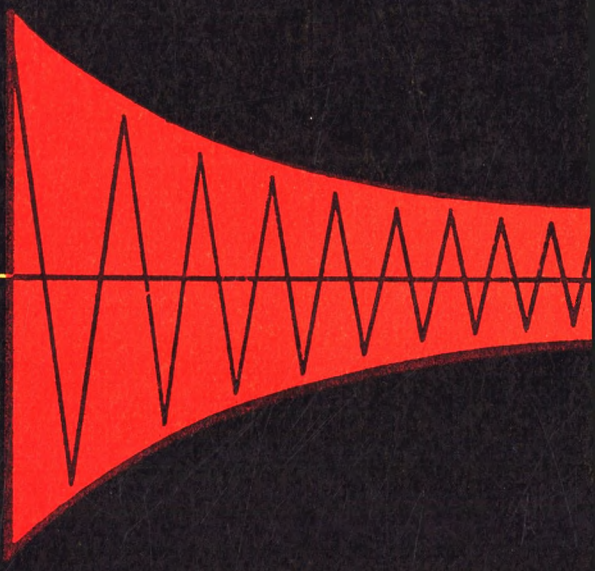
Методи обчислень

І. П. ГАВРИЛЮК
В. Л. МАКАРОВ

Підручник для вузів

1



І.П. ГАВРИЛЮК
В.Л. МАКАРОВ

МЕТОДИ ОБЧИСЛЕНЬ

У двох частинах

ЧАСТИНА І

Затверджено
Міністерством
освіти України
як підручник для студентів
вузів, які навчаються
за спеціальністю
«Прикладна математика»

КИЇВ
«ВИЩА ШКОЛА»
1995

ББК 22.193я73
Г12
УДК 51 (075.8)

Рецензенти: д-р фіз.-мат. наук *В. Я. Скоробагатько* (Інститут прикладних проблем механіки і математики НАН України), д-р фіз.-мат. наук *М. Ф. Кириченко* (Чернівецький університет)

Редакція літератури з математики, фізики, інформатики

Редактор *Є. В. Бондарчук*

Гаврилюк І. П., Макаров В. Л.

Г12 Методи обчислень: Підручник: У 2 ч.— К. : Вища шк., 1995.— Ч. 1. — 367 с. : іл.
ISBN 5-11-004111-2 (ч. 1)
ISBN 5-11-004029-X

Розглядаються традиційні розділи чисельних методів: аналіз похибок та стійкість алгоритмів, розв'язування лінійних та нелінійних рівнянь, інтерполювання, сплайн-наближення функцій, наближення функцій в лінійних нормованих просторах, чисельне диференціювання та інтегрування. Коротко викладено загальну теорію наближених методів.

Для студентів вузів, які навчаються за спеціальністю «Прикладна математика».

Г $\frac{1602012000-019}{211-95}$ 36—94

ББК 22.193я73

ISBN 5-11-004111-2 (ч. 1)
ISBN 5-11-004029-X

1. Чисельні методи в обчислювальному експерименті. Предмет чисельних методів

Історія прикладної математики почалась кілька тисячоліть тому, коли були розв'язані найпростіші математичні задачі з обчислення площ, об'ємів та ін. За час, що минув, у прикладній математиці відбулося багато змін, які позначались на її можливостях і впливі на життя суспільства. Дійсно революційне перетворення науки взагалі і математики зокрема пов'язане з появою в 40-х роках нинішнього століття електронних обчислювальних машин (ЕОМ). Ця подія привела до зміни технології наукових досліджень, до розширення можливостей вивчення складних явищ природи і суспільства, проектування сучасних технічних систем тощо. Прикладом може бути оволодіння ядерною енергією та освоєння космічного простору. Серед складних задач, які зараз стоять перед наукою, можна назвати моделювання людини, її взаємодії з природою, моделювання клімату та багато інших.

Як же в наш час розв'язуються складні задачі? Для того щоб вивчити проблему за допомогою математичних методів та ЕОМ, на п е р ш о м у етапі формулюють її в термінах тих об'єктів, які вивчає сучасна математика — систем лінійних чи нелінійних рівнянь, диференціальних рівнянь і т. п. Іншими словами, створюють математичну модель (ММ) явища, яке вивчається, чи технічної системи, яка проектується.

Далі звертаються за допомогою до ЕОМ. Але, як відомо, ЕОМ виконує лише найпростіші арифметичні і логічні операції, хоча і робить це з величезною швидкістю. Тому на д р у г о м у етапі математичну модель перетворюють до такого вигляду, щоб до неї входили лише ті операції, які може виконувати ЕОМ. Таке перетворення виконують за допомогою методів, які називають «чисельні методи» (ЧМ) або «методи обчислень» (МО). Як наслідок дістають нову модель, яка називається (на відміну від вихідної неперервної моделі) дискретною моделлю (ДМ). Далі (т р е т і й етап) за дискретною моделлю складають програму (П) для ЕОМ. Зауважимо, що рівень математичного забезпечення (МЗ) сучасних ЕОМ дає змогу програмісту уникнути трудомісткого і виснажливого шляху, коли при програмуванні дискретну модель доводиться «розписувати» аж до елементарних арифметичних і логічних операцій. В МЗ ЕОМ є так звані паке-

ти прикладних програм (ППП), і якщо в дискретній моделі, наприклад, потрібно розв'язати систему лінійних алгебраїчних рівнянь, то в програмі, яка реалізує цю дискретну модель, досить з ППП викликати відповідну підпрограму.

На четвертому етапі перевіряють (налагоджують) програму і розраховують за нею різні варіанти (цей етап позначатимемо ВР). Слід зауважити, що налагодження програми є дуже трудомістким етапом, про що свідчить жартівливий «принцип програміста»: «В будь-якій, навіть налагодженій, програмі є принаймні одна помилка». Розробка теоретичних методів перевірки правильності (верифікації) програм є важливим розділом теорії програмування, який бурхливо розвивається, проте поширена і експериментальна перевірка, або тестування.

Останнім (п'ятим) етапом розв'язування задачі є обробка результатів розрахунків на ЕОМ (ОР). Оскільки часто результати розрахунків являють собою тисячі чисел і займають десятки метрів паперу, то і тут потрібні спеціальні методи — побудова графіків, перерізів, статистична обробка тощо. Математики та спеціалісти в тій галузі знань, до якої належить практична задача, порівнюють результати ЕОМ з іншою інформацією про досліджуване явище чи об'єкт, з його фізичною моделлю (ФМ) і роблять висновок про те, чи достатньо математична модель описує реальність. Якщо це не так, то математична модель уточнюється і весь процес дослідження починається спочатку. Таким чином, схематично дістаємо технологічну схему розв'язування задачі (рис. 1). На цій же схемі показано місце різних розділів математики, в тому числі і чисельних методів, в такій технології розв'язування задач. Цю технологію прийнято називати обчислювальним експериментом (ОЕ); опис його дав академік О. А. Самарський.

Підкреслимо деякі переваги обчислювального експерименту: 1) ОЕ дешевший, швидший, простіший, ним легше керувати, ніж натурним експериментом; 2) в ОЕ можна моделювати умови, які неможливо створити в лабораторії або які призводять до загибелі об'єкта (моделювання людини, екологічних явищ і т. п.); 3) ОЕ дає змогу розв'язувати великі комплексні проблеми і приймати науково обгрунтовані рішення, планувати дослідження на відміну від класичних математичних методів, за допомогою яких можна було описувати багато фізичних явищ лише якісно, а точно можна було розв'язувати окремі найпростіші задачі; 4) ОЕ легко перебудувати для розв'язування різних задач, оскільки багато фізичних явищ описуються одними й тими самими рівняннями (наприклад, процес дифузії і поширення теплоти).

Суттєвий недолік ОЕ в тому, що придатність результатів розрахунків обмежена рамками математичної моделі, яка будується на основі вивчених на дослідах фізичних закономірностей. Тому ОЕ ніколи не замінить повністю натурний експеримент і майбутнє за їх

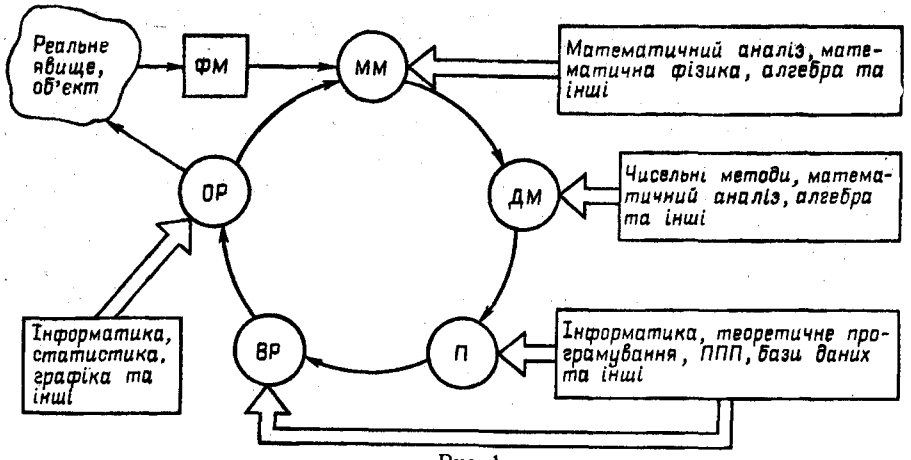


Рис. 1

розумним поєднанням. Таким чином, цикл обчислювального експерименту «об'єкт — модель — алгоритм — програма — ЕОМ — управління об'єктом» відображає основні етапи пізнання, в якому ЕОМ є інструментом для добування нових знань. Звичайно, наведена нами схема не описує всіх деталей технології ОЕ і може бути уточнена в різних напрямках. Нам важливо зрозуміти, яке місце посідає в ній розділ математики, що розглядатиметься далі.

Як можна визначити предмет математичного курсу «Методи обчислень» (МО), тобто коло питань, які в ньому вивчаються? У книзі «Математический энциклопедический словарь» (М.: Сов. энцикл., 1988.— 846 с.) обчислювальна математика розглядається як розділ математики, що включає коло питань, пов'язаних з використанням ЕОМ. Зміст терміна «МО» не можна вважати усталеним. Спочатку МО розглядали як прикладну математику. Термін «МО» застосовується і тоді, коли мають на увазі теорію чисельних методів і алгоритмів розв'язування типових математичних задач. У рамках сучасної термінології МО — частина інформатики, яка належить до методології застосування ЕОМ для розв'язування задач науки, техніки, виробництва і практично всіх областей людської діяльності.

Інформатика — це наука, яка перебуває в процесі становлення і вивчає закони і методи накопичення, передачі і обробки інформації за допомогою ЕОМ.

Як фундаментальна наука інформатика пов'язана з філософією через вчення про інформацію як загальнонаукову категорію і теорію пізнання; з математикою — через поняття математичної моделі, математичну логіку і теорію алгоритмів; з лінгвістикою — через вчення про формальні мови і про знакові системи. Інформатика тісно пов'я-

зана з такими спеціальними науками, як теорія інформації, кібернетика, системотехніка...

Щоб конкретизувати наведені вище означення розглянемо приклад конкретного фізичного явища і основні математичні питання при його вивченні за допомогою тріади «модель — алгоритм — програма».

Нехай треба сконструювати пристрій, наприклад руку маніпулятора, який в момент часу t_0 має почати прямолінійний рух із заданим прискоренням $f(t)$ і в момент часу t_1 досягти заданого положення. Якщо цей маніпулятор є частиною деякого робота, який має аналізувати координату положення маніпулятора в кожен момент часу за допомогою вмонтованого мікропроцесора (ЕОМ), то він повинен розв'язувати задачу

$$\frac{d^2 s(t)}{dt^2} = f(t), \tag{1}$$

$$s(t_0) = 0, \quad s(t_1) = s_1, \tag{2}$$

де $s(t)$ — шлях, пройдений кінцівкою маніпулятора на момент t, причому вважається, що шлях в момент часу t_0 дорівнює нулю і величина s_1 задана. Система співвідношень (1) являє собою математичну модель і називається ще крайовою задачею для звичайного диференціального рівняння. Одразу постає запитання: $З_1$) Чи має (1), (2) розв'язок, скільки таких розв'язків, які їх аналітичні властивості і т. п.? У випадку моделі (1), (2) відповідь можна дати порівняно просто. Так, з рівняння (1) послідовним інтегруванням знаходимо

$$\begin{aligned} \frac{ds(\xi)}{d\xi} &= \int_{t_0}^{\xi} f(\eta)\, d\eta + C_1, \\ s(t) &= \int_{t_0}^{t} \int_{t_0}^{\xi} f(\eta)\, d\eta d\xi + C_1 (t - t_0) + C_2, \end{aligned} \tag{3}$$

де C_1, C_2 — довільні сталі. Щоб задовольнити умови (2), треба покласти

$$C_2 = 0, \quad C_1 = (t_1 - t_0)^{-1} \left[s_1 - \int_{t_0}^{t_1} \int_{t_0}^{\xi} f(\eta)\, d\eta d\xi \right], \tag{4}$$

що і вирішує питання про існування розв'язку моделі (1), (2) за умови інтегровності функції $f(t)$. Вивчення цієї моделі можна продовжити. Але якщо, скажімо, модель якогось явища має схожий до (1), (2) вигляд

$$\frac{d^2 u}{dx^2} + q(x) u = f(x), \quad u(0) = u(1) = 0, \tag{5}$$

де $q(x)$, $f(x)$ — задані функції, то відповідь на запитання $З_1$ знайти значно важче, а для ще складніших моделей — взагалі неможливо.

Отже, задача розв'язування моделі (1), (2) звелася до обчислення інтегралів в (3), (4). Проте, якщо функція $f(t)$ досить складна, це може бути неможливо зробити аналітично. Постає запитання: $З_2$) Як обчислювати інтеграли на ЕОМ, знаючи основи математичного аналізу? Можемо запропонувати такий алгоритм $\tilde{A}_1$ обчислення інтеграла (при фіксованому ξ):

$$I = \int_{t_0}^{\xi} f(\eta)\, d\eta. \tag{6}$$

Розіб'ємо інтервал (t_0, ξ) на N частин точками $\eta_i = ih + t_0$, $i = \overline{1, N-1}$, $h = (\xi - t_0)/N$, і замінимо наближено інтеграл квадратурною сумою:

$$I \approx I_h = \sum_{i=1}^{N-1} h f(\eta_i) \tag{7}$$

(припускаємо, що маємо алгоритм для обчислення $f(\eta_i)$). Формула (7) є частиною дискретної моделі для математичної моделі (1), (2). Послідовно застосовуючи алгоритм $\tilde{A}_1$ до обчислення інтегралів в (3), знайдемо деякий алгоритм A_1 для розв'язання задачі. Зрозуміло, що розв'язок задачі (1), (2) маємо знайти із заданою наперед точністю, яка характеризується числом ε. Від цього залежить і характеристика точності ε_1, з якою треба вміти обчислювати інтеграл (6) при будь-якому ξ. Тоді постають запитання:

$З_3$) Яким треба вибрати N, а від цього залежить кількість операцій при обчисленні суми в (7), щоб виконувалась нерівність

$$|I - I_h| < \varepsilon_1? \tag{8}$$

$З_4$) Якою має бути функція $f(\eta)$, щоб при $h \to 0$ (або при $N \to \infty$) мала місце гранична рівність

$$\lim_{h \to 0} I_h = I? \tag{9}$$

При цьому треба врахувати, що числа в ЕОМ подаються наближено (бо комірки пам'яті ЕОМ мають лише скінченну кількість розрядів), а тому замість точного значення суми I_h ми дістанемо значення $\tilde{I}_h$, на яке вплинуть похибки заокруглення на ЕОМ. Знову виникає запитання: $З_5$) Яка сила цього впливу і як його зменшити?

Маючи початкові відомості з аналізу, а саме, пам'ятаючи, що

$$s'(t) = \lim_{\Delta t \to 0} \frac{s(t + \Delta t) - s(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{s(t) - s(t - \Delta t)}{\Delta t}, \tag{10}$$

можна запропонувати іншу дискретну модель і інший обчислювальний алгоритм A_2 розв'язування задачі (1), (2).

Розіб'ємо інтервал (t_0, t_1) на частини точками $t_j = j\tau + t_0$, $j = \overline{1, M-1}$, $\tau = (t_1 - t_0)/M$, $t_M = t_1$. Позначаючи наближене значення функції $s(t)$ в точці t_j через s_j і користуючись формулою (10), можна при малих τ покласти

$$s'(t_j) \approx \frac{s_{j+1} - s_j}{\tau} \text{ або } s'(t_j) \approx \frac{s_j - s_{j-1}}{\tau} \qquad (11)$$

і, застосовуючи це саме правило рекурентно до правої частини останньої рівності, покладемо

$$s''(t_j) \approx \frac{s'(t_j) - s'(t_{j-1})}{\tau} \approx \frac{s_{j+1} - 2s_j + s_{j-1}}{\tau^2}. \qquad (12)$$

Тоді (1), (2) можна наближено замінити на систему співвідношень

$$\frac{s_{j+1} - 2s_j + s_{j-1}}{\tau^2} = f(t_j), \quad j = \overline{1, M-1}, \quad s_0 = 0, \quad s_M = s_1. \qquad (13)$$

Маємо систему лінійних алгебраїчних рівнянь відносно невідомих s_i, $i = \overline{1, M-1}$, яка називається *різницевою схемою*. Вона ж є дискретною моделлю для (1), (2). Знову виникає ряд запитань:

$З_6$) Чи завжди

$$\lim_{\tau \to 0} \frac{s_{j+1} - s_j}{\tau} = s'(t_j), \quad \lim_{\tau \to 0} \frac{s_{j+1} - 2s_j + s_{j-1}}{\tau^2} = s''(t_j)?$$

$З_7$) За яких умов на вхідні дані розв'язок задачі (13) наближує розв'язок задачі (1), (2) в точках t_j, як визначити математично міру цієї близькості і як її оцінити через величину τ, якою ми можемо довільно оперувати?

$З_8$) Як розв'язувати систему лінійних алгебраїчних рівнянь (13)?

$З_9$) Який з алгоритмів A_1 і A_2 кращий: а) з точки зору кількості арифметичних операцій; б) з точки зору точності; в) з точки зору нечутливості до неминучих похибок заокруглення?

$З_{10}$) Якщо ми обчислили наближені значення $s_j \approx s(t_j)$, то як за ними обчислити (кажуть також інтерполювати) значення $s(t)$ при $t \neq t_j$? Конструювання алгоритмів A_i, відповідь на поставлені запитання $З_j$ належать до чисельних методів.

Можна поставити ще багато запитань, які виникають при дискретизації надзвичайно простої математичної моделі (1), (2), і на багато з них дати відповіді, але наведеного досить, щоб дати наступне означення предмета курсу «Методи обчислень», яке вважаємо за робоче: це *сукупність методів, технічних прийомів і теоретичних результатів, необхідних для розв'язування на ЕОМ математичних моделей задач науки і техніки.*

Наведений приклад дає нам також уявлення про конкретні розділи, які мають бути в цьому курсі — чисельне інтегрування, розв'язування систем лінійних алгебраїчних рівнянь, інтерполювання, або

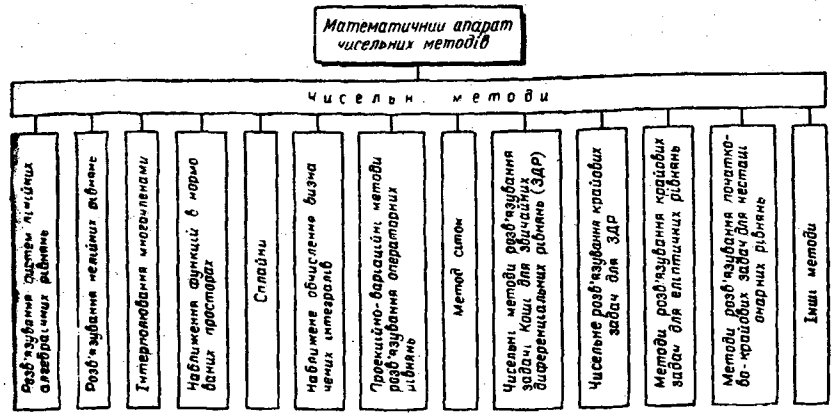


Рис. 2

наближення функцій, відомих лише в дискретній множині точок, та ін. Обмежимося класифікацією розділів чисельних методів (рис. 2), яка покладена в основу цієї книги.

Структура курсу. Автори прагнули до викладення сучасних понять теорії чисельних методів на найпростішому рівні, який би потребував мінімуму попередніх знань. Тому в перших главах ч. 1 нагадуються деякі необхідні далі поняття з математичного і функціонального аналізу. Розглядаються такі об'єкти чисельного аналізу, як різницеві (рекурентні) рівняння, системи лінійних алгебраїчних рівнянь, перше ознайомлення з якими відбувається у читача ще в шкільному курсі математики та на початку навчання у вищій школі. Але велике значення цих найпростіших об'єктів полягає в тому, що до них зводиться розв'язування багатьох складних задач. Наводяться також деякі відомості про абстрактні операторні схеми, які використовуються потім для обгрунтування ітераційних методів розв'язування лінійних операторних рівнянь та методів розв'язування нестаціонарних рівнянь з частинними похідними.

У книзі розглядаються різні способи апроксимації таких функціоналів і операторів, як функції, похідні, інтеграли, диференціальні та інтегральні оператори. Щоб дістати скінченновимірну задачу, треба апроксимувати не лише самі функціонали і оператори, але й простори, в яких вони діють, замінити деякими скінченновимірними просторами. Часто це підпростори скінченновимірного простору сіткових функцій або, як це іноді кажуть, функцій дискретного аргументу. Зауважимо, що такий спосіб апроксимації функціональних просторів не єдиний і з багатьох точок зору не є найкращим. Наприклад, функцію $f(x)$, задану на деякому відрізку $[a, b]$, можна апроксимувати вектором $f(x_i)$, $x_i \in \bar{\omega}_N = \{x_i : x_i \in [a, b], i = \overline{0, N}, x_i \neq x_j \text{ при } i \neq$

$\neq j\}$, який і є сітковою функцією, заданою на сітці ω_N. Альтернативним способом апроксимації функції $f(x)$ може бути заміна її набором чисел $a_0, a_1, \ldots, a_N$, які є коефіцієнтами розвинення $f(x)$ за деякою лінійно незалежною системою функцій $\varphi_0(x), \ldots, \varphi_N(x)$, $x \in [a, b]$. Перевагою і однією з причин поширеності першого способу є його простота. У гл. 1 першої частини вводиться поняття про сіткові функції як «елементарні частини» дискретного (скінченновимірного) аналізу і деякі операції над ними, наприклад, дискретні аналоги операцій диференціювання та інтегрування.

Оцінка якості наближення залежить, крім способу апроксимації, від вхідних даних вихідної задачі, тобто від областей визначення відповідних операторів. Розглядатимемо в основному області визначення, які містяться в класах функцій $C^k(\bar{\Omega})$ та $W_2^m(\Omega)$, де Ω — область зміни аргументу. Основні «інструменти», якими будемо користуватися,— це норма простору $C^k(\bar{\Omega})$

$$\|u\|_{C^k(\bar{\Omega})} = \sum_{i=0}^{k} \max_{x \in \bar{\Omega}} |u^{(i)}(x)|, \quad k \geqslant 0,$$

і норма простору $W_2^m(\Omega)$

$$\|u\|_{W_2^m(\Omega)} = \left(\sum_{i=0}^{m} |u|^2_{W_2^i(\Omega)} \right)^{1/2},$$

де

$$|u|_{W_2^i(\Omega)} = \left(\int_{\Omega} (u^{(i)}(x))^2 \, dx \right)^{1/2}$$

— напівнорма простору $W_2^i(\Omega)$, відповідні норми в просторах сіткових функцій та такі фундаментальні твердження про властивості цих просторів і операторів (функціоналів) в них, як лема Брембла — Гільберта, формула Тейлора, теореми вкладання, теорема Соболєва про еквівалентне нормування та ін. При дослідженні апроксимації в просторах $C^k(\bar{\Omega})$ зручно користуватися формулою Тейлора, а в просторах $W_2^m(\Omega)$ — лемою Брембла — Гільберта або теоремою Соболєва про еквівалентне нормування. Зосередження уваги на просторах $C^k(\bar{\Omega})$ та $W_2^m(\Omega)$ обумовлено тим, що вони широко використовуються при побудові математичних моделей реальних практичних задач. Крім того, вивчаючи чисельні методи в цих просторах, можна простежити, як залежить якість однієї й тієї самої апроксимації від гладкості вхідних даних.

В існуючих навчальних посібниках з методів обчислень при аналізі якості наближень основою є формула Тейлора. Техніка її застосування для кожної окремої задачі чисельного аналізу різна, що робить теорію чисельних методів більш схожою на «мозаїку» різних

прийомів, ніж на стрункну математичну конструкцію. Цей недолік певною мірою дає змогу усунути застосування леми Брембла — Гільберта, яка зводить доведення теорем про точність усіх традиційних чисельних методів до перевірки трьох умов: лінійності функціонала (оператора) похибки або його частин, обмеженості в просторі $W_2^m(\Omega)$ і перетворення його в нуль на многочленах до $(m-1)$-го степеня. Слід зауважити, що і така схема дослідження має недолік: сталі множники в оцінках похибки тут, як правило, більші, ніж при застосуванні формули Тейлора. Тому ці два підходи не виключають один одного.

Далі (гл. 2—5) розглядається апроксимація таких лінійних функціоналів, як значення функції $f(x)$ чи її похідних у фіксованій точці x^*, визначеного інтеграла від заданої функції $f(x)$ та одиничного оператора в нормованих просторах функцій (наближення функцій в лінійних нормованих просторах).

У другій частині книги описано проекційно-ітераційні методи і метод сіток для розв'язування операторних рівнянь, задача Коші та крайові задачі для звичайних диференціальних рівнянь, різні методи розв'язування цих задач.

Зробимо деякі зауваження щодо термінології. Так, для обчислення значення деякого оператора $L(x)$ автори використовують алгоритм, за яким обчислюється деяке значення $\xi(x, h)$, що і є наближенням для $L(x)$, залежним від малого параметра h (наприклад, від кроку сітки h), причому в деякій нормі $\xi(x, h) \to L$ при $h \to 0$, тобто $\|\xi(h, x) - L(x)\| \to 0$. Величину

$$R(x, h) = \xi(x, h) - L(x)$$

називатимемо *залишковим членом*. Якщо

$$\|R(x, h)\| = \|L(x) - \xi(x, h)\| \leqslant \varphi(x) h^p, \tag{14}$$

де числа $\varphi(x)$, p не залежать від h, причому $\varphi(x) > 0$, то порядок точності $\xi(x, h)$ або порядок точності формули $\xi(x, h) \approx L(x)$ дорівнює p, а вираз (14) називається *оцінкою залишкового члена*. Часто кажуть також, що точність величини $\xi(x, h)$ або точність формули $\xi(x, h) \approx L(x)$ є $0(h^p)$ або $\|\xi(x, h) - L(x)\| = 0(h^p)$. Зрозуміло, що чим більше значення $p > 0$ при $h \to 0$, тим точніше $\xi(x, h)$ наближає $L(x)$. Оцінки залишкового члена можуть бути апріорними і апостеріорними. *Апріорні оцінки* — це оцінки, які можна дати за постановкою і вихідними даними задачі без попередніх обчислень. *Апостеріорні оцінки* — це оцінки, які можна дати після проведення деяких обчислень. Зауважимо, що апріорні оцінки, як правило, мають вигляд нерівності $\|R(x, h)\| \leqslant \psi(x) h^p$, які дістають за допомогою строгих математичних міркувань. Апостеріорні оцінки часто мають вигляд наближеної рівності $\|R(x, h)\| \approx \psi(x) h^p$ і грунтуються як на строгих математичних міркуваннях, так і на «здоро-

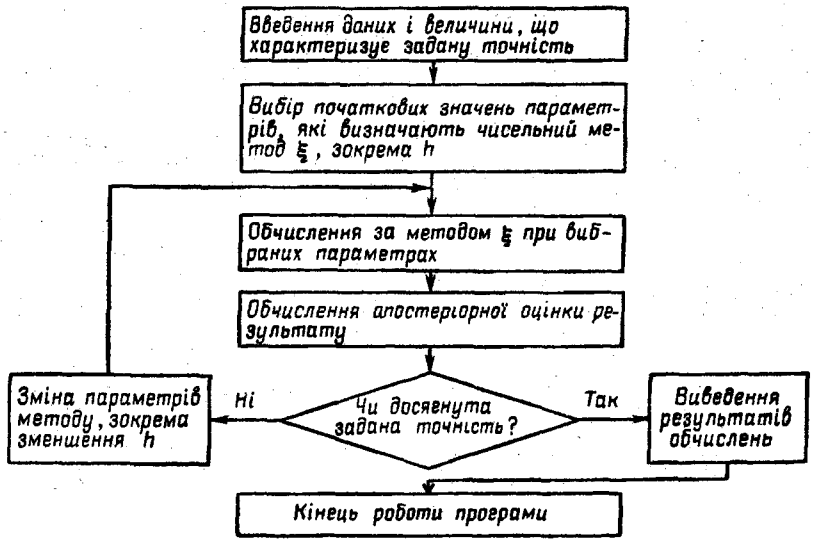


Рис. 3

вому глузді». Апостеріорні оцінки відіграють важливу роль у практичних обчисленнях і особливо при побудові програм розв'язування задач чисельного аналізу. Одним з основних принципів побудови цих програм є такий: на вхід програми надходить величина, що характеризує бажану точність результату, а на виході програма видає готовий результат із заданою точністю. Тому структура таких програм часто має вигляд, як на рис. 3.

Традиційно спочатку вивчається найпростіший спосіб апроксимації — інтерполювання. Розглядається умова однозначного розв'язку задачі інтерполювання, алгоритми побудови єдиного інтерполяційного алгебраїчного многочлена, різні форми його запису, оцінка похибки інтерполювання при фіксованому x, тобто коли йдеться про наближення функціонала $f(x)$. Новою порівняно з існуючими посібниками є техніка оцінки залишкового члена за допомогою леми Брембла — Гільберта та оцінка не тільки при належності функції $f(x)$ до класів $C^m[a, b]$, але й до класу $W_2^m(\Omega)$, причому m і кількість вузлів n незалежні.

Побудову інтерполяційного многочлена $p_n(x; f)$ можна розглядати також як апроксимацію неперервної функції $f(x)$ в лінійному нормованому просторі $C[a, b]$ або $L_{2,\rho}(a, b)$ з нормою $\|f\|_{L_{2,\rho}} = \left(\int_a^b f^2(x)\,\rho(x)dx\right)^{1/2}$. У зв'язку з цим аналізуються оцінки апроксимації в нормах цих просторів, зокрема важлива нерівність Лебега. Показано, що стала Лебега є коефіцієнтом підсилення неусувної похибки (або числом обумовленості задачі обчислення значення інтерполяційного многочлена в точці), а також досліджено вплив похибок заокруглення на результат при обчисленні значення інтерполяційного многочлена за схемою Горнера. Розглядаються умови збіжності інтерполяції та апостеріорні оцінки похибки. Як застосування інтерполювання розглядаються обернене інтерполювання та чисельне диференціювання.

Гл. 3 «Наближення функцій в лінійних нормованих просторах» починається класифікацією методів наближення функцій. Така класифікація певною мірою суб'єктивна, але, на думку авторів, вона є важливою з методичної точки зору, бо відіграє роль «опорного символу», який у вигляді «картинки» краще запам'ятовується і встановлює логічні зв'язки між задачами та методами чисельного аналізу, утворюючи деякий «каркас» курсу.

Такому важливому апарату наближення функцій, як інтерполяційний та згладжуючий сплайни, відведено окрему главу (гл. 4). Наводяться нові елементарні доведення додатної визначеності тридіагональної матриці системи лінійних алгебраїчних рівнянь, до якої зводиться задача побудови інтерполяційного кубічного сплайна, а також рівномірна збіжність згладжуючого кубічного сплайна до інтерполяційного при прямуванні до нескінченності вагових коефіцієнтів згладжуючого сплайна. Лема Брембла — Гільберта дає змогу легко довести оцінку похибки наближення функцій кубічними сплайнами.

Наближене обчислення визначених інтегралів розглянуто у гл. 5.

2. Аналіз похибок та стійкість алгоритмів

Аналіз похибок. Більшість задач обчислювальної математики можна сформулювати як задачу обчислення образу $f(x)$ деякого відображення $f: U \subset X \to Y$ на фіксованому елементі $x \in U$, де X, Y — деякі нормовані простори; U — деякий окіл x в X. Цю задачу позначатимемо також як задачу (f, x). Наприклад, задачу розв'язування системи лінійних алгебраїчних рівнянь

$$Ax = b$$

можна розглядати як задачу обчислення значення відображення $f(A, b) = A^{-1}b$ на елементі $(A, b) \in \mathrm{Mat}_n(R) \times R^n$, де R^n — евклідів n-вимірний простір векторів з дійсними компонентами; $\mathrm{Mat}_n(R)$ — простір дійсних $(n \times n)$-матриць.

Задача (f, x) розв'язується за допомогою деякого алгоритму, який, можливо, виробляє, крім результату $(f(x))$, ще деякі проміжні

результати за схемою:

Похибка в результаті $f(x)$ є, отже, наслідком похибок у вхідних даних та похибок в алгоритмі:

У той час як похибки вхідних даних належать до даної задачі і мало піддаються впливу з боку математика, що розв'язує цю задачу, похибки в алгоритмі можна регулювати, змінюючи метод розв'язування задачі.

Зупинимося на джерелах похибок у вхідних даних.

Класифікація джерел похибок у вхідних даних.

1. *Похибки за рахунок зображення дійсних чисел в ЕОМ.* В сучасних комп'ютерах прийнято так зване нормалізоване зображення з плаваючою комою, при якому дійсне число подається машинним числом

$$z = ad^e, \tag{1}$$

де a — мантиса; d — основа системи числення (як правило, 2, 8 або 16); e — порядок або експонента. Експонента змінюється в деякій множині цілих чисел:

$$e \in \{e_{\min}, \ldots, e_{\max}\} \subset Z.$$

Мантиса a є або нулем, або числом з проміжку $[d^{-1}, 1]$, яке має зображення

$$a = v \sum_{i=1}^{l} a_i d^{-i},$$

де $v \in \{\pm 1\}$ — знак числа; $a_i \in \{0, 1, \ldots, d-1\}$ — цифри; l — довжина мантиси.

Числа вигляду (1) утворюють у кожній ЕОМ деяку скінченну підмножину $F \subset R$ раціональних чисел. Кожне дійсне число x з модулем із проміжку $[d^{e_{\min}-1}, d^{e_{\max}}(1-d^{-l})]$ зображується в даній ЕОМ за допомогою машинного числа $\mathrm{fl}(x) \in F$, причому зауважимо, що відносна похибка цього зображення має оцінку

$$\left|\frac{x - \mathrm{fl}(x)}{x}\right| \leqslant \mathrm{eps} \equiv \frac{d^{1-l}}{2}, \tag{2}$$

де число eps називається *відносною машинною похибкою* (при single precision в мові FORTRAN або float в мові *C* в більшості комп'ютерів $\mathrm{eps} \approx 10^{-7}$). З виразу (2) для $\mathrm{fl}(x)$ випливає, що

$$\mathrm{fl}(x) = x(1 + \varepsilon(x)), \quad |\varepsilon(x)| \leqslant \mathrm{eps}. \tag{3}$$

Якщо $|x|$ менше, ніж найменше можливе в даній ЕОМ число $d^{e_{\min}-1}$, то його замінюють нулем і при цьому виникає ситуація, яка називається зникненням порядку (експоненти). Якщо ж

$$|x| > d^{e_{\max}}(1 - d^{-l}),$$

то виникає ситуація «переповнення» або ж автоматична зупинка (АЗ) і ЕОМ зупиняється.

Для подальшого аналізу важливо те, що дійсне число x після запису в ЕОМ не відрізняється від усіх інших дійсних чисел $\tilde{x}$, для яких $\mathrm{fl}(\tilde{x}) = \mathrm{fl}(x)$, тобто $\mathrm{fl}(x)$ є деякою множиною E дійсних чисел $\tilde{x}$ (окіл числа x (рис. 4)).

2. *Неусувна похибка, або похибка вимірювання.* Часто вхідні дані є результатом деякого експерименту і супроводжуються деякою абсолютною похибкою δx:

$$|x - \bar{x}| < \delta x,$$

де $\bar{x}$ — справжнє (невідоме!) значення. На практиці часто відносна похибка $|\delta x/x|$ лежить в межах $10^{-2} \ldots 10^{-3}$ (так звана «технічна точність») і є значно більшою від машинної похибки.

3. *Похибки в алгоритмі.* Будь-який алгоритм виражається через елементарні операції $(+, -, \cdot, /)$. Якщо позначати через $\circ$ будь-яку з цих операцій, то насправді в ЕОМ виконуються деякі наближені операції $\hat{\circ} \in \{\hat{+}, \hat{-}, \hat{\cdot}, \hat{/}\}$, які супроводжуються похибкою заокруглення, причому з виразу (3) маємо

$$x \hat{\circ} y = (x \circ y)(1 + \varepsilon(x, y)), \quad x, y \in F, \quad |\varepsilon(x, y)| \leqslant \mathrm{eps}.$$

При цьому для операцій $\hat{\circ}$ можуть не виконуватися відомі закони арифметики, зокрема асоціативні (наведіть приклад!), тобто при реалізації алгоритму на ЕОМ може мати значення порядок виконання операцій.

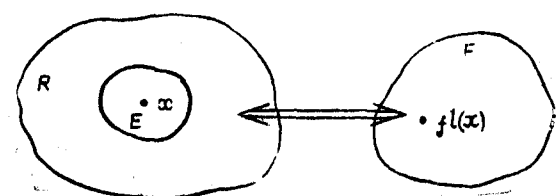


Рис. 4

4. *Похибки за рахунок наближених алгоритмів (похибки методу).* Ці похибки виникають

тоді, коли, наприклад, якась функція не може бути обчислена точно і обчислюється за допомогою деякої наближеної формули:

$$\cos x \approx 1 - \frac{x^2}{2!} + \frac{x^4}{4!}.$$

Обумовленість задачі і число обумовленості. Поставимо спочатку таке питання: як впливає похибка вхідних даних на похибку в результаті незалежно від вибраного алгоритму? Ми знаємо, що кожне машинне число $x \in \mathbb{R}^1$, яке є входом задачі (f, x) при розв'язанні на ЕОМ, насправді являє собою множину E дійсних чисел $\tilde{x}$:

$$E = \{\tilde{x} \in \mathbb{R}^1 \,|\, |x - \tilde{x}| \leqslant \mathrm{eps}\,|x|\}.$$

Якщо ж x знайдено з абсолютною точністю $\delta(x)$ з експерименту в результаті вимірювань, то воно являє собою множину дійсних чисел

$$E = \{\tilde{x} \in \mathbb{R}^1 \,|\, |x - \tilde{x}| \leqslant \delta(x)\}.$$

Отже, ми маємо розглядати відображення f в задачі (f, x) не як поточкове відображення, а як відображення множини E в деяку результуючу множину $R = f(E)$ (рис. 5). Нас цікавить, таким чином, співвідношення між множинами E та R; деяку характеристику цього співвідношення називатимемо *обумовленістю задачі* (f, x).

Щоб не ускладнювати аналіз, вважатимемо, що $f : U \subset \mathbb{R}^n \to \mathbb{R}^m$ є відображенням деякої відкритої множини U з евклідового простору $\mathbb{R}^n$ в евклідів простір $\mathbb{R}^m$. Задача (f, x) насправді задається відображенням f, точкою $x \in U$ і точністю вхідних даних δ. Ця точність може вимірюватися в деякій нормі простору $\mathbb{R}^n$:

$$\|\tilde{x} - x\| \leqslant \delta \quad \text{(абсолютна похибка)}, \tag{4}$$

або покомпонентно:

$$|\tilde{x}_i - x_i| \leqslant \delta \quad \text{(абсолютна покомпонентна похибка } \delta\text{)}, \tag{5}$$

$$|\tilde{x}_i - x_i| \leqslant \delta\,|x_i| \quad \text{(відносна покомпонентна похибка } \delta\text{)}, \quad i = \overline{1, n}.$$

Кожна з нерівностей (4), (5) виражає деяку множину вхідних даних $E \equiv E_\delta(x)$ задачі (f, x), що характеризується числом δ. Ми дістанемо бажану характеристику співвідношення множин E та R, якщо знайдемо оцінку різниці $f(\tilde{x}) - f(x)$ через різницю $\tilde{x} - x$ для всіх $\tilde{x}$ з множини $E_\delta(x)$. Ці різниці можна оцінювати в деяких нормах просторів $\mathbb{R}^m$ та $\mathbb{R}^n$ або

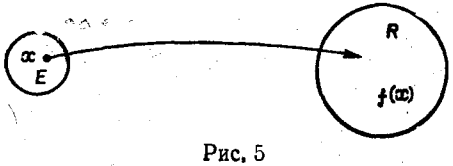


Рис. 5

покомпонентно. Залежно від цього говоритимемо про аналіз обумовленості за нормою чи про покомпонентний аналіз обумовленості. Залежно від задачі та цілей аналізу може бути зручнішим один або інший підхід. Почнемо з короткого огляду аналізу обумовленості за нормою.

Означення 1. Задача (f, x) називається *коректно поставленою* в $E_\delta(x)$, якщо існує стала $L_{\mathrm{abs}} > 0$ така, що

$$\|f(\tilde{x}) - f(x)\| \leqslant L_{\mathrm{abs}}\|\tilde{x} - x\| \;\forall\, \tilde{x} \in E_\delta(x). \tag{6}$$

В іншому разі (коли такої сталої L_{abs} не існує) задача називається *некоректно поставленою.*

Якщо задача (f, x) поставлена коректно, то позначимо через $L_{\mathrm{abs}}(\delta)$ мінімальну із сталих L_{abs}, для яких виконується (6).

Якщо $x \neq 0$, $f(x) \neq 0$, то означимо $L_{\mathrm{rel}}(\delta)$ як мінімальну сталу, для якої виконується нерівність

$$\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \leqslant L_{\mathrm{rel}}\frac{\|\tilde{x} - x\|}{\|x\|} \;\forall\, \tilde{x} \in E_\delta(x).$$

Іншими словами, сталі $L_{\mathrm{abs}}(\delta)$ та $L_{\mathrm{rel}}(\delta)$ характеризують «підсилення» абсолютної та відносної похибок при відображенні E в $f(E)$. В лінеарізованій теорії похибок розглядаються «досить малі» величини δ, інакше кажучи, розглядається $\delta \to 0$. Це приводить до таких означень.

Означення 2. Абсолютним числом обумовленості задачі (f, x) називається

$$k_{\mathrm{abs}} = \lim_{\delta \to 0} L_{\mathrm{abs}}(\delta),$$

а *відносним числом обумовленості* є

$$k_{\mathrm{rel}} = \lim_{\delta \to 0} L_{\mathrm{rel}}(\delta).$$

Задача (f, x) називається *добре (погано) абсолютно обумовленою,* якщо число k_{abs} є малим (великим). Задача (f, x) називається *добре (погано) відносно обумовленою,* якщо число k_{rel} є малим (великим).

Зауважимо, що відповідь на запитання, чи є число k_{abs} (k_{rel}) малим або великим, залежить від багатьох обставин, зокрема від ЕОМ, на якій розв'язується задача (точніше, від довжини мантиси l), від вимог точності і т. п. Одна і та сама задача залежно від цих обставин може бути як добре, так і погано обумовленою.

Якщо відображення f диференційовне, то в першому наближені (відносно $\tilde{x} - x$) маємо

$$f(\tilde{x}) - f(x) \approx f'(x)(\tilde{x} - x) \text{ при } \tilde{x} \to x,$$

звідки

$$\|f(\tilde{x}) - f(x)\| \leqslant \|f'(x)\| \|\tilde{x} - \tilde{x}\|,$$

де $f'(x) \in \mathrm{Mat}_{m,n}(\mathbb{R})$ — матриця Якобі відображення f,

$$\|f'(x)\| = \sup_{y \neq 0} \frac{\|f'(x)\, y\|}{\|y\|} = \sup_{\|y\|=1} \|f'(x)\, y\|$$

— матрична норма, узгоджена з векторною нормою $\|\cdot\|$ в $\mathbb{R}^n$. Звідси одразу дістанемо такі вирази для оцінки чисел обумовленості:

$$k_{\text{abs}} = \|f'(x)\|, \quad k_{\text{rel}} = \frac{\|x\|}{\|f(x)\|} \|f'(x)\|. \qquad 7)$$

Приклад 1. *Обумовленість додавання (віднімання).* Операція додавання є лінійним відображенням

$$f: \mathbb{R}^2 \to \mathbb{R}, \quad \begin{pmatrix} a \\ b \end{pmatrix} \to f(a, b) = a + b.$$

Звідси $f'(a, b) = (1, 1) \in \mathrm{Mat}_{1,2}(\mathbb{R})$. Вибравши в $\mathbb{R}^2$ норму

$$\|(a, b)^T\|_1 = |a| + |b|,$$

а в $\mathbb{R}$ взявши за норму абсолютну величину числа, маємо

$$\|f'(a, b)\|_1 = \sup_{|a|+|b|=1} \left\| (1,1) \begin{pmatrix} a \\ b \end{pmatrix} \right\|_1 = \sup_{|a|+|b|=1} |a + b| = 1,$$

звідки

$$k_{\text{abs}} = 1, \quad k_{\text{rel}} = \frac{|a| + |b|}{|a + b|}.$$

Як бачимо, у випадку чисел a, b з однаковими знаками $k_{\text{abs}} = k_{\text{rel}}$, і тому операцію додавання слід вважати добре обумовленою. Якщо ж числа a, b мають різні знаки (тобто йдеться про віднімання), то $|a + b| < |a| + |b|$ і тому $k_{\text{rel}} > 1$. Отже, задачу віднімання близьких за абсолютною величиною і різних за знаками чисел слід вважати погано відносно обумовленою. У цьому разі може настати явище, яке називається зникненням ведучих знаків (цифр). Суть його з'ясуємо на прикладі 2.

Приклад 2. *Зникнення ведучих знаків.* Нехай на ЕОМ з $\text{eps} = 10^{-7}$ віднімаються числа $x = 0{,}123456*$ та $y = 0{,}123445*$ (зірочкою позначимо непевний знак). Дістанемо $x - y = 0{,}000011*$ і після нормалізації $x - y = 0{,}11* \cdot 10^{-4}$, тобто непевний знак з'являється в третьому розряді. Це означає, що $k_{\text{rel}} \approx 10^4$ і задача погано обумовлена. *Звідси можна зробити такий загальний висновок: при конструюванні чи програмуванні алгоритму слід уникати віднімання близьких за абсолютною величиною чисел. Цього можна досягти за допомогою простого перетворення формул.*

Приклад 3. *Обумовленість системи лінійних алгебраїчних рівнянь* $Ax = b$. Вважатимемо спочатку, що лише вектор b являє собою вхідні дані відображення

$$f: \mathbb{R}^n \to \mathbb{R}^n, \quad b \to f(b) = A^{-1}b.$$

Неважко знайти, що $f'(b) = A^{-1}$, і тому для цієї задачі (f, b)

$$k_{\text{abs}} = \|A^{-1}\|, \quad k_{\text{rel}} = \frac{\|b\|}{\|A^{-1}b\|} \|A^{-1}\| = \frac{\|Ax\|}{\|x\|} \|A^{-1}\| \leqslant \|A\| \|A^{-1}\|.$$

Число

$$k(A) = \|A\| \|A^{-1}\|$$

називається *числом обумовленості матриці* A.

Нехай тепер вхідними даними є матриця A:

$$f: GL(n) \subset \mathrm{Mat}_n(\mathbb{R}) \to \mathbb{R}^n, \quad A \to f(A) = A^{-1}b,$$

(тут через $GL(n)$ позначено множину дійсних невироджених $(n \times n)$-матриць). Щоб знайти похідну відображення f, скористаємося розвиненням

$$(I - C)^{-1} = I + C + C^2 + \cdots,$$

яке справедливе для $C \in \mathrm{Mat}_n(\mathbb{R})$ таких, що $\|C\| < 1$. Тоді матимемо (порівняйте з похідною Гато нелінійного оператора!)

$$f'(A)\, C = \lim_{t \to 0} \frac{f(A + tC) - f(A)}{t} = \lim_{t \to 0} \frac{(A + tC)^{-1} b - A^{-1}b}{t} =$$

$$= \lim_{t \to 0} \frac{(I - (-tA^{-1}C))^{-1} A^{-1}b - A^{-1}b}{t} =$$

$$= \lim_{t \to 0} \frac{A^{-1}b - tA^{-1}CA^{-1}b - A^{-1}b + 0(t)}{t} = -A^{-1}CA^{-1}b = -A^{-1}Cx,$$

звідки

$$k_{\text{abs}} = \|f'(A)\| = \sup_{\|C\|=1} \|f'(A)\, C\| = \sup_{\|C\|=1} \|A^{-1}Cx\| \leqslant \|A^{-1}\| \|x\|,$$

$$k_{\text{rel}} = \frac{\|A\|}{\|A^{-1}b\|} \|f'(A)\| = \frac{\|A\|}{\|x\|} \|A^{-1}\| \|x\| = \|A\| \|A^{-1}\|.$$

Нехай, нарешті, вхідними даними для f є A та b:

$$f: GL(n) \times \mathbb{R}^n \to \mathbb{R}^n, \quad (A, b) \to f(A, b) = A^{-1}b.$$

Тоді для $C \in \mathrm{Mat}_n(R)$

$$f'(A, b)\, C = \lim_{t \to 0} \frac{f(A + tC, b + td) - f(A, b)}{t} =$$

$$= \lim_{t \to 0} \frac{(A + tC)^{-1}(b + td) - A^{-1}b}{t} =$$

$$= \lim_{t \to 0} \frac{(I - tA^{-1}C + o(t))\, A^{-1}(b + td) - A^{-1}b}{t} = -A^{-1}CA^{-1}b + A^{-1}d.$$

Введемо в просторі $GL(n) \times \mathbb{R}^n$ норму

$$\|(A, b)\| = \|A\|_m + \|b\|_v,$$

де $\|\cdot\|_v$ — векторна норма в $\mathbb{R}^n$, $\|\cdot\|_m$ — узгоджена з нею матрична норма. Тоді

$$k_{\text{abs}} = \|f'(A, b)\| = \sup_{\|C\|_m + \|d\|_v = 1} \|-A^{-1}CA^{-1}b + A^{-1}d\|_v \leqslant \|A^{-1}\| \|x\| + \|A^{-1}\|,$$

$$k_{\text{rel}} = \frac{\|A\|_m + \|b\|_v}{\|x\|_v} \|f'(A, b)\| \leqslant \frac{\|A\| + \|A\| \|x\|}{\|x\|} (\|A^{-1}\| \|x\| + \|A^{-1}\|) =$$

$$= k(A) \frac{(1 + \|x\|)^2}{\|x\|}.$$

Приклад 4. *Обумовленість системи нелінійних рівнянь.* Нехай f є неперервно диференційовним відображенням $f : \mathbb{R}^n \to \mathbb{R}^n$ і за заданим $y \in \mathbb{R}^n$ треба знайти $x \in \mathbb{R}^n$ таке, що

$$f(x) = y$$

(найчастіше $y = 0$). Ми бачимо, що задача є коректно поставленою, якщо існує $[f'(x)]^{-1}$. У цьому разі з теореми про обернену функцію випливає, що в деякому околі y існує f^{-1}, тобто $x = f^{-1}(y)$, причому

$$(f^{-1})'(y) = [f'(x)]^{-1}.$$

Тому числами обумовленості задачі (f^{-1}, y) є

$$k_{\text{abs}} = \|(f^{-1})'(y)\| = \|[f'(x)]^{-1}\|, \quad k_{\text{rel}} = \frac{\|f(x)\|}{\|x\|} \|[f'(x)]^{-1}\|.$$

У випадку одного скалярного рівняння $f(x) = y^* \neq 0$, $x \neq 0$, маємо $k_{\text{abs}} = \frac{1}{|f'(x)|}$, $k_{\text{rel}} = \frac{|y|}{|x||f'(x)|}$, тобто корені x_0 рівняння, в яких похідна близька до нуля, є погано обумовленими, а інші (x_1) добре обумовленими. Геометрично це означає, що пряма $y = y^*$ (рис. 6) в погано обумовлених коренях є «майже дотичною» до кривої $y = f(x)$.

Приклад 5. *Обумовленість задачі обчислення деякої суми.* Нехай треба обчислити суму

$$S(f_1, \ldots, f_n) = \sum_{i=1}^{n} \alpha_i f_i, \quad S : \mathbb{R}^n \to \mathbb{R},$$

де α_i — задані коефіцієнти, а компоненти вектора $f = (f_1, \ldots, f_n)^T \in \mathbb{R}^n$ — вхідні дані. Виберемо в $\mathbb{R}^n$ норму

$$\|f\|_\infty = \max_{i=\overline{1,n}} |f_i|,$$

а в $\mathbb{R}$ за норму вважатимемо модуль, тоді

$$S' = (\alpha_1, \ldots, \alpha_n) \in \text{Mat}_{1,n}(\mathbb{R}),$$

$$k_{\text{abs}} = \|S'\|_\infty = \sup_{x \neq 0} \frac{|S'x|}{\|x\|_\infty} = \sum_{i=1}^{n} |\alpha_i|.$$

Приклад 6. *Обумовленість однократного кореня поліноміального рівняння.* Нехай потрібно знайти однократний корінь ξ рівняння $p(x) = 0$, де

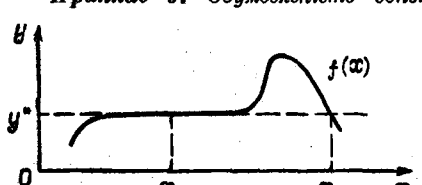


Рис. 6

$$p(x) = \sum_{i=0}^{n} a_i x^{n-i}$$

— многочлен n-го степеня, заданий своїми коефіцієнтами a_i, $i = \overline{0, n}$. Вхідними даними задачі $(p^{-1}, a) \equiv (\xi, a)$ є вектор $a = (a_0, \ldots, a_n) \in \mathbb{R}^{n+1}$ коефіцієнтів многочлена. Якщо ξ є простим коренем, то для досить малих $|\varepsilon|$ існує аналітична функція $\xi(\varepsilon)$ така, що $\xi(0) = \xi$ і $\xi(\varepsilon)$ є простим коренем «збуреного» рівняння $p_\varepsilon(x) \equiv p(x) + \varepsilon g(x) = 0$, де $g(x)$ — деякий многочлен n-го степеня, тобто

$$p(\xi(\varepsilon)) + \varepsilon g(\xi(\varepsilon)) \equiv 0.$$

Диференціюючи цю тотожність, дістаємо

$$kp'(\xi(0)) + g(\xi(0)) = 0, \quad k = \frac{d\xi(\varepsilon)}{d\varepsilon}\bigg|_{\varepsilon=0}.$$

Звідси

$$k = -g(\xi)/p'(\xi),$$

тобто в першому наближенні

$$\xi(\varepsilon) \approx \xi - \varepsilon \frac{g(\xi)}{p'(\xi)}.$$

Це означає, що $k_{\text{abs}} = \frac{g(\xi)}{p'(\xi)}$.

Якщо, наприклад, в многочлені $p(x)$ лише один коефіцієнт a_i замінити на збурений $a_i(1 + \varepsilon)$ (це означає, що $g(x) = a_i x^{n-i}$), то $k_{\text{abs}} \equiv k(i, \xi) = \left|\frac{a_i \xi^{n-i}}{p'(\xi)}\right|$ і нуль ξ буде *погано обумовленим, якщо* $k(i, \xi)$ *є значним* порівняно з ξ числом.

Розглянемо приклад многочлена

$$p(x) = (x-1)(x-2) \ldots (x-n) = \sum_{i=0}^{n} a_i x^{n-i},$$

який будемо вважати заданим своїми коефіцієнтами a_i, $i = \overline{0, n}$. Корені цього многочлена $\xi_k = k$, $k = \overline{1, n}$, є добре відокремленими і, здавалося б, їх можна легко знайти за допомогою будь-якого чисельного методу. Проте насправді це не так, що показує *аналіз на прикладі кореня* $\xi_n = n$. Припустимо, що лише коефіцієнт

$$a_1 = -\sum_{k=1}^{n} k = -(n+1)\, n/2$$

многочлена $p(x)$ збурюється, тобто замість a_1 маємо $a_1(1 + \varepsilon)$. Тоді, оскільки $p'(n) = (n-1)!$, дістанемо

$$k(1, n) = \left|\frac{a_1 n^{n-1}}{p'(n)}\right|.$$

Це означає, що

$$\xi_n(\varepsilon) - \xi_n \approx -\varepsilon \frac{(n+1)\, n \cdot n^{n-1}}{2(n-1)!}.$$

Із формули Стірлінга $n! \sim \sqrt{2\pi n}\,(n/e)^n$ випливає

$$|\xi_n(\varepsilon) - \xi_n| \approx \varepsilon \frac{(n+1)\, nn^{n-1}e^{n-1}}{2\sqrt{2\pi(n-1)}\,(n-1)^{n-1}} \approx O(\varepsilon n^{3/2} e^n) \to \infty, \quad n \to \infty.$$

Наприклад, при $n = 20$

$$\xi_{20}(\varepsilon) - \xi_{20} \approx -\varepsilon \frac{210 \cdot 20^{19}}{19!} \approx -\varepsilon \cdot 0{,}9 \cdot 10^{10}.$$

Приклад многочлена

$$p(x) = \sum_{i=0}^{20} a_i x^{20-i} = \prod_{j=0}^{20} (x - 2^{-j})$$

показує, що корені полінома, які близько розміщені один від одного, не обов'язково є погано обумовленими. Дійсно, цей многочлен має корені $\xi_j = 2^{-j}$, і якщо замінити $a_{20} = 2^{-1} \cdot 2^{-2} \cdot \ldots \cdot 2^{-20}$ на $a_{20}(1 + \varepsilon)$, то неважко знайти

$$\left| \frac{\xi_{20}(\varepsilon) - \xi_{20}}{\xi_{20}} \right| \approx \left| \frac{\varepsilon}{(2^{-1} - 1) \ldots (2^{-20} - 1)} \right| \leqslant 4 |\varepsilon|.$$

Можна довести, що навіть коли збурити всі коефіцієнти, то відносна похибка залишиться обмеженою. Проте це не стосується абсолютної похибки. Наприклад, якщо $a_{20} = 2^{-210}$ замінити на $\bar{a}_{20} = a_{20} + \Delta a_{20}$, $\Delta a_{20} = 2^{-48} \approx 10^{-14}$, то збурений многочлен матиме корені $\bar{\xi}_i$, причому

$$\bar{\xi}_1 \cdots \bar{\xi}_{20} = 2^{-210} + 2^{-48} = (2^{162} + 1)(\xi_1 \cdots \xi_{20}).$$

Останнє означає, що знайдеться принаймні одне r, для якого

$$|\bar{\xi}_r / \xi_r| \geqslant (2^{162} + 1)^{1/20} > 2^8 = 256.$$

Приклад 7. *Обумовленість многократного кореня поліноміального рівняння.*

Нехай ξ є m-кратним коренем рівняння $p(x) \equiv \sum_{i=0}^{n} a_i x^{n-i} = 0$. Тоді можна довести, що рівняння $p(x) + \varepsilon g(x) = 0$ має корінь вигляду $\xi(\varepsilon) = \xi + h(\varepsilon^{1/m})$, де $h(t)$ при досить малих $|t|$ є аналітичною функцією і $h(0) = 0$. Оскільки $p(\xi) = p'(\xi) = \ldots = p^{(m-1)}(\xi) = 0$, $p^{(m)}(\xi) \neq 0$, то, поклавши $t^m = \varepsilon$ і продиференціювавши тотожність

$$p(\xi + h(t)) + t^m g(\xi + h(t)) \equiv 0$$

m разів по t і поклавши потім $t = 0$, для $k = \left.\frac{dh(t)}{dt}\right|_{t=0}$, дістанемо

$$k = \left[-\frac{m!\, g(\xi)}{p^{(m)}(\xi)} \right]^{1/m}$$

і в першому наближенні

$$\xi(\varepsilon) - \xi \approx \varepsilon^{1/m} \left[-\frac{m!\, g(\xi)}{p^{(m)}(\xi)} \right]^{1/m}.$$

Зокрема, якщо в $p(x)$ збурити лише один коефіцієнт a_i, тобто $g(x) = a_i x^{n-i}$, то

$$\xi(\varepsilon) - \xi \approx \varepsilon^{1/m} \left[-\frac{m!\, a_i \xi^{n-i}}{p^{(m)}(\xi)} \right]^{1/m}.$$

Стійкість алгоритмів. При розв'язуванні задачі (f, x) за допомогою деякого алгоритму відображення f зображується у вигляді деякої суперпозиції відображень:

$$f = f_k \circ \cdots \circ f_1, \tag{8}$$

де $f_j : U_j \subset R^{n_j} \to U_{j+1} \subset R^{n_{j+1}}$. Кожному алгоритму для розв'язування однієї і тієї самої задачі (f, x) відповідає своє розвинення виду (8). Дійсно, f_j являють собою елементарні операції $\circ \in \{+, -, \cdot, /\}$, але оскільки в реальних алгоритмах їх може налічуватись мільйони, то для теоретичного аналізу застосовуються розвинення (8), в яких f_j може означати цілий блок елементарних операцій. Ми знаємо, що в арифметиці з плаваючою комою арифметичні операції а отже і f_j, виконуються неточно, зокрема замість операцій $\circ$ виконуються машинні операції $\hat{\circ}$. Отже,

$$a \hat{\circ} b = (a \circ b)(1 + \varepsilon(a, b)), \quad |\varepsilon(a, b)| \leqslant \mathrm{eps}. \tag{9}$$

Таким чином, на ЕОМ виконується не алгоритм (8), а деякий збурений похибками виду (9) алгоритм

$$\tilde{f} = \tilde{f}_k \circ \cdots \circ \tilde{f}_1. \tag{10}$$

Важливо підкреслити, що тепер нас цікавить не конкретний вигляд $\tilde{f}_j$, а лише те, що ці операції дають наближений результат з оцінкою типу (9). Отже, ми маємо аналізувати цілий клас $F = \{\tilde{f}\}$ різних наближених відображень $\tilde{f}$, проміжні кроки $\tilde{f}_j$ яких характеризуються оцінками типу (9). До класу F належить зокрема і відображення f.

Запитання про стійкість алгоритму (8) можна сформулювати тепер так: «Чи влаштовує нас результат $\tilde{f}(x)$, знайдений за допомогою алгоритму (10) замість алгоритму (8)?» Для відповіді ми, очевидно, повинні мати оцінку різниці $\tilde{f}(x) - f(x)$. Але оскільки замість вхідних даних x ми маємо справу також лише з деяким наближенням $\tilde{x}$, то насправді нас цікавить різниця

$$\tilde{f}(\tilde{x}) - f(x),$$

яка і становить повну похибку.

Відомо, що кожному частинному відображенню f_j можна поставити у відповідність частинне число обумовленості k_j, яке фактично є коефіцієнтом підсилення похибки даним частинним відображенням. Число обумовленості k всього відображення f вигляду (8) називатимемо *субмультиплікативним*, якщо

$$k \leqslant k_1 \cdot \ldots \cdot k_k. \tag{11}$$

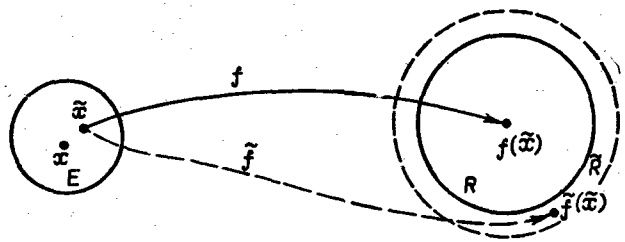


Рис. 7

Зрозуміло, що алгоритм слід вважати хорошим (стійким), якщо похибка $\tilde{f}(x) - f(x)$ лежить у розумних межах похибки $f(\tilde{x}) - f(x)$, спричиненої вхідними даними.

Щоб оцінити повну похибку, застосовуються два підходи, в яких використано дві теоретичні моделі поведінки похибок. Ці підходи відомі як *прямий* і *зворотний аналізи похибок.*

При прямому аналізі похибок розглядається множина (рис. 7)

$$\tilde{R} = \bigcup_{\tilde{f} \in F} \tilde{f}(E)$$

всіх образів, що виникають внаслідок неточності вхідних даних та неточностей в алгоритмі (оскільки $\tilde{f} \in F$, то $\tilde{f}(R) \in \tilde{R}$). Стійкість алгоритму при прямому аналізі характеризується співвідношенням множин $\tilde{R}$ та $R = f(E)$. Щоб описати це співвідношення кількісно, введемо такі припущення:

$$1)\ |\tilde{x} - x|_i \leqslant \delta |x|_i, \quad i = \overline{1, n}, \tag{12}$$

(покомпонентна відносна похибка вхідних даних не перевищує δ);

$$2)\ |\tilde{f}_j(u) - f_j(u)|_i \leqslant \delta |f_j(u)|_i \ \forall\, u \in U_j, \quad i = \overline{1, n_{j+1}}, \tag{13}$$

(покомпонентна відносна похибка при виконанні проміжних операцій $\tilde{f}_j$ не перевищує δ).

Зауважимо, що умови (12), (13) будуть, зокрема, виконані з $\delta \leqslant$ $\leqslant$ eps, якщо йдеться про зображення дійсних чисел в ЕОМ та про елементарні арифметичні операції.

Лема. *Нехай виконано умови* (12), (13). *Тоді для повної похибки при розв'язуванні задачі* (f, x) *за допомогою алгоритму* (10) *справедлива оцінка*

$$|\tilde{f}(\tilde{x}) - f(x)| \leqslant \tilde{\varepsilon} |f(x)| \text{ при } \delta \to 0,$$

де

$$\tilde{\varepsilon} = (1 + k_k + \cdots + k_k \cdot \ldots \cdot k_1)\delta,$$

k_j — *покомпонентне відносне число обумовленості відображення* f_j, *тобто* $\max\limits_i \dfrac{|\tilde{f}_j(u) - f_j(u)|_i}{|f_j(u)|_i} \leqslant k_j \max \dfrac{|\tilde{u} - u|_i}{|u|_i}$.

Доведення. Для спрощення викладу розглянемо випадок, коли всі f_j є скалярними фукціями, $f_j : U_j \subset \mathbb{R}^1 \to \mathbb{R}^1$. Позначимо $x_1 = x$, $\tilde{x}_1 = \tilde{x}$, $x_{j+1} = f_j(x_j)$, $\tilde{x}_{j+1} = \tilde{f}_j(\tilde{x}_j)$, $j = \overline{1, k}$. Тоді

$$|\tilde{f}(\tilde{x}) - f(x)| = |\tilde{x}_{k+1} - x_{k+1}| = |\tilde{f}_k(\tilde{x}_k) - f_k(x_k)| \leqslant$$

$$\leqslant |\tilde{f}_k(\tilde{x}_k) - f_k(\tilde{x}_k)| + |f_k(\tilde{x}_k) - f_k(x_k)| \leqslant$$

$$\leqslant \delta |f_k(\tilde{x}_k)| + k_k |\tilde{x}_k - x_k| |f_k(x_k)| / |x_k| \leqslant$$

$$\leqslant (\delta + k_k |\tilde{x}_k - x_k| / |x_k|) |f(x)| \leqslant \cdots \leqslant (\delta +$$

$$+ k_k\delta + \cdots + k_k \cdot \ldots \cdot k_1\delta) |f(x)| \text{ при } \delta \to 0.$$

З іншого боку, похибку, що є наслідком неточності вхідних даних $(f(E))$, можна виразити через число обумовленості k задачі (f, x):

$$|f(\tilde{x}) - f(x)| \leqslant \varepsilon |f(x)|, \quad \varepsilon = k\delta.$$

Вплив алгоритму (співвідношення між $\tilde{R}$ та R) можна тепер виразити числом

$$\frac{\tilde{\varepsilon}}{\varepsilon} = \frac{1}{k}(1 + k_k + \cdots + k_k \cdot \ldots \cdot k_1),$$

в якому, проте, не видно впливу частинних відображень f_j, бо невідомо чи k залежить від k_j і як. Якщо ж припустити, що k має властивість субмультиплікативності (11), то

$$|f(\tilde{x}) - f(x)| \leqslant k\delta |f(x)| \leqslant k_1 \cdot \ldots \cdot k_k\delta |f(x)|,$$

і тоді для $\varepsilon = k_1 \cdot \ldots \cdot k_k\delta$ маємо таку оцінку співвідношення $\tilde{R}$ та R:

$$\frac{\tilde{\varepsilon}}{\varepsilon} = 1 + \frac{1}{k_1} + \cdots + \frac{1}{k_1 \cdot \ldots \cdot k_k} \geqslant 1.$$

Означення 3. Нехай $f = f_k \circ \ldots \circ f_1$ є алгоритмом розв'язування задачі (f, x), а k_j — числом обумовленості проміжної задачі (f_j, x_j). Тоді число

$$\sigma = 1 + \frac{1}{k_1} + \cdots + \frac{1}{k_1 \cdot \ldots \cdot k_k} \tag{14}$$

називається *індикатором (показником) стабільності алгоритму* $f = f_k \circ \ldots \circ f_1$.

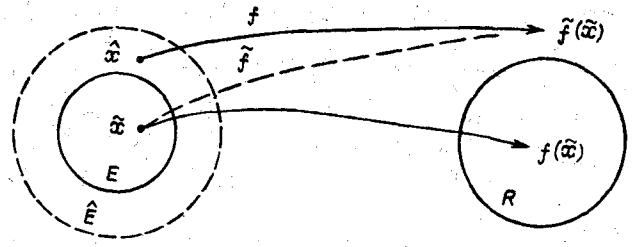


Рис. 8

З леми дістаємо таку оцінку повної похибки через індикатор стабільності:

$$|\tilde{f}(\tilde{x}) - f(x)| \leqslant \tilde{\varepsilon}\,|f(x)| = \sigma\varepsilon\,|f(x)| = k_1 \cdot \ldots \cdot k_k \sigma\delta\,|f(x)|. \quad (15)$$

Проте у цій оцінці ми не бачимо впливу числа обумовленості k задачі (f, x) на повну похибку, хоча «здоровий глузд» підказує, що цей вплив має бути одного порядку з впливом стійкості алгоритму, який виражений через σ. Щоб знайти відповідну оцінку, нам доведеться змінити концепцію (модель) оцінки повної похибки і звернутись до так званого *зворотного аналізу похибок,* що був запроваджений Дж. Х. Вілкінсоном.

Ідея зворотного аналізу полягає в тому, що похибки результату, обумовлені алгоритмом, інтерпретуються як похибки, викликані додатковою похибкою вхідних даних. Іншими словами, «збурений» алгоритмом результат $\tilde{y} = \tilde{f}(\tilde{x})$ вважається точним результатом відображення f (рис. 8) на «збуреному» вхідному значенні $\hat{x}$: $\tilde{f}(\tilde{x}) = f(\hat{x})$. Це можливо лише тоді, коли $\tilde{f}(E)$ належить множині значень f. В іншому разі зворотний аналіз неможливий і алгоритм вважається нестійким щодо зворотного аналізу. Може статися так, що існує більш ніж одне $\hat{x} \in f^{-1}(\tilde{y})$. Тоді береться те $\hat{x}$, яке міститься найближче до точного значення x задачі (f, x), тобто $\|\hat{x} - x\| = \min$. Розглянемо множину

$$\hat{E}(\tilde{x}, \tilde{f}) = \{\hat{x} \mid f(\hat{x}) = \tilde{f}(\tilde{x}) \text{ і } \|\hat{x} - x\| = \min\}, \quad \tilde{x} \in E, \ \tilde{f} \in F.$$

Припускаємо, що f є неперервним відображенням, тому ця множина не порожня. Кожен збурений результат $\tilde{f}(\tilde{x})$ можна тепер інтерпретувати як точний результат $f(\hat{x})$ на деякому елементі $\hat{x}$ з множини

$$\hat{E} = \bigcup_{\substack{\tilde{x}\in E \\ \tilde{f}\in F}} \hat{E}(\tilde{x}, \tilde{f}).$$

Співвідношення між множинами $\hat{E}$ та E і є мірою стійкості алгоритму щодо зворотного аналізу. Щоб оцінити цю міру кількісно, введемо величину

$$\tilde{\delta}(\tilde{x}, \tilde{f}) = \min_{f(\hat{x})=\tilde{f}(\tilde{x})} \|\hat{x} - x\|$$

і дамо таке означення.

Означення 4. Зворотною похибкою машинного алгоритму, що характеризується множиною F, називається величина

$$\tilde{\delta} = \max_{\tilde{x}\in E,\ \tilde{f}\in F} \tilde{\delta}(\tilde{x}, \tilde{f}) = \max_{\tilde{x}\in E,\ \tilde{f}\in F} \min_{f(\hat{x})=\tilde{f}(\tilde{x})} \|\hat{x} - x\|.$$

Наступна теорема показує, що індикатор стійкості σ характеризує співвідношення між зворотною похибкою і похибкою вхідних даних і при зворотному аналізі (зверніть увагу, що на відміну від прямого аналізу тут використовуються абсолютні числа обумовленості і оцінки не покомпоненті, а за нормою).

Теорема. *Нехай* $f = f_k \circ \ldots \circ f_1 : U \subset R^n \to R^n$ *є зображення деякого алгоритму розв'язування задачі* (f, x), *в якому частинні відображення* $f_j : U_j \subset R^n \to U_{j+1}$, $x_1 = x \in U$, $x_{j+1} = f_j(x_j)$, *є неперервно диференційовними і існують обернені відображення для* $f_j(x_j)$. *Нехай далі*

$$F_\delta = \{\tilde{f} = \tilde{f}_k \circ \ldots \circ \tilde{f}_1 \mid \|\tilde{f}_j(u) - f_j(u)\| \leqslant \delta \ \forall\, u \in U_j\}$$

є сім'єю збурених відображень, для яких частинні похибки за нормою обмежені величиною δ *і* $E_\delta = \{\tilde{x} \mid \|\tilde{x} - x\| \leqslant \delta\}$ *— відповідна множина збурених вхідних даних. Тоді для зворотної похибки* $\tilde{\delta}$ *щодо множин* E_δ *та* F_δ *має місце оцінка*

$$\tilde{\delta} \leqslant \sigma\delta \text{ при } \delta \to 0,$$

де σ *є індикатором стійкості з частинними числами обумовленості*

$$k_j = \|f_j'(x_j)\|.$$

Доведення. Розглянемо спочатку одне відображення f і клас $F = \{\tilde{f}\}$ збурених відображень, для яких виконується нерівність

$$\|\tilde{f}(u) - f(u)\| \leqslant \eta\delta \ \forall\, u \in U,$$

де η — незалежна від δ стала. Оскільки існує $f'^{-1}(x)$, то

$$\hat{x} - x \approx f'^{-1}(x)(f(\hat{x}) - f(x)) \text{ при } \hat{x} \to x,$$

звідки

$$\|\hat{x} - x\| \leqslant k^{-1} \|f(\hat{x}) - f(x)\|,$$

де $k = \|f'(x)\|$. Далі маємо

$$\|\hat{x} - x\| \leqslant k^{-1} \|f(\hat{x}) - f(x)\| = k^{-1} \|\tilde{f}(\tilde{x}) - f(x)\| \leqslant$$

$$\leqslant k^{-1} (\|f(\tilde{x}) - f(\tilde{x})\| + \|f(\tilde{x}) - f(x)\|) \leqslant k^{-1} (\eta\delta + k\delta) = \delta\left(1 + \frac{\eta}{k}\right).$$

Застосовуючи тепер цю нерівність рекурсивно до частинних відображень f_j і враховуючи, що $\eta_k = 1$, дістаємо

$$\tilde{\delta} \leqslant \delta\left(1 + \frac{\eta_1}{k_1}\right) \leqslant \delta\left(1 + \frac{1}{k_1}\left(1 + \frac{\eta_2}{k_2}\right)\right) \leqslant \cdots \leqslant$$

$$\leqslant \delta\left(1 + \frac{1}{k_1} + \frac{1}{k_1 k_2} + \cdots + \frac{\eta_k}{k_1 \cdot \ldots \cdot k_k}\right) = \delta\sigma.$$

Наслідок. *За умов теореми 1 для повної похибки має місце оцінка*

$$\|\tilde{f}(\tilde{x}) - f(x)\| \leqslant k\sigma\delta \text{ при } \delta \to 0.$$

Доведення. З теореми 1 випливає, що

$$\|\tilde{f}(\tilde{x}) - f(x)\| = \|f(\hat{x}) - f(x)\| \leqslant k\|\hat{x} - x\| \leqslant k\tilde{\delta} \leqslant k\sigma\delta. \qquad (16)$$

У цій оцінці (на відміну від (15)) враховується як вплив обумовленості задачі (через k), так і вплив похибок в алгоритмі (через індикатор стабільності σ), що відповідає нашим інтуїтивним уявленням. До того ж твердження теореми залишається правильним для всіх типів похибок та їхніх оцінок (абсолютної, відносної, за нормою, покомпонентної), причому не потребує субмультиплікативного числа обумовленості задачі k.

Означення 5. Алгоритм $f = f_k \circ \ldots \circ f_1$ називається *стійким*, якщо $\sigma \approx 1$, *слабо стійким*, якщо $\sigma \leqslant k + 1$, і *нестійким*, якщо $\sigma \gg k$.

Зазначимо, що при $k_j \geqslant 1 \ \forall \ j = \overline{1, k}$, маємо $\sigma \leqslant k + 1$, тобто якщо проміжні відображення f_j підсилюють похибку, то алгоритм є принаймні слабо стійким. Для таких алгоритмів допускається робити на кожному кроці похибку порядку похибки вхідних даних.

Приклад 8. *Обчислення скалярного добутку.* Для обчислення скалярного добутку

$$f(x, y) = \langle x, y\rangle = \sum_{i=1}^{n} x_i y_i, \ y, \ x \in \mathbb{R}^n,$$

застосуємо такий алгоритм:

$$f : \mathbb{R}^n \times \mathbb{R}^n \xrightarrow{f_1} \mathbb{R}^n \xrightarrow[+]{f_2} \mathbb{R}^{n-1} \xrightarrow[+]{f_3} \ldots \xrightarrow[+]{f_n} \mathbb{R},$$

де

$$f_1(x, y) = (x_1 y_1, \ldots, x_n y_n)^T,$$

$$f_j(z_1, \ldots, z_{n-j+2}) = (z_1 + z_2, z_3, \ldots, z_{n-j+2})^T, \quad j \geqslant 2.$$

Відносна покомпонентна похибка на кожному кроці обмежена величиною $\delta = \mathrm{eps}$, а для частинних чисел обумовленості, як ми бачили раніше, виконуються нерівності $k_1 = 2$, $k_j \geqslant 1$, $j \geqslant 2$. Тому для індикатора стабільності маємо

$$\sigma = 1 + \sum_{j=1}^{n} \prod_{l=1}^{j} \frac{1}{k_l} \leqslant 1 + \frac{n}{2} = \frac{n+2}{2},$$

що означає слабку стійкість алгоритму.

Розглянемо випадок, коли хоча б один з проміжних кроків має число обумовленості $k_j \ll 1$. Цю ситуацію називають «втратою інформації». Якщо попередні кроки алгоритму не компенсують її в тому розумінні, що $k_1 \cdot \ldots \cdot k_j \geqslant 1$, то це призводить до нестабільності алгоритму. Вираз для σ чітко вказує на те, що «втрату інформації» не можна компенсувати після її втрати.

Приклад 9. *Обчислення* $\cos mx$. У гармонічному аналізі часто доводиться обчислювати величини $\cos mx$. Для цього, на перший погляд, придатна рекурентна формула

$$c_{k+1} = 2\cos x \cdot c_k - c_{k-1}, \quad k = 1, 2, \ldots; \quad c_0 = 1, \ c_1 = \cos x, \qquad (17)$$

яка випливає з відомої формули

$$\cos(k+1)x = 2\cos x \cos kx - \cos(k-1)x.$$

Для використання цієї формули потрібно обчислювати відображення $g(x) = \cos x$, яке, очевидно, має числа обумовленості

$$k_{\mathrm{abs}} = \sin x, \quad k_{\mathrm{rel}} = x \operatorname{tg} x.$$

Якщо x близьке до 0 чи до π, то k_{rel} близьке до нуля і ми маємо «втрату інформації». В індикаторі стабільності алгоритму (10) в кожен доданок входитиме множник

$$\frac{1}{k_{\mathrm{rel}}} = \frac{1}{x \operatorname{tg} x},$$

який прямує до нескінченності, якщо x прямує до 0 або до π. Отже, при x близьких до 0 чи π алгоритм (17) є нестійким. Цю нестійкість можна усунути, застосувавши *алгоритм Райнша*

$$\Delta c_k = -4\sin^2\frac{x}{2} c_k + \Delta c_{k-1},$$

$$c_{k+1} = c_k + \Delta c_k, \quad k = 1, 2, \ldots; \quad c_0 = 1, \quad \Delta c_0 = -2\sin^2\frac{x}{2}.$$

У цьому алгоритмі обчислення відображення $h(x) = \sin^2\frac{x}{2}$ для малих $x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ не призведе до нестабільності, бо

$$\frac{1}{k(x, h)} = \left|\frac{\operatorname{tg}(x/2)}{x}\right| \to \frac{1}{2} \text{ при } x \to 0.$$

Аналогічно можна стабілізувати алгоритм і для $x \to \pi$. Зауважимо, що, як показує оцінка (16), на стабільність кінцевого результату впливає ще і число обумовленості всієї задачі (17), але можна довести, що вона є добре обумовленою.

Вигляд індикатора стійкості вказує і на те, що на його величину впливає *порядок виконання проміжних операцій.*

Приклад 10. *Обчислення варіацій.* У статистиці досить часто треба обчислювати величину

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

де $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ — середнє значення.

Найбільший інтерес для практики становить ситуація, коли числа $\bar{x}$ та x_i, $i = \overline{1, n}$, є близькими. Проаналізуємо, який з алгоритмів

а) $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$;

б) $S^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)$

в цій ситуації буде стійким. Індикатори стабільності мають вигляд

$$\sigma^{(a)} = 1 + \Sigma\Pi \frac{1}{k_i^{(a)}}, \quad \sigma^{(б)} = 1 + \Sigma\Pi \frac{1}{k_i^{(б)}}.$$

Віднімання в обох формулах є погано обумовленим, бо віднімаються близькі числа, тобто $k_{\text{відн}} \gg 1$. Тому у формулі для $\sigma^{(a)}$ в кожному доданку присутній множник

$$\frac{1}{k_{\text{відн}}} \ll 1,$$

що приводить до $\sigma^{(a)} \approx 1$. В $\sigma^{(б)}$ цей множник присутній лише в останньому доданку, тому $\sigma^{(б)} \gg 1$. Отже алгоритм а) є більш стійким, ніж б), хоча в ньому виконується n операцій віднімання замість однієї в алгоритмі б).

ГЛАВА 1

МАТЕМАТИЧНИЙ АПАРАТ ТЕОРІЇ ЧИСЕЛЬНИХ МЕТОДІВ

Класичний курс математичного аналізу починається з введення поняття функції, означення операцій диференціювання функцій, інтегрування і т. п., а потім на основі цих означень будується вся теорія. Природно діяти так і при побудові теорії чисельних методів. Тому далі вводяться поняття про сіткову функцію, або функцію дискретного аргументу, про скінченні та розділені різниці (дискретний аналог диференціювання), підсумовування (дискретний аналог інтегрування). Далі розглядаються деякі властивості цих операцій, наводяться додаткові відомості з різних розділів математики і потім будується теорія чисельних методів. Зауважимо, що хоча в неперервному випадку диференціювання та інтегрування якраз і визначаються через різниці та суми, наслідки цих означень виявляються часто простішими, ніж у дискретному аналізі, в чому ми матимемо змогу переконатися пізніше.

1.1. Сіткові функції та операції над ними

1.1.1. Сіткові функції. Сіткова функція, або функція дискретного аргументу,— це функція $y = y(x)$, $x \in \omega$, задана на дискретній множині чисел (точок) $\omega = \{x_i : i = 0, \pm 1, \pm 2, \ldots; x_i < x_{i+1} \forall i\}$. Множина ω називається *сіткою*, а точки x_i — *вузлами сітки.*

Якщо вузли сітки задаються формулою $x_i = x_0 + ih$, $i = 0, \pm 1, \ldots, h = \text{const}$, то сітка називається *рівномірною.* Часто використовуються сітки, вузли яких розміщені на деякому відрізку $[a, b]$: рівномірні сітки $\omega_h = \{x_i = a + ih : i = \overline{1, N-1}, h = (b-a)/N\}$, $\bar{\omega}_h = \left\{x_i = a + ih : i = \overline{0, N}, h = \frac{b-a}{N}\right\}$ та *нерівномірні* сітки $\hat{\omega}_h = \{x_i : x_i \in (a, b), i = \overline{1, N-1}, x_i < x_{i+1} \forall i = \overline{1, N-2}\}$, $\bar{\hat{\omega}}_h = \{x_i : x_i \in [a, b], i = \overline{0, N}, x_0 = a, x_N = b, x_i < x_{i+1}, i = \overline{0, N-1}\}$. Сітки ω_h, $\hat{\omega}_h$ та $\bar{\omega}_h$, $\bar{\hat{\omega}}_h$ є дискретними аналогами відповідно інтервалу (a, b) та відрізка $[a, b]$. Індекс h вказує, що сітка ω_h належить деякій послідовності сіток $\{\omega_h\}_{h>0}$, а якщо сітка нерівномірна, то під h розуміємо вектор $h = (h_2, h_3, \ldots, h_{N-1})$, $h_i = x_i - x_{i-1}$. Щоб підкреслити, що сітка має скінченну кількість вузлів, іноді позначають $\omega_N = \{x_i : x_i \in (a, b), i = \overline{1, N-1}, x_i < x_{i+1}\}$,

$\bar{\omega}_N = \{x_i : x_i \in [a, b],\ i = \overline{0, N},\ x_0 = a,\ x_N = b,\ x_i < x_{i+1}\}$. Якщо вузли сітки розміщені на проміжку $[a, b]$, то кажуть також, що сітка покриває відрізок $[a, b]$. Вузли сітки $\omega_h = \{x_i = a + ih : i = \overline{1, N-1}, h = (b-a)/N\}$ та вузли сітки $\bar{\omega}_h$, які збігаються з вузлами ω_h, називаються *внутрішніми* вузлами, а вузли $x_0 = a$, $x_N = b$ сітки $\bar{\omega}_h$ називаються *граничними*. Аналогічні терміни прийнято також для нерівномірних сіток $\hat{\omega}_h$ та $\tilde{\tilde{\omega}}_h$, що покривають відповідно інтервал (a, b) та проміжок $[a, b]$.

