведіть, що метод має третій порядок апроксимації.

▷ Враховуючи співвідношення

$$\begin{aligned} u_{n\pm 1} &= u_n \pm \tau u'_n + \frac{\tau^2}{2} u''_n \pm \frac{\tau^3}{6} u'''_n + \frac{\tau^4}{24} u_n^{IV} + O(\tau^5), \\ f(t_n, u_n) &= u'_n, \\ f(t_{n-1}, u_{n-1}) &= u'_{n-1} = u'_n - \tau u''_n + \frac{\tau^2}{2} u'''_n - \frac{\tau^3}{6} u_n^{IV} + O(\tau^4) \end{aligned}$$

похибку апроксимації цього методу розкладемо у ряд Тейлора в околі точки $t = t_n$

$$\begin{aligned} \psi_n &= -\frac{u_{n+1} + 4u_n - 5u_{n-1}}{\tau} + 4f(t_n, u_n) + 2f(t_{n-1}, u_{n-1}) = \\ &= -\frac{1}{\tau} \left[5u_n + \tau u'_n + \frac{\tau^2}{2} u''_n + \frac{\tau^3}{6} u'''_n + \frac{\tau^4}{24} u_n^{IV} - \right. \\ &\quad \left. - 5 \left(u_n - \tau u'_n + \frac{\tau^2}{2} u''_n - \frac{\tau^3}{6} u'''_n + \frac{\tau^4}{24} u_n^{IV} \right) \right] + \\ &\quad + 4u'_n + 2 \left(u'_n - \tau u''_n + \frac{\tau^2}{2} u'''_n - \frac{\tau^3}{6} u_n^{IV} \right) + O(\tau^4) = \\ &= -6u'_n + 2\tau u''_n - \tau^2 u'''_n + \frac{\tau^3}{6} u_n^{IV} + 6u'_n - 2\tau u''_n + \\ &\quad + \tau^2 u'''_n - \frac{\tau^3}{3} u_n^{IV} + O(\tau^4) = \\ &= -\frac{\tau^3}{6} u_n^{IV} + O(\tau^4) = O(\tau^3). \end{aligned}$$



5.5.4. Стійкість багатокрокових методів

Високого порядку апроксимації та малої локальної похибки ще недостатньо для того, щоб багатокроковий метод був придатний для практичних розрахунків. Чисельний метод може бути чутливим до малих збурень початкових даних y_0, y_1, \dots, y_{k-1} , тобто бути нестійким. Це означає, що як завгодно малі зміни початкових даних можуть викликати як завгодно великі зміни розв'язку навіть при $\tau \rightarrow 0$ і фіксованому $n\tau$.

Розглянемо явний лінійний двокроковий метод третього порядку апроксимації вигляду

$$y_{n+1} + 4y_n - 5y_{n-1} = \tau(4f_n + 2f_{n-1}).$$

Застосуємо його до задачі

$$u' = u, \quad u(0) = 1,$$

тоді будемо мати

$$y_{n+1} + 4(1 - \tau)y_n - (5 + 2\tau)y_{n-1} = 0. \quad (5.42)$$

За початкові візьмемо значення точного розв'язку $y_0 = 1, y_1 = e^\tau$.

За аналогією з лінійними диференціальними рівняннями будемо шукати розв'язок рівняння (5.42) у вигляді $y_n = q^n$, де q — деяка невідома константа. У результаті одержимо характеристичне рівняння

$$q^2 + 4(1 - \tau)q - (5 + 2\tau) = 0. \quad (5.43)$$

Загальний розв'язок різницевого рівняння (5.42) обчислюється за формулою

$$y_n = c_1 q_1^n + c_2 q_2^n, \quad (5.44)$$

де $q_1(\tau) = 1 + \tau + O(\tau^2)$, $q_2(\tau) = -5 + O(\tau)$ — корені рівняння (5.43), а коефіцієнти c_1, c_2 визначаються з початкових умов. Оскільки $q_1(\tau)$ апроксимує e^τ , перший член в (5.44) апроксимує значення точного розв'язку $u(t) = e^t$ в точці $t = n\tau$. Збурення в розв'язок різницевого рівняння вносить другий член, який називають “паразитним”, тому що $|q_2(\tau)| > 1$ при $\tau \rightarrow 0$. Цей “паразитний” розв'язок з ростом n стає великим і починає переважати у розв'язку y_n .

Звернемося тепер до питання стійкості багатокрокового методу (5.38). Важливу роль відіграє поведінка розв'язку при $\tau \rightarrow 0$. Очевидно, що (5.38) при $\tau \rightarrow 0$ зводиться до формули

$$\alpha_0 y_{n+1} + \alpha_1 y_n + \dots + \alpha_k y_{n-k+1} = 0. \quad (5.45)$$

Її можна розглядати як чисельний метод (5.38), застосований до розв'язання диференціального рівняння

$$u' = 0.$$

Підставляючи в (5.45) $y_n = q^n$, одержимо *характеристичне рівняння*

$$\alpha_0 q^k + \alpha_1 q^{k-1} + \dots + \alpha_k = 0. \quad (5.46)$$

Багатокроковий метод (5.38) називається *стійким (нуль-стійким)*, якщо корені характеристичного рівняння (5.46) задовольняють *кореневу умову*, тобто всі корені за модулем не більші за одиницю, а серед коренів модуль яких дорівнює одиниці немає кратних.

Для явного і неявного методів Адамса характеристичне рівняння (5.46) має вигляд $q^k - q^{k-1} = 0$. Крім простого кореня, рівного 1, характеристичне рівняння має нульовий корінь кратності $k - 1$. Отже, методи Адамса — стійкі.

Можна показати, що k -крокова формула диференціювання назад стійка при $k \leq 6$ і нестійка при $k \geq 7$.

Оскільки методи Рунге–Кутта є однокроковими, то для них характеристичне рівняння має вигляд $q - 1 = 0$. Це рівняння має лише один корінь $q = 1$, а отже, методи Рунге–Кутта — стійкі.

Справджується твердження.

► **ТЕОРЕМА 5.1.** *Якщо багатокроковий метод (5.38) стійкий і має p -й порядок апроксимації, то він збіжний з p -м порядком точності.*

Доведення. Див., наприклад, [23]. ■

5.6. Методи Нордсіка

Усі лінійні багатокрокові методи чисельного інтегрування еквівалентні знаходженню многочлена, який апроксимує розв'язок $u(t)$. Ідея методів Нордсіка полягає у тому, щоб зобразити цей многочлен через похідні від нульового до k -го порядку включно, тобто з допомогою вектора

$$\mathbf{z}_n = \left(y_n, \tau y'_n, \frac{\tau^2}{2!} y''_n, \dots, \frac{\tau^k}{k!} y_n^{(k)} \right)^T$$

величини $y_n^{(j)}$ мають зміст наближених значень $u_n^{(j)}$.

Щоб визначити процедуру чисельного інтегрування, необхідно задати правило знаходження \mathbf{z}_{n+1} за відомим \mathbf{z}_n і диференціальним рівнянням (5.4). При використанні розкладу в ряд Тейлора (наприклад, при $k = 3$) таке правило має вигляд

$$\begin{aligned} y_{n+1} &= y_n + \tau y'_n + \frac{\tau^2}{2!} y''_n + \frac{\tau^3}{3!} y'''_n, \\ \tau y'_{n+1} &= \tau y'_n + 2 \frac{\tau^2}{2!} y''_n + 3 \frac{\tau^3}{3!} y'''_n, \\ \frac{\tau^2}{2!} y''_{n+1} &= \frac{\tau^2}{2!} y''_n + 3 \frac{\tau^3}{3!} y'''_n, \\ \frac{\tau^3}{3!} y'''_{n+1} &= \frac{\tau^3}{3!} y'''_n. \end{aligned} \quad (5.47)$$

У загальному випадку прогнозоване значення

$$\mathbf{z}_{n+1}^{(0)} = P \mathbf{z}_n, \quad (5.48)$$

де P — трикутна матриця Паскаля, елементи якої визначаються формулою

$$P_{ij} = \begin{cases} C_j^i, & 0 \leq i \leq j \leq k, \\ 0, & i > j. \end{cases}$$

Оскільки ця апроксимація не використовує рівняння (5.4), то рівність $\tau y'_{n+1} = \tau f(t_{n+1}, y_{n+1})$, взагалі кажучи, не буде виконуватися, тому коректуємо $\mathbf{z}_{n+1}^{(0)}$, додаючи до нього величину, кратну $\tau f(t_{n+1}, y_{n+1}) - \tau y'_{n+1}$, і будемо вимагати, щоб виконувалася рівність

$$\mathbf{z}_{n+1} = \mathbf{z}_{n+1}^{(0)} + \mathbf{l}(\tau f(t_{n+1}, y_{n+1}) - \tau y'_{n+1}), \quad (5.49)$$

де $\mathbf{l} = (l_0, l_1, \dots, l_k)^T$. Друга компонента вектора \mathbf{l} повинна бути рівна 1 ($l_1 = 1$), щоб виконувалася умова

$$\tau y'_{n+1} = \tau f(t_{n+1}, y_{n+1}).$$

Співвідношення (5.48) і (5.49) можуть бути об'єднані в одну формулу

$$\mathbf{z}_{n+1} = P \mathbf{z}_n + \mathbf{l}(\tau f(t_{n+1}, y_{n+1}) - \mathbf{e}_1^T P \mathbf{z}_n), \quad (5.50)$$

де $\mathbf{e}_1 = (0, 1, 0, \dots, 0)^T$. Процедuru (5.50) називають *процедурою Нордсіка*.

Дослідимо метод Нордсіка на стійкість. Для цього підставимо в (5.50) $f(t, u) \equiv 0$, тоді одержимо

$$\mathbf{z}_{n+1} = M\mathbf{z}_n, \quad M = P - \mathbf{l}\mathbf{e}_1^T P.$$

Наприклад, при $k = 3$ ця матриця має вигляд

$$M = \begin{pmatrix} 1 & 1 - l_0 & 1 - 2l_0 & 1 - 3l_0 \\ 0 & 0 & 0 & 0 \\ 0 & -l_2 & 1 - 2l_2 & 3 - 3l_2 \\ 0 & -l_3 & -2l_3 & 1 - 3l_3 \end{pmatrix}.$$

Для стійкості методу необхідно, щоб всі власні числа матриці M задовольняли кореневу умову. Зауважимо, що 1 і 0 є власними значеннями матриці M , а її характеристичний многочлен не залежить від l_0 . Величини l_2, l_3, \dots, l_k зручно вибрати таким чином, щоб всі решта власні значення M були рівні нулю. У випадку $k = 3$ це виконується при $l_2 = 3/4$, $l_3 = 1/6$. Коефіцієнт l_0 можна вибрати з умови перетворення в нуль константи похибки методу. У нашому випадку $l_0 = 3/8$, а весь метод задається вектором

$$\mathbf{l} = \left(\frac{3}{8}, 1, \frac{3}{4}, \frac{1}{6} \right)^T.$$

Зазначимо, що цей метод еквівалентний трикроковому методу Адамса. Дійсно, запишемо розклад (5.47) для точки t_{n-1}

$$y_{n-1} = y_n - \tau y'_n + \frac{\tau^2}{2!} y''_n - \frac{\tau^3}{3!} y'''_n,$$

тоді

$$\begin{aligned} y'_{n-1} &= y'_n - \tau y''_n + \frac{\tau^2}{2!} y'''_n, \\ y'_{n-2} &= y'_n - 2\tau y''_n + 4\frac{\tau^2}{2!} y'''_n. \end{aligned}$$

Розв'яжемо цю систему рівнянь відносно y''_n , y'''_n , тоді будемо мати

$$\begin{aligned} \frac{\tau^2}{2!} y''_n &= \tau \left(\frac{3}{4} y'_n - y'_{n-1} + \frac{1}{4} y'_{n-2} \right), \\ \frac{\tau^3}{3!} y'''_n &= \tau \left(\frac{1}{6} y'_n - \frac{2}{6} y'_{n-1} + \frac{1}{6} y'_{n-2} \right). \end{aligned} \tag{5.51}$$

З рівності (5.50) для першої компоненти вектора Нордсіка отримаємо

$$y_{n+1} = y_n + \frac{5}{8}\tau y'_n + \frac{1}{8}\tau^2 y''_n - \frac{1}{48}\tau^3 y'''_n + \frac{3}{8}\tau f_{n+1}.$$

Підставивши (5.51) в останню рівність, одержимо трикроковий неявний метод Адамса.

$$y_{n+1} = y_n + \frac{\tau}{24}(9f(t_{n+1}, y_{n+1}) + 19f(t_n, y_n) - \\ - 5f(t_{n-1}, y_{n-1}) + f(t_{n-2}, y_{n-2})).$$

Можна показати, що метод Нордсіка еквівалентний деякій багатокроковій формулі порядку не нижче k . За допомогою вектора Нордсіка зручно змінювати крок чисельного інтегрування (див. розділ 5.8).

5.7. Чисельне інтегрування жорстких систем звичайних диференціальних рівнянь

Чисельні методи розв'язання задачі Коші для звичайних диференціальних рівнянь покоординатно без будь-яких змін переносяться на системи звичайних диференціальних рівнянь.

Наприклад, лінійні багатокрокові методи у випадку систем рівнянь можна записати у вигляді

$$\sum_{j=0}^k \alpha_j \mathbf{y}_{n-j+1} = \tau \sum_{j=0}^k \beta_j \mathbf{f}_{n-j+1}, \quad (5.52)$$

де \mathbf{y}_n , \mathbf{f}_n — вектори наближень до розв'язку $\mathbf{u} = (u_1, u_2, \dots, u_N)^T$ і правої частини $\mathbf{f} = (f_1, f_2, \dots, f_N)^T$ системи диференціальних рівнянь у відповідних точках. Однак при чисельному розв'язанні систем звичайних диференціальних рівнянь можуть виникнути додаткові труднощі, пов'язані з жорсткістю системи.

5.7.1. Поняття жорсткої задачі

Суть явища жорсткості полягає в тому, що розв'язок, який потрібно обчислити, змінюється повільно, однак існують швидкі згасаючі збурення. Наявність таких збурень перешкоджає знаходженню чисельного розв'язку, який повільно змінюється.

Розглянемо приклад жорсткої задачі

$$u' = \lambda(u - e^{-t}) - e^{-t}, \quad u(0) = 0, \quad t \in [0, T], \quad \lambda \ll 0,$$

точний розв'язок якої має вигляд

$$u(t) = e^{-t} - e^{\lambda t}.$$

Цей розв'язок складається з двох компонент, одна з яких $e^{\lambda t}$ змінюється значно швидше (оскільки $\lambda \ll 0$, наприклад $\lambda = -10^3$), ніж друга e^{-t} . Після дуже малого відрізка часу $[0, \theta]$, де $\theta = -1/\lambda$ — найменша часова константа, швидка (жорстка) компонента прямує до нуля і суттєвий вклад у розв'язок робить повільна (гладка) компонента.

Для виявлення деяких особливостей чисельних методів застосуємо до цієї задачі явний та неявний методи Ейлера:

$$\begin{aligned} y_{n+1} &= y_n + \tau f(t_n, y_n), \\ y_{n+1} &= y_n + \tau f(t_{n+1}, y_{n+1}). \end{aligned}$$

Тоді для явного методу Ейлера будемо мати

$$y_{n+1} = (1 + \tau\lambda)(y_n - e^{-t_n}) + e^{-t_n} - \tau e^{-t_n},$$

а для неявного

$$\begin{aligned} y_{n+1} &= (1 - \tau\lambda)^{-1}(y_n - e^{-t_n}) + \\ &+ (1 - \tau\lambda)^{-1}(e^{-t_n} - \tau e^{-t_{n+1}} - \tau\lambda e^{-t_{n+1}}). \end{aligned}$$

Одержані чисельні розв'язки порівняємо з точним

$$u_{n+1} = e^{-t_{n+1}} - e^{\lambda t_{n+1}} = e^{\lambda\tau}(u_n - e^{-t_n}) + e^{-t_{n+1}}.$$

Вираз $e^{\lambda\tau}(u_n - e^{-t_n})$ можна інтерпретувати як збурення повільного розв'язку e^{-t} в момент $t = t_{n+1}$. Це збурення швидко спадає завдяки $e^{\lambda\tau}$, де $\tau\lambda \ll 0$, тобто $u_{n+1} \rightarrow e^{-t_{n+1}}$ при $\tau\lambda \rightarrow -\infty$. Чисельний метод повинен бути в змозі зменшувати різницю $y_n - e^{-t_n}$ для значень $\tau\lambda \ll 0$. Різниця $y_n - e^{-t_n}$ у випадку неявного методу Ейлера швидко спадає і для будь-якого фіксованого $\tau > 0$, $y_{n+1} \rightarrow e^{-t_{n+1}}$ при $\tau\lambda \rightarrow -\infty$. Для явного методу різниця $y_n - e^{-t_n}$ буде зменшуватися тільки у випадку $|1 + \lambda\tau| \leq 1$, тобто при $\tau < -2/\lambda$. Ця умова накладає сильне обмеження на величину кроку τ при $\lambda \ll 0$. З іншого боку, якщо $y_n \approx e^{-t_n}$,

то $e^{-t_{n+1}}$ апроксимується виразом $e^{-t_n} - \tau e^{-t_n}$ досить точно для значно більших τ , ніж ті, що задовольняють умову $\tau < -2/\lambda$. Отже, величина кроку обмежується чисельною стійкістю, а не точністю.

Для лінійних систем диференціальних рівнянь зі сталими коефіцієнтами подібні компоненти розв'язку, які сильно відрізняються, виникають, коли матриця системи містить сильно розкидані власні значення. Розглянемо, наприклад, систему

$$\begin{cases} u_1' = 998u_1 + 1998u_2 \\ u_2' = -999u_1 - 1999u_2 \end{cases}$$

з початковими умовами $u_1(0) = u_2(0) = 1$. Власні значення матриці коефіцієнтів цієї системи

$$\begin{pmatrix} 998 & 1998 \\ -999 & -1999 \end{pmatrix}$$

дорівнюють: $\lambda_1 = -1$, $\lambda_2 = -10^3$. Тоді розв'язок буде мати вигляд

$$\begin{aligned} u_1(t) &= 4e^{-t} - 3e^{-10^3t}, \\ u_2(t) &= -2e^{-t} + 3e^{-10^3t}, \end{aligned}$$

де e^{-10^3t} — швидка компонента розв'язку, а e^{-t} — повільна.

Розглянемо систему нелінійних звичайних диференціальних рівнянь (5.2). Нехай $\mathbf{u}(t)$ і $\mathbf{v}(t)$ — два різні розв'язки системи. Утворимо різницю $\mathbf{z}(t) = \mathbf{u}(t) - \mathbf{v}(t)$, яка задовольняє рівняння

$$\mathbf{z}'_m = f_m(t, \mathbf{v} + \mathbf{z}) - f_m(t, \mathbf{v}), \quad m = \overline{1, N}. \quad (5.53)$$

Проведемо розклад правої частини (5.53) за формулою Тейлора. Оскільки

$$f_m(t, \mathbf{u}) = f_m(t, u_1, u_2, \dots, u_N),$$

то маємо

$$f_m(t, \mathbf{v} + \mathbf{z}) - f_m(t, \mathbf{v}) = \sum_{j=1}^N \frac{\partial f_m(t, \mathbf{v})}{\partial u_j} z_j(t) + o(\|\mathbf{z}\|).$$

У результаті розкладу система (5.53) набуває вигляду

$$\mathbf{z}' = J(t, \mathbf{v}(t))\mathbf{z}(t) + o(\|\mathbf{z}\|), \quad (5.54)$$

де через $J(t, \mathbf{v}(t)) = \partial \mathbf{f}(t, \mathbf{v}(t)) / \partial \mathbf{u}$ позначено матрицю Якобі з елементами

$$J_{ij}(t, \mathbf{v}(t)) = \frac{\partial f_i(t, \mathbf{v}(t))}{\partial u_j}, \quad i, j = \overline{1, N}.$$

Знехтувавши в (5.54) величиною $o(\|\mathbf{z}\|)$, одержимо так звану систему рівнянь першого наближення

$$\mathbf{w}' = J(t, \mathbf{v}(t))\mathbf{w},$$

яка є системою лінійних диференціальних рівнянь відносно $\mathbf{w}(t)$, оскільки функція $\mathbf{v}(t)$ задана. Припустимо, що в околі точки $(\tilde{t}, \mathbf{v}(\tilde{t}))$ матриця $J(t, \mathbf{v}(t))$ змінюється достатньо мало, тоді одержимо систему диференціальних рівнянь зі сталими коефіцієнтами

$$\mathbf{w}' = J(\tilde{t}, \mathbf{v}(\tilde{t}))\mathbf{w}, \quad t \geq \tilde{t}.$$

Припустимо, що $\lambda_j, j = \overline{1, N}$ — власні числа матриці $J(\tilde{t}, \mathbf{v}(\tilde{t}))$, які в загальному випадку є комплексними числами. Задача (5.1) називається *жорсткою* в околі точки \tilde{t} , якщо:

- 1) $\operatorname{Re} \lambda_j < 0, j = \overline{1, N}$;
- 2) $\frac{\max_{1 \leq j \leq N} |\operatorname{Re} \lambda_j|}{\min_{1 \leq j \leq N} |\operatorname{Re} \lambda_j|} \gg 1$.

Нехай λ — власне число з найбільшою за модулем негативною дійсною частиною, тоді величина $\theta = -1/\operatorname{Re}(\lambda)$ буде найменшою часовою константою. Мірою жорсткості задачі є величина

$$s = (T - t_0)/\theta, \quad (5.55)$$

яка називається *коефіцієнтом жорсткості*. Якщо $s \geq 10^3$, то система буде сильно жорсткою, якщо $s < 10$, тоді система нежорстка.

На практиці жорсткі задачі розпізнають, виходячи з фізичних міркувань. Будь-яка фізична система, яка моделюється звичайними диференціальними рівняннями і має фізичні компоненти з часовими константами, які сильно відрізняються, зводиться до жорсткої задачі. Фізичні компоненти з найменшими часовими сталими дуже швидко змінюються і роблять задачу жорсткою, тоді як дослідник проявляє найбільший інтерес до компонент з великими часовими константами. У жорсткій задачі розв'язки, які змінюються повільно, визначаються саме цими останніми компонентами. Існуючі методи чисельного розв'язання жорстких

задач дозволяють інтегрування з розмірами кроків такого ж порядку, як і великі часові константи. Класичні явні методи можуть застосовуватися тільки з розмірами кроків порядку найменших часових констант і тому є надзвичайно неефективними.

Однією з властивостей жорстких задач є наявність великої константи Ліпшиця

$$L(t) = \sup_{\mathbf{u}} \left\| \frac{\partial \mathbf{f}(t, \mathbf{u})}{\partial \mathbf{u}} \right\| \gg 0. \quad (5.56)$$

Якщо вважати, що часовий масштаб задачі зв'язаний з розв'язком, який змінюється повільно на інтервалі $[t_0, T]$, то вираз (5.56) можна замінити на більш конкретну умову $\tau L(t) \gg 1$.

5.7.2. Абсолютна стійкість чисельних методів

Для дослідження стійкості чисельних методів інтегрування жорстких систем диференціальних рівнянь, розглянемо лінійну систему зі сталими коефіцієнтами

$$\mathbf{u}' = A\mathbf{u}. \quad (5.57)$$

Застосовуючи лінійний багатокроковий метод (5.52) до системи (5.57), одержимо

$$\sum_{j=0}^k \alpha_j \mathbf{y}_{n-j+1} = \tau \sum_{j=0}^k \beta_j A \mathbf{y}_{n-j+1}. \quad (5.58)$$

Припустимо, що матриця A має N різних власних значень $\lambda_1, \lambda_2, \dots, \lambda_N$, тоді існує невироджена матриця H така, що $H^{-1}AH = \Lambda$, де Λ — діагональна матриця з елементами $\lambda_1, \lambda_2, \dots, \lambda_N$. Якщо покласти $\mathbf{u} = H\mathbf{v}$ і помножити систему (5.57) зліва на H^{-1} , то одержимо

$$\mathbf{v}' = H^{-1}AH\mathbf{v}$$

або

$$v'_m = \lambda_m v_m, \quad m = \overline{1, N}. \quad (5.59)$$

Зробимо аналогічну заміну в (5.58), тоді

$$\sum_{j=0}^k \alpha_j v_{m,n-j+1} - \tau \sum_{j=0}^k \beta_j \lambda_m v_{m,n-j+1} = 0. \quad (5.60)$$

Співвідношення (5.60) збігається зі скінченно-різницевою апроксимацією лінійним багатокроковим методом (5.52) рівнянь (5.59). Аналогічні висновки можна зробити відносно методів Рунге–Кутта. Отже, стійкість чисельних методів будемо досліджувати на скалярному модельному рівнянні

$$u' = \lambda u, \quad (5.61)$$

де λ — комплексне число, оскільки власні значення матриці A , взагалі кажучи, комплексні числа.

Багатокроковий метод, застосований до модельного рівняння (5.61), має вигляд

$$\sum_{j=0}^k (\alpha_j - z\beta_j) y_{n-j+1} = 0,$$

де $z = \tau\lambda$ — комплексний параметр. Якщо розв'язок цього рівняння шукати у вигляді $y_n = q^n$, то одержимо *характеристичне рівняння*

$$\sum_{j=0}^k (\alpha_j - z\beta_j) q^{k-j} = 0. \quad (5.62)$$

Зауважимо, що якщо однокроковий метод застосовувати до модельного рівняння (5.61), то $y_{n+1} = R(z)y_n$, де $R(z)$ називається *функцією стійкості* однокрокового методу. Характеристичне рівняння для однокрокового методу має вигляд $q = R(z)$.

Чисельний метод називається *абсолютно стійким* (стійким при будь-яких $\tau > 0$) для $z = \tau\lambda$, якщо для цього $z = \tau\lambda$ корені характеристичного рівняння (5.62) задовольняють кореневу умову, тобто всі корені за модулем не більші за одиницю, а серед коренів модуль яких рівний одиниці немає кратних. *Областю абсолютної стійкості* називається множина всіх точок комплексної площини $z = \tau\lambda$, для яких даний метод є абсолютно стійким.

Однокроковий метод буде *абсолютно стійким*, якщо $|R(z)| \leq 1$.

Метод називається *A-стійким*, якщо його область абсолютної стійкості включає всю півплощину $\operatorname{Re} z \leq 0$. Суть наведеного означення полягає в тому, що A-стійкий чисельний метод є абсолютно стійким, якщо стійким є розв'язок вихідного диференціального рівняння.

Явний метод Ейлера, застосований до модельного рівняння (5.61), має вигляд

$$y_{n+1} = (1 + z)y_n.$$

Вимога абсолютної стійкості приводить до оцінки

$$|1 + z| \leq 1$$

або

$$(1 + \omega_1)^2 + \omega_2^2 \leq 1,$$

де $\omega_1 = \operatorname{Re} z$, $\omega_2 = \operatorname{Im} z$. Область абсолютної стійкості в цьому випадку буде всередині круга одиничного радіуса з центром у точці $(-1, 0)$.

Для неявного методу Ейлера

$$y_{n+1} = y_n + \tau f(t_{n+1}, y_{n+1})$$

область абсолютної стійкості — зовні круга одиничного радіуса з центром в точці $(1, 0)$.

Отже, неявний метод Ейлера є A -стійкий, а явний — не A -стійкий.

Розглянемо формулу трапецій

$$y_{n+1} = y_n + \frac{\tau}{2}(f(t_{n+1}, y_{n+1}) + f(t_n, y_n)).$$

Для рівняння (5.61) цей метод має вигляд

$$y_{n+1} = R(z)y_n, \quad R(z) = \frac{1 + z/2}{1 - z/2}.$$

Звідси випливає, що $|R(z)| \leq 1$ тоді і тільки тоді, коли $\operatorname{Re} z \leq 0$. Отже, формула трапецій теж A -стійка.

Для знаходження області стійкості лінійного багатокрокового методу виразимо з рівняння (5.62) параметр z через змінну q , тоді дістанемо

$$z(q) = \frac{\sum_{j=0}^k \alpha_j q^{k-j}}{\sum_{j=0}^k \beta_j q^{k-j}}. \quad (5.63)$$

Покладемо q у рівнянні (5.63) таким, що приймає всі значення на одиничному колі $|q| = 1$ і обчислимо відповідні значення $z(q)$. Оскільки будь-яке q на одиничному колі можна записати у вигляді $q = e^{i\varphi}$, $i = \sqrt{-1}$, $0 \leq \varphi \leq 2\pi$, то $e^{i\varphi}$ можна підставити замість q у рівняння (5.63), що приведе до

$$z(q) = \frac{\sum_{j=0}^k \alpha_j e^{i(k-j)\varphi}}{\sum_{j=0}^k \beta_j e^{i(k-j)\varphi}}. \quad (5.64)$$

Множина точок, породжених (5.64), є геометричним місцем точок Γ , для яких $|q| = 1$. Оскільки коло є замкненою кривою, то Γ також замкнена. Геометричне місце точок Γ є симетричним відносно осі $\operatorname{Re} z$, тому достатньо обчислити $z(q)$ тільки для $0 \leq \varphi \leq \pi$.

При розв'язуванні жорстких систем диференціальних рівнянь було б бажано використовувати саме A -стійкі методи, тому що умови їх стійкості не накладають обмежень на крок τ . Однак серед явних методів (Рунге–Кутта, багатокрокових) не існує A -стійких.

Дальквіст довів, що не існує A -стійкого неявного лінійного багатокрокового методу порядку апроксимації вище другого.

У зв'язку з цим було введено ще одне означення стійкості.

Чисельний метод називається $A(\alpha)$ -стійким, якщо область його стійкості містить кут $|\arg(-z)| < \alpha$ (рис. 5.1).

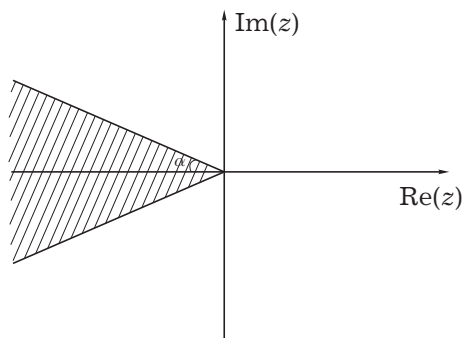


Рис. 5.1.

Зокрема, $A(\pi/2)$ -стійкість збігається з A -стійкістю.

Доведено, що ні для якого α не існує явного $A(\alpha)$ -стійкого методу.

Найбільш поширеним класом методів розв'язування жорстких задач є формули диференціювання назад, області стійкості яких зображені на рис. 5.2 (області стійкості розташовані зовні відповідних кривих Γ).

Формули диференціювання назад від 1-го до 6-го порядків апроксимації — $A(\alpha)$ -стійкі, причому чим більший порядок, тим менше α .

Приклад 5.2. При яких значеннях параметра $\sigma \geq 0$ чисельний метод

$$y_{n+1} = y_n + \tau(\sigma f(t_{n+1}, y_{n+1}) + (1 - \sigma)f(t_n, y_n))$$

буде A -стійким?

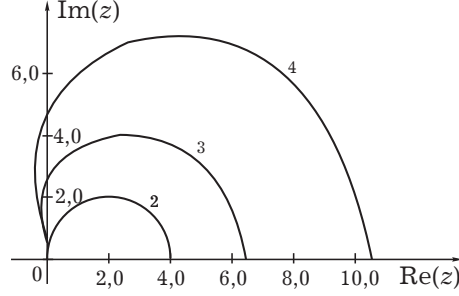


Рис. 5.2.

► Застосуємо цей метод до модельного рівняння (5.61), тоді одержимо

$$y_{n+1} = R(z)y_n,$$

де

$$R(z) = \frac{1 + (1 - \sigma)z}{1 - \sigma z}, \quad z = \tau\lambda$$

є аналітичною функцією в лівій півплощині комплексної площини z . Згідно з теоремою про максимум модуля аналітичної функції, максимум $|R(z)|$ досягається на границі області, тобто на уявній осі $\text{Im } z$. Отже, умова A -стійкості $|R(z)| \leq 1 \quad \forall \text{Re } z \leq 0$ буде виконуватися, якщо $|R(i\omega)| \leq 1 \quad \forall \omega \in \mathbb{R}^1$. Остання умова для заданої функції $R(z)$ буде мати вигляд $|1 + i(1 - \sigma)\omega| \leq |1 - i\sigma\omega|$. Звідси $\omega^2(1 - 2\sigma) \leq 0$. Отже, чисельний метод A -стійкий за умови $\sigma \geq 1/2$. ◀

5.8. Реалізація лінійних неявних багатокрокових методів

Для обчислення y_{n+1} за формулою (5.52) необхідно розв'язати систему N нелінійних алгебраїчних рівнянь з N невідомими вигляду

$$y_{n+1} = \frac{\beta_0}{\alpha_0} \tau \mathbf{f}(t_{n+1}, y_{n+1}) + \varphi_n, \quad (5.65)$$

де

$$\varphi_n = -\frac{1}{\alpha_0} \sum_{j=1}^k \alpha_j y_{n-j+1} + \frac{\tau}{\alpha_0} \sum_{j=1}^k \beta_j \mathbf{f}_{n-j+1}.$$

У найпростішому випадку для розв'язування системи (5.65) можна використати метод простої ітерації

$$y_{n+1}^{(s+1)} = \frac{\beta_0}{\alpha_0} \tau \mathbf{f}(t_{n+1}, y_{n+1}^{(s)}) + \varphi_n, \quad s = \overline{0, M-1}, \quad (5.66)$$

де $\mathbf{y}_{n+1}^{(0)}$ — задане початкове наближення до \mathbf{y}_{n+1} , для обчислення якого можна використати явні багатокрокові методи, s — номер ітерації. Цей ітераційний процес ефективний тільки у випадку нежорстких задач. Для жорстких задач він збіжний при τ менших, ніж найменша часова константа системи диференціальних рівнянь.

У випадку жорсткої задачі систему (5.65) розв'язують за допомогою методу Ньютона. Запишемо (5.65) у вигляді

$$\mathbf{F}(\mathbf{y}) = \mathbf{y} - \frac{\beta_0}{\alpha_0} \tau \mathbf{f}(t_{n+1}, \mathbf{y}) - \boldsymbol{\varphi}_n = 0.$$

Тоді ітераційний метод Ньютона буде мати вигляд

$$\mathbf{y}_{n+1}^{(s+1)} = \mathbf{y}_{n+1}^{(s)} - \left[\frac{\partial \mathbf{F}(\mathbf{y}_{n+1}^{(s)})}{\partial \mathbf{y}} \right]^{-1} \mathbf{F}(\mathbf{y}_{n+1}^{(s)})$$

або

$$\left(I - \frac{\beta_0}{\alpha_0} \tau J(t_{n+1}, \mathbf{y}_{n+1}^{(s)}) \right) \mathbf{z}_{n+1}^{(s)} = -\mathbf{F}(\mathbf{y}_{n+1}^{(s)}),$$

$$\mathbf{y}_{n+1}^{(s+1)} = \mathbf{y}_{n+1}^{(s)} + \mathbf{z}_{n+1}^{(s)},$$

де I — одинична матриця порядку N , $J = \partial \mathbf{f} / \partial \mathbf{u}$ — матриця Якобі правої частини системи, $s = \overline{0, M-1}$. Обчислення $\mathbf{y}_{n+1}^{(s+1)}$ за цією формулою вимагає значних затрат, пов'язаних з необхідністю обчислення матриці Якобі, і розв'язанням системи N лінійних алгебраїчних рівнянь з N невідомими для кожного s .

Ці затрати можуть бути зменшені за рахунок застосування модифікованого методу Ньютона і за рахунок врахування структури матриці Якобі. Для обчислення $\mathbf{y}_{n+1}^{(s+1)}$ з допомогою модифікованого методу Ньютона необхідно розв'язати систему лінійних алгебраїчних рівнянь

$$\left(I - \frac{\beta_0}{\alpha_0} \tau J(t_{n+1}, \mathbf{y}_{n+1}^{(0)}) \right) \mathbf{z}_{n+1}^{(s)} = -\mathbf{F}(\mathbf{y}_{n+1}^{(s)}),$$

$$\mathbf{y}_{n+1}^{(s+1)} = \mathbf{y}_{n+1}^{(s)} + \mathbf{z}_{n+1}^{(s)}, \quad s = \overline{0, M-1}.$$

Щоб розв'язати M таких систем лінійних рівнянь методом Гаусса необхідно один раз розкласти матрицю

$$P = I - \frac{\beta_0}{\alpha_0} \tau J(t_{n+1}, \mathbf{y}_{n+1}^{(0)})$$

у вигляді $P = LU$ (L — ліва трикутна матриця, U — права трикутна матриця), що відповідає прямому ходу методу Гаусса і розв'язати $2M$ систем з трикутними матрицями, що відповідає оберненому ходу методу Гаусса. Матрицю Якобі $J(t_{n+1}, \mathbf{y}_{n+1}^{(0)})$ можна обчислювати як чисельно, так і аналітично. Величину M часто вибирають рівною 3.

Ефективна програма чисельного інтегрування повинна передбачати зміну величини кроку. Однак у випадку багатокрових методів процедура зміни кроку викликає ускладнення, оскільки ці методи вимагають значень чисельного розв'язку в точках, які розташовані на однаковій віддалі. Існують дві можливості виходу з цього:

- 1) визначення з допомогою інтерполяції многочленами значень у точках, які необхідні при зміні кроку;
- 2) побудова багатокрових методів для нерівномірної сітки (див., напр., [23]).

Розглянемо детальніше перший підхід. Лінійні багатокрокові методи можна використовувати у формі з розділеними різницями (див. розділи 5.5.1, 5.5.2) або за допомогою зображення Нордсіка $Z_n = \left(\mathbf{y}_n, \frac{\tau}{1!} \mathbf{y}'_n, \frac{\tau^2}{2!} \mathbf{y}''_n, \dots, \frac{\tau^k}{k!} \mathbf{y}_n^{(k)} \right)^T$, кожна компонента $\mathbf{y}_n^{(j)}$ якого апроксимує відповідну похідну точного розв'язку (див. розділ 5.6). Довільна зміна кроку у зображенні Нордсіка досягається множенням i -го коефіцієнта на θ^i , де $\theta = \tau_H/\tau$, τ_H — новий крок, тобто $Z_{n+1} = Z_n \text{diag}(1, \theta, \theta^2, \dots, \theta^k)$. Однак, у цьому випадку при зміні порядку необхідно обчислити нове зображення Нордсіка. У формі з розділеними різницями зручніше змінювати порядок методу, але важче змінювати крок.

При застосуванні лінійних багатокрових методів важливим є питання контролю похибки, вибору кроку чисельного інтегрування, порядку методу, який є найбільш точним і ефективним у даній точці t_{n+1} .

Головний член локальної похибки лінійного багатокрового методу порядку апроксимації p має вигляд

$$\mathbf{l}_{n+1}^{[p]} = C_p \tau^{p+1} \mathbf{u}^{(p+1)}(t_n) + O(\tau^{p+2}),$$

де C_p — константа, яка залежить від порядку p .

Припустимо, що до точки t_n чисельне інтегрування проходило успішно, а для знаходження \mathbf{y}_{n+1} ми вибрали крок τ і порядок p . Щоб вирішити, чи придатне \mathbf{y}_{n+1} , необхідно оцінити локальну похибку. Для практичної оцінки локальної похибки методу порядку p можна викори-

стати, наприклад, формулу

$$\mathbf{l}_{n+1}^{[p]} = \mathbf{y}_{n+1} - \mathbf{y}_{n+1}^*, \quad (5.67)$$

де \mathbf{y}_{n+1}^* — розв'язок, одержаний методом порядку $p+1$. На основі формули (5.67) обчислюється похибка

$$E_p = \frac{\|\mathbf{l}_{n+1}^{[p]}\|}{d},$$

де d — масштабний множник. Для обчислення абсолютної похибки $d = 1$, а для відносної $d = \|\mathbf{y}_{n+1}^*\|$. Можна використати змішане масштабування (див. розділ 5.4).

Далі необхідно вибрати новий крок і порядок. Ідея вибору кроку полягає в знаходженні найбільшого τ_H , при якому похибка E_p не перевищує ε . Похибку E_p в точці t_{n+2} можна наблизити її значенням в точці t_{n+1} , тоді крок $\tau_H = \theta\tau$ вибирається з умови

$$\frac{\|C_{p+1}\tau_H^{p+1}\mathbf{u}^{(p+1)}(t_n)\|}{d} \approx \theta^{p+1}E_p = \varepsilon.$$

$$\theta = \left(\frac{\varepsilon}{E_p}\right)^{\frac{1}{p+1}}, \quad \tau_H^{[p]} = \frac{1}{1,2} \left(\frac{\varepsilon}{E_p}\right)^{\frac{1}{p+1}} \tau.$$

Коефіцієнт $\frac{1}{1,2}$ дозволяє врахувати вплив величини $O(\tau^{p+2})$. Для вибору оптимального порядку вибирають кроки

$$\tau_H^{[p+1]} = \frac{1}{1,4} \left(\frac{\varepsilon}{E_{p+1}}\right)^{\frac{1}{p+2}} \tau, \quad \tau_H^{[p-1]} = \frac{1}{1,3} \left(\frac{\varepsilon}{E_{p-1}}\right)^{\frac{1}{p}} \tau,$$

де

$$E_{p+1} = \frac{\|\mathbf{l}_{n+1}^{[p+1]}\|}{d}, \quad E_{p-1} = \frac{\|\mathbf{l}_{n+1}^{[p-1]}\|}{d}.$$

Тоді новий крок $\tau_H = \max(\tau_H^{[p]}, \tau_H^{[p-1]}, \tau_H^{[p+1]})$, а порядок $p_H = q$, де $\tau_H = \tau_H^{[q]}$, $q = p-1, p, p+1$.

Необхідно зазначити, що в кожній з реалізацій багатокрокових методів описаний алгоритм, удосконалений і доповнений ще рядом процедур. Наприклад, крок зберігається сталим, якщо відношення τ_H/τ близьке до 1, що спрощує обчислення.

Контрольні завдання

✎ 5.1. Який зв'язок існує між локальною і глобальною похибками чисельного методу?

✎ 5.2. Запишіть метод Рунге–Кутта четвертого порядку (5.21) для системи (5.1) у векторній формі.

✎ 5.3. Розв'яжіть методом Ейлера задачу Коші:

$$u' = u, \quad u(0) = 1$$

на відрізку $[0, 1]$, якщо $\tau = 1/N$. Доведіть, що

$$|u(1) - y_N| \geq e/(2N + 2).$$

✎ 5.4. Побудуйте всі методи Рунге–Кутта другого порядку апроксимації вигляду

$$\begin{array}{ccc} 0 & & \\ c_2 & c_2 & \\ c_3 & 0 & c_3 \\ & 0 & 0 & 1 \end{array}$$

✎ 5.5. Доведіть, що метод Рунге–Кутта (5.12), (5.13) можна інтерпретувати як два кроки методу Ейлера (довжини $\tau/2$), які супроводжуються екстраполяцією Річардсона (5.25).

✎ 5.6. Побудуйте вкладені формули Рунге–Кутта 1-го та 2-го порядків апроксимації при $s = 2, 3$, які задовольняють умову $a_{si} = b_i$, $i = \overline{1, s-1}$.

✎ 5.7. Доведіть, що метод Рунге–Кутта (5.21) має четвертий порядок апроксимації.

✎ 5.8. Порівняйте однокрокові і багатокрокові методи розв'язування задачі Коші.

✎ 5.9. Визначте порядок апроксимації методу

$$y_{n+1} = y_{n-1} + \frac{\tau}{2}(f_{n+1} + 2f_n + f_{n-1}).$$

🖋 5.10. Дослідіть на стійкість метод Мілна

$$y_{n+1} = y_{n-3} + \frac{4\tau}{3}(2f_n - f_{n-1} + 2f_{n-2}).$$

🖋 5.11. Визначте порядок апроксимації та дослідіть на стійкість двокроковий метод

$$y_{n+1} = \frac{1}{2}y_n + \frac{1}{2}y_{n-1} + \frac{3\tau}{4}(3f_n - f_{n-1}).$$

🖋 5.12. Чи буде жорсткою задача

$$u'' + 101u' + 100u = 0, \quad t \in [0, 1],$$

$$u(0) = 1,01, \quad u'(0) = -2 ?$$

Знайдіть коефіцієнт жорсткості.

🖋 5.13. Знайдіть та побудуйте область абсолютної стійкості чисельного методу

$$y_{n+1} = y_n + \tau f\left(t_n + \frac{\tau}{2}, \frac{y_{n+1} + y_n}{2}\right).$$

Чи буде метод A -стійким?

🖋 5.14. Запишіть розрахункові формули алгоритму обчислення чисельного розв'язку для неявного методу Ейлера

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \tau \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}), \quad \mathbf{y}_0 = \mathbf{u}(t_0)$$

для випадку жорсткої задачі.

РОЗДІЛ 6

ЧИСЕЛЬНЕ РОЗВ'ЯЗУВАННЯ КРАЙОВИХ ЗАДАЧ ДЛЯ ЗВИЧАЙНИХ ДИФЕРЕНЦІАЛЬНИХ РІВНЯНЬ

6.1. Крайові задачі для звичайних диференціальних рівнянь

Крайова задача полягає у знаходженні розв'язку системи звичайних диференціальних рівнянь першого порядку

$$\mathbf{u}'(x) = \mathbf{f}(x, \mathbf{u}(x)),$$

де

$$\mathbf{u} = (u_1, u_2, \dots, u_n)^T, \quad \mathbf{f} = (f_1, f_2, \dots, f_n)^T,$$

на відрізку $a \leq x \leq b$, який задовольняє граничні (крайові) умови. У загальному випадку граничні умови мають вигляд нелінійних співвідношень між значеннями $u_m(x)$, $m = \overline{1, n}$ у двох або більше точках цього відрізка. Якщо граничні умови задаються лише на кінцях відрізка $[a, b]$, то крайова задача називається *двоточною*. Відомо, що диференціальне рівняння n -го порядку можна звести до системи n рівнянь першого порядку. Тому часто розглядають крайові задачі для рівнянь n -го порядку.

Розглянемо диференціальне рівняння другого порядку:

$$F(x, y, y', y'') = 0. \tag{6.1}$$

Двоточкова крайова задача для рівняння (6.1) — це задача знаходження функції $u(x)$, яка задовольняє рівняння (6.1) всередині відрізка $[a, b]$, а на кінцях — крайові умови

$$\begin{aligned} \varphi_1[y(a), y'(a)] &= 0; \\ \varphi_2[y(b), y'(b)] &= 0. \end{aligned} \tag{6.2}$$

Якщо диференціальне рівняння і крайові умови — лінійні, то крайова задача — лінійна.

Чисельні методи розв'язування крайових задач для звичайних диференціальних рівнянь поділяють на два основні типи: *методи зведення крайових задач до послідовності задач Коші*, до яких належить *метод стрільби* та *сіткові методи*, до яких належать *методи скінченних різниць* та *скінченних елементів*.

6.2. Метод стрільби

Основну ідею методу стрільби розглянемо на прикладі нелінійної крайової задачі

$$u'' = f(x, u, u'), \quad a < x < b, \quad (6.3)$$

$$u(a) = \mu_1, \quad u(b) = \mu_2. \quad (6.4)$$

Розв'язок рівняння (6.3) з початковими умовами $u(a) = \mu_1$, $u'(a) = s$ позначимо через $u(x; s)$. Тоді, якщо ми знайдемо потрібний *кут стрільби* $s = s^*$ такий, що $u(b, s^*) = \mu_2$, то розв'язок задачі Коші буде також розв'язком крайової задачі (6.3), (6.4). Таким чином, отримаємо нелінійне алгебраїчне рівняння, розв'язок якого $s = s^*$ задовольняє рівняння $u(b, s) = \mu_2$. Це нелінійне рівняння можна розв'язати за допомогою ітераційних методів (див. розділ 2.1). Розв'язавши задачу Коші ще раз з $s = s^*$ отримаємо розв'язок крайової задачі (6.3), (6.4).

6.2.1. Метод однократної стрільби

Розглянемо лінійну крайову задачу

$$\mathbf{u}' = A(x)\mathbf{u} + \mathbf{f}(x), \quad a < x < b, \quad (6.5)$$

$$B_a\mathbf{u}(a) + B_b\mathbf{u}(b) = \mathbf{d}, \quad (6.6)$$

де $\mathbf{u}(x)$, $\mathbf{f}(x)$, \mathbf{d} — вектори розміру n , $A(x)$, B_a , B_b — матриці розміру $n \times n$. Надалі будемо припускати, що $A(x)$, $\mathbf{f}(x)$ неперервні на $[a, b]$ та $\text{rank}[B_a, B_b] = n$.

Загальний розв'язок рівняння (6.5) може бути записаний як

$$\mathbf{u}(x) = U(x)\mathbf{s} + \mathbf{v}(x), \quad a \leq x \leq b, \quad (6.7)$$

де \mathbf{s} — вектор-параметр розміру n , $\mathbf{v}(x)$ — частинний розв'язок рівняння (6.5), $U(x)$ — фундаментальний розв'язок, тобто розв'язок матричної

задачі Коші

$$U' = A(x)U, \quad a < x \leq b, \quad (6.8)$$

$$U(a) = I. \quad (6.9)$$

Дійсно, якщо функцію (6.7) підставити в (6.5), то отримаємо тотожність

$$(U' - A(x)U) \mathbf{s} + \mathbf{v}' - A(x)\mathbf{v} - \mathbf{f}(x) \equiv \mathbf{0}.$$

Частинний розв'язок $\mathbf{v}(x)$ може бути визначений як розв'язок задачі Коші

$$\mathbf{v}' = A(x)\mathbf{v} + \mathbf{f}(x), \quad a < x \leq b, \quad (6.10)$$

$$\mathbf{v}(a) = \mathbf{0}. \quad (6.11)$$

Отже, стовпці матриці $U(x)$ і вектор $\mathbf{v}(x)$ можуть бути обчислені як розв'язки $n + 1$ задачі Коші.

Вектор-параметр \mathbf{s} визначимо з умови, щоб функція (6.7) задовольняла крайову умову (6.6). Підставимо (6.7) в (6.6), тоді будемо мати

$$B_a U(a)\mathbf{s} + B_a \mathbf{v}(a) + B_b U(b)\mathbf{s} + B_b \mathbf{v}(b) = \mathbf{d}.$$

Звідси, враховуючи початкові умови (6.9), (6.11), отримаємо систему лінійних алгебраїчних рівнянь відносно вектора \mathbf{s}

$$Q\mathbf{s} = \hat{\mathbf{d}}, \quad (6.12)$$

де

$$Q = B_a + B_b U(b), \quad \hat{\mathbf{d}} = \mathbf{d} - B_b \mathbf{v}(b).$$

Припускаючи, що крайова задача (6.5), (6.6) має єдиний розв'язок, можна показати, що матриця Q буде не виродженою.

При практичній реалізації задачі (6.8), (6.9) та (6.10), (6.11) розв'язують чисельно на деякій нерівномірній сітці $\hat{\omega}_h = \{x_i, i = \overline{0, N}, x_0 = a, x_N = b\}$, а далі розв'язують систему лінійних рівнянь (6.12) і обчислюють розв'язок крайової задачі (6.5), (6.6) у вузлах сітки $\hat{\omega}_h$ за формулою (6.7).

Розглянемо тепер систему нелінійних звичайних диференціальних рівнянь першого порядку

$$\mathbf{u}' = \mathbf{f}(x, \mathbf{u}), \quad a < x < b, \quad (6.13)$$

з загальними двоточковими крайовими умовами

$$\mathbf{g}(\mathbf{u}(a), \mathbf{u}(b)) = \mathbf{0}. \quad (6.14)$$

Через $\mathbf{u}(x; \mathbf{s})$ позначимо розв'язок системи (6.13), який задовольняє початкову умову

$$\mathbf{u}(a; \mathbf{s}) = \mathbf{s}. \quad (6.15)$$

Тоді задача (6.13), (6.14) зводиться до знаходження вектора $\mathbf{s} = \mathbf{s}^*$, який є розв'язком системи нелінійних алгебраїчних рівнянь

$$\mathbf{F}(\mathbf{s}^*) = \mathbf{0}, \quad (6.16)$$

де

$$\mathbf{F}(\mathbf{s}) = \mathbf{g}(\mathbf{s}, \mathbf{u}(b; \mathbf{s})). \quad (6.17)$$

Розв'язок системи (6.16), (6.17) можна обчислити за допомогою ітераційного методу Ньютона

$$\mathbf{s}^{(k+1)} = \mathbf{s}^{(k)} - \left[\frac{\partial \mathbf{F}(\mathbf{s}^{(k)})}{\partial \mathbf{s}} \right]^{-1} \mathbf{F}(\mathbf{s}^{(k)}), \quad k = 0, 1, \dots$$

Будемо припускати, що початкове наближення $\mathbf{s}^{(0)}$ задано. Таким чином, на кожній ітерації необхідно обчислити матрицю Якобі $\partial \mathbf{F}(\mathbf{s}^{(k)}) / \partial \mathbf{s}$ і розв'язати систему лінійних рівнянь

$$\begin{aligned} \frac{\partial \mathbf{F}(\mathbf{s}^{(k)})}{\partial \mathbf{s}} \Delta \mathbf{s}^{(k)} &= -\mathbf{F}(\mathbf{s}^{(k)}), \\ \mathbf{s}^{(k+1)} &= \mathbf{s}^{(k)} + \Delta \mathbf{s}^{(k)}. \end{aligned} \quad (6.18)$$

Для обчислення $\mathbf{F}(\mathbf{s}) = \mathbf{g}(\mathbf{s}, \mathbf{u}(b; \mathbf{s}))$ при $\mathbf{s} = \mathbf{s}^{(k)}$ потрібно обчислити $\mathbf{u}(b; \mathbf{s}^{(k)})$, тобто розв'язати задачу Коші (6.13), (6.15) при $\mathbf{s} = \mathbf{s}^{(k)}$. Введемо позначення

$$B_a(\mathbf{s}) = \left. \frac{\partial \mathbf{g}(\mathbf{y}, \mathbf{v})}{\partial \mathbf{y}} \right|_{\substack{\mathbf{y}=\mathbf{s} \\ \mathbf{v}=\mathbf{u}(b; \mathbf{s})}}, \quad B_b(\mathbf{s}) = \left. \frac{\partial \mathbf{g}(\mathbf{y}, \mathbf{v})}{\partial \mathbf{v}} \right|_{\substack{\mathbf{y}=\mathbf{s} \\ \mathbf{v}=\mathbf{u}(b; \mathbf{s})}}$$

і $U(x) = U(x; \mathbf{s}) \equiv \partial \mathbf{u}(x; \mathbf{s}) / \partial \mathbf{s}$ — матриця розміру $n \times n$. Якщо продиференціювати диференціальне рівняння (6.13) та початкову умову (6.15) по \mathbf{s} , то отримаємо, що функція $U(x)$ є розв'язком задачі Коші

$$U' = A(x)U, \quad a < x \leq b, \quad (6.19)$$

$$U(a) = I, \quad (6.20)$$

де

$$A(x) = A(x, \mathbf{s}) = \frac{\partial \mathbf{f}(x, \mathbf{u}(x; \mathbf{s}))}{\partial \mathbf{u}}. \quad (6.21)$$

Тоді з (6.17) випливає, що

$$\frac{\partial \mathbf{F}(\mathbf{s})}{\partial \mathbf{s}} = B_a(\mathbf{s}) + B_b(\mathbf{s})U(b; \mathbf{s}).$$

Зазначимо, що величини $\mathbf{u}(x; \mathbf{s})$ та $U(x)$ будемо заміняти їх чисельною апроксимацією, одержаною чисельним розв'язуванням задач Коші (6.13), (6.15) і (6.19), (6.20) відповідно.

Іноді корисною альтернативою є апроксимація матриці $\partial \mathbf{F}(\mathbf{s})/\partial \mathbf{s}$ скінченними різницями, тобто j -й стовпець $\partial \mathbf{F}(\mathbf{s})/\partial s_j$ апроксимується співвідношенням

$$\varepsilon_j^{-1} (\mathbf{F}(\mathbf{s} + \varepsilon_j \mathbf{e}_j) - \mathbf{F}(\mathbf{s})),$$

де \mathbf{e}_j — j -й одиничний вектор, тобто вектор в j -му рядку якого стоїть 1, а всі решта елементи рівні 0, $\varepsilon_j \approx s_j \sqrt{\varepsilon_M}$, s_j — j -та компонента вектора \mathbf{s} , ε_M — машинна точність. Це вимагає ще одного обчислення вектора $\mathbf{F}(\mathbf{s} + \varepsilon_j \mathbf{e}_j)$, тобто розв'язування $n + 1$ нелінійної задачі Коші на кожній ітерації, але дозволяє запобігти явного обчислення матриці Якобі $\partial \mathbf{F}(\mathbf{s})/\partial \mathbf{s}$ та матриці $A(x)$ за формулою (6.21).

6.2.2. Метод многократної стрільби

Головний недолік методу однократної стрільби — нагромадження похибок заокруглень, що виникають через нестійкість задач Коші. Щоб запобігти цьому, було розроблено метод многократної стрільби.

Припустимо, що інтервал $[a, b]$ розбитий точками сітки $a = x_0 < x_1 < \dots < x_M = b$ на частини і розглянемо задачі Коші для $i = 0, M - 1$

$$\mathbf{u}' = \mathbf{f}(x, \mathbf{u}), \quad x_i < x < x_{i+1}, \quad (6.22)$$

$$\mathbf{u}(x_i) = \mathbf{s}_i. \quad (6.23)$$

Розв'язок задач (6.22), (6.23) позначимо через $\mathbf{u}_i(x; \mathbf{s}_i)$.

Невідомі nM параметри

$$\mathbf{s}^T = (\mathbf{s}_0^T, \mathbf{s}_1^T, \dots, \mathbf{s}_{M-1}^T)$$

будуть визначатися з умов, що чисельний розв'язок задачі (6.13), (6.14) неперервний на інтервалі $[a, b]$ і такий, що крайові умови (6.14) задовольняються. Отже, шуканий розв'язок визначається рівністю

$$\mathbf{u}(x) = \mathbf{u}_i(x; \mathbf{s}_i), \quad x_i \leq x \leq x_{i+1}, \quad i = \overline{0, M-1},$$

де умови на \mathbf{s} такі

$$\begin{aligned} \mathbf{u}_i(x_{i+1}; \mathbf{s}_i) &= \mathbf{s}_{i+1}, \quad i = \overline{0, M-2}, \\ \mathbf{g}(\mathbf{s}_0, \mathbf{u}_{M-1}(b; \mathbf{s}_{M-1})) &= \mathbf{0}. \end{aligned} \quad (6.24)$$

Систему nM нелінійних рівнянь (6.24) запишемо у вигляді

$$\begin{aligned} \mathbf{F}(\mathbf{s}) &= \begin{pmatrix} \mathbf{F}_1(\mathbf{s}) \\ \mathbf{F}_2(\mathbf{s}) \\ \dots \\ \mathbf{F}_{M-1}(\mathbf{s}) \\ \mathbf{F}_M(\mathbf{s}) \end{pmatrix} = \\ &= \begin{pmatrix} \mathbf{s}_1 - \mathbf{u}_0(x_1; \mathbf{s}_0) \\ \mathbf{s}_2 - \mathbf{u}_1(x_2; \mathbf{s}_1) \\ \dots \\ \mathbf{s}_{M-1} - \mathbf{u}_{M-2}(x_{M-1}; \mathbf{s}_{M-2}) \\ \mathbf{g}(\mathbf{s}_0, \mathbf{u}_{M-1}(b; \mathbf{s}_{M-1})) \end{pmatrix} = \mathbf{0}. \end{aligned} \quad (6.25)$$

Для розв'язування системи (6.25) використаємо ітераційний метод Ньютона

$$\frac{\partial \mathbf{F}(\mathbf{s}^{(k)})}{\partial \mathbf{s}} \Delta \mathbf{s}^{(k)} = -\mathbf{F}(\mathbf{s}^{(k)}), \quad \mathbf{s}^{(k+1)} = \mathbf{s}^{(k)} + \Delta \mathbf{s}^{(k)} \quad (6.26)$$

з матрицею Якобі

$$\begin{aligned} \frac{\partial \mathbf{F}(\mathbf{s})}{\partial \mathbf{s}} &= \\ &= \begin{pmatrix} -U_0(x_1) & I & & & \\ & -U_1(x_2) & I & & \\ & & \ddots & \ddots & \\ & & & -U_{M-2}(x_{M-1}) & I \\ B_a(\mathbf{s}) & & & & B_b(\mathbf{s})U_{M-1}(b) \end{pmatrix}, \end{aligned}$$

де $U_i(x) \equiv U_i(x; x_i, \mathbf{s}_i) \equiv \partial \mathbf{u}_i(x; \mathbf{s}_i) / \partial \mathbf{s}$ — матриця розміру $n \times n$, яка є розв'язком задачі

$$\begin{aligned} U'_i &= A(x)U_i, \quad x_i < x \leq x_{i+1}, \\ U(x_i) &= I, \quad i = \overline{0, M-1}. \end{aligned}$$

Підставляючи матрицю Якобі в (6.26) отримаємо систему лінійних рівнянь:

$$\Delta \mathbf{s}_1^{(k)} - U_0(x_1) \Delta \mathbf{s}_0^{(k)} = -\mathbf{F}_1(\mathbf{s}^{(k)}),$$

$$\Delta \mathbf{s}_2^{(k)} - U_1(x_2) \Delta \mathbf{s}_1^{(k)} = -\mathbf{F}_2(\mathbf{s}^{(k)}),$$

.....

$$\Delta \mathbf{s}_{M-1}^{(k)} - U_{M-2}(x_{M-1}) \Delta \mathbf{s}_{M-2}^{(k)} = -\mathbf{F}_{M-1}(\mathbf{s}^{(k)}),$$

$$B_a(\mathbf{s}) \Delta \mathbf{s}_0^{(k)} + B_b(\mathbf{s}) U_{M-1}(b) \Delta \mathbf{s}_{M-1}^{(k)} = -\mathbf{F}_M(\mathbf{s}^{(k)}).$$

Тоді

$$\Delta \mathbf{s}_1^{(k)} = U_0(x_1) \Delta \mathbf{s}_0^{(k)} - \mathbf{F}_1(\mathbf{s}^{(k)}),$$

$$\Delta \mathbf{s}_2^{(k)} = U_1(x_2) U_0(x_1) \Delta \mathbf{s}_0^{(k)} - \mathbf{F}_2(\mathbf{s}^{(k)}) - U_1(x_2) \mathbf{F}_1(\mathbf{s}^{(k)}),$$

.....

$$\Delta \mathbf{s}_{M-1}^{(k)} = U_{M-2}(x_{M-1}) U_{M-3}(x_{M-2}) \dots U_0(x_1) \Delta \mathbf{s}_0^{(k)} - \quad (6.27)$$

$$- \sum_{j=0}^{M-2} \left(\prod_{l=1}^j U_{M-l-1}(x_{M-l}) \right) \mathbf{F}_{M-j-1}(\mathbf{s}^{(k)}),$$

$$(B_a(\mathbf{s}) + B_b(\mathbf{s}) U_{M-1}(b) U_{M-2}(x_{M-1}) \dots U_0(x_1)) \Delta \mathbf{s}_0^{(k)} = \mathbf{w}, \quad (6.28)$$

де

$$\begin{aligned} \mathbf{w} = & -\mathbf{F}_M(\mathbf{s}^{(k)}) + B_b U_{M-1}(b) \mathbf{F}_{M-1}(\mathbf{s}^{(k)}) + \\ & + B_b U_{M-1}(b) U_{M-2}(x_{M-1}) \mathbf{F}_{M-2}(\mathbf{s}^{(k)}) + \\ & + \dots + B_b U_{M-1}(b) U_{M-2}(x_{M-1}) \dots U_1(x_2) \mathbf{F}_1(\mathbf{s}^{(k)}). \end{aligned}$$

Система (6.28) — це система лінійних алгебраїчних рівнянь відносно невідомого вектора $\Delta \mathbf{s}_0^{(k)}$, яка може бути розв'язана методом Гаусса. Знайшовши $\Delta \mathbf{s}_0^{(k)}$ послідовно обчислимо

$$\Delta \mathbf{s}_{i+1}^{(k)} = U_i(x_{i+1}) \Delta \mathbf{s}_i^{(k)} - \mathbf{F}_{i+1}(\mathbf{s}^{(k)}), \quad i = \overline{0, M-2},$$

тоді

$$\mathbf{s}_i^{(k+1)} = \mathbf{s}_i^{(k)} + \Delta \mathbf{s}_i^{(k)}, \quad i = \overline{0, M-1}.$$

Як у випадку методу однократної стрільби елементи матриці Якобі $U_i(x_{i+1})$ можна замінити їх скінченно-різницевою апроксимацією. На

кожному інтервалі стрільби (x_i, x_{i+1}) обчислимо апроксимацію розв'язку задачі (6.22), (6.23) $\mathbf{u}_i(x_{i+1}; \mathbf{s}_i)$. Одночасно ще n розв'язків рівняння (6.22) з початковими умовами

$$\tilde{U}(x_i; \mathbf{s}_i) = (\mathbf{s}_i | \dots | \mathbf{s}_i) + \sqrt{\varepsilon_M} I,$$

де $(\mathbf{s}_i | \dots | \mathbf{s}_i)$ — матриця розміру $n \times n$, стовпцями якої є вектори \mathbf{s}_i , ε_M — машинна точність. Тоді апроксимацію $U_i(x_{i+1})$ обчислимо за формулою

$$U_i(x_{i+1}) \approx \varepsilon_M^{-1/2} \left[\tilde{U}_i(x_{i+1}; \mathbf{s}_i) - (\mathbf{u}_i(x_{i+1}; \mathbf{s}_i) | \dots | \mathbf{u}_i(x_{i+1}; \mathbf{s}_i)) \right].$$

6.3. Метод скінченних різниць

Метод скінченних різниць полягає в заміні диференціальних рівнянь різницевиими (дискретними) рівняннями, які називають різницевою схемою. Для побудови різницевої схеми множина, на якій розглядається задача, замінюється дискретною множиною точок (сіткою). Значення функцій, похідних, початкові і граничні умови подають через значення дискретних (сіткових) функцій у вузлах вибраної сітки, тобто здійснюється заміна диференціального оператора різницевим, а також будуються різницеві аналоги всіх додаткових умов. Отже, задача зводиться до розв'язування системи алгебраїчних рівнянь.

Різницеві схеми повинні відображати в просторі сіткових функцій основні властивості диференціальних рівнянь — такі, як самоспряженість, знаковизначеність оператора, виконання певних апіорних оцінок тощо. Важливою задачею є одержання різницевих схем із заданою якістю. Для побудови таких схем використовують ряд методів, про які йтиметься в цьому параграфі.

6.3.1. Метод заміни похідних скінченними різницями

Цей спосіб побудови різницевих схем полягає у тому, що всі похідні, які входять у диференціальне рівняння та крайові умови, замінюються деякими різницевиими відношеннями. Для цього використовують формули чисельного диференціювання.

Диференціальний оператор L , заданий в класі функцій неперервного аргументу, може бути наближено замінений (апроксимований) різницевим оператором L_h , заданим на сіткових функціях. Одним із методів апроксимації є заміна кожної з похідних різницевим відношенням, яке містить значення сіткової функції в декількох вузлах сітки.

Розглянемо декілька прикладів побудови L_h .

1. Нехай $Lu = u'(x)$, $x \in [a, b]$. Найпростішими різницеви́ми апроксимаціями на рівномірній сітці

$$\bar{\omega}_h = \{x_i = a + ih, i = 0, N, h = (b - a)/N\}$$

є різницеві похідні:

$$\begin{aligned} L_h^- u_i &= u_{\bar{x},i} = \frac{u_i - u_{i-1}}{h} \text{ — лівая;} \\ L_h^+ u_i &= u_{x,i} = \frac{u_{i+1} - u_i}{h} \text{ — права;} \\ L_h^0 u_i &= u_{\dot{x},i} = \frac{u_{i+1} - u_{i-1}}{2h} \text{ — центральна.} \end{aligned}$$

Множину вузлів, в яких значення сіткових функцій входять у вираз $L_h u_i$, називають *шаблоном оператора* L_h у точці x_i . Очевидно, що шаблони операторів L_h^-, L_h^+ складаються з двох точок (x_{i-1}, x_i або x_i, x_{i+1}), а L_h^0 — з трьох (x_{i-1}, x_i, x_{i+1}).

Похибкою апроксимації оператора L оператором L_h називається різниця

$$\psi_i = L_h u_i - Lu_i.$$

Кажуть, що L_h має p -й *порядок апроксимації* в точці x_i , якщо

$$\psi_i = L_h u_i - Lu_i = O(h^p).$$

Використовуючи формулу Тейлора

$$u_{i\pm 1} = u_i \pm hu'_i + \frac{h^2}{2}u''_i \pm \frac{h^3}{6}u'''_i + \frac{h^4}{24}u^{IV}_i + O(h^5), \quad (6.29)$$

одержимо

$$u_{\bar{x},i} = u'_i - \frac{h}{2}u''_i + O(h^2),$$

$$u_{x,i} = u'_i + \frac{h}{2}u''_i + O(h^2),$$

$$u_{\dot{x},i} = u'_i + \frac{h^2}{6}u'''_i + O(h^3).$$

Отже, якщо $u \in C^{(2)}[a, b]$, то ліва і права різницеві похідні апроксимують u' з першим порядком, а якщо $u \in C^{(3)}[a, b]$, то центральна різницева похідна — з другим порядком.

2. $Lu = u''(x)$, $x \in [a, b]$. Виберемо триточковий шаблон, який складається з вузлів x_{i-1}, x_i, x_{i+1} , і розглянемо різницевий оператор

$$L_h u_h = u_{\bar{x}x,i} = \frac{1}{h}(u_{x,i} - u_{\bar{x},i}) = \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2}.$$

Користуючись формулою Тейлора (6.29) для $u_{i\pm 1}$, знайдемо

$$u_{\bar{x}x,i} = u_i'' + \frac{h^2}{12}u_i^{IV} + O(h^4),$$

тобто, якщо $u \in C^{(4)}[a, b]$, то L_h має другий порядок апроксимації.

3. Нехай $\omega_h = \{x_i, i = \overline{0, N}, x_0 = 0, x_N = 1\}$ — нерівномірна сітка з кроками $h_i = x_i - x_{i-1}$, $i = \overline{1, N}$. Виберемо триточковий шаблон x_{i-1}, x_i, x_{i+1} і введемо позначення:

$$\bar{h}_i = \frac{1}{2}(h_i + h_{i+1}), \quad u_{\bar{x},i} = \frac{u_i - u_{i-1}}{\bar{h}_i},$$

$$u_{x,i} = \frac{u_{i+1} - u_i}{h_{i+1}}, \quad u_{\hat{x},i} = \frac{u_{i+1} - u_i}{\bar{h}_i}.$$

Оператору $Lu = u''$ поставимо у відповідність різницевий:

$$L_h u_i = \frac{1}{\bar{h}_i} \left[\frac{u_{i+1} - u_i}{h_{i+1}} - \frac{u_i - u_{i-1}}{h_i} \right].$$

Похибка апроксимації

$$\psi_i = L_h u_i - (Lu)_i = \frac{h_{i+1} - h_i}{3} u_i''' + O(\bar{h}_i^2).$$

Звідси маємо:

$$\|\psi\|_{C(\omega_h)} = \max_{1 \leq i \leq N-1} |\psi_i| = O(|h|), \quad |h| = \max_{1 \leq i \leq N} h_i$$

оператор L_h має перший порядок апроксимації у сітковій нормі $C(\omega_h)$.

Розглянемо приклад крайової задачі

$$u'' - q(x)u = -f(x), \quad a < x < b, \quad (6.30)$$

$$u(a) = \mu_1, \quad u(b) = \mu_2, \quad (6.31)$$

де μ_1, μ_2 — задані числа. Для чисельного розв'язування (6.30), (6.31) введемо на відріжку $[a, b]$ рівномірну сітку $\bar{\omega}_h = \{x_i = a + ih, i = \overline{0, N}, h = (b - a)/N\}$ і замінімо $u''(x_i)$ другою різницевою похідною

$u_{\bar{x}x,i}$. Тоді замість диференціального рівняння одержимо різницеве рівняння

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - q_i y_i = -f_i, \quad i = \overline{1, N-1}, \quad (6.32)$$

де $q_i = q(x_i)$, $f_i = f(x_i)$. Граничні умови замінимо співвідношеннями

$$y_0 = \mu_1, \quad y_N = \mu_2. \quad (6.33)$$

Розв'язуючи задачі (6.30), (6.31) різницеvim методом, необхідно знати, з якою точністю розв'язок різницевої задачі наближає розв'язок вихідної задачі.

Величина $z_i = y_i - u_i$ буде похибкою наближеного розв'язку y_i . Підставимо $y_i = z_i + u_i$ в (6.32), (6.33), тоді отримаємо

$$\begin{aligned} z_{\bar{x}x,i} - q_i z_i &= -\psi_i, \quad i = \overline{1, N-1}, \\ z_0 &= 0, \quad z_N = 0, \end{aligned}$$

де $\psi_i = u_{\bar{x}x,i} - q_i u_i + f_i$, $i = \overline{1, N-1}$ — похибка апроксимації (нев'язка) різницевого рівняння (6.32) на розв'язку $u(x)$ рівняння (6.30). Граничні умови (6.31) задовольняються точно.

Оскільки

$$\psi_i = u_{\bar{x}x,i} - u_i'' = \frac{h^2}{12} u_i^{IV} + O(h^4),$$

то різницева схема має другий порядок апроксимації.

Нехай при $x = a$ розв'язок рівняння (6.30) задовольняє граничну умову третього роду

$$u'(a) = \beta_1 u(a) - \mu_1.$$

Побудуємо різницеву апроксимацію цієї умови. Для цього замінимо похідну $u'(a)$ першою різницевою похідною $u_{x,0} = (u_1 - u_0)/h$. Апроксимація граничної умови буде мати вигляд

$$y_{x,0} = \beta_1 y_0 - \mu_1. \quad (6.34)$$

Підставимо $y_0 = z_0 + u_0$ в (6.34), тоді

$$z_{x,0} = \beta_1 z_0 - \psi_0,$$

де $\psi_0 = u_{x,0} - \beta_1 u_0 + \mu_1$ — похибка апроксимації граничної умови (6.34). Оскільки

$$u_{x,0} = u'_0 + 0,5hu''_0 + O(h^2), \quad (6.35)$$

то

$$\begin{aligned}\psi_0 &= u'(a) - \beta_1 u(a) + \mu_1 + 0,5hu''(a) + O(h^2) = \\ &= 0,5hu''(a) + O(h^2).\end{aligned}$$

Звідси видно, що $\psi_0 = O(h)$, тобто гранична умова (6.34) має перший порядок апроксимації.

Замінімо граничну умову різницевою так, щоб порядок апроксимації був $O(h^2)$. Оскільки $u(x)$ — розв'язок диференціального рівняння (6.30), то $u''(a) = q(a)u(a) - f(a)$ і з (6.35) одержимо

$$u_{x,0} - 0,5h(q_0 u_0 - f_0) = u'_0 + O(h^2), \quad (6.36)$$

тобто вираз у лівій частині (6.36) апроксимує похідну в точці $x = a$ з другим порядком апроксимації. Звідси випливає, що гранична умова

$$y_{x,0} - 0,5h(q_0 y_0 - f_0) = \beta_1 y_0 - \mu_1$$

має другий порядок апроксимації.

Аналогічно можна показати, що різницеве рівняння

$$y_{\bar{x},N} + 0,5h(q_N y_N - f_N) = -\beta_2 y_N + \mu_2 \quad (6.37)$$

апроксимує з другим порядком граничну умову

$$u'(b) = -\beta_2 u(b) + \mu_2.$$

Приклад 6.1. Методом заміни похідних скінченними різницями для крайової задачі

$$(1 + x^2)u'' + 2xu' - u = x^2, \quad 0 < x < 1, \quad (6.38)$$

$$u'(0) = 1, \quad u(1) = 0 \quad (6.39)$$

побудуйте різницеву схему другого порядку апроксимації.

▷ Введемо сітку $\bar{\omega}_h = \{x_i = ih, i = \overline{0, N}, h = 1/N\}$. Замінімо другу похідну $u''(x_i)$ другою різницевою похідною $u_{\bar{x}x,i}$, а першу похідну $u'(x_i)$ центральною різницевою похідною $u_{x,i}^\circ$, тоді будемо мати

$$(1 + x_i^2)y_{\bar{x}x,i} + 2x_i y_{x,i}^\circ - y_i = x_i^2, \quad i = \overline{1, N-1}. \quad (6.40)$$

Для апроксимації $u'(0)$ з другим порядком використаємо рівність (6.35) та диференціальне рівняння в точці x_0 : $u''_0 - u_0 = 0$, тоді отримаємо

$$u_{x,0} - 0,5hu_0 = u'_0 + O(h^2).$$

Отже, різницеві крайові умови будуть мати вигляд

$$y_{x,0} = 0,5hy_0 + 1, \quad y_N = 0. \quad (6.41)$$

Покажемо, що різницева схема (6.40), (6.41) має другий порядок апроксимації. Справді,

$$\begin{aligned} \psi_i &= (1 + x_i^2)u_{\bar{x}x,i} + 2x_i u_{x,i}^\circ - u_i - x_i^2 = \\ &= (1 + x_i^2)u_i'' + 2x_i u_i' - u_i - x_i^2 + O(h^2) = O(h^2), \quad i = \overline{1, N-1}, \\ \psi_0 &= u_{x,0} - 0,5hu_0 - 1 = u_0' + 0,5h(u_0'' - u_0) - 1 + O(h^2) = O(h^2). \end{aligned}$$

◀

6.3.2. Метод неозначених коефіцієнтів

Метод неозначених коефіцієнтів полягає в тому, що різницеву схему задають у вигляді лінійної комбінації значень різницевого розв'язку у вузлах шаблону. Коефіцієнти цієї лінійної комбінації визначають з умови, щоб похибка апроксимації схеми мала якомога вищий порядок малості.

Розглянемо, наприклад, першу крайову задачу для звичайного диференціального рівняння другого порядку

$$(k(x)u')' - q(x)u = -f(x), \quad 0 < x < l, \quad (6.42)$$

$$u(0) = \mu_1, \quad u(l) = \mu_2, \quad (6.43)$$

де $k(x)$, $q(x)$, $f(x)$ — задані достатньо гладкі функції, які задовольняють умови $k(x) \geq c_1 > 0$, $q(x) \geq 0$, μ_1, μ_2 — задані числа.

За сформульованих припущень, існує єдиний розв'язок задачі (6.42), (6.43). Будемо вважати, що $u \in C^{(3)}[0, l]$.

Рівняння (6.42) описує стаціонарний розподіл температури (стаціонарне рівняння теплопровідності) або концентрації (рівняння дифузії).

Введемо на відрізьку $[0, l]$ рівномірну сітку

$$\bar{\omega}_h = \{x_i = ih, \quad i = \overline{0, N}, \quad h = l/N\}$$

і виберемо триточковий шаблон x_{i-1} , x_i , x_{i+1} , на якому будемо будувати різницеву схему. Будь-яке різницеве рівняння на цьому шаблоні має вигляд

$$b_i y_{i+1} - c_i y_i + a_i y_{i-1} = -h^2 \varphi_i, \quad i = \overline{1, N-1}, \quad (6.44)$$

де a_i, b_i, c_i — коефіцієнти, які залежать від h . Ці коефіцієнти поки що неозначені. Перепишемо (6.44) інакше

$$\frac{1}{h} \left(b_i \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right) - d_i y_i = -\varphi_i, \quad i = \overline{1, N-1}, \quad (6.45)$$

де $d_i = (c_i - a_i - b_i)/h^2$.

Розв'язок різницевого рівняння (6.45) повинен задовольняти крайові умови

$$y_0 = \mu_1, \quad y_N = \mu_2.$$

Обчислимо похибку апроксимації

$$\psi_i = \frac{1}{h} (b_i u_{x,i} - a_i u_{\bar{x},i}) - d_i u_i + \varphi_i - (ku')'_i + q_i u_i - f_i,$$

схеми (6.45). Для цього розкладемо $u_{i\pm 1}$ в точці x_i за формулою Тейлора, тоді з урахуванням

$$\begin{aligned} u_{x,i} &= u'_i + \frac{h}{2} u''_i + \frac{h^2}{6} u'''_i + O(h^3), \\ u_{\bar{x},i} &= u'_i - \frac{h}{2} u''_i + \frac{h^2}{6} u'''_i + O(h^3), \end{aligned}$$

отримаємо

$$\begin{aligned} \psi_i &= \frac{1}{h} (b_i - a_i) u'_i + \frac{b_i + a_i}{2} u''_i + \frac{h(b_i - a_i)}{6} u'''_i - \\ &\quad - d_i u_i + \varphi_i - k'_i u'_i - k_i u''_i + q_i u_i - f_i = \\ &= \left(\frac{1}{h} (b_i - a_i) - k'_i \right) u'_i + \left(\frac{b_i + a_i}{2} - k_i \right) u''_i + \\ &\quad + \frac{h(b_i - a_i)}{6} u'''_i - (d_i - q_i) u_i + \varphi_i - f_i + O(h^2). \end{aligned}$$

Якщо вибрати a_i, b_i, d_i, φ_i так щоб

$$\begin{aligned} \frac{b_i - a_i}{h} &= k'_i + O(h^2), \quad \frac{b_i + a_i}{2} = k_i + O(h^2), \\ d_i &= q_i + O(h^2), \quad \varphi_i = f_i + O(h^2), \end{aligned} \quad (6.46)$$

то $\psi_i = O(h^2)$.

Наведемо приклад різницевої схеми другого порядку апроксимації

$$\frac{1}{h} \left(k_{i+1/2} \frac{y_{i+1} - y_i}{h} - k_{i-1/2} \frac{y_i - y_{i-1}}{h} \right) - q_i y_i = -f_i, \\ i = \overline{1, N-1},$$

де $k_{i\pm 1/2} = k(x_i \pm 0,5h)$.

6.3.3. Інтегро-інтерполяційний метод побудови різницевих схем

Розглянемо крайову задачу для звичайного диференціального рівняння другого порядку:

$$(k(x)u')' - q(x)u = -f(x), \quad 0 < x < l, \quad (6.47)$$

$$k(0)u'(0) = \beta_1 u(0) - \mu_1, \quad u(l) = \mu_2. \quad (6.48)$$

де $k(x) \geq c_1 > 0$, $q(x) \geq 0$, $\beta_1 \geq 0$, μ_1, μ_2 — задані числа.

На відрітку $0 \leq x \leq l$ введемо рівномірну сітку $\bar{\omega}_h = \{x_i = ih, i = \overline{0, N}, h = l/N\}$. Різницеві схеми, які виражають на сітці закони збереження, називаються *консервативними* (або *дивергентними*) [17]. Для одержання консервативних різницевих схем природно виходити з рівнянь балансу, записаних для елементарних об'ємів сіткової області. Інтеграл і похідні, які входять у ці рівняння балансу, необхідно замінити наближеними різницевиими виразами. Такий метод побудови різницевих схем будемо називати *інтегро-інтерполяційним методом* (методом балансу).

Позначимо $x_{i\pm 1/2} = x_i \pm 0,5h$, $w(x) = ku'$, $w_{i\pm 1/2} = w(x_{i\pm 1/2})$ і проінтегруємо диференціальне рівняння на відрітку $x_{i-1/2} \leq x \leq x_{i+1/2}$. Тоді одержимо рівняння

$$w_{i+1/2} - w_{i-1/2} - \int_{x_{i-1/2}}^{x_{i+1/2}} q(x)u(x)dx = - \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)dx, \quad (6.49)$$

Якщо задача (6.47), (6.48) описує стаціонарний розподіл температури, то (6.49) є рівнянням балансу тепла на відрітку $[x_{i-1/2}, x_{i+1/2}]$, де $(-w(x))$ — потік тепла, $q(x)u(x)$ — потужність стоків тепла, $f(x)$ — густина розподілу зовнішніх джерел тепла.

Замінімо інтеграл

$$\int_{x_{i-1/2}}^{x_{i+1/2}} q(x)u(x)dx$$

його наближеним значенням

$$u_i \int_{x_{i-1/2}}^{x_{i+1/2}} q(x)dx$$

і введемо позначення

$$d_i = \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} q(x)dx, \quad \varphi_i = \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)dx \quad (6.50)$$

У результаті замість (6.49) одержимо рівняння

$$\frac{w_{i+1/2} - w_{i-1/2}}{h} - d_i u_i + \varphi_i = 0. \quad (6.51)$$

Виразимо тепер $w_{i\pm 1/2}$ через значення функції $u(x)$ в точках сітки. Для цього проінтегруємо співвідношення $u'(x) = w(x)/k(x)$ на відрізки $x_{i-1} \leq x \leq x_i$:

$$u_i - u_{i-1} = \int_{x_{i-1}}^{x_i} \frac{w(x)}{k(x)} dx \approx w_{i-1/2} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)}.$$

Нехай

$$a_i = \left(\frac{1}{h} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} \right)^{-1}, \quad i = \overline{1, N}, \quad (6.52)$$

тоді

$$w_{i-1/2} \approx a_i \frac{u_i - u_{i-1}}{h} = a_i u_{\bar{x}, i}, \quad w_{i+1/2} = a_{i+1} u_{x, i}.$$

Підставляючи ці вирази в (6.51) і позначаючи через y_i шукану функцію, одержимо різницеве рівняння

$$\frac{1}{h} \left(a_{i+1} \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h} \right) - d_i y_i = -\varphi_i, \quad i = \overline{1, N-1},$$

або

$$(ay_{\bar{x}})_{x,i} - d_i y_i = -\varphi_i, \quad i = \overline{1, N-1}.$$

Апроксимація другої граничної умови (6.48) має вигляд

$$y_N = \mu_2,$$

а різницева апроксимація першої граничної умови може бути одержана інтегро-інтерполяційним методом. Для цього проінтегруємо рівняння (6.47) на відрізку $0 \leq x \leq x_{1/2}$, де $x_{1/2} = 0,5h$:

$$w_{1/2} - w_0 - \int_0^{x_{1/2}} q(x)u(x)dx = - \int_0^{x_{1/2}} f(x)dx,$$

де $w_{1/2} = a_1 u_{\bar{x},1}$, $w_0 = \beta_1 u(0) - \mu_1$. Нехай

$$\int_0^{x_{1/2}} q(x)u(x)dx \approx u_0 \int_0^{x_{1/2}} q(x)dx,$$

тоді будемо мати різницеве рівняння

$$a_1 y_{x,0} = \bar{\beta}_1 y_0 - \bar{\mu}_1,$$

де

$$\bar{\beta}_1 = \beta_1 + 0,5hd_0, \quad \bar{\mu}_1 = \mu_1 + 0,5h\varphi_0, \quad (6.53)$$

$$d_0 = \frac{1}{0,5h} \int_0^{x_{1/2}} q(x)dx, \quad \varphi_0 = \frac{1}{0,5h} \int_0^{x_{1/2}} f(x)dx. \quad (6.54)$$

Отже, для задачі (6.47), (6.48) побудовано різницеву схему

$$(ay_{\bar{x}})_{x,i} - d_i y_i = -\varphi_i, \quad i = \overline{1, N-1}, \quad (6.55)$$

$$a_1 y_{x,0} = \bar{\beta}_1 y_0 - \bar{\mu}_1, \quad y_N = \mu_2, \quad (6.56)$$

коефіцієнти якої визначаються за формулами (6.50), (6.52), (6.53), (6.54).