Будь-яку сіткову функцію $y(x_i) \equiv y_i$, задану на скінченній сітці $\bar{\omega}_h$, можна подати у вигляді вектора $y = (y_0, \ldots, y_N)$ розмірності $N+1$. Тому множина таких сіткових функцій утворює скінченновимірний, точніше $(N+1)$-вимірний, простір H. Оскільки сітка ω_h залежить від параметра h, то і сіткова функція $y(x)$, $x \in \bar{\omega}_h$, залежить від параметра h, що іноді позначається так: $y(x) = y_h(x)$, $x \in \omega_h$. Тому і простір H сіткових функцій природно позначати як H_h. Оскільки кожний вузол сітки x_i взаємно однозначно зв'язаний з його номером i, то сіткову функцію можна розглядати як функцію цілочисельного аргументу i: $y(x) \equiv y(x_i) \equiv y(i) \equiv y_i$, $i = 0, \pm 1, \ldots$, $x \in \omega_h$. Зокрема функцію, задану на сітці $\bar{\omega}_h$, можна розглядати як функцію, задану на сітці $\bar{\omega}_N = \{i : i = \overline{0, N}\}$. Перехід від однієї сітки до іншої очевидний, і ми їх не розрізнятимемо. Якщо функція неперервного аргументу $y = y(x)$ задана на відрізку $[a, b]$, який покривається деякою сіткою $\tilde{\tilde{\omega}}_h$, то цій функції природно можна поставити у відповідність сіткову функцію $y_h(x) = y(x)$, $x \in \tilde{\tilde{\omega}}_h$, яка називається проекцією функції неперервного аргументу $y(x)$, $x \in [a, b]$, на сітку $\tilde{\tilde{\omega}}_h$.

Для сіткових функцій, як і для функцій неперервного аргументу, можна ввести операції додавання і множення на число, утворивши тим самим з простору сіткових функцій лінійний простір сіткових функцій. Аналогічно просторам функцій неперервного аргументу в лінійному просторі сіткових функцій можна ввести додаткові «інструменти» для вивчення цих просторів: скалярний добуток, норму і т. п.

1.1.2. Різницеві аналоги операцій диференціювання та інтегрування. У просторі сіткових функцій, заданих на рівномірній сітці ω_h, аналогом першої похідної є різницеві похідні першого порядку:
$y_{x,i} = (y_{i+1} - y_i)/h$ — права різницева похідна у вузлі x_i;
$y_{\bar{x},i} = (y_i - y_{i-1})/h$ — ліва різницева похідна у вузлі x_i;
$y_{\mathring{x},i} = (y_{i+1} - y_{i-1})/(2h)$ — центральна різницева похідна у вузлі x_i.

Іноді як аналоги першої похідної використовуються скінченні різниці першого порядку:

$\Delta y_i = y_{i+1} - y_i$ — права різниця в точці x_i;
$\nabla y_i = y_i - y_{i-1}$ — ліва різниця в точці x_i;
$\delta y_i = \frac{1}{2}(y_{i+1} - y_{i-1}) = \frac{1}{2}(\Delta y_i + \nabla y_i)$ — центральна різниця в точці x_i. Мають місце очевидні формули

$$\sum_{j=k}^{i} \Delta y_j = y_{i+1} - y_k, \quad \sum_{j=k}^{i} \nabla y_j = y_i - y_{k-1}, \tag{1}$$

які є різницевими аналогами формули

$$\int_a^b f'(x)\,dx = f(b) - f(a).$$

Для довільних функцій y_i, v_i цілочисельного аргументу справедливі формули

$$\Delta(y_i v_i) = y_i \Delta v_i + v_{i+1} \Delta y_i = y_{i+1} \Delta v_i + v_i \Delta y_i, \tag{2}$$

$$\nabla(y_i v_i) = y_{i-1} \nabla v_i + v_i \nabla y_i = y_i \nabla v_i + v_{i-1} \nabla y_i, \tag{3}$$

які перевіряються безпосередньо. Наприклад,

$$\Delta(y_i v_i) = y_{i+1} v_{i+1} - y_i v_i = y_{i+1} v_{i+1} - y_{i+1} v_i + y_{i+1} v_i - y_i v_i =$$
$$= y_{i+1}(v_{i+1} - v_i) + v_i(y_{i+1} - y_i) = y_{i+1} \Delta v_i + v_i \Delta y_i.$$

Формули (2), (3) є аналогами формули диференціювання добутку

$$(y(x)\,v(x))' = yv' + vy'.$$

Можна знайти різницеві аналоги і інших формул диференціювання. Розглянемо, наприклад, формулу

$$(y^2(x))' = 2y(x)\,y'(x).$$

Для різницевого аналогу маємо

$$\Delta y_i^2 = y_{i+1}^2 - y_i^2 = (y_{i+1} - y_i)(y_{i+1} + y_i) = (y_{i+1} + y_i)\,\Delta y_i. \tag{4}$$

Аналогом формули інтегрування частинами

$$\int_a^b y(x)\,dv(x) = y(x)\,v(x)\Big|_a^b - \int_a^b v(x)\,dy(x) = y(b)\,v(b) -$$
$$- y(a)\,v(a) - \int_a^b v(x)\,dy(x)$$

є формула підсумовування частинами

$$\sum_{i=0}^{N-1} y_i \Delta v_i = -\sum_{i=1}^{N} v_i \nabla y_i + (yv)_N - (yv)_0, \tag{5}$$

яку також записують у вигляді

$$\sum_{i=1}^{N-1} y_i \Delta v_i = -\sum_{i=1}^{N-1} v_j \nabla y_i + y_{N-1} v_N - y_0 v_1. \qquad (6)$$

Для доведення формули (5) скористаємося формулою (2):

$$y_i \Delta v_i = \Delta (y_i v_i) - v_{i+1} \Delta y_i = \Delta (y_i v_i) - v_{i+1} \nabla y_{i+1}.$$

Враховуючи, що $\Delta y_i = \nabla y_{i+1}$, а також формулу (1), маємо

$$\sum_{i=0}^{N-1} y_i \Delta v_i + \sum_{i=1}^{N} v_i y_j = \sum_{i=0}^{N-1} \Delta (y_i v_i) = (yv)_N - (yv)_0,$$

що і дає формулу (5). Якщо, $y_0 = y_N = 0$, то з (5) дістаємо

$$\sum_{i=0}^{N-1} y_i \Delta v_i = -\sum_{i=1}^{N} v_i \nabla y_i. \qquad (7)$$

Аналогічно тому, як формула інтегрування частинами використовується для обчислення інтегралів, формула підсумовування частинами використовується для обчислення сум.

Приклад 1. Обчислити суму $S_N = \sum_{i=1}^{N} i a^i$, $a \neq 1$.

Розв'язання. Покладемо $v_i = i$, $\Delta y_i = a^i$. Тоді

$$y_i = y_{i-1} + a^i = y_{i-2} + a^{i-1} + a^i = \cdots = y_0 + \sum_{j=1}^{i} a^j.$$

Остання сума є сумою геометричної прогресії із знаменником a і дорівнює

$$\sum_{j=1}^{i} a^j = \frac{a^{i+1} - a}{a - 1}.$$

Тому

$$y_i = y_0 + \frac{a^{i+1} - a}{a - 1}. \qquad (8)$$

Зауважимо, що y_i визначається неоднозначно, і вибираючи різні y_0, ми дістаємо різні y_i. Виберемо y_0 так, щоб $y_N = 0$. Покладаючи в (8) $i = N$ і прирівнюючи y_N до нуля, матимемо

$$y_0 = -\frac{a^{N+1} - a}{a - 1}.$$

Легко помітити, що $v_0 = 0$, $\Delta v_i = 1$. Тому з (5) дістанемо

$$S_N = \sum_{i=1}^{N} i a^i = \sum_{i=1}^{N} v_i \nabla y_i = -\sum_{i=0}^{N-1} y_i \Delta v_i = -\sum_{i=0}^{N-1} y_i =$$

$$= -\sum_{i=0}^{N-1} \left(y_0 + \frac{a^{i+1} - a}{a - 1} \right) = N \frac{a^{N+1} - a}{a - 1} - \frac{1}{a - 1} \sum_{i=0}^{N-1} a^{i+1} +$$

$$+ N \frac{a}{a - 1} = \frac{N a^{N+2} - (N + 1) a^{N+1} + a}{(a - 1)^2}.$$

Приклад 2. Обчислити суму $S_N = \sum_{i=1}^{N} i^2$.

Розв'язання. Покладемо $v_i = i^2$, $\nabla y_i = 1$. Тоді $y_i = y_{i-1} + 1 = y_{i-2} + 1 + 1 = y_0 + i$. Покладаючи $y_0 = -N$, дістаємо $y_N = 0$. Далі

$$\Delta v_i = (i + 1)^2 - i^2 = (i + 1 - i)(i + 1 + i) = 2i + 1, \quad v_0 = 0.$$

Застосовуючи формулу (5) та формулу для суми арифметичної прогресії, маємо

$$S_N = \sum_{i=1}^{n} v_i \nabla y_i = -\sum_{i=0}^{N-1} y_i \Delta v_i = -\sum_{i=0}^{N-1} (i - N)(2i + 1) =$$

$$= -2 \sum_{i=0}^{N-1} i^2 - \sum_{i=0}^{N-1} i + 2N \sum_{i=0}^{N-1} i + N \sum_{i=0}^{N-1} 1 =$$

$$= -2 \left(\sum_{i=1}^{N} i^2 - N^2 \right) - \frac{N}{2} (N - 1) + N^2 (N - 1) + N^2 =$$

$$= -2S_N + \frac{N}{2} (6N - N + 1 + 2N^2 - 2N) = -2S_N + \frac{N (N + 1) (2N + 1)}{2},$$

звідки

$$S_N = \frac{N (N + 1) (2N + 1)}{6}.$$

1.2. Методи розв'язування алгебраїчних рівнянь

1.2.1. Метод Гаусса для розв'язування систем лінійних алгебраїчних рівнянь. *LR*-розвинення матриць. Розглянемо метод розв'язування системи лінійних алгебраїчних рівнянь (СЛАР)

$$\begin{aligned} a_{11} x_1 + a_{12} x_2 + \cdots + a_{1n} x_n &= b_1, \\ a_{21} x_1 + a_{22} x_2 + \cdots + a_{2n} x_n &= b_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1} x_1 + a_{n2} x_2 + \cdots + a_{nn} x_n &= b_n, \end{aligned} \qquad (1)$$

який вперше був опублікований німецьким математиком Карлом Фрідріхом Гауссом в його праці «Disqussitiones arithmetical» (1801 р.) Систему (1) запишемо у вигляді

$$Ax = b, \qquad (1')$$

де $A \in \mathrm{Mat}_n (\mathbb{R})$, тобто A є матрицею з множини $\mathrm{Mat}_n (\mathbb{R})$ всіх $(n \times n)$-матриць з дійсними елементами; $b \in \mathbb{R}^n$ — заданий вектор; $x \in \mathbb{R}^n$ — шуканий вектор.

З лінійної алгебри відомо, що за умови $\det A \neq 0$, тобто $A \in GL(n)$ ($GL(n)$ — множина всіх невироджених $(n \times n)$-матриць), система (1) має єдиний розв'язок, який можна знайти за відомим

правилом Крамера. Проте це правило має скоріше теоретичне значення, бо лише обчислення $\det A$ безпосередньо потребує $n \cdot n!$ операцій множення (з використанням розвинення за теоремою Лапласа — 2^n операцій) і розв'язування всієї системи є надзвичайно трудомістким.

Перш ніж розглянути алгоритм Гаусса в загальному випадку, розглянемо окремі випадки системи (1) з невиродженими верхньою трикутною матрицею

$$Rx = z;$$

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ & r_{22} & \dots & r_{2n} \\ & \cdot & & \cdot \\ & & \cdot & \cdot \\ & & & \cdot \\ 0 & & & r_{nn} \end{bmatrix} \tag{2}$$

(z — відомий, x — шуканий вектор) та нижньою трикутною матрицею $Lz = b$ (b — відомий, z — невідомий вектор),

$$L = \begin{pmatrix} l_{11} & & 0 \\ l_{21} & l_{22} & \\ \cdot\cdot & \cdot\cdot & \\ l_{n1} & l_{n2} & \dots\ l_{nn} \end{pmatrix}. \tag{3}$$

З невиродженості цих матриць випливає, що $r_{ii} \neq 0$, $l_{ii} \neq 0$ $\forall\, i = \overline{1, n}$, і тому їхні розв'язки легко знаходяться за формулами

$$z_1 = b_1/l_{11}, \quad z_i = \left(b_i - \sum_{j=1}^{i-1} l_{ij} z_j\right)/l_{ii}, \quad i = \overline{2, n}, \tag{4}$$

для системи (3) та

$$x_n = z_n/r_{nn}, \quad x_i = \left(z_i - \sum_{j=i+1}^{n} r_{ij} x_j\right)/r_{ii}, \quad i = n-1, n-2, \dots, 1, \tag{5}$$

для системи (2).

Алгоритм (4) називається *прямою підстановкою*, алгоритм (5) — *зворотною підстановкою*. Неважко підрахувати, що для обох алгоритмів кількість операцій множення та додавання (їх значно більше, ніж ділень) дорівнює

$$\sum_{i=1}^{n} (i-1) = \frac{n(n-1)}{2} \sim \frac{n^2}{2}.$$

Ці алгоритми не потребують додаткової пам'яті ЕОМ, бо, наприклад, після обчислення z_1 в алгоритмі (4) це число можна розмістити на місці b_1, z_2 — на місці b_2 і т. д. Ідея алгоритму Гаусса полягає тепер в тому, щоб розв'язування системи лінійних алгебраїчних рівнянь загального вигляду (1) звести до прямої та зворотної підстановок. Для цього на першому кроці спробуємо систему (1) еквівалентним перетворенням звести до вигляду (якщо це можливо!)

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1, \\ a_{22}^{(2)}x_2 + \dots + a_{2n}^{(2)}x_n &= b_2^{(2)}, \\ \dots\dots\dots\dots\dots\dots & \\ a_{n2}^{(2)}x_2 + \dots + a_{nn}^{(2)}x_n &= b_n^{(2)}, \end{aligned} \tag{6}$$

далі цю систему знову ж таки еквівалентним перетворенням зводимо до вигляду

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1, \\ a_{22}^{(2)}x_2 + \dots + a_{2n}^{(2)}x_n &= b_2^{(2)}, \\ a_{33}^{(3)}x_3 + \dots + a_{3n}^{(3)}x_n &= b_3^{(3)}, \\ \dots\dots\dots\dots\dots\dots & \\ a_{n3}^{(3)}x_3 + \dots + a_{nn}^{(3)}x_n &= b_n^{(3)} \end{aligned} \tag{7}$$

і т. д., поки не дістанемо систему з верхньою трикутною матрицею, яку потім розв'яжемо за допомогою зворотної підстановки.

Щоб виключити x_1 з усіх рівнянь, крім першого, і дістати систему (6), виконаємо над i-м рядком таке перетворення:

новий i-й рядок=старий i-й рядок $-$ l_{i1} $\cdot$ *перший рядок,* де l_{i1} — деяке число, тобто

$$(a_{i1} - l_{i1}a_{11})x_1 + (a_{i2} - l_{i1}a_{12})x_2 + \dots + (a_{in} - l_{i1}a_{1n})x_n = b_i - l_{i1}b_1.$$

З умови $a_{i1} - l_{i1}a_{11} = 0$ дістаємо (якщо $a_{11} \neq 0$)

$$l_{i1} = a_{i1}/a_{11}.$$

Після виконання цього перетворення для всіх рядків $i = \overline{2, n}$, позначивши $a_{ij}^{(2)} = a_{ij} - l_{i1}a_{1j}$, $i = \overline{2, n}$, $j = \overline{2, n}$, дістанемо систему (6). При цьому елемент $a_{11} \neq 0$ називається *головним елементом,* а перший рядок — *головним рядком.* Якщо $a_{22}^{(2)} \neq 0$, то далі описану процедуру можна повторити для $n-1$ рівнянь, крім першого, і дістати системи (7). Іншими словами, знайдемо послідовність матриць

$$A = A^{(1)} \to A^{(2)} \to \dots \to A^{(n)} = R,$$

де

$$A^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ & \dots & & \dots \\ & a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ & \dots & & \dots \\ & a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{bmatrix}.$$

Над цією матрицею далі виконуємо крок виключення за формулами

$$l_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}, \quad i = k+1, \ldots, n,$$
$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)}, \quad i, j = k+1, \ldots, n,$$
$$b_i^{(k+1)} = b_i^{(k)} - l_{ik}b_k^{(k)}, \quad i = k+1, \ldots, n,$$

якщо головний елемент $a_{kk}^{(k)}$ не дорівнює нулю. Легко помітити, що перехід від $A^{(k)}$, $b^{(k)}$ до $A^{(k+1)}$, $b^{(k+1)}$ можна записати в матричному вигляді

$$A^{(k+1)} = L_k A^{(k)}, \; b^{(k+1)} = L_k b^{(k)},$$

де нижня трикутна матриця

$$L_k = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{k+1,k} & 1 & \\ & & \cdots & & \ddots \\ & & -l_{n,k} & & 1 \end{bmatrix}$$

називається також *матрицею Фробеніуса*. Неважко перевірити такі властивості матриць Фробеніуса:

$$L_k^{-1} = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{k+1,k} & 1 & \\ & & \cdots & & \ddots \\ & & l_{n,k} & & 1 \end{bmatrix},$$

$$L = L_1^{-1} \ldots L_{n-1}^{-1} = \begin{bmatrix} 1 & & & & 0 \\ l_{21} & 1 & & & \\ l_{31} & l_{32} & 1 & & \\ \cdots & & & & \\ l_{n1} & l_{n2} & l_{n3} & \ldots & l_{n,n-1} \quad 1 \end{bmatrix}.$$

Таким чином, ми звели систему $Ax = b$ до $Rx = z$, де

$$R = L^{-1}A, \; z = L^{-1}b,$$

причому R є *верхньою трикутною матрицею*, а L — *унітотентною*, тобто *нижньою трикутною матрицею* з одиницями на головній діагоналі. Розвинення матриці A у вигляді

$$A = LR,$$

де L — *унітотентна трикутна матриця*, R — верхня трикутна матриця, називається *трикутним розвиненням Гаусса*, або *LR-розвиненням матриці A*.

Вправа 1. Доведіть, що LR-розвинення матриці $A \in \mathrm{Mat}_n(\mathbb{R})$ єдине, якщо воно існує.

Алгоритм виключення Гаусса тепер можна записати таким чином.

Алгоритм 1. *Метод виключення Гаусса.*

1. Виконати LR-розвинення матриці A.
2. Виконати пряму підстановку $Lz = b$.
3. Виконати обернену підстановку $Rx = z$. Цей алгоритм використовує лише ту пам'ять ЕОМ, в якій спочатку містилися матриця A та вектор b, тобто $n(n+1)$ чисел, оскільки стовпчики матриці L (крім одиниць) можна зберігати на місці елементів, які перетворюються в нуль. Неважко також підрахувати, що на кроці 1 алгоритму Гаусса виконується $\sim \sum_{j=1}^{n-1} j^2 \approx n^3/3$, а на кожному з кроків 2 та $3 \sim \sum_{j=1}^{n-1} j \approx n^2/2$ операцій множення, тобто в цілому виконується $O(n^3/3)$ операцій множення. Зауважимо, що вказаний варіант алгоритму виключення Гаусса доцільно використовувати, коли потрібно розв'язати кілька систем лінійних алгебраїчних рівнянь з тією самою матрицею і різними правими частинами. При цьому найтрудомісткіший крок 1 досить виконати лише один раз. Звернемо увагу на такі недоліки алгоритму 1. По-перше, найпростіший приклад матриці

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \det A = -1, a_{11} = 0$$

показує, що LR-розвинення можливе не для всіх невироджених матриць, хоча проста перестановка рядків або стовпчиків усуває цей недолік:

$$\overline{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = LR, \quad L = R = I.$$

По-друге, слід чекати неприємностей при діленні на малий головний елемент, на що вказує наступний приклад Дж. Форсайта.

Приклад 1. Нехай система

$$\begin{cases} 1{,}00 \cdot 10^{-4}x_1 + 1{,}00x_2 = 1{,}00, \\ 1{,}00x_1 + 1{,}00x_2 = 2{,}0 \end{cases}$$

розв'язується методом Гаусса на гіпотетичній ЕОМ, яка оперує тризначними десятковими числами. Розв'язавши систему з точністю до чотирьох значущих цифр, знайдемо «точний» розв'язок

$$x_1 = 1{,}000, \quad x_2 = 0{,}9999,$$

що після заокруглення до трьох цифр дає

$$x_1 = 1{,}00, \quad x_2 = 1{,}00.$$

На гіпотетичній ЕОМ дістаємо

$$l_{21} = a_{21}/a_{11} = 1{,}00/(1{,}00 \cdot 10^{-4}) = 1{,}00 \cdot 10^4,$$

$$(1{,}00 - 1{,}00 \cdot 10^4 \cdot 1{,}00 \cdot 10^{-4})\, x_1 + (1{,}00 - 1{,}00 \cdot 10^4 \cdot 1{,}00)\, x_2 =$$
$$= 2{,}00 - 1{,}00 \cdot 10^4 \cdot 1{,}00,$$

а також систему з трикутною матрицею

$$\begin{cases} 1{,}00 \cdot 10^{-4} x_1 + 1{,}00 x_2 = 1{,}00, \\ -1{,}00 \cdot 10^4 x_2 = -1{,}00 \cdot 10^4, \end{cases}$$

в якої знаходимо наближений розв'язок $x_2 = 1{,}00$, $x_1 = 0{,}00$.

Останнє число повністю не відповідає дійсності. Поміняв рівняння місцями і змінивши таким чином головний елемент, маємо

$$\begin{cases} 1{,}00 x_1 + 1{,}00 x_2 = 2{,}00, \\ 1{,}00 \cdot 10^{-4} x_1 + 1{,}00 x_2 = 1{,}00, \end{cases}$$

$$\tilde{l}_{21} = 1{,}00 \cdot 10^{-4},$$

$$\begin{cases} 1{,}00 x_1 + 1{,}00 x_2 = 2{,}00, \\ 1{,}00 x_2 = 1{,}00, \end{cases}$$

звідки

$$x_2 = 1{,}00, \quad x_1 = 1{,}00,$$

що відповідає дійсності.

Звідси випливає така стратегія вибору головного елемента: на кожному кроці алгоритму Гаусса за головний слід вибирати рядок з найбільшим за модулем елементом у головному стовпчику.

Алгоритм 2. *Метод виключення Гаусса з вибором головного елемента в стовпчику.*

1. На кроці $A^{(k)} \to A^{(k+1)}$ вибрати $p \in \{k, \ldots, n\}$, для якого

$$|a_{pk}^{(k)}| \geqslant |a_{jk}^{(k)}| \; \forall\, j = \overline{k,\, n}.$$

Рядок p — головний.

2. Поміняти функціями рядки p та k:

$$A^{(k)} \to \tilde{A}^{(k)} = (\tilde{a}_{ij}^{(k)}),\ \tilde{a}_{ij}^{(k)} = \begin{cases} a_{kj}^{(k)}, & i = p, \\ a_{pj}^{(k)}, & i = k, \\ a_{ij}^{(k)}, & i \neq p,\ i \neq k \end{cases}$$

(при цьому не обов'язково виконувати пересилки в пам'яті ЕОМ!). Тепер

$$|\tilde{l}_{ik}| = \left| \frac{\tilde{a}_{ik}^{(k)}}{\tilde{a}_{kk}^{(k)}} \right| = \left| \frac{\tilde{a}_{ik}^{(k)}}{\tilde{a}_{pk}^{(k)}} \right| \leqslant 1.$$

3. Виконати такий крок виключення відносно матриці $\tilde{A}^{(k)}$:

$$\tilde{A}^{(k)} \to A^{(k+1)}.$$

Легко помітити, що крім вибору головного елемента в стовпчику можлива стратегія вибору головного елемента в рядку (в гіршому випадку обидві ці стратегії потребують додатково $O(n^2)$ операцій), а також стратегія пошуку головного елемента по всій матриці (додатково $O(n^3)$ операцій). Остання через свою трудомісткість використовується рідко. Опишемо формально LR-розвинення з вибором головного елемента в стопчику. Нехай π — деяка перестановка з множини $\sigma_n = \{1, \ldots, n\}$ всіх можливих перестановок чисел $1, \ldots, n$. У відповідність їй поставимо матрицю перестановок

$$P_\pi = [e_{\pi(1)}, \ldots, e_{\pi(n)}],$$

де $e_j = (\delta_{1j}, \ldots, \delta_{nj})^T$ — j-й координатний вектор. Перестановка π рядків матриці A тепер може бути описана таким чином:

$$A \to P_\pi A,$$

а перестановка стовпчиків π — виразом

$$A \to A P_\pi.$$

Неважко переконатися, що

$$P_\pi^{-1} = P^T, \quad \det P_\pi = \operatorname{sign} \pi,$$

тобто дорівнює $+1$, якщо перестановка π є парною, і -1, якщо перестановка π є непарною.

Наведемо теорему, яка вказує на те, що метод виключення Гаусса з вибором головного елемента теоретично можна здійснити (тобто виконати до кінця) для будь-якої невиродженої матриці.

Теорема 1. *Для будь-якої невиродженої матриці A існує матриця перестановок P така, що матриця PA має LR-розвинення, тобто*

$$PA = LR,$$

причому P може бути вибрана так, що елементи матриці L за модулем не перевищуватимуть одиниці (символічно $|L| \leqslant 1$).

Доведення. Розглянемо для прикладу алгоритм Гаусса з вибором головного елемента в стовпчику. Оскільки $\det A \neq 0$, то існує така перестановка $\tau_1 \in \sigma_n$ (допускається $\tau_1 = \mathrm{id}$, тобто тотожна перестановка, яка нічого не змінює), що перший діагональний елемент $a_{11}^{(1)}$ матриці

$$A^{(1)} = P_{\tau_1} A$$

не дорівнює нулю і є найбільшим за модулем у першому стовпчику:

$$|a_{11}^{(1)}| \geqslant |a_{i1}^{(1)}| \; \forall\, i = \overline{1,\, n},\ a_{11}^{(1)} \neq 0.$$

Після першого кроку виключення дістаємо матрицю вигляду

$$A^{(2)} = L_1 A^{(1)} = L_1 P_{\tau_1} A = \left[\begin{array}{c|c} a_{11}^{(1)} & * \ \dots \ * \\ \hline 0 & \\ \dots & B^{(2)} \\ 0 & \end{array}\right],$$

причому $|L_1| \leqslant 1$, $\det L_1 = 1$. Оскільки $a_{11}^{(1)} \neq 0$, то

$$0 \neq \operatorname{sign}(\tau_1) \cdot \det A = \det A^{(2)} = a_{11}^{(1)} \det B^{(2)},$$

і тому $B^{(2)}$ є також невиродженою матрицею. Продовжуючи за індукцією далі, маємо

$$R = A^{(n)} = L_{n-1} P_{\tau_{n-1}} \dots L_1 P_{\tau_1} A,$$

де $|L_k| \leqslant 1$, а τ_k є або тотожною перестановкою, або транспозицією двох чисел, що не менші за k. Якщо $\pi \in \sigma_n$ — деяка перестановка, що переставляє лише числа, які більші або дорівнюють $k+1$, то для матриці Фробеніуса

$$L_k = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & -l_{k+1,k} & 1 & & \\ & \dots & & \ddots & \\ & -l_{n,k} & & & 1 \end{bmatrix}$$

маємо

$$\hat{L}_k \equiv P_\pi L_k P_\pi^{-1} = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{\pi(k+1),k} & 1 & \\ & & \dots & & \ddots \\ & & -l_{\pi(n),k} & & 1 \end{bmatrix}. \qquad (8)$$

Подамо матрицю R у вигляді

$$R = L_{n-1} P_{\tau_{n-1}} L_{n-2} P_{\tau_{n-1}}^{-1} P_{\tau_{n-1}} P_{\tau_{n-2}} L_{n-3} \dots L_1 P_{\tau_1} A =$$

$$= \hat{L}_{n-1} \dots \hat{L}_1 P_{\pi_0} A,$$

де $\hat{L}_k = P_{\pi_k} L_k P_{\pi_k}^{-1}$, $\pi_{n-1} = \mathrm{id}$, $\pi_k = \tau_{n-1} \dots \tau_{k+1}$ $\forall\, k = \overline{0, n-2}$. Перестановка π_k тут переставляє лише числа, які більші або дорівнюють $k+1$, тому матриці L_k мають вигляд (8). Звідси

$$P_{\pi_0} A = LR,$$

де

$$L = \hat{L}_1^{-1} \dots \hat{L}_{n-1}^{-1} = \begin{bmatrix} 1 & & & & \\ l_{\pi_1(2),1} & 1 & & & \\ l_{\pi_1(3),1} & l_{\pi_2(3),2} & 1 & & \\ \dots & \dots & \ddots & \ddots & \\ l_{\pi_1(n),1} & l_{\pi_2(n),2} & \dots & l_{\pi_{n-1}(n),n-1} & 1 \end{bmatrix},$$

$$|L| \leqslant 1.$$

Теорему доведено.

Внаслідок впливу похибок заокруглення на виході алгоритму Гаусса ми, як правило, дістанемо лише деякий наближений розв'язок $\tilde{x}$ системи (1). Якщо він нас не задовольняє, то можна було б повторити обчислення з подвійною точністю. Такий шлях підвищення точності розв'язку вимагає значних обчислювальних затрат, тому на практиці частіше застосовується так зване ітераційне уточнення розв'язку. Для цього використовується залишок

$$r(y) = b - Ay = A(x - y).$$

Абсолютна похибка $\Delta x_0 = x - x_0$, де x — точний розв'язок, а x_0 — наближений розв'язок, знайдений за методом Гаусса, задовольняє рівняння

$$A \Delta x_0 = r(x_0),$$

яке легко розв'язати, використовуючи вже обчислене LR-розвинення матриці A. Проте ми знову обчислимо лише деяке наближення $\tilde{\Delta} x_0 \neq \Delta x_0$, але слід чекати, що

$$x_1 = x_0 + \tilde{\Delta} x_0$$

є кращим наближенням точного розв'язку x, ніж x_0. Цей процес уточнення можна продовжити.

Метод прогонки розв'язування СЛАР з тридіагональною матрицею. Дуже часто при розв'язуванні крайових задач для звичайних диференціальних рівнянь другого порядку використовують заміну похідних відповідними різницевими співвідношеннями. Внаслідок цього дістають скінченно різницеву крайову задачу вигляду

$$a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \quad i = \overline{1, N-1}, \quad a_i \neq 0, \ b_i \neq 0,$$
$$y_0 = \varkappa_1 y_1 + \mu_1, \ y_N = \varkappa_2 y_{N-1} + \mu_2, \qquad (9)$$

яку можна записати у вигляді системи лінійних алгебраїчних рівнянь з тридіагональною матрицею розмірності $(N+1) \times (N+1)$

$$Ay = f, \qquad (10)$$

де

$$A=\begin{bmatrix} 1 & -\varkappa_1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ a_1 & -c_1 & b_1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & a_2 & -c_2 & b_2 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & a_3 & -c_3 & b_3 & \dots & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & a_{N-2} & -c_{N-2} & b_{N-2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & a_{N-1} & -c_{N-1} & b_{N-1} \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & -\varkappa_2 & 1 \end{bmatrix},$$

$y=(y_0, y_1, \dots, y_N)^T$, $f=(\mu_1, -f_1, \dots, -f_{N-1}, \mu_2)^T$. У випадку першої крайової задачі ($\varkappa_1=\varkappa_2=0$) за допомогою крайових умов $y_N=\mu_2$, $y_0=\mu_1$ можна виключити невідомі y_0 та y_N і дістати систему вигляду (10) з матрицею розмірності $(N-1)\times(N-1)$ вигляду

$$A=\begin{bmatrix} -c_1 & b_1 & 0 & 0 & \dots & 0 & 0 & 0 \\ a_2 & -c_2 & b_2 & 0 & \dots & 0 & 0 & 0 \\ 0 & a_3 & -c_3 & b_3 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & a_{N-2} & -c_{N-2} & b_{N-2} \\ 0 & 0 & 0 & 0 & \dots & 0 & a_{N-1} & -c_{N-1} \end{bmatrix},$$

правою частиною $f=(-f_1-a_1\mu_1, -f_2, \dots, -f_{N-2}, -f_{N-1}-b_{N-1}\mu_2)^T$ та невідомим вектором $y=(y_1, \dots, y_{N-1})^T$. Для розв'язування системи (10) або крайової задачі (9) існує ефективний метод, який є модифікацією методу Гаусса для СЛАР з тридіагональною матрицею — *метод прогонки.*

Ідея методу прогонки грунтується на такому спостереженні. Якщо прямий хід виконати за методом Гаусса для системи (10), то дістанемо матрицю з відмінними від нуля елементами лише на головній діагоналі та паралельній верхній діагоналі. Тому зворотний хід здійснюватиметься за формулою

$$y_i=\alpha_{i+1}y_{i+1}+\beta_{i+1}, \tag{11}$$

де α_{i+1}, β_{i+1} — деякі коефіцієнти. Щоб їх знайти, підставимо $y_{i-1}=\alpha_i y_i+\beta_i$ в (9):

$$(a_i\alpha_i-c_i)y_i+b_iy_{i+1}=-(f_i+a_i\beta_i).$$

Порівнюючи цю тотожність з (11), дістаємо

$$\alpha_{i+1}=\frac{b_i}{c_i-a_i\alpha_i}, \quad i=\overline{1, N-1}, \tag{12}$$

$$\beta_{i+1}=\frac{a_i\beta_i+f_i}{c_i-a_i\alpha_i}, \quad i=\overline{1, N-1}. \tag{13}$$

Для визначення α_1, β_1 використаємо крайову умову при $i=0$. Із формул (11) та (9) при $i=0$ знаходимо

$$\alpha_1=\varkappa_1, \quad \beta_1=\mu_1. \tag{14}$$

Визначивши α_1, β_1 за формулами (14), послідовно за допомогою рекурентних співвідношень (12), (13) знайдемо α_2, $\alpha_3, \dots, \alpha_{N-1}$ та β_2, $\beta_3, \dots, \beta_{N-1}$. Далі, щоб для визначення y_i можна було скористатися формулою (11), яка є також рекурентним співвідношенням (сітковим рівнянням) першого порядку, треба знати y_N. Визначимо його з крайової умови $y_N=\varkappa_2 y_{N-1}+\mu_2$ та з формули (11) при $i=N-1$: $y_{N-1}=\alpha_N y_N+\beta_N$. Із системи цих двох рівнянь знаходимо

$$y_N=\frac{\mu_2+\varkappa_2\beta_N}{1-\alpha_N\varkappa_2}. \tag{15}$$

Тепер за формулою (11) можемо послідовно знайти y_{N-1}, $y_{N-2}, \dots, y_0$.

Таким чином, метод прогонки зводиться до розв'язування трьох задач Коші для різницевих рівнянь першого порядку: 1) (12), (14); 2) (13), (14) — процес розв'язування цих двох задач Коші, в результаті якого знаходимо α_i, β_i, називається *прямим ходом методу прогонки*; 3) (11), (15) — це є *зворотний хід методу прогонки,* в результаті якого знаходимо розв'язок задачі (9).

Приклад 2. Розв'язати систему лінійних алгебраїчних рівнянь

$$Am=\bar{f},$$

де $m=(m_1, \dots, m_5)^T$, $\bar{f}=\left(0, 0, 0, 0, -\frac{1}{3}\right)^T$,

$$A=\begin{bmatrix} 3 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix}.$$

Розв'язання. Матриця системи тридіагональна, тому застосуємо метод прогонки. Сформулюємо задачу у вигляді (9). Для цього перепозначимо $y_i=m_{i+1}$, $i=\overline{0, 4}$. Тоді перше і останнє рівняння системи набудуть вигляду

$$3y_0+y_1=0, \quad y_3+3y_4=-\frac{1}{3},$$

а інші можна записати так:

$$y_{i-1}+4y_i+y_{i+1}=0, \quad i=\overline{1, 3},$$

тобто

$$a_i=b_i=1, \quad c_i=-4, \quad i=\overline{1, 3}, \quad N=4, \quad \varkappa_1=-\frac{1}{3},$$

$$\mu_1 = 0, \quad \varkappa_2 = -\frac{1}{3}, \quad \mu_2 = -\frac{1}{9},$$

$$f = (\mu_1, -f_1, -f_2, -f_3, \mu_2)^T = \left(0, 0, 0, 0, -\frac{1}{9}\right)^T.$$

Далі здійснюємо обчислення за формулами (12) — (14), заповнюючи таку таблицю:

i	1	2	3	4
$c_i - a_i\alpha_i$	$-\frac{11}{3}$	$-\frac{41}{11}$	$-\frac{153}{41}$	
α_i	$-\frac{1}{3}$	$-\frac{3}{11}$	$-\frac{11}{41}$	$\frac{41}{153}$
β_i	0	0	0	0

Тепер за формулою (15) маємо

$$y_4 = \frac{\mu_2 + \varkappa_2\beta_4}{1 - \alpha_4\varkappa_2} = \frac{-1/9 + 0}{1 - (-41/153)(-1/3)} = -\frac{51}{418}.$$

За формулою (11) знаходимо решту невідомих.

Стійкість методу прогонки. Формули прогонки можна застосовувати, коли знаменники в (12), (13), (15) не перетворюються на нуль. Доведемо, що достатніми умовами цього є нерівності

$$|c_i| \geqslant |a_i| + |b_i|, \quad i = \overline{1, N-1},$$
$$|\varkappa_1| \leqslant 1, \quad |\varkappa_2| \leqslant 1, \quad |\varkappa_1| + |\varkappa_2| < 2. \tag{16}$$

Доведемо, крім того, що за умов (16) виконуються нерівності

$$|\alpha_i| \leqslant 1, \quad i = \overline{1, N}, \tag{17}$$

які забезпечують стійкість розрахунків за формулою (11).

Неважко помітити, що в силу рівності $\alpha_1 = \varkappa_1$ та умов (16) при $i = 1$ має місце (17). Далі застосуємо метод математичної індукції. Припустимо, що (17) має місце для $i = k$, тобто $|\alpha_k| \leqslant 1$, і доведемо, що $|\alpha_{k+1}| \leqslant 1$. Для цього розглянемо ланцюжок нерівностей

$$|c_k - a_k\alpha_k| - |b_k| \geqslant |c_k| - |a_k||\alpha_k| - |b_k| \geqslant |a_k| + |b_k| - |a_k||\alpha_k| - |b_k| = |a_k|(1 - |\alpha_k|) \geqslant 0,$$

тобто

$$|c_k - a_k\alpha_k| \geqslant |b_k| > 0.$$

Але тоді

$$|\alpha_{k+1}| = |b_k| / |c_k - a_k\alpha_k| \leqslant 1,$$

що і треба було довести.

Доведемо, що знаменник у формулі (15) також не може перетворюватися на нуль за умов (16). Насамперед зауважимо, що коли $|c_{i0}| > |a_{i0}| + |b_{i0}|$ хоча б в одній точці $i = i_0$, то має місце строга нерівність $|\alpha_i| < 1$ для всіх $i > i_0$, в тому числі $|\alpha_N| < 1$. Тоді $|1 - \alpha_N\varkappa_2| \geqslant 1 - |\alpha_N|\,|\varkappa_2| \geqslant 1 - |\alpha_N| > 0$ і умова $|\varkappa_1| + |\varkappa_2| < 2$ є зайвою. Якщо $|\varkappa_1| < 1$, то $|\alpha_N| < 1$, і тоді знову знаменник в (15) відмінний від нуля. Якщо ж $|\varkappa_1| = 1$, то в силу (16) маємо $|\varkappa_2| < 1$, і оскільки $|\alpha_i| \leqslant 1$, то $|1 - \alpha_N\varkappa_2| \geqslant 1 - |\alpha_N|\,|\varkappa_2| \geqslant 1 - \varkappa_2 > 0$, що і треба було довести.

Обчислення за формулами (11) — (15) виконують на ЕОМ, розрядна сітка яких завжди скінченна, в результаті чого виникають похибки заокруглення. Фактично знаходять не розв'язок задачі (11) y_i, а деяку функцію $\tilde{y}_i$, яку можна розглядати як розв'язок тієї самої задачі, але із збуреними коефіцієнтами $\tilde{a}_i$, $\tilde{b}_i$, $\tilde{c}_i$, $\tilde{\varkappa}_1$, $\tilde{\varkappa}_2$ та правими частинами $\tilde{f}_i$, $\tilde{\mu}_1$, $\tilde{\mu}_2$. Чи не може це призвести до втрати точності і навіть до аварійного припинення обчислень в результаті сильного росту проміжних величин? Якщо припустити, що коефіцієнти прогонки α_i, β_i обчислюються точно, то відповідь на це запитання дістати просто. Тоді маємо $\tilde{y}_i = \alpha_{i+1}\tilde{y}_{i+1} + \beta_{i+1}$, звідки для $\delta y_i = \tilde{y}_i - y_i$ знаходимо співвідношення $\delta y_i = \alpha_{i+1}\delta y_{i+1}$. Помічаємо, що при виконанні нерівності (17), яка в свою чергу має місце за умов (16), дістаємо $|\delta y_i| \leqslant |\delta y_{i+1}|$, тобто похибка не збільшується.

Якщо врахувати, що в процесі обчислень збурюються і коефіцієнти α_{i+1}, β_{i+1}, то можна довести, що

$$\max_{i=\overline{1,N}} |\delta y_i| \leqslant \varepsilon_0 N^2,$$

де ε_0 — оцінка похибки заокруглень. Звідси за заданим ε, яке характеризує потрібну точність результату, та числом рівнянь N можна обчислити припустиму похибку заокруглення ε_0, оскільки має бути $\varepsilon_0 N^2 \approx \varepsilon$.

Інші варіанти методу прогонки. Варіант методу прогонки, який щойно розглянули, називається *правою прогонкою,* оскільки обчислення y_i за формулою (11) виконуються справа наліво. Аналогічно можна дістати і формули *лівої прогонки*:

$$\xi_i = a_i/(c_i - b_i\xi_{i+1}), \quad i = N-1, N-2, \ldots, 1; \quad \xi_N = \varkappa_2, \tag{18}$$

$$\eta_i = (b_i\eta_{i+1} + f_i)/(c_i - b_i\xi_{i+1}), \quad i = N-1, N-2, \ldots, 1, \quad \eta_N = \mu_2, \tag{19}$$

$$y_{i+1} = \xi_{i+1}y_i + \eta_{i+1}, \quad i = 0, 1, \ldots, N-1, \; y_0 = \frac{\mu_1 + \varkappa_1\eta_1}{1 - \xi_1\varkappa_1}. \tag{20}$$

Вправа 2. Вивести формули (18), (19) і довести, що за умови (16) знаменники в них не перетворюються на 0 і помилка в (20) не збільшується.

Комбінація лівої та правої прогонок дає метод зустрічних прогонок. У цьому методі для $i = \overline{0, i_0 + 1}$ за формулами (12) — (14) обчислюються прогоночні коефіцієнти α_i, β_i, а для $i = \overline{i_0, N}$ за формулами (18), (19) знаходять ξ_i, η_i. При $i = i_0$ розв'язки у вигляді (11) та (20) «склеюють» таким чином.

Із формул

$$y_{i_0} = \alpha_{i_0+1} y_{i_0+1} + \beta_{i_0+1}, \quad y_{i_0+1} = \xi_{i_0+1} y_{i_0} + \eta_{i_0+1}$$

знаходимо

$$y_{i_0} = \frac{\beta_{i_0+1} + \alpha_{i_0+1}\eta_{i_0}}{1 - \alpha_{i_0+1}\xi_{i_0+1}}.$$

Ця формула має зміст, бо хоча б одна з величин $|\xi_{i_0+1}|$ чи $|\alpha_{i_0+1}|$ в силу (16) менша за одиницю, а інша не перевищує 1. Знаючи y_{i_0}, за формулою (11) знаходимо y_i при $i < i_0$, а за формулою (20) — при $i > i_0$. Обчислення при $i < i_0$ та при $i > i_0$ можна виконувати незалежно. Метод зустрічних прогонок особливо зручний, коли треба знайти y_i лише в одній точці $i = i_0$.

1.2.2. Метод квадратного кореня розв'язування систем лінійних алгебраїчних рівнянь з симетричною матрицею. Розглянемо СЛАР

$$Au = f, \tag{21}$$

де A — дійсна симетрична матриця. Методом квадратного кореня система лінійних алгебраїчних рівнянь з симетричною матрицею розв'язується так: 1) матриця A розкладається на добуток

$$A = S^*DS, \tag{22}$$

де S — верхня трикутна, D — діагональна матриці; 2) розв'язується система $S^*Dy \equiv S^*DSu = f$ відносно вектора $y \equiv Su$, причому ця система має нижню трикутну матрицю; 3) розв'язується система з верхньою трикутною матрицею $Su = y$ відносно невідомого вектора u.

Нехай $S = (s_{ij})$, $D = (d_{ii}\delta_{ij})$, δ_{ij} — символ Кронекера. Тоді

$$(DS)_{ij} = \sum_{k=1}^{N} d_{ik}s_{kj} = d_{ii}s_{ij}, \quad (S^*DS)_{ij} = \sum_{k=1}^{N} s_{ki}d_{kk}s_{kj},$$

оскільки $S^* = (s^*_{ij}) = (s_{ji})$. В силу (22) дістаємо рівності

$$\sum_{k=1}^{N} s_{ki}d_{kk}s_{kj} = a_{ij}. \tag{23}$$

Систему рівнянь (23) можна розв'язувати рекурентно. Оскільки S — верхня трикутна матриця, то $s_{ki} = 0$ при $k > i$, $s^*_{ik} = 0$ при $k < i$. Тому

$$\sum_{k=1}^{N} s_{ki}s_{kj}d_{kk} = \sum_{k=1}^{i-1} s_{ki}s_{kj}d_{kk} + s_{ii}s_{ij}d_{ii} + \sum_{k=i+1}^{N} s_{ki}s_{kj}d_{kk} = \sum_{k=1}^{i-1} s_{ki}s_{kj}d_{kk} + s_{ii}s_{ij}d_{ii} = a_{ij}.$$

При $i = j$

$$s_{ii}^2 d_{ii} = a_{ii} - \sum_{k=1}^{i-1} s_{ki}^2 d_{kk}. \tag{24}$$

Виберемо

$$d_{ii} = \operatorname{sign}\left(a_{ii} - \sum_{k=1}^{i-1} s_{ki}^2 d_{kk}\right). \tag{25}$$

Тоді з (24) можна знайти

$$s_{ii} = \sqrt{\left|a_{ii} - \sum_{k=1}^{i-1} s_{ki}^2 d_{kk}\right|}. \tag{26}$$

При $i < j$ дістаємо

$$s_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} s_{ki}s_{kj}d_{kk}}{s_{ii}d_{ii}}. \tag{27}$$

При $i = 1$, як легко помітити, суми у (26), (27) слід покласти рівними нулю.

Випишемо окремо формули методу квадратного кореня для симетричної п'ятидіагональної матриці

$$A = \begin{bmatrix} a_1 & b_1 & c_1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ b_1 & a_2 & b_2 & c_2 & 0 & 0 & \dots & 0 & 0 & 0 \\ c_1 & b_2 & a_3 & b_3 & c_3 & 0 & \dots & 0 & 0 & 0 \\ 0 & c_2 & b_3 & a_4 & b_4 & c_4 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & a_{N-2} & b_{N-2} & c_{N-2} \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & b_{N-2} & a_{N-1} & b_{N-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & c_{N-2} & b_{N-1} & a_N \end{bmatrix}. \tag{28}$$

У цьому випадку систему (21) можна розглядати як крайову задачу для сіткового рівняння четвертого порядку вигляду

$$a_1u_1 + b_1u_2 + c_1u_3 = f_1,$$
$$b_1u_1 + a_2u_2 + b_2u_3 + c_2u_4 = f_2,$$
$$c_{i-2}u_{i-2} + b_{i-1}u_{i-1} + a_iu_i + b_iu_{i+1} + c_iu_{i+2} = f_i, \quad i = \overline{3,\ N-2},$$
$$c_{N-3}u_{N-3} + b_{N-2}u_{N-2} + a_{N-1}u_{N-1} + b_{N-1}u_N = f_{N-1},$$
$$c_{N-2}u_{N-2} + b_{N-1}u_{N-1} + a_Nu_N = f_N.$$

Легко помітити, що в цьому випадку матриця S має вигляд

$$S = \begin{bmatrix} s_1 & r_1 & q_1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & s_2 & r_2 & q_2 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & s_3 & r_3 & q_3 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & s_4 & r_4 & q_4 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & s_{N-2} & r_{N-2} & q_{N-2} \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & s_{N-1} & r_{N-1} \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & s_N \end{bmatrix}. \quad (29)$$

Для s_i, r_i, q_i дістаємо рекурентні формули (в порядку використання)

$$d_1 = \operatorname{sign} a_1,\ s_1 = \sqrt{|a_1|},\ r_1 = b_1/(s_1d_1),\ q_1 = c_1/(s_1d_1),$$
$$d_2 = \operatorname{sign}(a_2 - r_1^2d_1),\ s_2 = \sqrt{|a_2 - r_1^2d_1|},$$
$$r_2 = (b_2 - r_1q_1d_1)/(s_2d_2),\ q_2 = c_2/(s_2d_2), \quad (30)$$
$$d_i = \operatorname{sign}(a_i - q_{i-2}^2d_{i-2} - r_{i-1}^2d_{i-1}), \quad i = \overline{3,\ N},$$
$$s_i = \sqrt{|a_i - q_{i-2}^2d_{i-2} - r_{i-1}^2d_{i-1}|}, \quad i = \overline{3,\ N},$$
$$r_i = (b_i - r_{i-1}q_{i-1}d_{i-1})/(s_id_i), \quad i = \overline{3,\ N},$$
$$q_i = c_i/(s_id_i), \quad i = \overline{3,\ N-2}.$$

Щоб дістати ці формули, зауважимо, що при обчисленні $d_{ii} = d_i$, $s_{ii} = s_i$, в сумі $\sum_{k=1}^{i-1} s_{ki}^2d_{kk}$ (див. формулу (26)) при $i \geqslant 3$ використовуються елементи i-го стовпчика матриці S, які розміщені вище головної діагоналі, а до суми $\sum_{k=1}^{i-1} s_{ki}s_{kj}d_{kk}$ входять добутки елементів i-го та j-го стовпчиків матриці S, що розміщені вище головної діагоналі.

Зауважимо, що операції добування кореня можна уникнути, якщо змінити зміст формули (22). Шукатимемо зображення матриці A з комплексними елементами у вигляді

$$A = SDS^*, \quad (31)$$

де $S = (s_{ij})$ — нижня трикутна матриця, $S^* = (s_{ij}^*) = (\bar{s}_{ji})$; $D = (d_{ij})$ — діагональна матриця, причому $s_{ii} = 1$, $d_{ij} = d_i\delta_{ij}$. В силу (31) матимемо

$$a_{i1} = d_1s_{i1}, \quad i \geqslant 1,$$
$$a_{ii} = d_i + \sum_{k=1}^{i-1} d_k|s_{ik}|^2, \quad 2 \leqslant i \leqslant N,$$
$$a_{ij} = d_js_{ij} + \sum_{k=1}^{j-1} d_ks_{ik}\bar{s}_{jk}, \quad i \geqslant j+1,\ 2 \leqslant j \leqslant N-1, \quad (32)$$

де $\bar{s}_{ij}$ — комплексно-спряжене до s_{ij} число. В останній формулі можна вважати, що $3 \leqslant i \leqslant N$, $2 \leqslant j \leqslant i-1$. Поклавши $\hat{s}_{ik} = d_ks_{ik}$, формули (32) запишемо у вигляді

$$a_{i1} = \hat{s}_{i1}, \quad i \geqslant 1; \quad d_1 = a_{11};$$
$$d_i = a_{ii} - \sum_{k=1}^{i-1} \hat{s}_{ik}\bar{s}_{ik}, \quad 2 \leqslant i \leqslant N, \quad (33)$$
$$\hat{s}_{ij} = a_{ij} - \sum_{k=1}^{j-1} \hat{s}_{ik}\bar{s}_{jk}, \quad 3 \leqslant i \leqslant N, \quad 2 \leqslant j \leqslant i-1.$$

Таким чином, приходимо до такого алгоритму.

Алгоритм 3. *Розкладання матриці СЛАР* (21) *методом квадратного кореня за формулами* (33) *на добуток* (31).

1. $d_1 = a_{11}$.
2. Для i від 2 до N з кроком 1 виконати початок.
3. $\hat{s}_{i1} = a_{i1}$; $s_{i1} = \hat{s}_{i1}/d_1$.
4. Якщо $i = j$, то перейти на 6, інакше.
5. Для j від 2 до $i-1$ з кроком 1 виконати початок

$$\hat{s}_{ij} = a_{ij} - \sum_{k=1}^{j-1} \hat{s}_{ik}\bar{s}_{jk}; \quad s_{ij} = \hat{s}_{ij}/d_j;$$

кінець.

6. $d_i = a_{ii} - \sum_{k=1}^{i-1} \hat{s}_{ik}\bar{s}_{ik}$

кінець.

Метод квадратного кореня потребує порядку $N^3/3$ арифметичних дій, тобто при великих N він вдвічі швидший за метод Гаусса і потребує вдвічі менше комірок пам'яті. Стійкість алгоритму можна гарантувати для додатно означених матриць та для матриць $A = (a_{ij})$ з діагональною перевагою, тобто

$$|a_{ii}| \geqslant \sum_{j \neq i} |a_{ij}|, \quad i = \overline{1, N}.$$

Приклад 3. Розв'язати систему лінійних алгебраїчних рівнянь

$$Bm = F, \tag{34}$$

де

$$B = \begin{pmatrix} 21 & -4 & 1 & 0 & 0 \\ -4 & 10 & -3 & 1 & 0 \\ 1 & -3 & 10 & -3 & 1 \\ 0 & 1 & -3 & 10 & -4 \\ 0 & 0 & 1 & -4 & 17 \end{pmatrix}, \quad F = \begin{vmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{vmatrix}.$$

Розв'язання. У даному разі матриця S (див. (29)) має вигляд

$$S = \begin{pmatrix} s_1 & r_1 & q_1 & 0 & 0 \\ 0 & s_2 & r_2 & q_2 & 0 \\ 0 & 0 & s_3 & r_3 & q_3 \\ 0 & 0 & 0 & s_4 & r_4 \\ 0 & 0 & 0 & 0 & s_5 \end{pmatrix},$$

а елементами a_i, b_i, c_i матриці (28) є такі:

$$a = (21, 10, 10, 10, 17), \quad b = (-4, -3, -3, -4), \quad c = (1, 1, 1).$$

Згідно з формулами (30) маємо

$$d_1 = 1, \quad s_1 = \sqrt{21}, \quad r_1 = -4/\sqrt{21}, \quad q_1 = 1/\sqrt{21},$$

$$d_2 = \operatorname{sign}(10 - 16/21) = 1, \quad s_2 = \sqrt{|10 - 16/21|} = \sqrt{194/21};$$

$$r_2 = (-3 + 4/21)/\sqrt{194/21} = -59/\sqrt{4074},$$

$$q_2 = \sqrt{21}/\sqrt{194};$$

$$d_3 = \operatorname{sign}(10 - 1/21 - 59^2/(21, 194)) = 1;$$

$$s_3 = \sqrt{37\,055/4074};$$

$$r_3 = \left(-3 + \frac{59}{194}\right)/\sqrt{37\,055/4074} = -\frac{523\sqrt{21}}{\sqrt{7\,188\,670}};$$

$$q_3 = \sqrt{21 \cdot 194}/\sqrt{37\,055} = \sqrt{4074/37\,055};$$

$$d_4 = \operatorname{sign}\left(10 - \frac{21}{194} - \frac{21 \cdot 523^2}{194 \cdot 37\,055}\right) = 1;$$

$$s_4 = \sqrt{65\,364\,436/(194 \cdot 37\,055)} = \frac{\sqrt{65\,364\,436}}{\sqrt{7188670}};$$

$$r_4 = \left(-4 + \frac{\sqrt{21} \cdot 523\sqrt{21}\sqrt{194}}{\sqrt{194}\sqrt{37\,055}\sqrt{37\,055}}\right) = -\frac{137\,237\sqrt{194}}{\sqrt{2\,422\,079\,175\,980}};$$

$$d_5 = \operatorname{sign}\left(17 - \frac{21 \cdot 194}{37\,055} - \frac{194 \cdot (137\,237)^2}{37\,055 \cdot 65\,364\,436}\right) = 1;$$

$$s_5 = \sqrt{\frac{37\,255\,256\,410\,610}{37\,055 \cdot 65\,364\,436}} = \sqrt{\frac{37\,255\,256\,410\,610}{2\,422\,079\,175\,980}}.$$

Систему (34) записуємо у вигляді $S^*Sm = F$ (матриця D — одинична) і розв'язуємо відносно $y = Sm$ систему $S^*y = F$. Легко знаходимо, що $y_1 = y_2 = y_3 = y_4 = 0$, $y_5 = \dfrac{\sqrt{242\,207\,917\,598}}{\sqrt{3\,725\,525\,641\,061}}$. Далі розв'язуємо систему $Sm = y$, звідки маємо

$$m_5 = 1, \quad m_4 = \frac{13\,311\,989}{3\,267\,218}, \quad m_3 = \frac{21 \cdot 97 \cdot 13\,048\,103}{7411 \cdot 3\,267\,218} = \frac{26\,578\,985\,811}{24\,213\,352\,598},$$

$$m_2 = -\frac{21 \cdot 23\,980\,857\,010}{194 \cdot 3\,267\,218 \cdot 7411} = -\frac{251\,798\,998\,605}{2\,348\,695\,202\,006},$$

$$m_1 = -\frac{2 \cdot 23\,980\,857\,010 + 97^2 \cdot 13\,048\,103}{194 \cdot 1\,633\,609 \cdot 7411} = -\frac{170\,731\,315\,147}{2\,348\,695\,202\,006}.$$

1.2.3. Норми та обумовленість матриць систем лінійних алгебраїчних рівнянь. При розв'язуванні системи лінійних алгебраїчних рівнянь (1′) на ЕОМ обов'язково виникають похибки заокруглення. Тому фактично маємо розв'язок $\tilde{x}$ деякої іншої системи

$$\tilde{A}\tilde{x} = \tilde{b}. \tag{35}$$

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

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

при $p = 1, 2, \infty$ і узгоджені з ними норми матриць (для яких $\|Ax\| \leqslant \|A\|\|x\|$)

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

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

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

де $\lambda_{\max}(AA^*)$ — максимальне власне значення матриці AA^*, а $\sqrt{\lambda_{\max}}$ називається *максимальним сингулярним числом матриці* A. Якщо замість (35) брати модель обчислень

$$A\tilde{x} = \tilde{b}, \tag{36}$$

тобто вважати, що матриця A подається в ЕОМ точно, то з ланцюжка очевидних співвідношень

$$\tilde{x} - x = A^{-1}(\tilde{b} - b),$$

$$\|\tilde{x} - x\| \leqslant \|A^{-1}\| \|\tilde{b} - b\|, \quad b = Ax, \quad \|b\| \leqslant \|A\| \|x\|$$

випливає оцінка

$$\delta x \leqslant \operatorname{cond} A \cdot \delta b, \tag{37}$$

де $\operatorname{cond} A = \|A\| \|A^{-1}\|$ (відповідно $\operatorname{cond}_p A = \|A\|_p \|A^{-1}\|_p$). Величина $\operatorname{cond} A$ називається *числом обумовленості матриці* A і, як показує оцінка (37), це число є мірою невизначеності розв'язку системи (1) при неточних вхідних даних.

Якщо брати модель обчислень

$$\tilde{A}\tilde{x} = b,$$

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

$$\tilde{x} - x = [\tilde{A}^{-1} - A^{-1}]\,b = -A^{-1}(\tilde{A} - A)\,\tilde{A}^{-1}b = -A^{-1}(\tilde{A} - A)\,\tilde{x},$$

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

$$\frac{\|\tilde{x} - x\|}{\|\tilde{x}\|} \leqslant \operatorname{cond} A \frac{\|\tilde{A} - A\|}{\|A\|}, \tag{38}$$

тобто і в цьому випадку число $\operatorname{cond} A$ є мірою невизначеності розв'язку при неточних вхідних даних і інтервал цієї невизначеності тим ширший, чим більша величина $\operatorname{cond} A$. Можна довести, що таку саму роль відіграє число обумовленості і у випадку моделі обчислень (35). Для цього нам потрібне таке допоміжне твердження.

Лема 1. *Якщо C — матриця розмірності $n \times n$ така, що $\|C\| < 1$, то існує обернена матриця $(I + C)^{-1}$, причому*

$$\|(I + C)^{-1}\| \leqslant \frac{1}{1 - \|C\|}.$$

Доведення. З властивостей норми випливає нерівність

$$\|(I + C)x\| = \|x + Cx\| \geqslant \|x\| - \|Cx\| \geqslant (1 - \|C\|)\|x\|.$$

Оскільки $1 - \|C\| > 0$, то звідси $\|(I + C)x\| > 0$ для $x \neq 0$, тобто система лінійних алгебраїчних рівнянь $(I + C)x = 0$ має лише тривіальний розв'язок, що і означає невиродженість матриці $I + C$. Далі маємо

$$1 = \|I\| = \|(I + C)(I + C)^{-1}\| = \|(I + C)^{-1} + C(I + C)^{-1}\| \geqslant$$

$$\geqslant \|(I + C)^{-1}\| - \|C\| \|(I + C)^{-1}\| = \|(1 + C)^{-1}\|(1 - \|C\|) > 0,$$

що і доводить твердження леми.

Теорема 2. *Нехай A — невироджена $(n \times n)$-матриця, $\tilde{A} = A + \Delta A$, причому $\|\Delta A\| < 1/\|A^{-1}\|$. Тоді якщо x та $\tilde{x} = x + \Delta x$ є відповідно розв'язками систем $Ax = b$ та $\tilde{A}\tilde{x} = \tilde{b}$, $\tilde{b} = b + \Delta b$, то має місце оцінка*

$$\frac{\|\Delta x\|}{\|x\|} \leqslant \frac{\operatorname{cond} A}{1 - \operatorname{cond} A \dfrac{\|\Delta A\|}{\|A\|}} \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right).$$

Доведення. Оскільки $\|A^{-1}\Delta A\| \leqslant \|A^{-1}\| \|\Delta A\| < 1$, то в силу леми 1 існує $(I + A^{-1} \cdot \Delta A)^{-1}$, причому

$$\|(I + A^{-1} \cdot \Delta A)^{-1}\| \leqslant \frac{1}{1 - \|A^{-1}\Delta A\|} \leqslant \frac{1}{1 - \|A^{-1}\| \|\Delta A\|}.$$

Помножимо рівняння $\tilde{A}\tilde{x} = \tilde{b}$ зліва на A^{-1} і, враховуючи, що $Ax = b$, $x = A^{-1}b$, знайдемо Δx:

$$(I + A^{-1}\Delta A)\,x + (I + A^{-1}\Delta A)\,\Delta x = A^{-1}b + A^{-1}\Delta b,$$

$$\Delta x = (I + A^{-1}\Delta A)^{-1} A^{-1}(\Delta b - \Delta A x),$$

звідки

$$\frac{\|\Delta x\|}{\|x\|} \leqslant \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\Delta A\|} \left(\frac{\|\Delta b\|}{\|x\|} + \|\Delta A\| \right).$$

Враховуючи, що в правій частині цієї нерівності $\|x\| \geqslant \|b\| / \|A\|$, дістаємо шукану нерівність. Теорему доведено.

Системи лінійних алгебраїчних рівнянь, матриця яких має відносне велике число обумовленості, називаються *погано обумовленими*. Такі матриці також називають погано обумовленими. Зрозуміло, що це означення залежить від норми і від ЕОМ, на яких здійснюють обчислення: одна і та сама система на різних ЕОМ може бути добре чи погано обумовленою. Саме на це вказують слова «відносно велике число обумовленості».

Вправа 3. На прикладі матриці

$$A=\begin{pmatrix}1 & a & 0 & \dots & 0 & 0\\ 0 & 1 & a & \dots & 0 & 0\\ \dots & \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & \dots & 0 & a\\ 0 & 0 & 0 & \dots & 0 & 1\end{pmatrix}$$

довести, що оцінка (37) точна, тобто в ній досягається знак рівності. Для цього показати, що

$$A^{-1}=\begin{pmatrix}1 & -a & a^2 & \dots & \pm a^{n-1}\\ 0 & 1 & -a & \dots & \pm a^{n-2}\\ \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & \dots & -a\\ 0 & 0 & 0 & \dots & 1\end{pmatrix},$$

$$\operatorname{cond}_\infty A=(1+a)\,\frac{a^n-1}{a-1},$$

і використати формули $x_i+ax_{i+1}=b_i$, $i=n-1, \dots, 1$, $x_n=b_n$ для розв'язування системи $Ax=b$.

З тотожності $A^{-1}A=I$ і нерівності $\|AB\|_p\leqslant\|A\|_p\|B\|_p$ випливає, що $\operatorname{cond}_p A\geqslant 1$, причому рівність при $p=2$ досягається для ортогональних матриць. При $p=2$

$$\operatorname{cond}_2 A=\mu_1/\mu_n,$$

де μ_1, μ_n — відповідно найбільше і найменше сингулярні числа матриці A, а для симетричних матриць

$$\operatorname{cond} A=\lambda_1/\lambda_n,$$

де λ_1, λ_n — найбільше і найменше власні значення матриці A.

Зрозуміло, що чим менше число cond A, тим краще. Це число характеризує обчислювальні властивості системи (1′) незалежно від методу розв'язування, який може додатково погіршити ситуацію.

1.2.4. Ортогональні перетворення. QR-розвинення. Методи виключення, зокрема метод Гаусса, який ми розглянули раніше, можна подати схематично у вигляді

$$A\xrightarrow{f_1}B_1A\xrightarrow{f_2}B_2B_1A\xrightarrow{f_3}\dots\xrightarrow{f_k}B_k\cdot\dots\cdot B_1A\equiv R, \quad (39)$$

$$A=BR,\quad B=B_1^{-1}\cdot\dots\cdot B_k^{-1},$$

де матриці B_j описують деякі перетворення f_j у множині матриць $\operatorname{Mat}_n(\mathbb{R})$ (у методі Гаусса $k=n-1$, $B_j=L_j$ являють собою матриці Фробеніуса), R є верхньою трикутною матрицею). Стійкість алгоритму (39) характеризується її індикатором

$$\sigma=1+\sum_{j=1}^{k}\prod_{l=1}^{j}\frac{1}{k_l}=1+\sum_{j=1}^{k}\prod_{l=1}^{j}\|B_l\|^{-1}. \quad (40)$$

Матриці Фробеніуса $B_j=L_j$ в алгоритмі Гаусса залежать від матриці A і норми $\|B_j\|^{-1}=\|L_j\|^{-1}$, тому можуть бути необмежені зверху, отже, для таких матриць A алгоритм Гаусса може бути нестійким. Вигляд індикатора стабільності (40) підказує, як в алгоритмі вигляду (39) можна уникнути нестабільності. Для цього, очевидно, матриці B_j мають задовольняти дві умови: 1) давати змогу «легко» робити прямий хід алгоритму, тобто легким подібно до алгоритму Гаусса має бути обчислення матриць B_j та обернення матриці $B=B_1^{-1}\dots B_k^{-1}$; 2) норми $\|B_j\|^{-1}$ мають бути обмежені зверху і незалежні від A. Оскільки для ортогональних $(n\times n)$-матриць, клас яких позначимо через $\mathfrak{O}(n)$, за означенням $QQ^T=Q^TQ=I$, в результаті чого $\|Q\|_2=\max\limits_j\sqrt{\lambda_j(QQ^T)}=1$ такі матриці задовольняють умову 2). Вони також задовольняють і умову 1), бо якщо $B_j\equiv Q_j\in\mathfrak{O}(n)$, то $Q_j^{-1}=Q_j^T$, $Q=Q_1^{-1}\dots Q_k^{-1}\in\mathfrak{O}(n)$, $Q^{-1}=Q^T$. Розвинення

$$A=QR, \quad (41)$$

де $Q\in\mathfrak{O}(n)$, а R — є верхньою трикутною матрицею, називається *QR-розвиненням матриці* A. Якщо маємо QR-розвинення, то систему лінійних алгебраїчних рівнянь $Ax=b$ можна легко, як у випадку LR-розвинення Гаусса, розв'язати в два етапи: 1) $Qz=b$, звідки $z=Q^Tb$ (прямий хід); 2) $Rx=z$, звідки x легко знайти зворотною підстановкою (зворотний хід). Для такого алгоритму $k_j=\|Q_j\|_2=1$, тобто $\sigma=k+1$, і він за означенням є принаймні слабко стійким.

Серед різних ортогональних перетворень Q на практиці часто застосовуються так звані повороти (або повороти Гівенса) та відображення (відображення Хаусгольдера). З'ясуємо їхню суть на прикладі двовимірних векторів. Мета першого кроку алгоритму Гаусса (а також інших алгоритмів виключення вигляду (39)) полягає в тому, що вектор $A_1\in\mathbb{R}^n$ — перший вектор-стовпчик матриці A перетворюється на вектор $a_{11}e_1$, де $e_1=(1, 0, \dots, 0)^T$ — перший координатний вектор простору $\mathbb{R}^n$. Аналогічною є суть і наступних кроків з тією різницею, що вони виконуються послідовно для векторів $\mathbb{R}^n$, $\mathbb{R}^{n-1}$, ..., $\mathbb{R}^2$. Отже, якщо йдеться про простір $\mathbb{R}^2$, то довільний вектор a на площині за допомогою перетворення Q (іншими словами, перетворення $a\to b=Qa$) треба перевести у вектор αe_1. Оскільки довжина вектора при цьому має залишатися незмінною, то $\alpha=\|a\|$. Перша можливість полягає в тому, що вектор a повертаємо на кут θ (рис. 9). Аналітично це можна записати у вигляді

$$a\to\alpha e_1=Qa,\quad a=(a_1, a_2)^T=(\|a\|\cos\theta,\ \|a\|\sin\theta),$$

$$Q=\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}\in\mathfrak{O}(2). \quad (42)$$

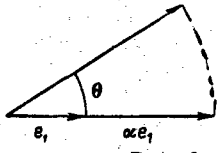


Рис. 9

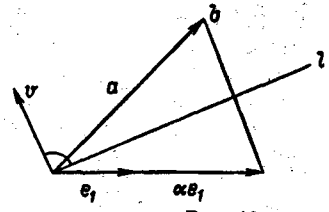


Рис. 10

Інша можливість полягає в тому, що вектор a (рис. 10) дзеркально відображається відносно прямої l — бісектриси кута θ (v — вектор, ортогональний до l):

$$b = 2\frac{\langle v, a\rangle}{\langle v, v\rangle} v = 2\frac{vv^T}{v^T v} a,$$

$$\alpha e_1 = a - b.$$

Аналітично це перетворення можна записати так:

$$a \to \alpha e_1 = Qa, \quad Q = I - 2\frac{vv^T}{v^T v}, \tag{43}$$

де I — одинична (2×2)-матриця, $v = (v_1, v_2)^T$; $\langle v, a\rangle = v_1 a_1 + v_2 a_2 = v^T a$ — скалярний добуток. Легко переконатися, що матриця Q є симетричною ($Q = Q^T$), ортогональною ($QQ^T = Q^T Q = I$) та інволютивною ($Q^2 = I$).

Перейдемо тепер до n-вимірного випадку. Розглянемо спочатку матриці

$$\Omega_{kl} = \begin{bmatrix} I & & & & \\ & c & & s & \\ & & I & & \\ & -s & & c & \\ & & & & I \end{bmatrix} \begin{matrix} \\ \leftarrow k \\ \\ \leftarrow l \\ \\ \end{matrix}, \quad \Omega_{kl} \in \mathrm{Mat}_n(\mathbb{R}), \tag{44}$$

$$\begin{matrix} & & \uparrow & & \uparrow & \\ & & k & & l & \end{matrix}$$

де $c, s \in \mathbb{R}$ — дійсні числа, що задовольняють співвідношення

$$c^2 + s^2 = 1,$$

I — одиничні матриці відповідної розмірності. Ці матриці введено в 1953 р. Г. Гівенсом і називаються *матрицями повороту Гівенса*. Вони нагадують матриці Q в (42), де $c = \cos\theta$, $s = \sin\theta$, і геометрично означають поворот на кут θ в площині (k, l) простору $\mathbb{R}^n$. Застосувавши матрицю (44) до деякого вектора $x \in \mathbb{R}^n$, дістанемо

$$x \to y = \Omega_{kl} x, \quad y_i = (\Omega_{kl} x)_i = \begin{cases} cx_k + sx_l, & i = k; \\ -sx_k + cx_l, & i = l; \\ x_i, & i \neq k, l. \end{cases} \tag{45}$$

Результат множення матриці Ω_{kl} на довільну матрицю $A \in \mathrm{Mat}_n(\mathbb{R})$ можна подати у вигляді

$$\Omega_{kl} A = [\Omega_{kl} A_1, \ \ldots, \ \Omega_{kl} A_n],$$

де A_j — j-й стовпчик матриці A. Із (45) випливає, що при такому множенні у матриці A будуть змінені лише k-й та l-й рядки, що важливо для збереження структури розміщення нульових елементів в A.

Виберемо у матриці Ω_{kl} числа c та s так, щоб після множення її на вектор x виключити l-й елемент, тобто дістати $y_l = (\Omega_{kl} x)_l = 0$. Оскільки Ω_{kl} оперує лише в (k, l)-площині, досить розглянути виключення другої компоненти вектора $a = (a_1, a_2)^T$ після множення на матрицю

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix}.$$

З рівності

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}$$

маємо

$$\begin{cases} ca_1 + sa_2 = r \\ -sa_1 + ca_2 = 0 \end{cases} \Rightarrow \begin{cases} c^2 a_1^2 + s^2 a_2^2 + 2ca_1 sa_2 = r^2 \\ sa_1 = ca_2 \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} a_1^2 (c^2 + s^2) + a_2^2 (c^2 + s^2) = r^2, \\ sa_1 = ca_2, \end{cases}$$

тобто

$$r = \pm\sqrt{a_1^2 + a_2^2}, \quad c = a_1/r, \quad s = a_2/r.$$

Отже, щоб дістати $y_l = (\Omega_{kl} x)_l = 0$ слід у виразі для Ω_{kl} покласти

$$r = \sqrt{x_k^2 + x_l^2}, \quad c = x_k/r, \quad s = x_l/r.$$

На практиці, щоб уникнути переповнення, використовуються формули:

$$\tau = x_k/x_l, \quad s = 1/\sqrt{1 + \tau^2}, \quad c = s\tau, \quad \text{якщо} \quad |x_l| > |x_k|,$$

$$\tau = x_l/x_k, \quad c = 1/\sqrt{1 + \tau^2}, \quad s = c\tau, \quad \text{якщо} \quad |x_l| \leqslant |x_k|.$$

Виключаючи таким чином стовпчик за стовпчиком ненульові елементи, що розміщені нижче головної діагоналі, можна довільну матрицю A звести до верхнього трикутного вигляду так, як показано нижче на прикладі (4×4)-матриці Над стрілками подаються індекси поворотів Гівенса і кожен раз обведено два елементи, за якими

будується таке перетворення (прослідкуйте за збереженням раніше утворених нулів!):

$$A = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix} \xrightarrow{(3,4)} \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \end{pmatrix} \xrightarrow{(2,3)}$$

$$\rightarrow \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix} \xrightarrow{(1,2)} \begin{pmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix} \xrightarrow{(3,4)}$$

$$\rightarrow \begin{pmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \end{pmatrix} \xrightarrow{(3,4)} \cdots \xrightarrow{(3,4)} \begin{pmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{pmatrix} = R.$$

Отже, $R = Q^T A$, $Q^T = \Omega^{(3)}_{n-1,n}\Omega^{(2)}_{n-2,n-1},\ldots,\Omega^{(n-1)}_{1,n}$ (Ω_{ij} в Q^T може бути відсутня, якщо відповідний елемент вже дорівнює нулю), і тому

$$A = QR,$$

де матриці Ω_{ij} і Q є ортогональними.

Кількість арифметичних операцій для такого QR-розвинення у загальному випадку становить $\sim n^2/2$ операцій добування квадратного кореня та $\sim 4n^3/3$ операцій множення порівняно з $\sim n^3/3$ операціями множення при LR-розвиненні Гаусса, але це порівняння може бути вигіднішим для розріджених нулями матриць.

У 1958 р. Хаусгольдер ввів матриці $Q \in \mathrm{Mat}_n(\mathbb{R})$

$$Q = I - 2\frac{vv^T}{v^T v}, \quad v \in \mathbb{R}^n,$$

які називаються *матрицями відображень Хаусгольдера*. Геометрично вони описують відображення відносно площини, ортогональної до вектора v. Легко перевірити, що ці матриці є симетричними ($Q = Q^T$), ортогональними ($QQ^T = Q^T Q = I$) та інволютивними ($Q^2 = I$). Якщо $v \in \mathbb{C}^n$, то $Q = I - \frac{2vv^*}{v^* v}$, $Q \in \mathrm{Mat}_n(\mathbb{C})$, $Q = Q^*$, Q — ермітова, $Q^*Q = QQ^* = I$ (унітарність), $Q^2 = I$. Довільний вектор $x \in \mathbb{R}^n$ множенням на Q перетворюється так:

$$x \to y = Qx = \left(I - 2\frac{vv^T}{v^T v}\right)x = x - 2\frac{\langle v, x\rangle}{\langle v, v\rangle}v.$$

Якщо ми хочемо таким чином вектор x перетворити на αe_1, тобто

$$\alpha e_1 = x - 2\frac{\langle v, x\rangle}{\langle v, v\rangle}v,$$

то має бути $|\alpha| = \|x\|_2 = \sqrt{\langle x, x\rangle}$, а вектор v, за яким будується матриця Q, має належати підпростору, натягнутому на вектор $x - \alpha e_1$. Оскільки для нас важливий лише напрям v, то можемо вибрати

$$v = x - \alpha e_1, \quad \alpha = \pm\|x\|_2,$$

тобто $v = (x_1 - \alpha, x_2, \ldots, x_n)^T$. Щоб уникнути віднімання близьких чисел і зникнення ведучих цифр, покладемо $\alpha = -\mathrm{sign}(x_1)\|x\|_2$. З використанням формули

$$\langle v, v\rangle = \langle x - \alpha e_1, x - \alpha e_1\rangle = \|x\|_2^2 - 2\alpha\langle x, e_1\rangle + \alpha^2 = -2\alpha(x_1 - \alpha)$$

перетворення $x \to \alpha e_1 = Qx$ для довільного $x \in \mathbb{R}^n$ можемо записати у вигляді

$$x \to Qx = x - 2\frac{\langle v, x\rangle}{\langle v, v\rangle}v = x + \frac{\langle v, x\rangle}{\alpha(x_1 - \alpha)}v.$$

За допомогою таких відображень Хаусгольдера ми можемо довільну матрицю $A \in \mathrm{Mat}_n(\mathbb{R})$ звести до верхнього трикутного вигляду. Для цього на першому кроці матрицю $A = [A_1, \ldots, A_n]$ із векторами-стовпчиками A_j помножимо зліва на матрицю

$$Q_1 = I - 2\frac{v_1 v_1^T}{v_1^T v_1}, \quad v_1 = A_1 - \alpha_1 e_1, \quad \alpha_1 = -\mathrm{sign}(a_{11})\|A_1\|_2.$$

В результаті дістанемо

$$A \to A^{(1)} = Q_1 A = \begin{pmatrix} \alpha_1 & & & \\ 0 & A_2^{(1)} & \ldots & A_n^{(1)} \\ \ldots & & & \\ 0 & & & \end{pmatrix}.$$

Після k аналогічних кроків знаходимо матрицю вигляду

$$A^{(k)} = \begin{bmatrix} * & \ldots & * \\ & \ddots & \vdots \\ & * & * \\ & 0 & \\ & \ldots & \\ & 0 & B^{(k+1)} \end{bmatrix}$$

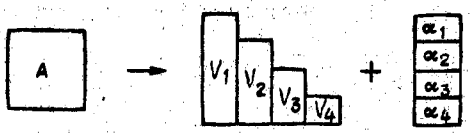


Рис. 11

і $(k+1)$-ше перетворення виконуємо за допомогою ортогональної матриці

$$Q_{k+1} = \begin{pmatrix} I_k & 0 \\ 0 & \bar{Q}_{k+1} \end{pmatrix},$$

в якій матриця $\bar{Q}_{k+1} \in \mathfrak{D}(n-k)$ будується за матрицею $B^{(k+1)}$ аналогічно тому, як на першому кроці матриця Q_1 будувалася за A. Після $n-1$ кроку дістанемо верхню трикутну матрицю

$$R = Q_{n-1} \dots Q_1 A$$

і в силу $Q_i^2 = I$ маємо розвинення

$$A = QR, \quad Q = Q_1 \dots Q_{n-1}.$$

При реалізації його на ЕОМ, крім матриці R, потрібно зберігати в пам'яті вектори $v_1, \dots, v_{n-1}$. Для цього можна діагональні елементи $r_{ii} = \alpha_i$, $i = \overline{1, n-1}$, $r_{nn} = a_{nn}$ зберігати як окремий вектор, а вектори v_i — на місці стовпчиків матриці A за схемою рис. 11.

Інша можливість полягає в тому, що вектори v_i нормуються так, щоб їхня перша компонента дорівнювала 1, і тоді потреба у додатковому векторі пам'яті відпадає.

Цей метод потребує $\sim 2n^3/3$ множень і за обсягом обчислювальної роботи наближено еквівалентний методу квадратного кореня.

Зауважимо, що для комплексних матриць QR-розвинення є неоднозначним, бо коли S є довільною матрицею (фазовою матрицею) вигляду

$$S = \operatorname{diag}(e^{i\varphi_1}, \dots, e^{i\varphi_n}),$$

то разом з Q є унітарною і матриця QS; тому

$$QSS^*R = QR.$$

Але можна довести, що з точністю до S це розвинення однозначне.

1.2.5. Алгебраїчна проблема власних значень. Алгебраїчна проблема власних значень формулюється таким чином: знайти числа $\lambda \in \mathbb{C}$ та вектори $x \in \mathbb{C}^n$, $x \neq 0$, для яких

$$Ax = \lambda x, \tag{46}$$

де A — задана матриця з множини $\operatorname{Mat}_n(\mathbb{C})$ $(n \times n)$-матриць з комплексними елементами; $\mathbb{C}$ — множина комплексних чисел. Числа λ називаються *власними числами* (значеннями), а відповідні вектори x — *правими власними векторами* матриці A.

Множина

$$L(\lambda) = \{x \mid (A - \lambda I)x = 0\}$$

утворює підпростір векторів простору $\mathbb{C}^n$, і цей підпростір має розмірність

$$\rho(\lambda) = n - \operatorname{rang}(A - \lambda I).$$

Число $\lambda \in \mathbb{C}$ є тоді і лише тоді власним числом матриці A, коли $L(\lambda) \neq 0$, тобто коли $\rho(\lambda) > 0$, і

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

Многочлен

$$\chi_A(\lambda) = \det(A - \lambda I) = (-1)^n(\lambda^n + \alpha_{n-1}\lambda^{n-1} + \dots + \alpha_0)$$

називається *характеристичним многочленом* матриці A, і його корені є власними значеннями A. Якщо $\lambda_1, \dots, \lambda_k$ є різними коренями $\chi_A(\lambda)$, то

$$\chi_A(\lambda) = (-1)^n(\lambda - \lambda_1)^{\sigma_1}(\lambda - \lambda_2)^{\sigma_2} \dots (\lambda - \lambda_k)^{\sigma_k}.$$

Число $\sigma(\lambda_i) = \sigma_i$ називається *кратністю власного значення*, точніше *алгебраїчною кратністю*.

Власні вектори матриці A визначаються неоднозначно: якщо x, y є власними векторами, що відповідають власному значенню λ, то і $\alpha x + \beta y \neq 0$ є відповідним до λ власним вектором. Нуль-вектор разом з власними векторами якраз і заповнюють підпростір $L(\lambda)$, а його розмірність, тобто максимальне число лінійно незалежних власних векторів, що відповідають λ, називається *геометричною кратністю* власного значення λ. Наступні приклади показують, що алгебраїчна та геометрична кратності не завжди збігаються, але можна довести, що

$$1 \leqslant \rho(\lambda) \leqslant \sigma(\lambda) \leqslant n.$$

Приклад 4. Діагональна матриця

$$D = \begin{pmatrix} d & & & 0 \\ & d & & \\ & & \ddots & \\ 0 & & & d \end{pmatrix} \in \operatorname{Mat}_n(\mathbb{C})$$

має характеристичний многочлен $\chi_D(\lambda) = (d - \lambda)^n$ і d є єдиним власним значенням, якому відповідає власний вектор x $\forall x \neq 0$, $x \in \mathbb{C}^n$. У даному випадку $\sigma(\lambda) = n = \rho(\lambda)$.

Приклад 5. Матриця

$$C_n(d) = \begin{pmatrix} d & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & d \end{pmatrix} \in \operatorname{Mat}_n(\mathbb{C})$$

має характеристичний многочлен $\chi(\lambda) = (d - \lambda)^n$ і $\lambda = d$ є єдиним власним значенням з алгебраїчною кратністю $\sigma(d) = n$. Але rang $(C_n(d) - dI) = n - 1$. Отже, геометрична кратність $\rho(d) = n - (n-1) = 1$

$$L(d) = \{\alpha e_1 \mid \alpha \in \mathbb{C}\},$$

$e_1 = (1, 0, \ldots, 0)^T$ — перший координатний вектор.