Нехай розв'язок диференціального рівняння (6.47) задовольняє крайову умову

$$-k(1)u'(1) = \beta_2 u(1) - \mu_2.$$

Різницевий аналог цієї умови, який також можна побудувати інтегро-інтерполяційним методом, буде мати вигляд

$$-a_N y_{\bar{x},N} = \bar{\beta}_2 y_N - \bar{\mu}_2,$$

де

$$\begin{aligned} \bar{\beta}_2 &= \beta_2 + 0,5hd_N, \quad \bar{\mu}_2 = \mu_2 + 0,5h\varphi_N, \\ d_N &= \frac{1}{0,5h} \int_{x_{N-1/2}}^1 q(x)dx, \quad \varphi_N = \frac{1}{0,5h} \int_{x_{N-1/2}}^1 f(x)dx. \end{aligned}$$

Встановимо порядок апроксимації різницевої схеми (6.55), (6.56). Похибка апроксимації різницевого рівняння (6.55)

$$\begin{aligned} \psi_i &= \frac{1}{h} (a_{i+1}u_{x,i} - a_i u_{\bar{x},i}) - d_i u_i + \\ &+ \varphi_i - (ku')'_i + q_i u_i - f_i, \quad i = \overline{1, N-1} \end{aligned}$$

буде величиною порядку $O(h^2)$, якщо виконуються умови (6.46):

$$\begin{aligned} \frac{a_{i+1} - a_i}{h} &= k'_i + O(h^2), \quad \frac{a_{i+1} + a_i}{2} = k_i + O(h^2), \\ d_i &= q_i + O(h^2), \quad \varphi_i = f_i + O(h^2), \end{aligned} \quad (6.57)$$

Перевіримо виконання цих умов для коефіцієнтів (6.50), (6.52), (6.53), (6.54). Для цього введемо позначення $p(x) = 1/k(x)$ і розкладемо $p(x)$ в ряд Тейлора в околі точки x_i , тоді

$$\begin{aligned} \frac{1}{a_i} &= \frac{1}{h} \int_{x_{i-1}}^{x_i} p(x)dx = \\ &= \frac{1}{h} \int_{x_{i-1}}^{x_i} [p_i + (x - x_i)p'_i + \frac{(x - x_i)^2}{2}p''_i + O(x - x_i)^3]dx = \\ &= p_i - \frac{h}{2}p'_i + \frac{h^2}{6}p''_i + O(h^3). \end{aligned}$$

Аналогічно

$$\frac{1}{a_{i+1}} = \frac{1}{h} \int_{x_i}^{x_{i+1}} p(x)dx = p_i + \frac{h}{2}p'_i + \frac{h^2}{6}p''_i + O(h^3).$$

Тоді

$$\begin{aligned}\frac{1}{2} \left(\frac{1}{a_{i+1}} + \frac{1}{a_i} \right) &= \frac{1}{k_i} + O(h^2), \\ \frac{1}{h} \left(\frac{1}{a_i} - \frac{1}{a_{i+1}} \right) &= -p'_i + O(h^2) = \frac{k'_i}{k_i^2} + O(h^2), \\ \frac{1}{a_{i+1}} \frac{1}{a_i} &= \frac{1}{k_i^2} + O(h^2).\end{aligned}$$

Помноживши останню рівність на $a_{i+1}a_i k_i^2$, отримаємо

$$a_{i+1}a_i = k_i^2 + O(h^2).$$

Отже,

$$\begin{aligned}\frac{a_{i+1} + a_i}{2} &= \frac{1}{2} \left(\frac{1}{a_{i+1}} + \frac{1}{a_i} \right) a_{i+1}a_i = k_i + O(h^2), \\ \frac{a_{i+1} - a_i}{h} &= \frac{1}{h} \left(\frac{1}{a_i} - \frac{1}{a_{i+1}} \right) a_{i+1}a_i = k'_i + O(h^2).\end{aligned}$$

Розкладемо функцію $q(x)$ в ряд Тейлора в околі точки x_i , тоді

$$\begin{aligned}d_i &= \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} q(x) dx = \\ &= \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} [q(x_i) + (x - x_i)q'(x_i) + O(x - x_i)^2] dx = \\ &= q(x_i) + O(h^2).\end{aligned}$$

Аналогічно можна показати, що

$$\varphi_i = f_i + O(h^2).$$

Дослідимо похибку апроксимації

$$\psi_0 = a_1 u_{x,0} - \bar{\beta}_1 u_0 + \bar{\mu}_1 - [k(0)u'(0) - \beta_1 u(0) + \mu_1]$$

різницевої граничної умови (6.56). З розкладу

$$\frac{1}{a_1} = \frac{1}{k(0)} - \frac{h}{2} \frac{k'(0)}{k^2(0)} + O(h^2)$$

випливає співвідношення

$$a_1 = k(0) + \frac{h}{2} \frac{k'(0)}{k(0)} a_1 + O(h^2) = k(0) + \frac{h}{2} k'(0) + O(h^2).$$

Тоді, з урахуванням $u_{x,0} = u'(0) + \frac{h}{2} u''(0) + O(h^2)$, одержимо

$$a_1 u_{x,0} = k(0) u'(0) + 0,5h (ku')'(0) + O(h^2).$$

Звідси будемо мати

$$\begin{aligned} \psi_0 &= 0,5h (ku')'(0) - \bar{\beta}_1 u_0 + \bar{\mu}_1 + \beta_1 u(0) - \mu_1 + O(h^2) = \\ &= 0,5h [(ku')'(0) - d_0 u_0 - \varphi_0] + O(h^2) = \\ &= 0,5h [(q(0) - d_0) u_0 - (f(0) - \varphi_0)] + O(h^2). \end{aligned}$$

Із співвідношення

$$d_0 = \frac{1}{0,5h} \int_0^{x_{1/2}} q(x) dx = q(0) + O(h), \quad \varphi_0 = f(0) + O(h)$$

отримаємо

$$\psi_0 = O(h^2).$$

Отже, при достатній гладкості коефіцієнтів $k(x)$, $q(x)$, $f(x)$ і розв'язку $u(x)$ різницева схема (6.55), (6.56) апроксимує задачу (6.47), (6.48) з другим порядком.

З практичного погляду зручно мати простіші формули для обчислення $a_i, i = \overline{1, N}$, $d_i, \varphi_i, i = \overline{0, N}$, які використовують значення $k(x)$, $q(x)$, $f(x)$ в окремих точках і задовольняють умови (6.57), наприклад

$$a_i = k_{i-1/2} = k(x_i - 0,5h) \quad a_i = 0,5(k_i + k_{i-1}),$$

або

$$d_i = q_i, \quad \varphi_i = f_i.$$

Використання коефіцієнтів (6.50), (6.52), (6.53), (6.54) різницевої схеми корисне при розв'язуванні задач з розривними функціями $k(x)$, $q(x)$, $f(x)$.

Побудовані різницеві схеми є системами лінійних алгебраїчних рівнянь відносно невідомих y_0, y_1, \dots, y_N з тридіагональними матрицями.

Наприклад, різницева схема (6.55), (6.56) може бути записана у вигляді системи триточкових різницевих рівнянь

$$\begin{aligned} -C_0 y_0 + B_0 y_1 &= -F_0, \\ A_i y_{i-1} - C_i y_i + B_i y_{i+1} &= -F_i, \quad i = \overline{1, N-1}, \\ A_N y_{N-1} - C_N y_N &= -F_N, \end{aligned} \quad (6.58)$$

де

$$\begin{aligned} C_0 &= \frac{2a_1}{h^2} + \frac{2\beta_1}{h} + d_0, \quad B_0 = \frac{2a_1}{h^2}, \quad F_0 = \frac{2\mu_1}{h} + \varphi_0, \\ A_i &= \frac{a_i}{h^2}, \quad C_i = \frac{a_i}{h^2} + \frac{a_{i+1}}{h^2} + d_i, \quad B_i = \frac{a_{i+1}}{h^2}, \quad F_i = \varphi_i, \\ A_N &= 0, \quad C_N = -1, \quad F_N = -\mu_2. \end{aligned}$$

З умов $a_i \geq c_1 > 0$, $d_i \geq 0$, $\beta_1 \geq 0$ випливає, що $|C_0| \geq |B_0|$, $|C_i| \geq |A_i| + |B_i|$, $|C_N| > |A_N|$. Тому різницева задача (6.58) однозначно розв'язна і її можна розв'язати методом прогонки (див. 1.1.2).

Приклад 6.2. Різницеву схему (6.40), (6.41) запишіть у вигляді системи триточкових різницевих рівнянь. Перевірте виконання умов стійкості методу прогонки для її розв'язування.

▷ Враховуючи вирази для різницевих похідних, схему (6.40), (6.41) запишемо у вигляді (6.58), де

$$\begin{aligned} A_i &= \frac{1+x_i^2}{h^2} - \frac{x_i}{h}, \quad C_i = \frac{2(1+x_i^2)}{h^2} + 1, \quad B_i = \frac{1+x_i^2}{h^2} + \frac{x_i}{h}, \\ F_i &= -x_i^2, \quad i = \overline{1, N-1}, \quad C_0 = \frac{1}{h} + 0,5h, \quad B_0 = \frac{1}{h}, \\ F_0 &= -1, \quad A_N = 0, \quad C_N = -1, \quad F_N = 0. \end{aligned}$$

Можна показати, що справджуються нерівності

$$|C_0| > |B_0|, \quad |C_i| > |A_i| + |B_i|, \quad i = \overline{1, N-1}, \quad |C_N| > |A_N|,$$

тобто метод прогонки стійкий. ◀

Приклад 6.3. Інтегро-інтерполяційним методом побудуйте різницеву схему другого порядку апроксимації для крайової задачі (6.38), (6.39).

▷ Запишемо крайову задачу (6.38), (6.39) у вигляді

$$\begin{aligned} ((1+x^2)u')' - u &= x^2, \quad 0 < x < 1, \\ u'(0) &= 1, \quad u(1) = 0. \end{aligned}$$

Тоді для її розв'язування можна використати схему (6.55), (6.56) з коефіцієнтами $a_i = 1 + x_{i-1/2}^2$, $i = \overline{1, N}$, $d_i = 1$, $\varphi_i = -x_i^2$, $i = \overline{1, N-1}$, $\beta_1 = 0,5h$, $\mu_1 = -1$, $\mu_2 = 0$. ◀

6.3.4. Збіжність різницевої схеми

Для того, щоб довести збіжність різницевої схеми (6.55), (6.56), нам потрібні будуть деякі різницеві тотожності та нерівності. Будемо розглядати сіткові функції, задані на сітці $\bar{\omega}_h$. Визначимо скалярні добутки

$$(y, v)_{\omega_h} = \sum_{i=1}^{N-1} y_i v_i h, \quad (y, v)_{\omega_h^+} = \sum_{i=1}^N y_i v_i h, \quad \omega_h^+ = \omega_h \cup x_N.$$

Справджується різницева тотожність:

$$(y, v_x)_{\omega_h} = -(v, y_{\bar{x}})_{\omega_h^+} + y_N v_N - y_0 v_1. \quad (6.59)$$

Справді

$$\begin{aligned} (y, v_x)_{\omega_h} &= \sum_{i=1}^{N-1} y_i v_{x,i} h = \sum_{i=1}^{N-1} y_i (v_{i+1} - v_i) h = \\ &= \sum_{i=2}^N y_{i-1} v_i h - \sum_{i=1}^{N-1} y_i v_i h = \\ &= \sum_{i=1}^N y_{i-1} v_i h - y_0 v_1 h - \sum_{i=1}^N y_i v_i h + y_N v_N h = \\ &= - \sum_{i=1}^N v_i (y_i - y_{i-1}) h + y_N v_N h - y_0 v_1 h. \end{aligned}$$

Рівність (6.59) називається *формулою сумування за частинами*.

Підставляючи в (6.59) замість v вираз $az_{\bar{x}}$ і замість y функцію z , одержимо *першу різницеву формулу Гріна*

$$(z, (az_{\bar{x}})_x)_{\omega_h} = -(a, z_{\bar{x}}^2)_{\omega_h^+} + a_N z_{\bar{x},N} z_N - a_1 z_{x,0} z_0,$$

де

$$(a, z_{\bar{x}}^2)_{\omega_h^+} = \sum_{i=1}^N a_i z_{\bar{x},i}^2 h.$$

Зокрема, якщо $z_N = 0$, то

$$(z, (az_{\bar{x}})_x)_{\omega_h} = -(a, z_{\bar{x}}^2)_{\omega_h^+} - a_1 z_{x,0} z_0. \quad (6.60)$$

Позначимо

$$\|z_{\bar{x}}\|_{\omega_h^+}^2 = \sum_{i=1}^N z_{\bar{x},i}^2 h$$

і доведемо, що для будь-якої сіткової функції z_i , яка задовольняє умову $z_N = 0$, справджується нерівність

$$\|z\|_{C(\omega_h)}^2 \leq l \|z_{\bar{x}}\|_{\omega_h^+}^2. \quad (6.61)$$

Для доведення використаємо тотожність

$$z_i = - \sum_{j=i+1}^N h z_{\bar{x},j} = - \sum_{j=i+1}^N \sqrt{h} \left(\sqrt{h} z_{\bar{x},j} \right), \quad i = \overline{0, N-1}$$

і застосуємо нерівність Коші–Буняковського

$$\left| \sum_{j=i+1}^N a_j b_j \right|^2 \leq \left(\sum_{j=i+1}^N a_j^2 \right) \left(\sum_{j=i+1}^N b_j^2 \right).$$

Тоді одержимо

$$\begin{aligned} |z_i|^2 &\leq \left(\sum_{j=i+1}^N h \right) \left(\sum_{j=i+1}^N h z_{\bar{x},j}^2 \right) = \\ &= (l - x_i) \sum_{j=i+1}^N h z_{\bar{x},j}^2 \leq l \sum_{j=1}^N h z_{\bar{x},j}^2. \end{aligned}$$

Звідси випливає нерівність (6.61).

Доведемо тепер збіжність різницевої схеми (6.55), (6.56). Запишемо рівняння для похибки $z_i = y_i - u(x_i)$. Для цього підставимо у рівняння (6.55), (6.56) вираз $y_i = z_i + u(x_i)$, тоді будемо мати

$$(az_{\bar{x}})_{x,i} - d_i z_i = -\psi_i, \quad i = \overline{1, N-1}, \quad (6.62)$$

$$a_1 z_{x,0} = \bar{\beta}_1 z_0 - \psi_0, \quad z_N = 0. \quad (6.63)$$

Кажуть, що різницева схема *збігається*, якщо $\|z\|_{C(\omega_h)} \rightarrow 0$ при $h \rightarrow 0$. Різницева схема *збігається з порядком p* (має p -й порядок точності), якщо

$$\|z\|_{C(\omega_h)} \leq M h^p \quad \text{або} \quad \|z\|_{C(\omega_h)} = O(h^p), \quad p > 0,$$

де $M > 0$ — стала, яка не залежить від h .

Помножимо рівняння (6.62) на hz_i і підсумуємо по i від 1 до $N-1$. Тоді одержимо

$$((az_{\bar{x}})_x, z)_{\omega_h} - (d, z^2)_{\omega_h} = -(\psi, z)_{\omega_h}.$$

Застосовуючи різницеву формулу Гріна (6.60), будемо мати

$$(a, z_{\bar{x}}^2)_{\omega_h^+} + a_1 z_{x,0} z_0 + (d, z^2)_{\omega_h} = (\psi, z)_{\omega_h}.$$

На підставі (6.63)

$$a_1 z_{x,0} z_0 = \bar{\beta}_1 z_0^2 - \psi_0 z_0.$$

Тоді

$$(a, z_{\bar{x}}^2)_{\omega_h^+} + \bar{\beta}_1 z_0^2 + (d, z^2)_{\omega_h} = (\psi, z)_{\omega_h} + \psi_0 z_0. \quad (6.64)$$

Оскільки $k(x) \geq c_1 > 0$, $\beta_1 \geq 0$, $q(x) \geq 0$, то коефіцієнти різницевої схеми (6.55), (6.56) задовольняють нерівності

$$a_i \geq c_1 > 0, \quad \bar{\beta}_1 \geq 0, \quad d_i \geq 0. \quad (6.65)$$

Використовуючи (6.65), оцінимо доданки, які входять у ліву частину тотожності (6.64), таким чином:

$$(a, z_{\bar{x}}^2)_{\omega_h^+} = \sum_{i=1}^N a_i z_{\bar{x},i}^2 h \geq c_1 \sum_{i=1}^N z_{\bar{x},i}^2 h = c_1 \|z_{\bar{x}}\|_{\omega_h^+}^2,$$

$$\bar{\beta}_1 z_0^2 \geq 0, \quad (d, z^2)_{\omega_h} \geq 0.$$

Отже, будемо мати нерівність

$$c_1 \|z_{\bar{x}}\|_{\omega_h^+}^2 \leq |(\psi, z)_{\omega_h}| + |\psi_0| \cdot |z_0|. \quad (6.66)$$

Оцінимо праву частину цієї нерівності, тоді

$$\begin{aligned} |(\psi, z)_{\omega_h}| + |\psi_0| \cdot |z_0| &\leq \sum_{i=1}^{N-1} |\psi_i| \cdot |z_i| h + |\psi_0| \cdot |z_0| \leq \\ &\leq \|z\|_{C(\omega_h)} \left(\sum_{i=1}^{N-1} |\psi_i| h + |\psi_0| \right). \end{aligned}$$

Підставляючи цю оцінку в (6.66) і враховуючи нерівність (6.61), одержимо

$$\frac{c_1}{l} \|z\|_{C(\omega_h)}^2 \leq \left(\sum_{i=1}^{N-1} |\psi_i| h + |\psi_0| \right) \|z\|_{C(\omega_h)},$$

тобто

$$\|z\|_{C(\omega_h)} \leq \frac{l}{c_1} \left(\sum_{i=1}^{N-1} |\psi_i| h + |\psi_0| \right).$$

Звідси

$$\|z\|_{C(\omega_h)} \leq \frac{l}{c_1} \left(l \|\psi\|_{C(\omega_h)} + |\psi_0| \right). \quad (6.67)$$

Оскільки $\|\psi\|_{C(\omega_h)} = O(h^2)$, $|\psi_0| = O(h^2)$, то з нерівності (6.67) випливає, що похибка $z_i = y_i - u(x_i)$ є величиною $O(h^2)$ при $h \rightarrow 0$.

Отже, справджується твердження. Нехай $k(x)$ — неперервно диференційовна і $q(x)$, $f(x)$ — неперервні функції при $x \in [0, l]$, розв'язок $u(x)$ задачі (6.47), (6.48) має четверту неперервну похідну і коефіцієнти різницевої схеми задовольняють умови (6.57). Тоді розв'язок різницевої задачі (6.55), (6.56) збігається до розв'язку диференціальної задачі (6.47), (6.48) з другим порядком точності так, що виконується оцінка

$$\|y - u\|_{C(\omega_h)} \leq M h^2,$$

де M — стала, яка не залежить від h .

6.4. Варіаційно-проекційні методи та метод скінченних елементів

6.4.1. Метод Рітца

Крайові задачі для звичайних диференціальних рівнянь можна розглядати в операторній формі

$$Au = f, \quad u \in D(A), \quad (6.68)$$

де A — диференціальний оператор, $D(A)$ — область визначення оператора A , тобто деяка множина функцій, які задовольняють граничні умови.

Будемо розглядати операторне рівняння (6.68) в гільбертовому просторі H зі скалярним добутком (u, v) за умов, що $f \in H$ і оператор A —

самоспряжений і додатно визначений, тобто $(Au, v) = (u, Av)$, $(Au, u) \geq \gamma^2 \|u\|^2 \forall u, v \in D(A)$, де $\gamma = \text{const} > 0$. Згідно з ідеєю варіаційних методів замінимо задачу розв'язування рівняння (6.68) *задачею мінімізації функціоналу*

$$F(u) = (Au, u) - 2(f, u), \quad u \in D(A), \quad (6.69)$$

який називається *функціоналом енергії*. Доведемо, що ці задачі еквівалентні.

► **ТЕОРЕМА 6.1.** Нехай A — лінійний самоспряжений додатно визначений оператор з областю визначення $D(A)$ цільною в гільбертовому просторі H зі скалярним добутком. Якщо рівняння (6.68) має розв'язок $u_0 \in D(A)$, то функціонал енергії (6.69) набуває на u_0 мінімального значення в $D(A)$. З іншого боку, якщо функціонал (6.69) набуває на u_0 мінімального значення в $D(A)$, то u_0 є розв'язком рівняння (6.68).

Доведення. Нехай u_0 — розв'язок рівняння (6.68), тобто $f = Au_0$. Підставимо Au_0 в рівність (6.69) замість f . Враховуючи самоспряженість оператора A , одержимо

$$\begin{aligned} F(u) &= (Au, u) - 2(Au_0, u) = (Au, u) - (Au_0, u) - (u, Au_0) = \\ &= (Au, u) - (Au_0, u) - (Au, u_0) = \\ &= (Au, u) - (Au_0, u) - (Au, u_0) + (Au_0, u_0) - (Au_0, u_0) = \\ &= (A(u - u_0), u) - (A(u - u_0), u_0) - (Au_0, u_0) = \\ &= (A(u - u_0), u - u_0) - (Au_0, u_0). \end{aligned}$$

Оскільки оператор A додатно визначений, то $F(u) > -(Au_0, u_0) = -F(u_0)$. Отже, функціонал $F(u)$ набуває свого мінімального значення в $D(A)$ на елементі $u = u_0$, а це доводить першу частину теореми.

Нехай тепер функціонал $F(u)$ набуває свого мінімального значення в $D(A)$ на елементі u_0 . Це означає, що якщо ми виберемо довільний елемент $v \in D(A)$ і довільне дійсне число t так, що $u_0 + tv \in D(A)$, то

$$F(u_0 + tv) \geq F(u_0). \quad (6.70)$$

З рівності

$$\begin{aligned} F(u_0 + tv) &= (A(u_0 + tv), u_0 + tv) - 2(u_0 + tv, f) = \\ &= (Au_0, u_0) - 2(u_0, f) + t(Av, u_0) + \\ &\quad + t(Au_0, v) - 2t(v, f) + t^2(Av, v), \end{aligned}$$

враховуючи самоспряженість оператора A , одержимо

$$F(u_0 + tv) = F(u_0) + 2t(Au_0 - f, v) + t^2(Av, v).$$

Тоді нерівність (6.70) буде мати вигляд

$$2t(Au_0 - f, v) + t^2(Av, v) \geq 0, \quad \forall v \in D(A), t \in \mathbb{R}^1.$$

Квадратний тричлен відносно t з додатнім коефіцієнтом $(Av, v) \geq \gamma^2 \|v\|^2$ буде невід'ємним $\forall t \in \mathbb{R}^1$, якщо його дискримінант недовід'ємний, тобто

$$[(Au_0 - f, v)]^2 \leq 0, \quad \forall v \in D(A).$$

Звідси

$$(Au_0 - f, v) = 0, \quad \forall v \in D(A).$$

Отже, оскільки елемент $Au_0 - f$ ортогональний в H до всіх елементів множини $D(A)$, яка є щільною в H , то

$$Au_0 = f,$$

що і треба було довести. ■

Розглянемо тепер *метод Рунца* для наближеного розв'язування задачі (6.68). Виберемо базис $\{\varphi_i\}_{i=1}^\infty$, $\varphi_i \in D(A)$, $i = 1, 2, \dots$ і побудуємо послідовність підпросторів H_n , які є лінійною оболонкою елементів φ_i , $i = \overline{1, n}$. Будемо шукати елемент $u_n \in H_n$, на якому

$$F(u_n) = \inf_{w \in H_n} F(w).$$

Оскільки $u_n \in H_n$, то його можна записати у вигляді

$$u_n = \sum_{i=1}^n y_i \varphi_i,$$

де y_i — дійсні коефіцієнти, які треба визначити. Підставимо u_n у формулу для $F(u)$, тоді

$$F(u_n) = \sum_{i=1}^n \sum_{j=1}^n \alpha_{ij} y_i y_j - 2 \sum_{i=1}^n \beta_i y_i,$$

де

$$\alpha_{ij} = (A\varphi_i, \varphi_j), \quad \beta_i = (f, \varphi_i).$$

Величини $\alpha_{ij} = \alpha_{ji}$, тому що A — самоспряжений оператор. Функція $F(u_n)$ є функцією n коефіцієнтів y_1, y_2, \dots, y_n . Необхідною умовою мінімуму $F(u_n)$ є система рівнянь

$$\frac{\partial F(u_n)}{\partial y_j} = 2 \sum_{i=1}^n \alpha_{ij} y_i - 2\beta_j = 0, \quad j = \overline{1, n}.$$

Отже, для визначення y_i , $i = \overline{1, n}$ ми одержали систему n лінійних алгебраїчних рівнянь

$$\sum_{i=1}^n \alpha_{ij} y_i = \beta_j, \quad j = \overline{1, n}. \quad (6.71)$$

Визначник системи (6.71) є визначник Грама системи лінійно незалежних елементів і тому відмінний від нуля. Це означає, що система лінійних алгебраїчних рівнянь (6.71) має єдиний розв'язок.

Для дослідження збіжності методу Рітца введемо на множині $D(A)$ скалярний добуток

$$(u, v)_A = (Au, v), \quad u, v \in D(A)$$

і відповідну норму

$$\|u\|_A = (u, u)_A^{1/2}.$$

Такі скалярний добуток і норму називають ще енергетичним скалярним добутком і енергетичною нормою. Поповнимо $D(A)$ за введеною нормою до гільбертового простору H_A , який називають *енергетичним простором*, породженим оператором A . Кожна функція з $D(A)$ належить простору H_A , однак в результаті поповнення в H_A можуть з'явитися елементи, які не входять в $D(A)$. З означення випливає, що $D(A)$ є скрізь щільною в H_A множиною.

Функціонал $F(u)$ можна розглядати не тільки в області визначення оператора A , але і в енергетичному просторі H_A . Тому розширимо функціонал (6.69) на весь простір H_A :

$$F(u) = (u, u)_A - 2(u, f), \quad u \in H_A. \quad (6.72)$$

Задача знаходження мінімуму функціоналу $F(u)$ в енергетичному просторі H_A має єдиний розв'язок. Дійсно, оскільки оператор A — додатно визначений, тобто $(Au, u) = (u, u)_A \geq \gamma^2 \|u\|^2$, $u \in D(A)$, $\gamma = \text{const} > 0$, то в результаті поповнення $D(A)$ до H_A співвідношення

$(u, u)_A \geq \gamma^2 \|u\|^2$ справджується для будь-якого елемента $u \in H_A$. Розглянемо лінійний функціонал (u, f) . Зауважимо, що цей функціонал обмежений в H_A :

$$|(u, f)| \leq \|u\| \cdot \|f\| \leq \frac{1}{\gamma} \|u\|_A \cdot \|f\| = C \|u\|_A.$$

Отже, на основі теореми Рісса [21] існує єдиний елемент $u_0 \in H_A$, такий, що для будь-якого $u \in H_A$ справджується тотожність

$$(u, f) = (u, u_0)_A.$$

Тоді функціонал $F(u)$ можна записати у вигляді

$$\begin{aligned} F(u) &= (u, u)_A - 2(f, u) = \\ &= (u, u)_A - 2(u, u_0)_A + (u_0, u_0)_A - (u_0, u_0)_A = \\ &= \|u - u_0\|_A^2 - \|u_0\|_A^2, \quad u \in H_A. \end{aligned} \quad (6.73)$$

З формули (6.73) випливає, що мінімум функціоналу досягається тоді і тільки тоді, коли $u = u_0$. Функцію u_0 називають *узагальненим розв'язком* рівняння (6.68). Якщо $u_0 \in D(A)$, то u_0 буде також класичним розв'язком задачі (6.68).

Отже, ми звели вихідну задачу до задачі мінімізації функціоналу (6.72) в енергетичному просторі H_A . Розглянемо тепер метод Рітца для наближеного розв'язування останньої варіаційної задачі. Введемо послідовність скінченновимірних підпросторів $H_n \subseteq H_A, n = 1, 2, \dots$. Припустимо, що послідовність $\{H_n\}, n = 1, 2, \dots$, повна в H_A , тобто для будь-яких $u \in H_A$ і $\varepsilon > 0$, $\exists N(u, \varepsilon)$ таке, що

$$\inf_{w \in H_n} \|u - w\|_A < \varepsilon \quad (6.74)$$

для всіх $n > N$. Інакше кажучи, повнота послідовності підпросторів $\{H_n\}$ означає, що будь-який елемент $u \in H_A$ може бути з будь-якою точністю апроксимований елементами простору H_n . Метод Рітца в енергетичному просторі H_A полягає в тому, щоб знайти елемент $u_n \in H_n$, який мінімізує (6.72) в просторі H_n .

► **ТЕОРЕМА 6.2.** Якщо послідовність $\{H_n\}$ повна в H_A , то наближений розв'язок u_n , побудований методом Рітца, збігається при $n \rightarrow \infty$ в H_A до узагальненого розв'язку задачі u_0 .

Доведення. Дійсно, кожне u_n мінімізує функціонал $F(u)$ на H_n , а тому з врахуванням (6.73) при $\forall w \in H_n$ маємо

$$\|u_0 - u_n\|_A^2 = F(u_n) - F(u_0) \leq F(w) - F(u_0) = \|u_0 - w\|_A^2.$$

Оскільки $w \in H_n$ довільне, то згідно з (6.74) одержимо

$$\|u_0 - u_n\|_A \leq \inf_{w \in H_n} \|u_0 - w\|_A \xrightarrow{n \rightarrow \infty} 0.$$

■

Розглянемо питання виділення головних і природних граничних умов у методі Рітца. При встановленні області визначення оператора A , тобто множини $D(A)$, на $u \in D(A)$ часто накладають деякі граничні умови. Виявляється, що при побудові енергетичного простору H_A в результаті поповнення $D(A)$ в метриці $\|\cdot\|_A$ в H_A можуть з'явитися елементи, які задовольняють не всі граничні умови. Якщо в H_A будуть елементи, які не задовольняють деяку граничну умову, то ця умова називається природною для оператора A . Гранична умова, яку задовольняють як елементи з $D(A)$, так і елементи з H_A , називається головною.

Оскільки в методі Рітца базисні функції $\{\varphi_i\}_{i=1}^n$ достатньо брати тільки з енергетичного простору (і не обов'язково з $D(A)$), то не обов'язково, щоб вони задовольняли природні граничні умови. Ця обставина полегшує вибір φ_i при розв'язуванні практичних задач, тому що φ_i повинні задовольняти тільки головні граничні умови.

Наведемо просту ознаку, яка дозволяє відрізнити природні граничні умови від головних. Нехай в (6.68) A — диференціальний оператор порядку $2m$, визначений на множині функцій $D(A)$, які задовольняють деякі однорідні граничні умови. Тоді граничні умови, що містять похідні до порядку $m-1$ включно, є головними. Граничні умови, що містять похідні порядку m та вище є природними граничними умовами.

6.4.2. Метод Гальоркіна

Розглянемо сепарабельний гільбертовий простір H і його щільну множину M . Якщо для деякого елемента $u \in H$ при виконується

$$(u, v) = 0,$$

то $u = 0$ в H . Нехай тепер $\{\varphi_i\}_{i=1}^\infty$ утворюють базис в H . Якщо $(u, \varphi_i) = 0$, $i = 1, 2, \dots$, то $u = 0$ в H . Дійсно, за припущенням $\{\varphi_i\}_{i=1}^\infty$ утворює

базис в H , а тому множина H_n всіх елементів вигляду

$$\sum_{i=1}^n y_i \varphi_i, \quad (6.75)$$

де y_i — довільні дійсні числа, є щільною в H . Оскільки для всіх i виконується $(u, \varphi_i) = 0$, то

$$\left(u, \sum_{i=1}^n y_i \varphi_i \right) = 0$$

для всіх елементів (6.75) з H_n . Звідси випливає, що $u = 0$.

Розглянемо тепер операторне рівняння (6.68) в гільбертовому просторі H . Якщо існує такий елемент $u_0 \in D(A)$, що для всіх $i = 1, 2, \dots$

$$(Au_0 - f, \varphi_i) = 0, \quad (6.76)$$

то

$$Au_0 - f = 0 \quad \text{в} \quad H. \quad (6.77)$$

Отже, u_0 є розв'язком рівняння (6.68) в H .

Метод Гальоркіна ґрунтується на твердженні, що з (6.76) випливає (6.77).

Нехай базис $\{\varphi_i\}_{i=1}^{\infty}$ і область визначення $D(A)$ оператора A такі, що будь-яка лінійна комбінація елементів цього базису належить $D(A)$, і нехай наближений розв'язок рівняння (6.68) шукається у вигляді

$$u_n = \sum_{i=1}^n y_i \varphi_i,$$

де y_i — невідомі коефіцієнти. У методі Гальоркіна ці коефіцієнти визначаються з умов ортогональності нев'язки $Au_n - f$ до $\varphi_1, \varphi_2, \dots, \varphi_n$:

$$(Au_n - f, \varphi_j) = 0, \quad j = \overline{1, n}$$

або

$$\sum_{i=1}^n (A\varphi_i, \varphi_j) y_i = (f, \varphi_j), \quad j = \overline{1, n}.$$

Зауважимо, що останні рівняння за формою збігаються з відповідними рівняннями (6.71) алгоритму Рітца (якщо $\varphi_i \in D(A)$). Отже, якщо

A — самоспряжений і додатно визначений оператор, то методи Гальоркіна і Рітца збігаються. Однак, можливості застосування методу Гальоркіна суттєво ширші, ніж методу Рітца. На відміну від методу Рітца метод Гальоркіна можна застосовувати і для задач з несамоспряженими і незначковизначеними операторами. Система рівнянь відносно невідомих y_i у методі Гальоркіна одержується проектуванням нев'язки $Au_n - f$ на базисні функції φ_j та прирівнюванням результату до нуля. Тому метод Гальоркіна називають *проекційним* методом.

6.4.3. Побудова сіткової схеми методом скінченних елементів

Метод скінченних елементів — це метод Гальоркіна (або Рітца), який використовує спеціальні базисні функції, що приводить до системи лінійних алгебраїчних рівнянь з розрідженою матрицею та добрими обчислювальними властивостями.

Нехай на $\Omega = (a, b)$ треба знайти розв'язок рівняння

$$(k(x)u')' - q(x)u(x) = -f(x), \quad (6.78)$$

який задовольняє граничні умови

$$u(a) = 0, \quad u(b) = 0. \quad (6.79)$$

Тут $f(x) \in L_2(a, b)$, $k(x), q(x)$ — обмежені функції $0 < c_1 \leq k(x) \leq c_2$, $0 \leq q(x) \leq c_3$.

Позначимо через A оператор задачі (6.78), (6.79), який визначається виразом $Au = -(k(x)u')' + q(x)u$ і областю визначення $D(A)$. Нехай $D(A)$ — це множина неперервних функцій $u(x)$, які мають похідні $u'(x) \in L_2(\Omega)$, таких, що $Au \in L_2(\Omega)$ і $u(a) = u(b) = 0$. Тепер задачу (6.78), (6.79) можна записати у вигляді операторного рівняння $Au = f$, яке будемо розглядати в гільбертовому просторі $H = L_2(\Omega)$ зі скалярним добутком

$$(u, v) = \int_a^b u(x)v(x)dx$$

і нормою

$$\|u\| = (u, u)^{1/2} = \left(\int_a^b u^2(x)dx \right)^{1/2}.$$

Зауважимо, що $\forall u, v \in D(A)$

$$\begin{aligned}
 (Au, v) &= \int_a^b [-(k(x)(u'(x)))' + q(x)u(x)] v(x) dx = \\
 &= - \int_a^b v(x) d(k(x)u'(x)) + \int_a^b q(x)u(x)v(x) dx = \\
 &= -k(x)u'(x)v(x)|_a^b + \int_a^b k(x)u'(x)v'(x) dx + \\
 &\quad + \int_a^b q(x)u(x)v(x) dx = \\
 &= \int_a^b k(x)u'(x)v'(x) dx + \int_a^b q(x)u(x)v(x) dx. \tag{6.80}
 \end{aligned}$$

Для наближеного розв'язування задачі (6.78), (6.79) застосуємо метод Гальоркіна.

Базисні функції зручно вибирати як функції зі скінченним (фінітним) носієм, тобто такі, кожна з яких тільки на деякому невеликому околі відмінна від нуля, а зовні цього околу тотожно рівна нулю. До найбільш поширених фінітних функцій належать *кусково-лінійні фінітні функції*.

Для побудови кусково-лінійних базисних функцій введемо на $\bar{\Omega} = [a, b]$ сітку $\bar{\omega}_h = \{x_i = a + ih, i = \overline{0, n}, h = (b - a)/n\}$, розбивши $[a, b]$ на n підобластей $\Omega_i = (x_{i-1}, x_i)$, $i = \overline{1, n}$, які називають скінченними елементами. Поставимо у відповідність кожному вузлу сітки неперервну функцію

$$\begin{aligned}
 \varphi_i(x) &= \begin{cases} \frac{x - x_{i-1}}{h}, & x \in (x_{i-1}, x_i), \\ -\frac{x - x_{i+1}}{h}, & x \in (x_i, x_{i+1}), \\ 0, & x \notin (x_{i-1}, x_{i+1}), \end{cases} \quad i = \overline{1, n-1}, \\
 \varphi_0(x) &= \begin{cases} -\frac{x - x_1}{h}, & x \in (x_0, x_1), \\ 0, & x \notin (x_0, x_1), \end{cases}
 \end{aligned}$$

$$\varphi_n(x) = \begin{cases} \frac{x - x_{n-1}}{h}, & x \in (x_{n-1}, x_n), \\ 0, & x \notin (x_{n-1}, x_n). \end{cases}$$

Зауважимо, що кусково-лінійна функція $\varphi_i(x)$ дорівнює одиниці у вузлі x_i і нулю в усіх інших вузлах, її похідна $\varphi'_i(x)$ — це кусково-стала функція

$$\varphi'_i(x) = \begin{cases} \frac{1}{h}, & x \in (x_{i-1}, x_i), \\ -\frac{1}{h}, & x \in (x_i, x_{i+1}), \\ 0, & x \notin (x_{i-1}, x_{i+1}), \end{cases} \quad i = \overline{1, n-1},$$

яка має розриви в точках x_{i-1}, x_i, x_{i+1} , але вона сумовна (інтегровна) з будь-яким степенем. Кусково-лінійні функції $\varphi_i(x)$, $i = \overline{0, n}$ — лінійно незалежні і утворюють базис в евклідовому просторі \mathbb{R}^{n+1} розміру $n+1$. Лінійну оболонку елементів $\{\varphi_i\}$ позначимо через H_n . Функції з H_n є неперервними кусково-лінійними функціями, які мають сумовну з будь-яким степенем першу похідну. Отже, $H_n \subset W_2^1$, де $W_2^1 = W_2^1(\Omega) = \{u \mid u, u' \in L_2(\Omega)\}$ — гільбертовий простір Соболева, скалярний добуток і норма в якому визначаються формулами:

$$(u, v)_{W_2^1} = \int_a^b [u(x)v(x) + u'(x)v'(x)] dx, \quad \|u\|_{W_2^1} = (u, u)_{W_2^1}^{1/2}.$$

Якщо функція u належить простору Соболева $W_2^1(\Omega)$ і, крім того, задовольняє умови (6.79), то такий простір будемо позначати $\overset{\circ}{W}_2^1(\Omega)$.

Наближений розв'язок задачі (6.78), (6.79) шукаємо у вигляді

$$u_n(x) = \sum_{i=0}^n y_i \varphi_i(x).$$

Оскільки $u_n(x_i) = y_i$, то невідомі коефіцієнти y_i збігаються зі значеннями шуканого наближеного розв'язку у вузлах сітки. Будемо вимагати, щоб розв'язок $u_n(x)$ задовольняв головні граничні умови задачі, тобто щоб $u_n(a) = y_0 = u_n(b) = y_n = 0$. Цим вимогам буде задовольняти лінійна комбінація вигляду

$$u_n(x) = \sum_{i=1}^{n-1} y_i \varphi_i(x). \quad (6.81)$$

Множину таких лінійних комбінацій позначимо через $\overset{\circ}{H}_n \subset \overset{\circ}{W}_2^1$.

Згідно з методом Гальоркіна коефіцієнти y_i наближеного розв'язку задачі (6.81) знаходяться з умов ортогональності нев'язки $Au_n - f$ до $\varphi_1, \varphi_2, \dots, \varphi_n$, тобто з СЛАР

$$\sum_{i=1}^{n-1} \alpha_{ij} y_i = \beta_j, \quad j = \overline{1, n-1}, \quad (6.82)$$

де

$$\alpha_{ij} = (A\varphi_i, \varphi_j) = \int_a^b [k(x)\varphi_i'(x)\varphi_j'(x) + q(x)\varphi_i(x)\varphi_j(x)] dx,$$

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

Зауважимо, що оскільки

$$\int_a^b \varphi_i(x)\varphi_j(x) dx \begin{cases} = 0, & |i-j| > 1, \\ \neq 0, & |i-j| \leq 1, \end{cases}$$

то $\alpha_{ij} \neq 0$ тільки при $j = i-1$, $j = i$, $j = i+1$. Звідси випливає, що (6.82) — це система лінійних алгебраїчних рівнянь з тридіагональною матрицею і її можна розв'язати методом прогонки.

Розглянемо тепер рівняння (6.78) з граничними умовами

$$u(a) = 0, \quad u'(b) = 0.$$

У цьому випадку наближений розв'язок повинен задовольняти умову $u(a) = 0$ (ця умова головна), але в H_A можуть виявитися функції, які не задовольняють другу граничну умову (ця умова — природна). Тоді

$$u_n(x) = \sum_{i=1}^n y_i \varphi_i(x).$$

У випадку задачі з граничними умовами

$$u'(a) = u'(b) = 0$$

функцію $u_n(x)$ записують у вигляді

$$u_n(x) = \sum_{i=0}^n y_i \varphi_i(x).$$

Для задачі з неоднорідними граничними умовами

$$u(a) = \mu_1, \quad u(b) = \mu_2$$

наближений розв'язок шукають у вигляді

$$u_n(x) = \sum_{i=0}^n y_i \varphi_i(x)$$

і вимагають, щоб $u_n(a) = \mu_1$, $u_n(b) = \mu_2$. Тоді $y_0 = \mu_1$, $y_n = \mu_2$.

Задачу з неоднорідними граничними умовами можна також звести до задачі з однорідними граничними умовами.

6.4.4. Збіжність методу скінченних елементів

Вивчимо властивості оператора A . Насамперед зауважимо, що множина $D(A)$ — щільна в L_2 і що оператор A — самоспряжений. Дійсно згідно з (6.80)

$$(Au, v) = \int_a^b k(x)u'(x)v'(x)dx + \int_a^b q(x)u(x)v(x)dx = (Av, u).$$

Покажемо, що оператор A додатно визначений, тобто для нього виконується умова

$$\gamma^2 \|u\|^2 \leq (Au, u), \quad u \in D(A), \quad (6.83)$$

де $\gamma = \text{const} > 0$. Для доведення цього факту використаємо *нерівність Фрідрікса*: якщо $u(x)$ має похідну $u'(x)$ на (a, b) , сумовну з квадратом і в одній з точок $x = a$, $x = b$ ця функція дорівнює нулю, тоді справджується оцінка

$$\|u\| \leq C \|u'\|. \quad (6.84)$$

Доведемо спочатку нерівність Фрідрікса. Нехай $u(a) = 0$. Тоді з тотожності

$$u(x) = \int_a^x u'(\xi)d\xi,$$

на основі нерівності Коші–Буняковського, будемо мати

$$\begin{aligned} u^2(x) &= \left(\int_a^x u'(\xi) d\xi \right)^2 \leq \\ &\leq \int_a^x d\xi \int_a^x (u'(\xi))^2 d\xi \leq (x-a) \int_a^b (u'(\xi))^2 d\xi. \end{aligned}$$

Інтегруючи ліву і праву частини останньої нерівності по x , отримаємо (6.84) при $C = (b-a)/\sqrt{2}$.

На основі нерівності Фрідрікса маємо

$$\begin{aligned} \|u\|^2 &\leq \frac{(b-a)^2}{2} (u', u') \leq \frac{(b-a)^2}{2c_1} (ku', u') \leq \\ &\leq \frac{(b-a)^2}{2c_1} [(ku', u') + q \|u\|^2] = \frac{(b-a)^2}{2c_1} (Au, u), \end{aligned}$$

тобто одержимо (6.83) при $\gamma = \sqrt{2c_1}/(b-a)$.

Властивості оператора A гарантують існування оберненого оператора A^{-1} , а отже, і однозначну розв'язність задачі (6.78), (6.79).

Дослідимо збіжність наближеного розв'язку $u_n(x)$ до точного $u(x)$ задачі (6.78), (6.79). Введемо енергетичний простір H_A , який відповідає оператору A . Скалярний добуток і норма в ньому будуть мати вигляд

$$\begin{aligned} (u, v)_A &= \int_a^b [k(x)u'(x)v'(x) + q(x)u(x)v(x)] dx, \\ \|u\|_A &= (u, u)_A^{1/2}. \end{aligned}$$

Оскільки A — самоспряжений додатно визначений оператор, то метод Гальоркіна збігається з методом Рітца. Для наближеного розв'язку за методом Рітца справджується нерівність $\|u - u_n\|_A \leq \|u - v_n\|_A$, де $v_n = \sum_{i=1}^{n-1} w_i \varphi_i(x)$ — довільна функція з $\overset{\circ}{H}_n$.

З нерівностей

$$\begin{aligned}
 \|u\|_A^2 &= \int_a^b [k(x)(u'(x))^2 + q(x)(u(x))^2] dx \leq \\
 &\leq \max(c_2, c_3) \int_a^b [(u'(x))^2 + (u(x))^2] dx = \\
 &= \max(c_2, c_3) \|u\|_{W_2^1}^2, \\
 \|u\|_A^2 &\geq c_1 \int_a^b [(u')^2 + u^2] dx \geq c_1 \|u\|_{W_2^1}^2
 \end{aligned}$$

випливає, що

$$\gamma_1 \|u\|_{W_2^1} \leq \|u\|_A \leq \gamma_2 \|u\|_{W_2^1}, \quad \gamma_1, \gamma_2 > 0. \quad (6.85)$$

Тоді для нашої задачі H_A збігається з простором $\overset{\circ}{W}_2^1(\Omega)$.
Враховуючи (6.85), одержимо

$$\begin{aligned}
 \|u - u_n\|_{W_2^1} &\leq c \|u - v_n\|_{W_2^1}, \\
 \|u - u_n\|_{W_2^1} &\leq c \inf_{v_n \in H_n} \overset{\circ}{\|} u - v_n \|_{W_2^1}.
 \end{aligned} \quad (6.86)$$

Отже, задача оцінки похибки методу скінченних елементів зводиться до задачі апроксимації функцій.

► **ТЕОРЕМА 6.3.** Якщо $u \in W_2^2$, то існує така функція $u_I \in H_n$, що

$$\|u - u_I\|_{W_2^1} \leq ch \|u\|_{W_2^2}, \quad (6.87)$$

де стала c не залежить від h і від $u(x)$, $W_2^2 = W_2^2(\Omega)$ — гільбертовий простір Соболева зі скалярним добутком та нормою

$$\begin{aligned}
 (u, v)_{W_2^2} &= \int_a^b [u(x)v(x) + u'(x)v'(x) + u''(x)v''(x)] dx, \\
 \|u\|_{W_2^2} &= (u, u)_{W_2^2}^{1/2}.
 \end{aligned}$$

Доведення. Розглянемо лінійну комбінацію

$$u_I(x) = \sum_{i=0}^n u(x_i) \varphi_i(x).$$

Оцінимо різницю $u - u_I$ в довільній точці $x \in (x_{i-1}, x_i)$. Для цього запишемо тотожність при $x \in (x_{i-1}, x_i)$:

$$\begin{aligned} u(x) - u_I(x) &= u(x) - u_{i-1} \frac{x_i - x}{h} - u_i \frac{x - x_{i-1}}{h} = \\ &= \frac{1}{h} \int_{x_{i-1}}^x d\xi \int_{x_{i-1}}^{x_i} d\eta \int_{\eta}^{\xi} u''(t) dt. \end{aligned} \quad (6.88)$$

Застосуємо до (6.88) нерівність Коші–Буняковського:

$$\begin{aligned} |u(x) - u_I(x)| &\leq \frac{1}{h} \int_{x_{i-1}}^x d\xi \int_{x_{i-1}}^{x_i} d\eta \int_{x_{i-1}}^{x_i} |u''(t)| dt \leq \\ &\leq h^{3/2} \left(\int_{x_{i-1}}^{x_i} |u''(t)|^2 dt \right)^{1/2}, \quad x \in (x_{i-1}, x_i). \end{aligned}$$

Отже,

$$\int_{x_{i-1}}^{x_i} |u(x) - u_I(x)|^2 dx \leq h^4 \int_{x_{i-1}}^{x_i} |u''(x)|^2 dx.$$

Сумуючи останню нерівність по $i = \overline{1, n}$, отримаємо

$$\begin{aligned} \sum_{i=1}^n \int_{x_{i-1}}^{x_i} |u(x) - u_I(x)|^2 dx &= \|u - u_I\|_{L_2}^2 \leq \\ &\leq h^4 \|u''\|_{L_2}^2 \leq h^4 \|u\|_{W_2^2}^2. \end{aligned} \quad (6.89)$$

Продиференціюємо (6.88), тоді будемо мати

$$u'(x) - u'_I(x) = \frac{1}{h} \int_{x_{i-1}}^{x_i} d\eta \int_{\eta}^x u''(t) dt.$$

Якщо тепер провести аналогічні міркування відносно $u'(x) - u'_I(x)$, то

$$\begin{aligned} |u'(x) - u'_I(x)| &\leq \frac{1}{h} \int_{x_{i-1}}^{x_i} d\eta \int_{x_{i-1}}^{x_i} |u''(t)| dt \leq \\ &\leq h^{1/2} \left(\int_{x_{i-1}}^{x_i} |u''(t)|^2 dt \right)^{1/2}, \end{aligned}$$

$$\int_{x_{i-1}}^{x_i} |u'(x) - u'_I(x)|^2 dx \leq h^2 \int_{x_{i-1}}^{x_i} |u''(t)|^2 dt.$$

Враховуючи (6.89), отримаємо

$$\begin{aligned} \|u - u_I\|_{W_2^2}^2 &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} [|u' - u'_I|^2 + |u - u_I|^2] dx \leq \\ &\leq h^2(1 + h^2) \sum_{i=1}^n \int_{x_{i-1}}^{x_i} |u''(t)|^2 dt \leq \\ &\leq c^2 h^2 \|u'\|_{L_2}^2 \leq c^2 h^2 \|u\|_{W_2^2}^2. \end{aligned}$$

■

На підставі (6.87) з нерівності (6.86) будемо мати

$$\|u - u_n\|_{W_2^1} \leq c h \|u\|_{W_2^2}.$$

Отже, $u_n(x)$ збігається до $u(x)$ при $h \rightarrow 0$.

Якщо припустити, що $k(x)$ має обмежену похідну $k'(x)$, то можна показати, що

$$\|u - u_n\|_{W_2^1} \leq c h \|u\|_{W_2^2} \leq c h \|f\|.$$

Контрольні завдання

🔗 **6.1.** Для якої точки відрізка $x \in [x_i, x_{i+1}]$ похибка апроксимації оператора u' різницеvim оператором u_x збільшується на порядок?

🔗 **6.2.** Дослідіть похибку апроксимації оператора u' різницеvim оператором $\sigma u_x + (1 - \sigma)u_{\bar{x}}$.

✎ 6.3. На сітці $\omega_h = \{x_i = a + ih, i = 0, 1, 2, \dots\}$ побудуйте різницеву апроксимацію для u''' та оцініть її похибку.

✎ 6.4. Те ж для $u^{(4)}$.

✎ 6.5. При яких значеннях α, β, γ різницева схема

$$y_{\bar{x}x,i} - (\alpha y_{i+1} + \beta y_i + \gamma y_{i-1}) = -f_i - \frac{h^2}{12} f_i'', \quad i = \overline{1, N-1},$$

$$y_0 = 0, \quad y_N = 0, \quad h = 1/N$$

апроксимує задачу

$$u'' - u = -f(x), \quad 0 < x < 1,$$

$$u(0) = u(1) = 0$$

з четвертим порядком?

✎ 6.6. Для диференціальної задачі

$$u'' - au = -\cos x, \quad 0 < x < \pi, \quad a > 0,$$

$$u(0) = 0, \quad u(\pi) = 1$$

на триточковому шаблоні x_{i-1}, x_i, x_{i+1} побудуйте схему десятого порядку апроксимації.

✎ 6.7. Дослідіть збіжність розв'язку різницевої схеми

$$y_{\bar{x}x,i} - (1 + ih)y_i = -(ih)^3 + ih + 2, \quad i = \overline{1, N-1},$$

$$y_0 = y_N = 0, \quad h = 1/N$$

до розв'язку диференціальної задачі

$$u'' - (1 + x)u = -x^3 + x + 2, \quad 0 < x < 1,$$

$$u(0) = u(1) = 0.$$

✎ 6.8. На нерівномірній сітці методом неозначених коефіцієнтів на триточковому шаблоні побудуйте різницеві схеми першого і другого порядків апроксимації для задачі

$$u'' = -f(x), \quad u(0) = \mu_1, \quad u(1) = \mu_2,$$

якщо $u \in C^{(4)}[0, 1]$.

 **6.9.** Доведіть, що система кусково-лінійних базисних функцій

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h}, & x \in (x_{i-1}, x_i), \\ -\frac{x - x_{i+1}}{h}, & x \in (x_i, x_{i+1}), \\ 0, & x \notin (x_{i-1}, x_{i+1}), \end{cases} \quad i = \overline{1, n-1},$$

$$\varphi_0(x) = \begin{cases} -\frac{x - x_1}{h}, & x \in (x_0, x_1), \\ 0, & x \notin (x_0, x_1), \end{cases}$$

$$\varphi_n(x) = \begin{cases} \frac{x - x_{n-1}}{h}, & x \in (x_{n-1}, x_n), \\ 0, & x \notin (x_{n-1}, x_n). \end{cases}$$

лінійно незалежна.


 **6.10.** Методом скінченних елементів побудуйте сіткову схему для задачі

$$(k(x)u')' = -1, \quad 0 < x < 1,$$

$$u(0) = u(1) = 0,$$


де

$$k(x) = \begin{cases} 3/2, & 0 \leq x < \pi/4, \\ 1, & \pi/4 \leq x \leq 1. \end{cases}$$

 **6.11.** Методом скінченних елементів побудуйте сіткову схему для задачі

$$u'' - au' - cu = -1, \quad c \geq 0, \quad 0 < x < 1,$$

$$u(0) = u(1) = 1.$$

 **6.12.** Доведіть, що для методу скінченних елементів оцінка швидкості збіжності в $L_2(a, b)$ має вигляд:

$$\|u - u_n\| \leq ch^2 \|f\|.$$

РОЗДІЛ 7

ЧИСЕЛЬНЕ РОЗВ'ЯЗУВАННЯ РІВНЯНЬ З ЧАСТИННИМИ ПОХІДНИМИ

7.1. Крайові задачі для рівнянь з частинними похідними

Для того, щоб повністю описати фізичний процес, необхідно, крім самого рівняння з частинними похідними, яке описує цей процес, задати початковий стан процесу (початкові умови) і режим на границі області (граничні умови). Додаткові умови (початкові та граничні) дозволяють виділити єдиний розв'язок диференціального рівняння.

Задачу, в якій є тільки початкові умови називають задачею Коші. Якщо задаються лише умови на границі області, то така задача називається крайовою. Задачу з початковими та граничними умовами називають мішаною задачею.

Розглянемо, наприклад, задачу: знайти функцію $u(x) \in C^{(2)}(\Omega)$, яка в області Ω задовольняє рівняння

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x), \quad x = (x_1, x_2) \in \Omega, \quad (7.1)$$

а на границі Γ області Ω одну з граничних умов:

$$u(x) = \mu(x), \quad x \in \Gamma, \quad (7.2)$$

$$\frac{\partial u(x)}{\partial \mathbf{n}} = \mu(x), \quad x \in \Gamma, \quad (7.3)$$

$$\frac{\partial u(x)}{\partial \mathbf{n}} + \alpha(x)u(x) = \mu(x), \quad x \in \Gamma, \quad (7.4)$$

де $\partial u / \partial \mathbf{n}$ — нормальна похідна за напрямком \mathbf{n} , $f(x)$, $\mu(x)$, $\alpha(x)$ — задані функції.

Рівняння (7.1) називається *рівнянням Пуассона*, яке, як відомо, є еліптичним рівнянням, а умови (7.2), (7.3), (7.4) називаються відповідно *крайовими умовами 1-го, 2-го та 3-го роду*. Крайова задача (7.1),

(7.2) ще називається *задачею Діріхле*, а задача (7.1), (7.3) — *задачею Неймана*.

Для рівняння параболічного типу (рівняння теплопровідності)

$$\frac{\partial u}{\partial t} = Lu + f(x, t), \quad x = (x_1, x_2) \in \Omega, \quad t \in (t_0, T] \quad (7.5)$$

мішана задача ставиться так: знайти функцію $u(x, t)$, яка задовольняє рівняння (7.5), початкову умову

$$u(x, t_0) = u_0(x), \quad x \in \bar{\Omega} = \Omega \cup \Gamma \quad (7.6)$$

та граничну умову

$$lu(x, t) = \mu(x, t), \quad x \in \Gamma, \quad (7.7)$$

де L — деякий еліптичний оператор (наприклад, $Lu = \Delta u$), l — оператор, який задає граничну умову 1-го, 2-го або 3-го роду.

Аналогічно ставиться *мішана задача для рівняння гіперболічного типу*: знайти функцію $u(x, t)$, яка задовольняє рівняння

$$\frac{\partial^2 u}{\partial t^2} = Lu + f(x, t), \quad x = (x_1, x_2) \in \Omega, \quad t \in (t_0, T], \quad (7.8)$$

початкові

$$u(x, t_0) = u_0(x), \quad \frac{\partial u(x, t_0)}{\partial t} = \bar{u}_0(x), \quad x \in \bar{\Omega} \quad (7.9)$$

та граничну умову (7.7).

Оскільки розв'язок задач (7.5) — (7.7) і (7.8), (7.9), (7.7) залежить від часу, то такі задачі називають нестационарними.

У цьому розділі розглядаються сіткові методи розв'язування рівнянь з частинними похідними. Їх можна застосовувати до широкого класу рівнянь і різних типів задач для них.

7.2. Основні поняття методу сіток

Для побудови сіткової схеми необхідно замінити область неперервної зміни аргументів дискретною множиною точок (сіткою), а диференціальне рівняння та додаткові умови — сітковими рівняннями, тобто системою алгебраїчних рівнянь.

Раніше ми вже розглядали приклади сіток в одновимірній області:

1) *рівномірна сітка* на відрізку $0 \leq x \leq l$:

$$\bar{\omega}_h = \{x_i = ih, \quad i = \overline{0, N}, \quad h = l/N\};$$

2) *нерівномірна сітка* на $[0, l]$:

$$\hat{\omega}_h = \{x_i, i = \overline{0, N}, x_0 = 0, x_N = l, h_{i+1} = x_{i+1} - x_i\}.$$

Розглянемо приклад сітки у двовимірній області. Нехай на площині $x = (x_1, x_2)$ задана область

$$\bar{\Omega} = \{0 \leq x_1 \leq l_1, 0 \leq x_2 \leq l_2\}$$

з границею Γ . На відрізках $0 \leq x_1 \leq l_1, 0 \leq x_2 \leq l_2$ побудуємо рівномірні сітки

$$\bar{\omega}_{h_1} = \{x_1^i = ih_1, i = \overline{0, N_1}, h_1 = l_1/N_1\},$$

$$\bar{\omega}_{h_2} = \{x_2^j = jh_2, j = \overline{0, N_2}, h_2 = l_2/N_2\}.$$

Множину вузлів $x_{ij} = (x_1^i, x_2^j)$ називають *сіткою у прямокутнику* $\bar{\Omega}$ і позначають

$$\bar{\omega}_h = \{x_{ij} = (ih_1, jh_2), i = \overline{0, N_1}, j = \overline{0, N_2}, \\ h_1 = l_1/N_1, h_2 = l_2/N_2\}.$$

Сітка $\bar{\omega}_h$ складається з точок перетину прямих $x_1 = x_1^i, x_2 = x_2^j$. Точки $\bar{\omega}_h$, які належать Γ , називають граничними і позначають через $\gamma_h = \{x_{ij} \in \Gamma\}$.

Нехай $\bar{\omega}_h$ — сітка, введена в одновимірній області, а x_i — вузли сітки. Функцію $y_i = y(x_i)$ дискретного аргумента x_i називають *сітковою функцією*, визначеною на $\bar{\omega}_h$. Аналогічно визначається сіткова функція на сітці $\bar{\omega}_h$ у двовимірній області. Якщо x_{ij} — вузол сітки $\bar{\omega}_h$, то $y_{ij} = y(x_{ij})$. Сіткові функції можна також розглядати як функції цілочисельного аргумента, який є номером вузла сітки.

Сіткову функцію y_i , задану на сітці $\bar{\omega}_h = \{x_i, i = \overline{0, N}\}$, можна записати у вигляді вектора розміру $N + 1$

$$y = (y_0, y_1, \dots, y_N).$$

Якщо $\bar{\omega}_h$ — сітка в прямокутнику, то сіткові функції y_{ij} , задані на $\bar{\omega}_h$, відповідає вектор

$$y = (y_{00}, \dots, y_{N_1 0}, y_{01}, \dots, y_{N_1 1}, \dots, y_{0 N_2}, \dots, y_{N_1 N_2})$$

розміру $(N_1 + 1)(N_2 + 1)$.

Як правило, розглядають множину сіток $\{\bar{\omega}_h\}$, які залежать від кроку h як від параметра, а тому сіткові функції $y = y_h(x)$ залежать від параметра h , якщо сітка рівномірна. У випадку нерівномірної сітки під h розуміють $h = (h_1, h_2, \dots, h_N)$, $|h| = \max_{1 \leq i \leq N} h_i$ або $|h| = \sqrt{\sum_{i=1}^N h_i^2}$.

Множина сіткових функцій $y_h(x)$ утворює простір H_h . У просторі H_h можна ввести норму $\|\cdot\|_h$. Вкажемо найпростіші типи норм:

$$\|y\| = \max_{x \in \bar{\omega}_h} |y(x)|$$

або

$$\|y\| = \max_{0 \leq i \leq N} |y_i|, \quad \|y\| = \left(\sum_{i=1}^{N-1} y_i^2 h \right)^{1/2}.$$

Нехай в області Ω евклідового простору \mathbb{R}^n з границею Γ необхідно знайти розв'язок лінійного диференціального рівняння

$$Lu = f(x), \quad x \in \Omega, \quad (7.10)$$

який задовольняє граничну умову

$$lu = \mu(x), \quad x \in \Gamma, \quad (7.11)$$

де $f(x), \mu(x)$ — задані функції, L, l — лінійні диференціальні оператори.

Введемо на $\bar{\Omega} = \Omega \cup \Gamma$ сітку $\bar{\omega}_h = \omega_h \cup \gamma_h$, $\omega_h \subset \Omega$, $\gamma_h \subset \Gamma$. Замінімо диференціальний оператор Lu в точці $x_i \in \omega_h$ лінійною комбінацією значень $u_h(x)$ сіткової функції на деякій множині вузлів сітки σ_i , яку називають *шаблоном*

$$(L_h u)_i = \sum_{x_j \in \sigma_i} a_{ij}^h u_h(x_j), \quad x_i \in \omega_h(\Omega),$$

де a_{ij}^h — коефіцієнти, $\sigma_i \in \bar{\omega}_h$. Поставимо у відповідність задачі (7.10), (7.11) сіткову (різницеву) задачу

$$L_h y_h = \varphi_h, \quad x \in \omega_h, \quad (7.12)$$

$$l_h y_h = \chi_h, \quad x \in \gamma_h, \quad (7.13)$$

де L_h, l_h — лінійні сіткові (різницеві) оператори, y_h, φ_h, χ_h — сіткові функції, які залежать від кроку сітки h . Змінюючи h , одержимо послідовності $\{y_h\}, \{\varphi_h\}, \{\chi_h\}$. Отже, ми розглядаємо множину задач (7.12), (7.13), яка залежить від параметра h . Цю множину задач називають *сітковою схемою*.

Зауважимо, що сіткову схему (7.12), (7.13) можна записати в еквівалентній формі

$$A_h y_h = \varphi_h, \quad (7.14)$$

якщо граничну умову (7.13) використати для виключення значень розв'язку в граничних точках області $\bar{\omega}_h$. Введемо для функцій y_h, φ_h відповідні сіткові норми $\|\cdot\|_{1h}, \|\cdot\|_{2h}$.

Кажуть, що сіткова схема (7.14) *стійка*, якщо існує така стала $M > 0$, яка не залежить від h і від вибору φ_h , що для розв'язку рівняння (7.14) справджується оцінка

$$\|y_h\|_{1h} \leq M \|\varphi_h\|_{2h} \quad (7.15)$$

при всіх достатньо малих $|h| : |h| \leq h_0$.

Сіткову схему (7.14) називають *коректною*, якщо розв'язок рівняння (7.14) існує і єдиний за будь-яких вхідних даних φ_h і якщо схема стійка, тобто виконується нерівність (7.15).

Стійкість означає неперервну залежність розв'язку y_h від вхідних даних, причому ця неперервна залежність рівномірна по h . Якщо \tilde{y}_h — розв'язок рівняння $A_h \tilde{y}_h = \tilde{\varphi}_h$, то $A_h (\tilde{y}_h - y_h) = \tilde{\varphi}_h - \varphi_h$ внаслідок лінійності A_h , тоді з (7.15) випливає

$$\|\tilde{y}_h - y_h\|_{1h} \leq M \|\tilde{\varphi}_h - \varphi_h\|_{2h}.$$

Отже, малим змінам вхідних даних відповідають малі зміни розв'язку.

Якщо схема (7.14) розв'язна, то існує обернений оператор A_h^{-1} і

$$y_h = A_h^{-1} \varphi_h, \quad \|y_h\|_{1h} \leq \|A_h^{-1}\| \|\varphi_h\|_{2h}.$$

Стійкість означає рівномірну по h обмеженість оберненого оператора

$$\|A_h^{-1}\| \leq M.$$

Схема нестійка, якщо $\|A_h^{-1}\|$ необмежено зростає при $|h| \rightarrow 0$.

Сіткова схема (7.12), (7.13) *стійка*, якщо для її розв'язку виконується оцінка

$$\|y_h\|_{1h} \leq M_1 \|\varphi_h\|_{2h} + M_2 \|\chi_h\|_{3h},$$

де $M_1 > 0$, $M_2 > 0$ — сталі, які не залежать від h і від вибору φ_h , χ_h .

При розв'язуванні задач (7.10), (7.11) сітковим методом необхідно встановити, з якою точністю розв'язок сіткової задачі наближає розв'язок вихідної задачі. Позначимо через $u_h(x)$ значення $u(x)$ на сітці ω_h . Величину $z_h = y_h - u_h$ називають *похибкою сіткової задачі*. Підставимо $y_h = z_h + u_h$ в (7.12), (7.13), тоді одержимо

$$L_h z_h = \psi_h, \quad x \in \omega_h, \quad l_h z_h = \nu_h, \quad x \in \gamma_h, \quad (7.16)$$

де $\psi_h = \varphi_h - L_h u_h$ називають *похибкою апроксимації (нев'язкою)* рівняння (7.12) на розв'язку рівняння (7.10), а $\nu_h = \chi_h - l_h u_h$ — *похибкою апроксимації (нев'язкою)* сіткової граничної умови (7.13) на розв'язку задачі (7.10), (7.11).

Для схеми (7.14) рівняння для похибки має вигляд

$$A z_h = \psi_h,$$

а похибка апроксимації

$$\psi_h = \varphi_h - A_h u_h.$$

Кажуть, що розв'язок сіткової задачі *збігається* до розв'язку задачі (7.10), (7.11), якщо $\|z_h\|_{1h} \rightarrow 0$ при $|h| \rightarrow 0$. Сіткова схема має *p -й порядок точності*, якщо

$$\|z_h\|_{1h} = O(|h|^p) \quad \text{або} \quad \|z_h\|_{1h} \leq M|h|^p,$$

де $M > 0$ — стала, яка не залежить від $|h|$.

Сіткова схема (7.12), (7.13) має *p -й порядок апроксимації*, якщо

$$\|\psi_h\|_{2h} = O(|h|^p), \quad \|\nu_h\|_{3h} = O(|h|^p).$$

У випадку схеми (7.14) сіткова задача має *p -й порядок апроксимації*, якщо

$$\|\psi_h\|_{2h} = O(|h|^p).$$

► **ТЕОРЕМА 7.1. (ЛАКСА—ФІЛПОВА)** Якщо сіткова схема коректна і апроксимує задачу (7.10), (7.11), то розв'язок сіткової задачі (7.14) збігається до розв'язку вихідної задачі (7.10), (7.11), причому порядок точності рівний порядку апроксимації.

Доведення. З коректності сіткової схеми випливає

$$\|z_h\|_{1h} \leq M_1 \|\psi_h\|_{2h}. \quad (7.17)$$

Якщо $\|\psi_h\|_{2h} \leq M_2 |h|^p$, то з (7.17) одержимо

$$\|z_h\|_{1h} \leq M_1 M_2 |h|^p,$$

тобто сіткова схема має p -й порядок точності. ■

7.3. Сіткові схеми як операторні рівняння

7.3.1. Запис сіткових схем у вигляді операторних рівнянь

Після заміни диференціальних рівнянь сітковими (різницевими) рівняннями на деякій сітці одержуємо систему лінійних алгебраїчних рівнянь, яку можна записати в матричній формі

$$Ay = \varphi, \quad (7.18)$$

де A — матриця системи, y — шуканий вектор, φ — заданий вектор, який визначається правими частинами сіткових рівнянь і граничними умовами.

Рівняння (7.18) можна розглядати як операторне, де A — лінійний оператор, який діє в скінченновимірному просторі H , y — шуканий елемент цього простору і $\varphi \in H$ — заданий елемент. Для сіткової схеми характерно, що кожна схема визначає не одне рівняння (7.18), а цілу множину рівнянь

$$A_h y_h = \varphi_h, \quad (7.19)$$

які залежать від кроку сітки h . При кожному значенні h оператор A_h діє в скінченновимірному просторі H_h . Розмірність простору H_h залежить від величини кроку сітки h .

Наведемо приклад запису різницевої схеми у вигляді операторного рівняння (7.19).

На рівномірній сітці $\bar{\omega}_h = \{x_i = ih, i = \overline{0, N}, hN = l\}$ розглянемо різницеву схему

$$y_{\bar{x},i} = -f_i, \quad i = \overline{1, N-1}, \quad y_0 = \mu_1, \quad y_N = \mu_2. \quad (7.20)$$

Перепишемо систему (7.20) у вигляді

$$\begin{aligned} -y_{\bar{x},i} &= f_i, \quad i = \overline{2, N-2}, \\ \frac{2y_1 - y_2}{h^2} &= \tilde{f}_1, \quad \frac{2y_{N-1} - y_{N-2}}{h^2} = \tilde{f}_{N-1}, \end{aligned} \quad (7.21)$$

де $\tilde{f}_1 = f_1 + \mu_1/h^2$, $\tilde{f}_{N-1} = f_{N-1} + \mu_2/h^2$.

Введемо $N-1$ -вимірний простір H_{N-1} , який складається з векторів $y = (y_1, y_2, \dots, y_{N-1})$, $y_i = y(x_i)$, $x_i \in \omega_h$. Визначимо в H_{N-1} оператор A і вектор $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_{N-1})$ так:

$$\begin{aligned} (Ay)_i &= -y_{\bar{x}x,i}, \quad i = \overline{2, N-2}, \\ (Ay)_1 &= \frac{2y_1 - y_2}{h^2}, \quad (Ay)_{N-1} = \frac{2y_{N-1} - y_{N-2}}{h^2}, \end{aligned} \quad (7.22)$$

$$\varphi_1 = \tilde{f}_1, \quad \varphi_i = f_i, \quad i = \overline{2, N-2}, \quad \varphi_{N-1} = \tilde{f}_{N-1}. \quad (7.23)$$

Тоді різницеву схему (7.21) можна записати в операторній формі (7.19). Матриця цього оператора є симетричною, тридіагональною і має вигляд

$$A = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 \end{pmatrix}.$$

Однак інколи зручніше вважати, що оператор A визначений на підпросторі $\overset{\circ}{H}_{N+1}$ функцій, заданих на сітці $\bar{\omega}_h$ і рівних нулю при $i = 0$, $i = N$. Функції з $\overset{\circ}{H}_{N+1}$ будемо позначати через $\overset{\circ}{y}_i$ ($\overset{\circ}{y}_i = y_i$, $i = \overline{1, N-1}$, $\overset{\circ}{y}_0 = \overset{\circ}{y}_N = 0$), тоді оператор A визначається формулою

$$(Ay)_i = -\overset{\circ}{y}_{\bar{x}x,i}, \quad i = \overline{1, N-1}, \quad (7.24)$$

а вектор φ згідно з (7.23). Різницєва схема (7.20) може бути записана у вигляді (7.19), де $A : \overset{\circ}{H}_{N+1} \rightarrow H_{N-1}$.

Оператор A , визначений формулами (7.22) або (7.24), будемо називати *оператором другої різницєвої похідної*.

7.3.2. Задача на власні значення для оператора другої різницєвої похідної

Задача на власні значення для оператора A полягає у знаходженні таких чисел λ (*власних чисел* або *власних значень*), для яких рівняння

$$Ay = \lambda y \quad (7.25)$$

має нетривіальні розв'язки (*власні функції*).

У випадку оператора другої різницевої похідної (7.24) рівняння (7.25) буде мати вигляд

$$-y_{\bar{x}x,i} = \lambda y_i, \quad i = \overline{1, N-1}, \quad y_0 = y_N = 0, \quad h = l/N \quad (7.26)$$

або

$$y_{i+1} + y_{i-1} = (2 - h^2 \lambda) y_i, \quad i = \overline{1, N-1}, \quad y_0 = y_N = 0. \quad (7.27)$$

Різницева задача (7.26) є апроксимацією диференціальної задачі

$$-u''(x) = \lambda u(x), \quad 0 < x < l, \quad u(0) = u(l) = 0, \quad (7.28)$$

розв'язком якої є власні числа

$$\lambda_k = \left(\frac{\pi k}{l} \right)^2, \quad k = 1, 2, \dots$$

і відповідні їм власні функції

$$u_k(x) = c \sin \frac{\pi k x}{l}, \quad k = 1, 2, \dots, \quad c \neq 0.$$

Власні функції задачі (7.26) будемо шукати у вигляді

$$y_k(x_i) = c \sin \frac{\pi k x_i}{l}, \quad x_i = ih, \quad k = 1, 2, \dots \quad (7.29)$$

Граничні умови $y_0 = y_N = 0$ при цьому виконуються. Підставляючи (7.29) у рівняння (7.27), одержимо

$$\sin \frac{\pi k (x_i + h)}{l} + \sin \frac{\pi k (x_i - h)}{l} = (2 - h^2 \lambda) \sin \frac{\pi k x_i}{l}$$

або

$$2 \sin \frac{\pi k x_i}{l} \cos \frac{\pi k h}{l} = (2 - h^2 \lambda) \sin \frac{\pi k x_i}{l}.$$

Звідси видно, що (7.29) є власними функціями оператора (7.24), якщо

$$2 \cos \frac{\pi k h}{l} = 2 - h^2 \lambda,$$

тобто

$$\lambda = \lambda_k = \frac{4}{h^2} \sin^2 \frac{\pi k h}{2l}.$$

При $k = \overline{1, N-1}$ одержимо $N-1$ різних дійсних чисел λ_k і відповідних їм власних функцій. Отже, розв'язок задачі (7.26) має вигляд

$$\begin{aligned} \lambda_k &= \frac{4}{h^2} \sin^2 \frac{\pi k h}{2l}, \quad y_k(x_i) = c \sin \frac{\pi k x_i}{l}, \\ x_i &= ih, \quad i = \overline{0, N}, \quad k = \overline{1, N-1}, \quad h = l/N. \end{aligned}$$

7.3.3. Властивості власних значень та власних функцій

Для власних значень λ_k , $k = \overline{1, N-1}$ справджуються нерівності

$$\frac{8}{l^2} \leq \lambda_1 < \lambda_2 < \dots < \lambda_{N-1} < \frac{4}{h^2}.$$

Справді, власні значення зростають з ростом k , оскільки $\sin \frac{\pi k h}{2l} < \sin \frac{\pi(k+1)h}{2l} < 1$ при $k \leq N-2$. Найбільше власне значення

$$\begin{aligned} \lambda_{N-1} &= \frac{4}{h^2} \sin^2 \frac{\pi h (N-1)}{2l} = \\ &= \frac{4}{h^2} \sin^2 \left(\frac{\pi}{2} - \frac{\pi h}{2l} \right) = \frac{4}{h^2} \cos^2 \frac{\pi h}{2l} < \frac{4}{h^2}. \end{aligned}$$

Найменше власне значення λ_1 запишемо у вигляді

$$\lambda_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2l} = \frac{\pi^2}{l^2} \left(\frac{\sin \xi}{\xi} \right)^2,$$

де $\xi = \frac{\pi h}{2l} \leq \frac{\pi}{4}$ (найбільше значення $h \leq l/2$). Оскільки функція $\frac{\sin \xi}{\xi}$ монотонно спадає при $\xi \in [0, \pi/4]$, то

$$\left(\frac{\sin \xi}{\xi} \right)^2 \geq \left(\frac{4\sqrt{2}}{2\pi} \right)^2 = \frac{8}{\pi^2},$$

тобто $\lambda_1 \geq 8/l^2$.

Перейдемо до вивчення властивостей власних функцій. Введемо в просторі $\overset{\circ}{H}_{N+1}$ скалярний добуток

$$(y, v) = \sum_{i=1}^{N-1} \overset{\circ}{y}_i \overset{\circ}{v}_i h$$

і норму

$$\|y\| = \sqrt{(y, y)} = \left(\sum_{i=1}^{N-1} \overset{\circ}{y}_i^2 h \right)^{1/2}. \quad (7.30)$$

Оператор A — самоспряжений, тобто $(Ay, v) = (y, Av)$ для $\forall y, v \in H_{N-1}$. Справді, використовуючи формулу сумування за частинами (див. (6.59))

$$\sum_{i=1}^{N-1} y_{x,i} v_i h = - \sum_{i=1}^N y_i v_{\bar{x},i} h + y_N v_N - v_0 y_1,$$

одержимо

$$(Ay, v) = - \sum_{i=1}^{N-1} \overset{\circ}{y}_{\bar{x},i} \overset{\circ}{v}_i h = \sum_{i=1}^N \overset{\circ}{y}_{\bar{x},i} \overset{\circ}{v}_{\bar{x},i} h.$$

В останній рівності враховано умову $\overset{\circ}{v}_0 = \overset{\circ}{v}_N = 0$. Якщо y і v поміняти місцями, то

$$(Av, y) = \sum_{i=1}^N \overset{\circ}{y}_{\bar{x},i} \overset{\circ}{v}_{\bar{x},i} h = (v, Ay). \quad (7.31)$$

Нехай

$$Ay_k = \lambda_k y_k, \quad Ay_m = \lambda_m y_m, \quad \lambda_k \neq \lambda_m.$$

$$(Ay_k, y_m) = \lambda_k (y_k, y_m), \quad (Ay_m, y_k) = \lambda_m (y_m, y_k)$$

і

$$(\lambda_k - \lambda_m) \cdot (y_m, y_k) = (Ay_k, y_m) - (Ay_m, y_k) = 0.$$

Звідси отримаємо, що $(y_m, y_k) = 0$, якщо $\lambda_k \neq \lambda_m$. Отже, система власних функцій (7.29) утворює ортогональний базис у просторі $\overset{\circ}{H}_{N+1}$.

Виберемо множник c так, щоб норма власної функції $y_k(x)$ була рівна одиниці

$$\|y_k\| = (y_k, y_k)^{1/2} = c \left(\sum_{i=1}^{N-1} \sin^2 \frac{\pi k x_i}{l} h \right)^{1/2} = 1.$$

Для цього обчислимо

$$\begin{aligned} \sum_{i=1}^{N-1} \sin^2 \frac{\pi k x_i}{l} h &= \sum_{i=1}^N \sin^2 \frac{\pi k x_i}{l} h = \\ &= \frac{1}{2} \sum_{i=1}^N \left(1 - \cos \frac{2\pi k x_i}{l} \right) h = \\ &= \frac{hN}{2} - \frac{h}{2} \sum_{i=1}^N \cos \frac{2\pi k x_i}{l}. \end{aligned}$$

З тотожності

$$\sin \frac{2\pi k (x_i + 0,5h)}{l} - \sin \frac{2\pi k (x_{i-1} + 0,5h)}{l} = 2 \sin \frac{\pi kh}{l} \cos \frac{2\pi k x_i}{l}$$

випливає

$$\begin{aligned} \cos \frac{2\pi k x_i}{l} &= \\ &= \frac{1}{2 \sin \frac{\pi kh}{l}} \left[\sin \frac{2\pi k (x_i + 0,5h)}{l} - \sin \frac{2\pi k (x_{i-1} + 0,5h)}{l} \right]. \end{aligned}$$

Тоді

$$\begin{aligned} \sum_{i=1}^{N-1} \sin^2 \frac{\pi k x_i}{l} h &= \frac{hN}{2} - \frac{h}{4 \sin \frac{\pi kh}{l}} \times \\ &\times \sum_{i=1}^N \left(\sin \frac{2\pi k (x_i + 0,5h)}{l} - \sin \frac{2\pi k (x_{i-1} + 0,5h)}{l} \right) = \\ &= \frac{l}{2} - \frac{h}{4 \sin \frac{\pi kh}{l}} \left(\sin \frac{2\pi k (x_N + 0,5h)}{l} - \sin \frac{\pi kh}{l} \right) = \\ &= \frac{l}{2} - \frac{h \sin \pi k \cdot \cos \frac{2\pi k (x_N + h)}{l}}{2 \sin \frac{\pi kh}{l}} = \frac{l}{2}. \end{aligned}$$

Отже, $c = \sqrt{2/l}$ і власні функції

$$\mu_k(x_i) = \sqrt{\frac{2}{l}} \sin \frac{\pi k x_i}{l}, \quad k = \overline{1, N-1}, \quad i = \overline{1, N-1}$$

утворюють ортонормований базис у просторі \mathring{H}_{N+1} .

7.3.4. Операторні нерівності

Встановимо оцінки для меж оператора другої різницевої похідної. Будь-який елемент $\mathring{y} \in \mathring{H}_{N+1}$ можна розкласти за власними функціями оператора A :

$$y_i = \sum_{k=1}^{N-1} c_k \mu_k(x_i), \quad i = \overline{1, N-1}, \quad (7.32)$$

де $c_k = (y, \mu_k)$ — коефіцієнти Фур'є. З ортонормованості системи $\{\mu_k\}$ випливає тотожність

$$\|y\|^2 = (y, y) = \sum_{k=1}^{N-1} c_k^2.$$

Використовуючи (7.32), отримаємо

$$Ay = \sum_{k=1}^{N-1} c_k A\mu_k = \sum_{k=1}^{N-1} c_k \lambda_k \mu_k$$

і

$$(Ay, y) = \sum_{k=1}^{N-1} c_k^2 \lambda_k. \quad (7.33)$$

На підставі (7.33) матимемо

$$\lambda_1 \|y\|^2 \leq (Ay, y) \leq \lambda_{N-1} \|y\|^2. \quad (7.34)$$

або згідно з (7.30), (7.31)

$$\lambda_1 \sum_{i=1}^{N-1} \overset{\circ}{y}_i^2 h \leq \sum_{i=1}^N \overset{\circ}{y}_{\bar{x},i}^2 h \leq \lambda_{N-1} \sum_{i=1}^{N-1} \overset{\circ}{y}_i^2 h. \quad (7.35)$$

Враховуючи доведені раніше оцінки для власних чисел, з (7.34) отримаємо нерівності

$$\frac{8}{l^2} \|y\|^2 \leq (Ay, y) < \frac{4}{h^2} \|y\|^2. \quad (7.36)$$

Звідси випливає, що оператор A — додатно визначений.

7.4. Різницеві схеми для одновимірного рівняння теплопровідності

7.4.1. Різницева схема з ваговими коефіцієнтами

Нехай в області $\bar{Q} = \{0 \leq x \leq l, 0 \leq t \leq T\}$ необхідно знайти розв'язок першої крайової задачі для рівняння теплопровідності

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad 0 < x < l, \quad 0 < t \leq T, \quad (7.37)$$

$$u(x, 0) = u_0(x), \quad 0 \leq x \leq l, \quad (7.38)$$

$$u(0, t) = \mu_1(t), \quad u(l, t) = \mu_2(t), \quad 0 \leq t \leq T. \quad (7.39)$$

Введемо сітки

$$\bar{\omega}_h = \{x_i = ih, \quad i = \overline{0, N}, \quad h = l/N\},$$

$$\omega_\tau = \{t_n = n\tau, \quad n = \overline{0, K}, \quad \tau = T/K\}$$

і сітку в \bar{Q}

$$\bar{\omega}_{h\tau} = \bar{\omega}_h \times \omega_\tau = \{(x_i, t_n), \quad i = \overline{0, N}, \quad n = \overline{0, K}\}.$$

Вузли (x_i, t_n) , $i = \overline{1, N-1}$, $n = \overline{1, K}$ називають *внутрішніми вузлами*, а решту — *граничними вузлами сітки*. Множину всіх внутрішніх вузлів сітки $\bar{\omega}_{h\tau}$ будемо позначати

$$\omega_{h\tau} = \{(x_i, t_n), \quad i = \overline{1, N-1}, \quad n = \overline{1, K}\}.$$

Для функції $y(x, t)$, визначеної на сітці $\bar{\omega}_{h\tau}$ введемо позначення

$$y_i^n = y(x_i, t_n), \quad y_{t,i}^n = \frac{y_i^{n+1} - y_i^n}{\tau}, \quad y_{\bar{x},i}^n = \frac{y_{i+1}^n - 2y_i^n + y_{i-1}^n}{h^2}. \quad (7.40)$$

Щоб апроксимувати рівняння (7.37) в точці (x_i, t_n) , розглянемо шеститочковий шаблон (рис. 7.1).