В силу рівностей

$$\det(A - \lambda I) = \det(A - \lambda I)^T = \det(A^T - \lambda I),$$

$$\det(A^* - \bar{\lambda} I) = \det(A - \lambda I)^* = \det\overline{(A - \lambda I)}^T = \overline{\det(A - \lambda I)}$$

маємо: якщо λ_0 є власним значенням матриці A, то λ_0 є власним значенням і матриці A^T, а $\bar{\lambda}_0$ є власним значенням A^*. Із співвідношень

$$Ax = \lambda_0 x, \quad A^T y = \lambda_0 y, \quad A^* z = \bar{\lambda}_0 z$$

випливає, що $\overline{y = z}$, але між x та y чи x та z не існує таких простих співвідношень. Вектор z називається спряженим до x. Оскільки $y^T = z^*$ і $z^* A = \lambda z^*$, то $z^* = y^T$ називається ще *лівим власним вектором* матриці A, який відповідає власному значенню λ_0. Неважко також помітити, що перетворення подібності

$$A \to B = T^{-1} A T,$$

де $T \in GL(n)$ зберігає власні значення, а відповідні власні вектори x змінює за формулою $y = T^{-1}x$. Не змінюється при цьому характеристичний поліном, а також числа $\rho(\lambda)$ та $\sigma(\lambda)$: для $\sigma(\lambda)$ це випливає з інваріантності характеристичного многочлена, а для $\rho(\lambda)$ — з того, що для $T \in GL(n)$ вектори $x_1, \ldots, x_\rho$ є лінійно незалежними тоді і тільки тоді, коли лінійно незалежними є $y_i = T^{-1}x_i$, $i = \overline{1, \rho}$. На еквівалентних перетвореннях грунтується багато чисельних методів розв'язування проблем власних значень і власних векторів.

Наведемо без доведення теореми про деякі форми представлення матриць.

Теорема 3. *Нехай A — довільна $(n \times n)$-матриця і $\lambda_1, \ldots, \lambda_k$ — її різні власні значення з геометричними $\rho(\lambda_i)$ та алгебраїчними $\sigma(\lambda_i)$, $i = \overline{1, k}$, кратностями. Тоді для кожного власного значення λ_i, $i = \overline{1, k}$, існують $\rho(\lambda_i)$ натуральних чисел $\nu_j^{(i)}$, $j = \overline{1, \rho(\lambda_i)}$, таких, що*

$$\sigma(\lambda_i) = \nu_1^{(i)} + \nu_2^{(i)} + \cdots + \nu_{\rho(\lambda_i)}^{(i)},$$

а також невироджена $(n \times n)$-матриця T така, що матриця $J = T^{-1} A T$, яка називається нормальною формою Жордана матриці A, має вигляд

$$J = \begin{bmatrix} C_{\nu_1^{(1)}}(\lambda_1) & & & & 0 \\ & \ddots & & & \\ & & C_{\nu_{\rho(\lambda_1)}^{(1)}}(\lambda_1) & & \\ & & & \ddots & \\ & & & & C_{\nu_{\rho(\lambda_k)}^{(k)}}(\lambda_k) \\ 0 & & & & \end{bmatrix},$$

де матриці

$$C_\nu(\lambda) = \begin{bmatrix} \lambda & 1 & & 0 \\ & \lambda & 1 & \\ & & \ddots & \ddots \\ & & & 1 \\ 0 & & & \lambda \end{bmatrix} = \mathrm{Mat}_\nu(\mathbb{C})$$

називаються жордановими клітинками. При цьому числа $\nu_j^{(i)}$, $j = \overline{1, \rho(\lambda_i)}$ (а з ними і матриця J), з точністю до перестановки однозначно визначені, а матриця T, взагалі кажучи, визначається неоднозначно.

Розщепимо матрицю T за стовпчиками відповідно до нормальної форми Жордана:

$$T = (T_1^{(1)}, \ldots, T_{\rho(\lambda_1)}^{(1)}, \ldots, T_1^{(k)}, \ldots, T_{\rho(\lambda_k)}^{(k)}).$$

Тоді із співвідношення $T^{-1}AT = J$, $AT = TJ$ маємо

$$AT_j^{(i)} = T_j^{(i)} C_{\nu_j^{(i)}}(\lambda_i), \quad i = \overline{1, k}, \quad j = \overline{1, \rho(\lambda_i)}. \tag{47}$$

Позначимо стовпчики $(n \times \nu_j^{(i)})$-матриці $T_j^{(i)}$ через t_m, $m = \overline{1, \nu_j^{(i)}}$, тобто

$$T_j^{(i)} = (t_1, t_2, \ldots, t_{\nu_j^{(i)}}).$$

Тоді із (47) та з означення матриць $C_{\nu_j^{(i)}}(\lambda_i)$ дістанемо

$$(A - \lambda_i I)(t_1, \ldots, t_{\nu_j^{(i)}}) = (t_1, \ldots, t_{\nu_j^{(i)}}) \begin{bmatrix} 0 & 1 & & 0 \\ & & \ddots & \\ & & & 1 \\ & & & \ddots \\ 0 & & & 0 \end{bmatrix}$$

або

$$(A - \lambda_i I)\, t_m = t_{m-1}, \quad m = \nu_j^{(i)}, \quad \nu_j^{(i)} - 1, \ \ldots, \ 2,$$
$$(A - \lambda_i I)\, t_1 = 0.$$

Зокрема помічаємо, що t_1 (перший стовпчик матриці $T_j^{(i)}$) є власним вектором матриці A, що відповідає власному значенню λ_i. Інші вектори t_m, $m = \overline{2, \nu_j^{(i)}}$, називаються *головними векторами*, відповідними значенню λ_i, і ми бачимо, що кожній клітинці Жордана $C_{\nu_j^{(i)}}(\lambda_i)$ відповідає один власний вектор і набір головних векторів. В цілому для кожної $(n \times n)$-матриці A можна знайти базис простору $\mathbb{C}^n$ (а саме стовпчики матриці T), який складається з власних та головних векторів матриці A.

Характеристичні поліноми $\det\,(C_{\nu_j^{(i)}}(\lambda_i) - \lambda I) = (\lambda_i - \lambda)^{\nu_j^{(i)}}$ окремих клітинок Жордана $C_{\nu_j^{(i)}}(\lambda_i)$ називаються *елементарними дільниками* матриці A. Таким чином, матриця A має лінійні елементарні дільники тоді і лише тоді, коли $\nu_j^{(i)} = 1 \ \ \forall\, i, j$, тобто жорданова нормальна форма матриці A є діагональною матрицею. У цьому разі A називається матрицею, що зводиться до діагонального вигляду, або матрицею, що припускає нормалізацію, тобто в $\mathbb{C}^n$ є базис, який складається лише з власних векторів матриці A, а головні вектори не з'являються. Отже, *кожну матрицю* $A \in \mathrm{Mat}_n(\mathbb{C})$, *яка має різні власні значення, можна звести до діагонального вигляду* за допомогою перетворення подібності.

Далі розглянемо деякі класи матриць, що зводяться до діагонального вигляду. Власні вектори таких матриць утворюють базис в $\mathbb{C}^n$.

Якщо в перетворенні подібності $T^{-1}AT$ припускати, що T не довільні несингулярні матриці, то A у загальному випадку не може бути зведена до форми Жордана. Для унітарних матриць T, тобто $T^*T = I$, має місце така теорема.

Теорема 4 (теорема Шура). *Для довільної матриці* $A \in \mathrm{Mat}_n(\mathbb{C})$ *існує унітарна матриця* U *така, що*

$$U^*AU = \begin{bmatrix} \lambda_1 & * & \ldots & * \\ & \ddots & \ddots & \\ & & \lambda_2 & * \\ & & & \ddots \\ 0 & & & \lambda_n \end{bmatrix},$$

де λ_i, $i = \overline{1, n}$,— *власні значення матриці* A *(не обов'язково різні).*

Якщо $A = A^*$, тобто A є ермітовою матрицею, то $(U^*AU)^* = U^*A^*U^{**} = U^*AU$ є також ермітовою і з теореми 4 випливає така теорема.

Теорема 5. *Для довільної ермітової матриці* A *існує унітарна матриця* U *така, що*

$$U^{-1}AU = U^*AU = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

При цьому власні значення λ_i, $i = \overline{1, n}$, *матриці* A *є дійсними, а* i*-й стовпчик* x_i *матриці* U *є власним вектором, відповідним* λ_i, *тобто* A *має* n *лінійно незалежних ортогональних власних векторів.*

Узагальненням ермітових є *нормальні матриці*, для яких $A^*A = AA^*$.

Теорема 6. *Матриця* $A \in \mathrm{Mat}_n(\mathbb{C})$ *є нормальною* $(A^*A = AA^*)$ *тоді і лише тоді, коли існує унітарна матриця* U *така, що*

$$U^{-1}AU = U^*AU = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix},$$

тобто нормальні матриці можна звести до діагонального вигляду, і вони мають n *лінійно незалежних ортогональних векторів* x_i, $i = \overline{1, n}$ $(Ax_i = \lambda_i x_i)$, *які є стовпчиками матриці* U.

Для довільної прямокутної $(m \times n)$-матриці A $(n \times n)$-матриця A^*A є невід'ємно визначеною з власними значеннями $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n \geqslant 0$. Числа $\sigma_k = \sqrt{\lambda_k} \geqslant 0$, $k = \overline{1, n}$, називаються *сингулярними числами* матриці A.

Теорема 7. *Нехай* $A \in \mathrm{Mat}_{m,n}(\mathbb{C})$. *Тоді:*

1) *існує унітарна* $(m \times m)$*-матриця* U *та унітарна* $(n \times n)$*-матриця* V *такі, що матриця* $U^*AV = \Sigma$ *є діагональною* $(m \times n)$*-матрицею вигляду*

$$\Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}, \quad D = \mathrm{diag}\,(\sigma_1, \ \ldots, \ \sigma_r),$$
$$\sigma_1 \geqslant \sigma_2 \geqslant \ \cdots \ \geqslant \sigma_r > 0,$$

де $\sigma_1, \ldots, \sigma_r$ *— відмінні від нуля сингулярні числа матриці* A, r *— ранг матриці* A;

2) *відмінними від нуля сингулярними числами матриці* A^* *є також* $\sigma_1, \ldots, \sigma_r$. *Розвинення* $A = U\Sigma V^*$ *називається сингулярнозначним зображенням матриці* A *(розвинення за сингулярними значеннями).*

Унітарні матриці U та V можна інтерпретувати таким чином: стовпчики матриці U являють собою m ортонормальних власних векторів

ермітової $(m \times m)$-матриці AA^*, а стовпчики V є ортонормальними власними векторами ермітової $(n \times n)$-матриці A^*A, бо $U^*AA^*U = \Sigma\Sigma^*$, $V^*A^*AV = \Sigma^*\Sigma$. Діагональна $(n \times m)$-матриця

$$\Sigma^+ = \begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

є псевдооберненою до Σ, або оберненою матрицею Мура — Пенроуза, тобто матрицею, що задовольняє співвідношення: а) $\Sigma^+\Sigma = (\Sigma^+\Sigma)^*$; б) $\Sigma\Sigma^+ = (\Sigma\Sigma^+)^*$; в) $\Sigma\Sigma^+\Sigma = \Sigma$, $\Sigma^+\Sigma\Sigma^+ = \Sigma^+$. Тому, як легко помітити, $(n \times m)$-матриця

$$A^+ = V\Sigma^+U^*$$

є в силу єдиності псевдооберненою до A.

Нехай власний вектор y_0 є спряженим до x:

$$A^*y_0 = \bar{\lambda}_0 y_0.$$

Заради простоти вважатимемо, що λ_0 — **просте власне значення**, тобто воно є простим коренем многочлена $\chi_A(\lambda)$.

Лема 1. *Нехай $\lambda_0, \bar{\lambda}_0 \in \mathbb{C}$ є простими власними значеннями матриць A, $A^* \in \mathrm{Mat}_n(\mathbb{C})$, що відповідають власним векторам x_0 та y_0. Тоді існує неперервно диференційовне відображення*

$$\lambda : V \subset \mathrm{Mat}_n(\mathbb{C}) \to \mathbb{C}, \quad B \to \lambda(B)$$

деякого околу V з простору матриць $\mathrm{Mat}_n(\mathbb{C})$ з центром в A таке, що $\lambda(A) = \lambda_0$ і $\lambda(B)$ є простим власним значенням матриці B для всіх $B \in V$, причому

$$\lambda'(A)\,C = \frac{\langle Cx_0, y_0\rangle}{\langle x_0, y_0\rangle} \quad \forall\, C \in \mathrm{Mat}_n(\mathbb{C}).$$

Доведення. Оскільки λ_0 є простим власним значенням характеристичного многочлена $\chi_A(\lambda)$, то $\forall\, C \in \mathrm{Mat}_n(\mathbb{C})$

$$0 \neq \chi'_A(\lambda_0) = \frac{\partial}{\partial\lambda}\chi_{A+tC}(\lambda)\,|_{t=0}.$$

З теореми про неявну функцію випливає, що в деякому околі нуля $(-\varepsilon, \varepsilon)$ $\exists\,\lambda$ існує неперервно диференційовне відображення

$$\lambda : (-\varepsilon, \varepsilon) \to \mathbb{C}, \quad t \to \lambda(t)$$

таке, що $\lambda_0 = \lambda(0)$ і $\lambda(t)$ є простим власним значенням матриці $A + tC$. Крім того, існує неперервно диференційовна функція

$$x : (-\varepsilon, \varepsilon) \to \mathbb{C}^n, \quad t \to x(t)$$

така, що $x(0) = x_0$ і $x(t)$ є власним вектором матриці $A + tC$, який відповідає власному значенню $\lambda(t) \equiv \lambda(A + tC)$, тобто

$$(A + tC)\,x(t) = \lambda(t)\,x(t).$$

Помноживши цю рівність, а також рівність $Ax_0 = \lambda_0 x_0$ скалярно на y_0, знайдемо

$$\langle Ax(0), y_0\rangle = \lambda(A)\langle x_0, y_0\rangle,$$
$$\langle Ax(t), y_0\rangle + t\langle Cx(t), y_0\rangle = \lambda(A + tC)\langle x(t), y_0\rangle,$$

звідки

$$[\lambda(A + tC) - \lambda(A)]\langle x_0, y_0\rangle + \lambda(A + tC)\langle x(t) - x(0), y_0\rangle =$$
$$= \langle A(x(t) - x(0)), y_0\rangle + t\langle Cx(t), y_0\rangle.$$

Після ділення на t і переходу до границі при $t \to 0$ дістанемо

$$\lambda'(A)\langle Cx_0, y_0\rangle + \lambda(A)\langle x'(0), y_0\rangle = \langle Ax'(0), y_0\rangle + \langle Cx_0, y_0\rangle.$$

Враховуючи, що $\langle Ax'(0), y_0\rangle = \langle x'(0), A^*y_0\rangle = \langle x'(0), \bar{\lambda}_0 y_0\rangle = \lambda_0\langle x'(0), y_0\rangle = \lambda(A)\langle x'(0), y_0\rangle$, остаточно дістаємо

$$\lambda'(A)\,C = \langle Cx_0, y_0\rangle / \langle x_0, y_0\rangle,$$

що і треба було довести.

Щоб оцінити обумовленість задачі (λ, A), маємо обчислити норму $\lambda'(A)$ як лінійного відображення

$$\lambda'(A) : \mathrm{Mat}_n(\mathbb{C}) \to \mathbb{C}, \quad C \to \frac{\langle Cx_0, y_0\rangle}{\langle x_0, y_0\rangle}.$$

Виберемо в просторі $\mathrm{Mat}_n(\mathbb{C})$ норму $\|C\|_2 = \max\limits_{j=\overline{1,n}} \sqrt{\lambda_j(CC^*)}$, а в просторі $\mathbb{C}$ за норму вважатимемо модуль. Тоді

$$|\langle Cx_0, y_0\rangle| \leqslant \|Cx_0\|_2\,\|y_0\|_2 \leqslant \|C\|_2\,\|x_0\|_2\,\|y_0\|_2,$$

причому якщо $Cx_0 = \alpha y_0$ ($\alpha = \mathrm{const}$), то в першій нерівності досягається знак рівності. Це матиме місце, зокрема, для $C = y_0x_0^*$, $x^* = \bar{x}^T$. Для матриці $C = y_0x_0^*$, очевидно, і в другій нерівності справжується знак рівності. Тому з урахуванням співвідношення $\|y_0x_0^*\|_2 = \|y_0\|_2\,\|x_0\|_2$ маємо

$$\|\lambda'(A)\| = \sup_{C \neq 0} \frac{|\langle Cx_0, y_0\rangle|\,|\langle x_0, y_0\rangle|^{-1}}{\|C\|_2} = \frac{\|x_0\|_2\,\|y_0\|_2}{|\langle x_0, y_0\rangle|} = \frac{1}{|\cos(\angle(x_0, y_0))|},$$

де через $\angle(x_0, y_0)$ позначено кут між власним вектором x_0 та спряженим вектором y_0. Для нормальних матриць A, для яких за означенням $A^*A = AA^*$, x_0 та y_0 збігаються, тобто $A^*x_0 = \bar{\lambda}_0 x_0$. Тому для таких матриць $\|\lambda'(A)\| = 1$.

Сформулюємо здобуті результати у вигляді теореми.

Теорема 8. *Абсолютне та відносне числа обумовленості задачі обчислення простого власного значення λ_0 матриці $A \in \mathrm{Mat}_n(\mathbb{C})$ стосовно матричної норми $\|\cdot\|_2$ виражаються формулами*

$$k_{\mathrm{abs}} = \|\lambda'(A)\| = \frac{\|x_0\|\,\|y_0\|}{|\langle x_0, y_0\rangle|} = \frac{1}{|\cos(\angle(x_0, y_0))|},$$

$$k_{\mathrm{rel}} = \frac{\|A\|_2}{|\lambda_0|}\,\|\lambda'(A)\| = \frac{\|A\|_2}{|\lambda_0\cos(\angle(x_0, y_0))|},$$

де x_0 — *власний вектор матриці* A, *що відповідає власному значенню* λ_0, y_0 — *спряжений до нього власний вектор матриці* A^*, *тобто* $Ax_0 = \lambda_0 x_0$, $A^* y_0 = \bar{\lambda}_0 y_0$. *Зокрема, проблема власних значень для нормальних матриць є добре обумовленою з числом обумовленості* $k_{\text{abs}} = 1$.

Приклад матриці

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

в власними значеннями $\lambda_1 = \lambda_2 = 0$ та збуреної матриці

$$\tilde{A} = \begin{pmatrix} 0 & 1 \\ \varepsilon & 0 \end{pmatrix}$$

з власними значеннями $\tilde{\lambda}_{1,2} = \pm \sqrt{\varepsilon}$ показує, що для несиметричних матриць проблема власних значень може бути, взагалі кажучи, погано обумовленою. Дійсно,

$$k_{\text{abs}} \geqslant \frac{|\tilde{\lambda}_1 - \lambda_1|}{\|\tilde{A} - A\|_\infty} = \frac{\sqrt{\varepsilon}}{\varepsilon} = \frac{1}{\sqrt{\varepsilon}} \xrightarrow{\varepsilon \to 0} \infty,$$

тобто задача обчислення власного значення $\lambda = 0$ матриці A стосовно абсолютної похибки взагалі не є коректно поставленою задачею.

Інший приклад, розглянутий у вступі (п. 2), показує, що обчислення власних значень як коренів характеристичного многочлена (навіть добре розділених і однократних) хоча теоретично і можливе, але потребує великої обережності. Тому на практиці, як правило, застосовуються інші методи, деякі з яких ми розглянемо далі.

Метод степенів. Цей метод застосовується для ітераційного обчислення найбільшого за модулем власного значення і відповідного власного вектора, а також власних значень, для яких відомі вже досить хороші наближення.

Нехай $A \in \mathrm{Mat}_n(\mathbb{C})$ — матриця, яку можна звести до діагонального вигляду, з власними значеннями λ_i, для яких

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|.$$

Припустимо додатково, що не існує відмінних від λ_1 власних значень, для яких $|\lambda_j| = |\lambda_1|$, тобто існує $r > 0$ таке, що

$$\lambda_1 = \lambda_2 = \cdots = \lambda_r,$$

$$|\lambda_1| = |\lambda_2| = \cdots = |\lambda_r| > |\lambda_{r+1}| \geqslant \cdots \geqslant |\lambda_n|.$$

Оскільки A зводиться до діагонального вигляду, то A має n лінійно незалежних власних векторів x_i ($Ax_i = \lambda_i x_i$), які утворюють базис в $\mathbb{C}^n$. Нехай $t_0 \in \mathbb{C}^n$ — довільний вектор, а послідовність $\{t_i\}_{i=0,1,2,\ldots}$ утворюється за правилом

$$t_{i+1} = At_i, \quad i = 0, 1, \ldots.$$

Вектор t_0 можна подати у вигляді

$$t_0 = \alpha_1 x_1 + \cdots + \alpha_n x_n,$$

і тому

$$t_i = A^i t_0 = \alpha_1 \lambda_1^i x_1 + \cdots + \alpha_n \lambda_n^i x_n.$$

Далі

$$\frac{1}{\lambda_1^i} t_i = \alpha_1 x_1 + \cdots + \alpha_r x_r + \alpha_{r+1} \left(\frac{\lambda_{r+1}}{\lambda_1}\right)^i x_{r+1} + \cdots + \alpha_n \left(\frac{\lambda_n}{\lambda_1}\right)^i x_n,$$

і тому в силу $|\lambda_j / \lambda_1| < 1$, $j \geqslant r+1$, маємо

$$\lim_{i \to \infty} \frac{1}{\lambda_1^i} t_i = \alpha_1 x_1 + \cdots + \alpha_r x_r.$$

Припустимо додатково, що для t_0 має місце $\alpha_1 x_1 + \ldots + \alpha_r x_r \neq 0$, тобто t_0 має «досить загальний вигляд». Нормуючи $t_i = (\tau_1^{(i)}, \ldots, \tau_n^{(i)})^T$ яким-небудь способом, наприклад,

$$z_i = t_i / \tau_{j_i}^{(i)}, \quad |\tau_{j_i}^{(i)}| = \|t_i\|_\infty = \max_s |\tau_s^{(i)}|,$$

дістаємо

$$\lim_{i \to \infty} \frac{\tau_{j_i}^{(i+1)}}{\tau_{j_i}^{(i)}} = \lambda_1, \quad \lim_{i \to \infty} z_i = \alpha (\alpha_1 x_1 + \cdots + \alpha_r x_r),$$

де $\alpha \neq 0$ — деяка нормуюча стала. Цей метод збігається до найбільшого за модулем власного значення λ_1 і відповідного власного вектора $t = \alpha (\alpha_1 x_1 + \ldots + \alpha_r x_r)$ (лінійна комбінація власних векторів, що відповідають λ_1, є також власним вектором).

Маємо, що для $r = 1$ (λ_1 — просте власне число) граничний вектор z є незалежним від вибору t_0 (якщо лише $\alpha_1 \neq 0$). Якщо $r > 1$ (λ_1 — многократне власне значення), то z залежить від співвідношення коефіцієнтів $\alpha_1, \ldots, \alpha_2$, тобто від початкового наближення t_0. Швидкість збіжності оцінюється величиною $|\lambda_{r+1}/\lambda_1|$ і є тим більшою, чим менше це число. З викладеного вище також випливає, що метод буде збігатися не до λ_1, а до λ_k і відповідного власного вектора, якщо вибрати в розвиненні для t_0 $\alpha_1 = \ldots = \alpha_{k-1} = 0$, $\alpha_k \neq 0$ (і якщо додатково не існує відмінних від λ_k власних значень з тим самим модулем). Проте це справедливо лише теоретично, бо внаслідок неминучих похибок заокруглення ми вже після першої ітерації для $\alpha_1 = 0$ дістанемо

$$\tilde{t}_1 = \mathrm{fl}(At_0) = \varepsilon \lambda_1 x_1 + \tilde{\alpha}_2 \lambda_2 x_2 + \cdots + \tilde{\alpha}_n \lambda_n x_n$$

з деяким $\varepsilon \neq 0$, $\tilde{\alpha}_i \approx \alpha_i$, $i = \overline{2, n}$, і метод, таким чином, збігатиметься до λ_1.

Якщо A не зводиться до діагонального вигляду і має єдине власне значення з найбільшим модулем, то ми можемо t_0 подати через власні і головні вектори матриці A і аналогічно довести, що для t_0 досить загального вигляду метод степенів збігається до λ_1 та відповідного власного вектора.

Практична обмеженість описаного вище методу полягає в тому, що він повільно збігається, якщо абсолютні значення власних чисел близькі, а також у тому, що за ним обчислюють лише одне (максимальне за модулем) власне значення і власний вектор. Цих недоліків, проте, можна позбутися за допомогою *методу зворотних ітерацій*, запропонованого в 1945 р. Віландтом. Цей метод застосовується тоді, коли для якогось із власних значень $\lambda_1, \ldots, \lambda_n$, наприклад λ_j, відоме «досить хороше наближення» λ, тобто

$$|\lambda_j - \lambda| \ll |\lambda_k - \lambda| \; \forall \lambda_k \neq \lambda_j.$$

Тоді для «досить довільного» початкового вектора $t_0 \in \mathbb{C}^n$ будується послідовність

$$(A - \lambda I) t_i = t_{i-1}.$$

Якщо $\lambda \neq \lambda_i$, $i = \overline{1, n}$, то існує $(A - \lambda I)^{-1}$ і остання рівність еквівалентна

$$t_i = (A - \lambda I)^{-1} t_{i-1},$$

тобто цей метод є звичайним методом степенів для матриці $(A - \lambda I)^{-1}$ та власних чисел $1/(\lambda_k - \lambda)$, $k = \overline{1, n}$, причому

$$\left|\frac{1}{\lambda_j - \lambda}\right| \gg \left|\frac{1}{\lambda_k - \lambda}\right| \; \forall \lambda_k \neq \lambda_j.$$

Припустивши знову, що A зводиться до діагонального вигляду і має власні вектори x_i, $i = \overline{1, n}$, і поклавши $t_0 = \alpha_1 x_1 + \ldots + \alpha_n x_n$, дістанемо для простого λ_j

$$t_i = (A - \lambda I)^{-i} t_0 = \sum_{k=1}^{n} \frac{\alpha_k}{(\lambda_k - \lambda)^i} x_k,$$

$$(\lambda_j - \lambda)^i t_i = \alpha_j x_j + \sum_{k \neq j} \left(\frac{\lambda_j - \lambda}{\lambda_k - \lambda}\right)^i \alpha_k x_k,$$

$$\lim_{i \to \infty} (\lambda_j - \lambda)^i t_i = \alpha_j x_j.$$

Швидкість збіжності тим більша, чим меншим є відношення $|\lambda_j - \lambda| / |\lambda_k - \lambda|$ для $\lambda_k \neq \lambda_j$, тобто чим кращим є наближення λ. Зауважимо, що матриця $A - \lambda I$ для досить хороших наближень λ є майже виродженою, проте можна довести, що оскільки ми шукаємо лише напрям власного вектора, то ця задача є добре обумовленою і не виникає жодних труднощів. Зазначимо також, що для обчислення t_i досить один раз обчислити LR-розвинення матриці $A - \lambda I$ і потім з його допомогою обчислювати розв'язки систем $(A - \lambda I) t_i = t_{i-1}$ з різними правими частинами.

Якщо нас цікавить не найбільше за модулем власне значення або ж потрібно розв'язати повну проблему власних значень, то розглянутий вище метод для цього непридатний. Оскільки остання задача взагалі складна і багатогранна, обмежимося далі лише повною проблемою власних значень для матриць з різними власними значеннями.

Найпростіша ідея, яка здавалось би веде до розв'язування повної проблеми власних значень, полягає в тому, щоб за допомогою перетворень подібності, скажімо, за допомогою унітарних матриць, звести матрицю A до діагонального вигляду. Перша ж спроба зробити це за допомогою матриць Хаусгольдера H показує, що це неможливо, бо нулі, які утворюються після множення на H зліва, «руйнуються» після множення на Q^* справа:

$$A = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix} \xrightarrow{H\,\cdot} \begin{pmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix} \xrightarrow{\cdot\, H^*} \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix}.$$

Проте легко помітити, що таким чином можна матрицю $A \in \mathrm{Mat}_n(\mathbb{C})$ звести до так званої форми Хессенберга:

$$A = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix} \xrightarrow{\tilde{H}\,\cdot} \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix} \xrightarrow{\cdot\, \tilde{H}^*} \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix},$$

де

$$\tilde{H} = \left(\begin{array}{c|c} 1 & \\ \hline & H^{(3)} \end{array}\right),$$

$H^{(3)}$ — (3×3)-матриця Хаусгольдера, побудована за обведеним тривимірним вектором. Якщо A — ермітова матриця, то замість згаданих вище ситуацій матимемо такі:

$$A = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix} \xrightarrow{H\,\cdot} \begin{pmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix} \xrightarrow{\cdot\, H^*} \begin{pmatrix} * & 0 & 0 & 0 \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix}$$

та

$$A = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix} \xrightarrow{\tilde{H}\,\cdot} \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix} \xrightarrow{\cdot\,\tilde{H}^*} \begin{pmatrix} * & * & 0 & 0 \\ * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix},$$

причому в силу ермітовості матриці A матриця $\tilde{H}A\tilde{H}^*$ також буде ермітовою:

$$(\tilde{H}A\tilde{H}^*)^* = \tilde{H}A^*\tilde{H}^* = \tilde{H}A\tilde{H}^*.$$

Сформулюємо ці міркування у вигляді такої леми.

Лема 2. *Нехай $A \in \mathrm{Mat}_n(\mathbb{C})$. Тоді існує унітарна матриця P, яка є добутком $(n-2)$ матриць Хаусгольдера і така, що*

$$PAP^* = \begin{bmatrix} * & \cdot & \cdot & \cdot & * \\ * & * & & & \cdot \\ & \cdot & \cdot & & \cdot \\ & & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot \\ 0 & & & * & * \end{bmatrix}$$

— форма Хессенберга для неермітових матриць та

$$PAP^* = \begin{bmatrix} \alpha_1 & \bar{\beta}_2 & & & 0 \\ \beta_2 & \alpha_2 & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \bar{\beta}_n \\ 0 & & & \beta_n & \alpha_n \end{bmatrix}$$

— тридіагональна ермітова матриця для ермітових матриць ($\alpha_i = \bar{\alpha}_i$, тобто α_i — дійсні числа).

Доведення. Продовжуючи описаний вище процес за допомогою матриць відображення Хаусгольдера $P_1, \ldots, P_{n-2}$, дістанемо

$$A \to \underbrace{P_{n-2} \ldots P_1}_{P} A \underbrace{P_1^* \ldots P_{n-2}^*}_{P^*} = \begin{cases} \text{форма Хессенберга для} \\ \text{неермітових матриць,} \\ \text{тридіагональна ермітова} \\ \text{форма для ермітових матриць} \end{cases}$$

QR-метод. На ідеї зведення матриці до «простішого» вигляду за допомогою перетворень подібності будується ряд чисельних методів. Проте розглянемо спочатку загальніший QR-метод, який ефективно використовується на практиці. Історично цей метод, запропонований Ф. Френсісом у 1959 р. і В. М. Кублановською в 1961 р., можна розглядати як розвинення LR-методу Рутісхаузера (1958 р.). Алгоритмічно вони дуже схожі, лише в останньому — замість QR-розвинення матриць застосовується LR-розвинення, що є недоліком, бо LR-розвинення існує не для всіх матриць (навіть невироджених). QR-розвинення матриці $A \in \mathrm{Mat}_n(\mathbb{C})$ існує завжди, але з точністю до так званої фазової матриці

$$S = \mathrm{diag}\,(e^{i\varphi_1}, \ldots, e^{i\varphi_n}), \quad i = \sqrt{-1}.$$

Дійсно, якщо $A = QR$, де Q — унітарна матриця $(QQ^* = Q^*Q = I)$, то разом з Q є унітарною матриця QS і $(QS)(S^*R) = QR$.

Суть QR-алгоритму полягає в побудові послідовності матриць $\{A_i\}$ за правилом: $A_1 = A$, $A_k = Q_kR_k$ (QR-розвинення), $A_{k+1} = R_kQ_k$, $k = 1, 2, \ldots$. Ми бачили, що QR-розвинення матриць A_i можна виконати, наприклад, за допомогою чисельно стійких перетворень Хаусгольдера $H_j^{(i)}$, $j = \overline{1, n-1}$:

$$H_{n-1}^{(i)} \ldots H_1^{(i)} A_i = R_i,$$

де R_i є верхньою трикутною матрицею. Оскільки для перетворень Хаусгольдера $H_j^{(i)}$ маємо $(H_j^{(i)})^* = (H_j^{(i)})^{-1} = H_j^{(i)}$, то

$$Q_i = H_1^{(i)} \ldots H_{n-1}^{(i)}, \quad A_{i+1} = R_iH_1^{(i)} \ldots H_{n-1}^{(i)}.$$

Лема 3. *Матриці A_i, Q_i, R_i, а також матриці*

$$P_i = Q_1Q_2 \ldots Q_i, \quad U_i = R_iR_{i-1} \ldots R_1, \quad P_0 = U_0 = I$$

мають такі властивості:

1) *A_{i+1} є подібною до A_i, тобто $A_{i+1} = Q_i^*A_iQ_i$;*
2) *$A_{i+1} = (Q_1 \ldots Q_i)^* A_1 (Q_1 \ldots Q_i) = P_i^*A_1P_i$;*
3) *$A^i = P_iU_i$.*

Доведення. 1) Оскільки $A_i = Q_iR_i$, $Q_i^*Q_i = I$, то

$$Q_i^*A_iQ_i = R_iQ_i \equiv A_{i+1}.$$

2) Це твердження є елементарним наслідком 1).

3) Із властивості 2) випливає, що

$$Q_1 \cdot \ldots \cdot Q_iA_{i+1} = A_1Q_1 \cdot \ldots \cdot Q_i, \quad i = 1, 2, \ldots,$$

тобто

$$P_iU_i = Q_1 \cdot \ldots \cdot Q_{i-1}(Q_iR_i)R_{i-1} \cdot \ldots \cdot R_1 =$$

$$= Q_1 \cdot \ldots \cdot Q_{i-1}A_iR_{i-1} \cdot \ldots \cdot R_1 = A_1Q_1 \cdot \ldots \cdot Q_{i-1}R_{i-1} \cdot \ldots \cdot R_1 =$$

$$= A_1P_{i-1}U_{i-1} = A_1^2P_{i-2}U_{i-2} = \cdots = A_1^i \equiv A^i, \quad i = 1, 2, \ldots.$$

Наступна теорема, доведення якої проведемо за Вілкінсоном, доводить збіжність QR-алгоритму.

Теорема 9. *Нехай матриця* $A \equiv A_1 \in \mathrm{Mat}_n(\mathbb{C})$ *має такі власні значення, що*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0, \tag{48}$$

і нормальну жорданову форму $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) = YAY^{-1}$, *в якій матриця* Y *має* LR*-розвинення*

$$Y = L_Y R_Y, \quad L_Y = \begin{bmatrix} 1 & & 0 \\ * & \ddots & \\ \cdots & \ddots & \ddots \\ * & & * \; 1 \end{bmatrix}, \quad R_Y = \begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ & & * \\ 0 & & \end{pmatrix}.$$

Тоді матриці A_i, Q_i, R_i, *визначені в* QR*-алгоритмі, мають такі властивості збіжності: існують фазові матриці*

$$S_i = \operatorname{diag}(\sigma_1^{(i)}, \ldots, \sigma_n^{(i)}), \quad |\sigma_k^{(i)}| = 1$$

такі, що

1) $\lim\limits_{i\to\infty} S_{i-1}^* Q_i S_i = I$;

2) $\lim\limits_{i\to\infty} S_i^* R_i S_{i-1} = \lim\limits_{i\to\infty} S_{i-1}^* A_i S_{i-1} = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ & \ddots & \ddots & \vdots \\ & & \ddots & * \\ 0 & & & \lambda_n \end{pmatrix}$,

зокрема

$$\lim_{i\to\infty} a_{jj}^{(i)} = \lambda_j, \quad j = \overline{1, n},$$

де $a_{jk}^{(i)}$ — *елементи матриці* A_i.

Доведення. Оскільки $A = XDY$, $X = Y^{-1}$ і існує D^{-1}, то

$$A^i = XD^iY = Q_X R_X D^i L_Y R_Y = Q_X R_X (D^i L_Y D^{-i}) D^i R_Y, \tag{49}$$

де $Q_X R_X = X$ є QR-розвиненням невиродженої матриці X в добуток унітарної матриці Q_X та невиродженої верхньої трикутної матриці R_X. Неважко помітити, що матриця $D^i L_Y D^{-i} \equiv (l_{jk}^{(i)})$ є нижньою трикутною матрицею, причому

$$l_{jk}^{(i)} = \left(\frac{\lambda_j}{\lambda_k}\right)^i l_{jk}, \quad L_Y = (l_{jk}), \quad l_{jk} = \begin{cases} 1, & j = k, \\ 0, & j < k. \end{cases}$$

Оскільки $|\lambda_j| < |\lambda_k|$ для $j > k$, то $\lim\limits_{i\to\infty} l_{jk}^{(i)} = 0$ для $j > k$, а також

$$D^i L_Y D^{-i} = I + E_i, \quad \lim_{i\to\infty} E_i = 0.$$

Тому із (49) маємо

$$A^i = Q_X R_X (I + E_i) D^i R_Y = Q_X (I + R_X E_i R_X^{-1}) R_X D^i R_Y = Q_X (I + F_i) R_X D^i R_Y, \tag{50}$$

де $F_i = R_X E_i R_X^{-1}$, $\lim\limits_{i\to\infty} F_i = 0$.

Розглянемо далі додатно визначену матрицю

$$(I + F_i)^* (I + F_i) = I + H_i, \quad H_i = F_i^* + F_i + F_i^* F_i,$$

в якій $\lim H_i = 0$ і скористаємося тим фактом, що будь-яка додатно визначена матриця має розвинення (див. метод квадратного кореня, п. 1.2.2)

$$I + H_i = \tilde{R}_i^* \tilde{R}_i,$$

де $\tilde{R}_i$ є верхньою трикутною матрицею з додатними діагональними елементами. При цьому зазначимо, що множник $\tilde{R}_i$ залежить неперервно від матриці $I + H_i$, тому

$$\lim_{i\to\infty} \tilde{R}_i = I.$$

Матриця

$$\tilde{Q}_i = (I + F_i) \tilde{R}_i^{-1}$$

є унітарною, бо

$$\tilde{Q}_i^* \tilde{Q}_i = (\tilde{R}_i^{-1})^* (I + F_i)^* (I + F_i) \tilde{R}_i^{-1} = (\tilde{R}_i^{-1})^* (I + H_i) \tilde{R}_i^{-1} = (\tilde{R}_i^{-1})^* (\tilde{R}_i^* \tilde{R}_i) \tilde{R}_i^{-1} = I.$$

Це означає, що матриця $I + F_i$ має QR-розвинення

$$I + F_i = \tilde{Q}_i \tilde{R}_i,$$

причому

$$\lim_{i\to\infty} \tilde{Q}_i = \lim_{i\to\infty} (I + F_i) \tilde{R}_i^{-1} = I, \quad \lim_{i\to\infty} \tilde{R}_i = I.$$

Тому з (50) маємо

$$A^i = (Q_X \tilde{Q}_i)(\tilde{R}_i R_X D^i R_Y),$$

де матриця $Q_X \tilde{Q}_i$ є унітарною, а матриця $\tilde{R}_i R_X D^i R_Y$ — верхньою трикутною. З іншого боку, з леми 4 випливає, що матриця A^i має також QR-розвинення

$$A^i = P_i U_i, \quad P_i = Q_1 \ldots Q_i, \quad U_i = R_i \ldots R_1.$$

Оскільки QR-розвинення матриць єдине з точністю до деякої фазової матриці, то існують матриці

$$S_i = \operatorname{diag}(\sigma_1^{(i)}, \ldots, \sigma_n^{(i)}), \quad |\sigma_k^{(i)}| = 1,$$

такі, що

$$P_i = Q_X \tilde{Q}_i S_i^*, \quad U_i = S_i \tilde{R}_i R_X D^t R_Y, \quad i \geqslant 1,$$

причому

$$\lim_{i\to\infty} P_i S_i = \lim_{i\to\infty} Q_X \tilde{Q}_i = Q_X,$$

$$Q_i = P_{i-1}^{-1} P_i = S_{i-1} \tilde{Q}_{i-1}^* \tilde{Q}_i S_i^*, \quad S_{i-1}^* Q_i S_i = \tilde{Q}_{i-1}^* \tilde{Q}_i,$$

$$\lim_{i\to\infty} S_{i-1}^* Q_i S_i = I,$$

$$R_i = U_i U_{i-1}^{-1} = S_i \tilde{R}_i R_X D^t R_Y \cdot R_Y^{-1} D^{-i+1} R_X^{-1} \tilde{R}_{i-1}^{-1} S_{i-1}^* =$$

$$= S_i \tilde{R}_i R_X D R_X^{-1} \tilde{R}_{i-1}^{-1} S_{i-1}^*,$$

$$S_i^* R_i S_{i-1} = \tilde{R}_i R_X D R_X^{-1} \tilde{R}_{i-1}^{-1}, \quad \lim_{i\to\infty} S_i^* R_i S_{i-1} = R_X D R_X^{-1},$$

і оскільки $A_i = Q_i R_i$, то

$$\lim_{i\to\infty} S_{i-1}^* A_i S_{i-1} = \lim_{i\to\infty} S_{i-1}^* Q_i S_i S_i^* R_i S_{i-1} = R_X D R_X^{-1}.$$

Матриця $R_X D R_X^{-1}$ є верхньою трикутною вигляду

$$R_X D R_X^{-1} = \begin{pmatrix} \lambda_1 & * & \dots & * \\ & \cdot & \cdot & \\ & & \cdot & \cdot \\ 0 & & & \lambda_n \end{pmatrix},$$

що і доводить теорему.

Збіжність QR-алгоритму можна довести і при слабкіших обмеженнях, ніж в теоремі 9. Якщо, наприклад,

$$|\lambda_1| > \dots > |\lambda_r| = |\lambda_{r+1}| > \dots > |\lambda_n|, \tag{51}$$

то для матриць $A_i = (a_{jk}^{(i)})$ можна довести таку теорему.

Теорема 10. *Нехай із умов теореми 9 замість (48) виконується (51). Тоді*

а) $\lim\limits_{i\to\infty} a_{jk}^{(i)} = 0 \; \forall\, j,\ k \notin \{r,\ r+1\},\quad j > k;$

б) $\lim\limits_{i\to\infty} a_{jj}^{(i)} = \lambda_j$ для $j \neq r,\ r+1;$

в) *хоча матриці*

$$\begin{pmatrix} a_{rr}^{(i)} & a_{r,r+1}^{(i)} \\ a_{r+1,r}^{(i)} & a_{r+1,r+1}^{(i)} \end{pmatrix}$$

при $i \to \infty$ не збігаються, а їхні власні значення збігаються до λ_r та λ_{r+1}, тобто

$$A_i \xrightarrow[i\to\infty]{} \begin{bmatrix} \lambda_1 & * & & \dots & * \\ & \cdot & \cdot & & \\ 0 & & \cdot & & \cdot \\ & \cdot & & \cdot & \\ & & \lambda_{r-1} & & \cdot \\ & & & (*)\ (*) & \cdot \\ & & & (*)\ (*) & \cdot \quad \cdot \\ & & & 0 \quad \lambda_{r+2} & \cdot \\ & & & \cdot \quad \cdot & * \\ & & & \cdot \quad \cdot & \\ & & & \cdot & \\ 0 & & & & 0 \quad \lambda_n \end{bmatrix},$$

де збіжність має місце для елементів, які позначені λ_j та 0, а також для власних значень (2×2)-матриць, позначених (*), які збігаються до λ_r та λ_{r+1}.

Можна також довести, що коли матриця A зводиться до діагонального вигляду

$$A = XDY, \quad Y = X^{-1}, \quad D = \operatorname{diag}(\lambda_1, \dots, \lambda_n),$$

але Y не має LR-розвинення, то QR-метод все одно збігається, лише в граничній матриці діагональні елементи, які є власними значеннями матриці A, не обов'язково будуть впорядкованими за значеннями модуля.

Звернемо увагу на такі недоліки QR-алгоритму в описаній вище формі: а) якщо матриця A повністю заповнена, то на кожному кроці $A_i \to A_{i+1}$ потрібно виконати $O(n^3)$ операцій; б) збіжність алгоритму невисока, якщо власні значення λ_j, $|\lambda_1| > \dots > |\lambda_n|$ погано розділені, тобто $|\lambda_j/\lambda_k| \approx 1$, $j > k$.

Щоб усунути перший недолік, тобто зменшити кількість арифметичних операцій, матрицю A за допомогою перетворень подібності зводять спочатку до форми Хессенберга (для матриць загального вигляду):

$$A \to A_0 \to A_1 \to \dots \to A_m = B,$$

$$A_i = T_i^{-1} A_{i-1} T_i, \quad B = A_m = T^{-1} A T, \quad T = T_1 \dots T_m,$$

$$B = \begin{bmatrix} * & \dots & * \\ * & \cdot & \cdot \\ & \cdot & \cdot \\ & \cdot & \cdot \\ 0 & * & * \end{bmatrix}$$

або до тридіагонального вигляду (для ермітових, зокрема симетричних,

матриць):

$$B = \begin{bmatrix} \alpha_1 & \bar{\beta}_2 & 0 \\ \beta_2 & \ddots & \bar{\beta}_n \\ 0 & \beta_n & \alpha_n \end{bmatrix}, \quad \alpha_i = \bar{\alpha}_i.$$

Це можна зробити, використовуючи лише унітарні матриці T_i, $T_i^{-1} = T_i^*$. Доведемо лему для симетричних матриць з дійсними компонентами, *якщо* $\{A_k\}$ — послідовність матриць з QR-алгоритму, тобто $A_1 = A$, $A_k = Q_k R_k$, $A_{k+1} = R_k Q_k$, $k = 1, 2, \ldots$.

Лема 4. *Матриці A_k мають такі властивості:*

1) *A_k подібні до A, тобто $A = Q_k A_k Q_k^T$, $Q_k \in \mathfrak{O}(n)$;*

2) *якщо A є симетричною, то симетричними є і матриці A_k;*

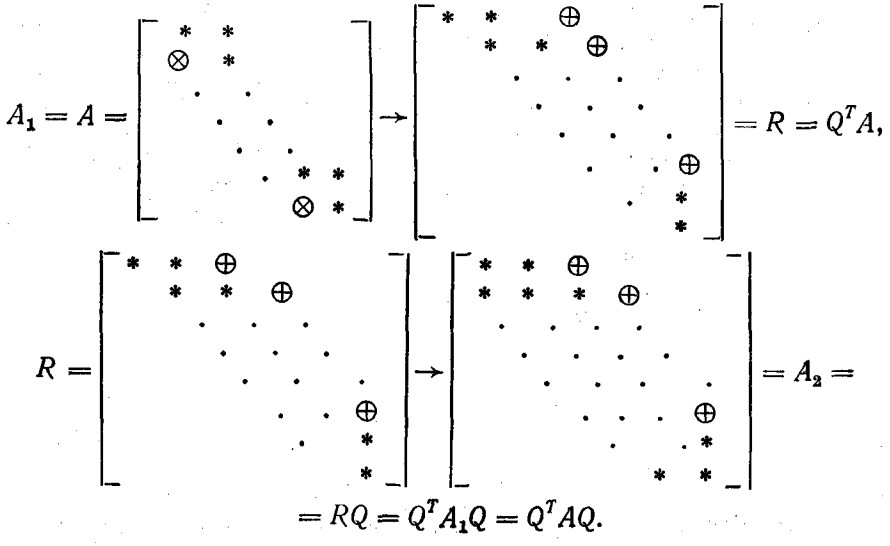
3) *якщо A є симетричною тридіагональною матрицею, то такими самими є і A_k, якщо Q_k реалізуються за допомогою поворотів Гівенса.*

Доведення. 1) Це твердження доведене в лемі 3.

2) Це твердження є наслідком властивості 1) і того, що для $B \in GL(n)$

$$(B^T A B)^T = B^T A^T B = B^T A B.$$

3) Нехай A є симетричною тридіагональною матрицею. Реалізуємо Q за допомогою $n-1$ поворотів Гівенса так, що $Q^T = \Omega_{n-1,n} \ldots \Omega_{12}$. Позначаючи через $\otimes$ елементи, які виключаються, а через $\oplus$ нові ненульові елементи, дістаємо:



$$= RQ = Q^T A_1 Q = Q^T A Q.$$

За властивістю 2) матриця A_2 має бути знову симетричною, тому всі елементи, які позначені $\oplus$, в A_2 насправді дорівнюють нулю. Отже, A_2 є тридіагональною.

Вказаний вище недолік б) QR-алгоритму, а саме повільна збіжність при $|\lambda_j/\lambda_k| \approx 1$, $j > k$, можна усунути за допомогою так званої стратегії зсувів (Shift-strategy). Для цього на кожному ітераційному кроці i застосовується параметр зсуву σ_i і послідовність $\{A_i\}$ визначається таким чином: $A_1 = A$, $A_i - \sigma_i I = Q_i R_i$ (QR-розвинення), $A_{i+1} = R_i Q_i + \sigma_i I$, $i \geqslant 1$. Неважко довести, що для тридіагональних симетричних матриць (див. леми 3, 4)

$$A_{i+1} = Q_i^T A_i Q_i,$$

$$(A - \sigma_i I) \ldots (A - \sigma_1 I) = Q_1 \ldots Q_i R_i \ldots R_1.$$

Послідовність $\{A_i\}$ збігається до $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ зі швидкістю

$$a_{jk}^{(i)} = O\left(\left|\frac{\lambda_j - \sigma_1}{\lambda_k - \sigma_1}\right| \ldots \left|\frac{\lambda_j - \sigma_{i-1}}{\lambda_k - \sigma_{i-1}}\right|\right), \quad j > k.$$

Параметри σ_i мають лежати якомога ближче до λ_j. Вілкінсон запропонував такий метод вибору параметрів зсуву для симетричної тридіагональної матриці A: якщо

$$A_i = \begin{bmatrix} \ddots & \ddots & \ddots & \\ \ddots & \ddots & \ddots & \\ & \ddots & d_{n-1}^{(i)} & e_n^{(i)} \\ & & e_n^{(i)} & d_n^{(i)} \end{bmatrix},$$

то за σ_i вибирається те власне значення (2×2)-матриці

$$\begin{pmatrix} d_{n-1}^{(i)} & e_n^{(i)} \\ e_n^{(i)} & d_n^{(i)} \end{pmatrix},$$

яке лежить ближче до $d_n^{(i)}$. Крім методу *явних зсувів* для прискорення збіжності застосовується метод *неявних зсувів*, який, проте, виходить за рамки цієї книги.

Для симетричних тридіагональних матриць потрібно виконати $O(n)$ операцій, щоб відшукати кожне власне значення, отже $O(n^2)$ операцій, щоб відшукати всі власні значення.

Крім власних значень, нас цікавлять також власні вектори, які для симетричних матриць можна знайти таким чином: якщо $Q \in \mathfrak{O}(n)$, $A \approx Q^T \Lambda Q$, $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, то стовпчики матриці Q апроксимують власні вектори η_i матриці A, тобто $Q \approx [\eta_1, \ldots, \eta_n]$.

Отже, для обчислення власних значень та власних векторів симетричної матриці A можна запропонувати такий QR-алгоритм.

1. Звести A до тридіагонального вигляду

$$A \to A_1 = PAP^T, \quad P \in \mathfrak{O}(n),$$

A_1 — симетрична і тридіагональна матриця;

2. Знайти власні наближені значення за допомогою QR-алгоритму з поворотами Гівенса: $\Omega A_1 \Omega^T \approx \Lambda$, Ω є добутком поворотів Гівенса $\Omega_{ij}^{(k)}$.

3. Стовпчики ΩP є апроксимаціями власних векторів A:

$$\Omega P \approx [\eta_1, \ldots, \eta_n].$$

Цей алгоритм потребує $\sim \frac{4}{3} n^3$ множень для перетворення до тридіагонального вигляду і $O(n^2)$ для виконання QR-алгоритму. Для загальних матриць A спочатку виконується перетворення до форми Хессенберга, а потім за допомогою QR-алгоритму — до нормальної форми Щура (комплексної верхньої трикутної матриці).

1.2.6. Розвинення матриці за сингулярними числами. Нехай $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$ — прямокутна матриця з дійсними елементами. Наступна теорема доводить існування розвинення матриці A, яке називається *розвиненням за сингулярними числами* і є корисним у багатьох практичних застосуваннях.

Теорема 11. *Нехай $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$ — довільна дійсна матриця. Тоді існують ортогональні матриці $U \in \mathfrak{O}(m)$ та $V \in \mathfrak{O}(n)$ такі, що*

$$U^T A V = \Sigma \equiv \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \in \mathrm{Mat}_{m,n}(\mathbb{R}),$$

де $p = \min(m, n)$, $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_p \geqslant 0$. *Розвинення*

$$A = U \Sigma V^T \text{ або } U^T A V = \Sigma$$

називається розвиненням матриці A за сингулярними числами σ_i, $i = \overline{1, p}$.

Доведення. Досить довести, що існують $U \in \mathfrak{O}(m)$ та $V \in \mathfrak{O}(n)$ такі, що

$$U^T A V = \begin{pmatrix} \sigma & 0 \\ 0 & B \end{pmatrix},$$

і далі застосувати метод математичної індукції. Нехай $\sigma = \|A\|_2 = \max_{\|x\|_2 = 1} \|Ax\|_2$. Оскільки максимум досягається, то існують вектори $v \in R^n$, $u \in R^m$, для яких

$$Av = \sigma u \text{ та } \|u\|_2 = \|v\|_2 = 1.$$

Розширимо $\{v\}$ до ортонормованого базису $\{v = V_1, \ldots, V_n\}$ простору R^n, $\{u\}$ — до $\{u = U_1, \ldots, U_m\}$ простору $\mathbb{R}^m$. Тоді

$$V = [V_1, \ldots, V_n] \text{ та } U = [U_1, \ldots, U_m]$$

є ортогональними матрицями, $V \in \mathfrak{O}(n)$, $U \in \mathfrak{O}(m)$, причому

$$A_1 = U^T A V = \begin{pmatrix} \sigma & w^T \\ 0 & B \end{pmatrix},$$

де $w \in \mathbb{R}^{n-1}$. Оскільки

$$\left\| A_1 \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 = \left\| \begin{matrix} \sigma^2 + \|w\|_2^2 \\ Bw \end{matrix} \right\|_2^2 \geqslant (\sigma^2 + \|w\|_2^2)^2,$$

$$\left\| \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 = \sigma^2 + \|w\|_2^2,$$

то

$$\sigma^2 + \|w\|_2^2 \leqslant \left\| A_1 \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 / [(\sigma)^2 + \|w\|_2^2] =$$

$$= \left\| A_1 \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 / \left\| \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 \leqslant \sup_{x \neq 0} \|A_1 x\|_2^2 / \|x\|_2^2 = \|A_1\|_2^2 = \|A\|_2^2 = \sigma^2,$$

звідки дістаємо $w = 0$, тобто насправді

$$U^T A V = \begin{pmatrix} \sigma & 0 \\ 0 & B \end{pmatrix}.$$

Розвинення за сингулярними числами матриці містить найважливішу інформацію про неї, що випливає з таких наслідків теореми 11.

Наслідок. *Нехай $U^T A V = \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$ — розвинення за сингулярними числами матриці A, де $p = \min(m, n)$. Тоді:*

1) *$AV_i = \sigma_i U_i$, $A^T U_i = \sigma_i V_i$ $\forall\, i = 1, p$,* де *U_i* та *V_i — стовпчики матриць U та V відповідно;*

2) *якщо $\sigma_1 \geqslant \ldots \geqslant \sigma_r > \sigma_{r+1} = \ldots = \sigma_p = 0$, то*

$\mathrm{rang}\, A = r$, $\ker A = \mathrm{span}\{V_{r+1}, \ldots, V_n\}$, $\mathrm{im}\, A = \mathrm{span}\{U_1, \ldots, U_r\}$, де $\ker A$ *— ядро оператора A*, $\mathrm{im}\, A$ *— доповнення ядра оператора A* (span *означає лінійну оболонку*);

3) *$\|A\|_2 = \sigma_1$, де σ_1 — найбільше сингулярне число;*

4) *для норми Фробеніуса $\|A\|_F = \left(\sum_{i=1}^{n} \|A_i\|_2^2 \right)^{1/2}$, де A_i — стовпчики матриці A, маємо*

$$\|A\|_F^2 = \sigma_1^2 + \ldots + \sigma_p^2;$$

5) *числом обумовленості матриці A відносно евклідової норми $\|\cdot\|_2$ є*

$$k_2(A) \equiv \mathrm{cond}_2 A = \sigma_1 / \sigma_p,$$

де σ_1 *— найбільше, а σ_p — найменше сингулярні числа;*

6) *квадрати сингулярних чисел $\sigma_1^2, \ldots, \sigma_p^2$ є власними значеннями матриці $A^T A$ та AA^T, які відповідають власним векторам $V_1, \ldots, V_p$ та $U_1, \ldots, U_p$.*

Оскільки норма $\|\cdot\|_2$ інваріантна щодо ортогональних перетворень U та V, розвинення за сингулярними значеннями матриці A приводить до зручного зображення псевдооберненої матриці A^+, про яку йтиметься в п. 8.

Ми бачили, що сингулярні числа матриці A пов'язані з власними числами матриці A^TA формулою

$$\sigma_i(A) = \sqrt{\lambda_i(A^TA)}. \tag{52}$$

Оскільки A^TA є симетричною матрицею, то задача обчислення її власних значень добре обумовлена і тому, здавалося б, формула (52) дає зручний метод обчислення сингулярних чисел. Наступний приклад показує, що такий шлях все-таки непридатний.

Приклад 6. Користуючись гіпотетичним комп'ютером, що оперує десятковими чотиризначними числами, дістаємо

$$A = A^T = \begin{pmatrix} 1{,}005 & 0{,}995 \\ 0{,}995 & 1{,}005 \end{pmatrix}, \quad \sigma_1 = \lambda_1(A) = 2, \quad \sigma_2 = \lambda_2(A) = 0{,}01,$$

$$\mathrm{fl}(A^TA) = \begin{pmatrix} 2{,}000 & 2{,}000 \\ 2{,}000 & 2{,}000 \end{pmatrix}, \quad \tilde{\sigma}_1^2 = 4, \quad \tilde{\sigma}_2^2 = 0.$$

Тому доцільно знайти алгоритм обчислення сингулярних чисел, який оперує лише з матрицею A. Для цього потрібно знати, які перетворення матриці A зберігають її сингулярні числа.

Лема 5. *Нехай $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$, $R \in \mathfrak{O}(n)$, $Q \in \mathfrak{O}(m)$. Тоді матриці A та $B = PAQ$ мають однакові сингулярні числа.*

Доведення очевидне.

Ідея алгоритму обчислення сингулярних чисел тепер полягає в тому, щоб за допомогою ортогональних перетворень такого вигляду, як в лемі 4, звести матрицю A до бідіагонального вигляду B, а потім до тридіагональної симетричної матриці B^TB застосувати QR-алгоритм. Для визначеності вважатимемо $m \geqslant n$.

Лема 6. *Для будь-якої матриці $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$, $m \geqslant n$, існують ортогональні матриці $P \in \mathfrak{O}(m)$ та $Q \in \mathfrak{O}(n)$, які є добутками матриць Хаусгольдера, такі, що*

$$PAQ = \begin{bmatrix} * & * & & \\ & \cdot & \cdot & \\ & & \cdot & \cdot \\ \cdots & & \cdot & * \\ & \cdots & & * \\ 0 & \cdots & & 0 \\ \cdots & & & \cdots \\ 0 & \cdots & & 0 \end{bmatrix} = \begin{bmatrix} B \\ 0 \end{bmatrix},$$

де B — квадратна бідіагональна матриця.

Доведення. Покажемо схематично перетворення за допомогою матриць Хаусгольдера:

$$A = \begin{bmatrix} * & \cdots & * \\ & & \\ \cdots & & \cdots \\ * & \cdots & * \end{bmatrix} \xrightarrow{P_1\cdot} \begin{bmatrix} * & & \cdots & * \\ 0 & * & \cdots & * \\ \cdots & \cdots & & \cdots \\ 0 & * & \cdots & * \end{bmatrix} \xrightarrow{\cdot Q_1} \begin{bmatrix} * & * & 0 & \cdots & 0 \\ 0 & * & * & \cdots & * \\ \cdots & \cdots & \cdots & & \cdots \\ 0 & * & * & \cdots & * \end{bmatrix} \xrightarrow{P_2\cdot}$$

$$\to \begin{bmatrix} * & * & 0 & \cdots & 0 \\ 0 & * & * & \cdots & * \\ 0 & 0 & * & \cdots & * \\ 0 & 0 & * & \cdots & * \end{bmatrix} \to \cdots \xrightarrow{P_{n-1}\cdot} \begin{bmatrix} * & * & & & \\ & * & * & & \\ & & \cdot & \cdot & \\ & & & \cdot & \cdot \\ \cdots & & & \cdot & * \\ & & \cdots & & * \\ 0 & & \cdots & & 0 \\ \cdots & & & & \cdots \\ 0 & & \cdots & & 0 \end{bmatrix} = \begin{pmatrix} B \\ 0 \end{pmatrix}.$$

Як видно,

$$\begin{pmatrix} B \\ 0 \end{pmatrix} = \underbrace{P_{n-1} \cdots P_1}_{P} A \underbrace{Q_1 \cdots Q_{n-2}}_{Q}.$$

Щоб спростити QR-алгоритм для тридіагональної симетричної матриці B^TB, намагатимемося знайти таку його версію, яка оперує лише матрицею B. Для цього виконаємо перше виключення за допомогою матриці Гівенса в матриці $C = B^TB$:

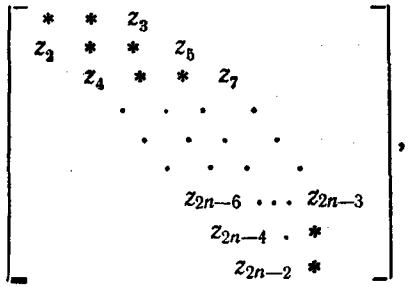
$$C \to \Omega_{12} B^T B \Omega_{12}^T = \underbrace{(B\Omega_{12}^T)^T}_{\tilde{B}^T} \underbrace{B\Omega_{12}^T}_{\tilde{B}},$$

де

$$\tilde{B} = B\Omega_{12}^T = \begin{bmatrix} * & * & & & \\ \oplus & * & * & & \\ & & \cdot & \cdot & \\ & & & \cdot & \cdot \\ & & & & * \\ & & & & * \end{bmatrix}$$

і на місці $\oplus$ утвориться новий ненульовий елемент. Продовжуючи цей процес виключення в матриці $C = B^TB$ помічаємо, що він відповідає

такому процесу виключення в матриці B:



бо виключення z_2 (множення на матрицю Гівенса зліва) приводить до появи z_3, виключення z_3 (множення на матрицю Гівенса справа) до появи z_4 і т. д., нарешті, виключення z_{2n-3} (множення на матрицю Гівенса справа) приводить до появи елемента z_{2n-2}, який виключається множенням на матрицю Гівенса зліва. Образно кажучи, ведеться «переслідування» нових елементів z_2, z_3, ... вздовж діагоналей і ці елементи виключаються поступово почережним множенням на відповідні матриці Гівенса зліва і справа (англійською мовою цей процес називається chasing). Зрештою матриця знову має бідіагональний вигляд і ми, таким чином, ітераційний крок QR-алгоритму для B^TB (він полягає в переході $B^TB \to R \equiv B_1 = CQ^T$ і далі в множенні $RQ = B_1^TB_1$) звели до операцій лише над B. Із теорем про збіжність QR-алгоритму маємо

$$B_k^TB_k \to \Lambda = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2) = \Sigma^2 \text{ при } k \to \infty.$$

Звідси випливає, що послідовність $\{B_k\}$ збігається до діагональної матриці сингулярних значень матриці B (а також A):

$$B_k \to \Sigma \text{ при } k \to \infty.$$

Отже, ми дістали такий *QR-алгоритм обчислення сингулярних значень матриці A*.

1. За допомогою ортогональних перетворень (наприклад, перетворень Хаусгольдера) $P \in \mathfrak{O}(m)$, $Q \in \mathfrak{O}(n)$ звести матрицю A до бідіагонального вигляду

$$PAQ = \begin{pmatrix} B \\ 0 \end{pmatrix},$$

де $B \in \mathrm{Mat}_n(\mathbb{R})$ — верхня бідіагональна матриця.

2. Виконати QR-алгоритм для B^TB, застосувавши chasing-метод для B. Визначати таким чином послідовність бідіагональних матриць $\{B_k\}$, яка збігається до діагональної матриці сингулярних значень Σ.

Якщо $m = n$, то кількість арифметичних операцій становить: $\sim \frac{4}{3}n^3$ множень на кроці 1 та $O(n^2)$ множень на кроці 2.

1.2.7. Розв'язування систем лінійних алгебраїчних рівнянь з прямокутною матрицею. Розглянемо таку задачу: знайти $x \in \mathbb{R}^n$ такий, що для заданих $b \in \mathbb{R}^m$ та матриці $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$

$$Ax = b. \tag{53}$$

Якщо такий вектор $x \in \mathbb{R}^n$ існує, то система лінійних алгебраїчних рівнянь (53) називається сумісною (тоді залишок $r \equiv b - Ax = 0$) в іншому разі — несумісною.

Нехай ранг матриці A дорівнює r, що стисло запишемо так: $\operatorname{rang} A = r$. З лінійної алгебри відомо (критерій Кронекера — Капеллі), що система (53) сумісна тоді і лише тоді, коли ранг r матриці A дорівнює рангу $\bar{r}$ розширеної матриці $\bar{A} = (A, b)$. При $m = n = r$ система (53) має єдиний розв'язок, методи відшукання якого ми розглянули раніше. Слід лише зазначити, що при великому числі обумовленості невиродженої матриці A її можна трактувати як вироджену і розв'язувати за допомогою алгоритмів, що спираються на розвинення матриці A за сингулярними числами.

Якщо $m < n$, $r = \bar{r} \leqslant m$, то розв'язок системи (53) існує, але не є єдиним. Він визначається через $n - r$ невідомих, які називаються *вільними*, бо можуть набувати довільних значень. Іншими словами, розв'язок визначається з точністю до елементів підпростору $\ker A \subset \mathbb{R}^n$ розмірності $n - r$, що утворюється множиною розв'язків однорідного рівняння $Ax = 0$. Як відомо, існують $n - r$ лінійно незалежних розв'язків цієї системи, що утворюють базис в $\ker A$. Цей підпростір називається ядром оператора $A : \mathbb{R}^n \to \mathbb{R}^m$.

Для розв'язування системи (53) і в цьому разі можна застосувати алгоритм Гаусса. Неможливість вибору ненульового головного елемента з пошуком по всій матриці на $(r+1)$-му кроці означає, що $\operatorname{rang} A = r$. Зауважимо, що перевірка на «нуль» максимального за модулем елемента (тобто головного елемента) має виконуватися його порівнянням з машинною точністю eps. В результаті алгоритм виключення Гаусса з вибором головного елемента по всій матриці зупиниться на кроці

$$(A, b) \to (R, \bar{b}), \quad R = \begin{bmatrix} * & & & \ldots & * \\ & \ddots & & & \vdots \\ & & \ddots & & \\ & & & * & \ldots & * \\ & & & 0 & \ldots & * \\ & & & \vdots & & \vdots \\ & & & 0 & \ldots & * \\ & & & \uparrow & & \\ & & & r+1 & & \end{bmatrix} \xleftarrow{r+1} \equiv \begin{pmatrix} R_r & B \\ 0 & A_{n-r} \end{pmatrix},$$

де R_r — невироджена верхня трикутна матриця, A_{n-r} — вироджена $(n-r)\times(n-r)$-матриця, B — прямокутна $r\times(n-r)$-матриця. Зауважимо, що в результаті перестановок стовпчиків робиться перейменування (перенумерація) невідомих. Після цього невідомі $x_{r+1}, \ldots, x_n$ оголошують вільними і розв'язки системи (53) задають формулою

$$x = (x^{(r)}, x^{(n-r)})^T,$$

де

$$x^{(r)} = R_r^{-1}(b^{(r)} - Bx^{(n-r)}),$$

$$x^{(r)} = (x_1, \ldots, x_r) \in \mathbb{R}^r, \quad x^{(n-r)} = (x_{r+1}, \ldots, x_n)^T \in \mathbb{R}^{n-r},$$

а вектор $b^{(r)} \in \mathbb{R}^r$ складається з перших r компонент вектора $\overline{b}$.

Якщо система (53) сумісна, то це означає, що існує принаймні один вектор $x \in \mathbb{R}^n$, для якого

$$r(x) = b - Ax = 0,$$

тобто $\|r(x)\| = \|r(x)\|_2 = 0$. Якщо ж система несумісна, то це означає, що в $\mathbb{R}^n$ не існує вектора x, для якого $\|r(x)\| = 0$. Але доцільно сформулювати таку задачу: знайти такий вектор $x \in \mathbb{R}^n$, що для заданих $b \in R^m$, $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$

$$\|r(x)\| = \|b - Ax\| = \min. \tag{54}$$

Вказаний вектор x і вважатимемо за розв'язок системи (53) в смислі методу найменших квадратів. Така постановка задачі має такий практичний зміст. Відомо, що деяка фізична величина, яку можна вимірювати, змінюється за законом

$$\varphi(t;\ x_1, \ldots, x_n) = a_1(t)x_1 + \cdots + a_n(t)x_n, \tag{55}$$

тобто лінійно залежить від невідомих параметрів $x_1, \ldots, x_n$, а функції $a_j(t)$ відомі. Оскільки кількість вимірювань може бути довільною, вимірювання на практиці завжди містять деяку похибку і, крім того, вигляд функції (55) нам відомий лише наближено, то система лінійних алгебраїчних рівнянь

$$\begin{aligned} a_1(t_1)x_1 + \cdots + a_n(t_1)x_n &= b_1, \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ a_1(t_m)x_1 + \cdots + a_n(t_m)x_n &= b_m, \end{aligned} \tag{56}$$

яку стисло записуватимемо у вигляді

$$Ax = b, \quad A = (a_{ij}) \in \mathrm{Mat}_{m,n}(\mathbb{R}), \quad a_{ij} = a_j(t_i),$$

як правило, несумісна. Задача полягає в тому, щоб знайти вектор x, який у певному розумінні (наприклад, у розумінні (54)) якнайкраще задовольняв рівність (56). З цієї причини задачу (54) називають також *лінійною задачею вирівнювання (вирівнювань)*. Як приклад найпростішої фізичної залежності вигляду (55) може служити закон Ома $\varphi = xt$, де t — сила струму, φ — напруга, x — опір провідника. Такий вигляд функція φ (напруга) має лише для певних (а саме середніх) температур.

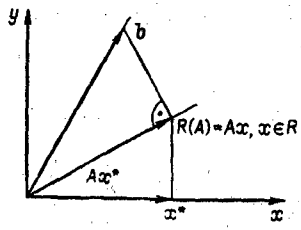


Рис. 12

Зауваження. Якщо замість норми $\|\cdot\|_2$ у постановці задачі (54) взяти $\|\cdot\|_1$, то дістанемо стандартну задачу лінійного програмування, яка відіграє важливу роль в математичній економіці. Можна вибрати і норму $\|\cdot\|_\infty$, але далі в цьому параграфі розглядатимемо норму $\|\cdot\|_2$, зокрема тому, що вона природно зв'язана зі скалярним добутком і припускає наочні геометричні інтерпретації.

Далі обмежимося випадком $m \geqslant n$.

При розв'язуванні задачі (54) шукаємо точку $z = Ax$ з множини образів $R(A)$ оператора A, яка лежить найближче до заданої точки b. При $m = 2$, $n = 1$ маємо $A \in \mathrm{Mat}_{2,1}(\mathbb{R})$, $R(A) \in \mathbb{R}^2$. Нехай $A = \begin{pmatrix}\alpha\\ \beta\end{pmatrix}$. Тоді $Ax = \begin{pmatrix}\alpha x\\ \beta x\end{pmatrix}$ і в площині (x, y) маємо $y = \frac{\beta}{\alpha}x$, тобто $R(A)$ являє собою або точку $(0, 0)$, якщо $A = \begin{pmatrix}0\\ 0\end{pmatrix}$, або ж деяку пряму, що проходить через початок координат (рис. 12). Помічаємо, що $\|b - Ax\|$ буде мінімальною, якщо $x = x^*$, де x^* — число, для якого Ax^* є ортогональною проекцією вектора b на підпростір $R(A)$. Сформулюємо цей результат в абстрактній формі.

Теорема 12. *Нехай V — скінченновимірний евклідів простір із скалярним добутком $\langle\cdot,\cdot\rangle$, U — деякий підпростір, $U \subset V$ і $U^\perp = \{v \in V \mid \langle v, u\rangle = 0\ \forall\, u \in U\}$ — його ортогональне доповнення до V. Тоді для норми $\|v\| = \sqrt{\langle v, v\rangle}$ і для всіх $v \in V$ співвідношення*

$$\|u - v\| = \min_{u' \in U} \|u' - v\| \tag{57}$$

та $u - v \in U^\perp$ еквівалентні.

Доведення. Нехай $u \in U$ — елемент (єдиний), для якого $v - u \in U^\perp$. Тоді для будь-якого $u' \in U$ маємо $\|v - u'\|^2 = \langle v - u', v - u'\rangle = \langle v - u + u - u',\ v - u + u - u'\rangle = \|v - u\|^2 + 2\langle v - u, u - u'\rangle + \|u - u'\|^2 = \|v - u\|^2 + \|u - u'\|^2 \geqslant \|v - u\|^2$, причому рівність має місце тоді і лише тоді, коли $u' = u$.

Теорема 12 дає єдиний розв'язок $u \in U$ задачі (57), який називається *ортогональною проекцією* елемента v на U. Відображення

$$P : V \to U, \quad v \mapsto Pv,$$

яке виражається рівністю

$$\|v - Pv\| = \min_{u \in U} \|v - u\|,$$

є лінійним і називається ортогональним проектором з простору V на підпростір U.

Теорема 12 залишається вірною, якщо U замінимо деяким афінним підпростором $W = w_0 + U \subset V$, де $w_0 \in V$, а U є паралельним до W підпростором. Тоді для будь-яких $v \in V$ $w \in W$ матимемо

$$\|u - w\| = \min_{w' \in W} \|v - w'\| \Leftrightarrow v - w \in U^{\perp}.$$

Відображення

$$P : V \to W,\ v \to Pv,\ \|v - Pv\| = \min_{w \in W} \|v - w\|$$

є афінним лінійним відображенням, яке також називається ортогональною проекцією V на афінний підпростір W.

За допомогою теореми 12 можна тепер довести існування та єдиність розв'язку лінійної задачі вирівнювання.

Теорема 13. *Вектор $x \in \mathbb{R}^n$ є розв'язком задачі* (54) *тоді і лише тоді, коли він задовольняє таку систему нормальних рівнянь:*

$$A^T A x = A^T b. \tag{58}$$

Цей розв'язок єдиний тоді і тільки тоді, коли ранг матриці A є максимальним, тобто $\operatorname{rang} A = n$.

Доведення. Застосовуючи теорему 12 до $V = \mathbb{R}^m$ та $U = R(A)$, маємо

$$\|b - Ax\| = \min \Leftrightarrow \langle b - Ax,\ Ax' \rangle = 0\ \forall\, x' \in \mathbb{R}^n \Leftrightarrow$$

$$\Leftrightarrow \langle A^T (b - Ax),\ x' \rangle = 0\ \forall\, x' \in \mathbb{R}^n \Leftrightarrow A^T (b - Ax) = 0 \Leftrightarrow$$

$$\Leftrightarrow A^T A x = A^T b,$$

тобто дістаємо перше твердження. Друге твердження випливає з того, що $A^T A$ тоді і лише тоді має обернену матрицю, коли $\operatorname{rang} A = n$.

Як випливає з доведення, геометрично нормальні рівняння виражають те, що $b - Ax$ є ортогональним до $R(A) \subset R^m$, звідки і походить назва.

Розглянемо задачу чисельного розв'язування нормальних рівнянь за умови однозначного розв'язку, тобто $\operatorname{rang} A = n$. Оскільки матриця $A^T A$ є симетричною додатно визначеною, то природно застосувати метод квадратного кореня (метод Холецького). При цьому для обчислення $A^T A$ потрібно $\sim \frac{1}{2} n^2 m$ операцій (множення), а для факторизації $A^T A$ — ще $\sim \frac{1}{6} n^3$ операцій. Отже, в цілому кількість операцій становить

$\sim \frac{1}{2} n^2 m$ при $m \gg n$ та $\sim \frac{2}{3} n^3$ при $n \approx m$. Недоліком такого методу є те, що вже $A^T A$ вимагає обчислення n^2 скалярних добутків і це призводить до додаткових похибок заокруглення. Тому краще було б побудувати алгоритм, який оперує лише матрицею A. Іншим недоліком вказаного вище підходу є те, що число обумовленості матриці $A^T A$ значно більше, ніж число обумовленості матриці A, про що йдеться в наступній теоремі.

Теорема 14. *Для будь-якої матриці $A \in \operatorname{Mat}_{m,n}(\mathbb{R})$ рангу $p = \min(m, n)$*

$$\operatorname{cond}_2 (A^T A) = \|A^T A\|_2 \|(A^T A)^{-1}\|_2 = [\operatorname{cond}_2 (A)]^2.$$

Доведення. Використовуючи розвинення за сингулярними числами матриці A

$$A = U \Sigma V^T,\ \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \operatorname{Mat}_{m,n}(\mathbb{R}),\ U \in \mathfrak{O}(m),$$

$$V \in \mathfrak{O}(n),\ \sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_p > 0,$$

дістаємо таке розвинення за сингулярними числами матриці $A^T A$:

$$A^T A = (U \Sigma V^T)^T U \Sigma V^T = V \Sigma U^T U \Sigma V^T = V \Sigma^2 V^T,$$

де

$$\Sigma^2 = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2) \in \operatorname{Mat}_{m,n}(\mathbb{R}),$$

тобто

$$\operatorname{cond}_2 (A^T A) = \sigma_1^2 / \sigma_p^2 = [\operatorname{cond}_2 (A)]^2.$$

Хоча через те що матриця $A^T A$ та права частина A_b^T зв'язані і тому обумовленість матриці $A^T A$ та розв'язку задачі (54) (відносно вхідних даних A та b) не однакові, все ж згадані вище міркування вимушують шукати інші методи розв'язування, бо на практиці вже сама матриця A, як правило, є погано обумовленою.

Розумною альтернативою методу квадратного кореня можуть бути методи, що грунтуються на ортогональних перетвореннях (див. п. 6) та QR-розвиненні матриці A. Нехай для $A \in \operatorname{Mat}_{m,n}(\mathbb{R})$ відома ортогональна матриця $Q \in \mathfrak{O}(m)$, для якої

$$Q^T A = \begin{bmatrix} * & \ldots & * \\ & \ddots & \\ & & * \\ 0 & \ldots & 0 \\ \ldots & & \ldots \\ 0 & \ldots & 0 \end{bmatrix} = \begin{pmatrix} R \\ 0 \end{pmatrix}, \tag{59}$$

де R — верхня трикутна матриця. Матрицю $\begin{pmatrix} R \\ 0 \end{pmatrix}$ будемо називати верхньою трикутною формою для прямокутної матриці A. Основою методу ортогональних перетворень для задачі (54) є така теорема.

Теорема 15. *Нехай $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$, $m \geqslant n$, має максимальний ранг n, $b \in \mathbb{R}^m$, $Q \in \mathfrak{O}(m)$ — матриця, для якої*

$$Q^T A = \begin{pmatrix} R \\ 0 \end{pmatrix}, \quad Q^T b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},$$

де $b_1 \in \mathbb{R}^n$, $b_2 \in \mathbb{R}^{m-n}$, $R \in \mathrm{Mat}_n(\mathbb{R})$ — невироджена верхня трикутна матриця. Тоді $x = R^{-1}b_1$ є розв'язком задачі (54).

Доведення. Оскільки ортогональні перетворення зберігають норму $\| \cdot \|_2$, то для будь-якого $x \in \mathbb{R}^n$ маємо

$$\|b - Ax\|^2 = \|Q^T(b - Ax)\|^2 = \left\| \begin{pmatrix} b_1 - Rx \\ b_2 \end{pmatrix} \right\|^2 =$$
$$= \|b_1 - Rx\|^2 + \|b_2\|^2 \geqslant \|b_2\|^2,$$

причому для $x = R^{-1}b_1$

$$\|b - Ax\|^2 = \|b_2\|^2.$$

Для зведення матриці A до вигляду (59) (QR-розвинення), тобто для побудови матриці Q, можна застосувати вже відомі нам повороти Гівенса та відображення Хаусгольдера (див. п. 4). Неважко підрахувати, що при застосуванні поворотів Гівенса на QR-розвинення потрібно $\sim \frac{n^2}{2}$ операцій добування квадратного кореня та $\sim \frac{4n^3}{3}$ операцій множення, якщо $m \approx n$; $\sim mn$ операцій добування квадратного кореня та $\sim 2mn^2$ операцій множення, якщо $m \gg n$.

Застосовуючи теорему 15, для розв'язування задачі $\|b - Ax\| = \min$ можна запропонувати такий алгоритм з використанням відображень Хаусгольдера.

Алгоритм 4. *Розв'язування задачі* (54) *за допомогою відображень Хаусгольдера.*

1. $A = QR$, де $Q = Q_1 \ldots Q_p$, $p = \min(m - 1, n)$,

$$Q_k = \begin{pmatrix} I_{k-1} & 0 \\ 0 & \overline{Q}_k \end{pmatrix},$$

$I_{k-1} \in \mathrm{Mat}_{k-1}(\mathbb{R})$ — одинична матриця, $\overline{Q}_k \in \mathfrak{O}(m - k + 1)$ — матриця Хаусгольдера.

2. $(b_1, b_2)^T = Q^T b$, де $b_1 \in \mathbb{R}^n$, $b_2 \in \mathbb{R}^{m-n}$.

3. $Rx = b_1$ — зворотна підстановка з верхньою трикутною матрицею R.

Обчислювальні затрати: $\sim 2n^2m$ операцій множення, якщо $m \gg n$; $\sim \frac{2}{3}n^3$ операцій множення, якщо $m \approx n$.

1.2.8. Узагальнена обернена матриця. Ми бачили вище, що розв'язок задачі

$$\|b - Ax\| = \min \tag{60}$$

у випадку $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$, $m \geqslant n$, $\operatorname{rang} A = n$, є розв'язком системи нормальних рівнянь

$$A^T Ax = A^T b$$

і лінійно залежить від b, що ми формально можемо записати

$$x = A^+ b,$$

де $A^+ = (A^T A)^{-1} A^T$. Оскільки $A^+ A = I$, то A^+ називається *псевдооберненою матрицею* для $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$. Поняття псевдооберненої матриці можна поширити на матриці $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$ довільного рангу. В цьому разі розв'язок задачі (60), взагалі кажучи, не єдиний. Нехай

$$\overline{P} : \mathbb{R}^m \to R(A) \subset \mathbb{R}^m$$

— ортогональний проектор простору $\mathbb{R}^m$ на множину значень оператора (матриці) A, тобто $\|u - \overline{P}u\| = \min\limits_{v \in R(A)} \|u - v\| \forall\ u \in \mathbb{R}^m$, $\overline{P}u \in R(A)$. Тоді з теореми 12 випливає, що розв'язки задачі (60) утворюють афінний підпростір

$$L(b) = \{x \in \mathbb{R}^n \mid \|b - Ax\| = \min\} = \{x \in \mathbb{R}^n \mid Ax = \overline{P}b\}.$$

Якщо $\overline{x} \in L(b)$ є деяким розв'язком задачі (60), то всі розв'язки визначаються формулою

$$L(b) = \overline{x} + N(A),$$

де $N(A)$ — нуль-простір (ядро) оператора A. Щоб визначити єдиний розв'язок, виберемо з множини $L(b)$ вектор x, який має найменшу норму $\| \cdot \|$, і позначимо його

$$x = A^+ b.$$

Неважко помітити, що x є ортогональною проекцією початку координат $0 \in \mathbb{R}^n$ на афінний підпростір $L(b)$ (рис. 13). Зрозуміло, що найменший розв'язок x має бути ортогональним до $N(A)$, тобто x — однозначно визначеним вектором $x \in N(A)^\perp$, для якого $\|b - Ax\| = \min$. Звідси маємо таке означення псевдооберненої матриці A^+.

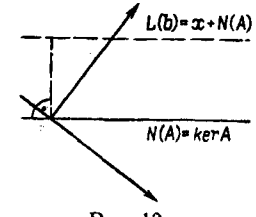


Рис. 13

Означення. *Псевдооберненою матрицею* для деякої матриці $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$ називається така матриця $A^+ \in \mathrm{Mat}_{n,m}(\mathbb{R})$, що для будь-

якого заданого $b \in \mathbb{R}^m$ вектор $x = A^+b$ є найменшим за нормою розв'язком задачі (60), тобто

$$A^+b \in N(A)^\perp, \quad \|b - AA^+b\| = \min.$$

Для наочності доцільно використовувати таку діаграму, в якій i означає включення:

$$\mathbb{R}^n \underset{A^+}{\overset{A}{\underset{\leftarrow}{\rightarrow}}} \mathbb{R}^m,$$

$$\begin{array}{ccc} P = A^+A & \downarrow\uparrow i, & i \uparrow\downarrow \bar{P} = AA^+. \\ R(A^+) = N(A)^\perp & & R(A) \end{array}$$

За означенням ортогонального проектора $\bar{P} = \mathbb{R}^m \to R(A)$

$$\bar{P} = AA^+,$$

а $P = A^+A$ є проектором з $\mathbb{R}^n$ на ортогональне доповнення $N(A)^\perp$ ($\| x - A^+Ax \| = \min_{y \in N(A)^\perp} \| x - y \|$, $x \in \mathbb{R}^n$ за означенням). Із властивостей проектора випливає, що

$$A^+AA^+ = A^+, \quad AA^+A = A. \tag{61}$$

Будь-який ортогональний проектор $Q : X \to Y \subset X$ є симетричним, бо

$$\langle x - Px, \ y \rangle = 0 \ \forall x \in X, \ y \in Y \Rightarrow \langle x, \ y \rangle -$$
$$- \langle x, \ P^Ty \rangle = 0 \Leftrightarrow \langle x, \ y - P^Ty \rangle = 0, \quad P^T : Y \to X.$$

Таким чином,

$$(A^+A)^T = A^+A, \quad (AA^+)^T = AA^+. \tag{62}$$

Властивості (61), (62) однозначно визначають псевдообернену матрицю, що стверджує така теорема.