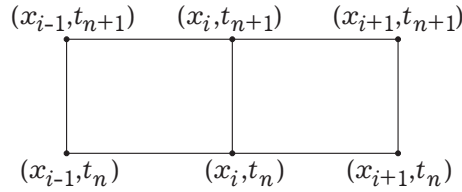


Рис. 7.1.

Замінюючи похідну $\partial u / \partial t$ в точці (x_i, t_n) першою різницевою похідною $y_{t,i}^n$, а $\partial^2 u / \partial x^2$ — лінійною комбінацією різницевої похідної $y_{\bar{x},i}^n$ на n -ому і на $n+1$ -ому ярусах, розглянемо різницеву схему з ваговими коефіцієнтами

$$\begin{aligned} \frac{y_i^{n+1} - y_i^n}{\tau} &= \sigma y_{\bar{x},i}^{n+1} + (1 - \sigma) y_{\bar{x},i}^n + \varphi_i^n, \\ i &= \overline{1, N-1}, \quad n = \overline{0, K-1}, \end{aligned} \quad (7.41)$$

де σ — дійсний параметр, φ_i^n — сіткова функція, яка апроксимує праву частину f рівняння (7.37), наприклад, $\varphi_i^n = f(x_i, t_{n+1/2})$, $t_{n+1/2} = t_n + \tau/2$. Оскільки схема (7.41) містить значення шуканої функції y_i на двох ярусах, то вона *двоярусна*.

Початкові і крайові умови апроксимуються точно

$$y_i^0 = u_0(x_i), \quad i = \overline{0, N}, \quad (7.42)$$

$$y_0^n = \mu_1(t_n), \quad y_N^n = \mu_2(t_n), \quad n = \overline{0, K}. \quad (7.43)$$

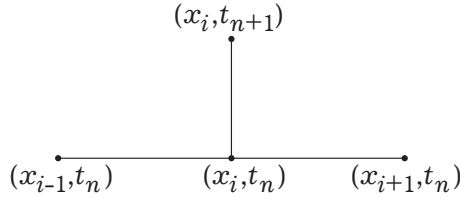


Рис. 7.2.

Розглянемо часткові випадки різницевої схеми (7.41). При $\sigma = 0$ одержуємо схему, визначену на чотириточковому шаблоні (рис. 7.2), яка має вигляд

$$\frac{y_i^{n+1} - y_i^n}{\tau} = \frac{y_{i+1}^n - 2y_i^n + y_{i-1}^n}{h^2} + \varphi_i^n, \quad i = \overline{1, N-1}, \quad n = \overline{0, K-1}$$

або

$$y_i^{n+1} = (1 - 2\gamma) y_i^n + \gamma (y_{i-1}^n + y_{i+1}^n) + \tau \varphi_i^n, \quad \gamma = \tau/h^2. \quad (7.44)$$

Значення y_i^{n+1} у кожній точці яруса $t = t_{n+1}$ (нового яруса) виражається за допомогою явної формули (7.44) через y_i^n на ярусі $t = t_n$ (на старому). Оскільки при $t = 0$ задана початкова умова $y_i^0 = u_0(x_i)$, то формула (7.44) дозволяє послідовно визначити y_i^n на будь-якому ярусі. Схему (7.44) називають *явною*.

Якщо $\sigma \neq 0$, то схему (7.41) називають *неявною* двоярусною схемою. При $\sigma \neq 0$ для визначення y_i^{n+1} на новому ярусі одержуємо систему лінійних алгебраїчних рівнянь

$$-\frac{1}{\sigma\tau} y_i^{n+1} + y_{xx,i}^{n+1} = -F_i, \quad i = \overline{1, N-1}, \quad n = \overline{0, K-1}.$$

$$y_0^{n+1} = \mu_1(t_{n+1}), \quad y_N^{n+1} = \mu_2(t_{n+1}), \quad n = \overline{0, K-1},$$

де $F_i = \frac{1}{\sigma\tau}y_i^n + \left(\frac{1}{\sigma} - 1\right)y_{\bar{x}x,i}^n + \frac{1}{\sigma}\varphi_i^n$. Використовуючи (7.40), зведемо цю схему до вигляду

$$\begin{aligned} -C_0y_0^{n+1} + B_0y_1^{n+1} &= -F_0, \\ A_iy_{i-1}^{n+1} - C_iy_i^{n+1} + B_iy_{i+1}^{n+1} &= -F_i, \quad i = \overline{1, N-1}, \\ A_Ny_{N-1}^{n+1} - C_Ny_N^{n+1} &= -F_N, \end{aligned} \quad (7.45)$$

де

$$B_0 = A_N = 0, \quad C_0 = C_N = 1, \quad F_0 = -\mu_1(t_{n+1}), \quad F_N = -\mu_2(t_{n+1}),$$

$$A_i = B_i = \frac{1}{h^2}, \quad C_i = \frac{2}{h^2} + \frac{1}{\sigma\tau}, \quad i = \overline{1, N-1}.$$

Систему (7.45) можна розв'язувати методом прогонки (див. розділ 1.1.2), якщо $|2/h^2 + 1/(\sigma\tau)| \geq 2/h^2$. З цієї нерівності випливає, що метод прогонки стійкий при $\sigma \geq -h^2/(4\tau)$.

При $\sigma = 1$ маємо *чисто неявну схему*

$$\frac{y_i^{n+1} - y_i^n}{\tau} = y_{\bar{x}x,i}^{n+1} + \varphi_i^n, \quad i = \overline{1, N-1}, \quad n = \overline{0, K-1},$$

визначену на чотириточковому шаблоні (рис. 7.3).

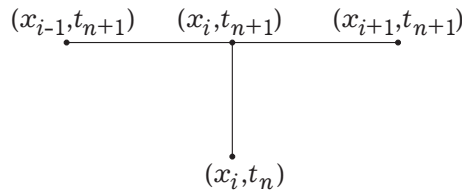


Рис. 7.3.

При $\sigma = 1/2$ одержуємо *шеститочкову симетричну схему*

$$\begin{aligned} \frac{y_i^{n+1} - y_i^n}{\tau} &= \frac{1}{2}(y_{\bar{x}x,i}^{n+1} + y_{\bar{x}x,i}^n) + \varphi_i^n, \\ i &= \overline{1, N-1}, \quad n = \overline{0, K-1}, \end{aligned}$$

яку ще називають *схемою Кранка–Нікольсона*.

7.4.2. Порядок апроксимації різницевої схеми з ваговими коефіцієнтами

Дослідимо похибку апроксимації різницевої схеми з ваговими коефіцієнтами. Запишемо розв'язок задачі (7.41) — (7.43) у вигляді $y_i^n = u(x_i, t_n) + z_i^n$, де $u(x_i, t_n)$ — точний розв'язок задачі (7.37) — (7.39). Тоді для похибки отримаємо систему

$$z_{t,i}^n = \sigma z_{\bar{x},i}^{n+1} + (1 - \sigma) z_{\bar{x},i}^n + \psi_i^n, \\ i = \overline{1, N-1}, \quad n = \overline{0, K-1},$$

$$z_0^n = z_N^n = 0, \quad n = \overline{1, K}, \quad z_i^0 = 0, \quad i = \overline{0, N}.$$

Величину

$$\psi_i^n = \sigma u_{\bar{x},i}^{n+1} + (1 - \sigma) u_{\bar{x},i}^n - u_{t,i}^n + \varphi_i^n$$

називають *похибкою апроксимації схеми* (7.41) — (7.43) на розв'язку задачі (7.37) — (7.39). Будемо розвивати всі функції, які входять у вираз для ψ_i^n , за формулою Тейлора в точці $(x_i, t_n + \tau/2)$. Враховуючи розклад

$$u_{t,i}^n = \dot{u}(x_i, t_{n+1/2}) + O(\tau^2), \\ u_{\bar{x},i}^n = u''(x_i, t_n) + \frac{h^2}{12} u^{IV}(x_i, t_n) + O(h^4),$$

де

$$u'' = \frac{\partial^2 u}{\partial x^2}, \quad u^{IV} = \frac{\partial^4 u}{\partial x^4}, \quad \dot{u} = \frac{\partial u}{\partial t}, \quad t_{n+1/2} = t_n + \tau/2,$$

одержимо

$$\psi_i^n = \sigma \left[u''(x_i, t_{n+1}) + \frac{h^2}{12} u^{IV}(x_i, t_{n+1}) \right] + \\ + (1 - \sigma) \left[u''(x_i, t_n) + \frac{h^2}{12} u^{IV}(x_i, t_n) \right] - \\ - \dot{u}(x_i, t_{n+1/2}) + \varphi_i^n + O(\tau^2 + h^4).$$

Використовуючи позначення $u = u(x_i, t_{n+1/2})$, запишемо розклад ψ_i^n в

ОКОЛІ ТОЧКИ $(x_i, t_{n+1/2})$

$$\begin{aligned}\psi_i^n &= \sigma \left(u'' + \frac{\tau}{2} \dot{u}'' + \frac{h^2}{12} u^{IV} + \frac{\tau h^2}{24} \dot{u}^{IV} \right) + \\ &+ (1 - \sigma) \left(u'' - \frac{\tau}{2} \dot{u}'' + \frac{h^2}{12} u^{IV} - \frac{\tau h^2}{24} \dot{u}^{IV} \right) - \\ &- \dot{u} + \varphi_i^n + O(\tau^2 + h^4) = \\ &= u'' - \dot{u} + \varphi_i^n + (\sigma - 0,5) \tau \left(\dot{u}'' + \frac{h^2}{12} \dot{u}^{IV} \right) + \\ &+ \frac{h^2}{12} u^{IV} + O(\tau^2 + h^4).\end{aligned}$$

З рівняння (7.37) $u'' - \dot{u} = -f$ і $u^{IV} - \dot{u}'' = -f''$, тоді

$$\begin{aligned}\psi_i^n &= \left[\varphi_i^n - f(x_i, t_{n+1/2}) - \frac{h^2}{12} f''(x_i, t_{n+1/2}) \right] + \\ &+ \left[(\sigma - 0,5) \tau + \frac{h^2}{12} \right] \dot{u}'' + (\sigma - 0,5) \tau \frac{h^2}{12} \dot{u}^{IV} + O(\tau^2 + h^4).\end{aligned}$$

Якщо

$$\sigma = \frac{1}{2} - \frac{h^2}{12\tau},$$

$$\varphi_i^n = f(x_i, t_{n+1/2}) + \frac{h^2}{12} f''(x_i, t_{n+1/2}) + O(\tau^2 + h^4),$$

то $\psi_i^n = O(\tau^2 + h^4)$ і схема (7.41) — (7.43) має другий порядок апроксимації по τ і четвертий по h . Таку схему називають схемою *підвищеного порядку апроксимації*. Якщо $\sigma = 0,5$, $\varphi_i^n = f(x_i, t_{n+1/2}) + O(\tau^2 + h^2)$, то схема (7.41) — (7.43) має порядок апроксимації $O(\tau^2 + h^2)$. При $\sigma \neq 0,5$, $\sigma \neq 1/2 - h^2/(12\tau)$ і $\varphi_i^n = f(x_i, t_{n+1/2}) + O(\tau + h^2)$ схема має перший порядок апроксимації по τ і другий по h .

7.4.3. Апроксимація крайових умов третього роду

Розглянемо тепер рівняння (7.37) з крайовими умовами третього роду

$$\begin{aligned}\frac{\partial u(0, t)}{\partial x} &= \beta_1 u(0, t) - \mu_1(t), \quad \beta_1 = \text{const} > 0, \\ \frac{\partial u(l, t)}{\partial x} &= -\beta_2 u(l, t) + \mu_2(t), \quad \beta_2 = \text{const} > 0.\end{aligned}\tag{7.46}$$

Якщо в крайових умовах $\partial u(0, t)/\partial x$ замінити правою різницевою похідною $y_{x,0}^n$, а $\partial u(l, t)/\partial x$ — лівою різницевою похідною $y_{\bar{x},N}^n$, то отримаємо різницеві граничні умови порядку апроксимації $O(h)$. Для одержання різницевих крайових умов другого порядку апроксимації продовжимо розв'язок задачі $u(x, t)$ зовні відрізка $[0, l]$ ще на один інтервал зліва і справа, тобто введемо додаткові вузли $x_{-1} = -h$ і $x_{N+1} = l + h$. Частинні похідні в крайових умовах замінимо центральними різницеви похідними, тоді отримаємо

$$\begin{aligned}\frac{y_1^n - y_{-1}^n}{2h} &= \beta_1 y_0^n - \mu_1(t_n), \quad n = \overline{0, K}, \\ \frac{y_{N+1}^n - y_{N-1}^n}{2h} &= -\beta_2 y_N^n + \mu_2(t_n), \quad n = \overline{0, K}.\end{aligned}\tag{7.47}$$

Запишемо різницеву схему (7.41) для $i = 0, i = N$

$$\begin{aligned}y_{t,0}^n &= \sigma \frac{y_1^{n+1} - 2y_0^{n+1} + y_{-1}^{n+1}}{h^2} + \\ &+ (1 - \sigma) \frac{y_1^n - 2y_0^n + y_{-1}^n}{h^2} + f(0, t_{n+1/2}), \\ y_{t,N}^n &= \sigma \frac{y_{N+1}^{n+1} - 2y_N^{n+1} + y_{N-1}^{n+1}}{h^2} + \\ &+ (1 - \sigma) \frac{y_{N+1}^n - 2y_N^n + y_{N-1}^n}{h^2} + f(l, t_{n+1/2}),\end{aligned}\tag{7.48}$$

З (7.47) матимемо

$$\begin{aligned}y_{-1}^n &= y_1^n - 2h(\beta_1 y_0^n - \mu_1(t_n)), \\ y_{N+1}^n &= y_{N-1}^n - 2h(\beta_2 y_N^n - \mu_2(t_n)), \quad n = \overline{0, K}.\end{aligned}$$

Підставимо y_{-1}^n , y_{-1}^{n+1} , y_{N+1}^n , y_{N+1}^{n+1} у рівняння (7.48), тоді

$$\begin{aligned}y_{t,0}^n &= \frac{2\sigma}{h} (y_{x,0}^{n+1} - \beta_1 y_0^{n+1}) + \frac{2(1-\sigma)}{h} (y_{x,0}^n - \beta_1 y_0^n) + \\ &+ \frac{2}{h} (\sigma \mu_1(t_{n+1}) + (1-\sigma) \mu_1(t_n)) + f(0, t_{n+1/2}), \\ y_{t,N}^n &= -\frac{2\sigma}{h} (y_{\bar{x},N}^{n+1} + \beta_2 y_N^{n+1}) - \frac{2(1-\sigma)}{h} (y_{\bar{x},N}^n + \beta_2 y_N^n) + \\ &+ \frac{2}{h} (\sigma \mu_2(t_{n+1}) + (1-\sigma) \mu_2(t_n)) + f(l, t_{n+1/2}).\end{aligned}$$

Помножимо ліву і праву частини цих рівностей на $\pm h/2$ відповідно, тоді одержимо різницеві крайові умови

$$\begin{aligned}\sigma(y_{x,0}^{n+1} - \beta_1 y_0^{n+1}) + (1 - \sigma)(y_{x,0}^n - \beta_1 y_0^n) &= 0,5h y_{t,0}^n - \varphi_0^n, \\ \sigma(y_{\bar{x},N}^{n+1} + \beta_2 y_N^{n+1}) + (1 - \sigma)(y_{\bar{x},N}^n + \beta_2 y_N^n) &= -0,5h y_{t,N}^n + \varphi_N^n,\end{aligned}\quad (7.49)$$

де

$$\begin{aligned}\varphi_0^n &= \sigma \mu_1(t_{n+1}) + (1 - \sigma) \mu_1(t_n) + 0,5h f(0, t_{n+1/2}), \\ \varphi_N^n &= \sigma \mu_2(t_{n+1}) + (1 - \sigma) \mu_2(t_n) + 0,5h f(l, t_{n+1/2}).\end{aligned}$$

Можна показати, що при $\sigma = 0,5$ різницеві крайові умови (7.49) апроксимують умови (7.46) з порядком $O(\tau^2 + h^2)$, а при $\sigma \neq 0,5$ з $O(\tau + h^2)$.

Приклад 7.1. Доведіть, що різницева схема

$$\begin{aligned}\frac{1}{12}y_{t,i+1}^n + \frac{5}{6}y_{t,i}^n + \frac{1}{12}y_{t,i-1}^n &= \frac{1}{2}y_{\bar{x}x,i}^{n+1} + \frac{1}{2}y_{\bar{x}x,i}^n, \\ i &= \overline{1, N-1}, \quad n = \overline{0, K-1},\end{aligned}$$

$$y_i^0 = u_0(x_i), \quad i = \overline{0, N}, \quad x_i = ih, \quad h = 1/N,$$

$$y_0^n = y_N^n = 0, \quad n = \overline{0, K}, \quad \tau = T/K$$

апроксимує задачу

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t \leq T,$$

$$u(x, 0) = u_0(x), \quad 0 \leq x \leq 1,$$

$$u(0, t) = u(1, t) = 0, \quad 0 \leq t \leq T$$

з порядком $O(\tau^2 + h^4)$.

▷ Запишемо похибку апроксимації різницевої схеми

$$\psi_i^n = \frac{1}{2}u_{\bar{x}x,i}^{n+1} + \frac{1}{2}u_{\bar{x}x,i}^n - \frac{1}{12}u_{t,i+1}^n - \frac{5}{6}u_{t,i}^n - \frac{1}{12}u_{t,i-1}^n.$$

Використовуючи співвідношення

$$u_t^n = \dot{u}^n + \frac{\tau}{2}\ddot{u}^n + O(\tau^2), \quad u_{\bar{x}x,i} = u_i'' + \frac{h^2}{12}u_i^{IV} + O(h^4)$$

розкладемо похибку апроксимації в околі точки (x_i, t_n) в ряд Тейлора. Тоді

$$\begin{aligned}\psi_i^n = & \frac{1}{2}[u''(x_i, t_{n+1}) + \frac{h^2}{12}u^{IV}(x_i, t_{n+1})] + \frac{1}{2}[u''(x_i, t_n) + \frac{h^2}{12}u^{IV}(x_i, t_n)] - \\ & - \frac{1}{12}\{\dot{u}(x_{i+1}, t_n) + \dot{u}(x_{i-1}, t_n) + \frac{\tau}{2}[\ddot{u}(x_{i+1}, t_n) + \ddot{u}(x_{i-1}, t_n)]\} - \\ & - \frac{5}{6}[\dot{u}(x_i, t_n) + \frac{\tau}{2}\ddot{u}(x_i, t_n)] + O(\tau^2 + h^4).\end{aligned}$$

Оскільки

$$\begin{aligned}v^{n+1} &= v^n + \tau \dot{v}^n + O(\tau^2), \\ w_{i\pm 1} &= w_i \pm h w'_i + \frac{h^2}{2}w''_i \pm \frac{h^3}{3}w'''_i + O(h^4), \\ w_{i+1} + w_{i-1} &= 2w_i + h^2 w''_i + O(h^4),\end{aligned}$$

то

$$\begin{aligned}\psi_i^n = & \frac{1}{2}(u'' + \tau \dot{u}'' + \frac{h^2}{12}u^{IV} + \frac{\tau h^2}{12}\dot{u}^{IV}) + \frac{1}{2}(u'' + \frac{h^2}{12}u^{IV}) - \\ & - \frac{1}{12}(2\dot{u} + h^2\dot{u}'' + \tau\ddot{u} + \frac{\tau h^2}{2}\ddot{u}'') - \frac{5}{6}(\dot{u} + \frac{\tau}{2}\ddot{u}) + O(\tau^2 + h^4) = \\ & = u'' - \dot{u} + \frac{h^2}{12}(u^{IV} - \dot{u}'') + \frac{\tau}{2}(\dot{u}'' - \ddot{u}) + \\ & + \frac{\tau h^2}{24}(\dot{u}^{IV} - \ddot{u}'') + O(\tau^2 + h^4),\end{aligned}$$

де $u = u(x_i, t_n)$.

Використаємо рівняння $\dot{u} = u''$, $\dot{u}'' = u^{IV}$, тоді з урахуванням рівностей $\ddot{u} = \dot{u}'' = u^{IV}$, $\dot{u}^{IV} = \ddot{u}''$ будемо мати $\psi_i^n = O(\tau^2 + h^4)$. \blacktriangleleft

Приклад 7.2. Різницеву схему

$$\begin{aligned}y_{i,i}^n &= \frac{1}{2}y_{\bar{x},i}^{n+1} + \frac{1}{2}y_{\bar{x},i}^n, \quad i = \overline{1, N-1}, \quad n = \overline{0, K-1}, \\ y_i^0 &= u_0(x_i), \quad i = \overline{0, N}, \\ hy_{i,0}^n &= y_{x,0}^{n+1} + y_{x,0}^n, \quad y_N^{n+1} = 0, \quad n = \overline{0, K-1}\end{aligned}$$

запишіть у вигляді системи триточкових різницевих рівнянь. Перевірте умови стійкості методу прогонки її розв'язування.

▷ Враховуючи (7.40) різницеву схему запишемо у вигляді

$$\begin{aligned}\frac{y_i^{n+1} - y_i^n}{\tau} &= \frac{y_{i+1}^{n+1} - 2y_i^{n+1} + y_{i-1}^{n+1}}{2h^2} + \frac{1}{2}y_{\bar{x},i}^n, \\ & i = \overline{1, N-1}, \quad n = \overline{0, K-1},\end{aligned}$$

$$\frac{y_0^{n+1} - y_0^n}{\tau} = \frac{y_1^{n+1} - y_0^{n+1}}{h^2} + \frac{1}{h} y_{x,0}^n, \quad y_N^{n+1} = 0, \\ n = \overline{0, K-1}.$$

Отже, різницеву схему можна записати у вигляді (7.45), де

$$C_0 = \frac{1}{h^2} + \frac{1}{\tau}, \quad B_0 = \frac{1}{h^2}, \quad F_0 = \frac{1}{h} y_{x,0}^n + \frac{y_0^n}{\tau}, \\ A_i = \frac{1}{2h^2}, \quad C_i = \frac{1}{h^2} + \frac{1}{\tau}, \quad B_i = \frac{1}{2h^2}, \quad F_i = \frac{1}{2} y_{xx,i}^n + \frac{y_i^n}{\tau}.$$

Оскільки

$$|C_i| = \frac{1}{h^2} + \frac{1}{\tau} > \frac{1}{h^2} = |A_i| + |B_i|, \quad i = \overline{1, N-1}, \\ |C_0| = \frac{1}{h^2} + \frac{1}{\tau} > \frac{1}{h^2} = |B_0|, \quad |C_N| = 1 > 0 = |A_N|,$$

то метод прогонки стійкий, і його можна застосувати для розв'язування цієї задачі. ◀

Приклад 7.3. Покажіть, що різницева схема з попереднього прикладу апроксимує задачу

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \quad 0 < t \leq T, \\ u(x, 0) = u_0(x), \quad 0 \leq x \leq 1, \\ \frac{\partial u(0, t)}{\partial x} = 0, \quad u(1, t) = 0$$

з порядком $O(\tau^2 + h^2)$.

▷ Запишемо похибку апроксимації різницевої схеми та крайові умови

$$\psi_i^n = \frac{1}{2} u_{xx,i}^{n+1} + \frac{1}{2} u_{xx,i}^n - u_{t,i}^n, \quad i = \overline{1, N-1}, \\ \psi_0^n = u_{x,0}^{n+1} + u_{x,0}^n - h u_{t,0}^n.$$

Оскільки різницева схема є схемою Кранка–Нікольсона, то $\psi_i^n = O(\tau^2 + h^2)$, $i = \overline{1, N-1}$. Розкладемо ψ_0^n в ряд Тейлора в околі точки $(0, t_n)$:

$$\psi_0^n = u'(0, t_{n+1}) + \frac{h}{2} u''(0, t_{n+1}) + u'(0, t_n) + \\ + \frac{h}{2} u''(0, t_n) - h \dot{u}(0, t_n) - \frac{\tau h}{2} \ddot{u}(0, t_n) + O(\tau^2 + h^2) = \\ = 2u' + \tau \dot{u}' + h u'' + \frac{\tau h}{2} \dot{u}'' - h \dot{u} - \frac{\tau h}{2} \ddot{u} + O(\tau^2 + h^2),$$

де $u = u(0, t_n)$. Використаємо крайову умову та диференціальне рівняння

$$u'(0, t) = 0, \quad \dot{u}'(0, t) = 0, \quad \dot{u} = u'', \quad \ddot{u} = \dot{u}'',$$

тоді $\psi_0^n = O(\tau^2 + h^2)$. ◀

7.5. Різницеві схеми для рівняння коливання струни

7.5.1. Різницева схема з ваговими коефіцієнтами

Розглянемо крайову задачу

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad 0 < x < l, \quad 0 < t \leq T, \quad (7.50)$$

$$u(x, 0) = u_0(x), \quad \frac{\partial u(x, 0)}{\partial t} = \bar{u}_0(x), \quad 0 \leq x \leq l, \quad (7.51)$$

$$u(0, t) = \mu_1(t), \quad u(l, t) = \mu_2(t), \quad 0 \leq t \leq T. \quad (7.52)$$

Для побудови різницевої схеми замінімо похідну $\partial^2 u / \partial t^2$ у вузлі (x_i, t_n) сітки $\bar{\omega}_{h\tau}$ другою різницевою похідною

$$y_{tt,i}^n = (y_i^{n+1} - 2y_i^n + y_i^{n-1}) / \tau^2,$$

а $\partial^2 u / \partial x^2$ — лінійною комбінацією різницевої похідної $y_{\bar{x}x,i}$ на $n+1$, n і $n-1$ ярусах. Тоді отримаємо різницеву схему з ваговими коефіцієнтами

$$y_{tt,i}^n = \sigma y_{\bar{x}x,i}^{n+1} + (1 - 2\sigma) y_{\bar{x}x,i}^n + \sigma y_{\bar{x}x,i}^{n-1} + \varphi_i^n, \quad (7.53)$$

$$i = \overline{1, N-1}, \quad n = \overline{1, K-1},$$

$$y_0^n = \mu_1(t_n), \quad y_N^n = \mu_2(t_n), \quad n = \overline{0, K}, \quad (7.54)$$

$$y_i^0 = u_0(x_i), \quad y_{t,i}^0 = \tilde{u}_0(x_i), \quad i = \overline{0, N}, \quad (7.55)$$

де $\varphi_i^n = f(x_i, t_n)$, а $\tilde{u}_0(x_i)$ визначимо далі.

Крайові умови і перша початкова умова на сітці $\bar{\omega}_{h\tau}$ задовольняються точно. Виберемо $\tilde{u}_0(x_i)$ так, щоб похибка апроксимації $\tilde{u}_0(x_i) - u_{t,i}^0$ була величиною другого порядку апроксимації. Для цього використаємо розклад

$$\begin{aligned} u_{t,i}^0 &= \frac{\partial u(x_i, 0)}{\partial t} + \frac{\tau}{2} \frac{\partial^2 u(x_i, 0)}{\partial t^2} + O(\tau^2) = \\ &= \frac{\partial u(x_i, 0)}{\partial t} + \frac{\tau}{2} \left(\frac{\partial^2 u(x_i, 0)}{\partial x^2} + f(x_i, 0) \right) + O(\tau^2). \end{aligned}$$

Отже,

$$\frac{\partial u(x_i, 0)}{\partial t} = u_{t,i}^0 - \frac{\tau}{2} \left(\frac{\partial^2 u(x_i, 0)}{\partial x^2} + f(x_i, 0) \right) + O(\tau^2),$$

а тому якщо

$$\tilde{u}_0(x_i) = \bar{u}_0(x_i) + \frac{\tau}{2} (u_{0\bar{x}x,i} + f(x_i, 0)),$$

то,

$$\tilde{u}_0(x_i) - u_{t,i}^0 = O(\tau^2 + h^2).$$

Схема (7.53) — *триярусна*. Її застосування передбачає, що при знаходженні значень y_i^{n+1} на верхньому ярусі значення на попередніх ярусах y_i^n, y_i^{n-1} , $i = \overline{0, N}$ зберігаються у пам'яті комп'ютера.

Якщо $\sigma = 0$, то одержимо *явну* різницеву схему

$$\frac{y_i^{n+1} - 2y_i^n + y_i^{n-1}}{\tau^2} = \frac{y_{i+1}^n - 2y_i^n + y_{i-1}^n}{h^2} + \varphi_i^n,$$

$$i = \overline{1, N-1}, \quad n = \overline{1, K-1}.$$

Розв'язок цієї схеми y_i^{n+1} виражається явно через значення на попередніх ярусах:

$$y_i^{n+1} = 2y_i^n - y_i^{n-1} + \gamma(y_{i+1}^n - 2y_i^n + y_{i-1}^n) + \tau^2 \varphi_i^n,$$

$$i = \overline{1, N-1}, \quad n = \overline{1, K-1}, \quad \gamma = \frac{\tau^2}{h^2}.$$

Для початку розрахунку за цією формулою повинні бути задані початкові значення y_i^0, y_i^1 , $i = \overline{0, N}$. З першої початкової умови одержуємо

$$y_i^0 = u_0(x_i), \quad i = \overline{0, N},$$

а з другої визначаємо y_i^1

$$y_i^1 = y_i^0 + \tau \bar{u}_0(x_i) + \frac{\tau^2}{2} (u_{0\bar{x}x,i} + f(x_i, 0)), \quad i = \overline{0, N}.$$

При $\sigma \neq 0$ схема (7.53) — (7.55) — *неявною*. Для визначення y_i^{n+1} одержимо систему рівнянь:

$$-\frac{1}{\sigma \tau^2} y_i^{n+1} + y_{\bar{x}x,i}^{n+1} = -F_i, \quad i = \overline{1, N-1}, \quad n = \overline{0, K-1},$$

$$y_0^{n+1} = \mu_1(t_{n+1}), \quad y_N^{n+1} = \mu_2(t_{n+1}), \quad n = \overline{0, K-1}.$$

Ця схема може бути зведена до системи триточкових різницевих рівнянь (див. (7.45)), де

$$B_0 = 0, \quad C_0 = 1, \quad F_0 = \mu_1(t_{n+1}),$$

$$\begin{aligned}
A_N &= 0, \quad C_N = 1, \quad F_N = \mu_2(t_{n+1}), \\
A_i &= B_i = \frac{1}{h^2}, \quad C_i = \frac{2}{h^2} + \frac{1}{\sigma\tau^2}, \quad i = \overline{1, N-1}, \\
F_i &= \frac{1}{\sigma\tau^2} (2y_i^n - y_i^{n-1}) + \left(\frac{1}{\sigma} - 2\right) y_{\bar{x}\bar{x},i}^n + y_{\bar{x}\bar{x},i}^{n-1} + \frac{1}{\sigma} \varphi_i^n, \\
&\quad i = \overline{1, N-1},
\end{aligned}$$

яка розв'язується методом прогонки. Зауважимо, що при $\sigma > 0$ прогонка стійка.

7.5.2. Порядок апроксимації різницевої схеми з ваговими коефіцієнтами

Обчислимо похибку апроксимації схеми (7.53) — (7.55). Нехай $u(x, t)$ — розв'язок задачі (7.50) — (7.52), $y(x_i, t_n)$ — розв'язок різницевої задачі (7.53) — (7.55). Підставимо $y_i^n = z_i^n + u_i^n$ в (7.53) — (7.55), тоді отримаємо

$$\begin{aligned}
z_{\bar{t}\bar{t},i}^n &= \sigma z_{\bar{x}\bar{x},i}^{n+1} + (1 - 2\sigma) z_{\bar{x}\bar{x},i}^n + \sigma z_{\bar{x}\bar{x},i}^{n-1} + \psi_i^n, \\
i &= \overline{1, N-1}, \quad n = \overline{1, K-1},
\end{aligned}$$

$$z_0^n = z_N^n = 0, \quad n = \overline{0, K}, \quad z_i^0 = 0, \quad z_{t,i}^0 = \nu_i, \quad i = \overline{0, N},$$

де $\psi_i^n = \sigma u_{\bar{x}\bar{x},i}^{n+1} + (1 - 2\sigma) u_{\bar{x}\bar{x},i}^n + \sigma u_{\bar{x}\bar{x},i}^{n-1} - u_{\bar{t}\bar{t},i}^n + \varphi_i^n$ — похибка апроксимації схеми (7.53) на розв'язку $u = u(x, t)$, $\nu_i = \tilde{u}_0(x_i) - u_{t,i}^0$, — похибка апроксимації для другої початкової умови $y_{t,i}^0 = \tilde{u}_0(x_i)$. З попереднього випливає, що $\nu_i = O(\tau^2 + h^2)$.

Враховуючи, що $u_i^{n+1} = u_i^n + \tau u_{t,i}^n$, $u_i^{n-1} = u_i^n - \tau u_{t,i}^n$, маємо

$$\sigma u_i^{n+1} + (1 - 2\sigma) u_i^n + \sigma u_i^{n-1} = u_i^n + \sigma\tau^2 u_{\bar{t}\bar{t},i}^n.$$

Тоді

$$\begin{aligned}
\psi_i^n &= u_{\bar{x}\bar{x},i}^n + \sigma\tau^2 u_{\bar{x}\bar{x}\bar{t}\bar{t},i}^n - u_{\bar{t}\bar{t},i}^n + \varphi_i^n = \\
&= u''(x_i, t_n) + \sigma\tau^2 \ddot{u}''(x_i, t_n) - \ddot{u}(x_i, t_n) + \varphi_i^n + O(\tau^2 + h^2) = \\
&= \varphi_i^n - f(x_i, t_n) + O(\tau^2 + h^2).
\end{aligned}$$

Отже, при $\varphi_i^n = f(x_i, t_n) + O(\tau^2 + h^2)$ і при довільному σ схема має порядок апроксимації $O(\tau^2 + h^2)$.

7.6. Стійкість двоярусних та триярусних сіткових схем

7.6.1. Канонічний вигляд та умови стійкості двоярусних сіткових схем

Розглянемо крайову задачу для рівняння параболічного типу

$$\begin{aligned} \frac{\partial u}{\partial t} &= Lu + f(x, t), \quad t \in (t_0, T], \\ x &= (x_1, x_2, \dots, x_n) \in \Omega, \quad \Omega \subset \mathbb{R}^n, \\ lu &= \mu(x, t), \quad x \in \Gamma, \quad t \in [t_0, T], \\ u(x, t_0) &= u_0(x), \quad x \in \bar{\Omega} = \Omega \cup \Gamma, \end{aligned} \quad (7.56)$$

де L, l — диференціальні оператори, які діють на $u = u(x, t)$ як функцію $x = (x_1, x_2, \dots, x_n) \in \Omega$, $f(x, t)$, $\mu(x, t)$, $u_0(x)$ — задані функції. Якщо диференціальний оператор L та граничні умови на сітці $\bar{\omega}_h$ апроксимувати деяким сітковим оператором A_h , то одержимо задачу Коші для системи звичайних диференціальних рівнянь

$$\begin{aligned} \frac{dv_h}{dt} &= A_h v_h + F_h(t), \quad t \in (t_0, T], \\ v_h(t_0) &= u_{h,0}. \end{aligned} \quad (7.57)$$

Далі для спрощення індекс h будемо опускати.

На відрізку $[t_0, T]$ введемо рівномірну сітку

$$\omega_\tau = \{t_n = t_0 + n\tau, \quad n = \overline{0, K}, \quad K\tau = T - t_0\}$$

з кроком $\tau > 0$ і будемо розглядати функції $y(t_n) \in H_h$ дискретного аргумента $t_n \in \omega_\tau$ зі значеннями в просторі H_h . Функції $y_n = y(t_n)$ можуть залежати від h і τ , тобто $y(t_n) = y_{h\tau}(t_n)$. Для розв'язування задачі (7.57) розглянемо однокрокову (двоярусну) сіткову схему вигляду

$$B_1 y_{n+1} + B_2 y_n = \varphi_n, \quad y_0 = u_0, \quad (7.58)$$

де B_1, B_2 — лінійні оператори в H_h , які можуть залежати від h, τ, n , а $\varphi_n = \varphi(t_n)$ може залежати також від h, τ . Враховуючи тотожність

$$y_{n+1} = y_n + \tau \frac{y_{n+1} - y_n}{\tau},$$

одержимо, що будь-яку двоярусну схему можна записати на сітці ω_τ у канонічному вигляді:

$$B \frac{y_{n+1} - y_n}{\tau} + Ay_n = \varphi_n, \quad y_0 = u_0, \quad (7.59)$$

де $A = B_1 + B_2$, $B = \tau B_1$. Надалі будемо вважати, що існує оператор B^{-1} . Це гарантує існування та єдиність розв'язку задачі (7.59).

Сіткову схему (7.59) називають *стійкою*, якщо існують сталі $M_1 > 0$, $M_2 > 0$, які не залежать від h , τ , n і такі, що при $\forall \varphi_n, y_0 \in H_h$ для розв'язку (7.59) справджується оцінка

$$\|y_{n+1}\|_{1h} \leq M_1 \|y_0\|_{1h} + M_2 \max_{0 \leq j \leq n} \|\varphi_j\|_{2h}. \quad (7.60)$$

Якщо виконується нерівність (7.60), то схему називають *стійкою за початковими даними і за правою частиною*.

Схему (7.59) називають *стійкою за початковими даними*, якщо для розв'язку однорідного рівняння

$$B \frac{y_{n+1} - y_n}{\tau} + Ay_n = 0, \quad y_0 = u_0 \quad (7.61)$$

виконується оцінка

$$\|y_{n+1}\|_{1h} \leq M_1 \|y_0\|_{1h}.$$

Сіткову схему (7.59) називають *стійкою за правою частиною*, якщо для розв'язку задачі

$$B \frac{y_{n+1} - y_n}{\tau} + Ay_n = \varphi_n, \quad y_0 = 0$$

справджується оцінка

$$\|y_{n+1}\|_{1h} \leq M_2 \max_{0 \leq j \leq n} \|\varphi_j\|_{2h}.$$

Схему (7.59) називають *рівномірно стійкою за початковими даними*, якщо існують стала $\rho > 0$ і стала M_1 , яка не залежить від h, τ, n , такі, що при будь-якому $n = \overline{0, K-1}$

$$\|y_{n+1}\|_{1h} \leq \rho \|y_n\|_{1h},$$

причому $\rho^{n+1} \leq M_1$.

Запишемо однорідне рівняння (7.61) у вигляді

$$y_{n+1} = Sy_n, \quad n = \overline{0, K-1},$$

де $S = I - \tau B^{-1}A$ називають оператором переходу.

Зауважимо, що вимога рівномірної стійкості за початковими даними еквівалентна обмеженості норми оператора S константою ρ

$$\|S\| \leq \rho.$$

► **ТЕОРЕМА 7.2.** Нехай схема (7.59) рівномірно стійка за початковими даними. Тоді ця схема стійка і за правою частиною, причому для її розв'язку виконується оцінка (7.60), де $\|\varphi_j\|_{2h} = \|B_j^{-1}\varphi_j\|_{1h}$ і $M_2 = M_1(T - t_0)$.

Доведення. Запишемо (7.59) при $n = j$ у вигляді

$$y_{j+1} = S_{j+1}y_j + \tau B_j^{-1}\varphi_j$$

і використаємо нерівність трикутника

$$\begin{aligned} \|y_{j+1}\|_{1h} &\leq \|S_{j+1}\| \cdot \|y_j\|_{1h} + \tau \|B_j^{-1}\varphi_j\|_{1h} \\ &\leq \rho \|y_j\|_{1h} + \tau \|B_j^{-1}\varphi_j\|_{1h}. \end{aligned}$$

Послідовно застосовуючи цю нерівність, отримаємо

$$\begin{aligned} \|y_{n+1}\|_{1h} &\leq \rho \|y_n\|_{1h} + \tau \|B_n^{-1}\varphi_n\|_{1h} \leq \\ &\leq \rho^2 \|y_{n-1}\|_{1h} + \tau \rho \|B_{n-1}^{-1}\varphi_{n-1}\|_{1h} + \tau \|B_n^{-1}\varphi_n\|_{1h} \leq \\ &\leq \dots \leq \rho^{n+1} \|y_0\|_{1h} + \sum_{j=0}^n \tau \rho^{n-j} \|B_j^{-1}\varphi_j\|_{1h}. \end{aligned}$$

Оскільки $\rho^{n+1} \leq M_1$, $\rho^{n-j} \leq M_1$, то

$$\begin{aligned} \|y_{n+1}\|_{1h} &\leq M_1 \left(\|y_0\|_{1h} + \sum_{j=0}^n \tau \|B_j^{-1}\varphi_j\|_{1h} \right) \leq \\ &\leq M_1 \left(\|y_0\|_{1h} + \max_{0 \leq j \leq n} \|B_j^{-1}\varphi_j\| (t_{n+1} - t_0) \right) \leq \\ &\leq M_1 \|y_0\|_{1h} + M_1 (T - t_0) \max_{0 \leq j \leq n} \|B_j^{-1}\varphi_j\|. \end{aligned}$$

■

Надалі будемо розглядати рівномірну стійкість лише для випадку $\rho = 1$.

Припустимо, що в H_h введений скалярний добуток $(y, v)_h$, тоді $\|y\|_h = \sqrt{(y, y)_h}$. Для спрощення запису індекс h у скалярному добутку і нормі надалі будемо опускати.

► **ТЕОРЕМА 7.3.** Нехай у схемі (7.59) оператор A — самоспряжений додатний і не залежить від n . Якщо виконується операторна нерівність

$$B \geq 0,5\tau A, \quad (7.62)$$

то схема (7.59) рівномірно стійка за початковими даними, причому для розв'язку однорідного рівняння (7.61) справджується оцінка

$$\|y_{n+1}\|_A \leq \|y_n\|_A, \quad n = \overline{0, K-1}. \quad (7.63)$$

Доведення. Позначимо $y_t = (y_{n+1} - y_n)/\tau$, $y = y_n$ і помножимо рівняння (7.61) скалярно на y_t . Тоді отримаємо тотожність

$$(By_t, y_t) + (Ay, y_t) = 0,$$

яку можна записати у вигляді

$$((B - 0,5\tau A)y_t, y_t) + (0,5\tau Ay_t + Ay, y_t) = 0. \quad (7.64)$$

Оскільки

$$0,5\tau Ay_t + Ay = 0,5A(y_n + y_{n+1}),$$

то (7.64) запишемо у вигляді

$$((B - 0,5\tau A)y_t, y_t) + 0,5\tau^{-1}(A(y_{n+1} + y_n), y_{n+1} - y_n) = 0. \quad (7.65)$$

Використовуючи умови самоспряженості і додатності оператора A , а також незалежності від n , одержимо

$$\begin{aligned} (A(y_{n+1} + y_n), y_{n+1} - y_n) &= (Ay_{n+1}, y_{n+1}) - (Ay_{n+1}, y_n) + \\ &+ (Ay_n, y_{n+1}) - (Ay_n, y_n) = (Ay_{n+1}, y_{n+1}) - (Ay_n, y_n) = \\ &= \tau (Ay_n, y_n)_t = \tau (\|y_n\|_A^2)_t. \end{aligned}$$

Звідси і з (7.65) приходимо до тотожності

$$2((B - 0,5\tau A)y_t, y_t) + (\|y_n\|_A^2)_t = 0. \quad (7.66)$$

Оскільки, згідно з (7.62)

$$((B - 0,5\tau A) y_t, y_t) \geq 0,$$

то з (7.66) отримаємо, що для розв'язку рівняння (7.61) справджується нерівність

$$(\|y_n\|_A^2)_t \leq 0.$$

Звідси випливає оцінка (7.63), яка гарантує рівномірну стійкість за початковими даними. ■

Теорема 7.3 дозволяє сформулювати таке правило дослідження стійкості двоярусних сіткових схем. Насамперед треба звести схему до канонічного вигляду (7.59) і визначити оператори B і A . Відтак дослідити властивості оператора A . Якщо цей оператор є самоспряженим, додатним і не залежить від n , то залишається перевірити виконання операторної нерівності (7.62).

7.6.2. Стійкість різницевої схеми з ваговими коефіцієнтами для рівняння теплопровідності

Для дослідження стійкості різницевої схеми (7.41) — (7.43) зведемо її до канонічного вигляду. Запишемо спочатку цю схему у вигляді

$$\begin{aligned} \frac{y_1^{n+1} - y_1^n}{\tau} = & \sigma \frac{y_2^{n+1} - 2y_1^{n+1}}{h^2} + (1 - \sigma) \frac{y_2^n - 2y_1^n}{h^2} + \varphi_1^n + \\ & + \frac{\sigma \mu_1(t_{n+1}) + (1 - \sigma) \mu_1(t_n)}{h^2}, \quad n = \overline{0, K-1}, \end{aligned}$$

$$\begin{aligned} \frac{y_i^{n+1} - y_i^n}{\tau} = & \sigma y_{\bar{x}x,i}^{n+1} + (1 - \sigma) y_{\bar{x}x,i}^n + \varphi_i^n, \quad i = \overline{2, N-2}, \\ & n = \overline{0, K-1}, \end{aligned}$$

$$\begin{aligned} \frac{y_{N-1}^{n+1} - y_{N-1}^n}{\tau} = & \sigma \frac{-2y_{N-1}^{n+1} + y_{N-2}^{n+1}}{h^2} + \\ & + (1 - \sigma) \frac{-2y_{N-1}^n + y_{N-2}^n}{h^2} + \varphi_{N-1}^n + \\ & + \frac{\sigma \mu_2(t_{n+1}) + (1 - \sigma) \mu_2(t_n)}{h^2}, \quad n = \overline{0, K-1}, \end{aligned}$$

$$y_i^0 = u_0(x_i), \quad i = \overline{0, N}.$$

Введемо простір H_{N-1} сіткових функцій, заданих на сітці ω_h , а також простір $\overset{\circ}{H}_{N+1}$ функцій, заданих на сітці $\bar{\omega}_h$ і рівних нулю при $i = 0, i = N$. Визначимо оператор A за формулами

$$(Ay)_i = -\overset{\circ}{y}_{\bar{x}x,i}, \quad i = \overline{1, N-1}, \quad y \in H_{N-1}, \quad \overset{\circ}{y} \in \overset{\circ}{H}_{N+1}.$$

Позначимо через $y_n \in H_{N-1}$ вектор $y_n = (y_1^n, y_2^n, \dots, y_{N-1}^n)$. Тоді різницеву схему (7.41) — (7.43) можна записати в операторному вигляді:

$$\frac{y_{n+1} - y_n}{\tau} + \sigma Ay_{n+1} + (1 - \sigma) Ay_n = \varphi_n,$$

де

$$\begin{aligned} \varphi_n &= (\tilde{\varphi}_1^n, \varphi_2^n, \dots, \varphi_{N-2}^n, \tilde{\varphi}_{N-1}^n), \\ \tilde{\varphi}_1^n &= \varphi_1^n + \frac{\sigma \mu_1(t_{n+1}) + (1 - \sigma) \mu_1(t_n)}{h^2}, \\ \tilde{\varphi}_{N-1}^n &= \varphi_{N-1}^n + \frac{\sigma \mu_2(t_{n+1}) + (1 - \sigma) \mu_2(t_n)}{h^2}. \end{aligned}$$

Використовуючи тотожність

$$y_{n+1} = y_n + \tau \frac{y_{n+1} - y_n}{\tau},$$

будемо мати

$$(I + \sigma \tau A) \frac{y_{n+1} - y_n}{\tau} + Ay_n = \varphi_n,$$

де I — одиничний оператор. Якщо ввести оператор

$$B = I + \sigma \tau A,$$

то одержимо канонічну форму двоярусних сіткових схем.

Оператор A — самоспряжений додатно визначений у просторі зі скалярним добутком

$$(y, v) = \sum_{i=1}^{N-1} \overset{\circ}{y}_i \overset{\circ}{v}_i h.$$

Умова стійкості має вигляд (див. теорема 7.3)

$$B = I + \sigma \tau A \geq 0,5 \tau A$$

і означає, що для $\forall y \in H_{N-1}$ повинна виконуватися нерівність

$$(\sigma - 0,5) \tau (Ay, y) + \|y\|^2 \geq 0.$$

Оскільки

$$(Ay, y) \leq \lambda_{N-1} \|y\|^2, \quad \lambda_{N-1} = \frac{4}{h^2} \cos^2 \frac{\pi h}{2l},$$

то

$$\begin{aligned} (\sigma - 0,5) \tau (Ay, y) + \|y\|^2 &\geq \\ &\geq (\sigma - 0,5) \tau (Ay, y) + \frac{1}{\lambda_{N-1}} (Ay, y) \geq 0 \end{aligned}$$

і схема стійка при

$$(\sigma - 0,5) \tau + \frac{1}{\lambda_{N-1}} \geq 0.$$

Явна різницева схема ($\sigma = 0$) стійка за умови $\tau \leq h^2/2$, тобто умовно стійка. Неявна різницева схема ($\sigma = 1$), схема Кранка–Нікольсона ($\sigma = 1/2$) та схема підвищеного порядку точності ($\sigma = 1/2 - h^2/(12\tau)$) — безумовно стійкі.

7.6.3. Канонічний вигляд та умови стійкості триярусних сіткових схем

Двокроковою (триярусною) сітковою схемою називають схему вигляду

$$\begin{aligned} B_2 y_{n+1} + B_1 y_n + B_0 y_{n-1} &= \varphi_n, \quad n = \overline{1, K-1}, \\ y_0 &= u_0, \quad y_1 = u_1. \end{aligned} \quad (7.67)$$

Оператори B_0, B_1, B_2 можуть залежати від h, τ, n , а функція $\varphi_n = \varphi(t_n)$ може залежати також від τ і h .

Введемо на сітці ω_τ різницеві співвідношення

$$\begin{aligned} y_t &= \frac{y_{n+1} - y_n}{\tau}, \quad y_{\bar{t}} = \frac{y_n - y_{n-1}}{\tau}, \quad y_t^\circ = \frac{y_{n+1} - y_{n-1}}{2\tau} \\ y_{\bar{t}t} &= \frac{y_t - y_{\bar{t}}}{\tau} = \frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2}. \end{aligned}$$

Безпосередньою перевіркою можна встановити справедливості таких тотожностей:

$$y_{n-1} = y - \tau y_t^\circ + 0,5\tau^2 y_{\bar{t}t}, \quad y_{n+1} = y + \tau y_t^\circ + 0,5\tau^2 y_{\bar{t}t},$$

де $y = y_n$. Підставляючи ці тотожності у рівняння (7.67), матимемо

$$\tau (B_2 - B_0) y_t^\circ + 0,5\tau^2 (B_2 + B_0) y_{\bar{t}t} + (B_2 + B_1 + B_0) y = \varphi,$$

де $\varphi = \varphi_n$. Позначимо

$$B = \tau(B_2 - B_0), \quad R = 0,5(B_2 + B_0), \quad A = B_2 + B_1 + B_0.$$

Тоді (7.67) можна записати у вигляді

$$By_t^\circ + \tau^2 Ry_{\bar{t}t} + Ay = \varphi, \quad (7.68)$$

Рівність (7.68) називають *канонічним виглядом* триярусної сіткової схеми.

Покажемо, що триярусну схему завжди можна записати у вигляді деякої еквівалентної двоярусної схеми.

Введемо простір $H^2 = H \oplus H$, який є прямою сумою двох просторів $H = H_h$. Простір H^2 визначається як множина векторів вигляду $Y = (Y^{(1)}, Y^{(2)})$, $Y^{(1)}, Y^{(2)} \in H$, а операція додавання і множення на число вводяться покоординатно, тобто

$$Y + \bar{Y} = (Y^{(1)} + \bar{Y}^{(1)}, Y^{(2)} + \bar{Y}^{(2)}), \quad \alpha Y = (\alpha Y^{(1)}, \alpha Y^{(2)}).$$

Якщо в H задано скалярний добуток $(\cdot, \cdot)_H$, то

$$(Y, \bar{Y})_{H^2} = (Y^{(1)}, \bar{Y}^{(1)})_H + (Y^{(2)}, \bar{Y}^{(2)})_H.$$

Визначимо вектори $Y_n, F_n \in H^2$ як

$$Y_n = \left(\frac{1}{2}(y_n + y_{n-1}), y_n - y_{n-1} \right), \quad F_n = (\varphi_n, 0),$$

де y_n — розв'язок рівняння (7.68), а φ_n — права частина (7.68). Введемо оператори

$$\begin{aligned} \bar{A} &= \begin{pmatrix} A & 0 \\ 0 & R - \frac{1}{4}A \end{pmatrix}, \\ \bar{B} &= \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \end{aligned} \quad (7.69)$$

так, щоб (7.68) можна було записати у вигляді

$$\bar{B}Y_t + \bar{A}Y = F, \quad (7.70)$$

де $F = F_n$. Оскільки

$$Y_t = (y_t^\circ, \tau y_{\bar{t}t}),$$

то (7.70) можна записати у вигляді системи рівнянь

$$B_{11}y_t^\circ + \tau B_{12}y_{\bar{t}t} + \frac{1}{2}A(y_n + y_{n-1}) = \varphi_n, \quad (7.71)$$

$$B_{21}y_t^\circ + \tau B_{22}y_{\bar{t}t} + \left(R - \frac{1}{4}A\right) \tau y_{\bar{t}} = 0. \quad (7.72)$$

Виберемо оператори B_{11} , B_{12} , B_{21} , B_{22} так, щоб рівняння (7.71) було еквівалентне (7.68), а (7.72) перетворилося тотожно в нуль. Перепишемо (7.72), враховуючи, що

$$y_{\bar{t}} = y_t^\circ - 0,5\tau y_{\bar{t}t}$$

у вигляді

$$\left[B_{21} + \tau \left(R - \frac{1}{4}A\right)\right] y_t^\circ + \tau \left[B_{22} - 0,5\tau \left(R - \frac{1}{4}A\right)\right] y_{\bar{t}t} = 0.$$

Очевидно, що остання рівність перетворюється в тотожність, якщо:

$$B_{21} = -\tau \left(R - \frac{1}{4}A\right), \quad B_{22} = 0,5\tau \left(R - \frac{1}{4}A\right).$$

Оскільки

$$y_{n-1} = y - \tau y_t^\circ + \frac{\tau^2}{2} y_{\bar{t}t},$$

то (7.71) еквівалентне рівнянню

$$\left(B_{11} - \frac{1}{2}\tau A\right) y_t^\circ + \tau^2 \left(\frac{B_{12}}{\tau} + \frac{1}{4}A\right) y_{\bar{t}t} + Ay = \varphi_n.$$

Порівнюючи цей вираз з рівнянням (7.68), одержимо

$$B_{11} = B + \frac{\tau}{2}A, \quad B_{12} = \tau \left(R - \frac{1}{4}A\right).$$

Отже, триярусна схема (7.68) еквівалентна двоярусній (7.70), де

$$\bar{B} = \begin{pmatrix} B + \frac{\tau}{2}A & \tau \left(R - \frac{1}{4}A\right) \\ -\tau \left(R - \frac{1}{4}A\right) & \frac{\tau}{2} \left(R - \frac{1}{4}A\right) \end{pmatrix}. \quad (7.73)$$

Розглянемо задачу Коші

$$By_t + \tau^2 Ry_{\bar{t}} + Ay = 0, \quad n = \overline{1, K-1}, \quad (7.74)$$

y_0, y_1 — задані. Припустимо, що існує оператор $(B+2\tau R)^{-1}$, а тому рівняння (7.74) однозначно розв'язне відносно y_{n+1} . Справджується така теорема про рівномірну стійкість схеми (7.68) за початковими даними.

► **ТЕОРЕМА 7.4.** Нехай A і R — самоспряжені додатні оператори, незалежні від n і такі, що:

$$R > \frac{1}{4}A, \quad B \geq 0,$$

то для будь-яких $y_0, y_1 \in H$ для розв'язку сіткової схеми (7.74) справджується нерівність

$$\|Y_{n+1}\|_{\bar{A}} \leq \|Y_n\|_{\bar{A}}, \quad n = \overline{1, K-1}, \quad (7.75)$$

де

$$\begin{aligned} \|Y_n\|_{\bar{A}}^2 = & \frac{1}{4} (A(y_n + y_{n-1}), y_n + y_{n-1}) + \\ & + \left(\left(R - \frac{1}{4}A \right) (y_n - y_{n-1}), y_n - y_{n-1} \right). \end{aligned}$$

Доведення. Запишемо схему (7.74) у вигляді двоярусної схеми

$$\bar{B} \frac{Y_{n+1} - Y_n}{\tau} + \bar{A} Y_n = 0, \quad n = \overline{1, K-1}, \quad (7.76)$$

де $Y_n \in H^2$, і оператори \bar{A}, \bar{B} визначаються формулами (7.69), (7.73), з початковою умовою $Y_1 = (0, 5(y_0 + y_1), y_1 - y_0)$.

Оскільки для будь-якого $Y = (Y^{(1)}, Y^{(2)}) \in H^2$

$$\bar{A}Y = \left(AY^{(1)}, \left(R - \frac{1}{4}A \right) Y^{(2)} \right),$$

то із самоспряженості A, R випливає рівність

$$\begin{aligned} (\bar{A}Y, \bar{Y}) &= (AY^{(1)}, \bar{Y}^{(1)}) + \left(\left(R - \frac{1}{4}A \right) Y^{(2)}, \bar{Y}^{(2)} \right) = \\ &= (Y^{(1)}, A\bar{Y}^{(1)}) + \left(Y^{(2)}, \left(R - \frac{1}{4}A \right) \bar{Y}^{(2)} \right) = \\ &= (Y, \bar{A}\bar{Y}), \end{aligned}$$

а з операторних нерівностей $A > 0$, $R > \frac{1}{4}A$ — нерівність

$$(\bar{A}Y, Y) = (AY^{(1)}, Y^{(1)}) + \left(\left(R - \frac{1}{4}A \right) Y^{(2)}, Y^{(2)} \right) > 0.$$

Отже, оператор \bar{A} — самоспряжений і додатний в H^2 .

У просторі H^2 визначимо норму $\|Y\|_{\bar{A}}$ за формулою

$$\|Y\|_{\bar{A}}^2 = (AY^{(1)}, Y^{(1)}) + \left(\left(R - \frac{1}{4}A \right) Y^{(2)}, Y^{(2)} \right).$$

тоді для $Y_n = (0,5(y_n + y_{n-1}), y_n - y_{n-1})$

$$\begin{aligned} \|Y_n\|_{\bar{A}}^2 &= \frac{1}{4} (A(y_n + y_{n-1}), y_n + y_{n-1}) + \\ &+ \left(\left(R - \frac{1}{4}A \right) (y_n - y_{n-1}), y_n - y_{n-1} \right). \end{aligned}$$

Перевіримо виконання умови стійкості $\bar{B} \geq \frac{\tau}{2}\bar{A}$ для схеми (7.76). З (7.69), (7.73) випливає, що

$$\bar{B} - \frac{\tau}{2}\bar{A} = \begin{pmatrix} B & \tau \left(R - \frac{1}{4}A \right) \\ -\tau \left(R - \frac{1}{4}A \right) & 0 \end{pmatrix}.$$

Для будь-якого елемента $Y = (Y^{(1)}, Y^{(2)}) \in H^2$ маємо

$$\left(\bar{B} - \frac{\tau}{2}\bar{A} \right) Y = \left(BY^{(1)} + \tau \left(R - \frac{1}{4}A \right) Y^{(2)}, -\tau \left(R - \frac{1}{4}A \right) Y^{(1)} \right).$$

Враховуючи самоспряженість оператора $R - \frac{1}{4}A$ та умову $B \geq 0$, одержимо

$$\begin{aligned} \left(\left(\bar{B} - \frac{\tau}{2}\bar{A} \right) Y, Y \right)_{H^2} &= (BY^{(1)}, Y^{(1)})_H + \\ &+ \tau \left(\left(R - \frac{1}{4}A \right) Y^{(2)}, Y^{(1)} \right)_H - \\ &- \tau \left(\left(R - \frac{1}{4}A \right) Y^{(1)}, Y^{(2)} \right)_H = \\ &= (BY^{(1)}, Y^{(1)})_H. \end{aligned}$$

Отже, на підставі теореми 7.3 справджується оцінка (7.75). ■

7.6.4. Стійкість різницевої схеми з ваговими коефіцієнтами для рівняння коливання струни

Дослідимо стійкість різницевої схеми (7.53) — (7.55). Введемо оператор

$$(Ay)_i = -\overset{\circ}{y}_{\bar{x}x,i}, \quad i = \overline{1, N-1}, \quad y \in H_{N-1}, \quad \overset{\circ}{y} \in \overset{\circ}{H},$$

де H_{N-1} — простір сіткових функцій y , заданих у вузлах сітки $\omega_h = \{x_i = ih, i = \overline{1, N-1}, h = l/N\}$, а $\overset{\circ}{H}_{N+1}$ — простір функцій $\overset{\circ}{y}$ заданих на сітці $\bar{\omega}_h = \{x_i = ih, i = \overline{0, N}, h = l/N\}$ і рівних нулю на границі при $i = 0, N$. Тоді різницеву схему (7.53) — (7.55) можна записати у вигляді

$$y_{\bar{t}t} + \sigma Ay_{n+1} + (1 - 2\sigma)Ay_n + \sigma Ay_{n-1} = \varphi_n, \quad (7.77)$$

де

$$y_n = (y_1^n, y_2^n, \dots, y_{N-1}^n), \quad y_{\bar{t}t} = (y_{n+1} - 2y_n + y_{n-1})/\tau^2,$$

$$\varphi_n = (\tilde{\varphi}_1^n, \varphi_2^n, \dots, \tilde{\varphi}_{N-1}^n),$$

$$\tilde{\varphi}_1^n = \varphi_1^n + \frac{\sigma\mu_1(t_{n+1}) + (1 - 2\sigma)\mu_1(t_n) + \sigma\mu_1(t_{n-1})}{h^2},$$

$$\tilde{\varphi}_{N-1}^n = \varphi_{N-1}^n + \frac{\sigma\mu_2(t_{n+1}) + (1 - 2\sigma)\mu_2(t_n) + \sigma\mu_2(t_{n-1})}{h^2}.$$

Використовуючи тотожності

$$y_{n-1} = y - \tau y_t^\circ + 0,5\tau^2 y_{\bar{t}t}, \quad y_{n+1} = y + \tau y_t^\circ + 0,5\tau^2 y_{\bar{t}t}$$

схему (7.77) зведемо до вигляду

$$(I + \sigma\tau^2 A) y_{\bar{t}t} + Ay_n = \varphi_n$$

де I — одиничний оператор. Якщо ввести оператори

$$B = 0, \quad R = \frac{1}{\tau^2}I + \sigma A,$$

то цю схему можна записати в канонічному вигляді (7.68).

Оператори A і R — самоспряжені. Тоді для стійкості різницевої схеми достатньо виконання умов $R > \frac{1}{4}A$, $B \geq 0$, тобто

$$\frac{1}{\tau^2}I + \left(\sigma - \frac{1}{4}\right)A > 0,$$

або

$$\frac{1}{\tau^2} \|y\|^2 + \left(\sigma - \frac{1}{4}\right) (Ay, y) > 0. \quad (7.78)$$

Оскільки

$$\|y\|^2 > \frac{1}{\Delta} (Ay, y) > 0, \quad \Delta = \frac{4}{h^2},$$

то нерівність (7.78) буде виконуватися, якщо вимагати

$$\frac{1}{\Delta \cdot \tau^2} + \sigma - \frac{1}{4} > 0.$$

Отже, різницева схема (7.53) — (7.55) стійка, коли

$$\sigma > \frac{1}{4} \left(1 - \frac{1}{\gamma}\right), \quad \gamma = \frac{\tau^2}{h^2}.$$

Наприклад, явна схема ($\sigma = 0$) стійка при $\tau^2/h^2 < 1$, тобто $\tau < h$. Якщо $\sigma = 1/4$, то різницева схема безумовно стійка.

Приклад 7.4. Зведіть до канонічного вигляду та дослідіть на стійкість різницеву схему

$$\begin{aligned} \frac{1}{12} y_{t,i+1}^n + \frac{5}{6} y_{t,i}^n + \frac{1}{12} y_{t,i-1}^n &= \frac{1}{2} y_{\bar{x}x,i}^{n+1} + \frac{1}{2} y_{\bar{x}x,i}^n, \\ i = \overline{1, N-1}, \quad n = \overline{0, K-1}, \quad \tau &= T/K, \quad h = 1/N \\ y_i^0 &= u_0(x_i), \quad i = \overline{0, N}, \quad y_0^n = y_N^n = 0, \quad n = \overline{0, K}. \end{aligned}$$

▷ Враховуючи рівність $v_{i+1} + 10v_i + v_{i-1} = v_{i+1} - 2v_i + v_{i-1} + 12v_i = 12v_i + h^2 v_{\bar{x}x,i}$ різницеву схему запишемо у вигляді

$$y_{t,i}^n + \frac{h^2}{12} y_{\bar{x}x,i}^n = \frac{1}{2} y_{\bar{x}x,i}^{n+1} + \frac{1}{2} y_{\bar{x}x,i}^n, \quad i = \overline{1, N-1}, \quad n = \overline{0, K-1}. \quad (7.79)$$

Визначимо оператор A за формулами

$$(Ay)_i = -\overset{\circ}{y}_{\bar{x}x,i}, \quad i = \overline{1, N-1}, \quad y \in H_{N-1}, \quad \overset{\circ}{y} \in \overset{\circ}{H}_{N+1},$$

де H_{N-1} — простір сіткових функцій y , заданих у вузлах сітки ω_h , а $\overset{\circ}{H}_{N+1}$ — простір функцій $\overset{\circ}{y}$ заданих на сітці $\bar{\omega}_h$ і рівних нулю на границі при $i = 0, N$. Позначимо через $y_n \in H_{N-1}$ вектор $y_n = (y_1^n, y_2^n, \dots, y_{N-1}^n)$, $y_t = (y_{t,1}^n, y_{t,2}^n, \dots, y_{t,N-1}^n)$. Тоді різницеву схему (7.79) можна записати у вигляді:

$$y_t - \frac{h^2}{12} Ay_t + \frac{1}{2} Ay_{n+1} + \frac{1}{2} Ay_n = 0.$$

Використовуючи тотожність

$$y_{n+1} = y_n + \tau y_t.$$

будемо мати

$$\left(I + \frac{1}{2} \left(\tau - \frac{h^2}{6} \right) \right) y_t + Ay = 0,$$

де I — одиничний оператор. Якщо ввести оператор

$$B = I + \frac{1}{2} \left(\tau - \frac{h^2}{6} \right) A$$

то одержимо канонічну форму двоярусних сіткових схем.

Умова стійкості $B \geq \frac{\tau}{2} A$ буде мати вигляд

$$I - \frac{h^2}{6} A \geq 0 \quad \text{або} \quad \|y\|^2 - \frac{h^2}{6} (Ay, y) \geq 0. \quad (7.80)$$

З урахуванням

$$(Ay, y) \leq \lambda_{N-1} \|y\|^2 < \frac{4}{h^2} \|y\|^2,$$

нерівність (7.80) буде справджуватися, оскільки

$$\|y\|^2 - \frac{h^2}{6} (Ay, y) > \frac{h^2}{12} (Ay, y) > 0.$$

◀

Приклад 7.5. Зведіть до канонічного вигляду та дослідіть на стійкість різницьову схему

$$y_{t,i}^n + \sigma \tau y_{tt,i}^n = y_{\bar{x}x,i}^{n+1} + \varphi_i^n, \quad i = \overline{1, N-1}, \quad n = \overline{1, K-1}, \quad (7.81)$$

$$y_i^0 = u_0(x_i), \quad i = \overline{0, N}, \quad y_0^n = \mu_1(t_n), \quad y_N^n = \mu_2(t_n), \quad n = \overline{0, K}. \quad (7.82)$$

▷ Задана різницьова схема — триярусна. Введемо оператор A

$$(Ay)_i = -\overset{\circ}{y}_{\bar{x}x,i}, \quad i = \overline{1, N-1}, \quad y \in H_{N-1}, \quad \overset{\circ}{y} \in H_{N-1},$$

а також вектори

$$y_n = (y_1^n, y_2^n, \dots, y_{N-1}^n), \quad y_t = (y_{t,1}^n, y_{t,2}^n, \dots, y_{t,N-1}^n),$$

$$y_{tt} = (y_{tt,1}^n, y_{tt,2}^n, \dots, y_{tt,N-1}^n), \quad \varphi_n = (\tilde{\varphi}_1^n, \varphi_2^n, \dots, \varphi_{N-2}^n, \tilde{\varphi}_{N-1}^n),$$

$$\tilde{\varphi}_1^n = \varphi_1^n + \frac{\mu_1(t_{n+1})}{h^2}, \quad \tilde{\varphi}_{N-1}^n = \varphi_{N-1}^n + \frac{\mu_2(t_{n+1})}{h^2}.$$

Тоді різницеву схему (7.81), (7.82) можна записати у вигляді

$$y_t + \sigma \tau y_{\bar{t}t} + Ay_{n+1} = \varphi_n, \quad y_0 = u_0.$$

Використовуючи тотожності

$$y_{n+1} = y_n + \tau y_t^\circ + 0,5\tau^2 y_{\bar{t}t},$$

$$y_t = y_t^\circ + 0,5\tau y_{\bar{t}t},$$

отримаємо

$$(I + \tau A)y_t^\circ + \tau^2 \left(\frac{\sigma + 0,5}{\tau} I + 0,5A \right) y_{\bar{t}t} + Ay_n = \varphi_n.$$

Отже, різницева схема (7.81), (7.82) зведена до канонічного вигляду (7.68), де

$$B = I + \tau A, \quad R = \frac{\sigma + 0,5}{\tau} I + 0,5A.$$

Оператори A, R — самоспряжені додатні. Тоді за теоремою 7.4 для стійкості різницевої схеми достатньо виконання умов $B \geq 0$, $R > \frac{1}{4}A$. На підставі властивості оператора A другої різницевої похідної $B > 0$, а умова $R > \frac{1}{4}A$ зводиться до

$$\frac{\sigma + 0,5}{\tau} I + \frac{1}{4}A > 0$$

або

$$\frac{\sigma + 0,5}{\tau} \|y\|^2 + \frac{1}{4}(Ay, y) > 0.$$

Ця нерівність справджується при $\sigma > -0,5 - \tau/h^2$. ◀

7.7. Різницева апроксимація задачі Діріхле для рівняння Пуассона

Розглянемо задачу Діріхле для рівняння Пуассона: знайти неперервну в $\bar{\Omega} = \Omega \cup \Gamma$ функцію $u(x_1, x_2)$, яка задовольняє рівняння

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x), \quad x = (x_1, x_2) \in \Omega, \quad (7.83)$$

і граничну умову

$$u(x) = \mu(x), \quad x \in \Gamma, \quad (7.84)$$

де $\Omega = \{0 < x_\alpha < l_\alpha, \alpha = 1, 2\}$, а Γ — границя області Ω , $f(x)$, $\mu(x)$ — задані функції. Припустимо, що $f(x)$, $\mu(x)$ такі, що розв'язок задачі (7.83), (7.84) існує та єдиний і є достатньо гладкою функцією.

Введемо в $\bar{\Omega}$ прямокутну сітку

$$\bar{\omega}_h = \left\{ x_{ij} = (x_1^i, x_2^j), x_1^i = ih_1, x_2^j = jh_2, \right. \\ \left. i = \overline{0, N_1}, j = \overline{0, N_2}, h_1 = l_1/N_1, h_2 = l_2/N_2 \right\}.$$

Для розв'язування задачі (7.83), (7.84) розглянемо різницеву схему

$$y_{\bar{x}_1 x_1, ij} + y_{\bar{x}_2 x_2, ij} = -f_{ij}, \quad i = \overline{1, N_1 - 1}, j = \overline{1, N_2 - 1}, \quad (7.85)$$

$$\begin{aligned} y_{i,0} &= \mu(x_1^i, 0), & y_{i,N_2} &= \mu(x_1^i, l_2), & i &= \overline{1, N_1 - 1}, \\ y_{0,j} &= \mu(0, x_2^j), & y_{N_1,j} &= \mu(l_1, x_2^j), & j &= \overline{1, N_2 - 1}, \end{aligned} \quad (7.86)$$

де

$$y_{ij} = y(x_{ij}), \quad y_{\bar{x}_1 x_1, ij} = \frac{y_{i+1,j} - 2y_{ij} + y_{i-1,j}}{h_1^2}, \\ y_{\bar{x}_2 x_2, ij} = \frac{y_{i,j+1} - 2y_{ij} + y_{i,j-1}}{h_2^2}.$$

Точки x_{ij} , в яких записується рівняння (7.85), належать підмножині $\omega_h = \{x_{ij}, i = \overline{1, N_1 - 1}, j = \overline{1, N_2 - 1}\}$ сітки $\bar{\omega}_h$, яку називають *множиною внутрішніх вузлів*. Множину точок $\gamma_h = \{x_{i0}, x_{iN_2}\}_{i=1}^{N_1-1} \cup \{x_{0j}, x_{N_1j}\}_{j=1}^{N_2-1}$, в яких задані різницеві граничні умови (7.86), називають *границею сітки $\bar{\omega}_h$* . Зауважимо, що кутові точки $(0,0)$, $(l_1,0)$, $(0,l_2)$, (l_1,l_2) не беруть участі у цій апроксимації і тому не належать ні до внутрішніх, ні до граничних.

Різницеве рівняння (7.85) записано на п'ятиточковому шаблоні $x_{i,j}, x_{i-1,j}, x_{i+1,j}, x_{i,j-1}, x_{i,j+1}$ (рис. 7.4), а тому схему (7.85) часто називають схемою “хрест”.

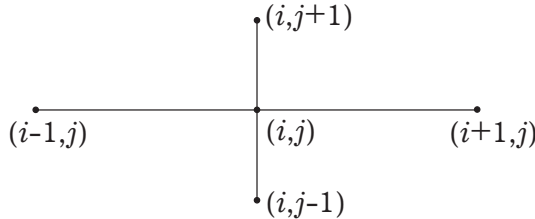


Рис. 7.4.

Нехай $u(x_1, x_2)$ — розв'язок задачі Діріхле (7.83), (7.84), а y_{ij} — розв'язок різницевої задачі (7.85), (7.86). Розглянемо похибку $z_{ij} = y_{ij} - u(x_1^i, x_2^j)$. Підставляючи $y_{ij} = z_{ij} + u_{ij}$ в (7.85), (7.86), одержимо для $z = z(x)$, $x \in \bar{\omega}_h$ різницеву задачу

$$z_{\bar{x}_1 x_1, ij} + z_{\bar{x}_2 x_2, ij} = -\psi_{ij}, \quad x_{ij} \in \omega_h, \quad (7.87)$$

$$z_{ij} = 0, \quad x_{ij} \in \gamma_h, \quad (7.88)$$

де $\psi_{ij} = u_{\bar{x}_1 x_1, ij} + u_{\bar{x}_2 x_2, ij} + f_{ij}$ — похибка апроксимації на розв'язку рівняння (7.83).

Враховуючи формули

$$u_{\bar{x}_1 x_1, ij} = \frac{\partial^2 u(x_1^i, x_2^j)}{\partial x_1^2} + \frac{h_1^2}{12} \frac{\partial^4 u(x_1^i, x_2^j)}{\partial x_1^4} + O(h_1^4),$$

$$u_{\bar{x}_2 x_2, ij} = \frac{\partial^2 u(x_1^i, x_2^j)}{\partial x_2^2} + \frac{h_2^2}{12} \frac{\partial^4 u(x_1^i, x_2^j)}{\partial x_2^4} + O(h_2^4),$$

отримаємо

$$\begin{aligned} \psi_{ij} &= \frac{\partial^2 u(x_1^i, x_2^j)}{\partial x_1^2} + \frac{\partial^2 u(x_1^i, x_2^j)}{\partial x_2^2} + \frac{h_1^2}{12} \frac{\partial^4 u(x_1^i, x_2^j)}{\partial x_1^4} + \\ &+ \frac{h_2^2}{12} \frac{\partial^4 u(x_1^i, x_2^j)}{\partial x_2^4} + f(x_1^i, x_2^j) + O(h_1^4 + h_2^4) = \\ &= O(|h|^2), \quad |h|^2 = h_1^2 + h_2^2. \end{aligned} \quad (7.89)$$

Отже, схема (7.85), (7.86) має другий порядок апроксимації.

7.8. Принцип максимуму для різницевих схем

7.8.1. Принцип максимуму

Нехай ω — скінченна множина точок (сітка) у деякій обмеженій області n -вимірного евклідового простору. Кожному вузлу $x \in \omega$ поставимо у відповідність підмножину точок сітки $\sigma'(x)$, яка не містить x і називається околom точки x .

Розглянемо рівняння

$$A(x) y(x) = \sum_{\xi \in \sigma'(x)} B(x, \xi) y(\xi) + F(x), \quad x \in \omega, \quad (7.90)$$

де $y(x)$ — невідома сіткова функція, задана на ω , а $A(x)$, $B(x, \xi)$, $F(x)$ — задані сіткові функції, які задовольняють умови

$$A(x) > 0, \quad B(x, \xi) > 0 \quad \forall x \in \omega, \quad \xi \in \sigma'(x), \quad (7.91)$$

$$D(x) = A(x) - \sum_{\xi \in \sigma'(x)} B(x, \xi) \geq 0. \quad (7.92)$$

Точку x називають *граничним* вузлом сітки, якщо в цій точці задано значення функції $y(x)$, тобто

$$y(x) = \mu(x), \quad x \in \gamma, \quad (7.93)$$

де γ — множина граничних вузлів, $\mu(x)$ — задана функція. Якщо на границі γ формально покласти

$$A(x) \equiv 1, \quad B(x, \xi) \equiv 0, \quad F(x) = \mu(x),$$

то граничну умову (7.93) можна записати у вигляді (7.90). Вузли, в яких розглядається рівняння (7.90), називають *внутрішніми вузлами сітки*. Надалі ω — множина внутрішніх вузлів сітки, а $\bar{\omega} = \omega \cup \gamma$ — множина всіх вузлів сітки.

Запис різницевої схеми у вигляді (7.90), називається *канонічною формою*. У канонічній формі можна записати будь-яке лінійне різницеве рівняння. В кожному конкретному випадку необхідно задати множину $\sigma'(x)$, коефіцієнти $A(x)$, $B(x, \xi)$ і функцію $F(x)$.

Будемо вважати, що сітка $\bar{\omega}$ зв'язна, якщо для будь-яких двох її вузлів x_0 , x'_0 таких, що хоча б один з них має непорожній окіл, існує така множина вузлів $x_i \in \omega$, $i = \overline{1, m}$, що $x_1 \in \sigma'(x_0)$, $x_2 \in \sigma'(x_1)$, \dots , $x_m \in \sigma'(x_{m-1})$, $x'_0 \in \sigma'(x_m)$, тобто кожен наступний вузол належить околу попереднього. Зв'язність сітки означає, що від будь-якого вузла $x_0 \in \omega$ можна перейти до будь-якого іншого вузла x'_0 , використовуючи тільки заданий шаблон.

Введемо позначення

$$Ly(x) = A(x)y(x) - \sum_{\xi \in \sigma'(x)} B(x, \xi)y(\xi), \quad (7.94)$$

і рівняння (7.90) запишемо у вигляді

$$Ly(x) = F(x). \quad (7.95)$$

Зауважимо, що вираз $Ly(x)$ можна записати ще так:

$$Ly(x) = D(x)y(x) + \sum_{\xi \in \sigma'(x)} B(x, \xi)(y(x) - y(\xi)). \quad (7.96)$$

► **ТЕОРЕМА 7.5. (ПРИНЦИП МАКСИМУМУ)** Нехай $y(x) \neq \text{const}$ — сіткова функція, визначена на зв'язній сітці $\bar{\omega}$, і виконуються умови (7.91), (7.92). Тоді, якщо $Ly(x) \leq 0$, ($Ly(x) \geq 0$) на ω , то $y(x)$ не може приймати найбільшого додатного (відповідно найменшого від'ємного) значення в усіх внутрішніх вузлах $x \in \omega$.

Доведення. Нехай $Ly(x) \leq 0$ для всіх $x \in \omega$. Припустимо, що $y(x)$ приймає найбільше додатне значення у внутрішньому вузлі $x_0 \in \omega$, тобто

$$y(x_0) = \max_{x \in \bar{\omega}} y(x) > 0. \quad (7.97)$$

У цій точці вираз

$$Ly(x_0) = D(x_0)y(x_0) + \sum_{\xi \in \sigma'(x_0)} B(x_0, \xi)(y(x_0) - y(\xi)) \quad (7.98)$$

невід'ємний. Дійсно згідно з умовами (7.91), (7.92) і припущенням (7.97) маємо $D(x_0) \geq 0$, $y(x_0) > 0$, $B(x_0, \xi) > 0$, $y(x_0) \geq y(\xi)$, так що $Ly(x_0) \geq 0$. А це суперечить умові $Ly(x) \leq 0$. Отже, якщо виконується умова (7.97) в точці $x_0 \in \omega$, то $Ly(x_0) = 0$. Враховуючи невід'ємність всіх доданків правої частини виразу (7.98), одержимо

$$D(x_0)y(x_0) = 0, \quad B(x_0, \xi)(y(x_0) - y(\xi)) = 0, \quad \xi \in \sigma'(x_0).$$

Тоді з припущення $y(x_0) > 0$ і умови $B(x_0, \xi) > 0$ випливає

$$y(\xi) = y(x_0) \quad \forall \xi \in \sigma'(x_0). \quad (7.99)$$

Далі, оскільки $y(x) \neq \text{const}$ на $\bar{\omega}$, знайдеться точка $x'_0 \in \bar{\omega}$, в якій $y(x'_0) < y(x_0)$. З припущення про зв'язність сітки $\bar{\omega}$ випливає існування системи вузлів x_1, x_2, \dots, x_m , які належать ω і задовольняють умови

$$x_1 \in \sigma'(x_0), \quad x_2 \in \sigma'(x_1), \dots, \quad x_m \in \sigma'(x_{m-1}), \quad x'_0 \in \sigma'(x_m).$$

З умови (7.97) і властивості (7.99) одержуємо $y(x_1) = y(x_0)$. Отже, для точки x_1 можна також довести, що

$$y(\xi) = y(x_1) \quad \forall \xi \in \sigma'(x_1).$$

Аналогічно доведемо, що

$$y(x_1) = y(x_2) = \dots = y(x_m) = y(x_0).$$

Оцінимо величину

$$Ly(x_m) = D(x_m)y(x_m) + \sum_{\xi \in \sigma'(x_m)} B(x_m, \xi)(y(x_m) - y(\xi)).$$

З умов (7.91), (7.92), рівності $y(x_m) = y(x_0)$ і припущення (7.97) одержуємо нерівність

$$Ly(x_m) \geq B(x_m, x'_0)(y(x_0) - y(x'_0)) > 0,$$

яка суперечить припущенню $Ly(x) \leq 0$. Якщо $Ly(x) \geq 0$ для всіх $x \in \omega$, то для доведення теореми досить замінити y на $-y$. ■

➔ **Наслідок 7.1.** Нехай ω — зв'язна сітка, виконані умови (7.91), (7.92), сіткова функція $y(x)$, задана на $\bar{\omega}$, невід'ємна на границі і $Ly(x) \geq 0$ на ω . Тоді $y(x)$ невід'ємна на $\bar{\omega}$. Якщо ж $y(x) \leq 0$ на γ , $Ly(x) \leq 0$ на ω , то $y(x) \leq 0$ на $\bar{\omega}$.

Доведення. Нехай $Ly(x) \geq 0$ на ω , $y(x) \geq 0$ на γ . Припустимо, що $y(x_0) < 0$ хоча б в одному внутрішньому вузлі $x_0 \in \omega$. Тоді $y(x)$ повинна приймати найменше від'ємне значення в середині ω , що неможливо за теоремою 7.5, тому що $y(x) \neq \text{const}$ на ω ($y(x_0) < 0$, $y|_\gamma \geq 0$). Аналогічно доводиться друга частина наслідку. ■

➔ **Наслідок 7.2.** Однорідне рівняння

$$Ly(x) = 0, \quad x \in \omega \tag{7.100}$$

з однорідною крайовою умовою

$$y(x) = 0, \quad x \in \gamma \tag{7.101}$$

має лише тривіальний розв'язок $y(x) \equiv 0$.

Доведення. Нехай існує розв'язок задачі (7.100), (7.101), $y(x) \neq 0$. Якщо $y(x) \neq 0$ хоча б в одній точці, то згідно з наслідком 7.1 на $\bar{\omega}$ повинні одночасно виконуватися нерівності $y(x) \geq 0$ і $y(x) \leq 0$, що можливе лише при $y(x) \equiv 0$. ■

Оскільки система (7.100), (7.101) — це система лінійних алгебраїчних рівнянь, у якій кількість рівнянь дорівнює кількості невідомих, то з наслідку 7.2 випливає таке твердження.

➔ **Наслідок 7.3.** Існує єдиний розв'язок задачі (7.90) — (7.93).

■ **ТЕОРЕМА 7.6. (ПОРІВНЯННЯ)** Нехай $y(x)$ — розв'язок задачі (7.90) — (7.93), а $Y(x)$ — розв'язок задачі

$$LY(x) = \bar{F}(x), \quad x \in \omega,$$

$$Y(x) = \bar{\mu}(x), \quad x \in \gamma.$$

Тоді, якщо

$$|F(x)| \leq \bar{F}(x), \quad x \in \omega,$$

$$|\mu(x)| \leq \bar{\mu}(x), \quad x \in \gamma,$$

то

$$|y(x)| \leq Y(x), \quad x \in \bar{\omega}.$$

Доведення. З наслідку 7.1 випливає $Y(x) \geq 0$, $x \in \bar{\omega}$. Функції $u(x) = Y(x) + y(x)$ і $v(x) = Y(x) - y(x)$ задовольняють рівняння

$$Lu = \bar{F}(x) + F(x) \geq 0, \quad Lv = \bar{F}(x) - F(x) \geq 0$$

та граничні умови

$$u|_{\gamma} = (Y + y)|_{\gamma} = \bar{\mu} + \mu \geq 0, \quad v|_{\gamma} = (Y - y)|_{\gamma} = \bar{\mu} - \mu \geq 0.$$

Оскільки умови наслідку 7.1 виконані, то $u(x) \geq 0$ або $y(x) \geq -Y(x)$, $v(x) \geq 0$ або $y(x) \leq Y(x)$. Звідси випливає, що $-Y(x) \leq y(x) \leq Y(x)$ або $|y(x)| \leq Y(x)$ на $\bar{\omega}$.