Теорема 16. *Псевдообернена матриця $A^+ \in \mathrm{Mat}_{n,m}(\mathbb{R})$ для довільної матриці $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$ однозначно визначається умовами* (61), (62).

Доведення. Для псевдооберненої матриці A^+ з означення умови (61), (62) виконуються, бо A^+A та AA^+ являють собою ортогональні проекції на $N(A)^\perp = R(A^+)$ та $R(A)$ відповідно. Навпаки, умови (61), (62) визначають ортогональні проектори $P = A^+A$ та $\bar{P} = AA^+$, бо $P^T = P = P^2$, $\bar{P}^T = \bar{P} = \bar{P}^2$, і з (61) випливає, що P та $\bar{P}$ є ортогональними проекторами на $N(A)$ та $R(A)$ відповідно (тобто $R(P) = N(A)$, $R(\bar{P}) = R(A)$ і проектори P та $\bar{P}$ умовами (61), (62) визначені незалежно від A^+). З умов (61), (62) випливає також єдиність

A^+, бо якби існували A_1^+, A_2^+, $A_1^+ \neq A_2^+$, що задовольняють (61), (62), то ми мали б

$$P = A_1^+A = A_2^+A, \quad \bar{P} = AA_1^+ = AA_2^+,$$
$$A_1^+ = A_1^+AA_1^+ = A_2^+AA_2^+ = A_2^+.$$

Теорему доведено.

Умови (61), (62) також можна було б взяти за основу означення псевдооберненої матриці A^+. Ці умови називаються *аксіомами Мура — Пенроуза,* а матриця A^+ називається також *псевдооберненою матрицею Мура — Пенроуза.* Якщо за основу означення A^+ взяти лише частину аксіом Мура — Пенроуза, то мова йтиме про *узагальнену обернену матрицю.*

Зупинимося далі на алгоритмі обчислення $x = A^+b$ з означення де використовується QR-розвинення. Нехай $p = \mathrm{rang}\, A \leqslant \min(m, n)$. Припустимо додатково, що матрицю A за допомогою ортогональних перетворень (наприклад, відображень Хаусгольдера) без перестановки рядків можна звести до верхнього трикутного вигляду, тобто існує $Q \in \mathfrak{O}(m)$ така, що

$$QA = \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix}, \tag{63}$$

де $R \in \mathrm{Mat}_p(\mathbb{R})$ — невироджена верхня трикутна матриця, $S \in \mathrm{Mat}_{p,n-p}(\mathbb{R})$. Вектори x та Qb подамо у вигляді

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad Qb = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad x_1, \ b_1 \in \mathbb{R}^p, \quad x_2 \in \mathbb{R}^{n-p}, \quad b_2 \in \mathbb{R}^{n-p}.$$

Звернемось до леми, яка дає характеристику розв'язків задачі $\| b - Ax \| = \min$.

Лема 7. *Вектор $x \in \mathbb{R}^n$ тоді і лише тоді є розв'язком задачі $\| b - Ax \| = \min$, коли*

$$x_1 = R^{-1}b_1 - R^{-1}Sx_2.$$

Доведення. Евклідова норма інваріантна відносно ортогональних перетворень, тому

$$\| b - Ax \|^2 = \| Qb - QAx \|^2 = \| Rx_1 + Sx_2 - b_1 \|^2 + \| b_2 \|^2.$$

Останній вираз є мінімальним тоді і лише тоді, коли $Rx_1 + Sx_2 - b_1 = 0$.

Випадок $p = \mathrm{rang}\, A = n$ відповідає перевизначеній системі лінійних алгебраїчних рівнянь максимального рангу, для якої

$$QA = \begin{pmatrix} R \\ 0 \end{pmatrix},$$

дістаємо розв'язок $x = x_1 = R^{-1}b_1$. Розв'язок при $p < n \leqslant m$ дає така теорема.

Теорема 17. *Нехай* $p < n$, $V = R^{-1}S \in \mathrm{Mat}_{p,n-p}(\mathbb{R})$, $u = R^{-1}b_1 \in \mathbb{R}^p$. *Тоді єдиний розв'язок задачі* $\| b - Ax \| = \min$, *який має мінімальну норму, визначається за формулами*

$$x = (x_1, x_2) \in \mathbb{R}^p \times \mathbb{R}^{n-p},$$
$$(I + V^TV)\, x_2 = V^Tu, \quad x_1 = u - Vx_2.$$

Доведення. З попередньої леми випливає, що для розв'язку $x = (x_1, x_2)$ задачі $\| b - Ax \| = \min$, маємо $x_1 = u - Vx_2$.
Тому

$$\| x \|^2 = \| x_1 \|^2 + \| x_2 \|^2 = \| u - Vx_2 \|^2 + \| x_2 \|^2 =$$
$$= \| u \|^2 - 2\langle u, Vx_2 \rangle + \langle Vx_2, Vx_2 \rangle + \langle x_2, x_2 \rangle =$$
$$= \| u \|^2 + \langle x_2, (I + V^TV)\, x_2 - 2V^Tu \rangle \equiv \varphi(x_2),$$

причому

$$\varphi'(x_2) = -2V^Tu + 2(I + V^TV)\, x_2, \quad \varphi''(x_2) = 2(I + V^TV).$$

Оскільки матриця $I + V^TV$ є симетричною і додатно визначеною, то при x_2 такому, що $\varphi'(x_2) = 0$, тобто $(I + V^TV)\, x_2 = V^Tu$, функція $\varphi(x_2)$ набуває мінімуму. Теорему доведено.

Для розвинення матриці $I + V^TV$ на трикутні множники доцільно застосувати метод квадратного кореня, і тому для обчислення розв'язку $x = A^+b$ задачі $\| b - Ax \| = \min$ з найменшою нормою можна запропонувати такий алгоритм.

Алгоритм 5. *Обчислення* $x = A^+b$ *для* $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$, $b \in \mathbb{R}^m$ *за допомогою QR-розвинення.*

1. QR-розвинення (63) матриці A рангу p.
2. Обчислення $V \in \mathrm{Mat}_{p,n-p}(\mathbb{R})$, $RV = S$, $V = R^{-1}S$.
3. Трикутне розвинення методом квадратного кореня

$$I + V^TV = LL^T,$$

де $L \in \mathrm{Mat}_{n-p}(\mathbb{R})$ — нижня трикутна матриця.

4. $(b_1, b_2)^T = Qb$, $b_1 \in \mathbb{R}^p$, $b_2 \in \mathbb{R}^{m-p}$.
5. Обчислення $u \in \mathbb{R}^p$ з $Ru = b_1$.
6. Обчислення $x_2 \in \mathbb{R}^{n-p}$ з $LL^Tx_2 = V^Tu$.
7. Обчислення $x_1 = u - Vx_2$.

Тоді маємо $x = (x_1, x_2)^T = A^+b$.

Тепер розглянемо обумовленість задачі вирівнювання в найпростішому випадку $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$, $m \geqslant n$, $\mathrm{rang}\, A = n$. Як ми вже бачили, за цих умов розв'язок задачі $\| b - Ax \| = \min$ дається формулою

$$x = A^+b, \quad A^+ = (A^TA)^{-1}A^T.$$

Якщо за вхідні дані вважати лише вектор $b \in \mathbb{R}^m$, а матрицю A вважати сталою, то задача (f, b), $f(b) = A^+b$ має аналогічно до систем лінійних алгебраїчних рівнянь відносне число обумовленості (за нормою)

$$k_N = \frac{\| A^+ \| \| b \|}{\| A^+b \|} = \frac{\| A^+ \| \| b \|}{\| x \|}$$

та покомпонентне відносне число обумовленості

$$k_C = \frac{\| A^+ \|_\infty \| b \|_\infty}{\| A^+b \|_\infty} = \frac{\| A^+ \|_\infty \| b \|_\infty}{\| x \|_\infty}.$$

Нехай тепер елементи матриці A розглядаються як вхідні дані задачі. Розглянемо відображення $f(A) = A^+b : \mathrm{Mat}_{m,n}(\mathbb{R}) \to R^n$ і обчислимо його похідну за Гато (похідну за напрямом $C \in \mathrm{Mat}_{m,n}(\mathbb{R})$). За означенням

$$f'(A)\, C = \lim_{t \to 0} \frac{f(A + tC) - f(A)}{t} = \lim_{t \to 0} \frac{(A + tC)^+ - A^+}{t}\, b =$$
$$= \lim_{t \to 0} \frac{[(A + tC)^T (A + tC)]^{-1} (A + tC)^T - (A^TA)^{-1} A^T}{t}\, b =$$
$$= \lim_{t \to 0} \frac{[I + t(A^TA)^{-1}(A^TC + C^TA) + 0(t^2)]^{-1} (A^TA)^{-1} (A^T + C^Tt) - A^+}{t}\, b.$$

Оскільки $t \to 0$, то при достатньо малих t матимемо $\| G \| < 1$, де $G = t(A^TA)^{-1}(A^TC + C^TA) + 0(t^2)$, і тому можна скористатися розвиненням

$$(I + G)^{-1} = I - G + G^2 - G^3 + \cdots, \quad \| G \| < 1.$$

Матимемо

$$f'(A)\, C =$$
$$= \lim_{t \to 0} \frac{[I - t(A^TA)^{-1}(A^TC + C^TA) + 0(t^2)](A^TA)^{-1}(A^T + tC^T) - A^+}{t}\, b =$$
$$= [-A^+CA^+ + (A^TA)^{-1} C^T (I - AA^+)]\, b.$$

Для розв'язку $x = f(A) \equiv A^+b$ задачі $\| b - Ax \| = \min$ знаходимо

$$f'(A)\, C = -A^+Cx + (A^TA)^{-1} C^T (b - Ax).$$

Звідси для відносного покомпонентного числа обумовленості маємо

$$k_C = \frac{\| f'(A) \|_\infty \| A \|_\infty}{\| x \|_\infty} = \frac{\| A^+ \|_\infty \| A \|_\infty \| x \|_\infty + \| (A^TA)^{-1} \|_\infty \| A \|_\infty \| b - Ax \|_\infty}{\| x \|_\infty}.$$

Як бачимо, до числа обумовленості лінійної задачі вирівнювання поряд з відомим уже з розгляду лінійних алгебраїчних рівнянь складовим

$\|A^+\|_\infty \|A\|_\infty \|x\|_\infty$ входить залишок (нев'язка) $r = b - Ax$. Хоча з нормальних рівнянь випливає, що $A^T r = 0$, проте $\|A^T\|_\infty \|r\|_\infty$ дорівнює нулю лише в спеціальних випадках. Таким чином, якщо залишок «досить великий», то лінійна задача вирівнювання поводить себе по відношенню до помилок вхідних даних (в матриці A) суттєво по-іншому, ніж задача розв'язування системи лінійних алгебраїчних рівнянь. З цієї причини у випадку «великого» залишку $r = b - Ax$, $m \geqslant n$, rang $A = n$, після відшукання x із системи нормальних рівнянь рекомендується виконати одну уточнюючу ітерацію для перетвореної системи нормальних рівнянь з симетричною матрицею

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix},$$

в якій r розглядається як змінна.

Якщо відоме розвинення матриці $A \in \mathrm{Mat}_{m,n}(\mathbb{R})$ за сингулярними числами, то легко знайти і матрицю A^+. Дійсно, нехай $U^T A V = \Sigma$ — розвинення за сингулярними числами, rang $A = p$,

$$\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p, 0, \ldots, 0).$$

Тоді псевдообернена матриця $A^+ \in \mathrm{Mat}_{n,m}(\mathbb{R})$ для A має вигляд

$$A^+ = V\Sigma^+ U^T, \quad \Sigma^+ = \mathrm{diag}(\sigma_1^{-1}, \ldots, \sigma_p^{-1}, 0),$$

бо, як легко помітити, Σ^+ є псевдооберненою до Σ і з урахуванням рівностей $V^T V = I$, $U^T U = I$ неважко переконатися, що $A^+ = V\Sigma^+ U^T$ задовольняє всі аксіоми Мура — Пенроуза. Систему

$$Ax = b, \quad m \geqslant n,$$

у цьому разі можна переписати у вигляді

$$\Sigma z = d,$$

де $z = V^T x$, $d = U^T b$. Оскільки

$$\|r\| = \|Ax - b\| = \|U^T(AVV^T x - b)\| = \|\Sigma z - d\|,$$

то координати вектора z, який мінімізує $\|r\|$, можна визначити так:

$$z_i = \begin{cases} d_i/\sigma_i, & \sigma_i > \tau, \\ \text{довільне значення, якщо } \sigma_i \leqslant \tau. \end{cases}$$

Тоді $\|r\|^2 = \sum_j d_j^2$, де підсумовування виконується за тими j, для яких $\sigma_j \leqslant \tau$ або $j > p$. Величина τ дорівнює відносній похибці вхідних даних або ж якщо вхідна інформація задана точно, то $\tau = \delta = \max_i \delta_i$, де $\delta < \mathrm{eps}$.

Якщо rang $A < n$, то розв'язок, що мінімізує $\|r\|$, не єдиний. Щоб виділити єдиний розв'язок, беруть вектор r мінімальної довжини (тобто $\|r\| = \min$). У цьому разі покладають $z_i = 0$, якщо $\sigma_i \leqslant \tau$.

Якщо rang $A = n$, то розв'язок єдиний.

1.2.9. Методи розв'язування нелінійних рівнянь. Нелінійні рівняння, як і лінійні, мають характеристику розв'язку, яка апріорі оцінює міру невизначеності при неточних вхідних даних. Це число обумовленості кореня, про яке йшлося в п. 2 вступу.

Розглянемо рівняння

$$f(x) = 0, \tag{64}$$

де $f(x)$ — неперервна функція. Виникають одразу запитання: чи має рівняння (64) корені, скільки і на яких інтервалах вони розміщені? Якщо на перше запитання часто буває відповісти просто, то на друге і третє — лише в окремих випадках. Універсальним методом визначення інтервалів, на яких розміщені корені (або кажуть відокремлення коренів), є побудова графіка за допомогою ЕОМ, тобто графічне відокремлення. Коли корені відокремлені, застосовують один з ітераційних методів.

Найпростішим методом розв'язування рівняння (64) є *метод ділення навпіл*, який називають також *методом дихотомії*, або *бісекції*. Цей метод дає змогу також довести існування кореня на відрізку $[x_0, x_1]$, якщо відомо, що $f(x_0) f(x_1) \leqslant 0$, тобто неперервна функція на кінцях відрізка набуває значення різних знаків або перетворюється в нуль. Обчислення виконують за такою схемою: визначається $f(x_2)$, де $x_2 = (x_0 + x_1)/2$ і за x_3 береться те із значень x_0 чи x_1, для якого $f(x_2) \times f(x_3) \leqslant 0$, далі обчислюється $f(x_4)$, $x_4 = (x_2 + x_3)/2$, і т. д. Цей процес продовжується доти, доки довжина відрізка, який містить корінь, не стане меншою ніж 2ε. Середина останнього відрізка дає значення кореня з заданою точністю ε. Такий ітераційний процес, очевидно, збігається зі швидкістю геометричної прогресії із знаменником 1/2, тобто

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

Неважко помітити, що кількість ітерацій, потрібних для виконання нерівності $|x_n - x_{n+1}| \leqslant \varepsilon$, є $\log_2(|x_1 - x_0|/\varepsilon) + 1$. Основний недолік цього методу — повільна збіжність.

Запишемо рівняння (64) у вигляді

$$x = \varphi(x), \tag{65}$$

тут $\varphi(x) = x + \rho(x) f(x)$, $\rho(x)$ — довільна функція, яка не має коренів на відрізку $[a, b]$, де розміщено корінь рівняння (64), (65).

Метод простої ітерації визначається формулою

$$x_{n+1} = \varphi(x_n), \quad n = 0, 1, 2, \ldots, \tag{66}$$

де n — номер ітерації; x_0 — початкове наближення. З принципу стискуючих відображень дістаємо таке твердження.

Теорема 18. *Нехай функція* $\varphi(x)$ *у деякому околі* $\Delta = \{x : |x - x_0| \leqslant \delta\}$ *задовольняє умову Ліпшиця*

$$|\varphi(x'') - \varphi(x')| \leqslant q|x'' - x'| \;\forall\, x',\ x'' \in \Delta$$

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

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

Тоді рівняння (65) *має в околі* Δ *єдиний корінь* x^*, *який є границею послідовності* $\{x_n\}$, *що визначається за формулою* (66).

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

$$|z_{n+1}| = |\varphi(x_n) - \varphi(x^*)| \leqslant q|x_n - x^*| = q|z_n| \leqslant q^{n+1}|z_0|,$$

тому кажуть, що метод простої ітерації збігається зі швидкістю геометричної прогресії із знаменником q. Число ітерацій, при якому виконується нерівність

$$|x_n - x^*| \leqslant \varepsilon |x_0 - x^*|,$$

дорівнює $n \geqslant n_0(\varepsilon) = \left|\frac{\ln(1/\varepsilon)}{\ln(1/q)}\right| + 1$, де $|a|$ — найменше ціле, яке більше або дорівнює a.

Якщо функція φ має похідну на Δ, то умова Ліпшиця виконується, коли $|\varphi'(x)| \leqslant q$, $x \in \Delta$, бо тоді

$$|\varphi(x'') - \varphi(x')| = |\varphi'(\xi)|\,|x'' - x'|.$$

Більшу швидкість збіжності для рівняння (64) має *метод Ньютона.* Якщо у розвиненні

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

$$\xi = x_n + \theta(x^* - x_n), \quad 0 < \theta < 1, \tag{67}$$

де x^* — точне значення кореня, відкинути останній член і замінити x^* на x_{n+1}:

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n) \tag{68}$$

або

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad f'(x_n) \neq 0, \quad n = 0, 1, 2, \ldots, \tag{69}$$

то дістанемо *метод Ньютона,* який ще називається *методом лінеаризації* (бо в (67) залишено тільки лінійний член), або *методом дотичних* (бо (68) є рівнянням дотичної до кривої $y = f(x)$ в точці x_n, а x_{n+1} — точка її перетину з віссю абсцис).

Записавши рівняння (64) у вигляді (65), де $\varphi(x) = x - \frac{f(x)}{f'(x)}$, помічаємо, що метод Ньютона є методом простої ітерації для (65). При-

пустимо, що відрізок $[a, b]$ містить єдиний корінь x^* рівняння $f(x) = 0$ і функція $f(x)$ має неперервні похідні $f'(x)$, $f''(x)$, які не перетворюються в нуль на $[a, b]$. Тоді

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

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 18 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 19. *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 наближення* $x_0 \in [a, b]$ *задовольняє умову* $f(x_0) \times f''(x_0) > 0$, *то послідовність* $\{x_n\}$ *методу Ньютона збігається до кореня* $x^* \in [a, b]$.

Доведення. За умов теореми рівняння $f(x) = 0$ має точно один корінь x^* на $[a, b]$. Розглянемо випадок $f(a) < 0$, $f(b) > 0$, $f'(x) > 0$, $f''(x) > 0$, $x \in [a, b]$. У цьому разі 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 $x_1 = x_0 - f(x_0)/f'(x_0)$. В силу умов теореми, очевидно, маємо $x_1 < x_0$. Застосовуючи формулу Тейлора, дістаємо

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

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

тобто $x_1 \in [x^*, x_0] \in [a, b]$. Припустимо, що $x_k \in [x^*, x_{k-1}] \in [a, b]$, і доведемо, що в такому разі $x_{k+1} \in [x^*, x_k]$ (метод математичної індукції). 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 далі

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

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

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

Таким чином, $x_{k+1} \in [x^*, x_k]$, що і треба було довести. Останнє означає, що послідовність $\{x_k\}$ монотонно спадає і обмежена знизу, тобто існує границя $\lim\limits_{k\to\infty} x_k = \tilde{x}$. Перейшовши до границі в (69), переконуємося, що $\tilde{x} = x^*$. Для повного доведення теореми досить аналогічно розглянути інші можливі випадки розміщення знаків $f(a)$, $f(b)$, $f'(x)$, $f''(x)$. Теорему доведено.

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 теоремою Лагранжа

$$f(x_k) = f(x^*) + (x_k - x^*) f'(\xi), \quad \xi \in (x_k, x^*)$$

або

$$|x_k - x^*| \leqslant \frac{|f(x_k)|}{m}.$$

За формулою Тейлора

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

$$\eta \in (x_{k-1}, x_k),$$

звідки

$$|f(x_k)| \leqslant \frac{1}{2} M_2 (x_k - x_{k-1})^2,$$

і тому

$$|x_k - x^*| \leqslant \frac{M_2}{2m} (x_k - x_{k-1})^2.$$

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.

Першого з указаних недоліків позбавлений метод січних, в якому замість $f'(x_n)$ використовують різницю

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

тобто замість (69) беруть

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

Зауважимо, що така зміна цілком природна, бо при $x_{n-1} \to x_n$ маємо $f[x_{n-1}, x_n] \to f'(x_n)$.

Ті самі ідеї переносяться на нелінійні операторні рівняння

$$F(x) = 0, \tag{70}$$

де $F(x)$ — нелінійний оператор з областю визначення $D(F)$ в банаховому просторі X із значеннями у банаховому просторі Y. Оператор F називається диференційовним у точці x_0 в розумінні Фреше, якщо існує лінійний обмежений оператор $A : X \to Y$ такий, що для всіх x з деякого околу S точки x_0

$$F(x) - F(x_0) = A(x - x_0) + \omega(x - x_0),$$

причому $\| \omega(x - x_0) \| = o(\| x - x_0 \|)$ при $x \to x_0$. Оператор A називається *похідною Фреше оператора F в точці x_0* і позначається $F'(x_0)$. Отже, при близьких значеннях x та x_0

$$F(x) - F(x_0) \approx F'(x_0)(x - x_0)$$

і рівняння (70) можна наближено замінити на

$$F(x_0) + F'(x_0)(x - x_0) = 0,$$

звідки знайдемо деяке наближення до розв'язку рівняння (70):

$$x_1 = x_0 - [F'(x_0)]^{-1} F(x_0).$$

Розглянувши ітераційний процес

$$x_{k+1} = x_k - [F'(x_k)]^{-1} F(x_k), \quad k = 0, 1, \ldots, \tag{71}$$

дістанемо деяку послідовність $\{x_k\}$, яка за певних умов, накладених на оператор F, збігатиметься до розв'язку x^* рівняння (70) у нормі простору X. Ітераційний процес (71) називається *ітераційним процесом Ньютона*. Його обчислювальну схему частіше записують у вигляді

$$F'(x_k) z^k = F(x_k), \tag{72}$$

$$x_{k+1} = x_k - z^k, \quad k = 0, 1, 2, \ldots, \tag{73}$$

тобто на кожному кроці розв'язують лінійне операторне рівняння (72) і за формулою (73) знаходять наступне наближення.

Переваги процесу (72), (73) — в його швидкій збіжності при вдалому виборі початкового наближення. Недоліки — в складності вибору початкового наближення та у великому обсязі обчислювальної роботи в зв'язку з необхідністю на кожному кроці обчислювати $F(x_k)$ і $F'(x_k)$. Тому застосовують також модифікацію методу Ньютона

$$x_{k+1} = x_k - [F'(x_0)]^{-1} F(x_k), \quad k = 0, 1, \ldots,$$

в якій обернений оператор $[F'(x_0)]^{-1}$ обчислюється один раз. Швидкість збіжності цього ітераційного процесу менша, ніж у методі Ньютона.

Окремим випадком рівняння (70) є система нелінійних рівнянь

$$f_1(x_1, \ldots, x_n) = 0,$$
$$\ldots\ldots\ldots\ldots$$
$$f_n(x_1, \ldots, x_n) = 0,$$

тобто $x = (x_1, \ldots, x_n)$ — елемент n-вимірного евклідового простору, а $F(x) = (f_i(x))_{i=\overline{1,n}}$. Похідною Фреше від цього оператора в точці x є матриця Якобі

$$J(x) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \ldots & \frac{\partial f_1(x)}{\partial x_n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \frac{\partial f_n(x)}{\partial x_1} & \frac{\partial f_n(x)}{\partial x_2} & \ldots & \frac{\partial f_n(x)}{\partial x_n} \end{pmatrix}.$$

У цьому разі кожен крок ітераційного методу Ньютона зводиться до розв'язування системи лінійних алгебраїчних рівнянь

$$J(x^{(k)})\, z^{(k)} = F(x^{(k)})$$

та відшукання наступної ітерації за простою формулою

$$x^{(k+1)} = x^{(k)} - z^{(k)}, \quad k = 0, 1, \ldots.$$

Обчислення можна припиняти, наприклад, при виконанні нерівності

$$\| x^{(k+1)} - x^{(k)} \| = \left(\sum_{i=1}^{n} (x_i^{(k+1)} - x_i^{(k)})^2 \right)^{1/2} \leqslant \varepsilon,$$

де ε — наперед задане число.

1.3. Відомості з теорії лінійних просторів

Нагадаємо лише деякі означення та твердження з теорії лінійних просторів, зосередивши основну увагу на їх значенні для теорії чисельних методів та на прикладах.

1.3.1. Лінійні простори, нормовані простори та простори із скалярним добутком. Збіжність, повнота. Якщо існує деяке максимальне число лінійно незалежних векторів лінійного простору H, то це число називається розмірністю простору H. Простір, який містить нескінченну множину лінійно незалежних векторів, називається нескінченновимірним.

Приклад 1. Розглянемо простір функцій однієї змінної $x(t)$, $y(t)$, $z(t)$, ..., заданих на відрізку $[a, b]$. Вводячи природним чином операції додавання $z(t) = x(t) + y(t)$ (при кожному фіксованому t — це додавання чисел) і множення на число $u(t) = \lambda v(t)$, $t \in [a, b]$, дістаємо лінійний простір функцій однієї змінної (перевірте аксіоми лінійного простору). Цей простір нескінченновимірний, бо функції $1, t, t^2, \ldots$ лінійно незалежні (чому?).

Приклад 2. Розглянемо простір сіткових функцій, заданих на скінченній сітці $\overline{\omega}_h = \{t_i : i = \overline{0, N},\ t_i < t_{i+1}\}$. Вводячи аналогічно до попереднього прикладу операції додавання $z(t) = x(t) + y(t)$ і множення на число $u(t) = \lambda x(t)$, $t \in \omega_h$, дістаємо лінійний простір сіткових функцій. Функції $y^{(i)}(t_j) = \delta_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$, $t_j \in \overline{\omega}_h$, $i = \overline{0, N}$, лінійно незалежні (це видно, якщо встановити відповідність між сітковою функцією $y^{(i)}(t_j)$, $t_j \in \overline{\omega}_h$ та вектором $(y^{(i)}(t_0), \ldots, y^{(i)}(t_N))$, тому простір таких функцій є $(N+1)$-вимірним. Простір сіткових функцій, заданих на нескінченій сітці $\omega_\infty = \{t_j, j = 0, \pm 1, \pm 2, \ldots, t_j < t_{j+1}\}$, є нескінченновимірним.

Простір H називається *нормованим,* якщо кожному $x \in H$ ставиться у відповідність дійсне число, яке позначається $\| x \|$ (або $\| x \|_H$) і називається *нормою елемента (вектора),* що задовольняє такі аксіоми: 1) $\| x \| > 0$ при $x \neq 0$; $\| x \| = 0$ тоді і лише тоді, коли $x = 0$; 2) $\| x + y \| \leqslant \| x \| + \| y \|$ (нерівність трикутника); 3) $\| Cx \| = | C | \| x \|$, де C — число. Якщо замість аксіоми 1) виконується аксіома $\| x \| \geqslant 0$ при $x \neq 0$, $\| 0 \| = 0$ (тобто з того, що $\| x \| = 0$, не обов'язково випливає $x = 0$), то це число називається напівнормою і позначається $| x |$ (або $| x |_H$).

Дві норми $\| \cdot \|_{(1)}$ і $\| \cdot \|_{(2)}$ у просторі H називаються *еквівалентними,* якщо існують сталі $C_1 > 0$, $C_2 > 0$ такі, що

$$C_1 \| x \|_{(1)} \leqslant \| x \|_{(2)} \leqslant C_2 \| x \|_{(1)}$$

для всіх $x \in H$. Із цього означення випливає, що всі оцінки для $\| x \|_{(1)}$ мають місце і для $\| x \|_{(2)}$, але з іншими сталими множниками. Якщо ж

$$\| x \|_{(1)} \leqslant C \| x \|_{(2)} \; \forall x \in H,$$

де $C > 0$ — стала, але протилежна нерівність $\| x \|_{(2)} \leqslant C_1 \| x \|_{(1)}$ не має місця, то кажуть, що норма $\| \cdot \|_{(2)}$ сильніша, ніж $\| \cdot \|_{(1)}$, а норма $\| \cdot \|_{(1)}$ слабкіша, ніж $\| \cdot \|_{(2)}$. У цьому разі кажуть також, що норма $\| \cdot \|_{(1)}$ підкорена нормі $\| \cdot \|_{(2)}$.

Нормований простір H називається *строго нормованим,* якщо з рівності $\| x_1 \| + \| x_2 \| = \| x_1 + x_2 \|$ випливає, що $x_2 = \alpha x_1$, де α — додатна стала.

Простір H називається простором *із скалярним добутком,* якщо кожній парі його векторів x, y ставиться у відповідність дійсне число (x, y), яке називається скалярним добутком елементів x, y, причому виконуються такі аксіоми: 1) $(x, y) = (y, x)$ (симетричність); 2) $(x_1 + x_2, y) = (x_1, y) + (x_2, y)$ (дистрибутівність); 3) $(\lambda x, y) = \lambda (x, y)$, де λ — довільне дійсне число (однорідність); 4) якщо $x \neq 0$, то $(x, x) > 0$.

Скінченновимірний дійсний простір H із скалярним добутком називається *евклідовим.*

Скалярний добуток породжує в H норму

$$\| x \| = \sqrt{(x, x)}. \qquad (1)$$

У просторі із скалярним добутком має місце нерівність Коші — Буняковського — Шварца

$$|(x, y)|^2 \leqslant (x, x)(y, y), \tag{2}$$

або

$$|(x, y)| \leqslant \|x\| \|y\|.$$

Маючи такі «інструменти», як норма і скалярний добуток, можна вводити поняття збіжності послідовності $x_1, x_2, \ldots$ елементів нормованого простору, чи простору із скалярним добутком.

Послідовність $\{x_n\}$ елементів нормованого простору H збігається в H, якщо існує елемент $x_0 \in H$ такий, що $\| x_n - x_0 \| \to 0$ при $n \to \infty$. У цьому разі кажуть, що послідовність $\{x_n\}$ збігається до x_0 в нормі H (або сильно збігається до x_0) і пишуть $x_0 = \lim\limits_{n\to\infty} x_n$ (або $x_n \underset{n\to\infty}{\to} x_0$).

Розглянемо нормовані простори R_0 раціональних чисел та дійсних чисел — R, в яких норма вводиться за формулою $\|x\| = |x|$. З математичного аналізу відомо, що не всяка послідовність $\{x_n\} \in R_0$ має границю в R_0. Наприклад, послідовність раціональних наближень до $\sqrt{2}$ з недостачею $x_1 = 1$, $x_2 = 1{,}4$, $x_3 = 1{,}41$, ... не має в R_0 границі, бо $\sqrt{2} \notin R_0$. Та сама послідовність в R має границю $\sqrt{2}$, тобто є збіжною. Чи існує критерій збіжності послідовностей в нормованому просторі? Для простору R відповідь на це запитання відома — це критерій Коші: для того щоб послідовність дійсних чисел $\{x_n\}$ була збіжною, необхідно і достатньо, щоб вона була фундаментальною, тобто щоб $\forall\, \varepsilon > 0$ існував номер $N = N(\varepsilon)$ такий, що $\forall\, n > N$, і для всіх натуральних p виконувалась нерівність $|x_{n+p} - x_n| < \varepsilon$. Можна довести, що це еквівалентно умові $|x_n - x_m| \underset{n,m\to\infty}{\to} 0$. Для R_0 відомо, що коли послідовність $\{x_n\} \subset R_0$ збігається в R_0, то вона фундаментальна, проте існують фундаментальні і не збіжні в R_0 послідовності (прикладом є наведена вище послідовність 1, 1,4, 1,41, ...). Справедливість критерію Коші в R означає, що вся дійсна вісь R заповнена точками (дійсними числами) і на ній немає «дірок», тобто вона «повна». Ідея фундаментальності послідовностей лежить також в основі поняття повноти нормованого простору.

Послідовність $\{x_n\}$ нормованого простору H називається *фундаментальною*, якщо для довільного $\varepsilon > 0$ існує $N = N(\varepsilon)$ таке, що для всіх $n > N$ і для будь-яких натуральних p виконується нерівність $\| x_{n+p} - x_n \| < \varepsilon$.

Вправа 1. Довести, що наведене означення еквівалентне такому:

$$\| x_n - x_m \| \to 0 \quad \text{при } m, n \to \infty.$$

Зауважимо, що, як і у випадку числових послідовностей, має місце таке твердження.

Лема 1. *Будь-яка збіжна послідовність нормованого простору H фундаментальна.*

Проте не в кожному нормованому просторі фундаментальна послідовність збігається. Саме тому нормовані простори розподіляються на повні і неповні.

Нормований простір називається *повним*, якщо в ньому будь-яка фундаментальна послідовність збігається.

Повний нормований простір називається *банаховим*, а повний простір із скалярним добутком — *гільбертовим*.

З цього означення випливає, що гільбертів простір є банаховим з нормою $\|x\| = \sqrt{(x, x)}$.

Повнота простору має важливе значення для здобуття закінчених математичних результатів (про збіжність), зокрема, в теорії чисельних методів. По суті, лише для таких просторів можна доводити збіжність багатьох наближених методів, в чому ми переконаємось далі.

Приклад 3. Нехай $\Omega = (a, b)$, $\bar{\Omega} = [a, b]$. Розглянемо простір функцій однієї змінної, визначених на відрізку $[a, b]$, які мають на ньому неперервні похідні до k-го порядку включно. Простір таких функцій з нормою

$$\|u\|_{C^k(\bar{\Omega})} \equiv \|u\|_{C^k} = \sum_{l=0}^{k} \max_{t\in[a,b]} |u^{(l)}(t)| \tag{3}$$

позначається $C^k(\bar{\Omega})$. Зокрема, нормований простір $C^0(\bar{\Omega}) \equiv C(\bar{\Omega})$ — це простір неперервних функцій з нормою

$$\|u\|_{C(\bar{\Omega})} \equiv \|u\|_C = \max_{t\in[a,b]} |u(t)|, \tag{4}$$

яка ще називається *чебишевською нормою*. Аналогічно вводиться простір $C^k(\bar{\Omega})$ функцій n змінних, заданих в області $\bar{\Omega}$ n-вимірного евклідового простору R^n.

Вправа 2. Перевірити, що для (3) і (4) виконуються аксіоми норми.

Вправа 3. Довести, що збіжність в $C^k(\bar{\Omega})$ — це рівномірна збіжність на $\bar{\Omega}$ послідовностей функцій $\{u^{(l)}(t)\}$, $l = \overline{0, k}$.

Простір $C^k(\bar{\Omega})$ є повним, а отже, і банаховим. Це є наслідком відомої теореми математичного аналізу, яка стверджує, що з рівномірної збіжності послідовності неперервних на $[a, b]$ функцій до функції $f(x)$ випливає неперервність $f(x)$ на $[a, b]$.

Приклад 4. Нормований простір $\tilde{W}_p^m(\bar{\Omega})$, $p > 1$, $m \geqslant 0$ — це простір неперервних функцій, заданих на $\bar{\Omega}$, які мають неперервні похідні до m-го порядку включно, інтегровні разом з їхніми p-ми степенями, норма в якому визначається за формулою

$$\|u\|^p_{\tilde{W}_p^m(\bar{\Omega})} \equiv \|u\|^p_{\tilde{W}_p^m} = \|u\|^p_{\tilde{W}_p^{m-1}} + |u|^p_{\tilde{W}_p^m}, \tag{5}$$

де

$$\|u\|^p_{\tilde{W}^{m-1}_p} \equiv \|u\|^p_{\tilde{W}^{m-1}_p(\bar{\Omega})} = \sum_{i\leqslant m-1} \int_a^b |u^{(i)}(t)|^p\, dt, \tag{6}$$

$$|u|^p_{\tilde{W}^m_p} \equiv |u|^p_{\tilde{W}^m_p(\bar{\Omega})} = \int_a^b |u^{(m)}(t)|^p\, dt. \tag{7}$$

При $p = 2$ норма (5) породжується скалярним добутком

$$(u, v) = \int_a^b \sum_{i=0}^m u^{(i)}(t)\, v^{(i)}(t)\, dt, \tag{8}$$

тобто $\tilde{W}^m_2(\bar{\Omega})$ є простором із скалярним добутком (8).

Вправа 4. Перевірити, чи виконуються для (8) аксіоми скалярного добутку, а для (5) — аксіоми норми. Переконатись, що (7) визначає напівнорму.

При $m = 0$ простір $\tilde{W}^0_p(\bar{\Omega})$ має спеціальне позначення $\tilde{L}_p(\bar{\Omega})$.

Простори $\tilde{W}^m_p(\bar{\Omega})$ не є повними. В наступній вправі, наприклад, стверджується, не є повним і простір $\tilde{W}^0_2(\bar{\Omega}) \equiv \tilde{L}_2[a, b]$.

Вправа 5. Довести, що послідовність функцій

$$x_n(t) = \begin{cases} x(t), \quad t \in [a, b] \setminus \bigcup_{k=1}^{l} \left(t_k - \dfrac{\delta}{n},\ t_k + \dfrac{\delta}{n}\right), \\ \dfrac{x\left(t_k + \dfrac{\delta}{n}\right) - x\left(t_k - \dfrac{\delta}{n}\right)}{2\dfrac{\delta}{n}} \left(t - t_k + \dfrac{\delta}{n}\right) + x\left(t_k - \dfrac{\delta}{n}\right), \\ t \in \left(t_k - \dfrac{\delta}{n},\ t_k + \dfrac{\delta}{n}\right), \quad k = \overline{1, l}, \end{cases}$$

де $x(t)$ — задана функція, що має в точках t_k, $k = \overline{1, l}$, розриви першого роду і неперервна в інших точках відрізка $[a, b]$, причому $x(t_k) = (x^-(t_k) + x^+(t_k))/2$, $x^{\pm}(t_k) = \lim\limits_{t \to t_k \pm 0} x(t)$, $x^+(t_k) \neq x^-(t_k)$, $k = \overline{1, l}$, є фундаментальною в $\tilde{L}_2(a, b]$, проте граничний елемент $x(t)$ не є елементом простору неперервних функцій $\tilde{L}_2[a, b]$, не існує також іншої функції $\tilde{x}(t) \in \tilde{L}_2[a, b]$, для якої $\|x_n - \tilde{x}\|_{\tilde{L}_2} \to 0$ при $n \to \infty$.

Вправа 6. Довести, що послідовність неперервних функцій $x_n(t) = \begin{cases} \dfrac{1}{\sqrt{t}}, & t \in \left[\dfrac{1}{n}, 2\right], \\ \sqrt{n}, & t \in \left[0, \dfrac{1}{n}\right], \end{cases}$ $x_n(t) \in C[0, 2]$, фундаментальна в нормі простору

неперервних на $[0, 2]$ функцій $\tilde{L}_1[0, 2] \equiv \tilde{L}[0, 2]$ і визначається формулою $\|u\|_{\tilde{L}[0,2]} = \int_0^2 |u(t)|\, dt$, крім того, $\left\|x_n(t) - \dfrac{1}{\sqrt{t}}\right\|_{\tilde{L}[0,2]} \to 0$, але $\dfrac{1}{\sqrt{t}} \notin \tilde{L}_1[0, 2]$, бо є розривною.

Часто на практиці використовують простори $Q^k[a, b]$, які складаються з функцій, що мають похідні до k-го порядку всюди на $[a, b]$, за винятком скінченної кількості точок, де ці похідні мають розриви першого роду.

Щоб ввести простір $\tilde{W}^m_p(\bar{\Omega})$ заданих в $\bar{\Omega} \subset \mathbb{R}^n$ функцій, введемо позначення $x = (x_1, \ldots, x_n)$,

$$D^\alpha u(x) \equiv \frac{\partial^{|\alpha|} u(x)}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}} \equiv \frac{\partial^\alpha u(x)}{\partial x^\alpha}, \tag{9}$$

$x^\alpha = x_1^{\alpha_1} \ldots x_n^{\alpha_n}$, $|\alpha| = \sum\limits_{i=1}^n \alpha_j$, $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\alpha_j \geqslant 0$, $j = \overline{1, n}$,

α — мультиіндекс, α_j — цілі числа. Тоді простір $\tilde{W}^m_p(\bar{\Omega})$ — це простір неперервно диференційовних в $\bar{\Omega}$ функцій, що мають інтегровні з p-м степенем похідні до m-го порядку, в якому норма вводиться за формулами

$$\|u\|_{\tilde{W}^m_p(\bar{\Omega})} \equiv \|u\|_{\tilde{W}^m_p} \equiv \|u\|_{m,p,\bar{\Omega}} = \left(\sum_{s=0}^m |u|^p_{s,p,\bar{\Omega}}\right)^{1/p}, \tag{10}$$

$$|u|_{m,p,\bar{\Omega}} \equiv |u|_{\tilde{W}^m_p} \equiv |u|_{\tilde{W}^m_p(\bar{\Omega})} = \left(\sum_{|\alpha|=m} |D^\alpha u|^p_{0,p,\Omega}\right)^{1/p}, \tag{11}$$

$$|f|_{0,p,\bar{\Omega}} = |f|_{\tilde{L}_p(\bar{\Omega})} = \begin{cases} \left(\int\limits_{\bar{\Omega}} |f(x)|^p\, dx\right)^{1/p}, & 1 \leqslant p \leqslant \infty, \\ \max\limits_{x \in \bar{\Omega}} |f(x)|, & p = \infty. \end{cases} \tag{12}$$

Якщо $p = 2$, то цей індекс у позначеннях пропускають. Неважко перевірити, що (11) визначає напівнорму в просторі $\tilde{W}^m_p(\bar{\Omega})$.

Зазначимо, що в просторі $C^k(\bar{\Omega})$ можна ввести еквівалентну норму

$$\|u\|_{C^k} \equiv \|u\|_{C^k(\bar{\Omega})} = \max_{|\alpha|\leqslant k} \max_{x \in \bar{\Omega}} |D^\alpha u(x)| \equiv \max_{0 \leqslant s \leqslant k} |f|_{C^s(\bar{\Omega})},$$

$$|f|_{C^s(\bar{\Omega})} = \max_{|\alpha|=s} \max_{x \in \bar{\Omega}} |D^\alpha u(x)|. \tag{13}$$

Вправа 7. Довести еквівалентність норми (13) і норми

$$\|u\|_{C^k(\bar{\Omega})} = \sum_{|\alpha|\leqslant k} \max_{x\in\bar{\Omega}} |D^\alpha u(x)| = \sum_{s=0}^{k} |u|_{C^s(\bar{\Omega})}. \tag{14}$$

Приклад 5. Розглянемо простір H_h сіткових функцій, заданих на скінченній рівномірній сітці $\bar{\omega}_h = \{t_i = t_0 + ih,\ i = \overline{0, N}\}$.

У ньому можна ввести скалярний добуток за формулою

$$(y, v) = \sum_{i=0}^{N} y(t_i)\, v(t_i)\, h = \sum_{t\in\bar{\omega}_h} y(t)\, v(t)\, h, \tag{15}$$

що є аналогом скалярного добутку в просторі $\tilde{L}_2[a, b]$:

$$(y, v) = \int_a^b y(t)\, v(t)\, dt.$$

Простір сіткових функцій, заданих на сітці $\bar{\omega}_h$, із скалярним добутком (15) позначимо через $L_2(\bar{\omega}_h)$. Норма в цьому просторі вводиться таким чином:

$$\|y\|_{L_2(\bar{\omega}_h)} = \|y\|_{L_2} = \|y\|_0 = \sqrt{(y, y)}.$$

Простір H_h з нормою

$$\|y\|_{C(\bar{\omega}_h)} = \|y\|_C = \max_{x\in\bar{\omega}_h} |y(x)| = \max_{i=\overline{0,N}} |y_i|$$

позначимо через $C(\bar{\omega}_h)$. Він є сітковим аналогом простору $C[a, b]$.

Якщо сітка $\bar{\omega}_h = \{x_i = a + ih : i = \overline{0, N},\ h = (b-a)/N\}$ покриває відрізок $[a, b]$ і $y(x)$, $v(x)$ — сіткові функції, що задані на $[a, b]$, M — деяка підмножина вузлів сітки $\bar{\omega}_h$, то іноді зручно користуватися позначенням

$$(y, v)_M = \sum_{x\in M} hy(x)\, v(x).$$

Зокрема, якщо позначити

$$\omega_h = \{x_i = a + ih : i = \overline{1, N-1},\ h = (b-a)/N\},$$

$$\omega_h^+ = \{x_i = a + ih : i = \overline{1, N},\ h = (b-a)/N\},\quad \omega_h^- = \{x_i = a + ih :$$

$$i = \overline{0, N-1},\ h = (b-a)/N\},$$

то використовуються також позначення

$$(y, v) \equiv (y, v)_{\omega_h},\quad [y, v] \equiv (y, v)_{\bar{\omega}_h},\quad (y, v] \equiv$$

$$\equiv (y, v)_{\omega_h^+},\quad [y, v) \equiv (y, v)_{\omega_h^-}.$$

Підпростір простору $C[a, b]$ неперервних функцій, заданих на $[a, b]$ і таких, що перетворюються в нуль в точках a, b, позначається $\mathring{C}[a, b]$.

Важливу роль в теорії наближень відіграє поняття щільної підмножини.

Лінійний многовид L (тобто підмножина, що має всі властивості лінійного простору), який міститься в нормованому просторі E ($L \subset E$), називається *щільним* в E, якщо для будь-якого $x \in E$ і будь-якого $\varepsilon > 0$ знайдеться елемент $u \in L$ такий, що $\|x - u\| < \varepsilon$.

Якщо множина L щільна в E і $x \in E$, то для кожного $\varepsilon_n = 1/n$ можна знайти $u_n \in L$ такий, що $\|x - u_n\| < 1/n$, $n = 1, 2, \ldots$. Таким чином, якщо множина L щільна в E, то для будь-якого $x \in E$ існує послідовність $\{u_n\} \subset L$ така, що $u_n \to x$ при $n \to \infty$. Це означає, що замикання L щільної множини L, яке складається з множини L і границь всіх збіжних послідовностей, збігається з простором E.

Прикладом щільної в $C[a, b]$ множини є лінійний многовид многочленів $p_n(t) = \sum_{k=0}^{n} a_k t^k$, $n = 0, 1, \ldots$ (друга теорема Вейерштрасса). Множина раціональних чисел щільна в R, тобто в просторі дійсних чисел.

Якщо L — лінійний многовид у просторі із скалярним добутком, то сукупність всіх елементів $x \in H$, які ортогональні до L (тобто $(x, u) = 0$ для всіх $u \in L$), називається *ортогональним доповненням* до L і позначається $L^\perp$.

Лінійний многовид L щільний в гільбертовому просторі H тоді і лише тоді, коли $L^\perp = \{0\}$. Звідси випливає, що коли $x \in H$ і $(x, v) = 0$ для всіх v із щільної в H множини, то $x = 0$.

Нехай $\{x_k\}$ — скінченна чи нескінченна система елементів лінійного простору E. Множина L всіх можливих лінійних комбінацій $\sum_{k=1}^{n} c_k x_k$ при різних n називається *оболонкою системи* $\{x_k\}$.

Нехай $\{\varphi_k\}$ — ортонормальна послідовність у гільбертовому просторі H, тобто $(\varphi_k, \varphi_j) = \delta_{jk}$, де $\delta_{jk} = \begin{cases} 1, & j = k, \\ 0, & j \neq k \end{cases}$ — символ Кронекера. Числа $c_k = \frac{(x, \varphi_k)}{\|\varphi_k\|^2}$, $k = 1, 2, \ldots$, називаються *коефіцієнтами Фур'є* елемента x за ортогональною системою $\{\varphi_k\}$, а ряд $\sum_{k=1}^{\infty} c_k \varphi_k$ — рядом Фур'є. Цей ряд збіжний, якщо існує $x_0 \in H$ такий, що $\|s_n - x_0\|_{n\to\infty} \to 0$, де $s_n = \sum_{k=1}^{n} c_k \varphi_k$ — часткова сума ряду Фур'є.

Ортогональна система $\{\varphi_k\}$ гільбертового простору H називається *повною*, якщо для будь-якого $x \in H$ ряд Фур'є, утворений для x, збігається до x, тобто $\sum_{k=1}^{\infty} c_k \varphi_k = x$. Критеріями повноти ортогональної

системи елементів $\{\varphi_k\}$ можуть бути такі твердження: а) для того щоб $\{\varphi_k\}$ була повною, необхідно і достатньо, щоб виконувалась *рівність Парсеваля — Стєклова*

$$\sum_{k=1}^{\infty} |c_k|^2 \|\varphi_k\|^2 = \|x\|^2;$$

б) ортогональна система $\{\varphi_k\}$ гільбертового простору H повна тоді і лише тоді, коли її лінійна оболонка L щільна в H, тобто $\overline{L} = H$.

Множина M нормованого простору X називається *компактною*, якщо з кожної послідовності $\{x_n\} \subset M$ можна виділити фундаментальну підпослідовність.

Множина M банахового простору X називається *бікомпактною*, якщо з кожної послідовності $\{x_n\} \subset M$ можна виділити збіжну підпослідовність, границя якої належить M.

Зазначимо, що коли X — банахів простір, то фундаментальна послідовність в силу повноти X збігається до деякого елемента $x_0 \in X$, але не обов'язково $x_0 \in M$, якщо $\{x_n\} \subset M$. Таким чином, поняття компактної множини слабкіше поняття бікомпактної множини. Очевидно, що для замкненої множини M з банахового простору X поняття компактності та бікомпактності збігаються.

Два лінійних простори X та $\tilde{X}$ називаються ізоморфними, якщо існує функція (відображення, оператор) $x = \tilde{J}(\tilde{x})$, $J : X \to \tilde{X}$, яка здійснює лінійну взаємно однозначну відповідність між X та $\tilde{X}$, тобто: 1) $J(\lambda x + \mu y) = \lambda J(x) + \mu J(y)$ для всіх $x, y \in X$ і чисел λ, μ; 2) якщо $J(x_1) = J(x_2)$, то $x_1 = x_2$; 3) для будь-якого $\tilde{x} \in \tilde{X}$ знайдеться $x \in X$ такий, що $\tilde{x} = J(x)$.

Два нормованих простори E і $\tilde{E}$ називаються *лінійно ізометричними*, якщо існує функція $J(x)$, яка здійснює ізоморфізм E та $\tilde{E}$ як лінійних просторів, і така, що $\|J(x)\| = \|x\|$. Ізометричні простори, очевидно, нічим не відрізняються за властивостями, зв'язаними з операціями додавання, множення на число, а також з нормою (наприклад, якщо послідовність $\tilde{x}_n \in \tilde{E}$ збігається в $\tilde{E}$, то $\{x_n\}$ збігається в E і т. п.).

Найпростішими прикладами компактних множин (компактів) є відрізок $[a, b] \subset R$ та замкнена область $\bar{\Omega}$ на площині (x_1, x_2) (в останньому прикладі $x_n = (x_{1n}, x_{2n})$, $\|x_n\| = \sqrt{x_{1n}^2 + x_{2n}^2}$). Як відомо, неперервна функція $f(x)$, задана на компакті $[a, b]$, набуває на ньому найбільшого і найменшого значень (теорема Коші). Аналогічні властивості мають компакти в функціональних просторах.

1.3.2. Поповнення нормованих просторів і просторів із скалярним добутком. Будь-який нормований простір можна перетворити в банахів, застосувавши так звану процедуру замикання. При цьому, якщо

йдеться про нормований простір звичайних функцій дійсної змінної, нам доведеться ввести до розгляду деякі нові математичні об'єкти — класи функцій, що і будуть елементами нового повного нормованого простору. Ідея такої побудови була застосована ще французьким математиком Коші при створенні теорії дійсних чисел, які він розглядав як класи еквівалентних фундаментальних послідовностей раціональних чисел. Наводимо теорему, яка ілюструє також конструкцію замикання (поповнення).

Теорема 1. *Будь-який нормований простір E можна розглядати як лінійний многовид, щільний у деякому банаховому просторі $\hat{E}$, який називається поповненням простору E.*

Доведення. Розглянемо множину всіх фундаментальних послідовностей простору E. Дві такі послідовності $\{x_n\}$ і $\{x'_n\}$ називатимемо еквівалентними, якщо $\|x_n - x'_n\| \to 0$ при $n \to \infty$. Якщо $\{x_n\}$ і $\{x'_n\}$ еквівалентні, то позначатимемо це $\{x_n\} \sim \{x'_n\}$.

Множину всіх фундаментальних послідовностей розіб'ємо на класи, що не перетинаються, таким чином. Дві такі послідовності $\{x_n\}$ та $\{x'_n\}$ включимо до одного класу тоді і лише тоді, коли $\{x_n\} \sim \{x'_n\}$. Множину всіх класів позначимо через $\hat{E}$, а самі класи позначатимемо через $\hat{x}, \hat{y}, \ldots$. Якщо $\{x_n\}$ входить до класу $\hat{x}$, то позначатимемо це так: $\{x_n\} \in \hat{x}$ і називатимемо послідовність $\{x_n\}$ представником класу $\hat{x}$.

Введемо в $\hat{E}$ операції додавання і множення на число, тобто перетворимо його в лінійний простір. Операцію додавання класів $\hat{x}$ та $\hat{y}$ означимо так: якщо $\{x_n\} \in \hat{x}$, $\{y_n\} \in \hat{y}$, то сумою класів $\hat{x} + \hat{y}$ називатимемо клас, що містить $\{x_n + y_n\}$ (неважко довести, що послідовність $\{x_n + y_n\}$ фундаментальна, якщо фундаментальні $\{x_n\}$ і $\{y_n\}$). Щоб ця операція додавання була коректною, треба довести, що означення суми $\hat{x} + \hat{y}$ не залежить від вибору представників класів $\hat{x}$ та $\hat{y}$. Дійсно, візьмемо інших представників $\{x'_n\} \in \hat{x}$, $\{y'_n\} \in \hat{y}$. З означення класів випливає, що $\{x'_n\} \sim \{x_n\}$, $\{y'_n\} \sim \{y_n\}$, а тому $\{x'_n + y'_n\} \sim \{x_n + y_n\}$. Це означає, що $\{x'_n + y'_n\} \in \hat{x} + \hat{y}$, тобто сума класів не залежить від вибору представників.

Операцію множення класу $\hat{x}$ на число λ введемо аналогічно: класом $\lambda\hat{x}$ називатимемо клас, який містить $\{\lambda x_n\}$, де $\{x_n\}$ — представник $\hat{x}$, тобто $\{x_n\} \in \hat{x}$. Коректність цього означення випливає з таких тверджень, що легко доводяться: 1) якщо $\{x_n\}$ фундаментальна, то $\{\lambda x_n\}$ також фундаментальна; 2) якщо $\{x'_n\} \sim \{x_n\}$, то $\{\lambda x'_n\} \sim \{\lambda x_n\}$; 3) означення класу $\lambda\hat{x}$ не залежить від вибору представника класу $\hat{x}$.

Оскільки ці означення операцій в $\hat{E}$ зводяться до операцій над елементами лінійного простору E, то для $\hat{E}$ виконуються всі аксіоми лінійного простору, тому $\hat{E}$ — лінійний простір. Роль нуля в E відіграє клас 0 з представником $\{0, 0, \ldots\}$.

Перетворимо $\hat{E}$ у нормований простір, вводячи норму за формулою

$$\|\hat{x}\|_{\hat{E}} = \lim_{n\to\infty} \|x_n\|_E, \tag{16}$$

де $\{x_n\} \in \hat{x}$. Ця границя існує, оскільки числова послідовність $\{\|x_n\|\}$ збігається. Дійсно, $\{\|x_n\|\}$ — фундаментальна послідовність, бо $|\|x_n\| - \|x_m\|| < \|x_n - x_m\| \to 0$ при $n, m \to \infty$, і в силу критерію Коші вона є збіжною. Неважко також помітити, що границя (16) не залежить від вибору представника класу $\hat{x}$: якщо $\{x_n'\} \in \hat{x}$, то $|\|x_n'\| - \|x_n\|| \leqslant \|x_n' - x_n\| \to 0$ при $n \to \infty$, тобто $\lim\limits_{n\to\infty} \|x_n'\|_E = \lim\limits_{n\to\infty} \|x_n\|_E$. Для остаточної впевненості в коректності формули (16) залишається перевірити виконання аксіом норми.

Таким чином, ми побудували нормований простір $\hat{E}$, елементами якого є нові об'єкти — класи фундаментальних еквівалентних послідовностей. Для доведення теореми 1 про поповнення нам тепер потрібно довести такі три твердження: А) E можна ототожнити з деяким лінійним многовидом в $\hat{E}$; Б) E — щільна в $\hat{E}$ в смислі вказаного ототожнення; В) $\hat{E}$ — повний, тобто банахів простір.

Почнемо з доведення твердження А). Елемент $x \in E$ ототожнимо з класом, який містить послідовність $\{x, x, \ldots, \}$, що називається *стаціонарною*. Позначимо цей клас через x. Означимо λx як клас, що містить стаціонарну послідовність $\{\lambda x, \lambda x, \ldots\}$, а $x + y$ — як клас, що містить $\{x + y, x + y, \ldots\}$. Таким чином, множина всіх класів, які містять стаціонарні послідовності, є лінійним многовидом в $\hat{E}$, для якого ми залишимо позначення E. Доведемо тепер твердження Б). Якщо клас $x \in E$, то

$$\|x\|_{\hat{E}} = \lim_{n\to\infty} \|x\|_E = \|x\|_E$$

як границя сталої.

Нехай $\hat{x} \in \hat{E}$. Доведемо, що існує така послідовність класів $\{x_n\} \subset E$, для якої $\|x_n - \hat{x}\|_E \to 0$ при $n \to \infty$. Це і означатиме щільність E в $\hat{E}$. Для доведення розглянемо $\{x_n\}$ як представника класу $\hat{x}$, $\{x_n\} \in \hat{x}$. Із фундаментальності $\{x_n\}$ випливає: для будь-якого $\varepsilon > 0$ існує такий номер $N = N(\varepsilon)$, що для всіх $n, m > N$ виконується нерівність

$$\|x_n - x_m\|_E < \varepsilon. \tag{17}$$

Зафіксуємо $n > N$ і спрямуємо m до нескінченності, внаслідок чого дістанемо

$$\lim_{m\to\infty} \|x_n - x_m\|_E = \|x_n - \hat{x}\|_{\hat{E}}, \tag{18}$$

де $x_n \in \hat{E}$ — клас, що містить стаціонарну послідовність $\{x_n, x_n, \ldots, x_n, \ldots\}$. Переходячи в (17) до границі при $m \to \infty$ і враховуючи (18), дістанемо нерівність $\|x_n - \hat{x}\|_{\hat{E}} \leqslant \varepsilon$ при $n > N$, а це означає, що $x_n \to \hat{x}$ при $n \to \infty$.

І нарешті, доведемо В). Нехай дано послідовність $\{\hat{x}_n\}$, яка є фундаментальною в $\hat{E}$. В силу Б) знайдеться послідовність класів $\{x_n\} \subset E$ така, що

$$\|\hat{x}_n - x_n\|_{\hat{E}} < \frac{1}{n}.$$

Послідовність класів $\{x_n\}$ буде також фундаментальною в $\hat{E}$, бо

$$\|x_n - x_m\|_{\hat{E}} \leqslant \|x_n - \hat{x}_n\|_E + \|\hat{x}_n - \hat{x}_m\|_{\hat{E}} + \|\hat{x}_m - x_m\|_{\hat{E}} < \frac{1}{n} + \\ + \|\hat{x}_n - \hat{x}_m\|_{\hat{E}} + \frac{1}{m} \xrightarrow[m,n\to\infty]{} 0.$$

Оскільки в смислі А) кожен клас $x_n \in E$ містить стаціонарну послідовність $\{\bar{x}_n, \bar{x}_n, \ldots\}$ елементів $\bar{x}_n$ вихідного нормованого простору E, то послідовність класів $\{x_n\}$ визначає також послідовність $\{\bar{x}_n\}$ з простору E. В силу фундаментальності послідовності класів $\{x_n\}$ в $\hat{E}$ послідовність $\{\bar{x}_n\}$ буде фундаментальною в E, бо

$$\|\bar{x}_n - \bar{x}_m\|_E = \|x_n - x_m\|_{\hat{E}}.$$

Але тоді існує клас $\hat{x}$, який має представника $\{\bar{x}_n\}$. Доведемо, що $\hat{x}_n \to \hat{x}$ в $\hat{E}$ при $n \to \infty$. Дійсно,

$$\|\hat{x}_n - \hat{x}\|_{\hat{E}} \leqslant \|\hat{x}_n - x_n\|_{\hat{E}} + \|\hat{x} - x_n\|_{\hat{E}} < \frac{1}{n} + \|\hat{x} - x_n\|_{\hat{E}}. \tag{19}$$

Представником класу $\hat{x} - x_n$ є послідовність $\{\bar{x}_m - \bar{x}_n\}_{m=1}^{\infty}$ і за означенням

$$\|\hat{x} - x_n\|_{\hat{E}} = \lim_{m\to\infty} \|\bar{x}_m - \bar{x}_n\|_E.$$

Звідси випливає, що в силу фундаментальності послідовності елементів $\{\bar{x}_n\}_{n=1}^{\infty}$ у вихідному просторі E

$$\lim_{n\to\infty} \|\hat{x} - x_n\|_{\hat{E}} = \lim_{n\to\infty} \lim_{m\to\infty} \|\bar{x}_m - \bar{x}_n\|_E = 0.$$

Тепер з (19) випливає, що клас $\hat{x}$ є границею послідовності класів $\{\hat{x}_n\}$, що і треба було довести в В). Теорему повністю доведено.

Нехай тепер вихідний простір E є простором із скалярним добутком $(,)$. Поповнюючи E як нормований простір з нормою $\|x\| = \sqrt{(x, x)}$, дістанемо банахів простір $\hat{E}$, елементами якого є класи $\hat{x}$ еквівалентних фундаментальних послідовностей $\{x_n\}$. У просторі $\hat{E}$ можна визначити скалярний добуток будь-яких двох елементів $\hat{x}, \hat{y} \in \hat{E}$ за формулою

$$(\hat{x}, \hat{y}) = \lim_{n\to\infty} (x_n, y_n), \tag{20}$$

де $\{x_n\}$, $\{y_n\}$ — представники відповідно елементів $\hat{x}$, $\hat{y}$, причому

$$(\hat{x}, \hat{x}) = \lim_{n\to\infty} (x_n, x_n) = \lim_{n\to\infty} \|x_n\|^2 = \|\hat{x}\|^2.$$

Неважко помітити, що виконуються всі аксіоми скалярного добутку. Отже, в силу повноти простір класів $\hat{E}$ із скалярним добутком (20) є гільбертовим простором.

Якщо тепер повернутися до прикладу 4, то стає зрозумілим, як зробити простори $\tilde{W}_p^m(\bar{\Omega})$ повними (поповнити їх). Такі простори називаються *просторами Соболєва* і позначаються $W_p^m(\Omega)$, $p \geqslant 1$. Поповнення простору $\tilde{W}_p^0(\bar{\Omega}) \equiv \tilde{L}_p(\bar{\Omega})$ позначається через $L_p(\Omega)$, і його елементами є не звичайні неперервні функції, а класи фундаментальних і еквівалентних в нормі $\tilde{L}_p(\bar{\Omega})$ (або фундаментальних і еквівалентних у середньому) послідовностей неперервних в $\bar{\Omega}$ функцій. Розглянута в попередньому пункті послідовність $\{x_n(t)\}$ — це представник деякого класу, який є елементом повного простору $L_2(a, b)$, і цей клас можна ототожнити з розривною функцією $x(t)$.

Відносно простору $L(\Omega)$ можна довести таке твердження: нехай $x(t)$ визначена на $\bar{\Omega}$ і має на $\bar{\Omega}$ скінченну кількість точок розриву, причому збігається інтеграл

$$\int_{\Omega} |x(t)|\, dt;$$

тоді існує фундаментальна в нормі $L(\Omega)$ послідовність неперервних на $\bar{\Omega}$ функцій $\{x_n(t)\}$ така, що

$$\|x_n - x\|_{L(\Omega)} = \int_{\Omega} |x_n(t) - x(t)|\, dt \to 0, \quad n \to \infty$$

(інтеграли треба розуміти як невласні), і тому $x(t) \in L(\Omega)$.

Аналогічно всі неперервні і деякі розривні на $\bar{\Omega}$ функції можна ототожнити з деякими класами в $L_p(\Omega)$, $p > 1$.

В силу повноти в класах Соболєва $W_p^m(\Omega)$ можна здобувати завершені математичні результати і процедура поповнення показує, що будь-який елемент цього простору можна як завгодно добре наблизити неперервно диференційовними функціями, що свідчить про зручність використання цих класів і в чисельному аналізі.

1.4. Відомості з теорії лінійних операторів

1.4.1. Лінійні оператори в лінійних просторах. Нехай H, V — деякі лінійні простори, D — підпростір H. Якщо кожному вектору $x \in D$ за деяким правилом поставлений у відповідність єдиний вектор $y = Ax \in V$, то кажуть, що задано відображення або оператор $A : H \to V$. Множина $D \subset H$ називається *областю визначення* оператора A і позначається $D(A)$. Множина всіх векторів вигляду $y = Ax$, $x \in D(A)$ називається *областю значень оператора* A і позначається $R(A)$. Якщо $V = R$, де R — множина дійсних чисел, то відображення A називається *функціоналом*. Якщо $D(A) = H$, то кажуть, що оператор A заданий на H. Найчастіше розглядатимемо оператори $A : H \to H$, які діють у просторі H.

Оператор (функціонал) A називається *лінійним*, якщо виконуються умови: 1) A є адитивним, тобто $A(x_1 + x_2) = Ax_1 + Ax_2$ для будь-яких $x_1, x_2 \in D(A)$; 2) A є однорідним, тобто для будь-яких λ і $x \in D(A)$ має місце рівність

$$A(\lambda x) = \lambda A x. \tag{1}$$

Умови 1), 2) еквівалентні умові $A(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1 A x_1 + \lambda_2 A x_2$ для будь-яких чисел λ_1, λ_2 і для будь-яких векторів $x_1, x_2 \in D(A)$.

Приклад 1. Нехай H є лінійним простором неперервних на $[a, b]$ функцій $f(x)$ і x^* — фіксована точка у проміжку $[a, b]$. Поставимо кожній функції $f(x) \in H$ у відповідність дійсне число y за правилом

$$y = F(f) \equiv f(x^*),$$

чим задамо на H лінійний функціонал $F : H \to R$. Ту саму відповідність можна розглядати як оператор в H, якщо областю значень F вважати множину функцій, які є сталими на $[a, b]$.

Вправа 1. Перевірити лінійність оператора (функціонала) F.

Приклад 2. Нехай H — множина неперервних на $[a, b]$ функцій, $D(A)$ — підмножина в H функцій, які мають неперервні похідні до k-го порядку включно. Тоді в H можна задати лінійний оператор за правилом

$$y = Af \equiv \frac{d^k f}{dx^k}.$$

Якщо зафіксувати x^*, то відображення $A : H \to R$

$$y = Af \equiv \frac{d^k f(x^*)}{dx^k}$$

є лінійним функціоналом.

Приклад 3. Нехай H — множина заданих на $[a, b]$ інтегровних функцій; $K \subset H$ — множина функцій, заданих на $[a, b]$, які є сталими. Тоді лінійний оператор на H можна задати правилом

$$y = Ax \equiv \int_a^b x(t)\, dt.$$

Тут $D(A) = H$, $R(A) = K$. Цей оператор можна розглядати також як функціонал.

Приклад 4. Нехай $H = R$; $f(x)$ — задана на $[a, b] \subset R$ функція. Тоді формула

$$y = Ax \equiv f(x), \quad x \in [a\ b]$$

задає в H оператор $A : H \to H$, причому $D(A) = [a, b]$, $R(A) \subset H$. Очевидно, що в цьому випадку оператор A буде лінійним лише тоді, коли $f(x) = \alpha x$, де α — стала.

Приклад 5. Нехай H — простір неперервних на $[a, b]$ функцій; $D(A) \subset H$ — множина k разів неперервно диференційованих функцій $y(x)$, які задовольняють умови $y(a) = y_0$, $y'(a) = y_1$, ..., $y^{(k-1)}(a) = y_{k-1}$, де y_i, $i = \overline{0, k-1}$,— задані числа. Тоді в H можна задати такий лінійний оператор з областю визначення $D(A)$:

$$v = Ay \equiv a_0(x)\, y^{(k)}(x) + a_1(x)\, y^{(k-1)}(x) + \cdots + a_{k-1}(x)\, y(x) + a_k(x),$$

де $a_i(x)$, $i = \overline{0, k}$,— задані неперервні функції. Операторне рівняння $Ay = f$, де f — задана неперервна функція, $y \in D(A)$ — невідома функція, являє собою відому в теорії диференціальних рівнянь задачу Коші.

Приклад 6. Замінивши в попередньому прикладі область визначення на множину $D(A)$, яка є множиною k разів неперервно диференційовних функцій $y(x)$, що задовольняють умови $y^{(i)}(a) = y_{a,i}$, $i = \overline{0, l-1}$, $y^{(j)}(b) = y_{b,j}$, $j = \overline{0, k-l-1}$, де $y_{a,i}$, $y_{b,j}$ — задані числа, дістанемо ще один лінійний оператор в H. Операторне рівняння $Ay = f$, де f — задана функція з H, являє собою *крайову задачу*.

Оператор A, який діє в лінійному нормованому просторі H, називається *обмеженим*, якщо існує така стала $M > 0$, що

$$\|Ax\| \leqslant M\|x\| \ \forall x \in H. \tag{2}$$

Точна нижня грань множини чисел M, які задовольняють (2), називається *нормою оператора* A і позначається $\|A\|$. Очевидно, що

$$\|Ax\| \leqslant \|A\|\|x\|. \tag{3}$$

Норму оператора $A : H \to V$, де H, V — нормовані простори, можна визначити також формулою

$$\|A\| = \sup_{\|u\|_H \neq 0} \frac{\|Au\|_V}{\|u\|_H} = \sup_{\|u\|_H = 1} \|Au\|_V. \tag{4}$$

Звідси випливає нерівність

$$\|Au\|_V \leqslant \|A\|\|u\|_H. \tag{5}$$

У скінченновимірному просторі будь-який лінійний оператор обмежений.

Нехай X, Y — нормовані простори і $A : X \to Y$ — лінійний оператор (функціонал), $D(A) = X$.

Оператор (функціонал) A називається *неперервним у точці* $x_0 \in X$, якщо $Ax \to Ax_0$ при $x \to x_0$, тобто $\|Ax - Ax_0\| \to 0$ при $\|x - x_0\|_X \to 0$.

Для лінійних операторів поняття неперервності та обмеженості випливають одне з одного, тобто, для того щоб лінійний оператор A був неперервним, необхідно і достатньо, щоб він був обмеженим.

Якщо $f(x)$ — дійсний неперервний функціонал, визначений на бікомпактній множині Q, то існують такі $x_0 \in Q$, $x^0 \in Q$, що

$$f(x_0) = \inf_{x \in Q} f(x), \quad f(x^0) = \sup_{x \in Q} f(x)$$

(порівняйте з теоремою Коші для неперервних функцій).

Якщо кожному $y \in H$ відповідає лише один вектор $x \in H$, для якого $Ax = y$, то цією відповідністю визначається оператор A^{-1}, який називається *оберненим*, $A^{-1} : H \to H$. З цього означення випливає, що

$$A^{-1}(Ax) = x, \quad A(A^{-1}y) = y \text{ для всіх } x, y \in H.$$

Очевидно, що A^{-1} є лінійним оператором, якщо лінійним є оператор A. Має місце таке твердження: для того щоб лінійний оператор $A : H \to H$ мав обернений, необхідно і достатньо щоб рівняння $Ax = 0$ мало єдиний розв'язок $x = 0$. Оператор D, який діє за правилом $Dx = A(Bx)$, називається *добутком операторів* A і B і позначається $D = AB$. Оператор I називається *одиничним (тотожним)*, якщо $Ix = x$ для всіх $x \in H$. Якщо існує A^{-1}, то $A^{-1}A = AA^{-1} = I$. Оператори A і B називаються *перестановочними*, якщо $AB = BA$.

Часто буває корисною така теорема.

Теорема Банаха. *Якщо X — банахів простір і v — лінійний оператор $v : X \to X$, то за умови $\|v\| \leqslant q < 1$ оператор $I - v$ має неперервний обернений, причому* $\|(I - v)^{-1}\| \leqslant \frac{1}{1-q}$.

Якщо H — простір із скалярним добутком і A — лінійний оператор у ньому, то оператор $A^* : H \to H$ називається *спряженим до* $A : H \to H$, якщо

$$(Ax, y) = (x, A^*y)$$

для всіх $x, y \in H$. Оператор A називається *самоспряженим (симетричним)*, якщо $A = A^*$ (або $(Ax, y) = (x, Ay)$). Називатимемо лінійний оператор A *додатним*, якщо $(Ax, x) > 0$, *невід'ємним*, якщо $(Ax, x) \geqslant 0$, і *додатно визначеним*, якщо $(Ax, x) \geqslant \delta(x, x) \equiv \delta\|x\|^2$ для всіх $x \in H$, $x \neq 0$, причому δ — додатна стала.

Довільний оператор A можна подати у вигляді суми $A = A_0 + A_1$, де $A_0 = \frac{1}{2}(A + A^*)$ — самоспряжений (симетричний) опера-

тор; $A_1 = \frac{1}{2}(A - A^*)$ — кососиметричний оператор, для якого в дійсному просторі $(A_1x, x) = -(x, A_1x) = -(A_1x, x)$, і, отже, $(A_1x, x) = 0$. Таким чином, для довільного оператора A в просторі H виконується рівність $(Ax, x) = (A_0x, x)$ для всіх $x \in H$.

Якщо виконуються нерівності $(Ax, x) \geqslant 0$, $(Ax, x) > 0$, $(Ax, x) \geqslant \delta \| x \|^2$ для всіх $x \in H$, $x \neq 0$, то відповідно записуємо $A \geqslant 0$, $A > 0$ і $A \geqslant \delta I$. Нерівність $B \geqslant \alpha A$ означає, що $B - \alpha A \geqslant 0$, тобто $((B - \alpha A)x, x) \geqslant 0$ для всіх $x \in H$. Якщо в дійсному просторі $A \neq A^*$, то нерівності $A \geqslant 0$, $A > 0$ еквівалентні відповідно нерівностям $A_0 \geqslant 0$, $A_0 > 0$.

При $A > 0$ існує оператор $A^{-1} : H \to H$, причому $A^{-1} > 0$, а при $A = A^*$ має місце рівність $(A^{-1})^* = A^{-1}$. Дійсно, A^{-1} існує лише тоді, коли рівняння $Ax = 0$ має лише тривіальний розв'язок. Припустивши, що A^{-1} не існує, ми мали б $Ax = 0$ для деякого $x \neq 0$, і тоді $0 = (Ax, x)$ при $x \neq 0$, що суперечить умові $A > 0$.

Нехай A — самоспряжений лінійний оператор в N-вимірному просторі H із скалярним добутком. Задача про власні значення оператора A ставиться таким чином: знайти значення λ_i параметра λ, для яких рівняння $Ax = \lambda x$ має нетривіальні розв'язки x_i. Числа λ_i називаються *власними значеннями* оператора A, а елементи (вектори) x_i — його *власними векторами*.

Наведемо основні елементи з лінійної алгебри для задачі про власні значення в N-вимірному просторі із скалярним добутком.

1. Самоспряжений оператор A має N власних значень λ_i, $i = \overline{1, N}$, $0 \leqslant |\lambda_1| \leqslant \ldots \leqslant |\lambda_N|$, яким відповідають ортонормовані власні вектори ξ_i, $i = \overline{1, N}$; $(\xi_i, \xi_j) = \delta_{ij}$; $\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$ — символ Кронекера.

2. Якщо $A > 0$, то $0 < \lambda_1 \leqslant \ldots \leqslant \lambda_N$.

3. Довільний вектор $x \in H$ можна розкласти за власними векторами самоспряженого оператора, тобто подати у вигляді

$$x = \sum_{k=1}^{N} c_k \xi_k,$$

причому $\| x \|^2 = \sum_{k=1}^{N} c_k^2$, де $c_k = (x, \xi_k)$ — коефіцієнти Фур'є вектора x.

4. Якщо $A = A^* > 0$, то розв'язок рівняння $Ax = f$ можна подати у вигляді

$$x = \sum_{k=1}^{N} \frac{f_k}{\lambda_k} \xi_k,$$

де $f_k = (f, \xi_k)$ — коефіцієнти Фур'є вектора f.

5. Норма самоспряженого оператора дорівнює максимальному з модулів його власних значень, тобто $\| A \| = \max_{k=\overline{1,N}} |\lambda_k| = |\lambda_N|$.

6. Якщо $A = A^*$, то

$$\| A \| = \sup_{\|x\|=1} |(Ax, x)|.$$

7. Якщо $A = A^* > 0$, то $\lambda_1 I \leqslant A \leqslant \lambda_N I$ або

$$\lambda_1 \| x \|^2 \leqslant (Ax, x) \leqslant \lambda_N \| x \|^2, \quad \lambda_1 > 0, \quad x \in H.$$

8. Якщо оператор A в N-вимірному просторі H додатний, то він і додатно визначений, тобто існує стала $\delta > 0$ така, що з умови $A > 0$ випливає нерівність $A \geqslant \delta I$.

9. Якщо існує оператор Q^{-1}, то операторні нерівності $C \geqslant 0$ та $Q^*CQ \geqslant 0$ еквівалентні, що є наслідком тотожності

$$(Q^*CQx, x) = (CQx, Qx) = (Cy, y),$$

де $y = Qx$, $x = Q^{-1}y$.

10. Якщо A_1, A_2 — лінійні самоспряжені додатні і перестановочні оператори в H, то оператори A_1, A_2, $A_1 + A_2$, A_1A_2 мають спільну систему власних векторів $\{\xi_k\}$, причому

$$\lambda(A_1 + A_2) = \lambda(A_1) + \lambda(A_2), \quad \lambda(A_1A_2) = \lambda(A_1)\lambda(A_2),$$

де через $\lambda(B)$ позначаються власні значення оператора B.

11. Якщо $A = A^* > 0$, то існує оператор A^{-1}, який є самоспряженим, додатно визначеним, має ті самі власні вектори, що й оператор A, і власні значення $\lambda(A^{-1}) = \lambda^{-1}(A)$.

1.4.2. Різницеві оператори в лінійному просторі сіткових функцій. Позначимо через Ω_{N+1} простір сіткових функцій, які задані на сітці $\bar{\omega}_N = \{i : i = \overline{0, N}\}$, а через $\mathring{\Omega}_{N+1}$ — підпростір — простору Ω_{N+1}, що складається з функцій, заданих на $\bar{\omega}_N$, які перетворюються на нуль у граничних вузлах сітки $\bar{\omega}_N$, тобто $y_0 = y_N = 0$. Функції з простору $\mathring{\Omega}_{N+1}$ позначатимемо $\mathring{y}_i = \mathring{y}(i)$. Розглянемо приклади найпростіших різницевих операторів.

Для оператора правої різниці $\Delta y_i = y_{i+1} - y_i$, $i = \overline{0, N-1}$, областю визначення є $D(\Delta) = \Omega_{N+1}$, а областю значень — простір Ω_N функцій, заданих на сітці $\omega_N^- = \{i : i = \overline{0, N-1}\}$. Простір Ω_N^- має розмірність N.

Для оператора лівої різниці $\nabla y_i = y_i - y_{i-1}$, $i = \overline{1, N}$, областю визначення є $D(\nabla) = \Omega_{N+1}$, а областю значень — N-вимірний підпростір Ω_N^+ функцій, заданих на сітці $\omega_N^+ = \{i : i = \overline{1, N}\}$.

Оператор

$$\Delta^2 y_i = \Delta(\Delta y_{i-1}) = \Delta(\nabla y_i) = y_{i+1} - 2y_i + y_{i-1}$$

визначений для сіткових функцій з Ω_{N+1} і відображає Ω_{N+1} у простір Ω_{N-1} функцій, визначених на сітці $\omega_N = \{i : i = \overline{1, N-1}\}$. Оператор

$$\Lambda y_i = b_i y_{i+1} - c_i y_i + a_i y_{i-1} = b_i \Delta(\nabla y_i) - (b_i - a_i)\nabla y_i - (c_i - a_i - b_i) y_i, \quad i = \overline{1, N-1},$$

також відображає Ω_{N+1} в Ω_{N-1}, тобто $\Lambda : \Omega_{N+1} \to \Omega_{N-1}$.

Розглянемо операторне рівняння

$$\Lambda y_i = -f_i, \ i = \overline{1, N-1}; \qquad y_0 = \mu_1, \quad y_N = \mu_2, \tag{6}$$

де f_i, $i = \overline{1, N-1}$, μ_1, μ_2 — задані числа і y_i — шукана сіткова функція. Задача розв'язування рівняння (6) називається також різницевою крайовою задачею. Її можна записати в матричному вигляді

$$Ay = \Phi, \tag{7}$$

де $\Phi = (f_1 + a_1\mu_1, f_2, \ldots, f_{N-2}, f_{N-1} + b_{N-1}\mu_2)$ — відомий, а $y = (y_1, \ldots, y_{N-1})$ — невідомий вектори розмірності $N-1$; A — квадратна тридіагональна матриця розмірності $(N-1) \times (N-1)$ вигляду

$$A = -\begin{Vmatrix} -c_1 & b_1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ a_2 & -c_2 & b_2 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & a_3 & -c_3 & b_3 & 0 & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & 0 & \ldots & a_{N-2} & -c_{N-2} & b_{N-2} \\ 0 & 0 & 0 & 0 & 0 & \ldots & 0 & a_{N-1} & -c_{N-1} \end{Vmatrix}. \tag{8}$$

Порівнюючи (6) та (7), помічаємо, що

$$\tilde{\Lambda} y_i = -\varphi_i, \quad i = \overline{1, N-1}, \tag{9}$$

де $\tilde{\Lambda} y_1 = -c_1 y_1 + b_1 y_2$, $\varphi_1 = f_1 + a_1\mu_1$, $\tilde{\Lambda} y_i = \Lambda y_i$, $\varphi_i = f_i$, $i = \overline{2, N-2}$, $\tilde{\Lambda} y_{N-1} = a_{N-1} y_{N-2} - c_{N-1} y_{N-1}$, $\varphi_{N-1} = f_{N-1} + b_{N-1}\mu_2$. Різницевий оператор $\tilde{\Lambda}$ відображає Ω_{N-1} в Ω_{N-1}, причому $\tilde{\Lambda} y_i = \overset{\circ}{\Lambda} y_i$, тобто замість (9) можна записати

$$\overset{\circ}{\Lambda} y_i = -\varphi_i, \quad i = \overline{1, N-1}. \tag{10}$$

Введемо оператор A, який відповідає матриці (8), покладаючи

$$Ay_i = -\tilde{\Lambda} y_i = -\overset{\circ}{\Lambda} y_i, \quad i = \overline{1, N-1}.$$

Тоді замість різницевої крайової задачі (6) дістанемо операторне рівняння

$$Ay = \varphi,$$

де $A : \Omega_{N-1} \to \Omega_{N-1}$, $\varphi \in \Omega_{N-1}$. Очевидно, що A є лінійним оператором, причому оскільки $Ay = -\overset{\circ}{\Lambda} y$, то можна також вважати, що A відображає Ω_{N-1} в Ω_{N-1}.

Простір $H = \Omega_{N-1}$ можна перетворити на простір із скалярним добутком, ввівши останній за формулою

$$(y, v) = N^{-1} \sum_{i=1}^{N-1} y_i v_i = \sum_{i=1}^{N-1} h y_i v_i.$$

Зауважимо, що аналогічно записуються в операторному вигляді друга ($\varkappa_1 = \varkappa_2 = 1$) і третя ($\varkappa_1 \neq 0, 1$; $\varkappa_2 \neq 0, 1$) крайові задачі. У цьому разі матриця A має розмірність $(N+1) \times (N+1)$, а оператор A визначається формулами

$$Ay_i = -\Lambda y_i = -(b_i y_{i+1} - c_i y_i + a_i y_{i-1}), \quad i = \overline{1, N-1},$$
$$(Ay)_0 = -(\varkappa_1 y_1 - y_0), \quad (Ay)_N = -(y_N - \varkappa_2 y_{N-1})$$

і відображає Ω_{N+1} в Ω_{N+1}.