Функція $Y(x)$ називається *мажорантою* для розв'язку задачі (7.90) — (7.93). Якщо мажоранта відома, то з нерівності $|y(x)| \leq Y(x)$ отримаємо оцінку для $\|y\|_C$. ■

➔ **Наслідок 7.4.** Для розв'язку задачі

$$Ly(x) = 0, \quad x \in \omega,$$

$$y(x) = \mu(x), \quad x \in \gamma,$$

справджується оцінка

$$\max_{x \in \bar{\omega}} |y(x)| = \|y\|_{C(\bar{\omega})} \leq \|\mu\|_{C(\gamma)},$$

де

$$\|\mu\|_{C(\gamma)} = \max_{x \in \gamma} |\mu(x)|.$$

Доведення. Розглянемо задачу

$$LY(x) = 0, \quad x \in \omega,$$

$$Y(x) = \|\mu\|_{C(\gamma)}, \quad x \in \gamma.$$

Функція $Y(x) \geq 0 \forall x \in \bar{\omega}$ і набуває найбільшого значення у деякому вузлі сітки $\bar{\omega}$. Якщо $Y(x) \neq \text{const}$, то цей вузол не може бути внутрішнім, а тому $\|Y\|_{C(\bar{\omega})} = \|Y\|_{C(\gamma)} = \|\mu\|_{C(\gamma)}$. Якщо $Y(x) = \text{const}$, то $Y(x) = \|\mu\|_{C(\gamma)}$. У цих випадках $\|Y\|_{C(\bar{\omega})} = \|\mu\|_{C(\gamma)}$. Звідси з урахуванням нерівності $\|y\|_{C(\bar{\omega})} \leq \|Y\|_{C(\bar{\omega})}$ випливає оцінка $\|y\|_{C(\bar{\omega})} \leq \|\mu\|_{C(\gamma)}$. ■

7.8.2. Стійкість та збіжність різницевої задачі Діріхле

Покажемо, що різницеву задачу (7.83), (7.84) можна записати у вигляді (7.90). Запишемо різницеву схему (7.83) у вигляді

$$\left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) y_{ij} = \frac{y_{i+1,j} + y_{i-1,j}}{h_1^2} + \frac{y_{i,j+1} + y_{i,j-1}}{h_2^2} + f_{ij}. \quad (7.102)$$

Позначимо через x точку x_{ij} — центральну точку п'ятиточкового шаблону, на якому апроксимується рівняння (7.83), а через $\sigma'(x)$ — решту вузлів шаблону, тобто $\sigma'(x) = \{x_{i+1,j}, x_{i-1,j}, x_{i,j+1}, x_{i,j-1}\}$. Тоді рівняння (7.102) можна записати у вигляді (7.90), де коефіцієнти $A(x)$, $B(x, \xi)$ і функція $F(x)$ визначається так:

$$\begin{aligned} A(x) &= \frac{2}{h_1^2} + \frac{2}{h_2^2}, & B(x, x_{i\pm 1,j}) &= \frac{1}{h_1^2}, \\ B(x, x_{i,j\pm 1}) &= \frac{1}{h_2^2}, & F(x) &= f(x_{ij}). \end{aligned} \quad (7.103)$$

Відмітимо властивості цих коефіцієнтів:

$$A(x) > 0, \quad B(x, \xi) > 0, \quad A(x) = \sum_{\xi \in \sigma'(x)} B(x, \xi).$$

До рівняння (7.102) необхідно ще додати граничні умови (7.84). Різницеву задачу (7.83), (7.84) запишемо у вигляді

$$Ly(x) = F(x), \quad x \in \omega, \quad (7.104)$$

$$y(x) = \mu(x), \quad x \in \gamma, \quad (7.105)$$

де

$$L(x)y(x) = A(x)y(x) - \sum_{\xi \in \sigma'(x)} B(x, \xi)y(\xi),$$

а коефіцієнти $A(x)$, $B(x, \xi)$, $F(x)$, визначаються формулами (7.103).

Розв'язок задачі (7.104), (7.105) можна записати у вигляді $y(x) = \tilde{y}(x) + \bar{y}(x)$, де $\tilde{y}(x)$ — розв'язок однорідного рівняння з неоднорідною граничною умовою:

$$L\tilde{y}(x) = 0, \quad x \in \omega, \quad \tilde{y}(x) = \mu(x), \quad x \in \gamma \quad (7.106)$$

і $\bar{y}(x)$ — розв'язок неоднорідного рівняння з однорідною граничною умовою:

$$L\bar{y}(x) = F(x), \quad x \in \omega, \quad \bar{y}(x) = 0, \quad x \in \gamma. \quad (7.107)$$

Зауважимо, що для задачі (7.104), (7.105) виконуються всі умови принципу максимуму. Застосуємо наслідок 7.4 до задачі (7.106), тоді отримаємо оцінку

$$\|\tilde{y}\|_{C(\bar{\omega})} \leq \|\mu\|_{C(\gamma)}. \quad (7.108)$$

Для розв'язку задачі (7.107) побудуємо мажоранту і застосуємо теорему порівняння. Розглянемо функцію

$$Y(x) = K(l_1^2 + l_2^2 - x_1^2 - x_2^2), \quad (7.109)$$

де K — довільна додатна стала, а l_1 , l_2 — довжини сторін прямокутника Ω . Зауважимо, що $Y(x) \geq 0$ за побудовою і

$$\begin{aligned} LY(x) &= -Y_{\bar{x}_1 x_1} - Y_{\bar{x}_2 x_2} = K \left((x_1^2)_{\bar{x}_1 x_1} + (x_2^2)_{\bar{x}_2 x_2} \right) = \\ &= K \left[\frac{1}{h_1^2} ((x_1 + h_1)^2 - 2x_1^2 + (x_1 - h_1)^2) + \right. \\ &\quad \left. + \frac{1}{h_2^2} ((x_2 + h_2)^2 - 2x_2^2 + (x_2 - h_2)^2) \right] = 4K. \end{aligned}$$

Отже, функція $Y(x)$ є розв'язком крайової задачі

$$LY(x) = \bar{F}(x), \quad x \in \omega, \quad Y(x) = \bar{\mu}(x), \quad x \in \gamma, \quad (7.110)$$

де $\bar{F}(x) = 4K$ і $\bar{\mu}(x) \geq 0$. Якщо

$$K = \frac{1}{4} \|F\|_{C(\omega)},$$

то для задач (7.107), (7.110) будуть виконуватись умови теореми порівняння. З цієї теореми випливає оцінка

$$\|\bar{y}\|_{C(\bar{\omega})} \leq \max_{x \in \bar{\omega}} Y(x) \leq K(l_1^2 + l_2^2).$$

Тоді

$$\|\bar{y}\|_{C(\bar{\omega})} \leq \frac{l_1^2 + l_2^2}{4} \|F\|_{C(\omega)}. \quad (7.111)$$

З нерівності трикутника і оцінок (7.108), (7.111) випливає оцінка для розв'язку задачі (7.83), (7.84)

$$\|y\|_{C(\bar{\omega})} \leq \|\tilde{y}\|_{C(\bar{\omega})} + \|\bar{y}\|_{C(\bar{\omega})} \leq \frac{l_1^2 + l_2^2}{4} \|f\|_{C(\omega)} + \|\mu\|_{C(\gamma)}. \quad (7.112)$$

Оскільки константи, які входять в оцінку (7.112), не залежать від кроків h_1 , h_2 , то дана оцінка гарантує стійкість різницевої схеми за правою частиною і за граничною умовою μ .

Отже, ми довели коректність (однозначну розв'язність і стійкість) різницевої схеми (7.83), (7.84). Перейдемо тепер до дослідження збіжності різницевої схеми.

Для розв'язку задачі (7.87), (7.88) справджується оцінка, аналогічна до (7.112), а саме

$$\|z\|_{C(\bar{\omega})} \leq \frac{l_1^2 + l_2^2}{4} \|\psi\|_{C(\omega)}.$$

Звідси і з (7.89) отримаємо оцінку

$$\|z\|_{C(\bar{\omega})} \leq M_2(h_1^2 + h_2^2),$$

де $M_2 = \frac{1}{4}M_1(l_1^2 + l_2^2)$ — стала, яка не залежить від h_1 і h_2 . Отже, схема (7.83), (7.84) збігається з другим порядком точності.

7.8.3. Монотонні різницеві схеми

Лінійний оператор L , який визначається формулою (7.94) називається *монотонним оператором*, якщо з умови $Ly(x) \geq 0 \forall x \in \omega$ випливає, що $y(x) \geq 0 \forall x \in \omega$. Різницеві схеми, які задовольняють $\forall x \in \omega$ умови (7.91), (7.92), називають *монотонними різницевими схемами*. Схеми, для яких умови (7.91), (7.92) не виконанні хоча б в одній точці $x \in \omega$, називаються *немонотонними*. Згідно з принципом максимуму,

якщо виконуються умови (7.91), (7.92) і $Ly(x) \geq 0 \quad \forall x \in \omega$, то $y(x)$ не може приймати найменшого від'ємного значення в усіх внутрішніх вузлах $x \in \omega$, тобто $y(x) \geq 0 \quad \forall x \in \omega$. Отже, умови (7.91), (7.92) забезпечують монотонність оператора L та коректність різницевої задачі (7.95) в сітковій нормі C :

$$\|y\|_{C(\omega)} = \max_{x \in \omega} |y(x)|.$$

Звідси не випливає, що немонотонна схема обов'язково некоректна. Виконання умов (7.91), (7.92) є достатньою умовою коректності.

Наведемо приклад монотонної різницевої схеми. Розглянемо схеми з ваговими коефіцієнтами для рівняння теплопровідності

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad 0 < x < l, \quad 0 < t \leq T, \\ u(x, 0) &= u_0(x), \quad 0 \leq x \leq l, \\ u(0, t) &= 0, \quad u(l, t) = 0, \quad 0 \leq t \leq T. \end{aligned}$$

Ця схема має вигляд

$$\begin{aligned} \frac{y_i^{n+1} - y_i^n}{\tau} &= \sigma y_{\bar{x}x, i}^{n+1} + (1 - \sigma) y_{\bar{x}x, i}^n + \varphi_i^n, \\ i &= \overline{1, N-1}, \quad n = \overline{0, K-1}, \end{aligned} \quad (7.113)$$

$$y_i^0 = u_0(x_i), \quad i = \overline{0, N}, \quad y_0^n = y_N^n = 0, \quad n = \overline{0, K}.$$

Знайдемо, при яких значеннях параметрів τ , h , σ схема (7.113) буде монотонною. Щоб записати схему (7.113) у вигляді (7.90) розв'яжемо її відносно y_i^{n+1} . Тоді одержимо

$$\begin{aligned} y_i^{n+1} &= \sigma \gamma (y_{i+1}^{n+1} - 2y_i^{n+1} + y_{i-1}^{n+1}) + \\ &+ (1 - \sigma) \gamma (y_{i+1}^n - 2y_i^n + y_{i-1}^n) + y_i^n + \tau \varphi_i^n, \quad \gamma = \frac{\tau}{h^2} \end{aligned}$$

або

$$\begin{aligned} (1 + 2\sigma\gamma) y_i^{n+1} &= (1 - 2(1 - \sigma)\gamma) y_i^n + \sigma\gamma (y_{i+1}^{n+1} + y_{i-1}^{n+1}) + \\ &+ (1 - \sigma)\gamma (y_{i+1}^n + y_{i-1}^n) + \tau\varphi_i^n. \end{aligned} \quad (7.114)$$

Звідси видно, що окіл точки $x = (x_i, t_{n+1})$ складається з п'яти точок $(x_{i\pm 1}, t_{n+1})$, (x_i, t_n) , $(x_{i\pm 1}, t_n)$. Умови додатності коефіцієнтів (7.91),

(7.92) зводяться до нерівностей $0 < \sigma < 1$, $\sigma > 1 - 1/(2\gamma)$. Зауважимо, що схема залишається монотонною і в тому випадку, якщо ці нерівності замінити на нестрогі, тобто вимагати

$$0 \leq \sigma \leq 1, \quad \sigma \geq 1 - \frac{1}{2\gamma}. \quad (7.115)$$

Дійсно, виконання однієї з умов (7.115) зі знаком рівності означає лише, що окіл $\sigma'(x)$ складається не з п'яти, а з меншого числа вузлів. Наприклад, при $\sigma = 0$ (явна схема) окіл $\sigma'(x)$ складається з точок (x_i, t_n) , $(x_{i\pm 1}, t_n)$ і умова монотонності (7.115) набуває вигляду

$$\frac{\tau}{h^2} \leq \frac{1}{2}.$$

Якщо, $\sigma = 0$, $\tau/h^2 = 0,5$, то дві з трьох нерівностей (7.115) виконані зі знаком рівності. В цьому випадку треба вважати, що окіл $\sigma'(x)$ складається з двох вузлів $(x_{i\pm 1}, t_n)$.

Отже, схема з вагами (7.113) є монотонною за умов (7.115), а чисто неявна схема ($\sigma = 1$) монотонна за $\forall \tau, h$. Схема Кранка–Нікольсона ($\sigma = 1/2$) монотонна за умов $\tau \leq h^2$. Позначимо

$$\|y^n\|_{C(\omega_h)} = \max_{1 \leq i \leq N-1} |y_i^n|.$$

Тоді з урахуванням невід'ємності коефіцієнтів рівняння (7.114), одержимо

$$\begin{aligned} (1 + 2\sigma\gamma) |y_i^{n+1}| &\leq (1 - 2(1 - \sigma)\gamma) |y_i^n| + \sigma\gamma (|y_{i+1}^{n+1}| + |y_{i-1}^{n+1}|) + \\ &\quad + (1 - \sigma)\gamma (|y_{i+1}^n| + |y_{i-1}^n|) + \tau |\varphi_i^n| \leq \\ &\leq \|y^n\|_{C(\omega_h)} + 2\sigma\gamma \|y^{n+1}\|_{C(\omega_h)} + \tau \|\varphi^n\|_{C(\omega_h)}, \end{aligned}$$

а, отже

$$\begin{aligned} (1 + 2\sigma\gamma) \|y^{n+1}\|_{C(\omega_h)} &\leq \|y^n\|_{C(\omega_h)} + \\ &\quad + 2\sigma\gamma \|y^{n+1}\|_{C(\omega_h)} + \tau \|\varphi^n\|_{C(\omega_h)}. \end{aligned}$$

Звідси

$$\begin{aligned} \|y^{n+1}\|_{C(\omega_h)} &\leq \|y^n\|_{C(\omega_h)} + \tau \|\varphi^n\|_{C(\omega_h)} \leq \\ &\leq \dots \leq \|y^0\|_{C(\omega_h)} + \sum_{j=0}^n \tau \|\varphi^j\|_{C(\omega_h)} \leq \\ &\leq \|y^0\|_{C(\omega_h)} + M_2 \max_{0 \leq j \leq n} \|\varphi^j\|_{C(\omega_h)}. \end{aligned}$$

Тобто схема з ваговими коефіцієнтами (7.113) за умов (7.115) стійка в просторі $C(\omega_h)$. Відомо, що достатньою умовою стійкості схеми (7.113) в просторі з середньоквадратичною нормою є умова $\sigma \geq 1/2 - 1/(4\gamma)$. Умова (7.115) є сильнішою вимогою.

7.9. Метод скінченних елементів розв'язування задачі Діріхле для рівняння Пуассона

В області $\bar{\Omega} = \{0 \leq x_1 \leq l_1, 0 \leq x_2 \leq l_2\}$ розглянемо задачу

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x), \quad x = (x_1, x_2) \in \Omega, \quad (7.116)$$

$$u = 0, \quad x \in \Gamma, \quad (7.117)$$

де Γ — границя області $\bar{\Omega}$. Задачу (7.116), (7.117) будемо розглядати в гільбертовому просторі $H = L_2(\Omega)$.

Нехай

$$Au = -\Delta u$$

з областю визначення оператора $D(A) = \{u : u \in W_2^2(\Omega), u = 0 \text{ на } \Gamma\}$. Задачу (7.116), (7.117) можна записати у вигляді

$$Au = f, \quad f \in L_2(\Omega). \quad (7.118)$$

Покажемо, що оператор A самоспряжений:

$$\begin{aligned} (Au, v) &= - \iint_{\Omega} \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} \right) v(x_1, x_2) dx_1 dx_2 = \\ &= - \int_0^{l_2} dx_2 \int_0^{l_1} \frac{\partial^2 u}{\partial x_1^2} v(x_1, x_2) dx_1 - \\ &\quad - \int_0^{l_1} dx_1 \int_0^{l_2} \frac{\partial^2 u}{\partial x_2^2} v(x_1, x_2) dx_2 = \\ &= - \int_0^{l_2} \left(\frac{\partial u}{\partial x_1} v(x_1, x_2) \Big|_0^{l_1} - \int_0^{l_1} \frac{\partial u}{\partial x_1} \frac{\partial v}{\partial x_1} dx_1 \right) dx_2 - \end{aligned}$$

$$\begin{aligned}
& - \int_0^{l_1} \left(\frac{\partial u}{\partial x_2} v(x_1, x_2) \Big|_0^{l_2} - \int_0^{l_2} \frac{\partial u}{\partial x_2} \frac{\partial v}{\partial x_2} dx_2 \right) dx_1 = \\
& = \iint_{\Omega} \left(\frac{\partial u}{\partial x_1} \frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2} \frac{\partial v}{\partial x_2} \right) dx_1 dx_2 = (Av, u), \\
& u, v \in D(A).
\end{aligned}$$

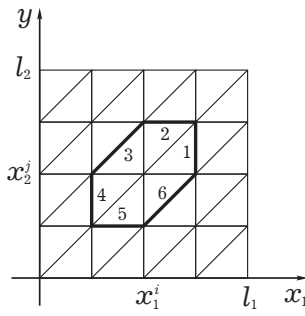


Рис. 7.5.

Для розв'язування задачі (7.118) застосуємо метод Гальоркіна. Введемо рівномірну сітку

$$\begin{aligned}
\bar{\omega}_h = \left\{ x_{ij} = (x_1^i, x_2^j), x_1^i = ih_1, x_2^j = jh_2, \right. \\
i = \overline{0, N_1}, j = \overline{0, N_2}, h_1 = l_1/N_1, \\
\left. h_2 = l_2/N_2 \right\}.
\end{aligned}$$

розбивши область Ω на прямокутники. Діагоналю, яка з'єднує точки (x_1^i, x_2^j) і (x_1^{i+1}, x_2^{j+1}) розділимо кожен з прямокутників на два трикутники, тобто здійснимо *триангуляцію* області Ω . Об'єднання трикутників, які прилягають до вузла (x_1^i, x_2^j) , позначимо через Ω_{ij} , а кожен з трикутників через Ω_{ij}^m , $m = \overline{1, 6}$ (рис. 7.5).

Кожному вузлу сітки (x_1^i, x_2^j) поставимо у відповідність базисну функцію $\varphi_{ij}(x_1, x_2)$, рівну одиниці в заданому вузлі і нулю в усіх інших вузлах (x_1^k, x_2^l) , і лінійну в кожному трикутнику області Ω . Такі функції мають вигляд

$$\varphi_{ij}(x_1, x_2) = \begin{cases} 1 - \frac{x_1 - x_1^i}{h_1}, & x_1, x_2 \in \Omega_{ij}^1, \\ 1 - \frac{x_2 - x_2^j}{h_2}, & x_1, x_2 \in \Omega_{ij}^2, \\ 1 + \frac{x_1 - x_1^i}{h_1} - \frac{x_2 - x_2^j}{h_2}, & x_1, x_2 \in \Omega_{ij}^3, \\ 1 + \frac{x_1 - x_1^i}{h_1}, & x_1, x_2 \in \Omega_{ij}^4, \\ 1 + \frac{x_2 - x_2^j}{h_2}, & x_1, x_2 \in \Omega_{ij}^5, \\ 1 - \frac{x_1 - x_1^i}{h_1} + \frac{x_2 - x_2^j}{h_2}, & x_1, x_2 \in \Omega_{ij}^6. \end{cases}$$

Функція $\varphi_{ij}(x_1, x_2)$, відмінна від нуля лише в області. Зауважимо, що $\varphi_{ij}(x_1, x_2) \in \overset{\circ}{W}_2^1(\Omega)$.

Множина лінійних комбінацій

$$u_h(x_1, x_2) = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} y_{ij} \varphi_{ij}(x_1, x_2) \quad (7.119)$$

утворює $(N_1 - 1)(N_2 - 1)$ -вимірний підпростір $\overset{\circ}{H}_h$ простору $\overset{\circ}{W}_2^1(\Omega)$ з базисом $\varphi_{ij}(x_1, x_2)$.

Наближений розв'язок $u_h(x_1, x_2)$ будемо шукати у вигляді (7.119), де y_{ij} знаходять (згідно з методом Гальоркіна) з системи лінійних алгебраїчних рівнянь

$$(Au_h, \varphi_{ij}) = (f, \varphi_{ij}), \quad i = \overline{1, N_1 - 1}, \quad j = \overline{1, N_2 - 1}. \quad (7.120)$$

Оскільки

$$u_h(x_1, x_2) = \begin{cases} y_{ij} + y_{x_1, ij}(x_1 - x_1^i) + y_{x_2, i+1, j}(x_2 - x_2^j), & (x_1, x_2) \in \Omega_{ij}^1, \\ y_{ij} + y_{x_1, i, j+1}(x_1 - x_1^i) + y_{x_2, ij}(x_2 - x_2^j), & (x_1, x_2) \in \Omega_{ij}^2, \\ y_{ij} + y_{\bar{x}_1, ij}(x_1 - x_1^i) + y_{x_2, ij}(x_2 - x_2^j), & (x_1, x_2) \in \Omega_{ij}^3, \\ y_{ij} + y_{\bar{x}_1, ij}(x_1 - x_1^i) + y_{\bar{x}_2, i-1, j}(x_2 - x_2^j), & (x_1, x_2) \in \Omega_{ij}^4, \\ y_{ij} + y_{\bar{x}_1, i, j-1}(x_1 - x_1^i) + y_{\bar{x}_2, ij}(x_2 - x_2^j), & (x_1, x_2) \in \Omega_{ij}^5, \\ y_{ij} + y_{x_1, ij}(x_1 - x_1^i) + y_{\bar{x}_2, ij}(x_2 - x_2^j), & (x_1, x_2) \in \Omega_{ij}^6, \end{cases}$$

то

$$\begin{aligned} (Au_h, \varphi_{ij}) &= \iint_{\Omega_{ij}} \left(\frac{\partial \varphi_{ij}}{\partial x_1} \frac{\partial u_h}{\partial x_1} + \frac{\partial \varphi_{ij}}{\partial x_2} \frac{\partial u_h}{\partial x_2} \right) dx_1 dx_2 = \\ &= - \iint_{\Omega_{ij}^1 \cup \Omega_{ij}^6} \frac{y_{x_1, ij}}{h_1} dx_1 dx_2 - \iint_{\Omega_{ij}^2 \cup \Omega_{ij}^3} \frac{y_{x_2, ij}}{h_2} dx_1 dx_2 + \\ &\quad + \iint_{\Omega_{ij}^3 \cup \Omega_{ij}^4} \frac{y_{\bar{x}_1, ij}}{h_1} dx_1 dx_2 + \iint_{\Omega_{ij}^5 \cup \Omega_{ij}^6} \frac{y_{\bar{x}_2, ij}}{h_2} dx_1 dx_2 = \\ &= -h_2 y_{x_1, ij} - h_1 y_{x_2, ij} + h_2 y_{\bar{x}_1, ij} + h_1 y_{\bar{x}_2, ij} = \\ &= -h_1 h_2 (y_{\bar{x}_1 x_1, ij} + y_{\bar{x}_2 x_2, ij}). \end{aligned}$$

Отже, система рівнянь (7.120) буде мати вигляд

$$y_{\bar{x}_1 x_1, ij} + y_{\bar{x}_2 x_2, ij} = -\frac{1}{h_1 h_2} \iint_{\Omega_{ij}} f(x_1, x_2) \varphi_{ij}(x_1, x_2) dx_1 dx_2,$$

$$i = \overline{1, N_1 - 1}, \quad j = \overline{1, N_2 - 1},$$

де

$$y_{i0} = y_{iN_2} = y_{0j} = y_{N_1 j} = 0.$$

Аналогічно як в одновимірному випадку можна показати, що наближений розв'язок $u_h(x_1, x_2)$ збігається до точного $u(x_1, x_2)$ при $|h| \rightarrow 0$, причому справджуються оцінки

$$\|u - u_h\|_{W_2^1(\Omega)} \leq c|h| \cdot \|u\|_{W_2^2(\Omega)} \leq c|h| \cdot \|f\|_{L_2(\Omega)},$$

$$\|u - u_h\|_{L_2(\Omega)} \leq c|h|^2 \|f\|_{L_2(\Omega)}, \quad |h| = \max(h_1, h_2).$$

7.10. Методи розв'язування сіткових рівнянь

При розв'язуванні крайових задач для еліптичних рівнянь методом сіток одержуємо систему лінійних алгебраїчних рівнянь

$$Ay = f.$$

Матриця A цієї системи має порядок, який дорівнює числу вузлів сітки. Крім того, матриця системи має багато нульових елементів, стрічкову структуру і, нарешті, є погано обумовленою. Метод Гаусса для розв'язування таких систем неефективний, оскільки кількість арифметичних операцій у ньому пропорційна кубу кількості невідомих.

Ці особливості еліптичних сіткових рівнянь вимагають спеціальних економних алгоритмів для їх чисельного розв'язування. Спеціальні прямі методи застосовують, як правило, для розв'язування вузького класу сіткових рівнянь. Ітераційні методи застосовують до більш загальних задач у випадку довільних областей, рівнянь загального вигляду зі змінними коефіцієнтами.

7.10.1. Запис сіткових схем розв'язування задачі Діріхле для рівняння Пуассона в операторному вигляді

Методи розв'язування двовимірних сіткових крайових задач будемо розглядати на прикладі задачі Діріхле для рівняння Пуассона

$$\begin{aligned} \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} &= -f(x), \quad x = (x_1, x_2) \in \Omega, \\ u(x) &= \mu(x), \quad x \in \Gamma \end{aligned} \quad (7.121)$$

в прямокутнику $\bar{\Omega} = \{0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2\}$ з границею Γ . Введемо в $\bar{\Omega}$ прямокутну сітку з кроками h_1 і h_2 :

$$\begin{aligned} \bar{\omega}_h &= \{x_{ij} = (x_1^i, x_2^j), x_1^i = ih_1, x_2^j = jh_2, \\ & i = \overline{0, N_1}, j = \overline{0, N_2}, h_1 = l_1/N_1, h_2 = l_2/N_2\}. \end{aligned}$$

Різницева задача Діріхле, яка відповідає задачі (7.121), має вигляд

$$\begin{aligned} y_{\bar{x}_1 x_1, ij} + y_{\bar{x}_2 x_2, ij} &= -f_{ij}, \quad x_{ij} \in \omega_h, \\ y_{ij} &= \mu(x_{ij}), \quad x_{ij} \in \gamma_h. \end{aligned} \quad (7.122)$$

Визначимо оператор A

$$(Ay)_{ij} = -\overset{\circ}{y}_{\bar{x}_1 x_1, ij} - \overset{\circ}{y}_{\bar{x}_2 x_2, ij}, \quad x_{ij} \in \omega_h, \quad (7.123)$$

де $y \in H$, $\overset{\circ}{y} \in \overset{\circ}{H}$, H — простір сіткових функцій, заданих на ω_h , $\overset{\circ}{H}$ — простір сіткових функцій, заданих на сітці $\bar{\omega}_h$ таких, що $\overset{\circ}{y}(x) = y(x) \forall x \in \omega_h$ і рівних нулю на границі γ_h сітки $\bar{\omega}_h$. Позначаючи

$$\begin{aligned} \varphi_{1j} &= f_{1j} + \frac{\mu(0, x_2^j)}{h_1^2}, \quad \varphi_{N_1-1, j} = f_{N_1-1, j} + \frac{\mu(l_1, x_2^j)}{h_1^2}, \\ j &= \overline{1, N_2-1}, \end{aligned} \quad (7.124)$$

$$\begin{aligned} \varphi_{i1} &= f_{i1} + \frac{\mu(x_1^i, 0)}{h_2^2}, \quad \varphi_{i, N_2-1} = f_{i, N_2-1} + \frac{\mu(x_1^i, l_2)}{h_2^2}, \\ i &= \overline{1, N_1-1}, \end{aligned} \quad (7.125)$$

$$\varphi_{ij} = f_{ij}, \quad i = \overline{2, N_1-2}, \quad j = \overline{2, N_2-2}, \quad (7.126)$$

запишемо різницеву схему (7.122) в операторному вигляді

$$Ay = \varphi, \quad y, \varphi \in H. \quad (7.127)$$

Введемо в $\overset{\circ}{H}$ скалярний добуток

$$(y, v) = \sum_{i=1}^{N_1-1} \sum_{j=1}^{N_2-1} \overset{\circ}{y}_{ij} \overset{\circ}{v}_{ij} h_1 h_2$$

і покажемо, що оператор A самоспряжений. Запишемо A у вигляді суми $A = A_1 + A_2$, де $(A_1 y)_{ij} = -\overset{\circ}{y}_{\bar{x}_1 x_1, ij}$, $(A_2 y)_{ij} = -\overset{\circ}{y}_{\bar{x}_2 x_2, ij}$. Покажемо, що кожен з операторів A_1 і A_2 є самоспряженим. Достатньо показати це для оператора A_1 . Розглянемо скалярний добуток

$$(A_1 y, v) = - \sum_{j=1}^{N_2-1} h_2 \left(\sum_{i=1}^{N_1-1} \overset{\circ}{y}_{\bar{x}_1 x_1, ij} \overset{\circ}{v}_{ij} h_1 \right). \quad (7.128)$$

Використаємо формулу сумування за частинами

$$\sum_{i=1}^{N_1-1} \overset{\circ}{y}_{\bar{x}_1 x_1, ij} \overset{\circ}{v}_{ij} h_1 = - \sum_{i=1}^{N_1} \overset{\circ}{y}_{\bar{x}_1, ij} \overset{\circ}{v}_{\bar{x}_1, ij} h_1.$$

Підставляючи цей вираз в (7.128), одержимо

$$(A_1 y, v) = \sum_{j=1}^{N_2-1} h_2 \left(\sum_{i=1}^{N_1} \overset{\circ}{y}_{\bar{x}_1, ij} \overset{\circ}{v}_{\bar{x}_1, ij} h_1 \right) = (y, A_1 v). \quad (7.129)$$

Аналогічно можна показати, що $A_2^* = A_2$, і, отже

$$\begin{aligned} (Ay, v) &= ((A_1 + A_2)y, v) = (A_1 y, v) + (A_2 y, v) = \\ &= (y, A_1 v) + (y, A_2 v) = (y, Av) \end{aligned}$$

тобто $A^* = A$.

З (7.129) випливає

$$(A_1 y, y) = \sum_{j=1}^{N_2-1} \sum_{i=1}^{N_1} \overset{\circ 2}{y}_{\bar{x}_1, ij} h_1 h_2 > 0,$$

аналогічно $(A_2 y, y) > 0$. Отже, A — самоспряжений і додатно визначений.

Знайдемо границі δ і Δ оператора A , тобто числа, для яких виконуються нерівності $\delta I \leq A \leq \Delta I$, де I — одиничний оператор. У розділі 7.3.4 показано, що

$$\delta_1 \sum_{i=1}^{N_1-1} \overset{\circ 2}{y}_{ij} h_1 \leq \sum_{i=1}^{N_1} \overset{\circ 2}{y}_{\bar{x}_1, ij} h_1 \leq \Delta_1 \sum_{i=1}^{N_1-1} \overset{\circ 2}{y}_{ij} h_1, \quad (7.130)$$

де

$$\delta_1 = \frac{4}{h_1^2} \sin^2 \frac{\pi h_1}{2l_1}, \quad \Delta_1 = \frac{4}{h_1^2} \cos^2 \frac{\pi h_1}{2l_1}.$$

Помножимо нерівності (7.130) на h_2 та підсумуємо по $j = \overline{1, N_2 - 1}$, тоді одержимо

$$\delta_1 \sum_{j=1}^{N_2-1} \sum_{i=1}^{N_1-1} \overset{\circ}{y}_{ij}^2 h_1 h_2 \leq \sum_{j=1}^{N_2-1} \sum_{i=1}^{N_1} \overset{\circ}{y}_{x_1, ij}^2 h_1 h_2 \leq \Delta_1 \sum_{j=1}^{N_2-1} \sum_{i=1}^{N_1-1} \overset{\circ}{y}_{ij}^2 h_1 h_2$$

або

$$\delta_1(y, y) \leq (A_1 y, y) \leq \Delta_1(y, y).$$

Аналогічно знаходимо

$$\delta_2(y, y) \leq (A_2 y, y) \leq \Delta_2(y, y),$$

де

$$\delta_2 = \frac{4}{h_2^2} \sin^2 \frac{\pi h_2}{2l_2}, \quad \Delta_2 = \frac{4}{h_2^2} \cos^2 \frac{\pi h_2}{2l_2}.$$

Звідси випливає

$$\delta \|y\|^2 \leq (Ay, y) \leq \Delta \|y\|^2,$$

де

$$\begin{aligned} \delta &= \delta_1 + \delta_2 = \frac{4}{h_1^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{\pi h_2}{2l_2}, \\ \Delta &= \Delta_1 + \Delta_2 = \frac{4}{h_1^2} \cos^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \cos^2 \frac{\pi h_2}{2l_2}. \end{aligned}$$

7.10.2. Швидке дискретне перетворення Фур'є

Нехай необхідно обчислити суми вигляду

$$s_j = \sum_{k=1}^{N-1} z_k \sin \frac{\pi k j}{N}, \quad j = \overline{1, N-1}, \quad (7.131)$$

де z_k — задані числа.

Для безпосереднього обчислення всіх s_j , $j = \overline{1, N-1}$ потрібно $(N-1)^2$ множень. Розглянемо метод обчислення сум вигляду (7.131), який вимагає $O(N \log_2 N)$ дій множення. Таке прискорене обчислення

сум ґрунтується на тому, що серед чисел $\sin \frac{\pi k j}{N}$, $k = \overline{1, N-1}$, $j = \overline{1, N-1}$, є багато однакових. Тому можна перегрупувати доданки, і таким чином зменшити кількість множень.

Перетворимо суму (7.131) до вигляду

$$s_j = \sum_{k=1}^{N-1} z_k \sin \frac{2\pi k j}{2N} = \sum_{k=0}^{2N-1} z_k \sin \frac{2\pi k j}{2N},$$

де будемо вважати, що $z_0 = z_N = z_{N+1} = \dots = z_{2N-1} = 0$. Позначимо $M = 2N$ і розглянемо більш загальну задачу обчислення суми

$$v_j = \sum_{k=0}^{M-1} z_k e^{i \frac{2\pi k j}{M}} = \sum_{k=0}^{M-1} z_k w^{kj},$$

де

$$w = e^{i \frac{2\pi}{M}} \quad (7.132)$$

і i — уявна одиниця. Отже, будемо обчислювати суми

$$v_j = \sum_{k=0}^{M-1} z_k w^{kj}, \quad j = \overline{0, M-1}. \quad (7.133)$$

Зауважимо, що $\text{Im } v_j = s_j$ для дійсних z_k . Надалі суттєвою є умова $M = 2^m$, $m > 0$, яка означає, що число точок сітки $N = 2^{m-1}$ є степенем двійки.

Зобразимо число k у формулі (7.133) в двійковій системі

$$k = k_0 + 2k_1 + 2^2k_2 + \dots + 2^{m-1}k_{m-1}, \quad (7.134)$$

де k_i — або 0, або 1. Позначимо

$$z_k = z(k_0, k_1, \dots, k_{m-1}).$$

Тоді суму (7.133) можна записати у вигляді

$$\begin{aligned} v_j &= \sum_{k_0, k_1, \dots, k_{m-1}} z(k_0, k_1, \dots, k_{m-1}) w^{(k_0 + 2k_1 + \dots + 2^{m-1}k_{m-1})j} = \\ &= \sum_{k_0=0}^1 w^{k_0 j} \left[\sum_{k_1=0}^1 w^{2k_1 j} \dots \times \right. \\ &\quad \left. \times \dots \sum_{k_{m-1}=0}^1 w^{2^{m-1}k_{m-1} j} z(k_0, k_1, \dots, k_{m-1}) \right]. \end{aligned} \quad (7.135)$$

Перетворимо останню суму в виразі (7.135)

$$\sum_{k_{m-1}=0}^1 w^{2^{m-1}k_{m-1}j} z(k_0, k_1, \dots, k_{m-1}). \quad (7.136)$$

Зобразимо число j в двійковій системі

$$j = j_0 + 2j_1 + \dots + 2^{m-1}j_{m-1}, \quad (7.137)$$

де j_i — або 0, або 1. Тоді отримаємо

$$\begin{aligned} w^{2^{m-1}k_{m-1}j} &= \\ &= \left(w^{2^{m-1}k_{m-1}j_0} \right) \left(w^{2^m k_{m-1}j_1} \right) \dots \left(w^{2^m k_{m-1}j_{m-1} 2^{m-2}} \right). \end{aligned}$$

В цьому добутку всі множники, починаючи з другого, дорівнюють 1. Дійсно, оскільки

$$w^{2^m k_{m-1}j_i 2^{i-1}} = (w^M)^{k_{m-1}j_i 2^{i-1}}, \quad i = \overline{1, m-1},$$

то з урахуванням (7.132) і того, що $k_{m-1}j_i$ дорівнює або 0, або 1, отримаємо $w^M = 1$, $w^{2^m k_{m-1}j_i 2^{i-1}} = 1$, $i = \overline{1, m-1}$. Отже, $w^{2^{m-1}k_{m-1}j} = w^{2^{m-1}k_{m-1}j_0}$. Позначимо суму (7.136) через $z_1(j_0, k_0, k_1, \dots, k_{m-2})$ і запишемо у вигляді

$$\begin{aligned} z_1(j_0, k_0, k_1, \dots, k_{m-2}) &= \\ &= \sum_{k_{m-1}=0}^1 w^{2^{m-1}k_{m-1}j_0} z(k_0, k_1, \dots, k_{m-1}). \end{aligned} \quad (7.138)$$

Передостанню суму в виразі (7.135) запишемо у вигляді

$$\sum_{k_{m-2}=0}^1 w^{2^{m-2}k_{m-2}j} z_1(j_0, k_0, k_1, \dots, k_{m-2}). \quad (7.139)$$

Зобразимо число $2^{m-2}k_{m-2}j$ у вигляді

$$2^{m-2}k_{m-2}(j_0 + 2j_1) + 2^m k_{m-2}(j_2 + \dots + 2^{m-3}j_{m-1}),$$

звідки отримаємо $w^{2^{m-2}k_{m-2}j} = w^{2^{m-2}k_{m-2}(j_0+2j_1)}$. Отже, сума (7.139) дорівнює

$$\begin{aligned} z_2(j_0, j_1, k_0, k_1, \dots, k_{m-3}) &= \\ &= \sum_{k_{m-2}=0}^1 w^{2^{m-2}k_{m-2}(j_0+2j_1)} z_1(j_0, k_0, k_1, \dots, k_{m-2}). \end{aligned} \quad (7.140)$$

Процес послідовного обчислення сум продовжується до тих пір, поки в (7.135) не вичерпаються всі суми. Оскільки, друга сума в (7.135) дорівнює

$$z_{m-1}(j_0, j_1, \dots, j_{m-2}, k_0) = \sum_{k_1=0}^1 w^{2k_1(j_0+2j_1+\dots+2^{m-2}j_{m-2})} z_{m-2}(j_0, j_1, \dots, j_{m-3}, k_0, k_1), \quad (7.141)$$

то

$$v_j = \sum_{k_0=0}^1 w^{k_0 j} z_{m-1}(j_0, j_1, \dots, j_{m-2}, k_0). \quad (7.142)$$

Отже, можна запропонувати такий алгоритм обчислення сум вигляду (7.133), який називається алгоритмом швидкого дискретного перетворення Фур'є. Числа k і j зображуються в двійковій системі за формулами (7.134), (7.137). Далі послідовно обчислюються суми (7.138), (7.140) — (7.142), що складаються з двох дододанків.

Підрахуємо кількість множень, необхідних для знаходження сум v_j , $j = \overline{0, M-1}$ при вказаному способі обчислень. Функція $z_1(j_0, k_0, k_1, \dots, k_{m-2})$ використовується тільки при обчисленні $z_2(j_0, j_1, k_0, k_1, \dots, k_{m-3})$. При цьому необхідно обчислити значення $z_1(j_0, k_0, k_1, \dots, k_{m-2})$ два рази, при $k_{m-2} = 0$ і $k_{m-2} = 1$. Обчислення $z_1(j_0, k_0, k_1, \dots, k_{m-2})$ при кожному значенні k_{m-2} вимагає двох множень. Отже, загальна кількість множень, потрібних для обчислення $z_1(j_0, k_0, k_1, \dots, k_{m-2})$ дорівнює чотирьом. Така ж кількість множень потрібна для обчислення кожної з сум $z_l(j_0, \dots, j_{l-1}, k_0, k_1, \dots, k_{m-l-1})$. Всього маємо m таких сум. Тому кількість множень, необхідних для обчислення v_j при кожному фіксованому j , дорівнює $4m$, а для обчислення всіх v_j , $j = \overline{0, M-1}$, це число дорівнює $4nM = 4M \log_2 M$. При великих $N = 2^{m-1}$ це призводить до значного зменшення кількості множень у порівнянні з кількістю множень $(N-1)^2$, потрібних для безпосереднього обчислення сум вигляду (7.131). Так, при $N = 128$ кількість множень буде майже в два рази меншою.

7.10.3. Прямі методи. Метод розділення змінних

Серед прямих методів розв'язування сіткових рівнянь виділимо *метод матричної прогонки*, *метод редукції (декомпозиції)* та *метод розділення змінних*. Метод матричної прогонки є узагальненням методу

прогонки (див. розділ 1.1.2) на випадок, коли x_i і f_i , які входять в (1.10), — вектори, а коефіцієнти a_i , b_i , c_i — матриці. Операцію ділення в алгоритмі прогонки слід замінити множенням на обернену матрицю зліва. Метод редукції — це модифікація методу Гаусса, в основі якого лежить спеціальний спосіб виключення невідомих (див., напр., [20]).

Розглянемо метод розділення змінних розв'язування різницевої задачі Діріхле для рівняння Пуассона в прямокутнику. Запишемо задачу (7.122) у вигляді

$$\begin{aligned} y_{\bar{x}_1 x_1, ij} + y_{\bar{x}_2 x_2, ij} &= -\varphi_{ij}, \quad x_{ij} \in \omega_h, \\ y_{ij} &= 0, \quad x_{ij} \in \gamma_h, \end{aligned} \quad (7.143)$$

де φ_{ij} визначається за формулами (7.124) — (7.126).