1.4.3. Різницеві формули Гріна. Умова самоспряженості різницевого рівняння другого порядку. Неважко помітити, що матриця A з попереднього пункту симетрична лише за умови

$$b_i = a_{i+1}, \quad i = \overline{1, N-1}, \tag{11}$$

тоді Λy_i можна переписати у вигляді

$$\begin{aligned} \Lambda y_i &= b_i y_{i+1} - c_i y_i + a_i y_{i-1} = a_{i+1} y_{i+1} - c_i y_i + a_i y_{i-1} = \\ &= a_{i+1}(y_{i+1} - y_i) - a_i (y_i - y_{i-1}) - (c_i - a_i - a_{i+1}) y_i = \\ &= a_{i+1} \nabla y_{i+1} - a_i \nabla y_i - (c_i - a_i - a_{i+1}) y_i = \\ &= \Delta(a_i \nabla y_i) - (c_i - a_i - a_{i+1}) y_i. \end{aligned} \tag{12}$$

Нехай сітка $\bar{\omega}_h \{x_i = ih : i = \overline{0, N}, h = 1/N\}$ покриває відрізок $[0, 1]$. Розділимо (12) на $h^2 = 1/N^2$ і скористаємося позначеннями

$$y_i = y(x_i) = y(i), \quad y_{x,i} = \Delta y_i / h,$$
$$y_{\bar{x},i} = \nabla y_i / h, \quad y_{\bar{x}x,i} = \Delta(\nabla y_i)/h^2.$$

Дістанемо різницевий оператор

$$\tilde{\Lambda} y_i \equiv \Lambda y_i / h^2 = (a y_{\bar{x}})_{x,i} - d_i y_i,$$

$$d_i = (c_i - a_i - a_{i+1})/h^2, \quad i = \overline{1, N-1}.$$

Іноді зручно те саме записати в безіндексних позначеннях

$$\tilde{\Lambda} y = (a y_{\bar{x}})_x - dy \equiv (a(x) y_{\bar{x}}(x))_x - d(x) y(x), \quad x \in \omega_h,$$

де

$$\omega_h = \{x_i = ih : i = \overline{1, N-1},\quad h = 1/N\},$$
$$y_{\bar{x}}(x) = (y(x) - y(x-h))/h,\quad y_x(x) = (y(x+h) - y(x))/h.$$

Користуючись цими позначеннями, формулу підсумовування частинами (див. 1.1.2)

$$\sum_{i=0}^{N-1} y_i \Delta v_i = -\sum_{i=1}^{N} v_i \nabla y_i + (yv)_N - (yv)_0$$

можна переписати у вигляді

$$\sum_{i=0}^{N-1} y_i v_{x,i} h = -\sum_{i=1}^{N} v_i y_{\bar{x},i} h + (yv)_N - (yv_0) \tag{13}$$

або

$$\sum_{x\in\omega_h^-} hy(x)\, v_x(x) = -\sum_{x\in\omega_h^+} hv(x)\, y_{\bar{x}}(x) + y(x_N)\, v(x_N) - y(x_0)\, v(x_0),$$

де

$$\omega_h^+ = \omega_h \cup \{x_N = 1\},\quad \omega_h^- = \omega_h \cup \{x_0 = 0\}.$$

Іноді буває зручно в лівій частині (13) підсумовування вести від $i = 1$ до $i = N-1$. Перенісши доданок при $i = 0$ з лівої частини (13) в праву, дістанемо формулу

$$\sum_{i=1}^{N-1} y_i v_{x,i} h = -\sum_{i=1}^{N} v_i y_{\bar{x},i} h + (yv)_N - y_0 v_1, \tag{14}$$

або

$$\sum_{x\in\omega_h} hyv_x = -\sum_{x\in\omega_h^+} hy_{\bar{x}} v + (yv)_N - y_0 v_1. \tag{15}$$

Підставляючи в (14) $v_i = a_i z_{\bar{x},i}$, дістанемо першу різницеву формулу Гріна

$$\sum_{i=1}^{N-1} y_i (az_{\bar{x}})_{x,i} h = -\sum_{i=1}^{N} a_i y_{\bar{x},i} z_{\bar{x},i} h + (ayz_{\bar{x}})_N - y_0 (az_{\bar{x}})_1. \tag{16}$$

Поміняємо в цій формулі місцями y та z:

$$\sum_{i=1}^{N-1} z_i (ay_{\bar{x}})_{x,i} h = -\sum_{i=1}^{N} a_i z_{\bar{x},i} y_{\bar{x},i} h + (azy_{\bar{x}})_N - z_0 (ay_{\bar{x}})_1.$$

Віднявши це співвідношення від (16), дістанемо другу різницеву формулу Гріна

$$\sum_{i=1}^{N-1} y_i (az_{\bar{x}})_{x,i} h = \sum_{i=1}^{N-1} z_i (ay_{\bar{x}})_{x,i} h + a_N (yz_{\bar{x}} - zy_{\bar{x}})_N - (y_0 (az_{\bar{x}})_1 - z_0 (ay_{\bar{x}})_1). \tag{17}$$

Якщо $y_0 = z_0 = y_N = z_N = 0$, тобто $y = \mathring{y}$, $z = \mathring{z}$, $y, z \in \mathring{\Omega}_{N+1}$, то з (17) випливає, що

$$\sum_{i=1}^{N-1} \mathring{y}_i (a\mathring{z}_{\bar{x}})_{x,i} h = \sum_{i=1}^{N-1} \mathring{z}_i (a\mathring{y}_{\bar{x}})_{x,i} h. \tag{18}$$

Віднімаючи у (18) від обох частин суму $\sum_{i=1}^{N-1} d_i \mathring{y}_i \mathring{z}_i h$, дістаємо другу різницеву формулу Гріна для різницевого оператора $\Lambda y_i = (a\mathring{y}_{\bar{x}})_{x,i} - d_i \mathring{y}_i$:

$$\sum_{i=1}^{N-1} \mathring{y}_i \Lambda \mathring{z}_i h = \sum_{i=1}^{N-1} \mathring{z}_i \Lambda \mathring{y}_i h,\quad \mathring{y}, \mathring{z} \in \mathring{\Omega}_{N+1}. \tag{19}$$

Введемо скалярні добутки

$$(y, v) = \sum_{i=1}^{N-1} hy_i v_i = \sum_{x\in\omega_h} hy(x)\, v(x),$$
$$(y, v] = \sum_{i=1}^{N} hy_i v_i = \sum_{x\in\omega_h^+} hy(x)\, v(x),$$
$$[y, v) = \sum_{i=0}^{N-1} hy_i v_i = \sum_{x\in\omega_h^-} hy(x)\, v(x),$$
$$[y, v] = \sum_{i=0}^{N} hy_i v_i = \sum_{x\in\bar{\omega}_h} hy(x)\, v(x)$$

і зв'язані з ними норми

$$\|y\| = (y, y)^{1/2},\quad \|y\| = (y, y]^{1/2},\quad \|[y\| = [y, y)^{1/2},\quad \|[y]\| = [y, y]^{1/2}.$$

Позначимо через $H = \Omega_{N-1}$ простір сіткових функцій із скалярним добутком $(,)$. У цих позначеннях формулу (15), наприклад, можна записати таким чином:

$$(y, v_x) = -(y_{\bar{x}}, v] + (yv)_N - y_0 v_1.$$

Якщо оператор A задати формулою

$$Ay = -\Lambda \mathring{y},\quad y \in H, \tag{20}$$

то другу різницеву формулу Гріна (19) можна записати у вигляді

$$(y, Az) = (Ay, z),$$

що означає самоспряженість оператора A (і, отже, $A^* = A$). Якщо $\overset{\circ}{y}, \overset{\circ}{z} \in \overset{\circ}{\Omega}_{N+1}$, то з першої різницевої формули Гріна знаходимо

$$-\sum_{i=1}^{N-1} \overset{\circ}{y}_i (a\overset{\circ}{y}_{\bar{x}})_{x,i} h = \sum_{i=1}^{N} a_i (\overset{\circ}{y}_{\bar{x},i})^2 \cdot h > 0$$

при умовах $\overset{\circ}{y}_i \not\equiv 0$, $a_i > 0$. Враховуючи означення оператора A, маємо

$$(Ay, y) = \sum_{i=1}^{N} a_i (y_{\bar{x},i})^2 h + \sum_{i=1}^{N-1} d_i y_i^2 h \equiv (ay_{\bar{x}}^2, 1] + (dy^2, 1) > 0,$$

якщо $a_i > 0$, $d_i \geqslant 0$. Таким чином, різницевий оператор A, визначений формулою (20), є самоспряженим і додатним за умов $a_i > 0$, $d_i \geqslant 0$, $i = \overline{1, N-1}$, $a_N > 0$.

Відомо, що умова $b_i = a_{i+1}$ є достатньою для самоспряженості оператора $\Lambda y_i = b_i y_{i+1} - c_i y_i + a_i y_{i-1}$. Покажемо, що вона є і необхідною, тобто випливає із самоспряженості Λ.

Дійсно, подамо Λ у вигляді суми $\Lambda y_i = \Lambda_1 y_i + \Lambda_2 y_i$, де

$$\Lambda_1 y_i = a_{i+1}(y_{i+1} - y_i) - a_i (y_i - y_{i-1}) - (c_i - a_i - b_i) y_i =$$
$$= h^2 [(ay_{\bar{x}})_{x,i} - d_i y_i],$$
$$\Lambda_2 y_i = (b_i - a_{i+1}) y_{i+1}.$$

Оскільки Λ_1 в силу викладеного вище є самоспряженим у просторі $H = \overset{\circ}{\Omega}_{N+1}$, або $H = \Omega_{N-1}$ із скалярним добутком $(y, v) = \sum_{i=1}^{N-1} h y_i v_i$, то

$$(h^{-2}\Lambda \overset{\circ}{y}, \overset{\circ}{v}) - (\overset{\circ}{y}, h^{-2}\Lambda \overset{\circ}{v}) = (h^{-2}\Lambda_1 \overset{\circ}{y}, \overset{\circ}{v}) -$$
$$- (\overset{\circ}{y}, h^{-2}\Lambda_1 \overset{\circ}{v}) + (h^{-2}\Lambda_2 \overset{\circ}{y}, \overset{\circ}{v}) - (\overset{\circ}{y}, h^{-2}\Lambda_2 \overset{\circ}{v}) =$$
$$= (h^{-2}\Lambda_2 \overset{\circ}{y}, \overset{\circ}{v}) - (\overset{\circ}{y}, h^{-2}\Lambda_2 \overset{\circ}{v}) =$$
$$= \sum_{i=1}^{N-1} h^{-2} (b_i - a_{i+1})(y_{i+1} v_i - y_i v_{i+1}) h.$$

Звідси випливає $\Lambda = \Lambda^*$ лише за умови, що

$$\sum_{i=1}^{N-1} h^{-2} (b_i - a_{i+1})(y_{i+1} v_i - y_i v_{i+1}) h = 0$$

для довільних $y, v \in H$. Виберемо $y_i = \delta_{i,i_0+1}$, $v_i = \delta_{i,i_0}$, де i_0 — довільний фіксований вузол сітки, δ_{i,i_0} — символ Кронекера. Тоді $y_{i+1} v_i - y_i v_{i+1} = \delta_{i,i_0}$ і остання умова набере вигляду

$$b_{i_0} - a_{i_0+1} = 0,$$

що і треба було довести.

Зауважимо, що рівняння

$$\Lambda y_i = b_i y_{i+1} - c_i y_i + a_i y_{i-1} = -f_i \qquad (21)$$

завжди можна звести до вигляду

$$\tilde{\Lambda} y_i \equiv \Delta (A_i \nabla y_i) - D_i y_i = -F_i, \qquad (22)$$

де $\tilde{\Lambda}$ — самоспряжений оператор. Дійсно, помножимо обидві частини (21) на $\mu_i \not\equiv 0$:

$$\tilde{\Lambda} y_i = \mu_i a_i y_{i-1} - \mu_i c_i y_i + \mu_i b_i y_{i+1} = -\mu_i f_i.$$

Умова самоспряженості буде виконана, якщо

$$b_i \mu_i = (\mu a)_{i+1} = \mu_{i+1} a_{i+1} = A_{i+1}.$$

Звідси

$$\mu_{i+1} = \frac{b_i}{a_{i+1}} \mu_i = \mu_1 \prod_{k=1}^{i} \frac{b_k}{a_{k+1}}$$

і дістанемо (22), де $A_i = a_i \mu_i$, $D_i = \mu_i (c_i - a_i - b_i)$, $F_i = \mu_i f_i$. Іншими словами, якщо матриця (8) не є симетричною, то її можна звести до симетричного вигляду множенням зліва на діагональну матрицю

$$M = \begin{pmatrix} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_{N-1} \end{pmatrix}.$$

1.4.4. Власні значення різницевого оператора другого порядку. Розглянемо різницеву задачу на власні значення

$$(ay_{\bar{x}})_{x,i} - d_i y_i + \lambda y_i = 0, \quad i = \overline{1, N-1}, \quad y_0 = y_N = 0$$

або

$$Ay = \lambda y, \quad y \in \Omega_{N-1},$$

де $Ay = -(a\overset{\circ}{y}_{\bar{x}})_{x,i} + d_i \overset{\circ}{y}_i$, $y \in \overset{\circ}{\Omega}_{N+1}$. Оператор A є самоспряженим і додатним, тому він має $N-1$ дійсних власних значень і відповідних їм ортонормованих власних векторів (функцій).

У найпростішому випадку $a_i = 1$, $d_i = 0$ їх можна знайти в явному вигляді. Тоді рівняння

$$y_{\bar{x}x,i} + \lambda y_i = 0, \quad i = \overline{1, N-1}, \quad y_0 = y_N = 0,$$

можна переписати таким чином:

$$y_{i+1} - 2\cos\alpha y_i + y_{i-1} = 0,$$

де $2\cos\alpha = 2 - \lambda h^2$, $h = 1/N$. Загальний розв'язок має вигляд

$$y_i = c_1 \cos i\alpha + c_2 \sin ih.$$

З крайових умов знаходимо $y_0 = c_1 = 0$, $y_N = c_2 \sin N\alpha = 0$. Оскільки потрібно знайти нетривіальний розв'язок, то $c_2 \neq 0$, і тому має бути $\sin N\alpha = 0$, тобто $N\alpha = m\pi$, $m = 0, 1, 2, \ldots$. Звідси $\alpha = \alpha_m = m\pi/N = m\pi h$. З рівності $2\cos\alpha = 2 - \lambda h^2$ знаходимо $\lambda h^2 = 2(1 - \cos\alpha) = 4\sin^2\frac{\alpha}{2}$, $\lambda = \lambda_m = \frac{4}{h^2}\sin^2\frac{m\pi h}{2}$. Цьому значенню λ відповідає власна функція $y_m(i) = c\sin m\pi x_i$, $c \neq 0$, $x_i = ih$, $i = \overline{0, N}$, яка визначається з точністю до сталого множника. Неважко помітити, що

$$y_N(j) = c\sin N\pi x_j = c\sin \pi j = 0, \quad j = \overline{0, N},$$
$$y_{N+1}(j) = c\sin(N+1)\pi x_j = c\sin(N\pi x_j + \pi x_j) =$$
$$= c\sin\pi x_j \cdot \cos\pi j = (-1)^j y_1(j), \ \ldots,$$
$$y_{N+m+1}(j) = (-1)^j y_m(j), \quad m = \overline{1, N-1}.$$

Отже, лінійно незалежні лише функції $y_m(i)$, $m = \overline{1, N-1}$. Виберемо множник c так, щоб норма функцій $y_m(i)$ дорівнювала одиниці. Позначаючи $\alpha = 2\pi mh$ і користуючись формулою $\cos 2\pi mx_k = \cos k\alpha = \operatorname{Re}(\cos k\alpha + i\sin k\alpha) = \operatorname{Re}(e^{ik\alpha})$, де $i = \sqrt{-1}$ — уявна одиниця, матимемо

$$\|y_m(j)\|^2 = c^2 \sum_{k=1}^{N-1} h\sin^2 m\pi x_k = \frac{c^2}{2}\sum_{k=1}^{N-1} h(1 - \cos 2m\pi x_k) =$$
$$= \frac{c^2(N-1)h}{2} - \frac{c^2}{2}\operatorname{Re}\sum_{k=1}^{N-1} he^{ik\alpha} = \frac{c^2(N-1)h}{2} -$$
$$- \frac{c^2h}{2}\operatorname{Re}\frac{e^{i\alpha} - e^{iN\alpha}}{1 - e^{i\alpha}} = \frac{c^2(N-1)h}{2} + \frac{c^2h}{2} = \frac{c^2Nh}{2} = \frac{c^2}{2}.$$

Звідси знаходимо $c = \sqrt{2}$ і, таким чином, ортонормована (у просторі із скалярним добутком $(y, v) = \sum_{i=1}^{N-1} hy_iv_i$) система власних функцій має вигляд

$$y_m(i) = \sqrt{2}\sin m\pi x_i.$$

Власні значення λ_s зростають із збільшенням s, бо $\sin\frac{\pi h}{2}s < \sin\frac{\pi h}{2}(s+1) < 1$ при $s \leqslant N$. Найменше і найбільше власні значення мають вигляд

$$\lambda_1 = \frac{4}{h^2}\sin^2\frac{\pi h}{2}, \quad \lambda_{N-1} = \frac{4}{h^2}\sin^2\frac{(N-1)\pi h}{2} =$$
$$= \frac{4}{h^2}\sin^2\left(\frac{\pi}{2} - \frac{\pi h}{2}\right) = \frac{4}{h^2}\cos^2\frac{\pi h}{2}.$$

Записавши λ_1 у вигляді $\lambda_1 = \pi^2(\sin\xi/\xi)^2$, де $\xi = \pi h/2 \leqslant \pi/4$ (найбільше, яке має зміст, значення $h = 1/2$), і враховуючи, що функція $f(\xi) = (\sin\xi)/\xi$ спадає і має мінімум при $\xi = \pi/4$, дістанемо

$$\lambda_1 \geqslant 8.$$

Для λ_{N-1} маємо оцінку $\lambda_{N-1} < 4/h^2$, і тому

$$8 < \lambda_k \leqslant 4/h^2, \quad k = \overline{1, N-1}.$$

Вправа 2. Довести, що розв'язком задачі $y_{\bar{x}x} + \lambda y = 0$, $x \in \omega_h$, $y(a) = y(b) = 0$, де $\omega_h = \{x_i = a + ih: i = \overline{1, N-1}\}$, є власні значення $\lambda_k = \frac{4}{h^2}\sin^2\frac{k\pi h}{2(b-a)}$, $k = \overline{1, N-1}$, і власні функції $y_k(x) = \sqrt{\frac{2}{b-a}}\sin\frac{k\pi x}{b-a}$, які ортонормовані у просторі із скалярним добутком

$$(y, v) = \sum_{x\in\omega_h} hy(x)v(x).$$

1.4.5. Одно- та багатокрокові операторні схеми. Нехай H — N-вимірний лінійний простір із скалярним добутком $(,)$; A, B — лінійні оператори в ньому. Розглянемо деяку обчислювальну схему (алгоритм) для знаходження послідовності $y_0, y_1, \ldots \in H$, яка задається рекурентною формулою

$$y_{k+1} = F(y_k, y_{k-1}, \ldots, y_{k-m+1}), \quad k = m-1, m, \ldots, \qquad (23)$$

де $y_0, y_1, \ldots, y_{m-1}$ — задані елементи H; F — деякий лінійний оператор. Таку схему називатимемо m-кроковою схемою або $(m+1)$-ярусною схемою. Найчастіше на практиці зустрічаються лінійні однокрокові схеми

$$y_{k+1} = Ay_k, \quad k = 0, 1, \ldots \ (y_0 \text{ задано}). \qquad (24)$$

Як ми переконаємось далі, важливу роль відіграє запис лінійної двох'ярусної схеми в канонічному вигляді:

$$B_k\frac{y_{k+1} - y_k}{\tau_{k+1}} + A_ky_k = \varphi_k, \quad k = 0, 1, \ldots, \qquad (25)$$

де лінійні оператори $B_k : H \to H$ мають обернені B_k^{-1}; $\tau_{k+1} > 0$ — деякі параметри. Такий запис неоднозначний, і, наприклад, можна покласти

$$A_k = -B_k\frac{A - I}{\tau_{k+1}}, \quad \varphi_k = 0,$$

де I — одиничний оператор. Далі розглядатимемо такі двох'ярусні схеми:

$$B\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = \varphi_k, \quad k = 0, 1, \ldots. \qquad (26)$$

Якщо $\tau_{k+1} = \tau$, тобто не залежить від k, то схему (26) називатимемо *стаціонарною*.

Якщо $B = I$, то схема (26) називається *явною;* тоді y_{k+1} визначається через y_k за явною формулою

$$y_{k+1} = y_k - \tau_{k+1}(Ay_k - \varphi_k).$$

Якщо $B \neq I$, то схема (26) називається *неявною,* бо для визначення y_{k+1} треба розв'язати операторне рівняння

$$By_{k+1} = \psi_k, \quad \psi_k = By_k - \tau_{k+1}(Ay_k - \varphi_k).$$

Неважко помітити, що послідовність $\{y_k\}$, яка задається формулою (26), можна побудувати як суму двох послідовностей, що задаються формулами

$$B\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = 0, \quad k = 0, 1, \ldots \quad (y_0 = u_0 \text{ — задано}) \tag{27}$$

та

$$B\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = \varphi_k, \quad k = 0, 1, \ldots, y_0 = 0. \tag{28}$$

Звідси випливає доцільність таких двох означень.

Схема (26) називається *стійкою за початковими даними,* якщо для задачі (27) виконується оцінка

$$\|y_k\|_{(1)} \leqslant M, \|y_0\|_{(1)}, \tag{29}$$

де $\|\cdot\|_{(1)}$ — деяка норма в просторі H; M_1 — абсолютна стала (не залежить від τ_k, k).

Схема (26) називається *стійкою за правою частиною,* якщо для розв'язання задачі (28) використовується оцінка

$$\|y_k\|_{(1)} \leqslant M_2 \max_{0 \leqslant k \leqslant n} \|\varphi_k\|_{(2)}, \tag{30}$$

де M_2 — абсолютна стала; $\|\cdot\|_{(1)}$, $\|\cdot\|_{(2)}$ — деякі норми в H.

Іноді користуються також поняттям ρ-*стійкості,* коли означенням передбачається виконання для (27) нерівностей

$$\|y_k\|_{(1)} \leqslant \rho\|y_{k-1}\|_{(1)} \leqslant \cdots \leqslant \rho^k\|y_0\|_{(1)},$$

де $\rho > 0$ — деяке число.

Нехай $D = D^* > 0$ — самоспряжений додатний оператор в H. Тоді в H можна ввести норму

$$\|y\|_{(1)} \equiv \|y\|_D = \sqrt{(Dy, y)}.$$

Простір H з такою нормою позначатимемо H_D. Користуватимемося також таким означенням стійкості за початковими умовами: схема (26) називається *стійкою в* H_D, якщо

$$\|y_{k+1}\|_D \leqslant \|y_k\|_D. \tag{31}$$

Записавши (26) у вигляді

$$y_{k+1} = s_k y_k, \quad s_k = I - \tau_{k+1}B^{-1}A, \tag{32}$$

де s_k — оператор переходу на наступний ярус, помічаємо, що умова (31) еквівалентна умові

$$\|s_k\|_D \leqslant 1 \; \forall k. \tag{33}$$

У свою чергу, (33) еквівалентна умові

$$J_D = \|y\|_D^2 - \|s_k y\|_D^2 = (Dy, y) - (Ds_k y, s_k y) \geqslant 0 \quad \forall y \in H, \quad \forall k. \tag{34}$$

Отже, три умови стійкості в H_D, а саме (31), (33), (34), еквівалентні.

Теорема 1. *Якщо $A = A^*$ — самоспряжений додатний оператор і існує B^{-1}, то для стійкості стаціонарної схеми* (26) *за початковими умовами (тобто при $\tau_k = \tau$, $s_k = S$) в H_A необхідно і достатньо, щоб*

$$(By, y) - \frac{\tau}{2}(Ay, y) \geqslant 0, \quad \forall y \in H, \tag{35}$$

що записуватимемо так: $B \geqslant \frac{\tau}{2}A$.

Доведення. Досить переконатися в еквівалентності виразів (34), (35), або, точніше, в еквівалентності (35) нерівності

$$J_A = (Ay, y) - (ASy, Sy) = (Ay, y) - (Ay - \tau AB^{-1}Ay, y - \tau B^{-1}Ay) = 2\tau(AB^{-1}Ay, y) - \tau^2(AB^{-1}Ay, B^{-1}Ay) \geqslant 0 \forall y \in H.$$

Позначивши $B^{-1}Ay = x$, $Ay = Bx$, дістанемо

$$J_A = 2\tau\left((Bx, x) - \frac{\tau}{2}(Ax, x)\right) \geqslant 0 \quad \forall x \in H, \tag{36}$$

звідки випливає, що нерівність (35) обумовлює нерівність (34) при $D = A$, що і є доведенням достатності умови (35).

Доведемо необхідність, тобто, що із стійкості стаціонарної схеми (26) випливає (35). Дійсно, якщо схема стійка в H_A, то виконується нерівність (34) (при $\tau_k = \tau$, $s_k = S$, $D = A$), а (36) вказує, що $J_A \geqslant 0$, звідки випливає (35). Теорему доведено.

Приклад 7. Розглянемо рекурентну послідовність чисел

$$b\frac{y_{k+1} - y_k}{\tau} + ay_k = 0, \quad k = 0, 1, \ldots,$$

де y_0, b, τ, a — задані числа, $a > 0$, $b > 0$, $\tau > 0$. Із формули $y_{k+1} = (1 - \tau a/b)\, y_k$ випливає, що умова стійкості

$$|y_{k+1}| \leqslant |y_k| \leqslant \cdots \leqslant |y_0|$$

виконується при $|1 - \tau a/b| \leqslant 1$. Розв'язуючи цю нерівність, маємо

$$-1 \leqslant 1 - \tau a/b \leqslant 1, \quad b \geqslant \frac{\tau a}{2}.$$

що цілком відповідає твердженню теореми 1.

Теорема 2. *Якщо* $A = A > 0$, $B = B^* > 0$, *то для стійкості стаціонарної схеми* (26) *в* H_B *(тобто* $\|y_{k+1}\|_B \leqslant \|y_k\|_B$*) необхідно і достатньо виконання умови* (35), *тобто* $B \geqslant \frac{\tau}{2} A$.

Доведення. Запишемо схему (26) у вигляді (32) і покажемо, що умова (33) еквівалентна умові (35).

Нехай y — довільний вектор з H; $\{\xi_k\}$ — власні вектори задачі

$$A\xi_k = \lambda_k B\xi_k, \quad \lambda_k > 0, \tag{37}$$

причому

$$(B\xi_k, \xi_m) = \delta_{km} = \begin{cases} 1, & k = m, \\ 0, & k \neq m, \end{cases} \quad k, m = \overline{1, N}.$$

Враховуючи, що $S\xi_k = \xi_k - \tau B^{-1} A\xi_k = (1 - \tau\lambda_k)\,\xi_k$, $BS\xi_k = (1 - \tau\lambda_k)\,B\xi_k$ і розкладаючи вектор y за системою $\{\xi_k\}$, тобто $y = \sum\limits_{k=1}^{N} \alpha_k \xi_k$ (N — розмірність H), маємо

$$(By, y) = \sum_{k=1}^{N} \alpha_k^2, \quad (Ay, y) = \sum_{k=1}^{N} \lambda_k \alpha_k^2,$$

$$(BSy, Sy) = \sum_{k=1}^{N} \alpha_k^2 (1 - \tau\lambda_k)^2 \leqslant \|S\|_B^2 \sum_{k=1}^{N} \alpha_k^2 = \|S\|_B (By, y),$$

де

$$\|S\|_B^2 = \max_{1 \leqslant k \leqslant N} (1 - \tau\lambda_k)^2.$$

Звідси випливає, що нерівність $\|S_B\| \leqslant 1$ еквівалентна умові

$$\tau\lambda_k \leqslant 2, \quad k = \overline{1, N},$$

яка в силу рівності

$$(By, y) - \frac{\tau}{2}(Ay, y) = \sum_{k=1}^{N} \alpha_k^2 \left(1 - \frac{\tau\lambda_k}{2}\right)$$

еквівалентна умові (35). Теорему доведено.

Теорема 3. *Якщо* $A = A^* > 0$, $B = B^* > 0$, *то необхідною і достатньою умовою* ρ*-стійкості стаціонарної схеми* (26) *за початковими умовами з будь-яким* $\rho > 0$, *тобто* $\|y_{k+1}\|_D \leqslant \rho \|y_k\|_D$, $D = A$, B *в виконання операторних нерівностей*

$$\frac{1 - \rho}{\tau} B \leqslant A \leqslant \frac{1 + \rho}{\tau} B. \tag{38}$$

Доведення. Нерівності (38) еквівалентні умовам

$$\frac{1 - \rho}{\tau} \leqslant \lambda_k \leqslant \frac{1 + \rho}{\tau}, \quad k = \overline{1, N}, \tag{39}$$

де λ_k — власні числа задачі (37).

Доведемо спочатку достатність умов (38), (39). Нехай $D = B$ і маємо нерівності (38), (39). З умови (39) випливає, що $-\rho \leqslant \tau\lambda_k - 1 \leqslant \rho$, $|1 - \tau\lambda_k| \leqslant \rho$ і в силу (37) $\|S\|_B \leqslant \rho$, тобто

$$\|y_{k+1}\|_B \leqslant \|S\|_B \|y_k\|_B \leqslant \rho \|y_k\|_B,$$

що і доводить достатність.

Нехай тепер має місце ρ-стійкість, тобто $\|y_{k+1}\|_B \leqslant \rho \|y_k\|_B$. Тоді, оскільки $\|S\|_B$ — найменша стала, для якої виконується нерівність $\|y_{k+1}\|_B^2 = (By_{k+1}, y_{k+1}) = (BSy_k, Sy_k) \leqslant M\,(By_k, y_k) \equiv M \|y_k\|_B^2$, то $\|S\|_B \leqslant \rho$. Тому в силу рівності (37) маємо $|1 - \tau\lambda_k| \leqslant \rho$, тобто виконуються (38), (39). Необхідність доведено.

Аналогічно доведемо теорему при $D = A$, коли врахуємо, що

$$(ASy, Sy) = \sum_{k=1}^{N} \alpha_k^2 \lambda_k (1 - \tau\lambda_k)^2 \leqslant \max_{1 \leqslant k \leqslant N} (1 - \tau\lambda_R)^2 (Ay, y).$$

Теорему доведено.

На практиці важливу роль відіграє ρ-стійкість при $\rho < 1$. Умови такої стійкості дає наступна теорема.

Теорема 4. *Якщо* $A = A^* > 0$, $B = B^* > 0$, $\gamma_1 B \leqslant A \leqslant \gamma_2 B_2$, $\gamma_1 > 0$, $\tau \leqslant \tau_0$, $\tau_0 = \frac{2}{\gamma_1 + \gamma_2}$, *то має місце оцінка*

$$\|y_{k+1}\|_D \leqslant \rho \|y_k\|_D, \tag{40}$$

де y_{k+1} *знаходять за стаціонарною схемою* (27) *при* $\tau_{k+1} = \tau$, $D = A$, B, $\rho = 1 - \tau\gamma_1 < 1$.

Доведення. Для доведення досить оцінити норми $\|S\|_B = \|S\|_A = \max\limits_{1 \leqslant k \leqslant N} |1 - \tau\lambda_k|$ за умов $\gamma_1 \leqslant \lambda_k \leqslant \gamma_2$, $0 < \gamma_1 = \lambda_1 \leqslant \ldots \leqslant \lambda_N = \gamma_2$. Для цього розглянемо величину

$$\varphi_k = (1 - \tau\lambda_1)^2 - (1 - \tau\lambda_k)^2 = 2\tau\,(\lambda_k - \lambda_1)\left(1 - \frac{\tau}{2}(\lambda_k + \lambda_1)\right).$$

Неважко помітити, що $\varphi_k \geqslant 0$, коли

$$1 - \frac{\tau}{2}(\lambda_k + \lambda_1) \geqslant 1 - \frac{\tau}{2}(\gamma_2 + \gamma_1) \geqslant 1 - \frac{\tau_0}{2}(\gamma_1 + \gamma_2) = 0,$$

тобто $\max\limits_{1 \leqslant k \leqslant N} |1 - \tau\lambda_k| = 1 - \tau\gamma_1$ при $\tau \leqslant \tau_0$. Це і доводить теорему.

Розглянемо тепер стійкість за правою частиною стаціонарної схеми (28), тобто з'ясуємо, за яких умов для послідовності $\{y_k\}$, що визначається співвідношенням

$$B\frac{y_{k+1} - y_k}{\tau} + Ay_k = \varphi_k, \quad k = 0, 1, \ldots; \quad y_0 = 0, \tag{41}$$

має місце нерівність (30).

Теорема 5. *Нехай* $A = A^* > 0$, $B = B^* > 0$ *і виконується умова*

$$B \geqslant \frac{\tau}{2} A. \tag{42}$$

Тоді для послідовності $\{y_k\}$, *яку знаходять за схемою* (41), *має місце стійкість за правою частиною, а саме*

$$\|y_k\|_B \leqslant \sum_{j=0}^{k-1} \tau \|\varphi_j\|_{B^{-1}}. \tag{43}$$

Доведення. Запишемо (41) у вигляді

$$y_{k+1} = Sy_k + \tau B^{-1}\varphi_k, \quad k = 0, 1, \ldots; \quad S = I - \tau B^{-1}A, \quad y_0 = 0.$$

З нерівності трикутника маємо

$$\|y_{k+1}\|_D \leqslant \|Sy_k\|_D + \tau \|B^{-1}\varphi_k\|_D \leqslant \|S\|_D \|y_k\|_D + \tau \|B^{-1}\varphi_k\|_D.$$

З теореми 2 випливає, що за умови (42)

$$\|S\|_B \leqslant 1, \quad \|B^{-1}\varphi_k\|_B^2 = (B(B^{-1}\varphi_k), B^{-1}\varphi_k) =$$
$$= (B^{-1}\varphi_k, \varphi_k) = \|\varphi_k\|_{B^{-1}}^2,$$

і з попередньої нерівності знаходимо

$$\|y_{k+1}\|_B \leqslant \|y_k\|_B + \tau \|\varphi_k\|_{B^{-1}}.$$

Враховуючи, що $y_0 = 0$ і застосовуючи рекурентно цю нерівність, дістаємо (43). Теорему доведено.

Доведемо ще одну оцінку, яка виражає стійкість стаціонарної схеми (26) за правою частиною та початковими умовами. Крім самої оцінки, важливим є метод її доведення, який є досить загальним і називається *методом енергетичних оцінок.*

Теорема 6. *Нехай* $A = A^* > 0$ *і послідовність* $\{y_k\}$ *визначається за стаціонарною двох'ярусною схемою*

$$B\frac{y_{k+1} - y_k}{\tau} + Ay_k = \varphi_k, \quad k = 0, 1, \ldots; \quad y_0 = u_0. \tag{44}$$

Тоді за умови $B \geqslant \frac{1}{2}A$ *при* $\varphi_k = 0$ *має місце нерівність*

$$\|y_{k+1}\|_A \leqslant \|y_k\|_A, \tag{45}$$

тобто схема стійка за початковими умовами в H_A.

Якщо виконується більш сильна нерівність

$$B \geqslant \varepsilon I + \frac{\tau}{2}A, \tag{46}$$

де $\varepsilon > 0$ — *довільне число, то має міцне нерівність*

$$\|y_k\|_A^2 \leqslant \|y_0\|_A^2 + \frac{1}{2\varepsilon}\sum_{j=0}^{k-1} \tau \|\varphi_j\|^2, \tag{47}$$

яка виражає стійкість схеми (44) *за початковими умовами і за правою частиною.*

Доведення. Підставляючи в (44) вираз

$$y_k = \frac{1}{2}(y_k + y_{k+1}) - \frac{\tau}{2}\frac{y_{k+1} - y_k}{\tau},$$

маємо

$$\left(B - \frac{\tau}{2}A\right)\frac{y_{k+1} - y_k}{\tau} + \frac{1}{2}A(y_{k+1} + y_k) = \varphi_k.$$

Помножимо цю рівність скалярно на $2(y_{k+1} - y_k)$ і врахуємо, що

$$(A(y_{k+1} + y_k), y_{k+1} - y_k) = (Ay_{k+1}, y_{k+1}) + (Ay_k, y_{k+1}) -$$
$$- (Ay_{k+1}, y_k) - (Ay_k, y_k) = (Ay_{k+1}, y_{k+1}) - (Ay_k\, y_k),$$

бо в силу самоспряженості A

$$(Ay_n, y_{n+1}) = (Ay_{n+1}, y_n).$$

Як наслідок дістанемо так звану «енергетичну тотожність»:

$$2\tau\left(\left(B - \frac{\tau}{2}A\right)\frac{y_{k+1} - y_k}{\tau}, \frac{y_{k+1} - y_k}{\tau}\right) + (Ay_{k+1}, y_{k+1}) =$$
$$= (Ay_k, y_k) + 2(\varphi_k, y_{k+1} - y_k). \tag{48}$$

Звідси при $B \geqslant \frac{\tau}{2}A$ маємо нерівність (45), тобто першу частину теореми доведено. Перетворимо другий доданок в правій частині (48), тобто

$$2(\varphi_k, y_{k+1} - y_k) = 2\tau\left(\varphi_k, \frac{y_{k+1} - y_k}{\tau}\right).$$

За допомогою ε-нерівності

$$|ab| = (\sqrt{2\varepsilon}\, a)\left(\sqrt{\frac{1}{2\varepsilon}}\, b\right) \leqslant \varepsilon a^2 + \frac{1}{4\varepsilon}b^2,$$

де a, b, ε — будь-які додатні числа, знаходимо

$$2(\varphi_k, y_{k+1} - y_k) \leqslant 2\tau\|\varphi_k\|\left\|\frac{y_{k+1} - y_k}{\tau}\right\| \leqslant$$
$$\leqslant 2\tau\varepsilon\left\|\frac{y_{k+1} - y_k}{\tau}\right\|^2 + \frac{\tau}{2\varepsilon}\|\varphi_k\|^2.$$

Підставляючи цю оцінку в «енергетичну тотожність» (48), матимемо

$$2\tau\left(\left(B - \varepsilon I - \frac{\tau}{2}A\right)\frac{y_{k+1} - y_k}{\tau}, \frac{y_{k+1} - y_k}{\tau}\right) + \|y_{k+1}\|_A^2 \leqslant$$
$$\leqslant \|y_k\|_A^2 + \frac{\tau}{2\varepsilon}\|\varphi_k\|^2,$$

звідки за умови (46)

$$\|y_{k+1}\|_A^2 \leqslant \|y_k\|_A^2 + \frac{\tau}{2\varepsilon}\|\varphi_k\|^2,$$

і далі рекурентно дістаємо (45). Теорему доведено.

Для дослідження стійкості багатоярусних операторних схем зручно користуватися прямою сумою просторів.

Простір $H^m = \underbrace{H \oplus H \oplus \ldots \oplus H}_{m}$ (пряма сума m просторів H) визначається як множина векторів вигляду

$$x = (x^1, x^2, \ldots, x^m), \quad x^\alpha \in H, \quad \alpha = \overline{1, m},$$

в якій додавання і множення на число визначаються покоординатно. Нульовим елементом простору H^m є вектор (0, 0, ..., 0), кожна компонента якого — це нуль простору H. Якщо H є простором із скалярним добутком (,), то в H^m можна ввести скалярний добуток за формулою

$$(y, v) = \sum_{\alpha=1}^{m} (y^\alpha, v^\alpha),$$

де $y = (y_1, \ldots, y_m)$, $v = (v_1, \ldots, v_m)$, $y, v \in H^m$, $y^\alpha, v^\alpha \in H$, $\alpha = \overline{1, m}$. Нехай задані оператори $C_{\alpha\beta} : H \to H$, $\alpha, \beta = \overline{1, m}$. Тоді оператор

$$C = \begin{pmatrix} C_{11} & \ldots & C_{1m} \\ \ldots & & \ldots \\ C_{m1} & \ldots & C_{mm} \end{pmatrix}$$

діє в просторі H^m за правилом

$$Cx = \{(Cx)^\beta\}_{\beta=\overline{1,m}} = \left\{\sum_{\gamma=1}^{m} C_{\beta\gamma}x^\gamma\right\}_{\beta=\overline{1,m}},$$

де $x = (x^1, \ldots, x^m) \in H^m$, $x^\alpha \in H$, $\alpha = \overline{1, m}$. Неважко помітити, що для таких операторних матриць справедливі звичні правила додавання матриць і множення матриці на число. Нульовим є оператор

$$A = (A_{\alpha\beta})_{\alpha,\beta=\overline{1,m}}, \quad A_{\alpha\beta} = 0, \quad \alpha, \beta = \overline{1, m},$$

а одиничним — оператор

$$A = (A_{\alpha\beta})_{\alpha\beta=\overline{1,m}}, \quad A_{\alpha\beta} = \delta_{\alpha\beta}I,$$

де $\delta_{\alpha\beta}$ — символ Кронекера; I — одиничний оператор простору H. Зберігається і правило множення матриць: якщо $A = (A_{\alpha\beta})_{\alpha,\beta=\overline{1,m}}$, $B = (B_{\alpha\beta})_{\alpha,\beta=\overline{1,m}}$, то компоненти оператора $C = AB \equiv (C_{\alpha\beta})_{\alpha\beta=\overline{1,m}}$ визначаються формулою

$$C_{\alpha\beta} = \sum_{\gamma=1}^{m} A_{\alpha\gamma}B_{\gamma\beta}.$$

З'ясуємо умови самоспряженості операторів, які діють у H^m. Нехай $\mathscr{P} = (\mathscr{P}_{\alpha\beta})_{\alpha\beta=\overline{1,m}} : H^m \to H^m$, $y = (y^1, \ldots, y^m)$, $v = (v^1, \ldots, v^m)$ — елементи H^m. Тоді

$$(\mathscr{P}y, v) = \sum_{\alpha,\beta=1}^{m} (\mathscr{P}_{\alpha\beta}y^\beta, v^\alpha),$$

$$(\mathscr{P}y, y) = \sum_{\alpha,\beta=1}^{m} (\mathscr{P}_{\alpha\beta}y^\beta, y^\alpha).$$

Лема 1. *Нехай* $\mathscr{P} = (\mathscr{P}_{\alpha\beta}) : H^m \to H^m$. *Тоді спряжений оператор* $\mathscr{P}^* = (\mathscr{P}^*)_{\alpha,\beta}$, *для якого* $(\mathscr{P}y, v) = (y, \mathscr{P}^*v)$, *має компоненти*

$$(\mathscr{P}^*)_{\alpha\beta} = \mathscr{P}^*_{\beta\alpha}, \quad \alpha, \beta = \overline{1, m},$$

зокрема, для самоспряженості оператора $\mathscr{P}$ *необхідно і достатньо, щоб*

$$\mathscr{P}_{\beta\alpha} = \mathscr{P}^*_{\alpha\beta}, \quad \alpha, \beta = \overline{1, m}, \tag{49}$$

де $\mathscr{P}^*_{\alpha\beta}$ — *оператор, спряжений до* $\mathscr{P}_{\alpha\beta}$ *в* H.

Доведення. Маємо

$$(\mathscr{P}y, v) = \sum_{\alpha,\beta=1}^{m} (\mathscr{P}_{\alpha\beta}y^\beta, v^\alpha) = \sum_{\alpha,\beta=1}^{m} (y^\beta, \mathscr{P}^*_{\alpha\beta}v^\alpha) =$$

$$= \sum_{\alpha,\beta=1}^{m} (y^\alpha, \mathscr{P}^*_{\beta\alpha}v^\beta) = (y, \mathscr{P}^*, v),$$

що і доводить лему 1.

Наслідок. *Для самоспряженого оператора* $\mathscr{P} : H^m \to H^m$ *справедлива тотожність*

$$(\mathscr{P}y, y) = \sum_{\alpha=1}^{m} (\mathscr{P}_{\alpha\alpha}y^\alpha, y^\alpha) + 2\sum_{1\leqslant\alpha<\beta\leqslant m} (\mathscr{P}_{\alpha\beta}y^\beta, y^\alpha). \tag{50}$$

У випадку $m = 2$ *співвідношення* (49), (50) *мають вигляд*

$$\mathscr{P}_{11} = \mathscr{P}^*_{11}, \quad \mathscr{P}_{21} = \mathscr{P}^*_{12}, \quad \mathscr{P}_{22} = \mathscr{P}^*_{22},$$
$$(\mathscr{P}x, x) = (\mathscr{P}_{11}x^1, x^1) + 2(\mathscr{P}_{12}x^2, x^1) + (\mathscr{P}_{22}x^2, x^2). \tag{51}$$

Розглянемо $(m+1)$-ярусну операторну схему

$$B_m y^{n+m} + B_{m-1}y^{n+m-1} + \ldots + B_1 y^{n+1} + B_0 y^n = \varphi^n, \tag{52}$$

де $y^{n+\alpha} \in H$, $B_\alpha : H \to H$, $\alpha = \overline{0, m}$. Можна розглянути сітку (іноді її називають часовою)

$$\omega_\tau = \{t_n = n\tau, \ n = 0, 1, 2, \ldots\}$$

і вважати y^n функцією дискретного аргументу $t_n = n\tau$ із значеннями в просторі H, тобто $y^n = y(t_n) \in H$. Будемо припускати, що елементи $y^0, y^1, \ldots, y^{m-1}$ задані й існує оператор B_m^{-1}. Розв'язком схеми (52) в момент $t_n = n\tau$ називатимемо вектор $y_n = \{y^n, y^{n+1}, \ldots, y^{n+m-1}\}$,

компоненти якого задовольняють (52). Неважко помітити, що (52) можна записати у вигляді

$$y_{n+1} = Sy_n + \varphi_n, \tag{53}$$

де $y_n = (y^n, y^{n+1}, \ldots, y^{n+m-1})$, $\varphi_n = (0, 0, \ldots, 0, B_m^{-1}\varphi^n)$,

$$S = \begin{Vmatrix} 0 & I & 0 & \ldots & 0 \\ 0 & 0 & I & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & I \\ -B_m^{-1}B_0 & -B_m^{-1}B_1 & -B_m^{-1}B_2 & \ldots & -B_m^{-1}B_{m-1} \end{Vmatrix}.$$

Таким чином, кожну $(m+1)$-ярусну схему можна розглядати як двох'ярусну в просторі H^m. Відповідно стійкість $(m+1)$-ярусної схеми (52) можна визначити як стійкість еквівалентної двох'ярусної схеми (53).

Означення 1. Схему (52) називатимемо стійкою, якщо для її розв'язку $y_n = (y^n, y^{n+1}, \ldots, y^{n+m-1})$ при будь-яких початкових даних $y_0 = (y^0, y^1, \ldots, y^{m-1})$ і будь-яких правих частинах $\varphi^n \in H$ виконується оцінка

$$\|y_n\|_{(1_n)} \leqslant M_1 \|y_0\|_{(1_0)} + M_2 Q_n[\varphi],$$

де $\|y_k\|_{(1_k)}$ — деяка норма у просторі H^m; $Q_n[\varphi] = Q_n[\varphi^0, \varphi^1, \ldots, \varphi^{n-1}]$ — функціонал від $\varphi = (\varphi^0, \ldots, \varphi^{n-1})$ з властивостями норми; $M_1 > 0$, $M_2 > 0$ — сталі, які не залежать від n (отже, і від τ; якщо $H = H_h$ — простір сіткових функцій, заданих на сітці ω_h, то сталі M_1, M_2 не повинні залежати від h, τ).

Так само, як і у випадку двох'ярусних схем, можна розрізняти окремо стійкість за початковими умовами і стійкість за правою частиною багатоярусних схем, тобто коли виконуються відповідно оцінки

$$\|y_n\|_{(1_n)} \leqslant M_1 \|y_0\|_{(1_0)} \text{ при } \varphi \equiv 0$$

і

$$\|y_n\|_{(1_n)} \leqslant M_2 Q_n[\varphi] \text{ при } y_0 = 0.$$

У двох'ярусних схемах достатні умови стійкості виражалися у вигляді операторних нерівностей типу $A \geqslant 0$ або $A - B \geqslant 0$, де A, B — деякі оператори в H. У просторі H^m такі нерівності записуватимуться для матричних операторів виду $\mathscr{P} = (\mathscr{P}_{\alpha\beta})_{\alpha,\beta=\overline{1,m}}$, і нам важливо знати, за яких умов, накладених на оператори $\mathscr{P}_{\alpha\beta}$, вони мають місце.

Розглянемо випадок простору H^2 і оператора $\mathscr{P} : H^2 \to H^2$:

$$\mathscr{P} = \begin{pmatrix} \mathscr{P}_{11} & \mathscr{P}_{12} \\ \mathscr{P}_{21} & \mathscr{P}_{22} \end{pmatrix}, \quad \mathscr{P}_{\alpha\beta} : H \to H.$$

Доведення грунтуються на приведенні тотожності (51) до вигляду

$$(\mathscr{P}y, y) = (D_1 y^1, y^1) + (D_2 y^2, y^2), \tag{54}$$

де y^1, y^2 — деякі елементи простору H, а D_1, D_2 — оператори в H. Таким чином, питання про додатність чи невід'ємність оператора $\mathscr{P}$ в H^2 зводиться до аналогічного питання для D_1 і D_2 в H.

Нехай C_1, C_2 — оператори в H такі, що існують оператори $(I - C_2C_1)^{-1}$, $(I - C_1C_2)^{-1}$. Виконаємо в (51) заміну

$$x^1 = y^1 - C_2 y^2, \quad y^1 = (I - C_2C_1)^{-1}(x^1 + C_2 x^2),$$

$$x^2 = y^2 - C_1 y^1, \quad y^2 = (I - C_1C_2)^{-1}(C_1 x^1 + x^2),$$

яка перетворює (51) в рівність

$$(\mathscr{P}x, x) = (Dy, y),$$

де

$$y = (y^1, y^2), \quad D = (D_{\alpha\beta}),$$

$$\begin{aligned} D_{11} &= \mathscr{P}_{11} - 2\mathscr{P}_{12}C_1 + C_1^*\mathscr{P}_{22}C_1, \\ D_{12} = D_{21}^* &= \mathscr{P}_{12} - \mathscr{P}_{11}C_2 + C_1^*\mathscr{P}_{12}^*C_2 - C_1^*\mathscr{P}_{22}, \\ D_{22} &= \mathscr{P}_{22} - 2C_2^*\mathscr{P}_{12} + C_2^*\mathscr{P}_{11}C_2. \end{aligned} \tag{55}$$

Оператори C_1, C_2 підбиратимемо так, щоб

$$D_{12} = \mathscr{P}_{12} - \mathscr{P}_{11}C_2 + C_1^*\mathscr{P}_{12}^*C_2 - C_1^*\mathscr{P}_{22} = 0. \tag{56}$$

Лема 2. *Нехай існує оператор $C_2 : H \to H$ такий, що*

$$\mathscr{P}_{12} = \mathscr{P}_{11}C_2. \tag{57}$$

Тоді невід'ємність самоспряженого оператора $\mathscr{P} = (\mathscr{P}_{\alpha\beta}) : H^2 \to H^2$ еквівалентна умовам

$$\mathscr{P}_{11} \geqslant 0, \quad \mathscr{P}_{22} \geqslant C_2^*\mathscr{P}_{11}C_2.$$

Доведення. Покладемо в (56) $C_1 \equiv 0$. Тоді з урахуванням (57) матимемо $D_{12} = 0$ і рівності (55) набирають вигляду

$$D_{11} = \mathscr{P}_{11}, \quad D_{22} = \mathscr{P}_{22} - C_2^*\mathscr{P}_{11}C_2.$$

Тому за умов леми

$$(\mathscr{P}x, x) = (\mathscr{P}_{11}y^1, y^1) + ((\mathscr{P}_{22} - C_2^*\mathscr{P}_{11}C_2)x^2, x^2),$$

де $y^1 = x^1 + C_2x^2$, звідки і випливає твердження леми.

Наслідок 1. *Нехай існує оператор $\mathscr{P}_{11}^{-1}$, тоді для невід'ємності $\mathscr{P} = (\mathscr{P}_{\alpha\beta})$ необхідно і достатньо, щоб виконувалися умови*

$$\mathscr{P}_{11} \geqslant 0, \quad \mathscr{P}_{22} \geqslant \mathscr{P}_{12}^*\mathscr{P}_{11}^{-1}P_{12}.$$

Для доведення досить переписати (57) у вигляді $C_2 = \mathscr{P}_{11}^{-1}\mathscr{P}_{12}$ і скористатися лемою 2.

Наслідок 2. *Нехай існує оператор* $\mathscr{P}_{22}^{-1}$, *тоді для невід'ємності* $\mathscr{P} = (\mathscr{P}_{\alpha\beta})$ *необхідно і достатньо, щоб*

$$\mathscr{P}_{22} \geqslant 0, \quad \mathscr{P}_{11} \geqslant \mathscr{P}_{12}\mathscr{P}_{22}^{-1}\mathscr{P}_{12}^{*}.$$

Аналогічно можна довести й інші ознаки знакосталості операторів в H^2.

Перейдемо до розгляду триярусних схем вигляду

$$B_2 y^{n+2} + B_1 y^{n+1} + B_0 y^n = \varphi^n, \quad n = 0, 1, 2, \ldots, \quad y^0, y^1 \text{ задано}, \tag{58}$$

де $y^{n+\alpha} = y(t_{n+\alpha}) \in H$, $t_{n+\alpha} = (n+\alpha)\tau$, $\varphi^n \in H$, $B_\alpha : H \to H$.
Якщо покласти $\varphi = \varphi^n$, $y = y^{n+1}$, $y_{\overset{\circ}{t}} = (y^{n+2} - y^n)/(2\tau)$,

$$y_{\overline{t}t} = (y^{n+2} - 2y^{n+1} + y^n)\tau^{-2}, \quad A = B_2 + B_1 + B_0,$$

$$B = \tau(B_2 - B_0), \quad R = 0{,}5(B_2 + B_0), \quad B_2 = \frac{1}{2\tau}B + R, \tag{59}$$

$$B_1 = A - 2R, \quad B_0 = -\frac{1}{2\tau}B + R,$$

то схему (58) можна записати в такому канонічному вигляді:

$$By_{\overset{\circ}{t}} + \tau^2 R y_{\overline{t}t} + Ay = \varphi, \quad y^0, y^1 \text{ задано}. \tag{60}$$

При дослідженні двох'ярусних схем ми використовували норми просторів H_A чи H_B, пов'язаних з операторами A та B схеми. У триярусних схемах вибір норм не такий очевидний, але метод енергетичних нерівностей залишається ефективним. Проілюструємо його на схемі (60) при $\varphi = 0$, тобто дослідимо цю схему на стійкість за початковими умовами.

Нехай H — простір із скалярним добутком (y, v) і нормою $\|y\| = \sqrt{(y, y)}$, а оператори A та R схеми (60) самоспряжені. Помножимо рівняння (60) при $\varphi = 0$ на $y_{\overset{\circ}{t}}$ скалярно:

$$(By_{\overset{\circ}{t}}, y_{\overset{\circ}{t}}) + \tau^2 (Ry_{\overline{t}t}, y_{\overset{\circ}{t}}) + (Ay, y_{\overset{\circ}{t}}) = 0. \tag{61}$$

З означень $y_{\overline{t}t}$ та $y_{\overset{\circ}{t}}$ випливає, що

$$(Ry_{\overline{t}t}, y_{\overset{\circ}{t}}) = \frac{1}{2\tau}((Ry_t, y_t) - (Ry_{\overline{t}}, y_{\overline{t}}) + (Ry_t, y_{\overline{t}}) - (Ry_{\overline{t}}, y_t)),$$

а в силу самоспряженості R

$$(Ry_t, y_{\overline{t}}) - (Ry_{\overline{t}}, y_t) = 0,$$

звідки

$$(Ry_{\overline{t}t}, y_{\overset{\circ}{t}}) = 0{,}5(Ry_{\overline{t}}, y_{\overline{t}})_t. \tag{62}$$

З тотожності

$$(Ay, y_t) = \left(A\left(\frac{\hat{y}+y}{2} - \frac{\hat{y}-y}{2}\right), \frac{\hat{y}-y}{\tau}\right) =$$

$$= \frac{1}{2\tau}(A(\hat{y}+y), \hat{y}-y) - 0{,}5\tau(Ay_t, y_t), \quad \hat{y} = y^{n+2}, \quad y = y^{n+1}$$

і самоспряженості A випливає рівність

$$(Ay, y_t) = 0{,}5(Ay, y)_t - 0{,}5\tau(Ay_t, y_t).$$

Аналогічно знаходимо

$$(Ay, y_{\overline{t}}) = 0{,}5(Ay, y)_{\overline{t}} + 0{,}5\tau(Ay_{\overline{t}}, y_{\overline{t}}).$$

Тому

$$(Ay, y_{\overset{\circ}{t}}) = 0{,}5(Ay, y_t) + 0{,}5(Ay, y_{\overline{t}}) =$$

$$= \frac{1}{4}((Ay, y) + (A\check{y}, \check{y}))_t - \frac{\tau^2}{4}(Ay_{\overline{t}}, y_{\overline{t}})_t, \quad \check{y} = y^n.$$

Використовуючи тотожність

$$(Ay, y) + (A\check{y}, \check{y}) = 0{,}5(A(y+\check{y}), y+\check{y}) + 0{,}5(A(y-\check{y}), y-\check{y}) =$$

$$= 0{,}5(A(y+\check{y}), y+\check{y}) + 0{,}5\tau^2(Ay_{\overline{t}}, y_{\overline{t}}),$$

дістаємо

$$(Ay, y_{\overset{\circ}{t}}) = \frac{1}{8}(A(y^n + y^{n+1}), y^n + y^{n+1})_t - \frac{\tau^2}{8}(Ay_{\overline{t}}, y_{\overline{t}})_t \tag{63}$$

Підставляючи (62), (63) в (61), знайдемо енергетичну тотожність

$$(By_{\overset{\circ}{t}}, y_{\overset{\circ}{t}}) + \frac{\tau^2}{2}\left(\left(R - \frac{1}{4}A\right)y_{\overline{t}}, y_{\overline{t}}\right)_t + \frac{1}{8}(A(y^n + y^{n+1}),$$
$$y^n + y^{n+1})_t = 0. \tag{64}$$

Нехай $B \geqslant 0$, тоді з останньої тотожності випливає нерівність

$$\mathscr{E}_{n+1} \leqslant \mathscr{E}_n,$$

де квадратична форма $\mathscr{E}_n$ визначається рівністю

$$\mathscr{E}_n \equiv \mathscr{E}_n[y^n, y^{n+1}] = \frac{1}{4}(A(y^n + y^{n+1}), y^n + y^{n+1}) +$$
$$+ \left(\left(R - \frac{1}{4}A\right)(y^{n+1} - y^n), y^{n+1} - y^n\right).$$

Неважко помітити, що при

$$A > 0, \quad R > \frac{1}{4}A \tag{65}$$

вираз $\sqrt{\mathscr{E}_n[y^n, y^{n+1}]}$ визначає норму в H^2. Тому нерівність

$$B \geqslant 0 \tag{66}$$

разом з (65) забезпечує стійкість схеми (60) в нормі

$$\|y_n\|_A = \sqrt{\mathscr{E}_n[y^n, y^{n+1}]} = \sqrt{\frac{1}{4}\|y^n + y^{n+1}\|_A^2 + \|y^{n+1} - y^n\|^2_{R-\frac{1}{4}A}}. \tag{67}$$

Отже, ми довели таке твердження.

Теорема 7. *Нехай в схемі* (60) $A = A^* > 0$, $R = R^* > 0$ *— сталі оператори. Тоді умови*

$$B = B(t) \geqslant 0 \quad \forall t \in \omega_\tau, \quad R > \frac{1}{4}A$$

є достатніми для стійкості схеми (60) *за початковими умовами, тобто для виконання нерівності*

$$\|y_{n+1}\|_A \leqslant \|y_n\|_A,$$

де $y_n = (y^n, y^{n+1})$.

Такий самий результат можна дістати за допомогою теорії стійкості двох'ярусних схем, записавши схему (60) у вигляді

$$\mathscr{B}y_t + \mathscr{A}y = \mathscr{F}, \quad y_0 \text{ — задано}, \tag{68}$$

де $\mathscr{F} = (f^1, f^2) \in H^2$, $y_t = (y_{n+1} - y_n)/\tau$, $\mathscr{A}$, $\mathscr{B} : H^2 \to H^2$ (зазначимо, що запис багатоярусної схеми у вигляді двох'ярусної не єдиний). Якщо ввести оператор

$$\mathscr{A} = \begin{pmatrix} A & 0 \\ 0 & R - \frac{1}{4}A \end{pmatrix} \tag{69}$$

і вектор

$$y_n = \left(\frac{1}{2}(y^n + y^{n+1}), y^{n+1} - y^n\right)^T, \tag{70}$$

$y_n \in H^2$, то неважко помітити, що

$$\mathscr{E}_n[y^n, y^{n+1}] = (\mathscr{A}y_n, y_n).$$

Знайдемо оператор $\mathscr{B} = (B_{\alpha,\beta}) : H^2 \to H^2$, щоб схема (68), де y_n та $\mathscr{A}$ визначаються формулами (70), (69), була еквівалентною схемі (60). Оскільки

$$y_t = (y^{n+1}_{\circ t}, \tau y^{n+1}_{\bar{t}t}),$$

то (68) можна записати у вигляді системи

$$B_{11}y^{n+1}_{\circ t} + \tau B_{12}y^{n+1}_{\bar{t}t} + \frac{1}{2}A(y^n + y^{n+1}) = f_1, \tag{71}$$

$$B_{21}y^{n+1}_{\circ t} + \tau B_{22}y^{n+1}_{\bar{t}t} + \left(R - \frac{1}{4}A\right)\tau y^{n+1}_{\bar{t}} = f_2. \tag{72}$$

Виберемо оператори $B_{\alpha\beta}$ так, щоб рівняння (71) було еквівалентне рівнянню (60), а ліва частина рівняння (72) перетворювалася тотожно на нуль. Для цього, враховуючи формулу

$$y^{n+1}_{\bar{t}} = y^{n+1}_{\circ t} - 0{,}5\tau y^{n+1}_{\bar{t}t},$$

запишемо (72) у вигляді

$$\left[B_{21} + \tau\left(R - \frac{1}{4}A\right)\right] y_{\circ t} + \tau\left[B_{22} - 0{,}5\tau\left(R - \frac{1}{4}A\right)\right] y_{\bar{t}t} = 0.$$

Звідси випливає, що рівняння (72) перетвориться на тотожність, якщо вибрати $f_2 = 0$,

$$B_{21} = -\tau\left(R - \frac{1}{4}A\right), \quad B_{22} = \frac{\tau}{2}\left(R - \frac{1}{4}A\right).$$

Оскільки

$$y^n = y^{n+1} - \tau y^{n+1}_{\circ t} + \frac{\tau^2}{2}y^{n+1}_{\bar{t}t},$$

то рівняння (71) еквівалентне рівнянню

$$(B_{11} - 0{,}5\tau A)\, y_{\circ t} + \tau^2\left(\tau^{-1}B_{12} + \frac{1}{4}A\right) y_{\bar{t}t} + Ay = 0.$$

Порівнюючи це з (60), дістаємо

$$B_{11} = B + \frac{\tau}{2}A, \quad B_{12} = \tau\left(R - \frac{1}{4}A\right), \quad f_1 = \varphi.$$

Отже, триярусна схема (60) еквівалентна двох'ярусній (68), де оператор A визначений формулою (69), вектор y_n — формулою (70) і

$$\mathscr{B} = \begin{pmatrix} B + 0{,}5\tau A & \tau\left(R - \frac{1}{4}A\right) \\ -\tau\left(R - \frac{1}{4}A\right) & \frac{\tau}{2}\left(R - \frac{1}{4}A\right) \end{pmatrix}.$$

Неважко перевірити, що коли $A = A^*$, $R = R^*$, то оператор $\mathscr{A} : H^2 \to H^2$ самоспряжений, а оператор $\mathscr{B}$ — несамоспряжений, причому $\mathscr{A} > 0$, якщо $A > 0$ і $R > \frac{1}{4}A$. Умова стійкості двох'ярусної схеми в H^2

$$\mathscr{B} \geqslant \frac{\tau}{2}\mathscr{A}$$

еквівалентна умовам

$$B = B(t) \geqslant 0 \quad \forall t \in \omega_\tau, \quad R \geqslant \frac{1}{4}A$$

і при цьому має місце твердження теореми 7.

Розглянемо тепер стійкість схеми (60) за правою частиною. Поклавши $y^0 = y^1 = 0$, ми замість енергетичної тотожності (64) матимемо

$$2\tau (By_{\overset{\circ}{t}}, y_{\overset{\circ}{t}}) + \|\hat{y}\|^2_{\mathcal{A}} = \|y\|^2_{\mathcal{A}} + 2\tau (\varphi, y_{\overset{\circ}{t}}), \qquad (73)$$

де $\hat{y} = y_{n+1}$, $y = y_n = (y^n, y^{n+1})$, $\varphi = \varphi^n$. За допомогою ε_0-нерівності $\left(|ab| \leqslant \varepsilon_0 a^2 + \frac{1}{4\varepsilon_0} b^2\right)$ маємо

$$2\tau (\varphi, y_{\overset{\circ}{t}}) \leqslant \tau\varepsilon_0 \|y_{\overset{\circ}{t}}\|^2 + \frac{\tau}{\varepsilon_0} \|\varphi\|^2,$$

де $\varepsilon_0 > 0$ — деяка стала. Припустимо, що

$$B \geqslant \varepsilon I, \quad \varepsilon = \text{const} > 0,$$

і покладемо $\varepsilon_0 = 2\varepsilon$. Тоді з (73) матимемо

$$\|\hat{y}\|^2_{\mathcal{A}} \leqslant \|y\|^2_{\mathcal{A}} + \frac{\tau}{2\varepsilon} \|\varphi\|^2.$$

Звідси, враховуючи, що $y^0 = 0$, $y^1 = 0$, тобто $y_0 = 0$, дістанемо таке твердження.

Теорема 8. *Нехай $A = A^* > 0$, $R = R^* > 0$ — сталі оператори. Тоді за умов*

$$B \geqslant \varepsilon I, \quad R \geqslant \frac{1}{4} A, \quad \varepsilon = \text{const} > 0$$

схема (60) *стійка і має місце нерівність*

$$\|y_{n+1}\|_{\mathcal{A}} \leqslant \|y_0\|_{\mathcal{A}} + \frac{1}{\sqrt{2\varepsilon}} \left[\sum_{j=0}^{n} \tau \|\varphi^j\|^2\right]^{1/2}.$$

При $y_0 = 0$ ця нерівність означає стійкість за правою частиною.

Норма, яку ми використовували вище, має не очевидну структуру і залежить від розв'язку задачі (60) на n-му та $(n+1)$-му ярусах, що можна пояснити залежністю розв'язку на кожному ярусі від розв'язку на двох попередніх ярусах. Можна знайти оцінки і в простіших нормах просторів H_A, H_R, що проілюструємо на триярусній схемі вигляду

$$(I + \tau^2 R) y_{\overset{\circ}{t}\overset{\circ}{t}} + By_{\overset{\circ}{t}} + Ay = \varphi, \quad y^0, \ y^1 \text{ — задані.} \qquad (74)$$

Ця схема формально утворюється з (60) заміною R на $\tilde{R} = R + \tau^{-2} I$. Беручи до уваги цю заміну, дійдемо висновку, що схема (74) стійка при $R \geqslant \frac{1}{4} A$, і зможемо записати відповідні оцінки. Триярусні схеми також можна записувати у вигляді

$$Dy_{\overset{\circ}{t}\overset{\circ}{t}} + By_{\overset{\circ}{t}} + Ay = \varphi(t), \quad 0 < t \in \omega_\tau, \quad y^0, \ y^0_t \text{ — задано,} \qquad (75)$$

де $y = y^n$, $y_{\overset{\circ}{t}\overset{\circ}{t}} = \tau^{-2}(y^{n+1} - 2y^n + y^{n-1})$, $y_{\overset{\circ}{t}} = (y^{n+1} - y^{n-1})/(2\tau)$, $\varphi = \varphi^n = \varphi(n\tau)$, $n = 1, 2, \ldots$, $D = D(t)$, $A = A(t)$, $B = B(t)$ — лінійні оператори. Зокрема, для схеми (60) $D = \tau^2 R$, а для схеми (74) $D = I + \tau^2 R$.

Розглянемо також задачі

$$Dy_{\overset{\circ}{t}\overset{\circ}{t}} + By_{\overset{\circ}{t}} + Ay = 0, \quad y^0 = y(0), \quad y^0_t = y_t(0) \text{ задано,} \qquad (76)$$

$$Dy_{\overset{\circ}{t}\overset{\circ}{t}} + By_{\overset{\circ}{t}} + Ay = \varphi(t), \quad y^0 = y^0_t = 0. \qquad (77)$$

Для оцінок у просторах H_A, H_R будуть потрібні двосторонні оцінки для норми $\|y_n\|^2_{A,D} = \frac{1}{4}\|y^n + y^{n+1}\|^2_A + \left(\left(D - \frac{\tau^2}{4} A\right) y^n_t, \ y^n_t\right)$.

Лема 3. *Нехай виконуються умови*

$$A = A^* \geqslant 0, \quad D = D^* > 0, \quad B(t) \geqslant 0, \quad D \geqslant \frac{1+\varepsilon}{4} \tau^2 A, \qquad (78)$$

де $\varepsilon > 0$ — довільна стала. Тоді

$$\|y_n\|_{A,D} \leqslant \|y^n\|_A + \|y^n_t\|_D, \qquad (79)$$

$$\|y_n\|_{A,D} \geqslant \sqrt{\frac{\varepsilon}{1+\varepsilon}} \|y^{n+1}\|_A, \qquad (80)$$

$$\|y_n\|_{A,D} \geqslant \frac{1}{2} \sqrt{\frac{\varepsilon}{1+\varepsilon}} (\|y^{n+1}\|_A + \|y^n_t\|_D), \quad y_n = (y^n, y^{n+1}). \qquad (81)$$

Доведення. Позначимо $y = y^n$, $\hat{y} = y^{n+1}$,

$$J = \frac{1}{4}\|\hat{y} + y\|^2_A + \left(\left(D - \frac{\tau^2}{4} A\right) y_t, \ y_t\right).$$

Запишемо величину J у вигляді

$$J = \frac{1}{4}(\|y\|^2_A + 2(Ay, \hat{y}) + \|\hat{y}\|^2_A) - \frac{1}{4}(\|y\|^2_A - 2(A\hat{y}, y) + \|\hat{y}\|^2_A) + (Dy_t, y_t) = (Ay, \hat{y}) + \|y_t\|^2_D. \qquad (82)$$

В силу самоспряженості A вираз (82) можна подати також у вигляді

$$J = (A\hat{y}, y) + \|y_t\|^2_D. \qquad (83)$$

Підставимо $\hat{y} = y + \tau y_t$ у (82)

$$J = \|y\|^2_A + \tau(Ay, y_t) + \|y_t\|^2_D \leqslant \|y\|^2_A + \tau\|y\|_A \|y_t\|_A + \|y_t\|^2_D.$$

Використовуючи умову $D \geqslant \frac{1+\varepsilon}{4} \tau^2 A$, маємо

$$\|y_t\|_A \leqslant \frac{2}{\tau\sqrt{1+\varepsilon}} \|y_t\|_D,$$

отже,

$$J \leqslant \|y\|_A^2 + \frac{2}{\sqrt{1+\varepsilon}} \|y\|_A \|y_t\|_D + \|y_t\|_D^2 < (\|y\|_A + \|y_t\|_D)^2,$$

звідки і дістаємо (79).

Підставивши в (82) $y = \hat{y} - \tau y_t$, матимемо

$$J = (A\hat{y}, \hat{y}) - \tau (A\hat{y}, y_t) + \|y_t\|_D^2.$$

Звідси і з нерівностей Коші — Буняковського $(A\hat{y}, y_t) \leqslant \|\hat{y}\|_A \|y_t\|_A$ та $D \geqslant \frac{1+\varepsilon}{4} \tau A$ випливає

$$J \geqslant \|\hat{y}\|_A^2 - \tau \|\hat{y}\|_A \|y_t\|_A + \|y_t\|_D^2 \geqslant$$
$$\geqslant \|\hat{y}\|_A^2 - \frac{2}{\sqrt{1+\varepsilon}} \|\hat{y}\|_A \|y_t\|_D + \|y_t\|_D^2.$$

Застосуємо далі нерівність $|ab| \leqslant \delta a^2 + \frac{1}{4\delta} b^2$. Тоді

$$J \geqslant (1-\delta) \|\hat{y}\|_A^2 + \left(1 - \frac{1}{\delta(1+\varepsilon)}\right) \|y_t\|_D^2. \tag{84}$$

Поклавши $\delta = \frac{1}{1+\varepsilon}$, знаходимо

$$J \geqslant \frac{\varepsilon}{1+\varepsilon} \|\hat{y}\|_A^2,$$

що і доводить (80). Щоб дістати оцінку (81), виберемо δ за умови рівності коефіцієнтів при $\|\hat{y}\|_A^2$ і $\|y_t\|_D^2$ в (84), тобто

$$\delta = \frac{1}{\sqrt{1+\varepsilon}}, \quad 1-\delta = \frac{\sqrt{1+\varepsilon}-1}{\sqrt{1+\varepsilon}} = \frac{\varepsilon}{1+\varepsilon+\sqrt{1+\varepsilon}}.$$

Оскільки $\sqrt{1+\varepsilon} < 1+\varepsilon \; \forall \, \varepsilon > 0$, то $1-\delta > \frac{\varepsilon}{2(1+\varepsilon)}$, і тому

$$J \geqslant \frac{\varepsilon}{2(1+\varepsilon)} (\|\hat{y}\|_A^2 + \|y_t\|_D^2) \geqslant \frac{\varepsilon}{4(1+\varepsilon)} (\|\hat{y}\|_A + \|y_t\|_D)^2,$$

що повністю доводить лему.

Враховуючи, що оператор R в схемі (60) пов'язаний з оператором D в схемі (75) рівністю $D = \tau^2 R$, маємо $\|y_n\|_{\mathcal{A}} = \|y_n\|_{A,D}$ і з теореми 8 знаходимо оцінку

$$\|y_n\|_{A,D} \leqslant \|y_0\|_{A,D} + \frac{1}{\sqrt{2\varepsilon}} \left[\sum_{j=0}^{n-1} \tau \|\varphi^j\|^2\right]^{1/2},$$

або з урахуванням (79), (80)

$$\|y^{n+1}\|_A \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} \left\{\|y^0\|_A + \|y_t^0\|_D + \frac{1}{\sqrt{2\varepsilon}} \left[\sum_{i=0}^{n-1} \tau \|\varphi^j\|^2\right]^{1/2}\right\}. \tag{85}$$

Отже, ми довели таке твердження.

Теорема 9. *Нехай* $B \geqslant \varepsilon I$, $D \geqslant \frac{1+\varepsilon}{4} \tau^2 A$, $A = A^* > 0$, $R = R^* > 0$. *Тоді схема* (75) *стійка, тобто має місце нерівність* (85).

Нерівність $B \geqslant \varepsilon I$ можна послабити. Для цього розглянемо окремо задачі

$$Dy_{\bar{t}t} + By_{\bar{t}}^{\circ} + Ay = \varphi(t), \quad 0 < t \in \omega_\tau, \quad y^0 = y_t^0 = 0, \tag{86}$$

та

$$Dy_{\bar{t}t} + Ay_t^{\circ} + Ay = 0, \quad 0 < t \in \omega_\tau, \quad y^0, \; y_t^0 \text{ задано}. \tag{87}$$

Користуючись теоремою 7 (з урахуванням, що $D = \tau^2 R$) та нерівностями (79), (80) за умов

$$A = A^* > 0, \quad D = D^* > 0, \quad B \geqslant 0, \quad D \geqslant \frac{1+\varepsilon}{4} \tau^2 A, \tag{88}$$

для задачі (87) маємо

$$\|y^{n+1}\|_A \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} \|y_n\|_{A,D} \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} \|y_{n-1}\|_{A,D} \leqslant \ldots \leqslant$$
$$\leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} \|y_0\|_{A,D} \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} (\|y^0\|_A + \|y_t^0\|_D). \tag{89}$$

Це означає стійкість схеми (75) за початковими умовами.

Для оцінки розв'язку задачі (86) скористаємося принципом суперпозиції, згідно з яким шукатимемо цей розв'язок у вигляді

$$y^n = \sum_{s=1}^{n} \tau g_s^n, \quad n = 1, 2, \ldots, y^0 = 0,$$

де g_s^n як функція від n при будь-якому фіксованому s задовольняє рівняння

$$D(g_s^n)_{\bar{t}t} + B(g_s^n)_{\bar{t}}^{\circ} + Ag_s^n = 0 \tag{90}$$

і початкові умови

$$(0{,}5\tau B + D) \frac{g_s^{s+1} - g_s^s}{\tau} = \varphi^s, \quad g_s^s = 0. \tag{91}$$

Тоді, як неважко помітити,

$$y^{n+1} = \sum_{s=1}^{n+1} \tau g_s^{n+1}, \quad y^n = \sum_{s=1}^{n} \tau g_s^n, \quad y^{n-1} = \sum_{s=1}^{n-1} \tau g_s^{n-1},$$

$$y_t^\circ = 0{,}5\tau^{-1}(y^{n+1} - y^{n-1}) = \sum_{s=1}^{n-1} \tau (g_s^n)_t^\circ + (2\tau)^{-1}(\tau g_{n+1}^{n+1} + \tau g_n^{n+1}) =$$

$$= \sum_{s=1}^{n-1} \tau (g_s^n)_t^\circ + 0{,}5 g_n^{n+1},$$

$$y_{\bar{t}t} = \sum_{s=1}^{n-1} \tau (g_s^n)_{\bar{t}t} + \tau^{-2}(\tau g_{n+1}^{n+1} + \tau g_n^{n+1} - 2\tau g_n^n) = \sum_{s=1}^{n-1} \tau (g_s^n)_{\bar{t}t} + \tau^{-1} g_n^{n+1},$$

$$D y_{\bar{t}t} + B y_t^\circ + A y = \tau^{-1} D g_n^{n+1} + 0{,}5 B g_n^{n+1} = \varphi^n.$$

В силу умов (88) для розв'язку w рівняння

$$(0{,}5\tau B + D) w = \varphi$$

маємо

$$\| w \|_D^2 \leqslant 0{,}5\tau (Bw,\ w) + (Dw,\ w) = (\varphi,\ D^{1/2} D^{-1/2} w) =$$

$$= (D^{-1/2}\varphi,\ D^{1/2} w) \leqslant \| D^{-1/2}\varphi \| \| D^{1/2} w \| =$$

$$= \sqrt{(D^{-1}\varphi,\ \varphi)(Dw,\ w)} = \| \varphi \|_{D^{-1}} \| w \|_D,\ \ \| w \|_D \leqslant \| \varphi \|_{D^{-1}}.$$

Тому з (91) дістаємо

$$\| (g_s^s)_t \|_D \leqslant \| \varphi^s \|_{D^{-1}},$$

а застосовуючи оцінку (89) до задач (90), (91), маємо

$$\| g_s^n \|_A \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} \| (g_s^s)_t \|_D \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} \| \varphi^s \|_{D^{-1}}.$$

Користуючись нерівністю трикутника, для розв'язку y задачі (86) знаходимо оцінку

$$\| y^n \|_A \leqslant \sum_{s=1}^{n} \tau \| g_s^n \|_A = \sum_{s=1}^{n-1} \tau \| g_s^n \|_A \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} \sum_{s=1}^{n-1} \tau \| \varphi^s \|_{D^{-1}}.$$

Сформулюємо знайдені результати у вигляді наступної теореми.

Теорема 10. *Нехай виконуються умови* (88). *Тоді схема* (75) *стійка за початковими умовами і правою частиною, а саме, має місце оцінка*

$$\| y^n \|_A \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} \left(\| y^0 \|_A + \| y_t^0 \|_D + \sum_{s=1}^{n-1} \tau \| \varphi^s \|_{D^{-1}} \right). \qquad (92)$$

Наслідок. *Нехай* $D = I + \tau^2 R > I$, $D^{-1} < I$, *тоді* $\| \varphi^s \|_{D^{-1}} \leqslant \| \varphi^s \|$ *і для* (86) *правильна оцінка*

$$\| y^n \|_A \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} \sum_{s=1}^{n-1} \tau \| \varphi^s \|. \qquad (93)$$

Для рівняння

$$D y_{\bar{t}t} + A y = \varphi,\ \ 0 < t = n\tau < t_0,$$
$$y(0) = y^0,\ \ y_t^\circ(0) = \bar{y}^0 \qquad (94)$$

можна дістати більш сильні оцінки. Для цього припустимо, що

$$A = A^* > 0,\ \ D = D^* > 0,\ \ D \geqslant \frac{1+\varepsilon}{4} \tau^2 A, \qquad (95)$$

де ε *— абсолютна позитивна стала. Поклавши* $x = D^{1/2} y$, $C = D^{-1/2} A D^{-1/2}$, *перетворимо* (94)

$$x_{\bar{t}t} + Cx = \tilde{\varphi},\ \ x(0) = x_0,\ \ x_t(0) = \bar{x}_0. \qquad (96)$$

Застосувавши до (96) *оператор* C^{-1}, *дістанемо схему*

$$C^{-1} x_{\bar{t}t} + x = C^{-1}\tilde{\varphi},\ \ x(0) = x_0,\ \ x_t(0) = \bar{x}_0. \qquad (97)$$

Порівнюючи (97) зі схемою (75) і встановлюючи відповідність $C^{-1} \sim D$, $I \sim A \sim C^{-1}\tilde{\varphi} \sim \varphi$, помічаємо, що це схеми одного класу. Умова $D \geqslant \frac{1+\varepsilon}{4} \tau^2 A$ набирає вигляду

$$C^{-1} \geqslant \frac{1+\varepsilon}{4} \tau^2 I$$

або

$$I \geqslant \frac{1+\varepsilon}{4} \tau^2 C.$$

Скористаємося тепер оцінкою (92):

$$\| x^n \| \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} \left(\| x(0) \| + \| x_t(0) \|_{C^{-1}} + \sum_{s=1}^{n-1} \tau \| C^{-1}\tilde{\varphi}^s \|_C \right). \qquad (98)$$

Враховуючи, що $x = D^{1/2} y$, $\tilde{\varphi} = D^{-1/2}\varphi$,

$$\| x_t(0) \|_{C^{-1}}^2 = (C^{-1} x_t(0),\ x_t(0)) = (D^{1/2} A^{-1} D^{1/2} D^{1/2} y_t(0),\ D^{1/2} y_t(0)) =$$

$$= \| D y_t(0) \|_{A^{-1}}^2,$$

$$\| C^{-1}\tilde{\varphi} \|_C^2 = (C^{-1}\tilde{\varphi},\ \tilde{\varphi}) = (D^{1/2} A^{-1} D^{1/2} D^{-1/2}\varphi,\ D^{-1/2}\varphi) =$$

$$= (A^{-1}\varphi,\ \varphi) = \| \varphi \|_{A^{-1}}^2,$$

запишемо (98) у попередніх змінних:

$$\| y^n \|_D \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} \left(\| y(0) \|_D + \| D y_t(0) \|_{A^{-1}} + \sum_{s=1}^{n-1} \tau \| \varphi^s \|_{A^{-1}} \right). \qquad (99)$$

Таким чином, доведено наступне твердження.

Теорема 11. *Якщо для схеми* (94) *виконуються умови* (95), *то має місце оцінка* (99), *яка означає стійкість цієї схеми за початковими умовами і правою частиною. Зокрема, для схеми* (94) *при* $D = I$, $y^0 = \bar{y}^0 = 0$ *вірною є нерівність*

$$\|y^n\| \leqslant \sqrt{\frac{1+\varepsilon}{\varepsilon}} \sum_{s=1}^{n-1} \tau \|\varphi^s\|_{A^{-1}}.$$

Приклад 8. Дослідити на стійкість схему з ваговими коефіцієнтами

$$y_{\bar{t}t} + A(\sigma\hat{y} + (1-2\sigma)\, y + \sigma\check{y}) = \varphi, \tag{100}$$

де $A^* = A \geqslant \delta I$, $\delta > 0$, $y = y(t)$, $\hat{y} = y(t+\tau)$, $\check{y} = y(t-\tau)$.

Розв'язання. Неважко помітити, що

$$\sigma\hat{y} + (1-2\sigma)\, y + \sigma\check{y} = y + \sigma\tau^2 y_{\bar{t}t},$$

тому схема (100) запишеться у вигляді (94) з $D = I + \sigma\tau^2 A$. Умова стійкості (див. теорему 11)

$$D \geqslant \frac{1+\varepsilon}{4}\tau^2 A$$

або

$$I \geqslant \left(\frac{1+\varepsilon}{4} - \sigma\right)\tau^2 A$$

виконується при

$$\sigma \geqslant \frac{1+\varepsilon}{4} - \frac{1}{\tau^2\|A\|}. \tag{101}$$

Для явної схеми ($\sigma = 0$) звідси маємо таку умову стійкості:

$$\tau \leqslant \frac{2}{\sqrt{(1+\varepsilon)\|A\|}}, \tag{102}$$

де $\varepsilon > 0$ — будь-яке число.

Приклад 9. Дослідити на стійкість схему

$$y_t^{\circ} + \varkappa\tau^2 y_{\bar{t}t} + Ay = 0 \quad (y^0,\ y^1 \text{ задано}),$$

де $A = A^* \geqslant 0$.

Розв'язання. Порівнюючи цю схему з канонічною схемою (60), помічаємо, що $R = \varkappa I$, $B = I$. За теоремою 7 схема буде стійкою, коли $R > \frac{1}{4}A$, тобто $\varkappa I > \frac{1}{4}A$ або

$$\varkappa > \frac{1}{4}\|A\|.$$

1.4.6. Стаціонарні ітераційні методи розв'язування лінійних операторних рівнянь.

Розглянемо операторне рівняння

$$Au = f, \tag{103}$$

де $A : H \to H$ — заданий лінійний оператор у скінченновимірному просторі H розмірності N із скалярним добутком $(,)$ і нормою $\|u\| = \sqrt{(u, u)}$, f — заданий елемент простору H, u — шуканий елемент з H. Головна ідея ітераційних методів для рівняння (103) полягає в побудові послідовностей елементів $\{y_k\}$, яка в певному розумінні збігається до точного розв'язку рівняння (103). Член цієї послідовності y_k називається k-ю ітерацією, а k — номером ітерації. Послідовність ітерацій можна будувати, наприклад, за лінійною формулою (схемою):

$$B_k y_{k+1} = C_k^{(0)} y_k + C_k^{(1)} y_{k-1} + \cdots + C_k^{(m)} y_{k-m} + F_k, \tag{104}$$

де B_k, $C_k^{(i)}$ — задані лінійні оператори в H, $F_k \in H$ — елементи заданої послідовності. Спочатку задаються елементи $y_0, y_1, \ldots, y_m$, а потім за формулою (104) послідовно знаходять $y_{m+1}, y_{m+2}, \ldots$. Щоб за відомими $y_{k-m}, \ldots, y_k$ з (104) можна було визначити y_{k+1}, слід припустити, що існують B_k^{-1}. Формула (104) і задає лінійний ітераційний метод. Якщо при обчисленні y_{k+1} використовується лише попередня ітерація y_k, то ітераційний метод називається однокроковим або двох'ярусним. Аналогічно дається означення двохкрокових або триярусних методів (схем) і т. д.

Ітераційний метод називається збіжним, якщо для деякої норми

$$\lim_{k\to\infty} \|y_k - u\|_{(1)} = 0. \tag{105}$$

Щоб зупинити обчислення, задають деяке число $\varepsilon > 0$, яке є характеристикою похибки, і обчислення припиняють, коли виконується нерівність

$$\|y_n - u\|_{(1)} \leqslant \varepsilon \|y_0 - u\|_{(1)}. \tag{106}$$

Елемент y_n вважають наближеним (з точністю ε) значенням для точного розв'язку u. Оскільки u — невідомий елемент, то часто перевірити умову (106) неможливо. Тому на практиці обчислення припиняють, коли

$$\|Ay_n - f\|_{(1)} \leqslant \varepsilon \|Ay_0 - f\|_{(1)}, \tag{107}$$

де $Ay_n - f = Ay_n - Au = r_n$ — нев'язка, яка завжди може бути обчислена. Критерієм зупинки (107) слід користуватися обережно, оскільки можна навести приклади рівняння (103), для яких виконується (107), але не виконується (106). Підкреслимо, що в загальному випадку умова зупинки ітераційного процесу не є тривіальною. Часто норму $\|\cdot\|_{(1)}$ вибирають за формулою

$$\|u\|_{(1)} = \sqrt{(Du, u)}, \tag{108}$$

де $D = D^* > 0$, $D : H \to H$. Елементи простору H з нормою (108) утворюють нормований простір H_D, який називається *енергетичним простором оператора* D.

Розглянемо лінійний однокроковий ітераційний метод

$$B_k y_{k+1} = C_k y_k + F_k, \quad k = 0, 1, \ldots, \tag{109}$$

в якому початкове наближення (нульова ітерація) y_0 задане. Для зручності формування теорем збіжності і перевірки умов цих теорем метод (109) записують у канонічній формі:

$$B_k \frac{y_{k+1} - y_k}{\tau_{k+1}} + A y_k = f, \quad k = 0, 1, \ldots, \tag{110}$$

де $\tau_{k+1} > 0$ — деякі числові параметри. Покажемо, що це можна зробити.

Дійсно, зазначимо, що природно вимагати, щоб точний розв'язок рівняння (103) (а він не залежить від k) задовольняв (109) для будь-яких k:

$$(B_k - C_k) u = F_k.$$

Порівнюючи це з точним рівнянням (103), покладемо

$$F_k = \tau_{k+1} f, \quad A = \tau_{k+1}^{-1} (B_k - C_k). \tag{111}$$

Підставляючи (111) в (109), дістанемо (110).

В ітераційному процесі (110) в нашому розпорядженні маємо B_k та τ_{k+1}. Якщо y_k вже відоме, то y_{k+1} знаходимо як розв'язок операторного рівняння

$$B_k y_{k+1} = \Phi_k, \tag{112}$$

де $\Phi_k = B_k y_k - \tau_{k+1} (A y_k - f)$. Якщо $B_k = I$, то метод (109) називається *явним*, бо в цьому разі маємо явну формулу для $(k+1)$-ї ітерації: $y_{k+1} = \Phi_k$. При $B \neq I$ метод (109) називається *неявним*. Якщо $B_k \equiv B$, $\tau_{k+1} \equiv \tau$, тобто не залежать від k, то метод (110) називається *стаціонарним*, в іншому разі — *нестаціонарним*. Зрозуміло, що свободою у виборі B_k, τ_{k+1}, треба скористатися так, щоб, по-перше, рівняння (112) розв'язувалось простіше, ніж (103), і за мінімальну кількість арифметичних операцій, по-друге, щоб швидкість збіжності ітераційного процесу була максимальною. У цьому і полягає головна задача теорії ітераційних методів. Якщо $n = n(\varepsilon)$ — найменше з чисел, для яких виконується (106) або (107), то число арифметичних дій для визначення наближеного розв'язку дорівнює $Q_n(\varepsilon) = n(\varepsilon)\, q_0$, де q_0 — число дій при обчисленні одного y_k (однієї ітерації). Задача полягає в побудові такого ітераційного процесу (110), тобто виборі B_k, $\{\tau_k\}$, щоб $Q_n(\varepsilon)$ було при заданому ε мінімальним.

Ітераційний процес (110) можна записати також у вигляді

$$y_{k+1} = y_k - \tau_{k+1} w_k,$$

де $w_k = B_k^{-1} (A y_k - f) \equiv B_k^{-1} r_k$ — поправка.

Неважко помітити, що похибка $z_k = y_k - u$ задовольняє умови

$$B_k \frac{z_{k+1} - z_k}{\tau_{k+1}} + A z_k = 0, \quad k = 0, 1, 2, \ldots, \tag{113}$$

де $z_0 = y_0 - u$. Якщо $B_k \equiv B$, тобто не залежить від k, то поправка $w_k = B^{-1} r_k$ також задовольняє однорідні співвідношення

$$B \frac{w_{k+1} - w_k}{\tau_{k+1}} + A w_k = 0, \quad k = 0, 1, \ldots. \tag{114}$$

Дійсно, діючи на рівність $y_{k+1} - y_k = -\tau_{k+1} w_k$ оператором A і враховуючи, що

$$A y_{k+1} - A y_k = (A y_{k+1} - f) - (A y_k - f) = r_{k+1} - r_k,$$

$$r_{k+1} - r_k = B (B^{-1} r_{k+1} - B^{-1} \tau_k) = B (w_{k+1} - w_k),$$

дістаємо рівняння (114).

Враховуючи, що $y_{k+1} - y_k = A^{-1} (r_{k+1} - r_k)$, з виразу (110) маємо

$$B_k A^{-1} \frac{r_{k+1} - r_k}{\tau_{k+1}} + r_k = 0,$$

і якщо $B_k = I$, то звідси дістаємо рівняння для нев'язки

$$\frac{r_{k+1} - r_k}{\tau_{k+1}} + A r_k = 0.$$

Теорема 12. *Нехай $A = A^* > 0$ і виконується умова*

$$B > \frac{\tau}{2} A. \tag{115}$$

Тоді стаціонарний ітераційний метод

$$B \frac{y_{k+1} - y_k}{\tau} + A y_k = f \tag{116}$$

збігається в енергетичному просторі H_A, тобто $\| z_k \|_A = \| y_k - u \|_A \to 0$ при $k \to \infty$.

Доведення. Застосовуючи до задачі для похибки (113) першу частину теореми 6, дістаємо ланцюжок нерівностей

$$0 \leqslant \| z_{k+1} \|_A \leqslant \cdots \leqslant \| z_0 \|_A,$$

тобто послідовність $\{\| z_k \|_A\}$ незростаюча і обмежена знизу. За теоремою Вейєрштрасса існує границя $\lim \| z_k \|_A = z$. Доведемо, що $z = 0$. Оскільки оператор $P = B - \frac{\tau}{2} A$ додатний, то він і додатно визначений (див. 1.4.1), тобто існує стала $\delta > 0$ така, що

$$(Py, y) \geqslant \delta \| y \|^2 \; \forall \, y \in H.$$

Тому з енергетичної тотожності (48), записаної для z_k, маємо

$$\frac{2\delta}{\tau} \| z_{k+1} - z_k \|^2 + \| z_{k+1} \|_A^2 \leqslant \| z_k \|_A^2.$$

Переходячи в цій нерівності до границі при $k \to \infty$ і враховуючи, що $\| z_{k+1} \|_A \to z$, $\| z_k \|_A \to z$, дістаємо

$$\lim_{k\to\infty} \| z_{k+1} - z_k \| = 0. \tag{117}$$

З рівняння (113) знаходимо

$$Az_k = -\frac{1}{\tau} B (z_{k+1} - z_k), \quad z_k = -\frac{1}{\tau} A^{-1} B (z_{k+1} - z_k), \tag{118}$$

$$(Az_k, z_k) = \frac{1}{\tau^2} (B (z_{k+1} - z_k), A^{-1} B (z_{k+1} - z_k)),$$

$$\| z_k \|_A^2 \leqslant \frac{1}{\tau^2} \| A^{-1} \| \| B \|^2 \| z_{k+1} - z_k \|^2.$$

Звідси із (117) випливає $\lim\limits_{k\to\infty} \| z_k \|_A = 0$, що і треба було довести.

На практиці велике значення має швидкість збіжності. Розглянемо один з випадків, коли можна визначити швидкість збіжності ітераційного процесу (116).

Теорема 13. *Нехай* $A = A^* > 0$, $B = B^* > 0$, $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, $\gamma_2 \geqslant \gamma_1 > 0$. *Тоді за умови*

$$\tau \leqslant \tau_0 = \frac{2}{\gamma_1 + \gamma_2} \tag{119}$$

виконується нерівність

$$\| Ay_n - f \|_{B^{-1}} \leqslant \rho_0^n \| Ay_0 - f \|_{B^{-1}}, \tag{120}$$

яка характеризує швидкість збіжності ітераційного процесу (116), *причому*

$$\rho_0 = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2. \tag{121}$$

Доведення. Для поправки w_k маємо рівняння (див. (114))

$$B \frac{w_{k+1} - w_k}{\tau} + Aw_k = 0, \quad k = 0, 1, \ldots, \quad w_0 = B^{-1} (Ay_0 - f). \tag{122}$$

З теореми 4 і рівняння (122) при $\tau \leqslant \tau_0$ дістаємо

$$\| w_k \|_B \leqslant \rho^k \| w_0 \|_B,$$

де $\rho = |1 - \tau\gamma_1|$. Мінімум ρ досягається при $\tau = \tau_0$, причому

$$\rho \geqslant \rho_0 = 1 - \tau_0\gamma_1 = \frac{1 - \xi}{1 + \xi}.$$

Врахувавши, що $\| w_k \|_B = \| B^{-1} r_k \|_B = \| r_k \|_{B^{-1}}$, дістанемо твердження теореми.

Зауваження. На практиці задають число ε й ітерації припиняються, коли $\| Ay_n - f \|_{B^{-1}} \leqslant \varepsilon \| Ay_0 - f \|_{B^{-1}}$, що буде при виконанні нерівності

$$\rho_0^n \leqslant \varepsilon \text{ або } \left(\frac{1}{\rho_0}\right)^n \geqslant \frac{1}{\varepsilon}. \tag{123}$$

Звідси знаходимо оцінку для числа ітерацій

$$n \geqslant \frac{\ln (1/\varepsilon)}{\ln (1/\rho_0)}. \tag{124}$$

За умов теореми 13 цю оцінку можна знайти через вихідні величини γ_1, γ_2, ε. Дійсно, функція $\varphi(\xi) = \ln (1 + \xi)/(1 - \xi) - 2\xi$ додатна для всіх $\xi \in (0, 1)$, оскільки $\varphi'(\xi) = 2\xi^2/(1 - \xi^2) > 0$, $\varphi(0) = 0$. Тому $\ln^{-1} \frac{1}{\rho_0} < \frac{1}{2\xi}$ і умова (124) буде виконана, коли

$$n \geqslant n_0(\varepsilon) = \frac{1}{2\xi} \ln \frac{1}{\varepsilon}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \tag{125}$$

Зазначимо, що $n_0(\varepsilon)$ є оцінкою знизу для числа ітерацій і не обов'язково має бути цілим. Оцінка $\rho_0^n \leqslant \varepsilon$, очевидно, виконується, якщо $n_0(\varepsilon) \leqslant n < n_0(\varepsilon) + 1$. Тому за n (тобто достатню для виконання (123) кількість ітерацій) досить взяти цілу частину числа $n_0(\varepsilon) + 1$.

Розглянемо тепер деякі конкретні ітераційні методи для СЛАР

$$Au = f, \tag{126}$$

де $A = (a_{ij})_{i,j=\overline{1,N}}$ — квадратна матриця, $f = (f_i)_{i=\overline{1,N}}$ — заданий вектор, $u = (u_i)_{i=\overline{1,N}}$ — шуканий вектор.

Метод простої ітерації. Цей метод має вигляд

$$\frac{y_{k+1} - y_k}{\tau} + Ay_k = f, \quad k = 0, 1, \ldots, \tag{127}$$

де y_0 вибирається довільно. Порівнюючи (127) з (110), помічаємо що метод простої ітерації для СЛАР є явною двох'ярусною (однокроковою) стаціонарною схемою зі сталим параметром $\tau_k = \tau$. У координатній формі (127) має вигляд

$$y_{k+1}^{(i)} = y_k^{(i)} - \tau \left(\sum_{j=1}^{N} a_{ij} y_k^{(j)} - f^{(i)} \right), \quad i = \overline{1, N}. \tag{128}$$

Існують і інші варіанти простої ітерації, наприклад,

$$y_{k+1}^{(i)} = \frac{-1}{a_{ii}} \left(\sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij} y_k^{(j)} - f^{(i)} \right), \quad i = \overline{1, N}. \tag{129}$$

Підставивши сюди

$$\sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij} y_k^{(j)} = \sum_{j=1}^{N} a_{ij} y_k^{(j)} - a_{ii} y_k^{(i)} = (Ay_k)^{(i)} - (Dy_k)^{(i)},$$

де $D = \operatorname{diag} (a_{ii})_{i=\overline{1,N}}$ — діагональна матриця, дістанемо

$$y_{k+1}^{(i)} = y_k^{(i)} - \frac{1}{a_{ii}} \left(\sum_{j=1}^{N} a_{ij} y_k^{(j)} - f^{(i)} \right), \tag{130}$$

або в канонічному вигляді

$$D\frac{y_{k+1}-y_k}{\tau}+Ay_k=f,\quad k=0,\ 1,\ \ldots;\quad \tau=1. \tag{131}$$

Формально ця схема є неявною ($B=D\neq I$), але оскільки D — діагональна матриця, то y_{k+1} визначається за явними формулами (129).

Щоб з'ясувати умову збіжності методу простої ітерації (127), застосуємо теорему 12 при $B=I$. Враховуючи, що $I\geqslant\frac{1}{\|A\|}A$, маємо

$$B-\frac{\tau}{2}A=I-\frac{\tau}{2}A\geqslant\left(\frac{1}{\|A\|}-\frac{\tau}{2}\right)A>0$$

при $\frac{1}{\|A\|}-\frac{\tau}{2}>0$. Отже, метод простої ітерації для $A=A^*>0$ збігається, коли

$$\tau\leqslant\frac{2}{\|A\|}. \tag{132}$$

Якщо відомі границі спектра самоспряженого оператора $A>0$, тобто $\gamma_1 I\leqslant A\leqslant\gamma_2 I$, то для оцінки швидкості збіжності можна застосувати теорему 8.

Вправа 3. Враховуючи, що $B=I=B^*>0$, $A=A^*>0$, знайти умову збіжності методу простої ітерації з теореми 3.

Метод Зейделя. Цей метод досить поширений на практиці і застосовується в одній з двох форм:

$$\sum_{j=1}^{i}a_{ij}y_{k+1}^{(j)}+\sum_{j=i+1}^{N}a_{ij}y_k^{(j)}=f^{(i)},\quad a_{ii}\neq0,\ i=\overline{1,N}, \tag{133}$$

або

$$\sum_{j=1}^{i-1}a_{ij}y_k^{(j)}+\sum_{j=i}^{N}a_{ij}y_{k+1}^{(j)}=f^{(i)},\ i=\overline{1,N}. \tag{134}$$

На відміну від методу простої ітерації, де кожна компонента $(k+1)$-ї ітерації визначається лише через компоненти k-ї ітерації (тому всі компоненти можна обчислювати паралельно), в методі Зейделя щойно знайдена компонента $(k+1)$-ї ітерації застосовується для знаходження наступних компонент $(k+1)$-ї ітерації, тобто компоненти вектора y_{k+1} знаходять послідовно за формулами

$$\begin{aligned}y_{k+1}^{(1)}&=\frac{1}{a_{11}}\left(f^{(1)}-\sum_{j=2}^{N}a_{1j}y_k^{(j)}\right),\\ y_{k+1}^{(i)}&=\frac{1}{a_{ii}}\left(f^{(i)}-\sum_{j=i+1}^{N}a_{ij}y_k^{(j)}-\sum_{j=1}^{i-1}a_{ij}y_{k+1}^{(j)}\right),\quad i=\overline{2,N},\end{aligned} \tag{135}$$

для варіанта (133) і за формулами

$$\begin{aligned}y_{k+1}^{(N)}&=\frac{1}{a_{NN}}\left(f^{(N)}-\sum_{j=1}^{N-1}a_{Nj}y_k^{(j)}\right),\\ y_{k+1}^{(i)}&=\frac{1}{a_{ii}}\left(f^{(i)}-\sum_{j=1}^{i-1}a_{ij}y_k^{(j)}-\sum_{j=i+1}^{N}a_{ij}y_{k+1}^{(j)}\right),\\ i&=N-1,\ N-2,\ \ldots,\ 1,\end{aligned} \tag{136}$$

для варіанта (134). Запишемо цей метод у матричній формі. Для цього подамо матрицю A у вигляді

$$A=A^-+D+A^+,$$

де $D=\operatorname{diag}(a_{ii})_{i=\overline{1,N}}$ — діагональна матриця, A^- — нижня трикутна (піддіагональна) матриця з нулями на головній діагоналі $A^-=(\bar a_{ij})_{i,j=\overline{1,N}}$,

$$\bar a_{ij}=\begin{cases}a_{ij}, & j<i,\\ 0, & j\geqslant i,\end{cases}$$

$A^+=(a_{ij}^+)_{i,j=\overline{1,N}}$ — верхня трикутна (наддіагональна) матриця,

$$a_{ij}^+=\begin{cases}0, & j\leqslant i,\\ a_{ij}, & j>i.\end{cases}$$

За допомогою цих матриць рівняння (134) можна записати таким чином:

$$((A^++D)y_{k+1})^{(i)}+(A^-y_k)^{(i)}=f^{(i)},\quad i=\overline{1,N},$$

або в матричному вигляді

$$(A^++D)y_{k+1}+A^-y_k=f.$$

Після очевидних перетворень

$$\begin{aligned}(A^++D)y_{k+1}+A^-y_k&=(A^++D)(y_{k+1}-y_k)+(A^-+A^++D)y_k=\\ &=(A^++D)(y_{k+1}-y_k)+Ay_k\end{aligned}$$

запишемо метод Зейделя (134) в канонічному вигляді:

$$(D+A^+)(y_{k+1}-y_k)+Ay_k=f,\quad k=0,\ 1,\ \ldots. \tag{137}$$

Порівнюючи (137) з (110), помічаємо, що метод Зейделя відповідає однокроковій стаціонарній неявній схемі з $B=D+A^+$, $\tau_k\equiv1$. Але хоча схема неявна, в силу того, що $B=D+A^+$ — трикутна матриця, ітерацію y_{k+1} знаходять за явними формулами. Аналогічно записується і варіант (133) методу Зейделя:

$$(D+A^-)(y_{k+1}-y_k)+Ay_k=f,\quad k=0,\ 1,\ \ldots, \tag{138}$$

де $D+A^-$ — нижня трикутна матриця.

Умови збіжності методу Зейделя можна знайти з теореми 12. Дійсно, якщо $A = A^* > 0$, то ця теорема має місце і умова збіжності для (138) при $B = D + A^-$, $\tau \equiv 1$ після перетворення

$$B - \frac{1}{2}A = D + A^- - \frac{1}{2}(A^- + D + A^+) = \frac{D}{2} + \frac{1}{2}(A^- - A^+)$$

має вигляд

$$\left(\left(B - \frac{1}{2}A\right)y, y\right) = \frac{1}{2}(Dy, y) + \frac{1}{2}((A^+ - A^-)y, y) = \frac{1}{2}(Dy, y) > 0,$$

тобто $D > 0$. Але ця умова є наслідком умови $A > 0$. Дійсно, якщо $A > 0$ і $\xi = (\xi^{(1)}, 0, \ldots 0)$, то $(A\xi, \xi) = (D\xi, \xi) = a_{11}(\xi^{(1)})^2 \geq 0$, тобто $a_{11} > 0$. Аналогічно переконуємося, що $a_{ii} > 0$, $i = \overline{2, N}$. Отже, має місце така теорема збіжності методу Зейделя.

Теорема 14. *Якщо $A = A^* > 0$, то метод Зейделя збігається.*

Зазначимо, що умови збіжності методу Зейделя і методу простої ітерації перетинаються лише в тому розумінні, що можна побудувати СЛАР, для яких збігається метод простої ітерації і розбігається метод Зейделя, і навпаки.

Для оцінки швидкості збіжності методу Зейделя теорему 8 не можна застосувати, бо $B \neq B^*$. Але таку оцінку можна дістати за умови строгого діагонального переважання матриці $A = A^* > 0$, тобто

$$\sum_{\substack{j=1 \\ j \neq i}}^{N} |a_{ij}| \geqslant q\,|a_{ii}|, \quad i = \overline{1, N}, \quad q < 1. \tag{139}$$

Тоді для похибки $z_k = y_k - u$ матимемо (для варіанту (133), (138))

$$a_{ii}z_{k+1}^{(i)} = -\sum_{j<i} a_{ij}z_{k+1}^{(j)} - \sum_{j>i} a_{ij}z_k^{(j)},$$

$$|a_{ii}|\,|z_{k+1}^{(i)}| \leqslant \sum_{j<i} |a_{ij}|\,|z_{k+1}^{(i)}| + \sum_{j>i} |a_{ij}|\,|z_k^{(j)}|.$$

Якщо $\max |z_{k+1}^{(i)}|$ досягається при деякому $i = i_0$, то

$$\|z_{k+1}\|_C = |z_{k+1}^{(i_0)}|,$$

$$|a_{i_0i_0}|\,\|z_{k+1}\|_C \leqslant \sum_{j<i_0} |a_{i_0j}|\,\|z_{k+1}\|_C + \sum_{j>i_0} |a_{i_0j}|\,\|z_k\|_C,$$

$$\|z_{k+1}\|_C \leqslant \left[\sum_{j>i_0} |a_{i_0j}| \Big/ \left(|a_{i_0i_0}| - \sum_{j<i_0} |a_{i_0j}|\right)\right] \|z_k\|_C$$

і далі в силу умови (139)

$$\sum_{j>i_0} |a_{i_0j}| \leqslant q\,|a_{i_0i_0}| - \sum_{j<i_0} |a_{i_0j}| < q\left(|a_{i_0i_0}| - \sum_{j<i_0} |a_{i_0j}|\right), \tag{140}$$

$$\|z_{k+1}\|_C \leqslant q\,\|z_k\|_C \leqslant q^{k+1}\,\|z_0\|_C.$$

Остання нерівність означає, що при $A = A^* > 0$ і виконанні (139) метод Зейделя збігається зі швидкістю геометричної прогресії із знаменником $q < 1$.

Метод релаксації. З метою прискорення ітераційного процесу введемо в метод Зейделя деякий параметр ω:

$$(D + \omega A^-)\frac{y_{k+1} - y_k}{\omega} + Ay_k = f, \quad k = 0, 1, \ldots, \tag{141}$$

тобто дістаємо неявну стаціонарну однокрокову (двох'ярусну) схему виду (110) з $B = D + \omega A^-$, $\tau = \omega$.

Для того щоб ітераційний процес (141) мав смисл, тобто щоб можна було знайти y_{k+1} при відомому y_k, треба щоб існував оператор $B^{-1} = (D + \omega A^-)^{-1}$. Достатньою умовою існування B^{-1} є додатність B. При $A = A^* > 0$ маємо

$$(By, y) = ((D + \omega A^-)y, y) = (Dy, y) + \omega(A^-y, y),$$

$$(A^-y, y) = (A^+y, y), \quad (A^+)^* = A^-,$$

$$(Ay, y) = (Dy, y) + 2(A^-y, y),$$

$$(A^-y, y) = \frac{1}{2}((A - D)y, y).$$

Підставляючи останню рівність у першу, знаходимо

$$(By, y) = \left(1 - \frac{1}{2}\omega\right)(Dy, y) + \omega(Ay, y)$$

і помічаємо, що умова $(By, y) > 0$ виконується, коли

$$\omega \in (0, 2). \tag{142}$$

Саме при таких значеннях параметра релаксації ω і розглядатимемо метод релаксації. При $\omega \in [0, 1]$ цей метод називається *методом нижньої релаксації*, а при $\omega \in (1, 2)$ — *методом верхньої релаксації*.

Для СЛАР із самоспряженим додатним оператором A з теореми 12 дістаємо достатню умову збіжності методу релаксації:

$$(By, y) - \frac{\omega}{2}(Ay, y) = \left(1 - \frac{\omega}{2}\right)(Dy, y) + \frac{\omega}{2}(Ay, y) > 0.$$

Звідси випливає, що достатньою умовою збіжності є умова $A = A^* > 0$.

Параметром ω можна регулювати швидкість збіжності методу релаксації.

1.4.7. Нестаціонарні ітераційні методи розв'язування лінійних операторних рівнянь. Для рівняння (103) розглядатимемо ітераційний метод

$$B\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, \tag{143}$$

де y_0 — задане початкове наближення, τ_k — змінні параметри. Похибка $z_k = y_k - u$ та поправка $w_k = B^{-1}(Ay_k - f)$ задовольняють рівняння

$$B\frac{z_{k+1}-z_k}{\tau_{k+1}} + Az_k = 0, \quad k = 0, 1, \ldots; \; z_0 = y_0 - u. \tag{144}$$

Ітерації припиняються за умов

$$\|z_n\|_D \leqslant \varepsilon \|z_0\|_D$$

або

$$\|w_n\|_D \leqslant \varepsilon \|w_0\|_D, \; w_0 = B^{-1}(Ay_0 - f),$$

де D — додатний самоспряжений оператор, ε — задане додатне число. З (144) знаходимо

$$z_{k+1} = s_{k+1}z_k, \; s_{k+1} = I - \tau_{k+1}B^{-1}A, \tag{145}$$

де s_{k+1} — оператор переходу з яруса k на ярус $k+1$. Звідси

$$\|z_n\| = T_n z_0, \; T_n = s_n s_{n-1} \ldots s_1, \tag{146}$$

де T_n — *розв'язуючий оператор схеми* (145). Із (146) маємо

$$\|z_n\|_D \leqslant q_n \|z_0\|_D, \quad q_n = \|T_n\|_D.$$

Звідси випливає, що умова припинення ітерацій виконується, якщо

$$q_n = \|T_n\|_D \leqslant \varepsilon. \tag{147}$$

Розв'язуючи цю нерівність відносно $n = n(\varepsilon)$, знайдемо оцінку кількості ітерацій, необхідних для виконання умови припинення ітераційного процесу.

Чебишевська ітераційна схема (метод Річардсона). Розглянемо далі явну схему

$$\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, \tag{148}$$

за умов

$$A = A^* > 0, \quad \gamma_1 I \leqslant A \leqslant \gamma_2 I, \quad \gamma_1 > 0 \tag{149}$$

і виберемо параметри τ_k так, щоб мінімізувати кількість ітерацій $n(\varepsilon)$, необхідних для виконання умови виду (147). Для нев'язки $r_k = Ay_k - f$ виконується рівняння

$$\frac{r_{k+1}-r_k}{\tau_{k+1}} + Ar_k = 0, \quad k = 0, 1, \ldots, \; r_0 = Ay_0 - f$$

або

$$r_{k+1} = s_{k+1}r_k, \quad s_{k+1} = I - \tau_{k+1}A.$$

Звідси

$$r_n = T_n r_0, \quad T_n = s_1 s_2 \ldots s_n =$$
$$= (I - \tau_1 A)(I - \tau_2 A) \ldots (I - \tau_n A),$$

тобто T_n є поліномом степеня n відносно A, який ми позначимо $p_n(A) \equiv T_n$. Коефіцієнти цього полінома залежать лише від $\tau_1, \ldots, \tau_n$. Для нев'язки маємо оцінку

$$\|r_n\| \leqslant \|T_n\| \|r_0\| = \|p_n(A)\| \|r_0\|.$$

Якщо за умову припинення ітераційного процесу (148) вибрати нерівність

$$\|r_n\| \leqslant \varepsilon \|r_0\|, \tag{150}$$

де ε — задане додатне число, то вона виконується при

$$\|p_n(A)\| \leqslant \varepsilon. \tag{151}$$

Параметри $\tau_1, \tau_2, \ldots, \tau_m$ при заданому m в (148) виберемо так, щоб норма $\|p_m(A)\|$ була мінімальною, і покажемо, що, збільшуючи m, можна добитися умови (151).

Поліном

$$p_m(A) = \prod_{k=1}^{m}(I - \tau_k A) = c_0 + c_1 A + \cdots + c_m A^m, \; c_0 = 1,, \; p_m(0) = 1$$

є самоспряженим оператором. Нехай ξ_s, λ_s $(s = \overline{1, N})$ є відповідно власними функціями та власними значеннями оператора A, тобто

$$A\xi_s = \lambda_s \xi_s, \quad s = \overline{1, N}, \quad (\xi_s, \xi_m) = \delta_{sm}.$$

Оператор A^k має ті самі власні функції і власні значення λ_k^s. Тому

$$p_m(A)\xi_s = \sum_{k=0}^{m} c_k A^k \xi_s = \sum_{k=0}^{m} c_k \lambda_s^k \xi_s = p_m(\lambda_s)\xi_s,$$

тобто оператор $p_m(A)$ має власні функції ξ_s, $s = \overline{1, N}$, і відповідні власні значення $\lambda_s(p_m(A)) = p_m(\lambda_s)$. В силу самоспряженості оператора $p_m(A)$ його норма дорівнює найбільшому за модулем власному значенню:

$$\|p_m(A)\| = \max_{1 \leqslant s \leqslant N} |p_m(\lambda_s)|.$$

Оскільки за умовою власні числа оператора A розміщені на відрізку $[\gamma_1, \gamma_2]$, то

$$\max_{1 \leqslant s \leqslant N} |p_m(\lambda_s)| = \max_{x \in [\gamma_1, \gamma_2]} |p_m(x)|.$$

Отже, задача найкращого вибору параметрів $\tau_1, \tau_2, \ldots, \tau_m$ звелася до задачі відшукання

$$\min_{\tau_k, k=\overline{1,m}} \max_{x \in [\gamma_1, \gamma_2]} |p_m(x)|.$$

За допомогою заміни

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

відобразимо відрізок $x \in [\gamma_1, \gamma_2]$ на відрізок $[-1, 1]$ і позначимо $p_m(x) = p_m\left(\frac{1}{2}\left[(\gamma_1 - \gamma_2)\, t + \gamma_2 + \gamma_1\right]\right) = \tilde{p}_m(t)$. Умова нормування $p_m(0) = 1$ набере вигляду

$$\tilde{p}_m(t_0) = 1, \quad t_0 = \frac{1}{\rho_0}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \tag{152}$$

Оскільки параметри $\tau_1, \ldots, \tau_m$ цілком визначають поліноми $p_m(A)$, $p_m(x)$, $\tilde{p}_m(t)$, то нам треба знайти такий поліном $\tilde{p}_m(t)$, який мінімально відхиляється від нуля на відрізку $[-1, 1]$ (тобто має найменшу величину $\max\limits_{t\in[-1,1]} |\tilde{p}_m(t)|$) і задовольняє умову нормування (152). Таким поліномом є

$$\tilde{p}_m(t) = \frac{T_m(t)}{T_m(t_0)}, \tag{153}$$

де $T_m(t)$ — поліном Чебишева першого роду (див. п. 1.6, 2.4. 2), який має вигляд

$$T_m(t) = \begin{cases} \cos(m \arccos t), & |t| \leqslant 1, \\ \frac{1}{2}\left[(t + \sqrt{t^2-1})^m + (t - \sqrt{t^2-1})^m\right], & |t| > 1. \end{cases} \tag{154}$$

Неважко помітити, що поліном $T_m(t)$ має m нулів на відрізку $[-1, 1]$ і вони визначаються формулою

$$t_i = \cos\frac{2i-1}{2m}\pi, \quad i = \overline{1, m},$$

а поліном $p_m(x) = (1 - \tau_1 x) \ldots (1 - \tau_m x)$ має нулі $x_i = \frac{1}{\tau_i}$. Поліном $\tilde{p}_m(t)$ має нулі

$$t_i = \frac{2x_i}{\gamma_1 - \gamma_2} - \frac{\gamma_1 + \gamma_2}{\gamma_1 - \gamma_2}$$

і він збігатиметься з поліномом $T_m(t)/T_m(t_0)$, якщо вони матимуть однакові нулі, тобто

$$t_i = \frac{2}{\tau_i(\gamma_1 - \gamma_2)} - \frac{\gamma_1 + \gamma_2}{\gamma_1 - \gamma_2} = \frac{2/\tau_i - (\gamma_1 + \gamma_2)}{\gamma_1 - \gamma_2}.$$

Звідси

$$\frac{2}{\tau_i} = (\gamma_1 - \gamma_2)\, t_i + (\gamma_1 + \gamma_2),$$

$$\tau_i = \frac{2}{\gamma_1 + \gamma_2 - (\gamma_2 - \gamma_1)\, t_i} = \frac{2}{\gamma_2\left[1 + \xi - (1-\xi)\, t_i\right]} = \frac{2}{\gamma_2(1+\xi)\left[1 - \rho_0 t_i\right]} = \frac{\tau_0}{1 - \rho_0 t_i}, \tag{155}$$

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

Зазначимо, що коли $m = 1$, то $\tau_i = \tau_0$ (оптимальний параметр методу простої ітерації).

Якщо позначити (при заданому наперед m) через

$$\mathfrak{M}_m = \left\{-\cos\frac{2i-1}{2m}\pi, \quad i = \overline{1, m}\right\}$$

множину, що утворюється з нулів многочленів Чебишева $T_m(x) = \cos(m \arccos x)$, а через $\{\mu_k\}$ — будь-яку послідовність елементів цієї множини, то $\min\limits_{\{\tau_k\}} \|p_m(A)\|$ (а це означає, що і мінімальне число ітерацій $m(\varepsilon)$) досягається при значеннях параметрів

$$\tau_k = \frac{\tau_0}{1 + \rho_0 \mu_k}, \quad k = \overline{1, m}, \qquad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}. \tag{156}$$

Знайдемо тепер

$$q_m = \|p_m(A)\| = \max_{x\in[\gamma_1,\gamma_2]} |p_m(x)| = \max_{t\in[-1,1]} |\tilde{p}_m(t)| = \max_{t\in[-1,1]} \left|\frac{T_m(t)}{T_m(t_0)}\right| = \frac{1}{|T_m(t_0)|}.$$

Неважко помітити, що $t_0 > 1$. Тому, користуючись формулами (152), (154), маємо

$$t_0 \pm \sqrt{t_0^2 - 1} = \frac{1}{\rho_0} \pm \sqrt{\frac{1}{\rho_0^2} - 1} = \frac{1}{\rho_0}\left[1 \pm \sqrt{1 - \rho_0^2}\right] = \frac{1}{\rho_0}\left(1 \pm \frac{2\sqrt{\xi}}{1+\xi}\right) = \frac{1}{\rho_0}(1 \pm \sqrt{\xi})^2 (1+\xi)^{-1} =$$

$$= (1 \pm \sqrt{\xi})^2 (1-\xi)^{-1} = \frac{1 \pm \sqrt{\xi}}{1 \mp \sqrt{\xi}},$$

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

$$T_m(t_0) = \frac{1}{2}\left(\frac{1}{\rho_1^m} + \rho_1^m\right) = \frac{1 + \rho_1^{2m}}{2\rho_1^m} = \frac{1}{q_m},$$

$$q_m = \frac{2\rho_1^m}{1 + \rho_1^{2m}}.$$

Отже, ми довели, що для схеми (148) з ітераційними параметрами (156), яка називається *чебишевською ітераційною схемою,* після m ітерацій виконується оцінка

$$\|Ay_m - f\| \leqslant q_m \|Ay_0 - f\|, \qquad (157)$$

$$q_m = \frac{2\rho_1^m}{1+\rho_1^{2m}}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}} < 1$$

і вибором $m = m(\varepsilon)$ завжди можна досягти виконання нерівності

$$\|Ay_m - f\| \leqslant \varepsilon \|Ay_0 - f\| \quad \forall\, \varepsilon > 0. \qquad (158)$$

Вимога $q_m \leqslant \varepsilon$ або $2\rho_1^m \leqslant \varepsilon(1+\rho_1^{2m})$ виконується, якщо $\rho_1^m \leqslant \varepsilon/2$ або

$$m(\varepsilon) \geqslant \frac{\ln(2/\varepsilon)}{\ln(1/\rho_1)}. \qquad (159)$$

Зазначивши, що

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

замінимо (159) зручнішою для перевірки умовою

$$m(\varepsilon) > m_0(\varepsilon) = \frac{1}{2\sqrt{\xi}} \ln\frac{2}{\varepsilon}. \qquad (160)$$

Маючи γ_1, γ_2, за формулою (160) знаходимо найменше ціле m, для якого виконується (160), визначаємо параметри (156) і виконуємо m ітерацій за формулою (148); після чого справджується умова (158). Набір параметрів (156) *називають чебишевським.* Чебишевський ітераційний процес іноді називають також *методом Річардсона.* Він має, однак, одну особливість, яка до недавна заважала його практичному використанню. Ця особливість — збільшення проміжних значень, що призводить до автоматичної зупинки ЕОМ та накопичення похибок заокруглення. Це явище називається *обчислювальною нестійкістю.* Річ у тім, що при реальних розрахунках на ЕОМ завжди присутні похибки заокруглення, які при переході з k-ї до $(k+1)$-ї ітерації, очевидно, множаться на норму оператора переходу $\|s_{k+1}\| = \|I - \tau_{k+1}A\|$. З умови $\gamma_1 I \leqslant A \leqslant \gamma_2 I$ випливає, що

$$(\tau_{k+1}\gamma_1 - 1)I \leqslant \tau_{k+1}A - I \leqslant (\tau_{k+1}\gamma_2 - 1)I.$$

Підставляючи сюди вираз для τ_{k+1} і враховуючи, що $1 - \tau_0\gamma_1 = \tau_0\gamma_2 - 1 = \rho_0$, дістаємо

$$-\frac{\rho_0(1-\mu_k)}{1+\rho_0\mu_k} I \leqslant \tau_{k+1}A - I \leqslant \frac{\rho_0(1+\mu_k)}{1+\rho_0\mu_k} I.$$

Якщо припустити, що нерівності $\gamma_1 I \leqslant A \leqslant \gamma_2 I$ точні, тобто існують $y_1 \in H$, $y_2 \in H$ такі, що $\gamma_1\|y_1\|^2 = (Ay_1, y_1)$, $(Ay_2, y_2) = \gamma_2\|y_2\|^2$, то

із самоспряженості оператора s_{k+1} випливає, що

$$\|s_{k+1}\| = \|\tau_{k+1}A - E\| = \begin{cases} \dfrac{\rho_0(1+\mu_k)}{1+\rho_0\mu_k}, & \mu_k > 0, \\ \dfrac{\rho_0(1-\mu_k)}{1+\rho_0\mu_k}, & \mu_k < 0, \end{cases}$$

тобто $\|s_{k+1}\| < 1$ для $\mu_k > 0$ і $\|s_{k+1}\| > 1$ для $\mu_k < -(1-\rho_0)/(2\rho_0)$. Оскільки

$$-\cos\frac{\pi}{2m} \leqslant \mu_k \leqslant -\cos\frac{(2m-1)}{2m}\pi = \cos\frac{\pi}{2m}, \quad k = \overline{1, m},$$

і при великих m значення $\cos\frac{\pi}{2m}$ близьке до 1, то для великої кількості номерів k маємо $\|s_{k+1}\| > 1$. Тому якщо на багатьох ітераціях підряд використовуються багато параметрів τ_k, для яких $\|s_{k+1}\| > 1$, то відбувається накопичення похибок заокруглення, що і призводить до обчислювальної нестійкості.

Щоб послабити цей ефект, природно розмістити параметри τ_k в послідовності $\tau_1^*, \ldots, \tau_m^*$ так, щоб після параметра, для якого норма оператора переходу більша за 1, розміщався параметр, для якого вона менша 1. Такий набір параметрів $\{\tau_k^*\}$ називається *стійким.* Існують різні стійкі набори. Як приклад наведемо один з них. Нехай $m = 2^p$, $p > 0$ — ціле (існують набори для довільних m). Параметри τ_k однозначно визначаються нулями многочлена Чебишева μ_k, тому можна говорити про впорядкування μ_k. Стійкий набір нулів μ_k має вигляд

$$\mathfrak{M}_m^* = \left\{-\cos\beta_i,\ \beta_i = \frac{\pi}{2m}\Theta_i^{(m)},\quad i = \overline{1, m}\right\},\quad m = 2^p,$$

де $\Theta_i^{(m)}$ — одне з непарних чисел 1, 3, 5, ..., $2m-1$, тобто задача зводиться до впорядкування множини m непарних чисел

$$\Theta_m = \{\Theta_1^{(m)}, \ldots, \Theta_m^{(m)}\}.$$

Це виконується рекурентно: виходячи з множини $\theta_1 = \{1\}$, будується множина $\theta_m^* = \theta_{2^p}^*$ за формулами

$$\Theta_{2i-1}^{(2j)} = \Theta_i^{(j)},\ \Theta_{2i}^{(2j)} = 4j - \Theta_{2i-1}^{(2j)},\quad i = \overline{1, j},\ j = \overline{1, 2^{p-1}}.$$

Якщо, наприклад, $n = 16 = 2^4$, то послідовно знаходимо

$\Theta_1 = \{1\}$, $\Theta_2 = \{1, 3\}$, $\Theta_4 = \{1, 7, 3, 5\}$, $\Theta_8 = \{1, 15, 7, 9, 3, 13, 5, 11\}$,
$\Theta_{16} = \{1, 31, 15, 17, 7, 25, 9, 23, 3, 29, 13, 19, 5, 27, 11, 21\}$.

Здобуті вище результати для явної нестаціонарної схеми переносяться на неявну схему вигляду

$$B\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f,\quad k = 0, 1, \ldots \quad (y_0 \text{ — задано}), \qquad (161)$$

де

$$A = A^* > 0,\ B = B^* > 0,\ \gamma_1 B \leqslant A \leqslant \gamma_2 B,\ \gamma_1 > 0. \qquad (162)$$

Зведемо цю схему до еквівалентної явної схеми. Оскільки $B = B^* > 0$, то існує $B^{1/2} = (B^{1/2})^* > 0$. Подіявши оператором $B^{-1/2}$ на (161), дістанемо

$$\frac{x_{k+1} - x_k}{\tau_{k+1}} + Cx_k = 0,\quad k = 0,\ 1,\ \ldots;\ x_0 = B^{1/2} w_0, \qquad (163)$$

де $x_k = B^{1/2} w_k$, $C = B^{-1/2} A B^{-1/2}$, $w_k = B^{-1} r_k = B^{-1}(Ay_k - f)$.

Визначимо границі спектра оператора C. Для цього розглянемо функціонал

$$J = ((A - \gamma B) y,\ y) = (Ay,\ y) - \gamma (By,\ y) =$$
$$= (AB^{-1/2}(B^{1/2} y),\ B^{-1/2}(B^{1/2} y)) - \gamma (B^{1/2} y,\ B^{1/2} y) =$$
$$= (Cx,\ x) - \gamma (x,\ x) = ((C - \gamma I) x,\ x),\ x = B^{1/2} y.$$

Оскільки y (а значить і x) — довільний елемент з H, то з попередньої рівності випливає, що оператори $A - \gamma B$ і $C - \gamma I$ мають однакові знаки, тобто

$$\gamma_1 I \leqslant C \leqslant \gamma_2 I.$$

Тому задача вибору оптимальних параметрів в (161) розв'язується за допомогою (156), а з (157) знаходимо оцінку

$$\|Cx_m\| \leqslant q_m \|x_0\|.$$

Підставляючи $x_k = B^{1/2} w_k = B^{-1/2} r_k = B^{-1/2}(Ay_k - f)$, маємо

$$\|Ay_m - f\|_{B^{-1}} \leqslant q_m \|Ay_0 - f\|_{B^{-1}},$$

$$q_m = \frac{2\rho_1^m}{1 + \rho_1^{2m}},\ \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}},\ \xi = \frac{\gamma_1}{\gamma_2}. \qquad (164)$$

Зрештою зазначимо, що перехід до неявних схем часто буває виправданим, бо швидкість збіжності ітераційного процесу зростає. Оскільки вибір неявної схеми (а отже, і B) залежить від нас, то постає запитання: як це робити? Як вже зазначалося, основною є вимога, щоб число арифметичних дій $Q(\varepsilon)$ для відшукання розв'язку з точністю ε було мінімальним. А це число $Q(\varepsilon)$ залежить від двох факторів, які і формують вимоги до вибору B: 1) число ітерацій, яке залежить від $\{\tau_k\}$ і B, має бути мінімальним; 2) визначення наступної ітерації з рівняння

$$By_{k+1} = F_k$$

має виконуватися застосуванням мінімального числа арифметичних операцій (це вимога економності оператора B). Далі розглянемо приклад неявної схеми з економним оператором B.

Попереміннo-трикутний метод. Зазначимо спочатку, що коли в схемі (161) оператор B є добутком скінченного числа економних операторів, то він також економний. Наприклад, якщо $B = B_1 B_2$, де B_1, B_2 — нижня і верхня трикутні матриці, то B — економний. Дійсно, розв'язування СЛАР $By_{k+1} = F_k$ зводиться до розв'язування двох систем:

$$B_1 w = F_k,\quad B_2 y_{k+1} = w,$$

кожна з яких розв'язується за явними формулами за кількість арифметичних дій порядку $O(N^2)$ (відомо, що оптимальний алгоритм розв'язування СЛАР включає $O(N^2)$ арифметичних дій, а це означає, що для будь-якого алгоритму розв'язування СЛАР загального вигляду існує СЛАР, на який він витрачає $O(N^2)$ операцій).

Розглянемо ітераційний метод (161) з оператором (оператор такого виду називають ще факторизованим)

$$B = (D + \omega R_1) D^{-1} (D + \omega R_2), \qquad (165)$$

$D = D^* > 0$, $R_1^* = R_2$, $R_1 + R_2 = R$, $R = R^* > 0$, ω — параметр. Такий метод називається *попереміннo-трикутним* (назва стане зрозумілою пізніше).

Покажемо, що оператор B є самоспряженим і додатним. Справді,

$$(By,\ v) = ((D + \omega R_1) D^{-1} (D + \omega R_2) y,\ v) =$$
$$= ((D + \omega R_2) y,\ D^{-1} (D + \omega R_2) v) =$$
$$= (y,\ (D + \omega R_1) D^{-1} (D + \omega R_2) v) = (y,\ Bv),$$

тобто $B = B^*$. Звідси

$$(By,\ y) = ((D + \omega R_2) y,\ D^{-1} (D + \omega R_2) y) = \|(D + \omega R_2) y\|^2_{D^{-1}} > 0.$$

Це означає, що $B = B^* > 0$.

У скінченновимірному просторі оператору R відповідає матриця $R = (r_{ij})_{i,j=\overline{1,N}}$, причому якщо $R = R^*$, то ця матриця симетрична, тобто $r_{ij} = r_{ji}$. Тому якщо взяти

$$R_1 = (\bar{r}_{ij})_{i,j=\overline{1,N}},\ \bar{r}_{ij} = \begin{cases} r_{ii}/2, & i = j, \\ r_{ij}, & i > j, \\ 0, & j > i, \end{cases}$$

$$R_2 = (\overset{+}{r}_{ij})_{i,j=\overline{1,N}},\ \overset{+}{r}_{ij} = \begin{cases} r_{ii}/2, & i = j, \\ r_{ij}, & j > i, \\ 0, & j < i, \end{cases}$$

то $R_2 = R_1^*$. Нехай $D = (d_{ij})_{ij=\overline{1,N}}$ — діагональна матриця. Тоді матриця $D + \omega R_1$ буде нижньою трикутною, а $D + \omega R_2$ — верхньою трикутною і, таким чином, визначення кожної наступної ітерації з

рівняння $By_{k+1} = F_k$ зводиться до почережного обертання нижньої та верхньої трикутних матриць, тобто до розрахунків за явними формулами (звідси і назва методу). Дійсно, вказане рівняння має матрицю $B = (D + \omega R_1)\ D^{-1}\ (D + \omega R_2)$ та праву частину $F_k = By_k - \tau_{k+1}Ay_k + \tau_{k+1}f$. Позначивши $D^{-1}(D + \omega R_2)\ y_{k+1} = \bar{y}_{k+1}$, дістанемо дві системи з трикутними матрицями

$$(D + \omega R_1)\,\bar{y}_{k+1} = F_k,\ (D + \omega R_2)\,y_{k+1} = D\bar{y}_{k+1},\quad k = 0,\ 1,\ \ldots, \tag{166}$$

які при кожному k розв'язуються послідовно. Оскільки

$$(R_1 y,\ y) = (R_2 y,\ y) = (Ry,\ y)/2,\ D > 0,\ \omega > 0,\ R > 0, \text{ то}$$

$$((D + \omega R_1)y,\ y) = (Dy,\ y) + \omega\,(R_1 y,\ y) =$$

$$= \left(\left(D + \frac{\omega}{2}R\right)y,\ y\right) = ((D + \omega R_2)\,y,\ y) > 0,$$

тобто існують $(D + \omega R_1)^{-1}$, $(D + \omega R_2)^{-1}$ і системи (166) мають розв'язок.

Щоб скористатися загальною теорією, розвинутою вище для неявних стаціонарних методів, треба визначити сталі γ_1, γ_2 в нерівності

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \tag{167}$$

яка в силу обмеженості і додатності операторів A, B має місце. Але спочатку виберемо «найкраще» значення параметра ω.

Лема 4. *Нехай оператор B визначається за формулою* (165), *де $\omega > 0$ і R задовольняє умови*

$$R \geqslant \delta D,\ \delta > 0,\ R_1 D^{-1} R_2 \leqslant \frac{\Delta}{4}R,\quad \Delta > 0, \tag{168}$$

тоді має місце оцінка

$$\mathring{\gamma}_1(\omega)\,B \leqslant R \leqslant \mathring{\gamma}_2(\omega)\,B,$$

$$\mathring{\gamma}_1(\omega) = \frac{\delta}{1 + \omega\delta + 0{,}25\omega^2\delta\Delta},\ \mathring{\gamma}_2(\omega) = \frac{1}{2\omega}, \tag{169}$$

причому відношення $\xi(\omega) = \mathring{\gamma}_1(\omega)/\mathring{\gamma}_2(\omega)$ набуває найбільшого значення $\mathring{\xi} = \xi(\mathring{\omega})$ при

$$\omega = \mathring{\omega} = \frac{2}{\sqrt{\delta\Delta}} \tag{170}$$

і це найбільше значення дорівнює

$$\mathring{\xi} = \frac{\mathring{\gamma}_1}{\mathring{\gamma}_2} = \frac{2\sqrt{\eta}}{1 + \sqrt{\eta}}, \tag{171}$$

$$\eta = \frac{\delta}{\Delta},\ \mathring{\gamma}_1 = \frac{\delta}{2\,(1 + \sqrt{\eta})},\ \mathring{\gamma}_2 = \frac{\delta}{4\sqrt{\eta}}.$$

Доведення. Нерівності (168) означають, що

$$(Ry,\ y) \geqslant \delta\,(Dy,\ y),\ (D^{-1}R_2 y,\ R_2 y) \leqslant \frac{\Delta}{4}\,(Ry,\ y)\ \forall y \in H.$$

Подамо оператор B виразом

$$B = (D + \omega R_1)\,D^{-1}\,(D + \omega R_2) = D - \omega\,(R_1 + R_2) + \omega^2 R_1 D^{-1} R_2 +$$

$$+ 2\omega\,(R_1 + R_2) = (D - \omega R_1)\,D^{-1}\,(D - \omega R_2) + 2\omega R$$

і визначимо

$$(By,\ y) = (D^{-1}\,(D - \omega R_2)\,y,\ (D - \omega R_2)\,y) + 2\omega\,(Ry,\ y) =$$

$$= \|(D - \omega R_2)\,y\|^2_{D^{-1}} + 2\omega\,(Ry,\ y) \geqslant 2\omega\,(Ry,\ y),$$

тобто $B \geqslant 2\omega R$ або $R \leqslant \frac{1}{2\omega}B$, $\mathring{\gamma}_2(\omega) = \frac{1}{2\omega}$.

Враховуючи умову (168), маємо

$$B = D + \omega R + \omega^2 R_1 D^{-1} R_2 \leqslant \frac{1}{\delta}R + \omega R + \frac{\omega^2\Delta}{4}R =$$

$$= \frac{1}{\delta}\left(1 + \omega\delta + \frac{\omega^2\delta\Delta}{4}\right)R,$$

тобто знаходимо оцінку зверху для B:

$$\mathring{\gamma}_1(\omega)\,B \leqslant R,$$

$$\mathring{\gamma}_1(\omega) = \delta\left(1 + \omega\delta + \frac{\omega^2\delta\Delta}{4}\right)^{-1}.$$

Як ми бачили раніше (див. (164)), число ітерацій, необхідне для розв'язування рівняння $Ry = f$ методом (161), залежить від відношення

$$\xi(\omega) = \frac{\mathring{\gamma}_1(\omega)}{\mathring{\gamma}_2(\omega)} = 2\omega\delta\left(1 + \omega\delta + \frac{\omega^2\delta\Delta}{4}\right)^{-1}$$

і тим менше, чим більше $\xi(\omega)$. Виберемо ω з умови максимуму $\xi(\omega)$. З рівняння

$$\xi'(\omega) = 2\delta\left(1 - \frac{\omega^2\delta\Delta}{4}\right)\left(1 + \omega\delta + \frac{\omega^2\delta\Delta}{4}\right)^{-2} = 0$$

знаходимо

$$\omega = \mathring{\omega} = \frac{2}{\sqrt{\delta\Delta}},$$

причому при цьому значенні дійсно досягається максимум $\xi(\omega)$, бо $\xi''(\mathring{\omega}) < 0$. Підставляючи $\mathring{\omega}$ в формули для $\mathring{\gamma}_1(\omega)$, $\mathring{\gamma}_2(\omega)$, $\xi(\omega)$, дістаємо твердження леми.

До методу (161) з оператором B вигляду (165) належать і знайдені вище результати відносно найкращого вибору параметрів $\{\tau_k\}$, який визначається формулами (156). Розглянемо теорему про швидкість збіжності такого поперемінно-трикутного методу з чебишевським набором параметрів.

Теорема 15. *Нехай оператор $A = A^* > 0$ представляється у вигляді суми $A = A_1 + A_2$, $A_2 = A_1^*$ і виконуються умови*

$$A \geqslant \delta D, \quad A_1 D^{-1} A_2 \leqslant \frac{\Delta}{4} A, \quad \delta > 0, \quad \Delta > 0. \tag{172}$$

Тоді для поперемінно-трикутного методу (161) *з оператором* (165), *в якому $D = D^* > 0$, $\omega = \frac{2}{\sqrt{\delta\Delta}}$, $R = A$, $R_1 = A_1$, $R_2 = A_2$ з чебишевським набором параметрів*

$$\tau_k^* = \frac{\tau_0}{1 + \rho_0 \mu_k^*}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2} = \frac{2\sqrt{\eta}}{1 + \sqrt{\eta}}, \tag{173}$$

$$\gamma_1 = \frac{\delta}{2(1 + \sqrt{\eta})}, \quad \gamma_2 = \frac{\delta}{4\sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}, \quad \mu_k^* \in \mathfrak{M}_n^*,$$

де $\mathfrak{M}_n^*$ — множина нулів многочлена Чебишева $T_n(x)$, досить $n(\varepsilon)$ ітерацій, щоб виконувалась оцінка

$$\| Ay_n - f \|_{B^{-1}} \leqslant \varepsilon \| Ay_0 - f \|_{B^{-1}},$$

де

$$n_0(\varepsilon) \leqslant n(\varepsilon) < n_0(\varepsilon) + 1, \quad n_0(\varepsilon) = \frac{\ln 2/\varepsilon}{2\sqrt{2}\sqrt[4]{\eta}}. \tag{174}$$

Доведення. Поклавши в попередній лемі $R = A$, $R_1 = A_1$, $R_2 = A_2$ і скориставшись оцінками (164) та (160), дійдемо твердження теореми.

Приклад 10. Записати розрахункові формули поперемінно-трикутного методу (ПТМ), явного методу простої ітерації (ЯПІ) і явної чебишевської схеми (ЯЧС) для розв'язування СЛАУ (невідомі u_i, $i = \overline{1, N-1}$)

$$u_{\bar{x}x,i} \equiv \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} = -f_i, \quad i = \overline{1, N-1}, \quad u_0 = u_N = 0, \tag{175}$$

де $h = \frac{1}{N}$, N та f_i задані, а також порівняти кількість ітерацій, які потрібно затратити в цих методах, щоб домогтися виконання нерівностей

$$\| Ay_n - f \|_{B^{-1}} \leqslant \varepsilon \| Ay_0 - f \|_{B^{-1}} \quad \text{(ПТМ)},$$
$$\| Ay_n - f \| \leqslant \varepsilon \| Ay_0 - f \| \quad \text{(ЯПІ, ЯЧС)}.$$

Розв'язання. Нехай $\Lambda v = v_{\bar{x}x}$ і $\overset{\circ}{v}$ — сіткова функція, яка задана на сітці $\bar{\omega}_h = \{x_i = ih : i = \overline{0, N}\}$ і перетворюється на нуль при $i = 0, N$ (простір таких функцій позначимо через $\mathring{\Omega}_{N+1}$). Позначимо через $H = \Omega_{N-1}$ простір сіткових функцій, заданих на сітці $\omega_h = \{x_i = ih : i = \overline{1, N-1}\}$ із скалярним добутком

$$(y, v) = \sum_{i=1}^{N-1} h y_i v_i.$$

Ввівши матрицю A розмірності $(N-1) \times (N-1)$

$$A = -\frac{1}{h^2} \begin{bmatrix} -2 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -2 \end{bmatrix}$$

і вектори $u = (u_1, \dots, u_{N-1})$, $f = (f_1, \dots, f_{N-1})$, задачу (175) можемо записати в матричному вигляді

$$Au = f,$$

причому

$$Av = -\Lambda \overset{\circ}{v}, \quad v \in \Omega_{N-1} = H, \quad \overset{\circ}{v} \in \mathring{\Omega}_{N+1}.$$

Підсумовуючи частинами, неважко показати, що $(Ay, v) = (y, Av)$, тобто $A = A^*$. Неважко також знайти власні значення оператора A (див. п. 1.4.4), звідки

$$\lambda_{N-1}(v, v) \geqslant (Av, v) \geqslant \delta \| v \|^2, \quad \delta = \lambda_1 = \frac{4}{h^2} \sin^2 \frac{\pi h}{2},$$
$$\| A \| = \lambda_{N-1} = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}. \tag{176}$$

Покладемо $Dy = y$, тобто $D = I$,

$$(A_1 y)_i \equiv (R_1 y)_i = \frac{y_{\bar{x},i}}{h} = \frac{y_i - y_{i-1}}{h^2},$$

$$(A_2 y) \equiv (R_2 y) = -\frac{y_{x,i}}{h} = -\frac{y_{i+1} - y_i}{h^2}, \quad A = A_1 + A_2.$$

З рівності $y_{\bar{x},i+1} = y_{x,i}$ випливає, що $A_1^* = A_2$. Дійсно, оскільки $v_1 = v_0 + hv_{\bar{x},1} = hv_{\bar{x},1}$, то, підсумовуючи частинами, маємо

$$(A_2 y, v) = -\sum_{i=1}^{N-1} y_{x,i} v_i = -y_1 v_1 \frac{1}{h} - \sum_{i=1}^{N-1} y_{i+1} v_{x,i} =$$
$$= y_1 v_{\bar{x},1} + \sum_{i=2}^{N} y_i v_{\bar{x},i} = \sum_{i=1}^{N-1} y_i v_{\bar{x},i} = h \sum_{i=1}^{N-1} y_i \frac{v_{\bar{x},i}}{h} = (y, A_1 v).$$

Обчислимо сталу Δ:

$$(A_1A_2y,\ y) = (A_2y,\ A_2y) = \frac{1}{h^2}\sum_{i=1}^{N-1}(y_{x,i})^2 h = \frac{1}{h^2}\sum_{i=2}^{N}(y_{\bar{x},i})^2 h \leqslant$$

$$\leqslant \frac{1}{h^2}\sum_{i=1}^{N} h\,(y_{\bar{x},i})^2 = \frac{1}{h^2}\sum_{i=1}^{N-1} h\,(Ay)_i\,y_i = \frac{1}{h^2}(Ay,\ y),$$

тобто $\Delta = \frac{4}{h^2}$.

Ітерації (i — номер компоненти, k — номер ітерації) в ПТМ визначаються за формулами

$$((E+\omega A_1)\,\bar{y}^{(k+1)})_i = \bar{y}_i^{(k+1)} + \omega\frac{\bar{y}_i^{(k+1)} - \bar{y}_{i-1}^{(k+1)}}{h^2} = F_i^{(k)},$$

$$((E+\omega A_2)\,y^{(k+1)})_i = y_i^{(k+1)} - \omega\frac{y_{i+1}^{(k+1)} - y_i^{(k+1)}}{h^2} = \bar{y}_i^{(k+1)} \equiv (D\bar{y}^{(k+1)})_i,$$

$$F_i^{(k)} = (By^{(k)} - \tau_{k+1}Ay^{(k)} + \tau_{k+1}f)_i.$$

Перепишемо це так:

$$\bar{y}_i^{(k+1)} = \frac{\omega\bar{y}_{i-1}^{(k+1)} + h^2F_i^{(k)}}{h^2+\omega},\quad i = 1,\ 2,\ 3,\ \ldots,\ N-1,\ \bar{y}_0^{(k+1)} = 0;$$

$$y_i^{(k+1)} = \frac{\omega y_{i+1}^{(k+1)} + h^2\bar{y}_i^{(k+1)}}{h^2+\omega},\quad i = N-1,\ N-2,\ \ldots,\ 1;\quad y_N^{(k+1)} = 0. \tag{177}$$

Отже, кожна ітерація обчислюється за явними формулами (177): спочатку $\bar{y}_i^{(k+1)}$ послідовно для $i = 1, 2, \ldots, N-1$ і потім $y_i^{(k+1)}$ послідовно для $i = N-1, N-2, \ldots, 1$.

Враховуючи (176) і теорему 15, маємо при малих h

$$\eta = \frac{\delta}{\Delta} = \sin^2\frac{\pi h}{2} \approx \frac{\pi^2h^2}{4},\quad \sqrt{\eta} \approx \frac{\pi h}{2},$$

$$\xi = 2\sqrt{\eta}/(1+\sqrt{\eta}) \approx 2\sqrt{\eta} \approx \pi h,\quad \sqrt{\xi} \approx \sqrt{\pi h}$$

і оцінка для числа ітерацій ПТМ набирає вигляду

$$n_0(\varepsilon) \approx \frac{1}{2\sqrt{\pi h}}\ln\frac{2}{\varepsilon}. \tag{178}$$

Запишемо формули явного методу простої ітерації

$$y_1^{(k+1)} = y_1^{(k)} + \frac{\tau}{h^2}(-2y_1^{(k)} + y_2^{(k)} - h^2f_1),$$

$$y_i^{(k+1)} = y_i^{(k)} + \frac{\tau}{h^2}(y_{i-1}^{(k)} - 2y_i^{(k)} + y_{i+1}^{(k)} - h^2f_i),\quad i = \overline{2,\ N-2}, \tag{179}$$

$$y_{N-1}^{(k+1)} = y_{N-1}^{(k)} + \frac{\tau}{h^2}(y_{N-2}^{(k)} - 2y_{N-1}^{(k)} - h^2f_{N-1}),$$

де згідно з (119)

$$\tau = \frac{2}{\gamma_1+\gamma_2} = \frac{2}{\lambda_1+\lambda_{N-1}} = \frac{h^2}{2\left(\sin^2\frac{\pi h}{2} + \cos^2\frac{\pi h}{2}\right)} = \frac{h^2}{2}.$$

Враховуючи (125), (176), маємо таку оцінку кількості ітерацій ЯПІ:

$$\xi = \operatorname{tg}^2\frac{\pi h}{2} \approx \frac{\pi^2h^2}{4},$$

$$n_0(\varepsilon) = \frac{\ln(1/\varepsilon)}{2\xi} \approx \frac{2}{10h^2}\ln\frac{1}{\varepsilon}. \tag{180}$$

Розрахункові формули явного чебишевського методу мають вигляд (178), де замість τ стоїть τ_{k+1}. Відповідно до (160) маємо таку оцінку кількості ітерацій ЯЧС:

$$n_0(\varepsilon) = \frac{1}{2\sqrt{\xi}}\ln\frac{2}{\varepsilon} \approx \frac{1}{\pi h}\ln\frac{2}{\varepsilon}. \tag{181}$$

Вибравши $\varepsilon = 10^{-4}$, $N = 100$ і враховуючи, що $h = \frac{1}{N}$, для ПТМ, ЯПІ, ЯЧС відповідно матимемо такі значення $n_0(\varepsilon)$: 30, 20 000, 340. Отже, з розглянутих нами методів ПТМ виявився найкращим.

Варіаційно-ітераційні методи. Для ефективного використання розглянутих вище нестаціонарних методів треба було знати γ_1 та γ_2 (границі спектра оператора A). Часто буває так, що знайти чи оцінити ці границі неможливо. Тоді застосовують методи, які в явному вигляді не використовують γ_1, γ_2. Далі розглянемо деякі з них, а саме три методи варіаційного типу: метод мінімальних нев'язок (ММН), найшвидшого спуску (МНС) та спряжених градієнтів (МСГ).

Розглянемо *метод мінімальних нев'язок* для явної схеми

$$\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f,\quad k = 0,\ 1,\ \ldots. \tag{182}$$

Неважко бачити, що нев'язка $r_k = Ay_k - f$ задовольняє рівняння

$$\frac{r_{k+1}-r_k}{\tau_{k+1}} + Ar_k = 0,\quad k = 0,\ 1,\ \ldots,\ r_0 = Ay_0 - f.$$

Параметр τ_{k+1} вибирається з умови мінімуму норми нев'язки (звідси і назва):

$$\|r_{k+1}\|^2 = \|r_k - \tau_{k+1}Ar_k\|^2 = \|r_k\|^2 - 2\tau_{k+1}(r_k,\ Ar_k) + \tau_{k+1}^2\|Ar_k\|^2 \equiv$$
$$\equiv \varphi(\tau_{k+1}).$$

З рівняння

$$\varphi'(\tau_{k+1}) = -2(r_k,\ Ar_k) + 2\tau_{k+1}\|Ar_k\|^2 = 0$$

знаходимо

$$\tau_{k+1} = \frac{(Ar_k,\ r_k)}{\|Ar_k\|^2},\quad k = 0,\ 1,\ \ldots,$$

причому оскільки $\varphi''(\tau_{k+1}) = 2\|Ar_k\|^2 > 0$, то це значення параметра мінімізує норму $\|r_{k+1}\|$.

Якщо $A = A^* > 0$, то легко оцінити швидкість збіжності ММН. Дійсно, якщо ітераційний параметр τ_0 не збігається з (182), то

$$\|r_{k+1}\|^2 = \|r_k - \tau_{k+1}Ar_k\|^2 \leqslant \|r_k - \tau_0 Ar_k\|^2 \leqslant \|I - \tau_0 A\|^2 \|r_k\|^2.$$

Нехай відомо, що $\gamma_1 I \leqslant A \leqslant \gamma_2 I$, причому ці границі точні, тобто існують $y_1,\ y_2 \in H$ такі, що

$$(Ay_1,\ y_1) = \gamma_1 (y_1,\ y_1), \quad (Ay_2,\ y_2) = \gamma_2 (y_2,\ y_2).$$

Тоді при $\tau_0 \leqslant \dfrac{2}{\gamma_1 + \gamma_2}$

$$((I - \tau_0 A)y,\ y) = (y,\ y) - \tau_0 (Ay,\ y) \leqslant (1 - \tau_0\gamma_1)\|y\|^2,$$

причому

$$((I - \tau_0 A)y_1,\ y_1) = (1 - \tau_0\gamma_1)\|y_1\|^2,$$

що означає

$$\|I - \tau_0 A\| = 1 - \tau_0\gamma_1.$$

Неважко помітити, що мінімум цієї норми досягається при

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}$$

і дорівнює величині (див. доведення теореми 13)

$$\rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}.$$

Отже, для ММН при $A = A^* > 0$, $\gamma_1 I \leqslant A \leqslant \gamma_2 A$ маємо

$$\|r_{k+1}\| \leqslant \rho_0 \|r_k\|$$

і, порівнявши з теоремою 13, дійдемо висновку, що в цьому разі ММН збігається з такою самою швидкістю, як і метод простої ітерації.

Неявний метод мінімальних нев'язок, або *метод поправок*, має вигляд

$$B\frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0,\ 1,\ \ldots,\ \forall y_0 \in H.$$

Поправка $w_k = B^{-1}r_k$, $w_0 = B^{-1}(Ay_0 - f)$ задовольняє рівняння

$$B\frac{w_{k+1} - w_k}{\tau_{k+1}} + Aw_k = 0, \quad k = 0,\ 1,\ \ldots,$$

де згідно з тією самою ідеєю, що і в ММН, знаходимо

$$\tau_{k+1} = \frac{(Aw_k,\ w_k)}{(B^{-1}Aw_k,\ w_k)}, \quad k = 0,\ 1,\ \ldots,$$

а за умов $A = A^* > 0$, $B = B^* > 0$ замість (183) дістаємо оцінку

$$\|Ay_n - f\|_{B^{-1}} \leqslant \rho_0^n \|Ay_0 - f\|_{B^{-1}}.$$

Явний метод найшвидшого спуску ($B = E$) застосовується при $A = A^* > 0$ і відрізняється від (182) вибором τ_k.

За умови, що на кожній ітерації мінімізувалася норма похибки

$$\|z_{k+1}\|_A^2 = \|y_{k+1} - u\|_A^2 = (Az_{k+1},\ z_{k+1}) =$$
$$= (Az_k - \tau_{k+1}A^2 z_k,\ z_k - \tau_{k+1}Az_k) = (r_k - \tau_{k+1}Ar_k,$$
$$z_k - \tau_{k+1}r_k) = (r_k,\ z_k) - 2\tau_{k+1}(r_k,\ r_k) + \tau_{k+1}^2 (Ar_k,\ r_k),$$

дістаємо

$$\tau_{k+1} = \frac{(r_k,\ r_k)}{(Ar_k,\ r_k)}.$$

Для похибки маємо оцінку

$$\|z_{k+1}\|_A^2 = \|(I - \tau_{k+1}A)z_k\|_A^2 \leqslant \|(I - \tau_0 A)z_k\|_A^2 \leqslant$$
$$\leqslant \|I - \tau_0 A\|^2 \|z_k\|_A^2 \leqslant \rho_0^2 \|z_k\|_A,$$

тобто

$$\|z_{n+1}\| = \|y_{n+1} - u\|_A \leqslant \rho_0^{n+1} \|y_0 - u\|_A,$$

і МНС збігається з швидкістю методу простої ітерації. Більш швидкі методи можна знайти в класі багатоярусних схем. *Метод спряжених градієнтів* належить до класу триярусних схем виду

$$By_{k+1} = \alpha_{k+1}(B - \tau_{k+1}A)y_k + (1 - \alpha_{k+1})By_{k-1} + \alpha_{k+1}\tau_{k+1}f,$$
$$k = 1,\ 2,\ \ldots,$$
$$By_1 = (B - \tau_1 A)y_0 + \tau_1 f,\ \forall\ y_0 \in H,$$

де

$$\tau_{k+1} = \frac{(r_k,\ w_k)}{(Aw_k,\ w_k)}, \quad \alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k}\frac{(r_k,\ w_k)}{(r_{k-1},\ w_{k-1})} \cdot \frac{1}{\alpha_k}\right)^{-},$$

$A = A^* > 0$, $B = B^* > 0$, ${}_1B \leqslant A \leqslant \gamma_2 B$, $k = 0,\ 1,\ \ldots,$

$\gamma_1 > 0$. Формули для τ_{k+1}, α_{k+1} знаходимо за умови мінімуму норми $\|z_k\|_A$. При цих оптимальних значеннях ітераційних параметрів справедлива оцінка

$$\|y_n - u\|_A \leqslant q_n \|y_0 - u\|_A, \quad q_n = \frac{2\rho_1^n}{1 + \rho_1^{2n}},$$

$$\rho_1 = \frac{1 - \sqrt{\xi}}{\sqrt{\xi} + 1}, \quad \xi = \frac{\gamma_1}{\gamma_2},$$

тобто швидкість збіжності МСГ не менша, ніж у чебишевського двох'ярусного методу. Тут можна використати факторизований оператор ПТМ

$$B = (D + \omega A_1) D^{-1} (D + \omega A_2),$$
$$A_1 + A_2 = A > 0, \quad A_1^* = A_2, \quad D = D^* > 0.$$

(При такому виборі, як показують розрахунки, МСГ виконує менше ітерацій, ніж чебишевський двох'ярусний метод.)

1.5. Вкладення нормованих просторів та оцінки лінійних функціоналів у них

1.5.1. Теореми вкладення. Важливим інструментом для вивчення різних математичних об'єктів у нормованих просторах є так звані *теореми вкладення*.

Кажуть, що нормований простір X вкладений в нормований простір $\hat{X}$, якщо всюди на X задана лінійна функція (відображення, оператор) $J(x) : X \to \hat{X}$, причому існує стала $\beta > 0$ така, що

$$\| J(x) \|_{\hat{X}} \leqslant \beta \| x \|_X$$

для будь-якого $x \in X$. Якщо, зокрема, $\hat{X}$ та X визначаються введенням різних норм відповідно в лінійному просторі E та в його лінійному многовиді (підпросторі) D і за J вибирається відповідність, яка ототожнює елементи X та $\hat{X}$ як елементи E, то говорять про природне вкладення X в $\hat{X}$.

Приклад 1. Покажемо, що нормований простір $\tilde{W}_2^m[a, b]$ вкладений в нормований простір $C^{m-1}[a, b]$, $m \geqslant 1$. Оскільки $\tilde{W}_2^m[a, b] \subset C^{m-1}[a, b]$, то за J можна вибрати таку відповідність: кожній функції $u(t)$, що розглядається як елемент простору $\tilde{W}_2^m[a, b]$, відображення J ставить у відповідність ту саму функцію, але вона розглядається як елемент простору $C^{m-1}[a, b]$. Це означає, що йдеться про природне вкладення $\tilde{W}_2^m[a, b]$ в $C^{m-1}[a, b]$.

Внаслідок неперервності $u^{(i-1)}(t)$, $1 \leqslant i \leqslant m$, існує точка $\xi \in \bar{\Omega} = [a, b]$ така, що

$$u^{(i-1)}(\xi) = \frac{1}{b-a} \int_a^b u^{(i-1)}(s)\, ds, \quad i = \overline{1, m}.$$

Тому має місце тотожність

$$u^{(i-1)}(t) = \int_\xi^t u^{(i)}(s)\, ds + \frac{1}{b-a} \int_a^b u^{(i-1)}(s)\, ds, \quad i = \overline{1, m}. \tag{1}$$

Зауважимо, що для $u \in \tilde{W}_2^m[a, b]$ існування інтеграла $\int_\xi^t u^{(m)}(s)\, ds$ випливає з нерівності Коші — Буняковського

$$\left| \int_\xi^t u^{(m)}(s)\, ds \right| \leqslant \sqrt{t - \xi} \left(\int_\xi^t [u^{(m-1)}(s)]^2\, ds \right)^{1/2}.$$

За допомогою нерівності Коші — Буняковського з (1) дістаємо

$$|u^{(i-1)}(t)| \leqslant \int_\xi^t |u^{(i)}(s)|\, ds + \frac{1}{b-a} \int_a^b |u^{(i-1)}(s)|\, ds \leqslant$$
$$\leqslant \sqrt{b-a} \left(\int_a^b [u^{(i)}(s)]^2\, ds \right)^{1/2} + \frac{1}{\sqrt{b-a}} \left(\int_a^b [u^{(i-1)}(s)]^2\, ds \right)^{1/2}.$$

Застосовуючи елементарну нерівність $\sqrt{a} + \sqrt{b} < \sqrt{2}\sqrt{a+b}$, $a \geqslant 0$, $b \geqslant 0$, далі маємо

$$|u^{(i-1)}(t)| \leqslant \sqrt{2} K \left(|u^{(i)}|^2_{\tilde{W}_2^i(\bar{\Omega})} + |u|^2_{\tilde{W}_2^{i-1}(\bar{\Omega})} \right)^{1/2}, \quad i = \overline{1, m}, \tag{2}$$

де $K = \max\left(\sqrt{b-a}, \frac{1}{\sqrt{b-a}} \right)$, $|u|^2_{\tilde{W}_2^i(\bar{\Omega})} = \int_a^b [u^{(i)}(s)]^2\, ds$ — квадрат напівнорми в просторі $\tilde{W}_2^i[a, b]$. З нерівності Коші — Буняковського для сум із (2) маємо

$$\|u\|_{C^{m-1}[a,b]} = \max_{t \in [a,b]} \sum_{i=0}^{m-1} |u^{(i)}(t)| \leqslant \sqrt{2} K \sum_{i=0}^{m-1} \left(|u|^2_{\tilde{W}_2^{i+1}(\bar{\Omega})} + |u|^2_{\tilde{W}_2^i(\bar{\Omega})} \right)^{1/2} \leqslant$$
$$\leqslant \sqrt{2} K \sqrt{\sum_{i=0}^{m-1} 1^2} \sqrt{\sum_{i=0}^{m-1} \left(|u|^2_{\tilde{W}_2^{i+1}(\bar{\Omega})} + |u|^2_{\tilde{W}_2^i(\bar{\Omega})} \right)} =$$
$$= \sqrt{2m} K \sqrt{\sum_{i=1}^{m} |u|^2_{\tilde{W}_2^i(\bar{\Omega})} + \sum_{i=0}^{m-1} |u|^2_{\tilde{W}_2^i(\bar{\Omega})}}.$$

Оскільки

$$\|u\|_{\tilde{W}_2^m(\bar{\Omega})} = \left(\sum_{i=0}^{m} |u|^2_{\tilde{W}_2^i(\bar{\Omega})} \right)^{1/2},$$

то з останньої нерівності випливає

$$\|u\|_{C^{m-1}[a,b]} \leqslant M \|u\|_{\tilde{W}_2^m[a,b]}, \tag{3}$$

де

$$M = 2\sqrt{m} \max\left(\sqrt{b-a}, 1/\sqrt{b-a} \right). \tag{4}$$

Нехай тепер послідовність $\{u_n(x)\} \subset \tilde{W}_2^m[a, b]$ фундаментальна в нормі $\tilde{W}_2^m[a, b]$. Тоді з (3) матимемо

$$\|u_n - u_m\|_{C^{m-1}[a,b]} \leqslant M \|u_n - u_m\|_{W_2^m[a,b]} \to 0$$

при $m, n \to \infty$. Отже, послідовність $\{u_n(x)\}$ фундаментальна в розумінні рівномірної збіжності в $C^{m-1}[a, b]$ і за критерієм Коші вона збігається до $u(x) \in C^{m-1}[a, b]$. Тим більше $u_n(x) \to u(x)$ при $n \to \infty$ в середньому. Отже, в класі $L_2(a, b)$, що містить $\{u_n(x)\}$ як представника, є неперервна функція $u(x)$ і цей клас можна ототожнити з $u(x)$. Ототожнимо елементи простору $W_2^m(a, b)$ з $(m-1)$ разів неперервно диференційовними функціями. Нехай $\{u_n\} \to u(x)$. Переходячи в нерівності

$$\|u_n\|_{C^{m-1}[a,b]} \leqslant M \|u_n\|_{W_2^m[a,b]}$$

до границі при $n \to \infty$, дістаємо нерівність

$$\|u\|_{C^{m-1}[a,b]} \leqslant M \|u\|_{W_2^m(a,b)}.$$

Таким чином, доведено таку теорему вкладення.

Теорема 1. *Нормований простір $W_2^m(a, b)$, вкладений у простір $C^{m-1}[a, b]$, $m \geqslant 1$, причому для всіх $u \in W_2^m(a, b)$ має місце нерівність $\|u\|_{C^{m-1}} \leqslant M \|u\|_{W_2^m(a,b)}$, де стала M визначається формулою (4) і залежить від області $\bar{\Omega} = [a, b]$, m, але не від u.*

Це приклад однієї з найпростіших теорем вкладення. Перш ніж сформулювати загальнішу теорему вкладення для просторів Соболєва, введемо поняття *просторів Соболєва* $W_p^m(\Omega)$ з нецілим показником m, які називаються також *просторами Соболєва — Слободецького.*

Нехай $m > 0$ — не ціле число, $1 < p < +\infty$ і область $\Omega \subset R^n$. Число m можна представити у вигляді $m = \bar{m} + \lambda$, де $\bar{m}$ — ціле, $\lambda \in (0, 1)$. Простір Соболєва $W_p^m(\Omega)$ складається з функцій $u(x) \in W_p^{\bar{m}}(\Omega)$, які мають скінченну норму

$$\|u\|_{W_p^m(\Omega)} = \|u\|_{W_p^{\bar{m}}(\Omega)} + |u|_{W_p^m(\Omega)}, \tag{5}$$

де

$$|u|_{W_p^m(\Omega)} = \left(\sum_{|\alpha|=\bar{m}} \int\int_{\Omega\Omega} \frac{|D^\alpha u(x) - D^\alpha u(y)|^p}{|x-y|^{n+\lambda p}} dxdy \right)^{1/p} \tag{6}$$

— напівнорма простору $W_p^m(\Omega)$.

Приклад 2. Нехай $n = 1$, $\Omega = (0, 1)$. Розглянемо в Ω функцію

$$f(x) = \begin{cases} 1, & x < 0{,}5, \\ 0, & x \geqslant 0{,}5. \end{cases}$$

Оскільки

$$\int_0^1 \int_0^1 \frac{|f(x) - f(y)|^p}{|x-y|^{1+\lambda p}} dxdy =$$

$$= 2 \int_{0,5}^{1} \int_{0}^{0,5} \frac{dy}{(x-y)^{1+\lambda p}} dx \leqslant \frac{2}{\lambda p} \int_{0,5}^{1} (x - 0{,}5)^{-p\lambda} dx$$

і останній інтеграл збігається при $p\lambda < 1$, то $f(x) \in W_p^\lambda(\Omega)$ для $\lambda \in (0, 1/p)$, $p \in (1, \infty)$.

Не вдаючись до подробиць (з ними можна детально ознайомитись у спеціальній літературі), зауважимо, що можна означити слід функції $u(x) \in W_p^m(\Omega)$, $m > 0$, на границі Γ області Ω, для якого має місце таке твердження.

Теорема 2. *Нехай границя Γ області $\Omega \in R^n$ належить до класу C^m. Якщо $u \in W_p^m(\Omega)$, то її слід на границю $v \equiv u|\Gamma$ належить простору $W_p^{m-1/p}(\Gamma)$ і виконується оцінка*

$$\|v\|_{W_p^{m-1/p}(\Gamma)} \leqslant K_1 \|u\|_{W_p^m(\Omega)}. \tag{7}$$

Навпаки, якщо $v \in W_p^{m-1/p}(\Gamma)$, то існує функція $u \in W_p^m(\Omega)$ така, що $v = u|_\Gamma$ і виконується оцінка

$$\|u\|_{W_p^m(\Omega)} \leqslant K_2 \|v\|_{W_p^{m-1/p}(\Gamma)}, \tag{8}$$

де сталі K_1, K_2 не залежать від u, v.

Можна також означити слід функції $u(x) \in W_p^m(\Omega)$ на K-вимірну область Ω^k, яка визначається перетином Ω з k-вимірною гіперплощиною в $\mathbb{R}^k$, $k = \overline{1, n}$ $(\Omega^n \equiv \Omega)$. В наступній теоремі вкладення Соболєва формулюється загальний випадок вкладення класів функцій різних просторів і різних вимірів.

Теорема 3. *Нехай Ω — відкрита область в R^n з неперервною за Ліпшицем границею і нехай Ω^k — k-вимірна область, означена вище. Нехай далі $m \geqslant 0$ — дійсне число, $p \in [1, \infty]$. Тоді мають місце такі вкладення:*

1) *якщо $mp < n$ і $n - mp < k \leqslant n$, то*

$$W_p^m \subset L_q(\Omega^k), \quad p \leqslant q \leqslant kp/(n - mp),$$

зокрема, для $k = n$ має місце оцінка

$$\|u\|_{L_q(\Omega)} \leqslant C \|u\|_{W_p^m(\Omega)}, \quad p \leqslant q \leqslant np/(n - mp);$$

2) *якщо $mp = n$, то для довільного k, $k \in [1, n]$,*

$$W_p^m(\Omega) \subset L_q(\Omega^k), \quad p \leqslant q < \infty, \tag{9}$$

зокрема, для $k = n$ має місце оцінка

$$\|u\|_{L_q(\Omega)} \leqslant C \|u\|_{W_p^m(\Omega)}, \quad p \leqslant q < \infty.$$

Якщо $p = 1$ і $m = n$, *то вкладення* (9) *має місце для* $q = \infty$, *тобто справедлива оцінка*

$$\|u\|_{C(\bar{\Omega})} \leqslant C \|u\|_{W_1^m(\Omega)};$$

3) *якщо* $mp > n$, *то*

$$W_p^m(\Omega) \subset L_\infty(\Omega);$$

коли $mp > n > (m-1)\,p$, *тоді*

$$W_p^m(\Omega) \subset C^{0,\lambda}(\bar{\Omega}), \quad 0 < \lambda \leqslant m - \frac{n}{p};$$

якщо $n = (m-1)\,p$, *то*

$$W_p^m(\Omega) \subset C^{0,\lambda}(\bar{\Omega}), \quad 0 < \lambda < 1$$

(тут через C *позначено різні сталі, які не залежать від* u, $L_\infty(\Omega)$ — *простір класів функцій з нормою* $\|u\|_{L_\infty(\Omega)} = \operatorname{vrai}\max_{x\in\Omega} |f(x)|$; $C^{m,\lambda}(\bar{\Omega})$ — *простір функцій, що мають неперервні в* $\bar{\Omega}$ *похідні до порядку* m *включно, які задовольняють умову Гьольдера з показником* λ, *тобто функцій, для яких*

$$|f|_{C^{m,\lambda}(\bar{\Omega})} \equiv \max_{|\alpha|=m} \sup_{x,y\in\bar{\Omega}} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{|x-y|^\lambda} < +\infty,$$

де $|x-y|$ — *нижня грань довжин кривих, які сполучають точки* $x, y \in \bar{\Omega}$; *норма в* $C^{m,\lambda}(\bar{\Omega})$ *вводиться за формулою*

$$\|f\|_{C^m(\bar{\Omega})} = \max_{|\alpha|\leqslant m} \max_{x\in\bar{\Omega}} |D^\alpha f(x)| \equiv \max_{0\leqslant k\leqslant m} |f|_{C^k(\bar{\Omega})};$$

$$\|f\|_{C^{m,\lambda}(\bar{\Omega})} = \|f\|_{C^m(\bar{\Omega})} + |f|_{C^{m,\lambda}(\bar{\Omega})},$$

вкладення $W_p^m(\Omega) \subset C(\bar{\Omega})$ *означає, що будь-який клас* $u \in W_p^m(\Omega)$ *містить стаціонарну послідовність* $\{\bar{u}(x), \bar{u}(x), \ldots\}$, *де* $\bar{u}(x) \in C(\bar{\Omega})$).