Розглянемо одновимірну задачу на власні значення

$$\begin{aligned} \frac{\mu(j+1) - 2\mu(j) + \mu(j-1))}{h_2^2} + \lambda \mu(j) &= 0, \\ j = \overline{1, N_2 - 1}, \quad \mu(0) = \mu(N_2) &= 0, \end{aligned} \quad (7.144)$$

де

$$h_2 N_2 = l_2, \quad \mu(j) = \mu(x_2^j).$$

У розділі 7.3.2 було показано, що задача (7.144) має такий розв'язок:

$$\begin{aligned} \mu(j) = \mu_k(j) &= \sqrt{\frac{2}{l_2}} \sin \frac{\pi k x_2^j}{l_2}, \\ \lambda_k &= \frac{4}{h_2^2} \sin^2 \frac{\pi k h_2}{2l_2}, \quad k = \overline{1, N_2 - 1}. \end{aligned}$$

Зафіксуємо деяке значення індексу $i = \overline{1, N_1 - 1}$ і будемо розглядати y_{ij} , φ_{ij} як функції, які залежать тільки від $j = \overline{1, N_2 - 1}$. Розкладемо y_{ij} , φ_{ij} за власними функціями задачі (7.144), тобто запишемо їх у вигляді

$$y_{ij} = \sum_{k=1}^{N_2-1} c_k(i) \mu_k(j), \quad (7.145)$$

$$\varphi_{ij} = \sum_{k=1}^{N_2-1} \hat{\varphi}_k(i) \mu_k(j), \quad (7.146)$$

де $c_k(i)$ — невідомі коефіцієнти,

$$\hat{\varphi}_k(i) = (\varphi_i, \mu_k) = \sum_{j=1}^{N_2-1} \varphi_{ij} \mu_k(j) h_2, \quad i = \overline{1, N_1 - 1}, \quad (7.147)$$

$$\varphi_i = (\varphi_{i1}, \varphi_{i2}, \dots, \varphi_{i, N_2-1}), \quad \mu_k = (\mu_k(1), \mu_k(2), \dots, \mu_k(N_2 - 1)).$$

Підставимо розклади (7.145), (7.146) у рівняння (7.143), тоді

$$\begin{aligned} \sum_{k=1}^{N_2-1} \mu_k(j) (c_k(i))_{\bar{x}_1 x_1, i} + \sum_{k=1}^{N_2-1} c_k(i) (\mu_k(j))_{\bar{x}_2 x_2, j} = \\ = - \sum_{k=1}^{N_2-1} \hat{\varphi}_k(i) \mu_k(j). \end{aligned}$$

Звідси на підставі (7.144)

$$\sum_{k=1}^{N_2-1} \left[(c_k(i))_{\bar{x}_1 x_1, i} - \lambda_k c_k(i) + \hat{\varphi}_k(i) \right] \cdot \mu_k(j) = 0.$$

Враховуючи лінійну незалежність μ_k , для обчислення коефіцієнтів c_k , $k = \overline{1, N_2 - 1}$, одержимо систему різницьових рівнянь

$$\begin{aligned} \frac{c_k(i+1) - 2c_k(i) + c_k(i-1))}{h_1^2} - \lambda_k c_k(i) + \hat{\varphi}_k(i) = 0, \\ i = \overline{1, N_1 - 1}, \quad c_k(0) = c_k(N_1) = 0. \end{aligned} \quad (7.148)$$

Алгоритм розв'язування задачі (7.143) такий. Спочатку обчислюють коефіцієнти Фур'є $\hat{\varphi}_k(i)$, $k = \overline{1, N_2 - 1}$. При кожному фіксованому i суми вигляду (7.147) можна обчислити для $k = \overline{1, N_2 - 1}$ за допомогою швидкого дискретного перетворення Фур'є за число дій $O(N_2 \log_2 N_2)$, а обчислення цих сум для всіх $i = \overline{1, N_1 - 1}$ вимагає $O(N_1 N_2 \log_2 N_2)$ дій. Потім методом прогонки $N_2 - 1$ разів розв'язують систему (7.148) для $k = \overline{1, N_2 - 1}$, що вимагає $O(N_1 N_2)$ дій. Знайшовши $c_k(i)$, обчислюють y_{ij} за формулами (7.145), які аналогічні до формул (7.147) і вимагають також кількості дій $O(N_1 N_2 \log_2 N_2)$. Отже, метод розділення змінних з використанням швидкого перетворення Фур'є для розв'язування задачі (7.143) потребує $O(N_1 N_2 \log_2 N_2)$ операцій. Для порівняння зазначимо, що звичайний метод виключення Гаусса вимагав би $O(N_1^3 N_2^3)$ дій і великої машинної пам'яті.

Недоліком цього методу є необхідність знаходження в явному вигляді власних чисел і власних векторів одновимірної дискретної задачі. У випадку, коли розв'язок задачі на власні значення в явному вигляді записати неможливо (наприклад, для крайових умов третього роду або у випадку змінних нероздільних коефіцієнтів), цей метод застосувати не можна.

7.10.4. Застосування ітераційних методів для розв'язування задачі Діріхле

Для розв'язування системи рівнянь (7.127) (або (7.122)), з оператором (7.123) і правою частиною (7.124) — (7.126) розглянемо двоярусні ітераційні методи (див. розділ 1.3), записані у канонічному вигляді

$$B \frac{y_{n+1} - y_n}{\tau_{n+1}} + Ay_n = \varphi. \quad (7.149)$$

Метод Якобі ($B = D = \text{diag} \{a_{11}, a_{22}, \dots, a_{MM}\}$, $\tau_{n+1} = 1$) для системи (7.122) записується у вигляді

$$y_{ij}^{n+1} = \frac{y_{i+1,j}^n + y_{i-1,j}^n + (h_1/h_2)^2 (y_{i,j+1}^n + y_{i,j-1}^n) + h_1^2 f_{ij}}{2 [1 + (h_1/h_2)^2]}, \quad (7.150)$$

$$x_{ij} \in \omega_h, \quad y_{ij}^n = \mu(x_{ij}), \quad x_{ij} \in \gamma_h.$$

Початкове наближення y_{ij}^0 — довільна сіткова функція, яка приймає на границі γ_h задані значення $y_{ij}^0 = \mu(x_{ij})$, $x_{ij} \in \gamma_h$. В даному випадку метод Якобі збігається з методом простої ітерації за оптимального значення ітераційного параметра. Дійсно метод простої ітерації

$$\frac{y_{n+1} - y_n}{\tau} + Ay_n = \varphi$$

для системи (7.122) у випадку $A^* = A > 0$ володіє найбільшою швидкістю збіжності, якщо $\tau = \tau_0 = 2/(\delta + \Delta)$, де δ, Δ — найбільше та найменше власні числа оператора (7.123), тобто

$$\delta = \frac{4}{h_1^2} \sin^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \sin^2 \frac{\pi h_2}{2l_2},$$

$$\Delta = \frac{4}{h_1^2} \cos^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \cos^2 \frac{\pi h_2}{2l_2}, \quad \tau_0 = \frac{h_1^2}{2 [1 + (h_1/h_2)^2]}.$$

При цьому значенні параметра метод простої ітерації має вигляд

$$y_{ij}^{n+1} - y_{ij}^n = \frac{2[1 + (h_1/h_2)^2]}{h_1^2} (y_{x_1x_1,ij}^n + y_{x_2x_2,ij}^n + f_{ij}), \quad x_{ij} \in \omega_h,$$

$$y_{ij}^n = \mu(x_{ij}), \quad x_{ij} \in \gamma_h.$$

Останні рівняння збігаються з (7.150).

Швидкість збіжності методу (7.150), як методу простої ітерації з оптимальним параметром визначається числом $\rho = \frac{1-\xi}{1+\xi}$, $\xi = \frac{\delta}{\Delta}$. Кількість ітерацій $n_0(\varepsilon)$, необхідних для досягнення заданої точності ε , дорівнює

$$n_0(\varepsilon) = \frac{\ln(1/\varepsilon)}{\ln(1/\rho)} = \frac{\ln(1/\varepsilon)}{\ln\left(1 + \frac{2\xi}{1-\xi}\right)}.$$

При $l_1 = l_2 = l$, $h_1 = h_2 = h = l/N \rightarrow 0$ маємо $\xi = \operatorname{tg}^2 \frac{\pi h}{2l} \approx \frac{\pi^2 h^2}{4l^2}$, $\ln \frac{1}{\rho} \approx 2\xi = \frac{\pi^2 h^2}{2l^2}$ так, що

$$n_0(\varepsilon) \approx \frac{2l^2 \ln(1/\varepsilon)}{\pi^2 h^2}.$$

Отже, метод простої ітерації вимагає $O(h^{-2})$ ітерацій для досягнення заданої точності.

Розглянемо метод Зейделя розв'язування системи (7.122):

$$\frac{y_{i-1,j}^{n+1} - 2y_{ij}^{n+1} + y_{i+1,j}^n}{h_1^2} + \frac{y_{i,j-1}^{n+1} - 2y_{ij}^{n+1} + y_{i,j+1}^n}{h_2^2} = -f_{ij},$$

$$x_{ij} \in \omega_h, \quad y_{ij}^{n+1} = \mu(x_{ij}), \quad x_{ij} \in \gamma_h.$$

Реалізація методу Зейделя зводиться до такого ітераційного процесу:

$$\left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right) y_{ij}^{n+1} + \frac{1}{h_1^2} y_{i-1,j}^{n+1} + \frac{1}{h_2^2} y_{i,j-1}^{n+1} +$$

$$+ \frac{1}{h_1^2} y_{i+1,j}^n + \frac{1}{h_2^2} y_{i,j+1}^n +$$

$$+ f_{ij}, \quad x_{ij} \in \omega_h, \quad y_{ij}^{n+1} = \mu(x_{ij}), \quad x_{ij} \in \gamma_h. \quad (7.151)$$

Вкажемо послідовність проведення обчислень ітераційним методом (7.151). Спочатку, використовуючи відомі граничні значення $y_{01}^{n+1} = \mu(x_{01})$ і $y_{10}^{n+1} = \mu(x_{10})$, знаходять y_{11}^{n+1} за формулою (7.151). Якщо

y_{11}^{n+1} відоме, то можна знайти y_{12}^{n+1} і т.д. Отже, невідомі y_{ij}^{n+1} обчислюють у порядку зміни індексів: $(1, 1), (1, 2), \dots, (1, N_2 - 1), (2, 1), (2, 2), \dots, (2, N_2 - 1), \dots, (N_1 - 1, 1), (N_1 - 1, 2), \dots, (N_1 - 1, N_2 - 1)$.

Метод Зейделя збігається дещо швидше, ніж метод простої ітерації, однак число ітерацій, необхідних для досягнення заданої точності, при $h_1 = h_2 = h$, як і у методі Якобі є величиною порядку $O(h^{-2})$.

Метод верхньої релаксації визначається рівняннями:

$$\begin{aligned} & \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y_{ij}^{n+1} + (1 - \omega) \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y_{ij}^n + \\ & + \omega \left[\frac{1}{h_1^2} y_{i-1,j}^{n+1} + \frac{1}{h_2^2} y_{i,j-1}^{n+1} + \frac{1}{h_1^2} y_{i+1,j}^n + \frac{1}{h_2^2} y_{i,j+1}^n + f_{ij} \right], \\ & x_{ij} \in \omega_h, \quad y^{n+1}(x_{ij}) = \mu(x_{ij}), \quad x_{ij} \in \gamma_h. \end{aligned}$$

Спосіб знаходження y_{ij}^{n+1} на новій ітерації такий самий, як і у методі Зейделя. Оптимальне значення параметра ω знаходять за формулою (див.[20, С. 382])

$$\omega = \omega_0 = \frac{2}{1 + \sqrt{\lambda_{\min}(2 - \lambda_{\min})}},$$

де

$$\lambda_{\min} = \frac{2h_2^2}{h_1^2 + h_2^2} \sin \frac{\pi h_1}{2l_1} + \frac{2h_1^2}{h_1^2 + h_2^2} \sin \frac{\pi h_2}{2l_2}.$$

Можна показати, що при $l_1 = l_2 = l$, $h_1 = h_2 = h = l/N \rightarrow 0$ необхідне число ітерацій дорівнює:

$$n_0(\varepsilon) \approx \frac{2l \ln(1/\varepsilon)}{\pi h} = O(h^{-1}).$$

У випадку явної схеми (7.149) ($B = I$) з чебишевським набором параметрів

$$\begin{aligned} \tau_n &= \frac{\tau_0}{1 + \rho_0 t_n}, \quad \tau_0 = \frac{2}{\delta + \Delta}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \\ \xi &= \frac{\delta}{\Delta}, \quad t_n = \cos \frac{(2n-1)\pi}{2m}, \quad n = \overline{1, m}. \end{aligned} \quad (7.152)$$

Якщо вибрати τ_n , згідно (7.152), то для похибки буде справджуватися оцінка (див. розділ 1.3.6):

$$\|y_n - y\| \leq q_n \|y_0 - y\|,$$

де

$$q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\delta}{\Delta}.$$

Систему (7.122) можна розв'язати за допомогою чебишевського ітераційного методу. Обчислення

$$y_{n+1} = (y_{11}^{n+1}, y_{21}^{n+1}, \dots, y_{N_1-1, N_2-1}^{n+1})$$

доцільно організувати так: спочатку за відомими наближеннями y_{ij}^n знаходиться нев'язка

$$\begin{aligned} r_{ij}^n &= (Ay_n)_{ij} - f_{ij} - \\ &= - \left(\frac{y_{i-1,j}^n - 2y_{ij}^n + y_{i+1,j}^n}{h_1^2} + \frac{y_{i,j-1}^n - 2y_{ij}^n + y_{i,j+1}^n}{h_2^2} + f_{ij} \right), \\ i &= \overline{1, N_1 - 1}, j = \overline{1, N_2 - 1}, \end{aligned}$$

а потім обчислюються значення y_{ij}^{n+1} за формулою

$$y_{ij}^{n+1} = y_{ij}^n - \tau_{n+1} r_{ij}^n, \quad i = \overline{1, N_1 - 1}, \quad j = \overline{1, N_2 - 1},$$

при цьому покладемо

$$y_{i0}^{n+1} = \mu(x_1^i, 0), \quad y_{i, N_2}^{n+1} = \mu(x_1^i, l_2), \quad x_1^i = ih_1, \quad i = \overline{1, N_1 - 1},$$

$$y_{0,j}^{n+1} = \mu(0, x_2^j), \quad y_{N_1,j}^{n+1} = \mu(l_1, x_2^j), \quad x_2^j = jh_2, \quad j = \overline{1, N_2 - 1}.$$

Кількість ітерацій, необхідних для досягнення заданої точності ε визначається за формулою:

$$n \geq n_0(\varepsilon), \quad n_0(\varepsilon) = \frac{\ln(2/\varepsilon)}{\ln(1/\rho_1)}, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}.$$

Зокрема для $l_1 = l_2 = l$, $h_1 = h_2 = h = l/N$, і малих h маємо

$$\xi = \frac{\pi^2 h^2}{4l^2}, \quad \ln \frac{1}{\rho_1} \approx 2\sqrt{\xi} = \frac{\pi h}{l}, \quad n_0(\varepsilon) \approx \frac{l \ln(2/\varepsilon)}{\pi h}.$$

Отже, число ітерацій $n_0(\varepsilon)$, необхідних для отримання заданої точності ε , є величиною $O(h^{-1})$.

7.11. Чисельне розв'язування багатовимірних задач теплопровідності

7.11.1. Різницеві схеми для багатовимірних рівнянь теплопровідності

Розглянемо двовимірне рівняння теплопровідності в прямокутнику $\bar{\Omega} = \{0 \leq x_1 \leq l_1, 0 \leq x_2 \leq l_2\}$ з границею Γ

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + f(x, t), \quad x = (x_1, x_2) \in \Omega, \quad t \in (0, T], \quad (7.153)$$

розв'язок якого задовольняє початкові та крайові умови

$$u(x, 0) = u_0(x), \quad x \in \bar{\Omega}, \quad (7.154)$$

$$u(x, t) = \mu(x, t), \quad x \in \Gamma, \quad t \in [0, T]. \quad (7.155)$$

Введемо сітку

$$\omega_\tau = \{t_n = n\tau, \quad n = \overline{0, K}, \quad K\tau = T\}$$

і просторову сітку

$$\begin{aligned} \bar{\omega}_h &= \{x_{ij} = (x_1^i, x_2^j), \quad x_1^i = ih_1, \quad x_2^j = jh_2, \\ &\quad i = \overline{0, N_1}, \quad j = \overline{0, N_2}, \quad h_1 = l_1/N_1, \quad h_2 = l_2/N_2\} \end{aligned}$$

з границею γ_h .

Будемо позначати

$$y_{ij}^n = y(x_{ij}, t_n) = y(ih_1, jh_2, t_n), \quad x_{ij} = (ih_1, jh_2) \in \omega_h.$$

Розглянемо схему з параметром σ

$$\begin{aligned} \frac{y_{ij}^{n+1} - y_{ij}^n}{\tau} &= \Lambda(\sigma y_{ij}^{n+1} + (1 - \sigma)y_{ij}^n) + \varphi_{ij}^n, \\ x_{ij} &\in \omega_h, \quad t_n \in \omega_\tau, \end{aligned} \quad (7.156)$$

$$y_{ij}^0 = u_0(x_{ij}), \quad x_{ij} \in \bar{\omega}_h, \quad (7.157)$$

$$y_{ij}^n = \mu(x_{ij}, t_n), \quad x_{ij} \in \gamma_h, \quad t_n \in \omega_\tau, \quad (7.158)$$

де

$$\Lambda y_{ij} = \Lambda_1 y_{ij} + \Lambda_2 y_{ij}, \quad \Lambda_1 y_{ij} = y_{\bar{x}_1 x_1, ij} = \frac{y_{i+1, j} - 2y_{ij} + y_{i-1, j}}{h_1^2},$$

$$\Lambda_2 y_{ij} = y_{\bar{x}_2 x_2, ij} = \frac{y_{i,j+1} - 2y_{ij} + y_{i,j-1}}{h_2^2}, \quad \varphi_{ij}^n = f(x_{ij}, t_{n+1/2}).$$

Аналогічно до одновимірного випадку можна показати, що при $\sigma = 0,5$ схема має порядок апроксимації $O(\tau^2 + |h|^2)$, де $|h|^2 = h_1^2 + h_2^2$, а при $\sigma \neq 0,5$ схема має порядок апроксимації $O(\tau + |h|^2)$. Можна також показати, що при

$$\sigma \geq \frac{1}{2} - \frac{1}{4\tau} \left(\frac{1}{h_1^2} + \frac{1}{h_2^2} \right)^{-1}$$

така схема є стійкою. Зокрема, явна схема ($\sigma = 0$) умовно стійка, тобто стійка за умови

$$\tau \leq \frac{1}{2} \left(\frac{1}{h_1^2} + \frac{1}{h_2^2} \right)^{-1}.$$

Для визначення y_{ij}^{n+1} на сітці ω_h за допомогою явної схеми необхідно затратити число дій, пропорційне числу вузлів сітки ω_h .

Якщо $\sigma \neq 0$, то для знаходження y_{ij}^{n+1} необхідно розв'язати систему $(N_1 - 1) \times (N_2 - 1)$ п'ятиточкових різницьових рівнянь

$$-\frac{y_{ij}^{n+1}}{\sigma\tau} + \Lambda y_{ij}^{n+1} = -F_{ij}, \quad x_{ij} \in \omega_h, \quad (7.159)$$

$$y_{ij}^{n+1} = \mu_{ij}^{n+1}, \quad x_{ij} \in \gamma_h,$$

де

$$F_{ij}^n = \frac{y_{ij}^n}{\sigma\tau} + \left(\frac{1}{\sigma} - 1 \right) \Lambda y_{ij}^n + \frac{1}{\sigma} \varphi_{ij}^n.$$

Систему (7.159) треба розв'язувати на кожному часовому ярусі, що вимагає значно більшої кількості операцій, ніж у випадку явної схеми.

Виникає потреба в побудові різницьових схем, які поєднують у собі найкращі властивості явних і неявних схем, тобто ці схеми повинні бути безумовно стійкими і вимагати кількості операцій, пропорційної кількості вузлів сітки ω_h . Такі схеми називають *економними*.

7.11.2. Метод змінних напрямків побудови економних різницьових схем

Розглянемо різницьову схему методу змінних напрямків для рівняння (7.153), яка називається *поздовжньо-поперечною схемою Пісмента-Рекфорда*. В цій схемі перехід від яруса n до яруса $n + 1$ здійснюється

у два етапи. На першому етапі визначають проміжні значення $y_{ij}^{n+1/2}$ з системи рівнянь

$$\frac{y_{ij}^{n+1/2} - y_{ij}^n}{0,5\tau} = \Lambda_1 y_{ij}^{n+1/2} + \Lambda_2 y_{ij}^n + \varphi_{ij}^n, \quad (7.160)$$

а на другому етапі, використовуючи знайдені значення $y_{ij}^{n+1/2}$, знаходять y_{ij}^{n+1} із системи рівнянь

$$\frac{y_{ij}^{n+1} - y_{ij}^{n+1/2}}{0,5\tau} = \Lambda_1 y_{ij}^{n+1/2} + \Lambda_2 y_{ij}^{n+1} + \varphi_{ij}^n. \quad (7.161)$$

До рівнянь треба додати початкові

$$y_{ij}^0 = u_0(x_{ij})$$

і різницеві крайові умови

$$y_{ij}^{n+1} = \mu(x_{ij}, t_{n+1}) \quad \text{при} \quad j = 0, j = N_2,$$

$$y_{ij}^{n+1/2} = \bar{\mu}(x_{ij}) \quad \text{при} \quad i = 0, i = N_1,$$

де

$$\bar{\mu}(x_{ij}) = \frac{1}{2} (\mu^{n+1}(x_{ij}) + \mu^n(x_{ij})) - \frac{\tau}{4} \Lambda_2 (\mu^{n+1}(x_{ij}) - \mu^n(x_{ij})),$$

$$\varphi_{ij}^n = f(x_{ij}, t_{n+1/2}).$$

Зауважимо, що схема (7.160) неявна за напрямком x_1 і явна за напрямком x_2 , а схема (7.161) явна за напрямком x_1 і неявна за напрямком x_2 .

Рівняння (7.160), (7.161) можна розв'язати послідовним застосуванням одновимірних прогонок спочатку за напрямком x_1 , а відтак за напрямком x_2 . Це вимагає операцій порядку $O(N)$, де $N = (N_1 - 1) \times (N_2 - 1)$.

Щоб встановити порядок апроксимації і дослідити стійкість поздовжньо-поперечної схеми, виключимо з системи (7.160), (7.161) проміжне значення $y_{ij}^{n+1/2}$. Для цього від рівняння (7.161) віднімемо рівняння (7.160), тоді

$$\frac{y_{ij}^{n+1} - 2y_{ij}^{n+1/2} + y_{ij}^n}{0,5\tau} = \Lambda_2 (y_{ij}^{n+1} - y_{ij}^n)$$

або

$$y_{ij}^{n+1/2} = \frac{1}{2} (y_{ij}^{n+1} + y_{ij}^n) - \frac{1}{4} \tau \Lambda_2 (y_{ij}^{n+1} - y_{ij}^n). \quad (7.162)$$

Підставимо (7.162) в (7.161):

$$\begin{aligned} & \frac{y_{ij}^{n+1} - \frac{1}{2} (y_{ij}^{n+1} + y_{ij}^n) + \frac{\tau}{4} \Lambda_2 (y_{ij}^{n+1} - y_{ij}^n)}{0,5\tau} = \\ & = \frac{1}{2} \Lambda_1 (y_{ij}^{n+1} + y_{ij}^n) - \frac{\tau}{4} \Lambda_1 \Lambda_2 (y_{ij}^{n+1} - y_{ij}^n) + \\ & + \Lambda_2 y_{ij}^{n+1} + \varphi_{ij}^n. \end{aligned}$$

Після спрощень одержимо

$$\frac{y_{ij}^{n+1} - y_{ij}^n}{\tau} = \frac{1}{2} \Lambda (y_{ij}^{n+1} + y_{ij}^n) + \varphi_{ij}^n - \frac{\tau}{4} \Lambda_1 \Lambda_2 (y_{ij}^{n+1} - y_{ij}^n). \quad (7.163)$$

Оскільки

$$y_{ij}^{n+1} = y_{ij}^n + O(\tau),$$

то

$$\frac{\tau}{4} \Lambda_1 \Lambda_2 (y_{ij}^{n+1} - y_{ij}^n) = O(\tau^2).$$

Звідси випливає, що поздовжньо-поперечна схема відрізняється від схеми Кранка-Нікольсона ($\sigma = 1/2$) на величину порядку $O(\tau^2)$, тому вона має порядок апроксимації $O(\tau^2 + |h|^2)$.

Поздовжньо-поперечна схема (7.163) має канонічний вигляд:

$$\left(I + \frac{\tau}{2} A_1\right) \left(I + \frac{\tau}{2} A_2\right) \frac{y_{n+1} - y_n}{\tau} + A y_n = \varphi_n, \quad y_0 = u_0,$$

де $A_\alpha y = -\Lambda_\alpha \overset{\circ}{y} = -\overset{\circ}{y}_{\bar{x}_\alpha x_\alpha}$, $\alpha = 1, 2$, $\overset{\circ}{y} = y$ при $x \in \omega_h$ і $\overset{\circ}{y} = 0$, $x \in \gamma_h$, $A_\alpha = A_\alpha^* > 0$, $\alpha = 1, 2$, $A_1 A_2 = A_2 A_1$.

Оскільки

$$B = I + \frac{\tau A}{2} + \frac{\tau^2}{4} A_1 A_2 \geqslant I + \frac{\tau A}{2} > \frac{\tau A}{2},$$

то схема (7.160), (7.161) безумовно стійка.

7.11.3. Локально-одновимірний метод

Поздовжньо-поперечна схема для задачі розмірності $p \geq 3$ не узагальнюється. Економні багатовимірні різницеві схеми можна побудувати *локально-одновимірним методом*, який також використовує проміжні яруси. Ці схеми мають лише *сумарну апроксимацію*. На проміжних ярусах вони взагалі не апроксимують вихідне рівняння, але похибки апроксимації проміжних ярусів сумуються так, що на цілому ярусі є апроксимація. Такі схеми називають *адитивними*.

Розглянемо p -вимірне рівняння

$$\frac{\partial u}{\partial t} = \sum_{\alpha=1}^p \frac{\partial^2 u}{\partial x_\alpha^2} + f(x, t), \quad 0 < t \leq T, \quad (7.164)$$

$$x = (x_1, x_2, \dots, x_p) \in \Omega = \{0 < x_\alpha < l_\alpha, \alpha = \overline{1, p}\}$$

з початковими та граничними умовами

$$u(x, 0) = u_0(x), \quad x \in \overline{\Omega}, \quad (7.165)$$

$$u(x, t) = \mu(t), \quad x \in \Gamma. \quad (7.166)$$

Рівняння (7.164) запишемо у вигляді

$$\sum_{\alpha=1}^p \left(\frac{1}{p} \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x_\alpha^2} - f_\alpha(x, t) \right) = 0,$$

де

$$f_1(x, t) + f_2(x, t) + \dots + f_p(x, t) = f(x, t).$$

На відріzkі $0 \leq t \leq T$ введемо рівномірну сітку $\omega_\tau = \{t_n = n\tau, n = \overline{0, K}, \tau = T/K\}$. Кожен інтервал $[t_n, t_{n+1}]$ розіб'ємо на p частин точками $t_{n+\alpha/p} = t_n + \alpha\tau/p, \alpha = \overline{1, p-1}$. Будемо послідовно розв'язувати рівняння

$$\begin{aligned} \frac{1}{p} \frac{\partial v_{(\alpha)}}{\partial t} &= \frac{\partial^2 v_{(\alpha)}}{\partial x_\alpha^2} + f_\alpha(x, t), \\ t_{n+(\alpha-1)/p} &\leq t \leq t_{n+\alpha/p}, \quad \alpha = \overline{1, p} \end{aligned} \quad (7.167)$$

за додаткових припущень

$$\begin{aligned} v_{(1)}(x, 0) &= u_0(x), \quad v_{(1)}(x, t_n) = v(x, t_n), \\ v_{(\alpha)}(x, t_{n+(\alpha-1)/p}) &= v_{(\alpha-1)}(x, t_{n+(\alpha-1)/p}), \quad \alpha = \overline{1, p}. \end{aligned} \quad (7.168)$$

Розв'язком цієї задачі є функція $v(x, t_{n+1}) = v_{(p)}(x, t_{n+1})$, $n = \overline{0, K-1}$.

Для розв'язування (7.167) можна використати схему з ваговими коефіцієнтами

$$\begin{aligned} \frac{y^{n+\alpha/p} - y^{n+(\alpha-1)/p}}{\tau} = \\ = \Lambda_\alpha (\sigma_\alpha y^{n+\alpha/p} + (1 - \sigma_\alpha) y^{n+(\alpha-1)/p}) + \varphi^{n+\alpha/p}, \quad x \in \bar{\omega}_h, \end{aligned}$$

де σ_α — параметри, які вибираються з умов стійкості і порядку апроксимації.

Розглянемо детальніше випадок $p = 2$, тоді (7.167), (7.168) буде мати вигляд

$$\begin{aligned} \frac{1}{2} \frac{\partial v_{(1)}}{\partial t} = \frac{\partial^2 v_{(1)}}{\partial x_1^2} + f_1(x, t), \quad t_n \leq t \leq t_{n+1/2}, \\ v_{(1)}^n = v^n, \quad v_{(1)}^0 = u_0(x), \end{aligned} \quad (7.169)$$

$$\begin{aligned} \frac{1}{2} \frac{\partial v_{(2)}}{\partial t} = \frac{\partial^2 v_{(2)}}{\partial x_2^2} + f_2(x, t), \quad t_{n+1/2} \leq t \leq t_{n+1}, \\ v_{(2)}^{n+1/2} = v_{(1)}^{n+1/2}. \end{aligned} \quad (7.170)$$

Розв'язок задачі (7.169), (7.170) $v(x, t_{n+1}) = v_{(2)}(x, t_{n+1})$. Кожне з рівнянь будемо апроксимувати за допомогою двоярусної неявної різницевої схеми з кроком $\tau/2$

$$\frac{y_{ij}^{n+1/2} - y_{ij}^n}{\tau} = \Lambda_1 y_{ij}^{n+1/2} + \varphi_{1ij}^n, \quad x_{ij} \in \omega_h, \quad (7.171)$$

$$\frac{y_{ij}^{n+1} - y_{ij}^{n+1/2}}{\tau} = \Lambda_2 y_{ij}^{n+1} + \varphi_{2ij}^n, \quad x_{ij} \in \omega_h, \quad (7.172)$$

$$y_{ij}^0 = u_0(x_{ij}), \quad x_{ij} \in \omega_h,$$

$$y_{ij}^{n+1/2} = \mu_{ij}^{n+1/2}, \quad y_{ij}^{n+1} = \mu_{ij}^{n+1}, \quad x_{ij} \in \gamma_h.$$

Дослідимо похибку апроксимації схеми (7.171), (7.172). Підставимо $y_{ij}^n = z_{ij}^n + u_{ij}^n$, $y_{ij}^{n+1/2} = z_{ij}^{n+1/2} + u_{ij}^{n+1/2}$, $y_{ij}^{n+1} = z_{ij}^{n+1} + u_{ij}^{n+1}$, тоді для похибки z отримаємо рівняння

$$\frac{z_{ij}^{n+1/2} - z_{ij}^n}{\tau} = \Lambda_1 z_{ij}^{n+1/2} + \psi_{1ij}^n,$$

$$\frac{z_{ij}^{n+1} - z_{ij}^{n+1/2}}{\tau} = \Lambda_2 z_{ij}^{n+1} + \psi_{2ij}^n,$$

де u — розв'язок задачі (7.164) — (7.166), $\psi_{1ij}^n, \psi_{2ij}^n$ — похибки апроксимації рівнянь (7.171), (7.172). Оскільки

$$u_{ij}^n = u_{ij}^{n+1/2} - \frac{\tau}{2} \frac{\partial u(x_{ij}, t_{n+1/2})}{\partial t} + O(\tau^2),$$

$$u_{ij}^{n+1} = u_{ij}^{n+1/2} + \frac{\tau}{2} \frac{\partial u(x_{ij}, t_{n+1/2})}{\partial t} + O(\tau^2),$$

то

$$\begin{aligned} \psi_{1ij}^n &= \Lambda_1 u_{ij}^{n+1/2} + \varphi_{1ij}^n - \frac{u_{ij}^{n+1/2} - u_{ij}^n}{\tau} = \\ &= \frac{\partial^2 u(x_{ij}, t_{n+1/2})}{\partial x_1^2} - \frac{1}{2} \frac{\partial u(x_{ij}, t_{n+1/2})}{\partial t} + \varphi_{1ij}^n + O(\tau + h_1^2), \end{aligned}$$

$$\begin{aligned} \psi_{2ij}^n &= \Lambda_2 u_{ij}^{n+1} + \varphi_{2ij}^n - \frac{u_{ij}^{n+1} - u_{ij}^{n+1/2}}{\tau} = \\ &= \frac{\partial^2 u(x_{ij}, t_{n+1/2})}{\partial x_2^2} - \frac{1}{2} \frac{\partial u(x_{ij}, t_{n+1/2})}{\partial t} + \varphi_{2ij}^n + O(\tau + h_2^2). \end{aligned}$$

Кожна зі схем (7.171), (7.172) не апроксимує вихідне рівняння (7.164), однак наявна сумарна апроксимація

$$\begin{aligned} \psi_{ij}^n &= \psi_{1ij}^n + \psi_{2ij}^n = \\ &= \frac{\partial^2 u(x_{ij}, t_{n+1/2})}{\partial x_1^2} + \frac{\partial^2 u(x_{ij}, t_{n+1/2})}{\partial x_2^2} - \frac{\partial u(x_{ij}, t_{n+1/2})}{\partial t} + \\ &+ \varphi_{1ij}^n + \varphi_{2ij}^n + O(\tau + |h|^2) = O(\tau + |h|^2), \end{aligned}$$

якщо

$$\varphi_{1ij}^n + \varphi_{2ij}^n = f_{ij}^{n+1/2} + O(\tau^2).$$

Остання рівність справджується, якщо вибрати, наприклад,

$$\varphi_{1ij}^n = 0, \quad \varphi_{2ij}^n = f_{ij}^{n+1/2} \quad \text{або} \quad \varphi_{1ij}^n = \varphi_{2ij}^n = 0,5 f_{ij}^{n+1/2}.$$

Отже, локально-одновимірний метод дозволяє розщепляти складні задачі на послідовність простіших і суттєво спрощувати розв'язування багатовимірних задач математичної фізики.

Контрольні завдання

✎ 7.1. На дев'ятиточковому шаблоні $(x_1^i, x_2^j), (x_1^i \pm h_1, x_2^j), (x_1^i, x_2^j \pm h_2), (x_1^i \pm h_1, x_2^j \pm h_2)$ побудуйте різницеву схему, яка апроксимує з четвертим порядком задачу Діріхле для рівняння Пуассона в області $\bar{\Omega} = \{0 \leq x_1 \leq l_1, 0 \leq x_2 \leq l_2\}$:

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x_1, x_2), \quad (x_1, x_2) \in \Omega,$$

$$u = \mu(x_1, x_2), \quad (x_1, x_2) \in \Gamma.$$

✎ 7.2. Визначте порядок апроксимації диференціального оператора

$$Lu = \frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x}$$

різницеvim оператором

$$L_{h\tau} y_i^n = \frac{y_i^{n+1} - \frac{y_{i+1}^n + y_{i-1}^n}{2}}{\tau} + a \frac{y_{i+1}^n - y_{i-1}^n}{2h}.$$

✎ 7.3. Доведіть, що різницева схема:

$$\frac{y_i^{n+1} - y_i^n}{\tau} = \frac{y_{i+1}^n - 2y_i^n + y_{i-1}^n}{h^2}$$

апроксимує диференціальне рівняння

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

з другим порядком по τ і четвертим по h , якщо $\tau/h^2 = 1/6$.

✎ 7.4. Доведіть, що різницева схема:

$$y_{t,i}^n + \sigma y_{\bar{x},i}^{n+1} + (1 - \sigma) y_{\bar{x},i}^n = 0, \quad i = 1, 2, \dots, \quad n = \overline{0, K-1},$$

$$y_i^0 = u_0(x_i), \quad i = 1, 2, \dots, \quad y_0^n = \mu(t_n), \quad n = \overline{0, K},$$

апроксимує задачу

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, \quad 0 < x < \infty, \quad 0 < t \leq T,$$

$$u(x, 0) = u_0(x), \quad 0 < x < \infty, \quad u(0, t) = \mu(t), \quad 0 \leq t \leq T,$$

із порядком апроксимації $O(\tau^2 + h^2)$ при $\sigma = \sigma_0 = \frac{1}{2} - \frac{h}{2\tau}$ і порядком $O(\tau + h)$ при $\sigma \neq \sigma_0$.

 7.5. Доведіть, що схема:

$$y_{t,i}^n + \sigma \tau y_{tt,i}^n = y_{xx,i}^{n+1} + \varphi_i^n, \quad i = \overline{1, N-1}, \quad n = \overline{0, K-1},$$

$$y_i^0 = u_0(x_i), \quad i = \overline{0, N},$$

$$y_0^n = \mu_1(t_n), \quad y_N^n = \mu_2(t_n), \quad n = \overline{0, K},$$


апроксимує задачу

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad 0 < x < 1, \quad 0 < t \leq T,$$

$$u(x, 0) = u_0(x), \quad 0 \leq x \leq 1,$$

$$u(0, t) = \mu_1(t), \quad u(1, t) = \mu_2(t), \quad 0 \leq t \leq T$$


із порядком $O(\tau^2 + h^2)$ при $\sigma = 0,5$, $\varphi_i^n = f_i^{n+1}$ і $O(\tau + h^2)$ при $\sigma \neq 0,5$, $\varphi_i^n = f_i^{n+1}$.

 7.6. Різницеву схему

$$y_{t,i}^n + \frac{1}{2} \tau y_{tt,i}^n = y_{xx,i}^{n+1} + \varphi_i^n, \quad i = \overline{1, N-1}, \quad n = \overline{0, K-1},$$

$$y_i^0 = u_0(x_i), \quad i = \overline{0, N}, \quad y_0^n = \mu_1(t_n), \quad y_N^n = \mu_2(t_n), \quad n = \overline{0, K},$$

запишіть у вигляді системи триточкових різницевих рівнянь. Перевірте умови стійкості методу прогонки для її розв'язування.

 7.7. Побудуйте різницеву схему, яка апроксимує з другим порядком по τ і h диференціальну задачу

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + x \cos t, \quad 0 < x < 1, \quad 0 < t \leq T,$$

$$u(x, 0) = \sin(\pi x), \quad 0 \leq x \leq 1,$$

$$\frac{\partial u(0, t)}{\partial x} = 0, \quad \frac{\partial u(1, t)}{\partial x} + u(1, t) = 1, \quad 0 \leq t \leq T,$$

і запишіть її у вигляді системи триточкових різницевих рівнянь. Перевірте умови стійкості методу прогонки для її розв'язування.

✎ 7.8. Зведіть до канонічного вигляду та вкажіть умови стійкості різницевої схеми

$$2\gamma y_i^{n+1} = (\gamma - 0,5) (y_{i-1}^{n+1} + y_{i+1}^{n+1}) + 0,5 (y_{i-1}^n + y_{i+1}^n),$$

$$\gamma = \tau/h^2, \quad i = \overline{1, N-1}, \quad n = \overline{0, K-1},$$

$$y_i^0 = u_0(x_i), \quad i = \overline{0, N}, \quad y_0^n = y_N^n = 0, \quad n = \overline{0, K}.$$

✎ 7.9. Різницеву схему Дугласа–Рекфорда

$$\frac{y_{ij}^{n+1/2} - y_{ij}^n}{\tau} = \Lambda_1 y_{ij}^{n+1/2} + \Lambda_2 y_{ij}^n,$$

$$\frac{y_{ij}^{n+1} - y_{ij}^{n+1/2}}{\tau} = \Lambda_2 (y_{ij}^{n+1} - y_{ij}^n),$$

$$y_{ij}^0 = u_0(x_{ij}), \quad x_{ij} \in \omega_h$$

запишіть у вигляді двох систем триточкових різницевих рівнянь. Перевірте умови стійкості методу прогонки для її розв'язування.

✎ 7.10. Зведіть до канонічного вигляду та вкажіть умови стійкості різницевої схеми з вправи 7.9.

✎ 7.11. Встановіть порядок апроксимації різницевої схеми з вправи 7.9.

✎ 7.12. Запишіть різницеву схему

$$\frac{y_{i+1,j} - 2y_{ij} + y_{i-1,j}}{h^2} + \frac{y_{i,j+1} - 2y_{ij} + y_{i,j-1}}{h^2} = -f_{ij},$$

$$y_{i0} = y_{i6} = 0, \quad y_{0j} = y_{6j} = 0, \quad i, j = \overline{1, 5},$$

у матричному вигляді $Ay = f$, перенумерувавши двовимірний масив індексів $(i, j)_{i,j=1}^5$ в одновимірний масив, наприклад, за правилом: індексові (i, j) поставимо у відповідність індекс $k = 5(i-1) + j$ одновимірного масиву.

СПИСОК ЛІТЕРАТУРИ

1. Бабенко К. И. *Основы численного анализа*. — М.: Наука, 1986.
2. Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. *Численные методы*. — М.: Наука, 1987.
3. Березин И. С., Жидков Н. П. *Методы вычислений*: В 2 т. — М.: ГИ-ФМЛ, 1962.
4. Воеводин В. В. *Численные методы алгебры: теория и алгоритмы*. — М.: Наука, 1966.
5. Волков Е. А. *Численные методы*. — М.: Наука, 1987.
6. Гаврилюк І. П., Макаров В. Л. *Методи обчислень*. — К.: Вища школа, 1995, ч.1, ч.2.
7. Гаврилюк І. П., Макаров В. Л. *Збірник задач з методів обчислень*. — К.: Вища школа, 1996.
8. Данилович В., Кутнів М. *Чисельні методи*. — Львів: Кальварія, 1998.
9. Дробышевич В. И., Дымников В. П., Ривин Г. С. *Задачи по вычислительной математике*. — М.: Наука, 1980.
10. Калиткин Н. Н. *Численные методы*. — М.: Наука, 1978.
11. Крылов В. И., Бобков В. В., Монастырный П. И. *Вычислительные методы*. — М.: Наука, 1976—1977, Т.1, Т.2.
12. Ляшко І. І., Макаров В. Л., Скоробогатько А. А. *Методи вычислений*. — К.: Вища школа, 1977.
13. Марчук Г. И. *Методы вычислительной математики*. — М.: Наука, 1989.
14. Марчук Г. И., Агошков В. И. *Введение в проекционно—сеточные методы*. — М.: Наука, 1981.

15. Ортега Дж., Пул У. *Введение в численные методы решения дифференциальных уравнений*. — М.: Наука, 1986.
16. Самарский А. А. *Введение в численные методы*. — М.: Наука, 1982.
17. Самарский А. А. *Теория разностных схем*. — М.: Наука, 1989.
18. Самарский А. А., Гулин А. В. *Численные методы*. — М.: Наука, 1986.
19. Самарский А. А., Лазаров Р. Д., Макаров В. Л. *Разностные схемы для дифференциальных уравнений с обобщенными решениями*. — М.: Наука, 1987.
20. Самарский А. А., Николаев Е. С. *Методы решения сеточных уравнений*. — М.: Наука, 1978.
21. Треногин В. А. *Функциональный анализ*. — М.: Наука, 1980.
22. Форсайт Дж., Малькольм М., Моулер К. *Машинные методы математических вычислений*. — М.: Мир, 1980.
23. Хайрер Э., Нерсетт С., Ваннер Г. *Решение обыкновенных дифференциальных уравнений. Нежесткие задачи*. — М.: Мир, 1990.