Як бачимо з теореми 3, властивості вкладення різних просторів залежать не лише від параметрів цих просторів, але і від розмірності простору незалежних змінних R^n. З теореми 1, наприклад, випливає, що при $n = 1$ $W_2^1(\Omega) \subset C(\bar{\Omega})$, але вже при $n = 2$ це вкладення не має місця. З теореми 3 бачимо, що у випадку $n = 2$ має місце вкладення $W_2^2(\Omega) \subset C^{0,\lambda}$, $0 < \lambda < 1$. Зокрема, $\tilde{W}_2^2(\bar{\Omega}) \subset C^{0,\lambda}$, де $\lambda \in (0, 1)$, Ω — двовимірна обмежена область. Це вкладення ми будемо використовувати далі. Вкладення $W_2^m(\Omega) \subset C(\bar{\Omega})$ виконується для $n = 2$ при $m > 1$.

Далі не заглиблюватимемося у теорію просторів Соболєва. Зауважимо лише, що в одновимірному випадку, який далі найчастіше зустрічатиметься, елементи класу $W_k^2(a, b)$, $k \geqslant 1$, можна розглядати як

звичайні $k-1$ разів неперервно диференційовні функції, k-та похідна яких інтегрована на (a, b) з квадратом (можливо, інтеграл розуміємо як невласний).

Розглянемо простір сіткових функцій, заданих на сітці із скінченною кількістю вузлів. Очевидно, цей простір є скінченновимірним. Відомо, що в будь-якому скінченновимірному просторі всі норми еквівалентні. Тому за умови даного вище означення, всі нормовані простори, дістані з деякого скінченновимірного простору сіткових функцій введенням в ньому різних норм, природно, вкладені один в одного. У цьому разі важливо мати оцінку для константи β в нерівності вкладення (див. означення вкладення)

$$\|J(x)\|_{\hat{X}} \leqslant \beta \|x\|_X,$$

причому особливу роль грають вкладення, де стала β не залежить від розмірності простору сіткових функцій.

Нехай, наприклад, $W_2^1(\bar{\omega}_n)$ — простір сіткових функцій $y(i)$, заданих на сітці $\bar{\omega}_n = \{x_i = i : i = 0, \overline{n+1}\}$, в якому норма визначається формулою

$$\|y\|^2_{W_2^1(\bar{\omega}_n)} = \|y\|^2_{L_2(\bar{\omega}_n)} + |y|^2_{W_2^1(\omega_n^+)}, \tag{10}$$

де

$$\|y\|^2_{L_2(\bar{\omega}_n)} \equiv \sum_{i=0}^{n+1} y^2(i), \quad |y|^2_{W_2^1(\omega_n^+)} \equiv \sum_{i=1}^{n+1} [\nabla y(i)]^2 = \sum_{i=1}^{n+1} [y(i) - y(i-1)]^2, \tag{11}$$

$$\omega_n^+ = \bar{\omega}_n \setminus \{0\}.$$

Простір тих самих функцій з нормою $\|\cdot\|_{L_2(\bar{\omega}_n)}$ позначатимемо $L_2(\bar{\omega}_n)$. Неважко помітити, що в просторі сіткових функцій, заданих на сітці $\bar{\omega}_n$, величина $|\cdot|_{W_2^1(\omega_n^+)}$ є напівнормою. У просторі сіткових функцій, заданих на $\bar{\omega}_n$, які перетворюються на нуль, хоча б в одній із граничних точок 0 чи $n+1$ норма $\|\cdot\|_{W_2^1(\bar{\omega}_n)}$ та напівнорма $|\cdot|_{W_2^1(\omega_n^+)}$ еквівалентні із сталими еквівалентності 1 та $\sqrt{1 + (n+1)^2}$.

Дійсно, якщо, наприклад, $y(0) = 0$, то

$$y(i) = \sum_{j=0}^{i-1} [y(j+1) - y(j)],$$

а тому

$$y^2(i) = \left(\sum_{j=0}^{i-1} [y(j+1) - y(j)]\right)^2 \leqslant i \sum_{j=0}^{i-1} [y(j+1) - y(j)]^2$$

і далі маємо

$$\|y\|^2_{L_2(\bar{\omega}_n)} = \sum_{i=0}^{n+1} y^2(i) \geqslant \sum_{i=1}^{n+1} i \sum_{j=0}^{i-1} [y(j+1) - y(j)]^2 \leqslant$$

$$\leqslant (n+1)^2 \sum_{j=1}^{n+1} [y(j+1) - y(j)]^2 = (n+1)^2 |y|^2_{W_2^1(\omega_n^+)}. \qquad (12)$$

Отже,

$$|y|_{W_2^1(\omega_n^+)} \leqslant \|y\|_{W_2^1(\bar{\omega}_n)} \leqslant \sqrt{1+(n+1)^2}\, |y|_{W_2^1(\omega_n^+)}, \qquad (13)$$

що й треба було довести.

Простір сіткових функцій, заданих на сітці $\bar{\omega}_n$ і таких, що $y(0) = y(n+1) = 0$, з нормою (10) позначатимемо $\mathring{W}_2^1(\bar{\omega}_n)$.

Розглянемо також нормований простір $W_2^1(\bar{\omega}_h)$ сіткових функцій $y(x)$, заданих на сітці $\bar{\omega}_h = \{x_i = ih : i = \overline{0, N},\ h = 1/N\}$, з нормою

$$\|y\|^2_{W_2^1(\bar{\omega}_h)} = \|y\|^2_{L_2(\bar{\omega}_h)} + |y|^2_{W_2^1(\omega_h^+)}, \qquad (14)$$

де

$$\|y\|^2_{L_2(\bar{\omega}_h)} = \sum_{i=0}^{N} h y^2(x_i) = \sum_{x \in \bar{\omega}_h} h y^2(x),$$

$$|y|^2_{W_2^1(\omega_h^+)} = \sum_{i=1}^{N} h y^2_{\bar{x},i} = \sum_{x \in \omega_h^+} h y^2_{\bar{x}}(x),$$

$$\omega_h^+ = \bar{\omega}_h \setminus \{0\}.$$

Простір тих самих функцій з нормою $\|\cdot\|_{L_2(\bar{\omega}_h)}$ позначимо $L_2(\bar{\omega}_h)$, а підпростір простору $W_2^1(\bar{\omega}_h)$, який складається з функцій, що обертаються в нуль при $x = 0,1$, позначимо $\mathring{W}_2^1(\bar{\omega}_h)$.

Нехай $C(\bar{\omega}_h)$ — простір сіткових функцій, заданих на $\bar{\omega}_h$, з нормою $\|y\|_{C(\bar{\omega}_h)} = \max\limits_{x \in \bar{\omega}_h} |y(x)|$, а $C(\bar{\omega}_n)$ — простір функцій, заданих на $\bar{\omega}_n$, з нормою $\|y\|_{C(\bar{\omega}_n)} = \max\limits_{i=\overline{0,n+1}} |y(i)|$. Через $\mathring{C}(\bar{\omega}_h) \equiv C(\omega_h)$ позначимо підпростір функцій з $C(\bar{\omega}_h)$, які обертаються в нуль при $x = 0,1$. Аналогічно введемо підпростір $\mathring{C}(\bar{\omega}_n) \equiv C(\omega_n)$.

Кожну функцію $y(x)$, задану на сітці $\bar{\omega}_h$, можна розглядати і як функцію $y(x_i) \equiv y(i)$, $x_i \in \bar{\omega}_h$, $i = \overline{0, N}$, задану на сітці $\bar{\omega}_N = \{i : i = \overline{0, N}\}$.

Оскільки

$$|y|_{W_2^1(\omega_N^+)} = \left(\sum_{i=1}^{N} (y(i) - y(i-1))^2\right)^{1/2} =$$

$$= \left(\sum_{x \in \omega_h^+} h y^2_{\bar{x}}(x)\right)^{1/2} \sqrt{h} \equiv \sqrt{h}\, |y|_{W_2^1(\omega_h^+)},$$

$$\|y\|_{L_2(\bar{\omega}_N)} = \left(\sum_{i=0}^{N} y^2(i)\right)^{1/2} = \left(\sum_{x \in \bar{\omega}_h} h y^2(x)\right)^{1/2} h^{-1/2} \equiv h^{-1/2} \|y\|_{L_2(\bar{\omega}_h)},$$

$$\omega_h^+ = \bar{\omega}_h \setminus \{0\},$$

то з (12) випливає (для функцій з $\mathring{W}_2^1(\bar{\omega}_h)$)

$$\|y\|_{L_2(\bar{\omega}_h)} \equiv \|y\|_{L_2(\omega_h)} = h^{1/2} \|y\|_{L_2(\bar{\omega}_N)} \leqslant$$

$$\leqslant h^{1/2} h^{-1} |y|^2_{W_2^1(\omega_N^+)} = |y|_{W_2^1(\omega_h^+)}.$$

Тому аналог нерівності (13) для функцій з $\mathring{W}_2^1(\omega_h)$ набирає вигляду

$$|y|_{W_2^1(\omega_h^+)} \leqslant \|y\|_{W_2^1(\bar{\omega}_h)} \leqslant \sqrt{2}\, |y|_{W_2^1(\omega_h^+)}, \qquad (15)$$

тобто в просторі $\mathring{W}_2^1(\omega_h)$ норма $\|\cdot\|_{W_2^1(\bar{\omega}_h)}$ та напівнорма $|\cdot|_{W_2^1(\omega_h^+)}$ еквівалентні, причому сталі еквівалентності не залежать від h.

Простори $L_2(\bar{\omega}_h)$, $L_2(\bar{\omega}_n)$ є сітковими аналогами простору $\tilde{L}_2(\bar{\Omega})$ неперервних функцій, заданих на області $\bar{\Omega}$ з нормою

$$\|f\|_{\tilde{L}_2(\bar{\Omega})} = \left(\int_{\bar{\Omega}} f^2(x)\, dx\right)^{1/2}.$$

Простори $W_2^1(\bar{\omega}_h)$ та $W_2^1(\bar{\omega}_n)$ є сітковими аналогами простору $\tilde{W}_2^1(\bar{\Omega})$ неперервно диференційовних функцій з нормою

$$\|f\|_{\tilde{W}_2^1(\bar{\Omega})} = \left(\int_{\bar{\Omega}} [u^2(x) + (u'(x))^2]\, dx\right)^{1/2}.$$

Зауважимо, що в силу скінченновимірності простори $L_2(\bar{\omega}_n)$, $L_2(\bar{\omega}_h)$, $W_2^1(\bar{\omega}_n)$, $W_2^1(\bar{\omega}_h)$ є банаховими.

З очевидних нерівностей

$$\|y\|_{L_2(\bar{\omega}_n)} \leqslant \sqrt{n+2}\, \|y\|_{C(\bar{\omega}_n)},$$

$$\|y\|_{L_2(\bar{\omega}_h)} \leqslant \|y\|_{C(\bar{\omega}_h)}$$

випливає, що простори $C(\bar{\omega}_n)$ та $C(\bar{\omega}_h)$ вкладені відповідно в простори $L_2(\bar{\omega}_n)$ та $L_2(\bar{\omega}_h)$.

Покажемо, що простір $\overset{\circ}{W}{}_2^1(\bar{\omega}_n)$ вкладений в $\overset{\circ}{C}(\omega_n)$, а простір $\overset{\circ}{W}{}_2^1(\bar{\omega}_h)$ — в простір $\overset{\circ}{C}(\bar{\omega}_h)$. Відображення J (див. означення вкладення) виберемо природним чином, тобто воно кожній сітковій функції $y \in \overset{\circ}{W}{}_2^1(\bar{\omega}_n)$ $(y \in \overset{\circ}{W}{}_2^1(\bar{\omega}_h))$ ставить у відповідність ту саму сіткову функцію y, але як елемент простору $\overset{\circ}{C}(\bar{\omega}_n)$ $(\overset{\circ}{C}(\bar{\omega}_h))$. Нерівності вкладення даються наступними двома лемами.

Лема 1. *Для будь-якої сіткової функції y, заданої на сітці $\bar{\omega}_n$ і такої, що перетворюється на нуль при $i = 0$, $i = n + 1$, має місце нерівність*

$$\|y\|_{\overset{\circ}{C}(\bar{\omega}_n)} \leqslant \frac{\sqrt{n+1}}{2} |y|_{W_2^1(\omega_n^+)}. \tag{16}$$

Доведення. Запишемо тотожність

$$(n+1)\, y^2(x) = (n+1-x)\, y^2(x) + x y^2(x).$$

Оскільки $y(0) = y(n+1) = 0$, то

$$y^2(x) = \left(\sum_{x'=1}^{x} \nabla y(x')\right)^2, \quad y^2(x) = \left(\sum_{x'=x+1}^{n+1} \nabla y(x')\right)^2,$$

де $\nabla y(x') = y(x') - y(x'-1)$. З попередньої тотожності маємо

$$(n+1)\, y^2(x) = (n+1-x)\left(\sum_{x'=1}^{x} \nabla y(x')\right)^2 + x\left(\sum_{x'=x+1}^{n+1} \nabla y(x')\right)^2.$$

Звідси за допомогою нерівності Коші — Буняковського маємо

$$(n+1)\, y^2(x) \leqslant (n+1-x)\, x \sum_{x'=1}^{x} [\nabla y(x')]^2 +$$

$$+ x(n+1-x) \sum_{x'=x+1}^{n+1} [\nabla y(x')]^2 =$$

$$= x(n+1-x) \sum_{x'=1}^{n+1} [\nabla y(x')]^2 \equiv x(n+1-x)\, |y|^2_{W_2^1(\omega_n^+)}.$$

Максимального значення вираз $x(n+1-x)$ досягає при $x = \frac{n+1}{2}$, а тому

$$(n+1)\, y^2(x) \leqslant \left(\frac{n+1}{2}\right)^2 |y|^2_{W_2^1(\omega_n^+)},$$

звідки і випливає твердження леми.

Оскільки

$$\|y\|_{C(\bar{\omega}_h)} \equiv \|y(i)\|_{C(\bar{\omega}_N)}, \quad N = h^{-1},$$

то з леми 1 випливає таке твердження.

Лема 2. *Для будь-якої функції $y(x)$, заданої на сітці $\omega_h = \{x_i = ih : i = \overline{0, N},\ h = 1/N\}$ і такої, що перетворюється на нуль в точках $x_0 = 0$, $x_N = 1$, справедлива нерівність*

$$\|y\|_{\overset{\circ}{C}(\bar{\omega}_h)} \leqslant \frac{1}{2} |y|_{W_2^1(\omega_h^+)}, \tag{17}$$

де

$$|y|_{W_2^1(\omega_h^+)} = \left(\sum_{x \in \omega_h^+} h y_{\bar{x}}^2(x)\right)^{1/2}.$$

В силу нерівностей $|y|_{W_2^1(\omega_h^+)} \leqslant \|y\|_{W_2^1(\bar{\omega}_h)}$, $|y|_{W_2^1(\omega_n^+)} \leqslant \|y\|_{W_2^1(\bar{\omega}_n)}$ з леми 2 дістаємо, що простір $\overset{\circ}{W}{}_2^1(\bar{\omega}_h)$ вкладений у простір $\overset{\circ}{C}(\bar{\omega}_h)$, причому в нерівності вкладення стала не залежить від h. Лема 1 дає величину константи в нерівності вкладення простору $\overset{\circ}{W}{}_2^1(\bar{\omega}_n)$ в $\overset{\circ}{C}{}_2^1(\bar{\omega}_n)$.

Зауваження. Нерівність вигляду (17) неважко дістати і для сіткових функцій, заданих на сітці $\bar{\omega}_{h'} = \{x_i = a + ih' : i = \overline{0, N},\ h' = (b-a)/N\}$, яка покриває довільний відрізок $[a, b]$. Для цього треба виконати заміну змінних $x' = (b-a)\, x + a$. Тоді x' змінюватиметься на відрізку $[a, b]$, а $x \in [0, 1]$, причому

$$h' = (b-a)\, h, \quad h = 1/N, \quad y_{\bar{x}'} = y_{\bar{x}}/(b-a).$$

Підставивши $y_{\bar{x}} = (b-a)\, y_{\bar{x}'}$, $h = h'/(b-a)$ у вираз для $|y|_{W_2^1(\omega_h^+)}$, дістанемо $(y \in \overset{\circ}{W}{}_2^1(\bar{\omega}_h))$

$$|y|^2_{W_2^1(\omega_h^+)} = \sum_{x \in \omega_h^+} y_{\bar{x}}^2(x)\, h = \sum_{x' \in \omega_{h'}^+} (b-a)^2\, y_{\bar{x}'}^2 (b-a)^{-1} h' =$$

$$= (b-a)\, |y|^2_{W_2^1(\omega_{h'}^+)}.$$

Таким чином, для сітки $\bar{\omega}_{h'}$, яка покриває відрізок $[a, b]$, замість (17) має місце нерівність

$$\|y\|_{\overset{\circ}{C}(\bar{\omega}_{h'})} \leqslant \frac{\sqrt{b-a}}{2} |y|_{W_2^1(\omega_{h'}^+)}. \tag{18}$$

У підпросторі сіткових функцій, які перетворюються на нуль на границі, вираз $|y|_{W_2^1(\omega_{h'}^+)}$ також є нормою, яка еквівалентна нормі $\|y\|_{W_2^1(\bar{\omega}_{h'})}$.

Покажемо, що має місце таке твердження.

Лема 3. *Для будь-якої функції $y(x)$, заданої на рівномірній сітці $\bar{\omega}_{h'} = \{x_i = a + ih' : i = \overline{0, N},\ h' (b-a)/N\}$, яка обертається на нуль при $x = a$, $x = b$ справедливі оцінки*

$$\frac{h'^2}{4} |y]_{W_2^1(\omega_{h'}^+)} \leqslant \|y\|_{L_2(\bar{\omega}_{h'})} \equiv \|y\|_{L_2(\omega_{h'})} \leqslant \frac{(b-a)^2}{8} |y]_{W_2^1(\omega_{h'}^+)}. \quad (19)$$

Доведення. Розкладемо $y(x)$ по власних функціях оператора $\Lambda y = y_{\bar{x}x}$:

$$y(x) = \sum_{k=1}^{N-1} C_k \mu^{(k)}(x), \quad C_k = (y, \mu^{(k)}) = \sum_{x \in \omega_{h'}} h' y(x) \mu^{(k)}(x),$$

$$\|y\|^2_{L_2(\omega_{h'})} = \|y\|^2 = \sqrt{(y, y)} = \sum_{k=1}^{N-1} C_k^2.$$

З першої формули Гріна дістанемо

$$(-\Lambda y, y) = |y]^2_{W_2^1(\omega_{h'}^+)} = \left(-\Lambda \sum_{k=1}^{N-1} C_k \mu^{(k)}(x), y(x)\right) =$$

$$= \left(\sum_{k=1}^{N-1} C_k \lambda_k \mu^{(k)}(x), \sum_{k=1}^{N-1} C_k \mu^{(k)}(x)\right) = \sum_{k=1}^{N-1} \lambda_k C_k^2.$$

(ми врахували ортонормованість власних функцій $\mu^{(k)}(x)$). Звідси

$$\lambda_1 \|y\|_{L_2(\omega_{h'})} \leqslant |y]_{W_2^1(\omega_{h'}^+)} \leqslant \lambda_{N-1} \|y\|_{L_2(\omega_{h'})},$$

де

$$\lambda_1 = \frac{4}{(h')^2} \sin^2 \frac{\pi h'}{2(b-a)},$$

$$\lambda_{N-1} = \frac{4}{(h')^2} \sin^2 \frac{\pi h' (N-1)}{2(b-a)} = \frac{4}{(h')^2} \cos^2 \frac{\pi h'}{2(b-a)}.$$

Позначивши $\alpha = \pi h'/(2(b-a))$, дістанемо $\lambda_1 = \frac{\pi^2}{(b-a)^2} \left(\frac{\sin \alpha}{\alpha}\right)^2$. Оскільки $h' \leqslant 0{,}5(b-a)$, то α змінюється на інтервалі $\left(0, \frac{\pi}{4}\right]$. Неважко перевірити, що мінімум функції $\sin \alpha/\alpha$ на цьому інтервалі досягається в точці $\alpha = \pi/4$, тобто $\lambda_1(h')$ має мінімум при $h' = 0{,}5(b-a)$. Звідси випливає, що $\lambda_1 \geqslant 8/(b-a)^2$. Враховуючи також, що $\lambda_{N-1} \leqslant 4/(h')^2$, дістаємо твердження леми.

Оскільки

$$\|y\|^2_{W_2^1(\omega_{h'})} = \|y\|_{L_2(\omega_{h'})} + |y]_{W_2^1(\omega_{h'})},$$

то з леми 3 для функції з $\mathring{W}_2^1(\bar{\omega}_{h'})$ випливає еквівалентність норми $\|y\|_{W_2^1(\omega_{h'})}$ та $|y]_{W_2^1(\omega_{h'}^+)}$. Якщо функція обертається на нуль лише в одній з граничних точок, то справджується нерівність

$$\|y\|_{C(\bar{\omega}_{h'})} \leqslant \sqrt{b-a}\, |y]_{W_2^1(\omega_{h'}^+)}, \quad (20)$$

яку можна довести аналогічно до леми 1, використавши представлення

$$y^2(x) = \left(\sum_{x'=h'+a}^{x} h' y_{\bar{x}}(x')\right)^2, \quad y^2(x) = \left(\sum_{x'=x+h'}^{b} h' y_{\bar{x}}(x')\right)^2.$$

Очевидно, що для $y(x) \in \mathring{W}_2^1(\bar{\omega}_{h'})$,

$$\|y\|_{L_2(\bar{\omega}_{h'})} \leqslant \sqrt{b-a}\, \|y\|_{C(\bar{\omega}_{h'})} \leqslant \frac{b-a}{2} |y]_{W_2^1(\omega_{h'}^+)},$$

а тому

$$|y]_{W_2^1(\omega_{h'}^+)} \leqslant \|y\|_{W_2^1(\bar{\omega}_{h'})} \leqslant [1 + (b-a)^2/4]^{1/2} |y]_{W_2^1(\omega_{h'}^+)}, \quad (21)$$

що означає еквівалентність норми $\|y\|_{W_2^1(\bar{\omega}_{h'})}$ та напівнорми $|y]_{W_2^1(\omega_{h'}^+)}$ і для функцій, які обертаються на нуль на одному з кінців сіткової області.

1.5.2. Формула Тейлора. Нехай $f(x) \in \mathring{W}_2^m(\bar{\Omega})$, $\bar{\Omega} = [a, b]$. Інтегруючи послідовно частинами, дістаємо рівність

$$\frac{1}{(m-1)!} \int_a^x (x-t)^{m-1} f^{(m)}(t)\, dt = \frac{(x-t)^{m-1}}{(m-1)!} f^{(m-1)}(t) \Big|_a^x +$$

$$+ \frac{1}{(m-2)!} \int_a^x (x-t)^{m-2} f^{(m-1)}(t)\, dt = -\frac{(x-a)^{m-1}}{(m-1)!} f^{(m-1)}(a) +$$

$$+ \frac{(x-t)^{m-2}}{(m-2)!} f^{(m-2)}(t) \Big|_a^x + \frac{1}{(m-3)!} \int_a^x (x-t)^{m-3} f^{(m-2)}(t)\, dt = \ldots =$$

$$= -\frac{(x-a)^{m-1}}{(m-1)!} f^{(m-1)}(a) -$$

$$- \frac{(x-a)^{m-2}}{(m-2)!} f^{(m-2)}(a) - \ldots - f(a) + f(x).$$

Звідси знаходимо, що для будь-якої функції $f(x) \in \mathring{W}_2^m[a, b]$ справедлива формула Тейлора

$$f(x) = f(a) + \frac{x-a}{1} f'(a) + \ldots + \frac{(x-a)^{m-1}}{(m-1)!} f^{(m-1)}(a) + R_m(x) \quad (22)$$

із залишковим членом в інтегральній формі

$$R_m(x) \equiv R_m(x;\ f) = \frac{1}{(m-1)!}\int_a^x (x-t)^{m-1} f^{(m)}(t)\,dt. \qquad (23)$$

Якщо ввести функцію

$$K_m(u) = \begin{cases} u^{m-1}, & u \geqslant 0, \\ 0, & u < 0, \end{cases}$$

то вираз для залишкового члена можна записати ще у вигляді

$$R_m(x) = \frac{1}{(m-1)!}\int_a^b K_m(x-t) f^{(m)}(t)\,dt, \qquad (24)$$

бо $K_m(x-t)$ при будь-якому фіксованому x дорівнює нулю, якщо t змінюється від x до b.

Виконавши в (23) заміну $\frac{x-t}{x-a} = s$, $t = x + (a-x)s$, $ds = -\frac{dt}{x-a}$, знайдемо ще одну форму запису залишкового члена в інтегральній формі

$$R_m(x) = \frac{(x-a)^m}{(m-1)!}\int_a^x \left(\frac{x-t}{x-a}\right)^{m-1} f^{(m)}(t)\,\frac{dt}{x-a} =$$

$$= -\frac{(x-a)^m}{(m-1)!}\int_1^0 s^{m-1} f^{(m)}(x + s(a-x))\,ds =$$

$$= \frac{(x-a)^m}{(m-1)!}\int_0^1 s^{m-1} f^{(m)}(x + s(a-x))\,ds. \qquad (25)$$

Користуючись теоремою про середнє, з (25) дістаємо таке зображення залишкового члена в формі Лагранжа:

$$R_m(x;\ f) = \frac{(x-a)^m}{(m-1)!} f^{(m)}(x + \theta(a-x))\int_0^1 s^{m-1} ds =$$

$$= \frac{(x-a)^m}{m!} f^{(m)}(x + \theta(a-x)), \quad \theta \in (0, 1).$$

Виконаємо в (25) заміну $s = 1 - t$:

$$R_m(x;\ f) = \frac{(x-a)^m}{(m-1)!}\int_0^1 (1-t)^{m-1} f^{(m)}(a + t(x-a))\,dt.$$

Користуючись знову теоремою про середнє, знайдемо звідси вираз для залишкового члена в формі Коші:

$$R_m(x;\ f) = \frac{(x-a)^m (1-\theta)^{m-1}}{(m-1)!} f^{(m)}(a + \theta(x-a)), \quad \theta \in (0, 1).$$

1.5.3. Лема Брембла — Гільберта. Теорема Соболева про еквівалентне нормування. Нехай $\Omega = (0, 1)$, $u \in \tilde{W}_2^m(\bar{\Omega})$. Запишемо формулу Тейлора у вигляді

$$u(x) = Q_m u(x) + R_m u(x),$$

де

$$Q_m(u(x)) = \sum_{k=0}^{m-1} \frac{(x-y)^k}{k!} D^k u(y), \quad D^k \equiv D_x^k = \frac{d^k}{dx^k},$$

$$R_m u(x) = m\,\frac{(x-y)^m}{m!}\int_0^1 s^{m-1} D^m u(x + s(y-x))\,ds.$$

Лема Брембла — Гільберта формулюється таким чином.

Лема 4. *Нехай лінійний функціонал $l(u)$ обмежений в $\tilde{W}_2^m(\bar{\Omega})$, тобто $|l(u)| \leqslant M \|u\|_{\tilde{W}_2^m(\bar{\Omega})}$, і перетворюється на нуль на множині π_{m-1} многочленів степеня не вище $m-1$, тобто $l(p) = 0\ \forall\, p \in \pi_{m-1}$. Тоді існує стала $\overline{M} = \overline{M}(m, \Omega)$ така, що*

$$|l(u)| \leqslant M\overline{M}\,|u|_{\tilde{W}_2^m(\bar{\Omega})},$$

$$\overline{M} = \left(1 + \sum_{j=0}^{m-1} \frac{1}{(2m-2j-1)\,[(m-j-1)!]^2}\right)^{1/2}.$$

Доведення. В силу умов леми

$$|l(u)| = |l(u - Q_m u)| = |l(R_m u)| \leqslant M \|R_m u\|_{\tilde{W}_2^m(\bar{\Omega})}. \qquad (26)$$

Неважко помітити, що для $j \leqslant m-1$

$$D_x^j(R_m u(x)) = D_x^j(u(x) - Q_m(x)) = D_x^j u(x) - \sum_{k=j}^{m-1} \frac{(x-y)^{k-j}}{(k-j)!} D^k u(y) =$$

$$= D_x^j u(x) - \sum_{l=0}^{m-j-1} \frac{(x-y)^l}{l!} D^l D^j u(y) = R_{m-j} D_x^j u(x)$$

і

$$D_x^m(R_m u(x)) = D_x^m(u(x) - Q_m(x)) = D_x^m u(x).$$

Тому

$$\|R_m u\|_{\tilde{W}_2^m(\bar{\Omega})} = \left(\sum_{j=0}^{m} \|D^j R_m u\|^2_{\tilde{L}_2(\bar{\Omega})}\right)^{1/2} =$$

$$= \left(\sum_{j=0}^{m-1} \|R_{m-j} D^j u\|^2_{\tilde{L}_2(\bar{\Omega})} + \|D^m u\|^2_{\tilde{L}_2(\bar{\Omega})}\right)^{1/2}. \qquad (27)$$

Але для $j \leqslant m-1$ маємо ($y \in (0, 1)$)

$$\|R_{m-j} D^j u\|_{\tilde{L}_2(\bar{\Omega})} = \left\{\int_0^1 \left[\frac{(x-y)^{m-j}}{(m-j)!}(m-j)\int_0^1 s^{m-j-1} D^m u\,(x + s(y-x))\,ds\right]^2 dx\right\}^{1/2} \leqslant$$

$$\leqslant \left\{\int_0^1 \left[\frac{(x-y)^{m-j}}{(m-j-1)!}\right]^2 \int_0^1 s^{2(m-j-1)} ds \times \int_0^1 [D^m u\,(x + s(y-x))]^2\, ds dx\right\}^{1/2} =$$

$$= \left\{\int_0^1 \left[\frac{(x-y)^{m-j}}{(m-j-1)!}\right]^2 \frac{1}{2m-2j-1}\left(y-x\right)\int_x^y [D^m u\,(t)]^2\, dt dx\right\}^{1/2} \leqslant$$

$$\leqslant \frac{1}{\sqrt{2m-2j-1}\,(m-j-1)!}\,\|D^m u\|_{\tilde{L}_2(\bar{\Omega})}.$$

Таким чином,

$$\|R_m u\|_{\tilde{W}_2^m(\bar{\Omega})} \leqslant \left(1 + \sum_{j=0}^{m-1} \frac{1}{(2m-2j-1)\,[(m-j-1)!]^2}\right)^{1/2} \|D^m u\|_{L_2(\bar{\Omega})},$$

що разом з (26), (27) і доводить лему. Зрозуміло, що ця лема справедлива і для просторів Соболєва $W_2^n(a, b)$.

У багатовимірному випадку лема Брембла — Гільберта формулюється таким чином.

Лема 5. *Нехай Ω — відкрита обмежена множина в R^n з неперервною за Ліпшицем границею і лінійний функціонал $l(u)$ має властивості: а) $l(u)$ обмежений в $W_p^m(\Omega)$ для деякого цілого числа m і деякого $p \in [1, \infty)$, тобто $|l(u)| \leqslant M\|u\|_{W_p^m(\Omega)}$; б) $l(u)$ перетворюється в нуль на многочленах степеня $m-1$ по змінних $x_1, \ldots, x_n$, тобто $l(p) = 0$, якщо $p(x) = \sum_{|\alpha|<m} a_\alpha x^\alpha$, $\alpha = (\alpha_1, \ldots, \alpha_n)$, $x^\alpha = x_1^{\alpha_1}, \ldots, x_n^{\alpha_n}$. Тоді існує стала $\bar{M} = \bar{M}(\Omega)$ така, що*

$$|l(u)| \leqslant M\bar{M}\,|u|_{W_p^m(\Omega)},$$

де $|\cdot|_{W_p^m(\Omega)}$ — напівнорма в просторі $W_p^m(\Omega)$.

Зауважимо, що аналогічна лема має місце і для нецілих m. Іноді буває зручно замість норми

$$\|u\|_{W_2^k(\Omega)} = \left\{\int_\Omega \sum_{|\alpha|\leqslant k} (D^\alpha u)^2\, dx\right\}^{1/2},$$

$$D^\alpha u = \partial^{|\alpha|} u / \partial x_1^{\alpha_1} \partial x_2^{\alpha_2}, \quad \alpha = (\alpha_1, \alpha_2), \quad |\alpha| = \alpha_1 + \alpha_2$$

простору $W_2^k(\Omega)$ використовувати яку-небудь еквівалентну норму. Якщо $l_i(u)$, $i = \overline{1, M}$,— лінійні обмежені в $W_2^k(\Omega)$ функціонали, які не перетворюються одночасно на нуль на жодному відмінному від тотожного нуля многочлені степеня не вище $k-1$, то має місце така теорема Соболєва

Теорема 4. *Норми*

$$\|u\|_{W_2^k(\Omega)} \quad та \quad \|u\| = \left\{\int_\Omega (D^k u)^2\, dx + \sum_{i=1}^{M} |l_i(u)|^2\right\}^{1/2}$$

еквівалентні.

1.6. Ортогональні многочлени та функції гіпергеометричного типу

1.6.1. Загальні властивості ортогональних многочленів. Нехай $p_0(x)$, $p_1(x)$, ...— ортогональна в $\tilde{L}_{2,\rho}[a, b]$ система многочленів, тобто

$$(p_k, p_l) \equiv \int_a^b \rho(x)\, p_k(x)\, p_l(x)\, dx = 0, \quad k \neq l.$$

Розглянемо загальні властивості таких многочленів, що випливають з цього означення, під яке, зокрема, підпадають многочлени гіпергеометричного типу.

Теорема 1. *З точністю до сталих множників система ортогональних многочленів для заданого $\rho(x)$ єдина.*

Теорема 2. *Нехай $p_0, p_1, \ldots$— ортогональна в $\tilde{L}_{2,\rho}[a, b]$ система многочленів. Тоді справедливе рекурентне співвідношення*

$$a_n p_{n+1}(x) + (b_n - x)\, p_n(x) + c_n p_{n-1}(x) = 0, \quad n \geqslant 1, \qquad (1)$$

де

$$c_n = (xp_{n-1}, p_n)/\|p_{n-1}\|^2 = \frac{k_{0,n-1}}{k_{0,n}}\,\|p_n\|^2/\|p_{n-1}\|^2,$$

$$b_n = (xp_n, p_n)/\|p_n\|^2 = \frac{k_{1,n}k_{0,n+1} - k_{0,n}k_{1,n+1}}{k_{0,n+1}k_{0,n}}, \qquad (1')$$

$$a_n = (xp_n, p_{n+1})/\|p_{n+1}\|^2 = \frac{k_{0,n}}{k_{0,n+1}},$$

$k_{i,n}$ — коефіцієнти при степенях x^{n-i} многочлена $p_n(x)$.

Зауваження. Якщо покласти $p_{-1}(x) = 0$, то формула буде мати зміст і для $n = 0$.

Теорема 3. *Для системи ортогональних многочленів $p_0, p_1, \ldots$ справджується тотожність Крістоффеля — Дарбу*

$$\sum_{i=0}^{n} \|p_i\|^{-2} p_i(x) p_i(y) = \frac{k_{0,n} \|p_n\|^{-2}}{k_{0,n+1}} \frac{p_{n+1}(x) p_n(y) - p_n(x) p_{n+1}(y)}{x - y}.$$

Теорема 4. *Кожний многочлен $p_n(x)$ має на відрізку $[a, b]$ рівно n різних коренів.*

Теорема 5. *Нехай $x_1^{(n)} < \ldots < x_n^{(n)}$ — нулі многочлена $p_n(x)$. Тоді*

$$a < x_1^{(n)} < x_1^{(n-1)} < \cdots < x_{n-1}^{(n-1)} < x_n^{(n)} < b,$$

тобто нулі многочленів $p_{n-1}(x)$ і $p_n(x)$ чергуються.

Теорема 6. *Нехай вагова функція $\rho(x)$ задовольняє диференціальне рівняння $\rho'(x)/\rho(x) = (a + bx)/(c + dx + ex^2)$, причому в точках a і b вираз $\rho(x)(c + dx + ex^2)$ перетворюється на нуль. Тоді кожний многочлен ортогональної системи $p_0(x), p_1(x), \ldots$ задовольняє диференціальне рівняння*

$$(c + dx + ex^2) p_n''(x) + [(a + d) + (b + 2e)x] p_n'(x) - c_n p_n(x) = 0,$$

де c_n — стала.

1.6.2. Многочлени гіпергеометричного типу. Нагадаємо деякі властивості розв'язків рівняння гіпергеометричного типу

$$\sigma(z) y'' + \tau(z) y' + \lambda y = 0, \tag{4}$$

де $\sigma(z)$ — многочлен не вище другого степеня, $\tau(z)$ — многочлен не вище першого степеня. Розв'язки рівняння (4) називаються функціями гіпергеометричного типу.

Теорема 7. *Похідні будь-якого порядку від функцій гіпергеометричного типу також є функціями гіпергеометричного типу.*

Теорема 8. *Поліноміальні розв'язки рівняння (4), які називаються многочленами (поліномами) гіпергеометричного типу, визначаються формулою*

$$y(z) \equiv y_n(t) = \frac{B_n}{\rho(z)} [\sigma^n(z) \rho(z)]^{(n)}, \quad n = 0, 1, \ldots, \tag{5}$$

однозначно з точністю до нормуючого множника, де $\rho(z)$ є розв'язком рівняння

$$(\sigma\rho)' = \tau\rho, \tag{6}$$

B_n — сталі,

$$B_n = \frac{1}{A_n} y_n^{(n)}(z), \quad A_n = (-1)^n \prod_{k=0}^{n-1} \mu_k, \quad A_0 = 1,$$

$$\mu_k = \lambda + k\tau'(z) + \frac{k(k-1)}{2}\sigma''.$$

Ці розв'язки відповідають значенням $\mu_n = 0$, тобто

$$\lambda = \lambda_n = -n\tau' - \frac{n(n-1)}{2}\sigma'', \quad n = 0, 1, \ldots.$$

Співвідношення (5) було виведено Родрігом в 1814 р. для окремого випадку многочленів гіпергеометричного типу — многочленів Лежандра, для яких $\sigma(z) = 1 - z^2$, $\rho(z) = 1$. Тому (5) називається *формулою Родріга.*

Розв'язуючи рівняння (6), дістаємо залежно від степеня многочлена $\sigma(z)$ такі можливі значення функції $\rho(z)$ (з точністю до сталого множника):

$$\rho(z) = \begin{cases} (b - z)^\alpha (z - a)^\beta & \text{при } \sigma(z) = (b - z)(z - a), \\ (z - a)^\alpha e^{\beta z} & \text{при } \sigma(z) = z - a, \\ e^{\alpha z^2 + \beta z} & \text{при } \sigma(z) = 1, \end{cases}$$

де a, b, α, β — деякі сталі (взагалі кажучи, комплексні). Лінійною заміною незалежної змінної вирази для $\sigma(z)$ і $\rho(z)$ можна привести до таких канонічних виглядів (з точністю до сталого множника):

$$\rho(z) = \begin{cases} (1 - z)^\alpha (1 + z)^\beta & \text{при } \sigma(z) = 1 - z^2, \\ z^\alpha e^{-z} & \text{при } \sigma(z) = z, \\ e^{-z^2} & \text{при } \sigma(z) = 1. \end{cases}$$

При такій заміні рівняння (4) перейде в рівняння того самого вигляду, а відповідні многочлени гіпергеометричного типу $y_n(z)$ залишаться многочленами відносно нової змінної і будуть, як і раніше, визначатися формулою Родріга (5).

Залежно від вигляду функції $\rho(z)$ дістаємо такі системи многочленів.

1) Якщо $\rho(z) = (1 - z)^\alpha (1 + z)^\beta$, $\sigma(z) = 1 - z^2$, то $\tau(z) = -(\alpha + \beta + 2)z + \beta - \alpha$ і відповідні многочленам $y_n(z)$ при $B_n = \frac{(-1)^n}{2^n n!}$ (такий вибір нормуючого множника B_n склався історично і, взагалі кажучи, є довільним) називаються *многочленами Якобі* і позначаються $p_n^{(\alpha,\beta)}(z)$:

$$p_n^{(\alpha,\beta)}(z) = \frac{(-1)^n}{2^n n!} (1 - z)^{-\alpha} (1 + z)^{-\beta} \frac{d^n}{dz^n} [(1 - z)^{n+\alpha} (1 + z)^{n+\beta}].$$

Важливими окремими випадками многочленів Якобі є такі:

а) *многочлени Лежандра* $p_n(z) = p^{(0,0)}(z)$;

б) *многочлени Чебишева* першого і другого роду

$$T_n(z) = \frac{n!}{(1/2)_n} p_n^{(-1/2,-1/2)}(z),$$

$$U_n(z) = \frac{(n+1)!}{(3/2)_n} p_n^{(1/2,1/2)}(z),$$

де

$$(\alpha)_n = \alpha(\alpha+1)\dots(\alpha+n-1) = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)},$$

$\Gamma(z)$ — гама-функція Ейлера.

Нагадаємо, що

$$\Gamma(z+1) = z\Gamma(z), \quad \Gamma(n+1) = n!, \quad \Gamma\left(n+\frac{1}{2}\right) = \frac{\sqrt{\pi}\,(2n)!}{2^{2n}n!}.$$

2) Нехай $\rho(z) = z^\alpha e^{-z}$, $\sigma(z) = z$. Тоді $\tau(z) = -z+\alpha+1$ і відповідні многочлени $y_n(z)$ при $B_n = \frac{1}{n!}$ називаються *многочленами Лагерра* і позначаються $L_n^\alpha(z)$:

$$L_n^\alpha(z) = \frac{1}{n!}\, e^z z^{-\alpha} \frac{d^n}{dz^n}(z^{\alpha+n}e^{-z}).$$

3) Нехай $\rho(z) = e^{-z^2}$, $\sigma(z) = 1$. Тоді $\tau(z) = -2z$. Многочлени $y_n(z)$ при $B_n = (-1)^n$ називаються *многочленами Ерміта* і позначаються $H_n(z)$:

$$H_n(z) = (-1)^n e^{z^2} \frac{d^n}{dz^n}(e^{-z^2}).$$

Теорема 9. *Нехай функція* $\rho(z)$ *задовольняє на кінцях деякого інтервалу* (a, b) *умову*

$$\sigma(z)\rho(z)z^k|_{z=a,b} = 0, \quad k = 0, 1, \dots. \tag{7}$$

Тоді многочлени гіпергеометричного типу $y_n(z)$, *які відповідають різним значенням* λ_n, *будуть ортогональні на інтервалі* (a, b) *з вагою* $\rho(z)$, *тобто*

$$\int_a^b \rho(z)\, y_n(z)\, y_m(z)\, dz = 0, \quad \lambda_m \neq \lambda_n. \tag{8}$$

Многочлени гіпергеометричного типу, для яких функція $\rho(z)$ задовольняє умову (7), називаються *класичними ортогональними многочленами*. Ці многочлени розглядаються за додаткових умов $\rho(z) > 0$, $\sigma(z) > 0$ на інтервалі (a, b). Переліченим вимогам многочлени Якобі $p_n^{(\alpha,\beta)}(z)$ задовольняють при $a = -1$, $b = 1$, $\alpha > -1$, $\beta > -1$ многочлени Лагерра $L_n^\alpha(z)$ — при $b = +\infty$, $a = 0$, $\alpha > -1$, многочлени Ерміта — при $a = -\infty$, $b = +\infty$. Зауважимо, що в цих випадках у (8) умову $\lambda_m \neq \lambda_n$ можна замінити еквівалентною умовою $m \neq n$.

Для квадрата норми $d_n^2 = \int_a^b \rho(z)\, y_n^2(z)\, dz$ і коефіцієнтів $k_{0,n}$ і $k_{1,n}$ в зображенні класичного ортогонального многочлена

$$y_n(z) = k_{0,n}z^n + k_{1,n}z^{n-1} + \dots$$

Таблиця 1.

$y_n(z)$	$p_n^{(\alpha,\beta)}(z)$ $(\alpha > -1,\ \beta > -1)$	$L_n^\alpha(z)$ $(\alpha > -1)$	$H_n(z)$
(a, b)	$(-1, 1)$	$(0, \infty)$	$(-\infty, \infty)$
$\rho(z)$	$(1-z)^\alpha(1+z)^\beta$	$z^\alpha e^{-z}$	e^{-z^2}
$\sigma(z)$	$1-z^2$	z	1
$\tau(z)$	$\beta-\alpha-(\alpha+\beta+2)z$	$1+\alpha-z$	$-2z$
λ_n	$n(n+\alpha+\beta+1)$	n	$2n$
B_n	$\frac{(-1)^n}{2^n n!}$	$\frac{1}{n!}$	$(-1)^n$

Таблиця 2.

$k_{0,n}$	$\frac{\Gamma(2n+\alpha+\beta+1)}{2^n n!\,\Gamma(n+\alpha+\beta+1)}$	$\frac{(-1)^n}{n!}$	2^n
$k_{1,n}$	$\frac{(\alpha-\beta)\,\Gamma(2n+\alpha+\beta)}{2^n(n-1)!\,\Gamma(n+\alpha+\beta+1)}$	$(-1)^{n-1}\times$ $\times\frac{n+\alpha}{(n-1)!}$	0
d_n^2	$\frac{2^{\alpha+\beta+1}\Gamma(n+\alpha+1)\,\Gamma(n+\beta+1)}{n!\,(2n+\alpha+\beta+1)\,\Gamma(n+\alpha+\beta+1)}$	$\frac{\Gamma(n+\alpha+1)}{n!}$	$2^n n!\sqrt{\pi}$
a_n	$\frac{2(n+1)(n+\alpha+\beta+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}$	$-(n+1)$	$1/2$
b_n	$\frac{\beta^2-\alpha^2}{(2n+\alpha+\beta)(2n+\alpha+\beta+2)}$	$2n+\alpha+1$	0
c_n	$\frac{2(n+\alpha)(n+\beta)}{(2n+\alpha+\beta)(2n+\alpha+\beta+1)}$	$-(n+\alpha)$	n

справедливі формули

$$k_{0,n} = \frac{A_n B_n}{n!} = B_n \prod_{k=0}^{n-1}\left(\tau' + \frac{n+k-1}{2}\sigma''\right), \quad a_0 = B_0,$$

(B_n — нормуюча стала в формулі Родріга),

$$\frac{k_{1,n}}{k_{0,n}} = \frac{n\tau_{n-1}(0)}{\tau'_{n-1}(0)},$$

$$d_n^2 = (-1)^n n!\, k_{0,n} B_n \int_a^b \sigma^n(z)\,\rho(z)\,dz. \tag{9}$$

Інтеграл в (9) можна виразити через гама-функцію Ейлера $\Gamma(z)$.

Основні характеристики для класичних ортогональних многочленів зручно записати у вигляді таблиць (табл. 1, 2).

Вправа. Виражаючи значення гама-функції через факторіал, записати основні характеристики для многочленів Лежандра, Чебишева першого і другого роду, Лагерра.

ГЛАВА 2

ІНТЕРПОЛЮВАННЯ

2.1. Задача інтерполювання та системи Чебишева

2.1.1. Постановка задачі інтерполювання. Нехай $C(\bar{\Omega})$ — банахів простір функцій, заданих на компакті $\bar{\Omega}$ з нормою (її часто називають *рівномірною* або *чебишевською*) $\|f\|_C \equiv \|f\|_\infty = \max\limits_{x\in\bar{\Omega}} |f(x)|$. Нехай задано різні точки $x_i \in \bar{\Omega}$, $i = \overline{0, n}$, і значення $f(x_i)$ функції $f(x)$. Іншими словами, задано сітку (нерівномірну) $\omega_n = \{x_i : x_i \in \bar{\Omega}, i = \overline{0, n}, x_i \neq x_j\}$ і сіткову функцію $f(x)$, $x \in \omega_n$. Нехай $\{\varphi_i(x)\} \subset C(\bar{\Omega})$ — система лінійно незалежних функцій, а M_{n+1} — підпростір розмірності $n+1$ узагальнених многочленів виду $\psi(x) = \sum\limits_{i=0}^{n} a_i\varphi_i(x)$ степеня (порядку) n з дійсними коефіцієнтами a_i. Задача інтерполювання ставиться таким чином: за заданим $x \neq x_i$, $i = \overline{0, n}$, знайти наближене значення функції $f(x)$ як значення в точці x узагальненого многочлена $p(x, f) = \sum\limits_{i=0}^{n} a_i\varphi_i(x)$ такого, що $p(x_i, f) = f(x_i)$, $i = \overline{0, n}$. Многочлен $p(x; f)$, який задовольняє останню умову, називається *інтерполяційним* для функції $f(x)$. Таким чином, щоб розв'язати задачу інтерполювання (або інтерполяції) потрібно: 1) побудувати інтерполяційний многочлен; 2) обчислити його значення в точці $x \neq x_i$, $i = \overline{0, n}$. У розглянутій постановці маємо справу з *лінійним інтерполюванням*, бо у $p(x; f)$ параметри a_i, які і визначають конкретний многочлен, входять лінійно. Якщо ж $p(x; f) \equiv p(x, a_0, \ldots, a_n, f)$ нелінійно залежить від параметрів, то інтерполювання називається *нелінійним*. Де ж на практиці зустрічається задача інтерполювання? Наведемо такі приклади: 1) в дискретні моменти часу $t_0, t_1, \ldots, t_n$ спостерігаються значення функції $f(t)$ (наприклад, $f(t)$ — висота літака і потрібно відновити її при будь-якому $t \neq t_i$); 2) на ЕОМ потрібно багатократно обчислювати одну і ту саму складну функцію в різних точках. В останньому випадку доцільно обчислити раз і назавжди її значення в фіксованих точках $x_0, x_1, \ldots, x_n$, а в інших точках обчислювати її наближені значення, використовуючи її інтерполяційний многочлен, який є «простішим» для обчислень.

2.1.2. Система Чебишева. Необхідна і достатня умова однозначного розв'язку задачі побудови інтерполяційного многочлена. У випадку лінійного інтерполювання для однозначного визначення інтерполяційного узагальненого многочлена на систему функцій $\{\varphi_i\}$ доводиться накладати додаткові обмеження.

Означення 1. Система функцій $\{\varphi_i(x)\}$, $i = \overline{0, n}$, називається *системою Чебишева* (порядку n) на компакті $\bar{\Omega}$, якщо будь-який узагальнений многочлен по цій системі, в якого хоча б один коефіцієнт відмінний від нуля, має на $\bar{\Omega}$ не більш ніж n різних коренів.

У цьому випадку підпростір M_{n+1} з базисом $\varphi_i(x)$, $i = \overline{0, n}$, називається *підпростором, що задовольняє умову Хаара*. Очевидно, будь-яка система Чебишева є лінійно незалежною, але не навпаки.

Теорема 1. *Для того щоб для будь-якої функції $f(x) \in C(\bar{\Omega})$ і будь-якого набору з $n+1$ різних точок $x_i \in \bar{\Omega}$, $i = \overline{0, n}$, існував узагальнений інтерполяційний многочлен, необхідно і достатньо, щоб система функцій $\{\varphi_i(x)\}$ була системою Чебишева на $\bar{\Omega}$. При цьому узагальнений інтерполяційний многочлен — єдиний.*

Доведення. Очевидно, для справедливості теореми необхідно і достатньо, щоб система лінійних алгебраїчних рівнянь відносно a_i

$$\sum_{i=0}^{n} a_i\varphi_i(x_j) = f(x_j), \quad j = \overline{0, n}, \tag{1}$$

мала розв'язок при будь-якому виборі різних точок, x_i, $i = \overline{0, n}$, і будь-яких числах $f(x_j)$, $j = \overline{0, n}$. Це можливо тоді і лише тоді, коли

$$\Phi = \det\begin{vmatrix} \varphi_0(x_0) & \ldots & \varphi_n(x_0) \\ \ldots & \ldots & \ldots \\ \varphi_0(x_n) & \ldots & \varphi_n(x_n) \end{vmatrix} \neq 0, \quad \forall\, x_j \in \bar{\Omega}. \tag{2}$$

Покажемо, що, в свою чергу, для виконання (2) необхідно і достатньо, щоб система функцій $\{\varphi_i(x)\}$ була системою Чебишева на $\bar{\Omega}$.

Необхідність. Нехай $\Phi \neq 0$ $\forall\, x_j \in \bar{\Omega}$ ($x_j \neq x_j$ при $i \neq j$). Покажемо, що тоді $\{\varphi_i(x)\}$ — система Чебишева. Припустимо від супротивного, що система $\{\varphi_i(x)\}$ не є системою Чебишева на $\bar{\Omega}$. Це значить, що існує набір чисел b_i, серед яких хоча б одне не дорівнює нулю, і набір різних точок x_j, $j = \overline{0, n}$, таких, що

$$\sum_{i=0}^{n} b_i\varphi_i(x_j) = 0. \tag{3}$$

Останнє означає лінійну залежність стовпчиків визначника Φ і рівність його нулю, що суперечить початковій умові. Це і доводить її необхідність.

Достатність. Припустимо, що система $\{\varphi_i(x)\}$ утворює систему Чебишева на $\bar{\Omega}$, і покажемо, що тоді $\Phi \neq 0$. З означення системи

Чебишева маємо, що не існує набору чисел b_i, $i = \overline{0, n}$, серед яких хоча б одне було відмінне від нуля, і набору різних точок x_j, $j = \overline{0, n}$, для яких виконувалась би рівність (3). Це означає лінійну незалежність стовпчиків визначника і, отже, те, що він відмінний від нуля. Теорему доведено.

Легко помітити, що узагальнений інтерполяційний многочлен для функції $f(x)$ за системою Чебишева можна подати у вигляді

$$p(x; f) \equiv p_n(x; f) = \sum_{i=0}^{n} f(x_i)\, l_{i,n}(x), \qquad (4)$$

де $l_{i,n}(x)$ — фундаментальні узагальнені многочлени інтерполяції, які будуються за системою $\{\varphi_i(x)\}$ і мають такі властивості:

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

($\delta_{i,j}$ — символ Кронекера). Неважко також помітити, що коли

$$\Phi_i(x) = \begin{vmatrix} \varphi_0(x_0) & \dots & \varphi_n(x_0) \\ \vdots & & \vdots \\ \varphi_0(x_{i-1}) & \dots & \varphi_n(x_{i-1}) \\ \varphi_0(x) & \dots & \varphi_n(x) \\ \varphi_0(x_{i+1}) & \dots & \varphi_n(x_{i+1}) \\ \vdots & & \vdots \\ \varphi_0(x_n) & \dots & \varphi_n(x_n) \end{vmatrix}, \qquad (5)$$

то $l_{i,n}(x) = \Phi_i(x)/\Phi$. Величину

$$R(x; f) \equiv R_n(x; f) \equiv R(x) = f(x) - p(x; f)$$

називатимемо *залишковим членом інтерполяційного многочлена.*

2.1.3. Умови, за яких система функцій є системою Чебишева. Розглянемо питання про те, які системи функцій і на яких компактах можуть бути системами Чебишева.

Означення 2. Система трьох неперервних відображень f_i: $[0, 1] \to \overline{\Omega}$ $(i = 1, 2, 3)$ таких, що $f_i(t) \not\equiv \text{const}$, $f_i(0) = a \in \overline{\Omega}$ і $f_i(t) \neq f_j(t')$ $\forall\, t, t' > 0$, $i \neq j$, називається *триподом.*

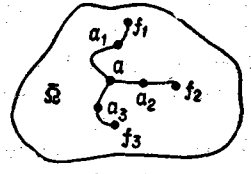


Рис. 14

Для двовимірного компакту $\overline{\Omega}$ область значень триподу показано на рис. 14.

Теорема 2. *Якщо компакт $\overline{\Omega}$ містить образ деякого триподу, то в $C(\overline{\Omega})$ немає підпростору*

розмірності більшої або рівної двом, який би задовольняв умову Хаара, тобто для такого компакту не існує системи Чебишева.

Доведення. Визначник $\Phi \equiv \Phi(x_0, x_1, \dots, x_n)$ неперервний за всіма змінними $x_0, x_1, x_2, \dots, x_n$. Зафіксуємо точки $x_2, x_3, x_4, \dots, x_n$ і розглянемо функцію $\varphi(a_1, a_2) = \Phi(a_1, a_2, x_2, x_3, \dots, x_n)$, де $a_1 = f_1(x_0)$, $a_2 = f_2(x_1)$. Виконаємо неперервно таке переміщення: 1) точку a_1 перемістимо на місце $a_3 : f_1(t_1) \to f_1(0) = a \to f_3(t_3)$; 2) точку a_2 перемістимо на місце $a_1 : f_2(t_1) \to f_2(0) = a \to f_1(t_1)$; 3) точку a_3 перемістимо на місце a_2: $(f_3(t_3) \to f_3(0) = a \to f_2(t_2)$. В результаті у визначнику Φ поміняються місцями два рядки і тому він змінить знак. Отже, існують проміжні точки $\bar{a}_1$, $\bar{a}_2$, для яких $\varphi(\bar{a}_1, \bar{a}_2) \equiv \Phi(\bar{a}_1, \bar{a}_2, x_2, \dots, x_n) = 0$, тобто система $\{\varphi_i(x)\}$ не може бути системою Чебишева.

Теорема 3. *Нехай $\overline{\Omega}$ — компактна множина в $\mathbb{R}^m$. Підпростір $M_n \subset C(\overline{\Omega})$ розмірності $n \geqslant 2$, що задовольняє умову Хаара, існує тоді і лише тоді, коли $\overline{\Omega}$ гомеоморфна замкненій частині кола* [1].

Теореми 1 і 2 говорять про те, що у випадку $\overline{\Omega} \subset \mathbb{R}^m$, $m \geqslant 2$, системи Чебишева існують на досить «екзотичних» компактах $\overline{\Omega}$.

Нехай $\overline{\Omega}$ є відрізком числової прямої. Виявимо достатні умови того, щоб система функцій $\{\varphi_i(x)\}$, $i = \overline{0, n}$, була системою Чебишева. Наведемо без доведення теорему, яка належить Прюферу і є однією з достатніх умов.

Теорема 4. *Нехай $p_0(x) > 0$, $r_0(x)$ і $q_0(x)$ — дійсні неперервні функції, що визначені на обмеженому відрізку $[a, b]$.*
Нехай $\lambda_0 < \lambda_1 < \dots$ — деякі числа і кожне рівняння

$$(p_0(x)u')' + [q_0(x) + \lambda_i r_0(x)]u = 0, \quad i = 0, 1, \dots,$$

має дійсний розв'язок $u_i(x)$, $x \in [a, b]$, який на $[a, b]$ має і нулів і такий, що існує $\lim u_i(x)/u_0(x)$ при $x \to a$ та $x \to b$. Тоді система функцій $u_i(x)$, $i = \overline{0, n}$, при будь-якому n утворює систему Чебишева на $[a, b]$.

Наступна теорема дає достатню умову того, щоб система досить гладких функцій була системою Чебишева.

Теорема 5. *Якщо $\varphi_i(x) \in C^{(n+1)}$ $[a, b]$, $i = \overline{0, n}$, і вронськіан $W[\varphi_0, \dots, \varphi_k] \neq 0$ для всіх $k = \overline{0, n}$ та всіх $x \in [a, b]$, то система $\{\varphi_i(x)\}$, $i = \overline{0, n}$, є системою Чебишева.*

Перш ніж перейти до доведення теореми 5, доведемо таке узагальнення теореми Ролля.

---

[1] *Гомеоморфізм* — взаємно однозначна відповідність між двома топологічними просторами, при якій обидва взаємно обернені відображення, що визначаються цією відповідністю, є неперервними.

Лема 1. *Нехай* $f(x) \in C^{(n+1)}[a, b]$ *і має на* $[a, b]$ $n+2$ *корені. Тоді існує точка* $\xi \in [a, b]$, *в якій вираз*

$$L_{n+1}(x;\ f) = W[\varphi_0,\ \varphi_1,\ \ldots,\ \varphi_n,\ f]/W[\varphi_0,\ \varphi_1,\ \ldots,\ \varphi_n]$$

перетворюється на нуль.

Доведення. З умов леми випливає, що для диференціального рівняння відносно функції φ

$$L_{k+1}(x;\ \varphi) = W[\varphi_0,\ \ldots,\ \varphi_k,\ \varphi]/W[\varphi_0,\ \ldots,\ \varphi_k] = 0,$$
$$k = 0,\ 1,\ \ldots,\ L_0(x;\ \varphi) = \varphi(x),$$

функції $\varphi_i(x)$, $i = \overline{0, k}$, утворюють фундаментальну систему, причому коефіцієнт при старшій похідній в цьому рівнянні дорівнює одиниці. Покажемо, що можна знайти такі функції $b_0(x)$, $b_1(x)$, ..., $b_n(x)$, що

$$L_{k+1}(x;\ \varphi) = \frac{d}{dx} L_k(x;\ \varphi) - b_k L_k(x;\ \varphi).$$

Дійсно, розглянемо лінійний диференціальний оператор порядку $k+1$ з коефіцієнтом при старшій похідній, що дорівнює одиниці:

$$\frac{d}{dx} L_k(x;\ \varphi) - b_k L_k(x;\ \varphi) = \frac{W[\varphi_0,\ \ldots,\ \varphi_{k-1}]\ \frac{d}{dx} W[\varphi_0, \ldots, \varphi_{k-1}, \varphi]}{W^2[\varphi_0,\ \ldots,\ \varphi_{k-1}]} -$$

$$- \frac{W[\varphi_0,\ \ldots,\ \varphi_{k-1},\ \varphi]\ \frac{d}{dx} W[\varphi_0, \ldots, \varphi_{k-1}]}{W^2[\varphi_0,\ \ldots,\ \varphi_{k-1}]} - b_k(x) L_k(x;\ \varphi).$$

Цей диференціальний оператор задовольняє умови

$$\frac{d}{dx} L_k(x;\ \varphi_j) - b_k(x) L_k(x;\ \varphi_j) \equiv 0, \quad j = \overline{0,\ k-1}. \qquad (6)$$

Визначимо $b_k(x)$ так, щоб рівність (6) виконувалась для $j = k$. Для цього досить покласти

$$b_k(x) = \frac{d}{dx} L_k(x;\ \varphi_k)/L_k(x;\ \varphi_k),$$

причому в силу умов леми $L_k(x;\ \varphi_k) \neq 0$, і тому $b_k(x)$ є неперервною функцією. При такому виборі $b_k(x)$ система $\{\varphi_i(x)\}$, $i = \overline{0, k}$, є фундаментальною для рівняння

$$\frac{d}{dx} L_k(x;\ \varphi) - b_k(x) L_k(x;\ \varphi) = 0,$$

а це означає, що

$$L_{k+1}(x;\ \varphi) \equiv \frac{d}{dx} L_k(x;\ \varphi) - b_k(x) L_k(x;\ \varphi).$$

Розглянемо тепер функцію

$$\psi_1(x) = f(x) \exp\left[-\int_a^x b_0(x)\,dx\right] = L_0(x;\ f) \exp\left[-\int_a^x b_0(x)\,dx\right],$$

що як і $f(x)$ перетворюється на нуль $n+2$ разів на $[a, b]$. Тому її похідна

$$\psi_1'(x) = [f'(x) - b_0(x) f(x)] \exp\left[-\int_a^x b_0(x)\,dx\right] =$$

$$= L_1(x;\ f) \exp\left[-\int_a^x b_0(x)\,dx\right],$$

і, отже, $L_1(x;f)$ перетворюється на $[a, b]$ на нуль принаймні $n+1$ разів.

Далі вводимо функцію

$$\psi_2(x) = L_1(x;\ f) \exp\left[-\int_a^x b_1(x)\,dx\right]$$

і аналогічно доводимо, що $L_2(x;\ f)$ перетворюється на нуль принаймні n разів. Продовжуючи цей процес, дійдемо висновку, що знайдеться хоча б одна точка $\xi \in [a, b]$, для якої

$$L_{n+1}(\xi;\ f) = 0.$$

Доведення теореми 5. Припустимо, що твердження теореми не справджується. Тоді знайдеться лінійна комбінація з дійсними коефіцієнтами c_i, серед яких хоча б один відмінний від нуля

$$f(x) = \sum_{i=0}^{n} c_i \varphi_i(x),$$

така, що $f(x)$ перетворюється на нуль принаймні в $n+1$ різних точках відрізка $[a, b]$. В силу щойно доведеної леми вираз $L_n(x;\ f)$ має перетворюватися на нуль хоча б в одній точці $\xi \in [a, b]$. Але

$$L_n(x;\ f) = \frac{W[\varphi_0,\ \varphi_1,\ \ldots,\ \varphi_{n-1},\ f]}{W[\varphi_0,\ \varphi_1,\ \ldots,\ \varphi_{n-1}]} = c_n \frac{W[\varphi_0,\ \varphi_1,\ \ldots,\ \varphi_n]}{W[\varphi_0,\ \varphi_1,\ \ldots,\ \varphi_{n-1}]}.$$

Оскільки за умовами теореми 5 вирази $W[\varphi_0, \ldots, \varphi_k]$ при всіх $k = \overline{0, n}$ не перетворюються на нуль на $[a, b]$, то $c_n = 0$. Значить, знайдеться $n+1$ різних точок відрізка $[a, b]$, в яких

$$f = \sum_{i=0}^{n-1} c_i \varphi_i(x)$$

перетворюється на нуль. Тоді, знову застосовуючи узагальнення теореми Ролля, знайдемо, що $L_{n-1}(x;\ f) = 0$ принаймні в одній точці

$\xi \in [a, b]$. Проводячи ті самі міркування, що й вище, знайдемо $c_{n-1} = 0$. Продовжуючи цей процес, дійдемо висновку, що всі коефіцієнти c_i дорівнюють нулю всупереч нашому припущенню. Теорему 5 доведено.

Означення 3. Система функцій $\{\varphi_i(x)\}_{i=0}^{n}$, що утворюють систему Чебишева на $[a, b]$ і для яких виконується рівність $\varphi_i(a) = \varphi_i(b)$, $i = \overline{0, n}$, називається *періодичною системою Чебишева*. Число n називається *порядком системи Чебишева*, а точки a та b вважаються за одну точку.

Лема 2. *Нехай* $a = x_0 < x_1 < \ldots < x_n < b = x_{n+1}$. *Тоді функція*

$$\tilde{\Phi}(x) \equiv \tilde{\Phi}(x, x_1, \ldots, x_n) = \begin{vmatrix} \varphi_0(x) & \varphi_1(x) & \ldots & \varphi_n(x) \\ \varphi_0(x_1) & \varphi_1(x_1) & \ldots & \varphi_n(x_1) \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_0(x_n) & \varphi_1(x_n) & \ldots & \varphi_n(x_n) \end{vmatrix},$$

де $\{\varphi_i(x)\}_{i=\overline{0,n}}$ — *система Чебишева, зберігає знак на кожному з інтервалів* (x_j, x_{j+1}), $j = \overline{0, n}$, *причому знаки* $\tilde{\Phi}(x)$ *в послідовних інтервалах чергуються.*

Доведення. Оскільки $\tilde{\Phi}(x)$ є узагальненим многочленом за системою Чебишева, який обертається на нуль в точках x_i, $i = \overline{1, n}$, то в жодній іншій точці відрізка $[a, b]$ він не може перетворюватись на нуль. Зауважимо також, що знак $\tilde{\Phi}(x)$ не зміниться, якщо як завгодно неперевно переміщувати точки $x, x_1, \ldots, x_n$ на відрізку $[a, b]$, не змінюючи їх взаємного розміщення.

Розглянемо два послідовних інтервали (x_k, x_{k+1}) та (x_{k+1}, x_{k+2}) і нехай $\xi \in (x_k, x_{k+1})$, $\eta \in (x_{k+1}, x_{k+2})$. В силу висловленого вище зауваження знак $\tilde{\Phi}(\xi, x_1, \ldots, x_n)$ залишиться незмінним, якщо ξ перемістити в положення x_{k+1}, а x_{k+1} — в положення η, тобто

$$\operatorname{sign}\tilde{\Phi}(\xi, x_1, \ldots, x_k, x_{k+1}, \ldots, x_n) = \operatorname{sign}\tilde{\Phi}(x_{k+1}, x_1, \ldots, x_k, \eta, x_{k+2}, \ldots, x_n) = -\operatorname{sign}\tilde{\Phi}(\eta, x_1, \ldots, x_k, x_{k+1}, x_{k+2}, \ldots, x_n).$$

Але визначники $\tilde{\Phi}(x_{k+1}, x_1, \ldots, x_k, \eta, x_{k+2}, \ldots, x_n)$ та $\tilde{\Phi}(\eta, x_1, \ldots, x_k, x_{k+1}, \ldots, x_n)$ відрізняються лише знаками. Таким чином, $\operatorname{sign}\tilde{\Phi}(\xi) = -\operatorname{sign}\tilde{\Phi}(\eta)$, що і треба було довести.

Теорема 6. *Порядок* n *періодичної системи Чебишева на відрізку* $[a, b]$ *має бути парним числом.*

Доведення. З леми 2 випливає, що при зростанні x від a до b визначник $\tilde{\Phi}(x, x_1, \ldots, x_n)$ змінює знак n разів і $\tilde{\Phi}(a) = \tilde{\Phi}(b)$, а це можливо лише при парному n.

Теорема 3 показує, що власні функції багатьох диференціальних операторів утворюють системи Чебишева і за ними можна будувати узагальнені інтерполяційні многочлени. Грунтуючись на цій теоремі, наведемо приклади простіших і найчастіше вживаних систем Чебишева.

Приклад 1. За допомогою теорем 4, 6 при $p_0(x) = 1$, $r_0(x) = 1$, $q_0(x_0) = 0$, $\lambda_n = n^2$, $a = 0$, $b = 2\pi$ неважко перевірити, що система функцій $1, \sin x, \cos x, \ldots, \sin kx, \cos kx$, є періодичною системою Чебишева на $[0, 2\pi]$.

Приклад 2. Поклавши $p_0(x) = x^2$, $q_0(x) = 0$, $r_0(x) = 1$, $\lambda_n = -(i+1)(i+2)$, дістанемо, що система функцій x^i, $i = \overline{0, n}$, утворює систему Чебишева на будь-якому відрізку $[a, b] \in (-\infty, \infty)$.

Вправа 1. Перевірити, що система функцій $1, e^{\alpha_1 x}, e^{\alpha_2 x}$, де α_i — різні числа, утворює систему Чебишева на будь-якому відрізку $[a, b] \in (-\infty, \infty)$.

Вказівка. Через визначник Вандермонда явно обчислити вронськіан $W[\varphi_0, \ldots, \varphi_k]$ і скористатися теоремою 5.

Вправа 2. Чи є системою Чебишева система $\{x^{n_j}\}_{j=0}^{\infty}$, де n_j — різні цілі невід'ємні числа?

2.2. Побудова інтерполяційного многочлена

Ми довели, що у випадку, коли система функцій $\varphi_i(x)$, $i = \overline{0, n}$, є системою Чебишева, існує єдиний інтерполяційний многочлен виду $\varphi(x) = \sum_{i=0}^{n} a_i \varphi_i(x)$. Як випливає з результатів п. 2.1, система функцій $\varphi_i(x) = x^i$, $i = \overline{0, n}$, є системою Чебишева на будь-якому скінченному відрізку $[a, b] \subset (-\infty, \infty)$, а тому згадане вище твердження справедливе і для неї. Інтерполяційний многочлен виду $\varphi(x) = \sum_{i=0}^{n} a_i x^i$, тобто алгебраїчний інтерполяційний многочлен, найчастіше використовується на практиці. Існують різні форми запису цього многочлена.

2.2.1. Інтерполяційний многочлен у формі Лагранжа. Неважко помітити, що у випадку системи функцій $\varphi_i(x) = x^i$ фундаментальні многочлени інтерполяції в формулі (4) (п. 2.1) можна записати у вигляді

$$l_{i,n} = \frac{(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)}{(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)}.$$

Отже, інтерполяційний многочлен за системою функцій x^i, $i = \overline{0, n}$, набирає вигляду

$$p(x; f) \equiv L_n(x; f) \equiv L_n(x) = \sum_{i=0}^{n} f_i \frac{(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)}{(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)}. \quad (1)$$

Інтерполяційний многочлен, записаний у вигляді (1), носить назву інтерполяційного многочлена в формі Лагранжа. Така форма запису не завжди зручна. Одним з її недоліків є те, що для обчислення $L_n(x)$ при фіксованому x потрібно виконати $O(n^2)$ арифметичних операцій. Далі побачимо, що запис того самого многочлена в формі Ньютона дає зробити це за $O(n)$ операцій.

2.2.2. Розділені різниці. Інтерполяційний многочлен у формі Ньютона. За означенням розділена різниця нульового порядку $f(x_i)$ від функції $f(x)$ по одному вузлу x_i збігається із значенням функції $f(x_i)$. Різниці першого порядку по вузлах x_i, x_j визначаються рівністю

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

різниці другого порядку — рівністю

$$f[x_i, x_j, x_k] = \frac{f[x_j, x_k] - f[x_i, x_j]}{x_k - x_i} \tag{2}$$

і т. д. Розділені різниці k-го порядку $f[x_1, \ldots, x_{k+1}]$ визначаються через різниці $(k-1)$-го порядку за формулою

$$f[x_1, \ldots, x_{k+1}] = \frac{f[x_2, \ldots, x_{k+1}] - f[x_1, \ldots, x_k]}{x_{k+1} - x_1}. \tag{3}$$

Іноді замість $f[x_1, \ldots, x_k]$ використовують позначення $(f)(x_1; \ldots; x_k)$ або $[x_1; \ldots; x_k]$.

Вправа 1. а) Довести справедливість формули

$$f[x_1, \ldots, x_k] = \sum_{j=1}^{k} \frac{f(x_j)}{\prod\limits_{i \neq j} (x_j - x_i)}; \tag{4}$$

б) Довести, що розділена різниця є лінійним оператором від функції f, тобто $(\alpha_1 f_1 + \alpha_2 f_2)[x_1, \ldots, x_k] = \alpha_1 f_1[x_1, \ldots, x_k] + \alpha_2 f_2[x_1, \ldots, x_k]$.

в) Довести, що розділена різниця є симетричною функцією своїх аргументів $x_1, \ldots, x_k$, тобто не змінюється при будь-якій перестановці їх.

Вказівка. Для доведення (4) застосувати індукцію по k. Інші дві властивості є наслідком (4).

Якщо функція задана на сітці $\omega_n = \{x_i : i = \overline{1, n},\ x_i \neq x_j \text{ при } i \neq j\}$, то таблицю

$$\begin{array}{llll}
f(x_1) & & & \\
f(x_2) & f[x_1, x_2] & & \\
f(x_3) & f[x_2, x_3] & f[x_1, x_2, x_3] & \\
\vdots & \vdots & \vdots & \ddots \\
f(x_n) & f[x_{n-1}, x_n] & f[x_{n-2}, x_{n-1}, x_n] & \ldots f[x_1; \ldots; x_n]
\end{array}$$

називають таблицею її розділених різниць.

Вправа 2. Побудувати алгоритм обчислення розділених різниць $f[x_1, \ldots, x_i]$, $i = \overline{1, n}$, який використовував би лише два масиви розмірності n, один з яких спочатку містить значення x_i, $i = \overline{1, n}$, а інший значення $f(x_i)$, $i = \overline{1, n}$.

Нехай $\mathscr{P}(x)$ — многочлен степеня n. Розглянемо його розділені різниці, віднімаючи від $\mathscr{P}(x)$ константу $\mathscr{P}(x_0)$, яка, очевидно, не впливає на розділені різниці, дістанемо многочлен $\mathscr{P}(x) - \mathscr{P}(x_0)$, що перетворюється на нуль при $x = x_0$ і тому ділиться на $x - x_0$. Отже, перша розділена різниця многочлена n-го степеня

$$\mathscr{P}[x;\ x_0] = [\mathscr{P}(x_0) - \mathscr{P}(x)]/(x_0 - x)$$

є многочленом степеня $n - 1$ відносно x. Із формули (2) видно, що чисельник другої різниці $\mathscr{P}[x, x_0, x_1]$ перетворюється на нуль при $x = x_1$ і, значить, націло ділиться на $x - x_1$. Це означає, що $\mathscr{P}[x, x_0, x_1]$, є многочленом степеня $n - 2$. Продовжуючи ці міркування, дійдемо висновку, що різниця $\mathscr{P}[x; x_0; \ldots; x_{n-1}]$ є многочленом нульового степеня, тобто стала, а розділені різниці більш високого порядку дорівнюють нулю.

З означення розділених різниць випливає

$$\begin{gathered}
\mathscr{P}(x) = \mathscr{P}(x_0) + (x - x_0)\mathscr{P}[x, x_0], \\
\mathscr{P}[x, x_0] = \mathscr{P}[x_0, x_1] + (x - x_1)\mathscr{P}[x, x_0, x_1], \\
\mathscr{P}[x, x_0, x_1] = \mathscr{P}[x_0, x_1, x_2] + (x - x_2)\mathscr{P}[x, x_0, x_1, x_2] \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots
\end{gathered}$$

т. д. Звідси для $\mathscr{P}(x)$ дістаємо формулу

$$\begin{aligned}
\mathscr{P}(x) = \mathscr{P}(x_0) + (x - x_0)\mathscr{P}[x_0, x_1] + (x - x_0)(x - x_1) \times \\
\times \mathscr{P}[x_0, x_1, x_2] + \cdots + (x - x_0)(x - x_1)\ldots \\
\ldots (x - x_{n-1})\mathscr{P}[x_0, x_1, x_2, \ldots, x_n].
\end{aligned} \tag{5}$$

Якщо $\mathscr{P}(x)$ — інтерполяційний многочлен для функції $f(x)$ за системою Чебишева $\varphi_i(x) = x^i$, $i = \overline{0, n}$, то його значення у вузлах сітки $\omega = \{x_0, x_1, \ldots, x_n\}$ збігаються із значеннями функції $f(x)$, а отже, збігаються і розділені різниці. Тому інтерполяційний многочлен за системою Чебишева $\varphi_i(x) = x^i$ для функції $f(x)$ можна записати у вигляді

$$\begin{aligned}
p_n(x;\ f) \equiv p(x;\ f) = f(x_0) + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1) \times \\
\times f[x_0, x_1, x_2] + \cdots + (x - x_0)(x - x_1)\ldots \\
\ldots (x - x_{n-1})f[x_0, x_1, \ldots, x_n].
\end{aligned} \tag{6}$$

Такий запис інтерполяційного многочлена називають інтерполяційним многочленом у формі Ньютона, а формулу (6) — *інтерполяційною формулою Ньютона.*

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

$$\begin{aligned}
p(x;\ f) = (\ldots((0 \cdot (x - x_n) + f[x_0, \ldots, x_n])(x - x_{n-1}) + \\
+ f[x_0, \ldots, x_{n-1}])(x - x_{n-2}) + \cdots + f[x_0, x_1])(x - x_0) + f(x_0).
\end{aligned} \tag{7}$$

Обчислення $p\,(x;\,f)$ для кожного x за цією схемою потребує n множень і $2n$ додавань та віднімань у той час, коли для обчислення значення цього многочлена, записаного в формі

$$\mathscr{P}(x) = u_n x^n + u_{n-1}x^{n-1} + \cdots + u_1 x + u_0,\ u_n \neq 0,$$

або в формі Лагранжа, потрібна кількість арифметичних дій порядку $O(n^2)$. Неважко помітити, що коли покласти $b_0 = 0$, $b_k = (x - x_{n-k+1}) \times b_{k-1} + f\,[x_0;\ \ldots;\ x_{n-k+1}]$, то $p_n\,(x;\,f) = b_{n+1}$. Це рекурентне співвідношення першого порядку легко програмується.

Вправа 3. Записати алгоритм, який за заданим x і масивом розділених різниць $f\,[x_0,\ \ldots,\ x_i]$, $i = \overline{0,\ n}$, обчислює $p_n\,(x;\,f)$ за $O\,(n)$ операцій множення і додавання.

Зауваження 1. За точністю розрахунків зручно слідкувати, оцінюючи швидкість спадання членів суми (6). Якщо вони спадають повільно, то на хорошу точність наближеної формули $f\,(x) \approx p\,(x;\,f)$ розраховувати важко (це стане зрозуміло далі). Якщо спадання швидке, то залишають лише ті члени, які більші, ніж величина допустимої похибки; тим самим визначають, скільки вузлів треба для розрахунку. Якщо потрібно контролювати точність розрахунків і залежно від цього вибирати кількість вузлів, то зручніша форма запису (6). Якщо ж число вузлів задане, то многочлен Ньютона зручніше обчислювати за схемою Горнера (7). Формула Ньютона зручна також тим, що в ній (байдуже, в якому порядку) перенумеровано вузли інтерполяції, тому при підключенні нового вузла до старого результату просто додається новий член.

Зауваження 2. Якщо $x_1 = x_0 + h$, $x_2 = x_0 + 2h$, ..., $x_n = x_0 + nh$, де h називається кроком таблиці, іноді використовують скінченні різниці, які визначаються так: 1) скінченні різниці першого порядку $\Delta f_i = f\,(x_{i+1}) - f\,(x_i)$, $i = \overline{0,\ n-1}$; 2) скінченні різниці k-го порядку $\Delta^k f_i = \Delta^{k-1} f_{i+1} - \Delta^{k-1} f_i$. Очевидно, що $\Delta^k f_i = k!\,h^k f\,[x_i,\ \ldots,\ x_{i+k}]$. Однак це скоріше данина традиціям, бо розділені різниці не менш зручні при розрахунках, ніж скінченні різниці.

Зауваження 3. Є багато інших форм запису інтерполяційного многочлена: Гаусса, Грегорі, Стірлінга, Лапласа — Евереттa та ін. Ці форми запису розраховані на певні окремі розміщення вузлів і ті переваги, які вони мають, часто не суттєві при розрахунках на ЕОМ.

Приклад. Побудувати інтерполяційний многочлен для функції $y = f\,(x)$, заданої таблицею:

x	0	1	2	3	5
$f\,(x)$	1	0	2	1	4

Розв'язання. Оскільки кількість вузлів дорівнює 5, інтерполяційний многочлен буде мати степінь 4. Складаємо таблицю розділених різниць:

x_i	$f\,(x_i)$	$f\,[x_i,\ x_k]$	$f\,[x_j,\ x_k,\ x_n]$	$f\,[x_j,\ x_k,\ x_m,\ x_n]$	$f\,[x_j,\ x_k,\ x_n,\ x_m,\ x_p]$
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

За таблицею розділених різниць складаємо інтерполяційний многочлен у формі Ньютона:

$$p_4\,(x) = 1 + (-1)\,x + \frac{3}{2}\,x\,(x-1) + (-1)\,x\,(x-1)\,(x-2) +$$

$$+ \frac{19}{60}\,x\,(x-1)\,(x-2)\,(x-3) = 1 - x + \frac{3}{2}\,x\,(x-1) - x\,(x-1)\,(x-2) +$$

$$+ \frac{19}{60}\,x\,(x-1)\,(x-2)\,(x-3).$$

2.3. Розділені різниці та інтерполювання з кратними вузлами

Нехай потрібно побудувати многочлен $p_{s-1}\,(x;\,f)$ степеня $s-1$, який задовольняє умови

$$p_{s-1}\,(x_1;\,f) = f\,(x_1),\ \ldots,\ p_{s-1}^{(m_1-1)}\,(x_1,\,f) = f^{(m_1-1)}\,(x_1),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad (1)$$
$$p_{s-1}\,(x_n;\,f) = f\,(x_n),\ \ldots,\ p_{s-1}^{(m_n-1)}\,(x_n;\,f) = f^{(m_n-1)}\,(x_n),$$

де x_i — різні вузли, $s = m_1 + \ldots + m_n$. Такий многочлен називають інтерполяційним многочленом з кратними вузлами, або *многочленом Ерміта,* а числа $m_1,\ \ldots,\ m_n$ — кратностями вузлів $x_1,\ \ldots,\ x_n$ відповідно.

Теорема 1. *Інтерполяційний многочлен, який задовольняє умови* (1), *єдиний.*

Доведення. Припустимо від супротивного, що є два многочлени степеня $s-1$, які задовольняють умови (1). Тоді різниця їх $Q_{s-1}\,(x)$ має властивості

$$Q_{s-1}\,(x_1) = \cdots = Q_{s-1}^{(m_1-1)}\,(x_1) = 0,\ \ldots,\ Q_{s-1}\,(x_n) = \cdots =$$
$$= Q_{s-1}^{(m_n-1)}\,(x_n) = 0,$$

тобто точки $x_1,\ \ldots,\ x_n$ є нулями многочлена $Q_{s-1}\,(x)$ кратності $m_1,\ \ldots$ $\ldots,\ m_n$ відповідно. Це означає, що многочлен $Q_{s-1}\,(x)$ степеня $s-1$ має s нулів, отже, $Q_{s-1}\,(x) \equiv 0$.

Теорему доведено.

Далі припускатимемо, що $f(x)$ неперервно диференційовна s разів. Побудуємо явно многочлен $p_{s-1}(x; f)$, доводячи тим самим і його існування.

Розглянемо послідовність точок x_{ij}^{ε}, $i = \overline{1, n}$, $j = \overline{1, m_i}$, причому всі точки x_{ij}^{ε} різні, $x_i \leqslant x_{i1}^{\varepsilon} < x_{i2}^{\varepsilon} < \ldots < x_{i,m_i}^{\varepsilon} < x_{i+1}$, $x_{ij}^{\varepsilon} \to x_i$ при $\varepsilon \to 0$. Зокрема, можна взяти $x_{ij}^{\varepsilon} = x_i + (j-1)\varepsilon$. Побудуємо інтерполяційний многочлен $p_{s-1}^{\varepsilon}(x; f)$ степеня $s-1$, який збігається з $f(x)$ в точках x_{ij}^{ε}, використовуючи при цьому таблицю розділених різниць

$$
\begin{array}{llll}
f(x_{11}^{\varepsilon}), & & & \\
f(x_{12}^{\varepsilon}) & f[x_{11}^{\varepsilon}, x_{12}^{\varepsilon}], & & \\
f(x_{13}^{\varepsilon}) & f[x_{12}^{\varepsilon}, x_{13}^{\varepsilon}] & f[x_{11}^{\varepsilon}, x_{12}^{\varepsilon}, x_{13}^{\varepsilon}], & \\
\ldots & \ldots & \cdot & \\
f(x_{1m_1}^{\varepsilon}) & f[x_{1m_1-1}^{\varepsilon}, x_{1m_1}^{\varepsilon}], & \cdot & \\
f(x_{21}^{\varepsilon}) & f[x_{1m_1}^{\varepsilon}, x_{21}^{\varepsilon}], & \cdot & \\
\ldots & \ldots & \cdot & \\
f(x_{nm_n}^{\varepsilon}) & f[x_{n,m_{n-1}}^{\varepsilon}, x_{n,m_n}^{\varepsilon}] & \cdots & f[x_{11}^{\varepsilon}, x_{12}^{\varepsilon}, \ldots, x_{nm_n}^{\varepsilon}].
\end{array}
\tag{2}
$$

Запишемо інтерполяційну формулу Ньютона з розділеними різницями

$$
p_{s-1}^{\varepsilon}(x; f) = A_0^{\varepsilon} + A_1^{\varepsilon}(x - x_{11}^{\varepsilon}) + A_2^{\varepsilon}(x - x_{11}^{\varepsilon})(x - x_{12}^{\varepsilon}) + \cdots + A_{s-1}^{\varepsilon}(x - x_{11}^{\varepsilon}) \ldots (x - x_{nm_n-1}^{\varepsilon}),
$$

де

$$
A_0^{\varepsilon} = f(x_{11}^{\varepsilon}),\ A_1^{\varepsilon} = f[x_{11}^{\varepsilon}, x_{12}^{\varepsilon}],\ A_2^{\varepsilon} = f[x_{11}^{\varepsilon}, x_{12}^{\varepsilon}, x_{13}^{\varepsilon}],\ \ldots,
$$

$$
A_{s-1}^{\varepsilon} = f[x_{11}^{\varepsilon}, x_{12}^{\varepsilon}, \ldots, x_{n,m_n}^{\varepsilon}].
$$

Подамо розділені різниці з кратними вузлами через похідні від функції f. Для цього скористаємося формулою (4) (п. 2.2):

$$
f[x, x_1, \ldots, x_n] = \frac{f(x)}{(x-x_1)(x-x_2)\ldots(x-x_n)} + \frac{f(x_1)}{(x_1-x)(x_1-x_2)\ldots(x_1-x_n)} + \cdots + \frac{f(x_n)}{(x_n-x)(x_n-x_1)\ldots(x_n-x_{n-1})}.
$$

Звідси

$$
f(x) = L_{n-1}(x) + R_{n-1}(x), \tag{3}
$$

де

$$
L_{n-1}(x) = f(x_1)\frac{(x-x_2)(x-x_3)\ldots(x-x_n)}{(x_1-x_2)\ldots(x_1-x_n)} + \cdots + f(x_n)\frac{(x-x_1)\ldots(x-x_{n-1})}{(x_n-x_1)\ldots(x_n-x_{n-1})}
$$

— інтерполяційний многочлен у формі Лагранжа, побудований по вузлах $x_1, \ldots, x_n$,

$$
R_{n-1}(x) \equiv f(x) - L_{n-1}(x) = (x-x_1)\ldots(x-x_n) f[x, x_1, \ldots, x_n] \tag{4}
$$

— залишковий член. Покладемо

$$
\varphi(z) = f(z) - L_{n-1}(z) - K\omega_n(z),
$$

де $\omega_n(z) = (z-x_1)\ldots(z-x_n)$. Виберемо K з умови $\varphi(x) = 0$, де $x \in [y_1, y_2]$ — точка, в якій оцінюється $R_{n-1}(x) = f(x) - L_{n-1}(x)$, $x \neq x_i$, $i = \overline{1, n}$, $y_1 = \min(x_1, \ldots, x_n)$, $y_2 = \max(x_1, \ldots, x_n)$. Очевидно,

$$
K = [f(x) - L_{n-1}(x)]/\omega_n(x),
$$

причому при такому виборі K функція $\varphi(z)$ перетворюється на нуль в $n+1$ точках: $x, x_1, \ldots, x_n$. Застосовуючи послідовно теорему Ролля, переконуємось, що існує точка $\xi \in [y_1, y_2]$ така, що

$$
\varphi^{(n)}(\xi) = f^{(n)}(\xi) - Kn! = 0,
$$

звідки

$$
K = \frac{f^{(n)}(\xi)}{n!}, \quad \xi \in [y_1, y_2].
$$

Співвідношення $\varphi(x) = 0$ тепер можна записати у вигляді

$$
R_{n-1}(x) \equiv f(x) - L_{n-1}(x) = \frac{f^{(n)}(\xi)\,\omega_n(x)}{n!}, \quad \xi \in [y_1, y_2]. \tag{5}
$$

Порівнюючи (4) і (5), дістаємо формулу

$$
f[x, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}, \quad \xi \in [y_1, y_2]. \tag{6}
$$

Таким чином,

$$
f[x_{il}^{\varepsilon}, \ldots, x_{im}^{\varepsilon}] = \frac{f^{(m-l)}(x_{ilm}^{\varepsilon})}{(m-l)!}, \tag{7}
$$

де x_{ilm}^{ε} міститься в найменшому проміжку, що містить всі точки $x_{il}^{\varepsilon}, \ldots, x_{im}^{\varepsilon}$. Перейшовши в (7) до границі при $\varepsilon \to 0$, дістанемо

$$
\lim_{\varepsilon \to 0} f[x_{il}^{\varepsilon}, \ldots, x_{im}^{\varepsilon}] = \frac{f^{(m-l)}(x_i)}{(m-l)!}. \tag{8}
$$

Це означає, що всі різниці таблиці (2) вигляду $f[x^{\varepsilon}_{il}, \ldots, x^{\varepsilon}_{im}]$ при $\varepsilon \to 0$ мають границі, які природно позначити

$$f[\underbrace{x_i, \ldots, x_i}_{m-l+1 \text{ раз}}],$$

причому з (8) випливає

$$f[\underbrace{x_i, \ldots, x_i}_{p+1 \text{ раз}}] = \frac{f^{(p)}(x_i)}{p!} \quad \forall p. \tag{9}$$

Покажемо, що всі різниці, які входять в таблицю (2), мають границі. Для цього застосуємо індукцію по порядку різниць. Для різниць нульового порядку, що збігаються із значенням функції, твердження очевидне. Нехай воно доведене для різниць порядку $q-1$.

Подамо кожну з різниць порядку q через різниці порядку $q-1$:

$$f[x^{\varepsilon}_{il}, \ldots, x^{\varepsilon}_{jp}] = \frac{f[x^{\varepsilon}_{i,l+1}, \ldots, x^{\varepsilon}_{jp}] - f[x^{\varepsilon}_{il}, \ldots, x^{\varepsilon}_{j,p-1}]}{x^{\varepsilon}_{jp} - x^{\varepsilon}_{il}}, \tag{10}$$

де для визначеності розглядається випадок $l < m_i$, $p > 1$, $i \neq j$. При $j = i$ існування границі різниць порядку q випливає з (8). При $j \neq i$ границя знаменника в правій частині (10) дорівнює $x_j - x_i \neq 0$, а границя чисельника існує за припущенням індукції. Отже, існує і границя лівої частини в (8), яку позначимо так:

$$f[\underbrace{x_i, \ldots, x_i}_{m_i-l+1}, \underbrace{x_{i+1}, \ldots, x_{i+1}}_{m_i+1}, \ldots, \underbrace{x_j, \ldots, x_j}_{p}].$$

Твердження індукції доведено. Аналогічно розглядаються інші випадки. Після переходу до границі таблиця (2) переходить у таку:

$$\begin{array}{lllll}
\left\{\begin{array}{l} f(x_1), \\ f(x_1) \quad f[x_1, x_1], \\ f(x_1) \quad f[x_1, x_1] \quad f[x_1, x_1, x_1], \\ \ldots \quad \ldots \quad \ldots \\ f(x_1) \quad f[x_1, x_1] \quad f[x_1, x_1, x_2], \ldots \end{array}\right. & & & & \\
\left.\begin{array}{llll} f(x_2) & f[x_1; x_2] & f[x_1, x_2, x_2], & \\ \ldots & \ldots & \ldots & \\ f(x_n) & f[x_n, x_n] & f[x_n, x_n, x_n] & \ldots f[\underbrace{x_1, \ldots, x_1}_{m_1}, \ldots, \underbrace{x_n, \ldots, x_n}_{m_n}]. \end{array}\right.
\end{array} \tag{10'}$$

Вона заповнюється за допомогою формули (9) і рекурентного співвідношення при $i \neq j$

$$f[\underbrace{x_i, \ldots, x_i}_{m_i-l+1}, \ldots, \underbrace{x_j, \ldots, x_j}_{p}] =$$

$$= \frac{f[\overbrace{x_i, \ldots, x_i}^{m_i-l}, \ldots, \overbrace{x_j, \ldots, x_j}^{p}] - f[\overbrace{x_i, \ldots, x_i}^{m_i-l+1}, \ldots, \overbrace{x_j, \ldots, x_j}^{p-1}]}{x_j - x_i}, \tag{11}$$

яке випливає з (10). Звідси бачимо, що коефіцієнти A^{ε}_k при $\varepsilon \to 0$ мають границі A_k, де

$$A_0 = f(x_1), \quad A_1 = f'(x_1), \quad A_2 = \frac{f''(x_1)}{2}, \ldots,$$
$$A_{s-1} = f[x_1, \ldots, x_1, \ldots, x_n, \ldots, x_n]. \tag{12}$$

Таким чином, многочлени $p^{\varepsilon}_{s-1}(x; f)$ прямують при $\varepsilon \to 0$ до деякого многочлена

$$p_{s-1}(x; f) = A_0 + A_1(x - x_1) + A_2(x - x_1)^2 + \ldots$$
$$\ldots + A_{s-1}(x - x_1)^{m_1} \ldots (x - x_{n-1})^{m_n - 1}(x - x_n)^{m_n - 1} =$$
$$= f(x_1) + f[x_1, x_1](x - x_1) + f[x_1, x_1, x_1](x - x_1)^2 + \ldots. \tag{13}$$

Враховуючи (12), многочлен (13) можна записати у вигляді

$$p_{s-1}(x; f) = \sum_{i=1}^{m_1} \frac{f^{(i-1)}(x_1)}{(i-1)!}(x - x_1)^{i-1} + (x - x_1)^{m_1} F(x - x_2, x - x_3, \ldots$$
$$\ldots, x - x_n),$$

де $F(t_1, \ldots, t_{n-1})$ — деякий многочлен від $t_1, \ldots, t_{n-1}$. Звідси випливає, що він задовольняє умови, які задані в точці x_1. В силу єдності інтерполяційного многочлена він не зміниться при перепозначенні x_1 на x_j і x_j на x_1 для довільного j. Тому граничний многочлен буде задовольняти задані умови в будь-якій точці x_j і, таким чином, він є шуканим многочленом

$$p(x; f) \equiv p_{s-1}(x; f).$$

Приклад. Побудувати інтерполяційний многочлен, який задовольняє умови, подані в таблиці

x	$f(x)$	$f'(x)$	$f''(x)$
0	3	1	
1	0		
2	1	2	1
3	5		

Р о з в' я з а н н я. Як випливає з таблиці, вузол $x_1 = 0$ має кратність 2, $x_2 = 1$ — кратність 1, $x_3 = 2$ — кратність 3, $x_4 = 3$ — кратність 1; отже, степінь многочлена буде 6. Складаємо таблицю розділених різниць вигляду (10′):

$$\begin{array}{llllll} 0 & (3), & & & & \\ 0 & (3) & (1), & & & \\ 1 & (0) & -3 & -4, & & \\ 2 & (1) & 1 & 2 & 3, & \\ 2 & (1) & (2) & -1/2 & -7/4, & \\ 2 & (1) & (2) & (1/2) & 0 & 7/8, \\ 3 & (5) & 4 & 2 & 3/2 & 1/3 \quad -13/72. \end{array}$$

За таблицею записуємо інтерполяційний многочлен

$$p_7(x; f) = 3 + 1 \cdot x - 4x^2 + 3x^2(x-1) - \frac{7}{4}x^2(x-1)(x-2) +$$

$$+ \frac{7}{8}x^2(x-1)(x-2)^2 - \frac{13}{72}x^2(x-1)(x-2)^3.$$

Дані, які згідно з формулою (9) беруться з вихідної таблиці, взято в дужки, інші — обчислюються за формулою (11).

2.4. Аналіз похибки інтерполяційних формул

При обчисленні $f(x)$ за інтерполяційним многочленом, тобто за формулою $f(x) \approx p(x; f)$, існують три джерела похибок: 1) похибка методу, за умови, що всі обчислення проводяться точно, а $x \neq x_i$; 2) неусувна похибка за рахунок того, що значення $f(x_i)$ виміряні або обчислені не точно; 3) похибка заокруглень, яка виникає при обчисленні $p(x; f)$ на ЕОМ через скінченність розрядної сітки.

2.4.1. Похибка методу для функцій з класу $C^n[a, b]$. Нехай функція $f(x)$ визначена в k вузлах інтерполяції $x_i \in [a, b]$, $i = \overline{1, k}$, а $T(x)$ — її інтерполююча функція з кратними вузлами (не обов'язково многочлен), тобто

$$f^{(j)}(x_i) = T^{(j)}(x_i), \quad i = \overline{1, k}, \quad j = \overline{0, m_i - 1}, \quad \sum_{i=1}^{k} m_i = k. \tag{1}$$

Теорема 1. *Нехай $f(x) \in C^n[a, b]$, $T(x)$ — її інтерполююча функція з кратними вузлами інтерполяції, причому $T(x) \in C^a[a, b]$. Якщо існує функція $g(x) \in C^n[a, b]$ така, що*

$$g^{(j)}(x_i) = 0, \quad i = \overline{1, k}, \quad j = \overline{0, m_i - 1}, \quad g^{(n)}(x) \neq 0 \quad \forall x \in [a, b],$$
$$g(x) \neq 0 \quad \forall x \neq x_i, \quad i = \overline{1, k}, \tag{2}$$

то для будь-якого $x \neq x_i$ залишковий член інтерполяційної формули визначається співвідношенням

$$R(x) = f(x) - T(x) = \frac{f^{(n)}(\xi) - T^n(\xi)}{g^{(n)}(\xi)} g(x), \tag{3}$$

де ξ — деяка точка з відрізка $[y_1, y_2]$, $y_1 = \min(x, x_1, \ldots, x_k)$, $y_2 = \max(x, x_1, \ldots, x_k)$.

Д о в е д е н н я. Розглянемо функцію

$$F(t) = f(t) - T(t) - \lambda g(t). \tag{4}$$

Неважко помітити, що $F(t) \in C^{(n)}[a, b]$, $F^{(j)}(x_i) = 0$, $j = \overline{0, m_i - 1}$, $i = \overline{1, k}$. Виберемо λ з умови $F(x) = 0$, тобто

$$\lambda = \frac{f(x) - T(x)}{g(x)}. \tag{5}$$

Тоді функція $F(t)$ перетворюється на нуль в $n + 1$ точці (з урахуванням кратності) відрізка $[y_1, y_2]$. За допомогою узагальненої теореми Ролля знаходимо, що існує точка $\xi \in [y_1, y_2]$, для якої $F^{(n)}(\xi) = 0$, звідки з урахуванням (4), (5) дістаємо твердження теореми.

Розглянемо окремі випадки формули (3): 1) нехай всі вузли прості (не кратні) і $g(x) = \omega_n(x) = \prod_{i=1}^{n}(x - x_i)$, тоді якщо $T(x)$ є алгебраїчним інтерполяційним $(n-1)$-го степеня многочленом (у формі Лагранжа, Ньютона чи іншій), то дістаємо вже відому нам формулу (5) (п. 2.3); 2) нехай $T(x) = p_{s-1}(x; f)$ — інтерполяційний алгебраїчний многочлен $(s-1)$-го степеня з кратними вузлами і

$$g(x) = \prod_{i=1}^{n}(x - x_i)^{m_i}, \quad \sum_{i=1}^{n} m_i = s;$$

тоді

$$R(x) = f(x) - T(x) = \frac{f^{(n)}(\xi)}{n!} \prod_{i=1}^{n}(x - x_i)^{m_i}; \tag{6}$$

3) якщо $x_1 = x_2 = \ldots = x_n$, то в попередньому прикладі $T(x)$ — многочлен Тейлора (відрізок ряду Тейлора), а формула (6) — залишковий член формули Тейлора в формі Коші.

2.4.2. Мінімізація похибки методу за рахунок вибору вузлів інтерполювання. Многочлени Чебишева першого роду. Якщо $f(x) \in C^{n+1}[a, b]$, то похибка методу для інтерполяційного многочлена за $n + 1$ вузлами (однократні вузли) має вигляд

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)\ldots(x - x_n), \quad \xi \in [a, b] \tag{7}$$

і її величина, очевидно, залежить від положення точки x і розміщення вузлів інтерполювання $x_0, \ldots, x_n$. Покладаючи тут $M_{n+1} = \sup_{x \in [a,b]} |f^{(n+1)}(x)|$, дістанемо

$$|R_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!} |(x - x_0)\ldots(x - x_n)|. \tag{8}$$

Як правило, для величини M_{n+1} маємо лише оцінку, але величину $\omega_{n+1}(x) = (x - x_0) \dots (x - x_n)$ ми можемо змінювати вибором вузлів інтерполювання. Природно вибрати $x_0, x_1, \dots, x_n$ так, щоб величина $\sup\limits_{x\in[a,b]} |\omega_{n+1}(x)|$ була мінімальною. Скористаємося для цього деякими властивостями многочленів Чебишева першого роду, які визначаються співвідношеннями

$$T_n(x) = \cos[n \arccos x], \quad |x| \leqslant 1, \quad n = 0, 1, \dots,$$

або

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n = 1, 2, \dots, \quad T_0(x) = 1, \quad T_1(x) = x.$$

Зазначимо, що коефіцієнт при x^n у $T_n(x)$ дорівнює 2^{n-1}. Неважко помітити, що нулі многочлена $T_n(x)$ визначаються формулою

$$x_m = \cos\frac{2m+1}{2n}\pi, \quad m = 0, 1, \dots, n-1,$$

а максимум $|T_n(x)|$ на відрізку $[-1, 1]$ дорівнює одиниці і досягається в $n+1$ точках

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

Звідси випливає, що коли за відрізок інтерполювання взяти $[-1, 1]$, а за вузли x_i, $i = \overline{0, n}$,— корені многочлена Чебишева $T_{n+1}(x)$, то

$$\omega_{n+1}(x) = 2^{-n}T_{n+1}(x), \quad \sup_{x\in[-1,1]} |\omega_{n+1}(x)| = 2^{-n}.$$

Покажемо, що має місце така лема.

Лема 1. *Серед всіх многочленів степеня n із старшим коефіцієнтом 1 многочлен*

$$\overline{T}_n(x) = 2^{-(n-1)}T_n(x)$$

найменше відхиляється від нуля на $[-1, 1]$ ($\overline{T}_n(x)$ називаються многочленами, що найменше відхиляються від нуля).

Доведення. Покажемо, що для будь-якого многочлена $p_n(x)$ із старшим коефіцієнтом 1 справедлива нерівність

$$\sup_{x\in[-1,1]} |p_n(x)| \geqslant 2^{-(n-1)}.$$

Справді, в іншому разі різниця $2^{-(n-1)}T_n(x) - p_n(x)$ була б многочленом степеня $n-1$, який набуває в точках $x_m = \cos\frac{m\pi}{n}$, $m = \overline{0, n}$, позмінно додатних і від'ємних значень, а це означало б, що цей многочлен має n нулів. Дійшли суперечності. Лему доведено.

З леми 1 безпосередньо випливає така теорема.

Теорема 2. *Нехай відрізком інтерполювання є $[-1, 1]$. Тоді величина $\sup\limits_{x\in[-1,1]} |\omega_{n+1}(x)|$ набуває найменшого значення, якщо за вузли інтерполювання взяти корені многочлена Чебишева $T_{n+1}(x)$, тобто*

$$x_i = \cos\frac{2i+1}{2n+2}\pi, \quad i = \overline{0, n},$$

а оцінка (8) у цьому випадку набирає вигляду

$$|R_n(x)| \leqslant \frac{M_{n+1}}{2^n(n+1)!}. \tag{9}$$

Якщо інтерполювання проводиться на довільному відрізку $[a, b]$, то заміною

$$x = \frac{1}{2}[(b-a)z + (b+a)], \quad z = \frac{1}{b-a}[2x - b - a]$$

він переходить в $[-1, 1]$, при цьому корені $T_{n+1}(x)$ перейдуть в точки

$$x_m = \frac{1}{2}\left[(b-a)\cos\frac{2m+1}{2n+2}\pi + (b+a)\right], \quad m = \overline{0, n}. \tag{10}$$

Многочлен $\overline{T}_n^{[a,b]}(x) = (b-a)^n 2^{1-2n} T_n\left(\frac{2x-b+a}{b-a}\right)$ із старшим коефіцієнтом 1 найменше відхиляється від нуля на $[a, b]$, а оцінка (9) має вигляд

$$|R_n(x)| \leqslant \frac{M_{n+1}}{(n+1)!}\frac{(b-a)^{n+1}}{2^{2n+1}}, \quad x \in [a, b].$$

Зазначимо, що

$$\max_{[a,b]} |p_n(x)| \geqslant \max_{[a,b]} |\overline{T}_n^{[a,b]}(x)| = (b-a)^n 2^{1-2n}. \tag{11}$$

Зауваження 1. Вузли многочлена Чебишева розміщені порівняно рідко в середині відрізка і згущаються біля його кінців, а за межами відрізка многочлен $\omega_{n+1}(x)$ швидко зростає. За допомогою інтерполяційного многочлена можна обчислювати значення $f(x)$ також і для точок x, які лежать за крайніми вузлами інтерполяції. У цьому разі говорять про екстраполювання, а термін інтерполювання, як правило, вживають для x, що лежать між крайніми вузлами інтерполювання. Таким чином, у випадку екстраполювання вибір вузлів у вигляді (10) мало що дає. Більше того, інтерполювання з вузлами (10) досить громіздке, а виграш у точності невеликий. Тому цей спосіб інтерполювання частіше використовують для спеціальних цілей, наприклад, при побудові різних апроксимуючих формул.

2.4.3. Поведінка залишкового члена (при фіксованих вузлах) залежно від вибору точки інтерполювання. Розглянемо тепер випадок, коли вузли інтерполювання фіксовані і треба відповісти на запитання:

для яких проміжків зміни x залишковий член $R(x; f)$ набуватиме більших значень, а для яких — менших. Щоб відповісти на це запитання, зауважимо, що многочлен $\omega_{n+1}(x)$ перетворюється на нуль в точках $x_0, \ldots, x_n$ і десь між ними набуває поперемінно максимального і мінімального значень, причому абсолютні значення цих екстремумів дорівнюватимуть один одному лише при виборі вузлів у нулях многочлена Чебишева $T_{n+1}(x)$. В інших випадках на більшу похибку інтерполювання слід сподіватися біля більших за абсолютною величиною екстремумів. Далі обмежимося розглядом рівновіддалених вузлів, тобто

$$x_1 - x_0 = x_2 - x_1 = \cdots = x_n - x_{n-1} = h. \qquad (12)$$

Якщо позначити $t = h^{-1}(x - x_0)$, то

$$\omega_{n+1}(x) = \omega_{n+1}(x_0 + th) = h^{n+1} t(t-1)(t-2) \ldots (t-n).$$

Зауважимо, що функція $\varphi(t) = t(t-1) \ldots (t-n)$ є парною чи непарною відносно точки $t = \frac{n}{2}$ залежно від непарності чи парності n, що випливає із співвідношення

$$\varphi(t) = (-1)^{n+1} \varphi(n-t),$$

$$\varphi(t+1) = (t+1)\, t(t-1) \ldots (t+1-n) = \frac{t+1}{t-n} \varphi(t).$$

Якщо розбити відрізок $[0, n]$ на частини $[0, 1], [1, 2], \ldots, [n-1, n]$, то видно, що на кожному такому відрізку значення $\varphi(t)$ знаходять множенням відповідного значення на попередньому відрізку на $\frac{t+1}{t-n}$. Цей множник завжди від'ємний при $t \in (0, n)$, тому знаки функції $\varphi(t)$ чергуються при переході від інтервалу до інтервалу. Абсолютна величина множника менша за одиницю на проміжку $\left[0, \frac{n-1}{2}\right]$. Таким чином, екстремальні значення $\varphi(t)$ спадатимуть за абсолютною величиною до середини відрізка $[0, n]$ і потім в силу симетрії знову зростатимуть. За межами відрізка $[0, n]$ функція $\varphi(t)$ швидко зростає за абсолютною величиною. Приблизний вигляд многочлена $\omega_{n+1}(x)$ при $n = 5$ зображений на рис. 15.

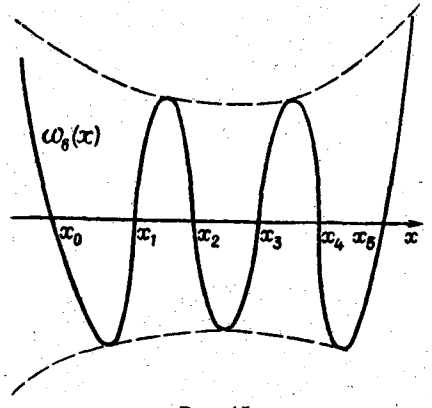


Рис. 15

Таким чином, можна дійти таких висновків:

1. При екстраполюванні слід чекати великої похибки.

2. При інтерполюванні для значень x, які лежать не близько до вузлів інтерполювання, точність буде більшою для середніх відрізків і меншою для крайніх.

3. Вигідно вибирати вузли так, щоб точка x попадала ближче до центра даної конфігурації вузлів, що забезпечить більшу точність.

2.4.4. Оцінка похибки інтерполяційного многочлена n-го степеня у випадку $f(x) \in C^{n+1}[0, 1]$ і рівномірного розміщення вузлів. Нехай вузли інтерполяції $x_i = x_0 + ih$, $i = \overline{0, n}$, розміщено рівномірно на $[0, 1]$. Запишемо інтерполяційний многочлен у формі Лагранжа

$$p_n(x; f) \equiv L_n(x) = \sum_{i=0}^{n} f(x_i) \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}, \qquad (13)$$

де $l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}$ — фундаментальні інтерполяційні многочлени Лагранжа. Як було показано, якщо $f(x) \in C^{n+1}[0, 1]$, то для $x \in [x_0, x_n]$ можливе таке зображення залишкового члена:

$$R_n(x) \equiv R_n(x; f) = f(x) - L_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)\, \omega_{n+1}(x), \qquad (14)$$

де $\xi \in (x_0, x_n)$, $\omega_{n+1}(x) = \prod_{j=0}^{n} (x - x_j)$.

Оцінимо $R_n(x)$ через степінь n. Насамперед зауважимо, що для $\omega_2(x) = (x - x_0)(x - x_1)$ максимальне значення модуля на відрізку $[x_0, x_1]$ досягається в середині відрізка і дорівнює $h^2/4$, тобто

$$|\omega_2(x)| \leqslant h^2/4, \quad x \in [x_0, x_1].$$

Доведемо, що

$$|\omega_{l+1}(x)| \leqslant \frac{h^{l+1} l!}{4}, \quad x \in [x_0, x_l]. \qquad (15)$$

Застосуємо метод математичної індукції. Для $l = 1$ ця нерівність доведена. Припустимо, що вона має місце для $l = k - 1$ і доведемо її для $l = k$.

Якщо $x \in [x_0, x_{k-1}]$, то

$$|\omega_{k+1}(x)| = |\omega_k(x)|\,|x - x_k| \leqslant \frac{h^k (k-1)!}{4} \cdot kh = \frac{h^{k+1} k!}{4}$$

і (15) доведено. У випадку $x \in [x_{k-1}, x_k]$ очевидні нерівності

$$|x - x_{k-1}|\,|x - x_k| \leqslant \frac{h^2}{4}, \quad |x - x_i| \leqslant (k-i)h, \quad i = \overline{0, k-1}.$$

Тому

$$|\omega_{k+1}(x)| = (x - x_0)(x - x_1) \ldots (x - x_{k-1})\,|x - x_k| \leqslant$$

$$\leqslant h^{k-1} k(k-1) \ldots 2 \cdot \frac{h^2}{4} = \frac{h^{k+1} k!}{4},$$

що і треба було довести. Тепер з (14) дістаємо

$$|f(x)-p_n(x;\ f)|\equiv|f(x)-L_n(x)|\leqslant\frac{h^{n+1}}{4\,(n+1)}\|f^{(n+1)}\|_{C[0,1]},\quad x\in(x_0,\ x_n).\tag{16}$$

2.4.5. Оцінка похибки інтерполяційного многочлена n-го степеня у випадку $f(x)\in C^{k+1}[0,1]$, $0\leqslant k<n$, і рівномірного розміщення вузлів. Доведемо справедливість наступного твердження.

Теорема 3. *Нехай $f(x)\in C^{k+1}[0,1]$, $0\leqslant k<n$, тоді*

$$|f(x)-p_n(x;\ f)|\equiv|f(x)-L_n(x)|\leqslant$$
$$\leqslant\frac{h^{k+1}}{(k+1)\,2^{k+2-n}}\|f\|_{C^{k+1}[0,1]},\quad x\in(x_0,\ x_n).\tag{17}$$

Доведення. Запишемо $L_n(x)$ у вигляді

$$L_n(x)=L_k(x)+(L_{k+1}(x)-L_k(x))+\ \cdots\ +(L_n(x)-L_{n-1}(x)).$$

Розглянемо $L_j(x)-L_{j-1}(x)$. Використовуючи (3) (п. 2.3), дістаємо

$$L_j(x)-L_{j-1}(x)=(x-x_0)(x-x_1)\ \ldots\ (x-x_{j-1})\,L_j[x,\ x_0,\ \ldots,\ x_{j-1}].$$

Оскільки $L_j(x)$ — многочлен степеня j, то розділена різниця j-го порядку $L_j[x,\ x_0,\ \ldots,\ x_{j-1}]$ є сталою, причому $L_j[x,\ x_0,\ \ldots,\ x_{j-1}]=L_j[x_0,\ x_1,\ \ldots,\ x_j]$. З іншого боку, оскільки многочлен $L_j(x)$ є інтерполяційним для функції $f(x)$ по вузлах $x_0,\ \ldots,\ x_j$, то

$$L_j[x_0,\ \ldots,\ x_j]=f[x_0,\ \ldots,\ x_j],$$

тобто

$$L_j(x)-L_{j-1}(x)=\omega_j(x)\,f[x_0,\ \ldots,\ x_j].$$

Тому

$$|f(x)-L_n(x)|\leqslant|f(x)-L_k(x)|+|L_{k+1}(x)-L_k(x)|+\ \cdots\ +$$
$$+|L_n(x)-L_{n-1}(x)|\leqslant\frac{h^{k+1}}{4\,(k+1)}\|f\|_{C^{k+1}[0,1]}+|f[x_0,\ \ldots,\ x_{k+1}]|\times$$
$$\times|\omega_{k+1}(x)|+\ \cdots\ +|f[x_0,\ \ldots,\ x_n]|\,|\omega_n(x)|.\tag{18}$$

Методом математичної індукції доведемо, що для функції $f(x)\in C^{k+1}[0,1]$ і для будь-якого $l=k+1,\ \ldots,\ n$ виконується нерівність

$$|f[x_0,\ \ldots,\ x_l]|\leqslant\frac{h^{k+1-l}}{2^{k+1-l}\,l!}\|f\|_{C^{k+1}[x_0,x_l]}.\tag{19}$$

Якщо $l=k+1$, то, як відомо,

$$f[x_0,\ \ldots,\ x_{k+1}]=\frac{f^{(k+1)}(\xi)}{(k+1)!},\quad \xi\in[x_0,\ x_{k+1}]$$

і оцінка (19) виконана. Припустимо, що вона має місце для цілого $l>k+1$ і доведемо її справедливість для $l+1$.

Нерівність

$$|f[x_1,\ \ldots,\ x_{l+1}]|\leqslant\frac{h^{k+1-l}}{2^{k+1-l}\,l!}\|f\|_{C^{k+1}[x_1,x_{l+1}]}$$

є наслідком нерівності (19) для функції $f(x+h)\equiv g(x)$, бо

$$|f[x_1,\ \ldots,\ x_{l+1}]|=|g[x_0,\ \ldots,\ x_l]|\leqslant\frac{h^{k+1-l}}{2^{k+1-l}\,l!}\|g\|_{C^{k+1}[x_0,x_l]}\leqslant$$
$$\leqslant\frac{h^{k+1-l}}{2^{k+1-l}\,l!}\|f\|_{C^{k+1}[x_1,x_{l+1}]}.$$

За означенням розділеної різниці

$$f[x_0,\ \ldots,\ x_{l+1}]=\frac{f[x_1,\ \ldots,\ x_{l+1}]-f[x_0,\ \ldots,\ x_l]}{x_{l+1}-x_0},$$

звідки

$$|f[x_0,\ \ldots,\ x_{l+1}]|\leqslant\frac{h^{k+1-l}}{2^{k+1-l}\,l!}\,\frac{\|f\|_{C^{k+1}[x_1,x_{l+1}]}+\|f\|_{C^{k+1}[x_0,x_l]}}{(l+1)\,h}\leqslant$$
$$\leqslant\frac{h^{k+1-(l+1)}}{2^{k+1-(l+1)}\,(l+1)!}\|f\|_{C^{k+1}[x_0,x_{l+1}]}$$

і оцінку (19) доведено.

Продовжимо оцінку (18), використовуючи (19) і (15):

$$|f(x)-L_n(x)|\leqslant\frac{h^{k+1}}{4\,(k+1)}\|f\|_{C^{k+1}[0,1]}+\frac{h^{k+1}k!}{4\,(k+1)!}\|f\|_{C^{k+1}[0,1]}+$$
$$+\frac{h^{-1}}{2^{-1}\,(k+2)!}\,\frac{h^{k+2}\,(k+1)!}{4}\|f\|_{C^{k+1}[0,1]}+$$
$$+\frac{h^{-2}}{2^{-2}\,(k+3)!}\,\frac{h^{k+3}\,(k+2)!}{4}\|f\|_{C^{k+1}[0,1]}+\ \cdots\ +$$
$$+\ \cdots\frac{h^{k+1-n}}{2^{k+1-n}n!}\,\frac{h^n\,(n-1)!}{4}\|f\|_{C^{k+1}[0,1]}\leqslant$$
$$\leqslant\frac{h^{k+1}}{4}\|f\|_{C^{k+1}[0,1]}\left(\frac{2^0}{k+1}+\frac{2}{k+2}+\ \cdots\ +\frac{2^{n-k-1}}{n}\right)<$$
$$<\frac{h^{k+1}}{4}\|f\|_{C^{k+1}[0,1]}\frac{1}{k+1}\,(1+2+\ \cdots\ +2^{n-k-1})=$$
$$=\frac{h^{k+1}}{4}\|f\|_{C^{k+1}[0,1]}\frac{1}{k+1}\,\frac{2^{n-k}-1}{2-1}<\frac{h^{k+1}}{(k+1)\,2^{k-n+2}}\|f\|_{C^{k+1}[0,1]},$$

що і треба було довести.

2.4.6. Оцінка похибки інтерполяційного многочлена n-го степеня у випадку $f(x) \in W_2^m(\Omega)$, $\Omega = (0, nh)$, $1 \leqslant m$, і рівномірного розміщення вузлів. Перш ніж перейти до оцінки залишкового члена, доведемо таке допоміжне твердження.

Лема 2. *Для $x \in [x_0, x_n]$ виконується нерівність*

$$\sum_{i=0}^{n} \left| \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \right| \leqslant 2^n.$$

Доведення. Зробимо заміну $t = (x - x_0)/h$, тоді $t \in [0, n]$, причому

$$\left| \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \right| = \left| \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x_0 + th - x_j}{x_0 + ih + x_0 - ih} \right| = \left| \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x_0 + th - x_0 - jh}{(i - j)\,h} \right| =$$

$$= \left| \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{t - j}{i - j} \right| = \left| \frac{t(t-1) \ldots (t - i + 1)(t - i - 1) \ldots (t - n)}{i(i-1) \ldots (i - i + 1)(i - i - 1) \ldots (i - n)} \right| =$$

$$= \frac{n(n-1) \ldots (n - i - 1)}{i!} \cdot \frac{t(t-1) \ldots (t - i + 1)(t - i - 1) \ldots (t - n)}{n(n-1) \ldots (n - i - 1)(n - i) \ldots 1} =$$

$$= C_n^i \frac{t(t-1) \ldots (t - i + 1)(t - i - 1) \ldots (t - n)}{n!}.$$

Неважко помітити, що дріб не перевищує одиницю, а тому

$$\sum_{i=0}^{n} \left| \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \right| \leqslant \sum_{i=0}^{n} C_n^i = 2^n.$$

Лему доведено.

Теорема 4. *Нехай $f(x) \in W_2^m(\Omega)$, $m \geqslant 1$, $\Omega = (0, nh)$, n — фіксоване, незалежне від h, $nh \leqslant 1$ і $L_n(x)$ — інтерполяційний многочлен степеня n, побудований по вузлах $x_i = ih$, $i = \overline{0, n}$. Тоді справджується*

$$|R_n(x)| \equiv |R_n(x; f)| = |f(x) - L_n(x)| \leqslant$$
$$\leqslant \overline{M}(m, n)\, n^{k-1/2} h^{k-1/2} |f|_{W_2^k(\Omega)}, \qquad (20)$$

де

$$k = \begin{cases} m, & \text{якщо } m \leqslant n \\ n + 1, & \text{якщо } m > n \end{cases},$$

$\overline{M}(m, n)$ — деяка стала, що залежить від m та n,

$$|f|_{W_2^k(\Omega)} = \left(\int_0^{nh} (f^{(k)}(x))^2\, dx \right)^{1/2}.$$

Доведення. За допомогою лінійної заміни $\xi = snh$ відобразимо $\overline{\Omega} = [0, nh]$ на відрізок $\overline{E} = [0, 1]$, $\xi \in \overline{\Omega}$, $s \in \overline{E}$, причому $E = (0, 1)$. Тоді інтерполяційний многочлен $L_n(x)$ запишеться у вигляді

$$L_n(x) = \tilde{L}_n(s) = \sum_{i=0}^{n} f(ih)\, l_i(x) = \sum_{i=0}^{n} \tilde{f}\left(\frac{i}{n}\right) l_i(x),$$

де $\tilde{f}(s) = f(\xi) = f(snh)$, $l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}$. Залишковий член $R_n(x) \equiv R_n(x; f) = f(x) - L_n(x) = \tilde{f}(s) - L_n(s) \equiv R_n(s; \tilde{f})$ при фіксованому x є лінійним функціоналом від $\tilde{f}$ в просторі $W_2^m(E)$. Спершу покажемо, що він обмежений у $W_2^m(E)$. В силу леми 2 і теореми вкладення маємо

$$|R_n(x; f)| = |R_n(s; \tilde{f})| \leqslant \|\tilde{f}\|_{C[0,1]} \left(1 + \sum_{i=0}^{n} |l_i(x)|\right) \leqslant (2^n + 1) \|\tilde{f}\|_{C[0,1]} \leqslant$$
$$\leqslant 2(2^n + 1) \|\tilde{f}\|_{W_2^1(0,1)} \leqslant 2(2^n + 1) \|\tilde{f}\|_{W_2^m(0,1)}.$$

З побудови інтерполяційного многочлена випливає, що $R_n(x; p) = R_n(s; \tilde{p}) = 0$ для довільного многочлена p не вище n-го степеня. Тому в силу леми Брембла — Гільберта маємо

$$|R_n(x; f)| = |R_n(s; \tilde{f})| \leqslant \overline{M}(m, n) |\tilde{f}|_{W_2^k(E)}, \qquad (21)$$

де k та $\overline{M}(m, n)$ мають той самий смисл, що й раніше.

Переходячи знову до змінної ξ, дістаємо

$$|\tilde{f}|_{W_2^k(E)} = \left(\int_0^1 [\tilde{f}^{(k)}(s)]^2\, ds \right)^{1/2} = (n^{-1}h^{-1}(nh)^{2k} \int_0^{nh} [f^{(k)}(x)]^2\, dx \Big)^{1/2} =$$
$$= n^{k-1/2} h^{k-1/2} |f|_{W_2^k(\Omega)},$$

звідки і дістаємо твердження теореми 4.

Зауваження 2. З доведення теореми 4 стає зрозуміло, як знайти оцінку залишкового члена у випадку, коли $f(x) \in C^m(\overline{\Omega})$, $m \geqslant 1$. У цьому разі маємо

$$|f|_{\tilde{W}_2^k(E)} = \left(n^{-1}h^{-1}(nh)^{2k} \int_0^{nh} [f^{(k)}(x)]^2\, dx \right)^{1/2} \leqslant (nh)^k \|f\|_{C^k[0,nh]},$$

що разом з (21) приводить до оцінки

$$|R_n(x; f)| \leqslant \overline{M}(m, n)(nh)^k \|f\|_{C^k(\overline{\Omega})}. \qquad (22)$$

З означення норм просторів $W_2^m(\Omega)$ та $C^m(\bar{\Omega})$ витікають оцінки

$$|f|_{W_2^k(\Omega)} \leqslant \|f\|_{W_2^k(\Omega)} \leqslant \|f\|_{W_2^m(\Omega)},$$

$$\|f\|_{C^k(\bar{\Omega})} \leqslant \|f\|_{C^m(\bar{\Omega})}$$

і оцінки (20), (22) можна записати у вигляді

$$|R_n(x;\ f)| \leqslant \overline{M}(m,\ n) \begin{cases} (nh)^{k-1/2}\|f\|_{W_2^m(\Omega)}, & f \in W_2^m(\Omega), \\ (nh)^k \|f\|_{C^m(\bar{\Omega})}, & f \in C^m(\bar{\Omega}), \end{cases} \tag{23}$$

де $m \geqslant 1$. Звідси видно, що для функції $f \in W_2^m(\Omega)$, тобто для функції меншої гладкості, ніж $C^m(\bar{\Omega})$, порядок залишкового члена інтерполяційного многочлена по h менший на 1/2. Порівняючи цю оцінку з оцінкою залишкового члена (17), бачимо, що при $f(x) \in C^m[0,\ 1]$, $1 \leqslant m < n + 1$ порядок h в них однаковий, але сталий множник $Mc = 2^n/(m2^{m+1})$ в (17) менший, ніж множник $M_W = \overline{M}(m,\ n)\, n^m$ в нерівності (23). Якщо точка x, в якій потрібно за допомогою інтерполяційного многочлена $p_n(x;\ f)$ фіксованого степеня n наближено обчислити значення функції $f(x) \in C^m[0,\ 1]$ (m фіксоване) за відомими її значеннями $f(x_i)$, $i = \overline{0,\ n}$, то при $h \to 0$

$$|R_n(x;\ f)| = O(h^k), \quad m \geqslant 1,$$

а для $f(x) \in W_2^m(0,\ 1)$

$$|R_n(x;\ f)| = O(h^{k-1/2}), \quad m \geqslant 1, \quad k = \begin{cases} m, & \text{якщо } m \leqslant n, \\ n+1, & \text{якщо } m > n. \end{cases}$$

Таким чином, зменшуючи h при фіксованих m, n, x, за допомогою $p_n(x;\ f)$, можна обчислити $f(x)$ як завгодно точно.

Оцінка (23) також показує, що при фіксованому x і фіксованій кількості вузлів n точність формули $f(x) \approx p_n(x;\ f)$ збільшується (по h) з ростом m, тобто з ростом гладкості функції, коли $m \leqslant n$. Якщо ж $m \geqslant n + 1$, то зростання гладкості не приводить до збільшення порядку точності, який для всіх $m \geqslant n + 1$ залишається рівним $n + 1$ для $f(x) \in C^m(\bar{\Omega})$ і $n + 1/2$ для $f(x) \in W_2^m(\Omega)$.

Приклад 1. За таблицею значень функції $f(x) = 1/x$, користуючись лінійною інтерполяцією, знайти $f(2,\ 718)$:

x	$f(x)$	$f(x_i,\ x_j)$
2,70	0,3704	
2,72	0,3676	—0,14
2,74	0,3650	—0,13

Р о з в' я з а н н я. Для даного випадку виберемо $x_0 = 2{,}70$, $x_1 = 2{,}72$, $h = x_1 - x_0 = 0{,}02$. Інтерполяційний многочлен Ньютона першого степеня має вигляд

$$p(x;\ f) = f(x_0) + (x - x_0)\, f(x_0,\ x_1),$$

тому

$$\frac{1}{2{,}718} = 0{,}3704 - 0{,}14 \cdot 0{,}018 = 0{,}3679.$$

Оскільки $f''(x) = 2/x^3$, то для залишкового члена

$$R_1(x) = \frac{f''(\xi)}{2!}(x - x_0)(x - x_1), \quad 2{,}70 < \xi < 2{,}72$$

і далі

$$|R_1(2{,}718)| < \frac{0{,}018 \cdot 0{,}02}{(2{,}7)^3} < 0{,}2 \cdot 10^{-4},$$

тобто в результаті всі знаки вірні, бо залишковий член може вплинути лише на п'ятий десятковий знак.

Приклад 2. Скільки вузлів таблиці з кроком $h = 5°$ потрібно взяти, щоб обчислити $\sin x$ для $x = 20°$ з точністю до 10^{-3}?

Р о з в' я з а н н я. Скористаємося формулою (16). Маємо $h = 5° \sim 0{,}0873$, $M_{n+1} = \sup\limits_{x \in (-\infty,\infty)} \left|\sin\left[x + (n+1)\frac{\pi}{2}\right]\right| = 1$. З оцінки (16) дістаємо нерівність

$$\frac{h^n}{4n}\|f\|_{C^n(-\infty,\infty)} = \frac{h^n}{4n} \cdot 1 < 10^{-3},$$

або

$$\frac{(0{,}0873)^n}{4n} < 10^{-3}.$$

Неважко перевірити, що при $n = 2$ ця нерівність не виконується, а при $n = 3$ має місце, бо

$$\frac{(0{,}0873)^3}{4 \cdot 3} < \frac{(0{,}09)^3}{12} < 0{,}000061 < 10^{-3}.$$

Отже, трьох вузлів таблиці буде достатньо. Зауважимо, що насправді при $n = 3$ така таблиця дасть чотири вірних знаки.

2.4.7. Стала Лебега. Оцінка відхилення інтерполяційного многочлена від функції в нормі простору $C[a,\ b]$. Крім оцінок залишкового члена в фіксованій точці x, знайдених вище, в багатьох випадках важливо мати оцінку, рівномірну за всіма x, або, іншими словами, оцінку відхилення інтерполяційного многочлена $p(x;\ f)$ від функції $f(x)$ в нормі простору $C[a,\ b]$, що ми і розглянемо далі.

Розглянемо інтерполяційний многочлен у формі

$$p(x) \equiv p(x;\ f) = \sum_{k=1}^{n} f(x_k)\, l_{k,n-1}(x). \tag{24}$$

Ця формула визначає оператор $P_n : C[a, b] \to \pi_{n-1}$, $P_n : f \to p(x; f)$, де π_{n-1} — лінійний підпростір простору $C[a, b]$, який складається з алгебраїчних многочленів степеня не вищого ніж $n - 1$.

Теорема 5. *Норма оператора P_n обчислюється за формулою*

$$\|P_n\| = \max_{x\in[a,b]} \sum_{k=1}^{n} |l_{a,n-1}(x)|.$$

Доведення. Очевидно,

$$|p(x; f)| \leqslant \max_i |f(x_j)| \sum_{k=1}^{n} |l_{k,n-1}(x)| \leqslant \|f\|_{C[a,b]} \sum_{k=1}^{n} |l_{k,n-1}(x)|,$$

тому

$$\|P_n\| = \sup_{f\neq 0} \frac{\|p(x; f)\|_{C[a,b]}}{\|f\|_{C[a,b}} \leqslant \max_{x\in[a,b]} \sum_{k=1}^{n} |l_{k,n-1}(x)|. \qquad (25)$$

З іншого боку, нехай $x_* \in [a, b]$ — точка, де досягається максимум функції $\sum_{k=1}^{n} |l_{k,n-1}(x)|$. Побудуємо функцію $f_0(x) \in C[a, b]$ таку, що $f_0(x_k) = \operatorname{sign} l_{k,n-1}(x_*)$, $k = \overline{1, n}$, $\|f_0\|_{C[a,b]} = 1$ (це може бути, наприклад, кусково-лінійна інтерполююча функція). Тоді

$$p(x_*; f_0) = \sum_{k=1}^{n} l_{k,n-1}(x_*) f_0(x_k) = \sum_{k=1}^{n} |l_{k,n-1}(x_*)|,$$

звідки

$$\frac{\|p(x; f_0)\|_{C[a,b]}}{\|f_0\|_{C[a,b]}} = \max_{x\in[a,b]} \sum_{k=1}^{n} |l_{k,n-1}(x)|,$$

що разом з (25) доводить теорему.

Норма $\|P_n\|$ називається *сталою Лебега* і позначається Λ_n. Стала Лебега не залежить від довжини відрізка інтерполювання $[a, b]$, а залежить лише від відносного розміщення вузлів на ньому.

Дійсно, покладаючи $x = a + (b - a)(y + 1)/2$, $y \in [-1, 1]$, $x_k = a + (b - a)(y_k + 1)/2$, $k = \overline{1, n}$, дістаємо

$$\max_{x\in[a,b]} \sum_{k=1}^{n} |l_{k,n-1}(x)| = \max_{y\in[-1,1]} \sum_{k=1}^{n} |\tilde{l}_{k,n-1}(y)|,$$

де $\tilde{l}_{k,n-1}(y) = \prod_{j\neq k}^{n} \frac{y - y_j}{y_k - y_j}$.

Теорема 6 (про ядро). *Нехай L — лінійне відображення, яке діє з нормованого простору B з нормою $\|\cdot\|$ в нормований простір G з нормою $\|\cdot\|_1$ і множина M міститься в множині* $\ker z \equiv \{z \in B : Lz = 0\}$. *Тоді для будь-якого $x \in B$*

$$\|Lx\|_1 \leqslant \|L\| \inf_{y\in M} \|x - y\|.$$

Доведення. Якщо $y \in M \leqslant \ker L$, то $L(x - y) = Lx$. Тому $\|Lx\|_1 \leqslant \|L\| \|x - y\|$, звідки в силу довільності y дістаємо твердження теореми.

Наслідок (нерівність Лебега). *Якщо $f \in C[a, b]$, то*

$$\|f(x) - p(x; f)\|_{C[a,b]} \leqslant (1 + \Lambda_n) E_{n-1}(f), \qquad (26)$$

де $E_{n-1}(f) = \inf_{q\in\pi_{n-1}} \|f(x) - q(x)\|_{C[a,b]}$ — величина найкращого рівномірного наближення функції $f(x)$ многочленами степеня не вищого $n - 1$.

Доведення. Покладемо $B = G = C[a, b]$ і розглянемо оператор $L = I - P_n$, де I — тотожний оператор. Очевидно, що $\|L\| \leqslant 1 + \Lambda_n$, а $\ker L = \pi_{n-1}$, тому нерівність Лебега є наслідком теореми про ядро.

Величина $E_{n-1}(f)$ залежить від гладкості функції f. Так, якщо $f(x)$ має похідні до r-го порядку включно, інтегровні з степенем p, причому

$$\left(\int_b^a |f^{(r)}(x)|^p dx\right)^{1/p} \leqslant M,$$

то

$$E_{n-1}(f) \leqslant A_r M n^{-r},$$

де A_r — стала, незалежна від M, n.

2.4.8. Оцінка відхилення інтерполяційного многочлена від функції в нормі $\tilde{L}_{2,\rho}$. Нехай $\tilde{L}_{2,\rho}$ — простір неперервних на $[a, b]$ функцій із скалярним добутком

$$(u, v) = \int_a^b \rho(x) u(x) v(x) dx, \qquad (27)$$

де $\rho(x) > 0$ — задана вагова функція. Нехай $\{p_n(x)\}$ — система многочленів, ортогональних у розумінні скалярного добутку (27). Розглянемо інтерполяційний процес у випадку, коли за вузли інтерполяції беруть нулі многочлена $p_n(x)$, які позначимо $x_k^{(n)}$, $k = \overline{1, n}$. Нагадаємо, що всі вони дійсні, різні і належать проміжку (a, b) (див. 1.6). Візьмемо інтерполяційний многочлен у вигляді

$$q_{n-1}(x) \equiv q_{n-1}(x; f) \equiv q(x; f) = \sum_{k=1}^{n} f(x_k^{(n)}) l_{k,n-1}(x). \qquad (28)$$

Неважко помітити, що фундаментальні інтерполяційні многочлени $l_{k,n-1}(x)$ можна подати у вигляді

$$l_{k,n-1}(x) = \frac{p_n(x)}{(x - x_k^{(n)}) p_n'(x_k^{(n)})}. \qquad (29)$$

Дійсно, при $x = x_j^{(n)}$, $j \neq k$, маємо $l_{k,n-1}(x_j^{(n)}) = 0$, бо $x_j^{(n)}$ — корінь многочлена $p_n(x)$, тобто $p_n(x_j^{(n)}) = 0$. Якщо ж $x = x_k^{(n)}$, то

$$l_{k,n-1}(x_k^{(n)}) = \lim_{x \to x_k^{(n)}} \frac{p_n(x)}{(x - x_k^{(n)})\, p_n'(x_k^{(n)})} =$$

$$= \frac{1}{p_n'(x_k^{(n)})} \lim_{x \to x_k^{(n)}} \frac{p_n(x) - p_n(x_k^{(n)})}{x - x_k^{(n)}} = \frac{p_n'(x_k^{(n)})}{p_n'(x_k^{(n)})} = 1.$$

Лема 3. *Фундаментальні інтерполяційні многочлени $l_{k,n-1}(x)$ і $l_{j,n-1}(x)$ при $k \neq j$ ортогональні.*

Доведення. З формули (29)

$$\int_a^b \rho(x)\, l_{k,n-1}(x)\, l_{j,n-1}(x)\, dx = \frac{1}{p_n'(x_k^{(n)})\, p_n'(x_j^{(n)})} \times$$

$$\times \int_a^b \rho(x)\, p_n(x)\, \frac{p_n(x)}{(x - x_k^{(n)})(x - x_j^{(n)})}\, dx.$$

Оскільки $\frac{p_n(x)}{(x - x_k^{(n)})(x - x_j^{(n)})}$ є многочленом степеня $n - 2$ ($x_k^{(n)}$, $x_j^{(n)}$ — корені $p_n(x)$!), то останній інтеграл дорівнює нулю, що і треба було довести.

Лема 4. *Має місце рівність*

$$\sum_{k=1}^{n} \int_a^b \rho(x)\, [l_{k,n-1}(x)]^2\, dx = \int_a^b \rho(x)\, dx.$$

Доведення. Побудувавши інтерполяційний многочлен для функції $f(x) \equiv 1$, дістанемо

$$\sum_{k=1}^{n} l_{k,n-1}(x) = 1.$$

Піднесемо цю рівність до квадрата, помножимо на $\rho(x)$ і проінтегруємо від a до b. В силу леми 3 інтеграли від подвоєних добутків, знайдених при піднесенні до квадрата лівої частини, перетворяться на нуль і дістанемо твердження леми.

У п. 2.4.7 розглядається оператор $P_n : C[a, b] \to C[a, b]$, який функції $f(x) \in C[a, b]$ ставить у відповідність її інтерполяційний многочлен як елемент простору $C[a, b]$. Але цей інтерполяційний многочлен можна розглядати і як елемент нормованого простору $\tilde{L}_{2,\rho}[a, b]$ з нормою $\|f\| = (f, f)^{1/2} = \left(\int_a^b \rho(x)\, f^2(x)\, dx\right)^{1/2}$. Отже, розглянемо оператор $Q_n : C[a, b] \to \tilde{L}_{2,\rho}[a, b]$, який функції $f(x) \in C[a, b]$ ставить у відповідність інтерполяційний многочлен (п. 2.1) як елемент простору $\tilde{L}_{2,\rho}[a, b]$.

Лема 5. *Має місце нерівність*

$$\|Q_n f\|_{\tilde{L}_{2,\rho}[a,b]} \leqslant \left(\int_a^b \rho(x)\, dx\right)^{1/2} \|f\|_{C[a,b]}.$$

Доведення. Підносячи до квадрата обидві частини рівності

$$Q_n f \equiv q_{n-1}(x;\, f) = \sum_{k=1}^{n} f(x_k^{(n)})\, l_{k,n-1}(x)$$

і інтегруючи результат по $[a, b]$ з вагою $\rho(x)$, в силу лем 3 і 4 дістанемо

$$\|Q_n f\|^2_{\tilde{L}_{2,\rho}[a,b]} = \sum_{k=1}^{n} (f(x_k^{(n)}))^2 \int_a^b \rho(x)\, [l_{n-1,k}(x)]^2\, dx \leqslant$$

$$\leqslant \|f\|^2_{C[a,b]} \sum_{k=1}^{n} \int_a^b \rho(x)\, [l_{n-1,k}(x)]^2\, dx = \int_a^b \rho(x)\, dx\, \|f\|^2_{C[a,b]},$$

що і доводить лему.

Теорема 7. *Для будь-якої функції $f(x) \in C[a, b]$ справджується нерівність*

$$\|f - Q_n f\|_{\tilde{L}_{2,\rho}[a,b]} \equiv \|f - q_{n-1}(x;\, f)\|_{\tilde{L}_{2,\rho}[a,b]} \leqslant 2\left(\int_a^b \rho(x)\, dx\right)^{1/2} E_{n-1}(f).$$

Доведення. Нехай $p_{n-1}^*(x)$ — довільний многочлен степеня $n - 1$. Тоді очевидна тотожність

$$f - Q_n f = (f - p_{n-1}^*) + Q_n(p_{n-1}^* - f). \tag{30}$$

Для першого доданку в правій частині очевидна нерівність

$$\|f - p_{n-1}^*\|_{\tilde{L}_{2,\rho}[a,b]} \leqslant \left(\int_a^b \rho(x)\, dx\right)^{1/2} \|f - p_{n-1}^*\|_{C[a,b]}.$$

Для другого доданку з леми 5 маємо

$$\|Q_n(p_{n-1}^* - f)\|_{\tilde{L}_{2,\rho}[a,b]} \leqslant \left(\int_a^b \rho(x)\, dx\right)^{1/2} \|p_{n-1}^* - f\|_{C[a,b]}.$$

Тому з (30)

$$\|f - Q_n f\|_{\tilde{L}_{2,\rho}[a,b]} \leqslant 2\left(\int_a^b \rho(x)\, dx\right)^{1/2} \|f - p_{n-1}^*\|_{C[a,b]},$$

звідки в силу довільності p_{n-1}^* дістаємо твердження теореми.

2.5. Неусувна похибка заокруглення. Обумовленість задачі інтерполювання

2.5.1. Неусувна похибка обчислення інтерполяційного многочлена. Нехай у вузлах x_k, $k = \overline{1, n}$, похибка при обчисленні (або при вимірюванні) $f(x_k) \in \varepsilon_k$, $k = \overline{1, n}$. Тоді, якщо використаємо інтерполяційний многочлен $p(x; f)$ для наближення функції $f(x)$, то неусувна похибка дорівнюватиме

$$\delta = \max_{x \in [a,b]} \left| \sum_{k=1}^{n} \varepsilon_k l_{n-1,k}(x) \right|. \tag{1}$$

Якщо припустити, що $|\varepsilon_k| \approx \varepsilon$, $k = \overline{1, n}$, то при несприятливому збігу знаків ε_k матимемо

$$\delta \approx \varepsilon \Lambda_n. \tag{2}$$

Ця нерівність, як і нерівність Лебега, показує, що вузли інтерполяційного многочлена бажано вибирати так, щоб стала Лебега була якомога меншою. Виявляється, що існує оптимальне розміщення вузлів, для якого константа Лебега набуває деякого мінімального значення Λ_n^*. Точні значення цих вузлів невідомі, але розглянемо теорему, що вказує асимптотично оптимальні вузли, для яких асимптотика сталої Лебега збігається з асимптотикою Λ_n^*.

Теорема 1. *Константа Лебега інтерполяційного многочлена з вузлами в нулях многочлена Чебишева $T_n(x)$ визначається рівністю*

$$\Lambda_n = \frac{2}{\pi} \ln n + 1 - \theta_n, \quad 0 \leqslant \theta_n < \frac{1}{4}, \tag{3}$$

причому

$$\Lambda_n^* > \frac{2}{\pi} \ln n + 1 - \theta_n - 0{,}201 > \frac{2}{\pi} \ln n + 0{,}549. \tag{4}$$

Для порівняння вкажемо границі сталої Лебега для рівновіддалених вузлів:

$$2^{n-3}(n-1)^{-1/2}(n-3/2)^{-1} < \Lambda_n < 2^{n-1}, \quad n \geqslant 4. \tag{5}$$

Із виразів (1) — (5) випливає важливий висновок: не слід використовувати алгебраїчні інтерполяційні многочлени високого степеня по рівновіддалених вузлах. Як правило, такими многочленами користуються для невеликих значень n.

Якщо вузли інтерполювання збігаються з нулями многочлена Якобі $P_n^{(\alpha,\beta)}(x)$, то відомий такий результат про поведінку сталої Лебега:

$$\Lambda_n \leqslant c n^{\sigma+0{,}5},$$

де c — незалежна від n стала, $\sigma > -0{,}5$, $\sigma = \max(\alpha, \beta)$.

Якщо ж вузли інтерполяційного многочлена є нулями ортогонального на $[a, b]$ з вагою $\rho(x) \geqslant \rho_0 > 0$ многочлена $p_n(x)$, то $\Lambda_n = 0$ (n).

З виразу (2) і теореми 1 випливає, що з ростом n неусувна похибка зростатиме, причому швидкість росту визначається величиною сталої Лебега Λ_n, тобто взаємним розміщенням вузлів. Константу Лебега Λ_n природно назвати числом обумовленості задачі обчислення інтерполяційного многочлена $p(x; f)$ в точці x. Чим більша величина Λ_n, тим гірше обумовлена задача обчислення многочлена $p(x; f)$ при заданому x (див. приклад 5, п. 2 вступу).

Розглянемо один з можливих підходів до аналізу похибок заокруглення (прямий аналіз).

2.5.2. Похибка заокруглення при обчисленні значення інтерполяційного многочлена Ньютона. Нехай відносна похибка представлення чисел в ЕОМ та заокруглення є ε. Нехай $\tilde{p}(\xi, f)$ — результат обчислення значення інтерполяційного многочлена $p(x; f)$ в точці ξ за схемою Горнера, яку запишемо у вигляді

$$\begin{aligned} q_n &= f[x_0; \ \ldots; \ x_n], \\ q_{n-1} &= (\xi - x_{n-1}) q_n + f[x_0; \ \ldots; \ x_{n-1}], \end{aligned} \tag{6}$$

тоді $p(\xi; f) = q_0$.

Припустимо, що розділені різниці $a_k \equiv f[x_0; \ \ldots; \ x_k]$, $k = \overline{0, n}$, обчислені точно. Обчислення за схемою Горнера виконуються в такій послідовності:

$$q_n = a_n; \quad q_{n-1} = (\xi \ominus x_{n-1}) \otimes q_n \oplus a_{n-1}, \ \ldots,$$
$$q_{n-r} = (\xi \ominus x_{n-r}) \otimes q_{n-r+1} \oplus a_{n-r}, \quad r = 2, 3, \ \ldots, \ n.$$

Враховуючи формулу $x \circledast y = (x * y)(1 + \delta)$, $|\delta| \leqslant \varepsilon$ (див. вступ, п. 2), дістаємо

$$q_{n-r} = \{[(\xi - x_{n-r})(1 + \tau_r) q_{n-r+1}](1 + \sigma_r) + a_{n-r}\}(1 + \delta_r),$$

де $|\tau_n| \leqslant \varepsilon$, $|\sigma_r| \leqslant \varepsilon$, $|\delta_r| \leqslant \varepsilon$, $\varepsilon = b^{1-t}/2$.
Тому

$$q_{n-1} = (\xi - x_{n-1}) q_n (1 + \tau_1)(1 + \sigma_1)(1 + \delta_1) + a_{n-1}(1 + \delta_1),$$

$$q_{n-2} = (\xi - x_{n-2})(\xi - x_{n-1}) a_n (1 + \tau_2)(1 + \sigma_2)(1 + \delta_2)(1 + \tau_1) \times$$
$$\times (1 + \sigma_1)(1 + \delta_1) + a_{n-1}(\xi - x_{n-2})(1 + \tau_2)(1 + \sigma_2)(1 + \delta_2)(1 + \delta_1) +$$
$$+ a_{n-2}(1 + \delta_2),$$

$$q_{n-3} = (\xi - x_{n-3})(\xi - x_{n-2})(\xi - x_{n-1}) a_n (1 + \tau_3)(1 + \sigma_3)(1 + \delta_3) \times$$
$$\times (1 + \tau_2)(1 + \sigma_2)(1 + \delta_2)(1 + \tau_1)(1 + \sigma_1)(1 + \delta_1) + a_{n-1}(\xi - x_{n-3}) \times$$
$$\times (\xi - x_{n-2})(1 + \tau_3)(1 + \sigma_3)(1 + \delta_3)(1 + \tau_2)(1 + \sigma_2)(1 + \delta_2)(1 + \delta_1) +$$
$$+ (\xi - x_{n-3}) a_{n-2}(1 + \tau_3)(1 + \sigma_3)(1 + \delta_3)(1 + \delta_2) + a_{n-3}(1 + \delta_3).$$

Зрештою

$$\tilde{p}(\xi;\ f) = A_n(\xi - x_{n-1}) \ldots (\xi - x_0) + A_{n-1}(\xi - x_{n-2}) \ldots$$
$$\ldots (\xi - x_0) + \ldots + A_1(\xi - x_0) + A_0,$$

де

$$A_r = a_r(1 + \delta_{n-r})(1 + \delta_{n-r+1})(1 + \sigma_{n-r+1})(1 + \tau_{n-r+1}) \ldots$$
$$\ldots (1 + \delta_n)(1 + \sigma_n)(1 + \tau_n),$$
$$A_1 = a_1(1 + \delta_{n-1})(1 + \delta_n)(1 + \sigma_n)(1 + \tau_n),$$
$$A_0 = a_0(1 + \delta_n).$$

Очевидно, що

$$(1 - \varepsilon)^{3r+1} \leqslant \mu_r \equiv (1 + \delta_{n-r}) \prod_{i=1}^{r} [(1 + \delta_{n-r+i})(1 + \sigma_{n-r+i}) \times$$
$$\times (1 + \tau_{n-r+i})] \leqslant (1 + \varepsilon)^{3r+1}. \qquad (7)$$

Далі, оскільки $\ln(1 + \varepsilon) < \varepsilon$, то

$$(1 + \varepsilon)^l = \exp\{l \ln(1 + \varepsilon)\} < \exp(l\varepsilon) = 1 + l\varepsilon\left(1 + \frac{l\varepsilon}{2!} + \right.$$
$$\left. + \frac{(l\varepsilon)^2}{3!} + \ldots\right) < 1 + l\varepsilon\left(1 + \frac{l\varepsilon}{2} + \frac{(l\varepsilon)^2}{2^2} + \frac{(l\varepsilon)^3}{2^3} + \ldots\right) =$$
$$= 1 + l\varepsilon \frac{1}{1 - l\varepsilon/2}.$$

Припустимо, що існує таке число $h > 0$, що

$$l\varepsilon \leqslant \frac{2h}{1+h}, \qquad (8)$$

тоді

$$\frac{1}{1 - \frac{l\varepsilon}{2}} \leqslant \frac{1}{1 - \frac{h}{1+h}} = 1 + h$$

і тому

$$(1 + \varepsilon)^l < 1 + l\varepsilon(1 + h). \qquad (9)$$

Доведемо за індукцією, що

$$(1 - \varepsilon)^l \geqslant 1 - l\varepsilon. \qquad (10)$$

Дійсно, при $l = 0$ нерівність справедлива. Нехай вона справедлива при $l = k$. Доведемо, що нерівність виконуватиметься при $l = k + 1$:

$$(1 - \varepsilon)^{k+1} = (1 - \varepsilon)(1 - \varepsilon)^k \geqslant (1 - \varepsilon)(1 - k\varepsilon) = 1 - \varepsilon - k\varepsilon + k\varepsilon^2 =$$
$$= 1 - (k + 1)\varepsilon + k\varepsilon^2 > 1 - (k + 1)\varepsilon.$$

З (7), (8), (9), (10) дістаємо, що при виконанні умови

$$n\varepsilon \leqslant \frac{2h}{1+h} \qquad (11)$$

матимемо для всіх $r = \overline{0, n}$

$$1 - (3r + 1)\varepsilon(1 + h) < 1 - (3r + 1)\varepsilon \leqslant \mu_r \leqslant 1 + (3r + 1)(1 + h)\varepsilon.$$

Це означає, що знайдеться таке число $\theta_r \in [-1, 1]$, що для всіх $r = \overline{0, n}$

$$\mu_r = 1 + \theta_r(3r + 1)\varepsilon(1 + h), \qquad (12)$$

тобто справджується рівність

$$A_r = a_r(1 + \varepsilon_r),$$

де

$$|\varepsilon_r| \leqslant (3r + 1)(1 + h)\varepsilon.$$

Сформулюємо доведене у вигляді такої теореми.

Теорема 2. *Нехай* $n\varepsilon \leqslant 2h/(1 + h)$, *де* h *— деяке додатне число,* $\varepsilon = b^{1-t}/2$ *— оцінка відносної похибки представлення чисел в ЕОМ,* b *— основа системи числення,* t *— кількість розрядів у мантисі. Нехай* $\tilde{p}(\xi;\ f)$ *— результат обчислення інтерполяційного многочлена в формі Ньютона* $p(\xi;\ f)$ *в точці* ξ *за схемою Горнера* (6) *за умови, що розділені різниці* $a_r \equiv f[x_0, \ldots, x_r]$ *і вузли* x_r *представлені точно. Тоді*

$$\tilde{p}(\xi;\ f) = A_0 + A_1(\xi - x_0) + \ldots +$$
$$+ A_n(\xi - x_0)(\xi - x_1)(\xi - x_{n-1}), \qquad (13)$$

де

$$A_r = a_r(1 + \varepsilon_r), \quad |\varepsilon_r| \leqslant (3r + 1)(1 + h)\varepsilon, \quad r = \overline{0, n}.$$

Які ж висновки можна зробити з доведеної теореми? По-перше, в деяких випадках похибка заокруглення може бути значною. Наприклад, якщо ξ — корінь многочлена $p(\xi;\ f)$, а розділені різниці $a_r \equiv f[x_0, \ldots, x_r]$ великі, то похибка заокруглення має вигляд

$$\tilde{p}(\xi;\ f) = a_0\varepsilon_0 + a_1\varepsilon_1(\xi - x_0) + \ldots + a_n\varepsilon_n(\xi - x_0) \ldots (\xi - x_{n-1})$$

і при несприятливому збігу знаків ε_i і розміщенні вузлів x_i може досягти значних величин. Те саме стосується і обчислення многочлена в формі $p(\xi) = \xi^n + a_1\xi^{n-1} + \ldots + a_n$ за схемою Горнера $q_n = 1$, $q_{n-r} = \xi q_{n-r+1} + a_{n-r}$, $r = \overline{1, n}$, $q_0 = p(\xi)$. У цьому випадку справедливе зображення

$$\tilde{p}(\xi) = (1 + \varepsilon_n)\xi^n + a_1(1 + \varepsilon_{n-1})\xi^{n-1} + \ldots + a_{n-1}(1 + \varepsilon_1)\xi + a_n,$$
$$|\varepsilon_j| \leqslant (2j + 1)\varepsilon(1 + h)$$

за умови $(2n+1)\varepsilon \leqslant h/(1+h)$. Для множини многочленів $p(x)$ степеня n, які задовольняють на $[0, 1]$ нерівність $|p(x)| \leqslant 1$, типова оцінка суми модулів коефіцієнтів має вигляд $\sum_{k=0}^{n} |a_k| \sim C^n$, де $C > 1$ — деяка стала (це пов'язано з переповненістю степеневого базису $\varphi_i(x) = x^i$, $i = 0, 1$, як наслідку теореми Мюнца). Тому якщо многочлен $p(x)$ задано в явному вигляді своїми коефіцієнтами або розділеними різницями (у формі Ньютона) і передбачається його багаторазове обчислення в різних точках, доцільно перейти до іншого представлення (до іншого базису), наприклад, через многочлени Чебишева, використовуючи при цьому багаторозрядну арифметику. Тоді

$$p(x) = \sum_{j=0}^{n-1} \alpha_j T_{n-j}(x) + \frac{1}{2}\alpha_n \tag{14}$$

і в силу ортогональності многочленів Чебишева з вагою $\frac{1}{\sqrt{1-x^2}}$ маємо

$$\sum_{j=0}^{n-1} \alpha_j^2 + 1/2\alpha_n^2 = \frac{2}{\pi} \int_{-1}^{1} \frac{p^2(x)}{\sqrt{1-x^2}}\, dx.$$

Якщо на $[-1, 1]$ виконується нерівність $|p(x)| \leqslant M$, то звідси $\sum_{j=0}^{n} |\alpha_j| \leqslant c\sqrt{n}$, де c — стала. Многочлени Чебишева задовольняють рекурентне співвідношення

$$T_n(x) - 2xT_{n-1}(x) + T_{n-2}(x) = 0, \quad n = 2, 3, \ldots,$$
$$T_0(x) = 1, \quad T_1(x) = x.$$

Тому (14) можна розглядати як окремий випадок суми

$$s_n(x) = \sum_{k=0}^{n} a_k p_k(x), \tag{15}$$

де функції $p_k(x)$, $k = \overline{1, n}$, $p_{-1}(x)$ задовольняють рекурентні співвідношення

$$p_{k+1}(x) + \alpha_k(x)\, p_k(x) + \beta_k(x)\, p_{k-1}(x) = 0, \quad k = \overline{0, n-1}, \tag{16}$$

$p_0(x)$ — задане. Для обчислення суми (15) можна застосувати таку узагальнену схему Горнера:

$$A_j = a_j - \alpha_j A_{j+1} - \beta_{j+1} A_{j+2}, \quad j = n, \ n-1, \ \ldots, \ 0,$$
$$A_{n+1} = A_{n+2} = 0. \tag{17}$$

Тоді

$$s_n(x) = A_0 p_0 \tag{18}$$

(при $\alpha_j = -x$, $\beta_j = 0$ дістаємо розглянуту раніше схему Горнера). Якщо попередньо обчислити $\alpha_j(x)$ і $\beta_j(x)$, то для обчислення $s_n(x)$ потрібно буде виконати $2n$ множень і $2n$ додавань.

Щоб переконатися в справедливості формули (17), проведемо індукцію по n. Якщо $n = 2$, то формула (17) очевидна. Нехай формула (17) правильна для $k = n$. Тоді

$$s_{n+1}(x) = \sum_{k=0}^{n} b_k p_k(x),$$

де $b_n = a_n - \alpha_n a_{n+1}$, $b_{n-1} = a_{n-1} - \beta_n a_{n+1}$, $b_k = a_k$, $k \leqslant n-2$. Побудуємо послідовність $A_{n+1}, A_n, \ldots, A_0$ за формулою (17), застосовуючи її для $j = n+1, n, \ldots, 0$ і вважаючи $A_{n+2} = A_{n+3} = 0$. Побудуємо також послідовність $B_n, \ldots, B_0$ за величинами $b_n, \ldots, b_0$. Тоді $s_{n+1}(x) = B_0 p_0(x)$, а з іншого боку $A_{n+1} = a_{n+1}$, $A_n = a_n - \alpha_n a_{n+1} = b_n = B_n$, $A_{n-1} = a_{n-1} - \alpha_{n-1}A_n - \beta_n A_{n+1} = b_{n-1} - \alpha_{n-1}B_n = B_{n-1}$ і, отже, $A_j = B_j$ при $0 \leqslant j < n-1$. Це означає $A_0 = B_0$, що і треба було довести.

Аналогічно попередньому можна провести аналіз похибок заокруглення при обчисленні (14) в точці x за схемою (17) і дістати подібний результат, який полягає в тому, що похибка результату визначатиметься сумою модулів величин α_j. Оскільки в (14) $\sum_j |\alpha_j| \leqslant cn^{1/2}$, то зрозуміло, що похибка в даному випадку буде набагато меншою, ніж при інших многочленах, і обчислення можна вести з одинарною точністю.

Існують також інші підходи до аналізу похибок заокруглення, і це становить предмет значної і цікавої частини чисельного аналізу, яка виходить за рамки цієї книги. Можна також виконувати арифметичні операції на ЕОМ точно, бо після запису чисел в ЕОМ вже оперуємо з цілими або раціональними числами. Проте при цьому потрібно використовувати досить складні програми арифметичних операцій. Це захоплюючий розділ арифметики, який частково викладено в книзі [13].

2.6. Апостеріорні оцінки похибки інтерполювання

2.6.1. Оцінки для випадку, коли різниці $(n+1)$-го та $(n+2)$-го порядків зберігають знак. Оцінки, розглянуті раніше, можна дати без будь-яких попередніх обчислень, тому вони називаються *апріорними*. Існують також *апостеріорні оцінки*, які можна зробити лише після певних обчислень. Розглянемо деякі з них.

Нехай відомо, що розділені різниці $(n+1)$-го і $(n+2)$-го порядків зберігають знаки на відрізку $[a, b]$, що містить вузли. Запишемо рівності

$$f(x) = \sum_{i=0}^{n} (x - x_0) \ldots (x - x_{i-1}) f[x_0, x_1, \ldots, x_i] + R_n(x),$$

$$R_n(x) = (x - x_0) \dots (x - x_n) f[x, x_0, \dots, x_n];$$

$$f(x) = \sum_{i=0}^{n+1} (x - x_0) \dots (x - x_{i-1}) f[x_0, x_1, \dots, x_i] + R_{n+1}(x);$$

$$R_{n+1}(x) = (x - x_0) \dots (x - x_{n+1}) f[x, x_0, \dots, x_{n+1}],$$

з яких маємо

$$R_n(x) = (x - x_0)(x - x_1) \dots (x - x_n) f[x_0, x_1, \dots, x_{n+1}] + R_{n+1}(x).$$

Для заданого x завжди можна підібрати x_{n+1} так, що $R_n(x)$ і $R_{n+1}(x)$ матимуть різні знаки. Дійсно, якщо, наприклад, $f[x, x_0, \dots, x_n]$ і $f[x, x_0, \dots, x_{n+1}]$ мають однакові знаки, то візьмемо $x_{n+1} > x$, в іншому разі $x_{n+1} < x$. Тоді

$$\operatorname{sign} R_n(x) = \operatorname{sign} [(x - x_0)(x - x_1) \dots (x - x_n) f[x_0, x_1, \dots, x_{n+1}]]$$

і

$$|R_n(x)| < |(x - x_0)(x - x_1) \dots (x - x_n) f[x_0, x_1, \dots, x_{n+1}]|.$$

Таким чином, знаючи $f(x_{n+1})$, можна оцінити $R_n(x)$.

2.6.2. **Схема Ейткена.** Позначимо через $L_{(k,k+1,\dots,l)}(x)$ інтерполяційний многочлен з вузлами інтерполювання $x_k, \dots, x_l$, тобто

$$L_{(k,k+1,\dots,l)}(x_j) = f(x_j), \quad j = k, k+1, \dots, l.$$

Зокрема, $L_{(k)}(x) = f(x_k)$ — многочлен нульового степеня. Справедлива рівність

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

Дійсно, права частина цієї рівності є многочленом степеня $l - k + 1$ і збігається з $f(x)$ у точках $x_k, \dots, x_{l+1}$, тому вона справедлива внаслідок єдиності інтерполяційного многочлена.

На рівності (1) грунтується схема Ейткена обчислення інтерполяційного многочлена $L_{(1,\dots,n)}(x)$. Ця схема полягає в послідовному обчисленні за допомогою формули (1) таких елементів:

$L_1(x)$,

$L_2(x)\, L_{(1,2)}(x)$,

$L_3(x)\, L_{(2,3)}(x)\, L_{(1,2,3)}(x)$,

. . . . . .

$L_n(x)\, L_{(n-1,n)}(x) \dots L_{(1,2,\dots,n)}(x)$.

Позначимо інтерполяційний многочлен $L_{(1,\dots,m)}(x)$ через $L_m(x)$, $m = \overline{1, n}$. Похибку можна подати так (див. формулу (4) (п. 2.3))

$$f(x) - L_m(x) = f[x, x_1, \dots, x_m]\, \omega_m(x),$$

$$\omega_m(x) = \prod_{i=1}^{m} (x - x_i).$$

Крім того,

$$f(x) - L_{m+1}(x) = f(x) - L_m(x) - (x - x_1) \dots$$
$$\dots (x - x_m) f[x_1, \dots, x_{m+1}],$$

звідки

$$L_{m+1}(x) - L_m(x) = f[x_1, \dots, x_{m+1}]\, \omega_m(x).$$

При малих $|x - x_k|$ і при достатній гладкості функції $f(x)$ маємо (див. п. 2.3)

$$f[x, x_1, \dots, x_m] \approx \frac{f^{(m)}(x)}{m!} \approx f[x_1, \dots, x_{m+1}],$$

тому можна дати таку апостеріорну оцінку похибки інтерполяційної формули $f(x) \approx L_m(x)$:

$$f(x) - L_m(x) \approx L_{m+1}(x) - L_m(x).$$

Схема Ейткена і ця оцінка лежать в основі алгоритму розв'язування такої задачі: за таблицею значень функції $f(x_k)$, $k = \overline{1, n}$, і за заданим x обчислити $f(x)$ із заданою точністю ε або з найкращою можливою точністю при заданій інформації.

Алгоритм будується так: 1) послідовно обчислюються значення $L_0(x)$, $L_1(x)$, $|L_1(x) - L_0(x)|$, $L_2(x)$, $|L_2(x) - L_1(x)|$, ..., якщо при деякому m буде $|L_{m+1}(x) - L_m(x)| \leqslant \varepsilon$, то обчислення припиняють і покладають $f(x) \approx L_m(x)$; 2) якщо ця нерівність не виконується при жодному m, то знаходиться $\varepsilon_{m_0} = \min_m |L_{m+1}(x) - L_m(x)|$ і покладається $f(x) \approx L_{m_0}(x)$; якщо цей мінімум досягається при декількох m, то вибирають найменше з них; 3) якщо величини $|L_{m+1}(x) - L_m(x)|$, починаючи з деякого m, мають стійку тенденцію до збільшення, то з цього моменту обчислення значень $L_m(x)$, $|L_{m+1}(x) - L_m(x)|$ припиняється.

2.7. Збіжність інтерполяційних многочленів

За яких умов при інтерполяції похибка методу прямує до нуля? На практиці є два способи переходу до границі: 1) зберігаючи степінь інтерполяційного многочлена, згущувати вузли інтерполювання (сітку вузлів); 2) збільшувати число вузлів, тобто підвищувати степінь многочлена. Розглянемо детальніше обидва способи.

1. *Зменшення кроку.* Якщо $f(x) \in C^{k+1}[0, 1]$, $0 \leqslant k < n$, і $p_n(x; f)$ — її інтерполяційний многочлен по рівновіддалених вузлах $x_i = x_0 + ih$, $i = \overline{0, n}$, $x_i \in [0, 1]$, то, як було показано в п. 2.4.4,

$$|f(x) - p_n(x; f)| \leqslant \frac{h^{k+1}}{(k+1)\, 2^{k+2-n}} \|f^{(k+1)}\|_{C[0,1]}. \qquad (1)$$

Звідси видно, що при $h \to 0$ і фіксованих n, x маємо

$$\lim_{h\to 0} p_n(x;\ f) = f(x).$$

Для заданої похибки методу ε і степеня n з нерівності (1) можна визначити крок h, при якому

$$|f(x) - p_n(x;\ f)| < \varepsilon.$$

2. *Збільшення степеня многочлена (кількості вузлів).* Розглянемо інтерполювання на $[a,\ b]$, коли число вузлів, які використовуються для побудови інтерполяційного многочлена, необмежено зростає.

Для кожного натурального n виберемо на відрізку $[a,\ b]$ $n+1$ різних точок $x_i^{(n)}$ $(i = \overline{0,\ n};\ n = 0,\ 1,\ \ldots)$ і побудуємо за цими наборами точок послідовність інтерполяційних многочленів $p_n(x;\ f)$ для функції $f(x) \in C[a,\ b]$. Інтерполяційний процес називатимемо збіжним, якщо $\lim\limits_{n\to\infty} p_n(x;\ f) = f(x)$, $x \in [a,\ b]$, і рівномірно збіжним, якщо ця збіжність рівномірна по x.

Можна довести справедливість такої теореми.

Теорема 1. *Нехай $f(x)$ — будь-яка ціла функція, тобто $f(x)$ можна подати у вигляді всюди збіжного степеневого ряду*

$$f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k.$$

Тоді послідовність її інтерполяційних многочленів $p_n(x;\ f)$ по вузлах $x_k^{(n)} \in [a,\ b]$, $k = \overline{0,\ n}$, $n = 0,\ 1,\ \ldots$, рівномірно збігається на $[a,\ b]$ до $f(x)$.

Однак практична цінність цього твердження невелика, бо клас цілих функцій дуже вузький. А якщо навіть $f(x)$ має на $[a,\ b]$ неперервні похідні як завгодно високих порядків, то це не гарантує збіжності при довільному розміщенні вузлів. Наприклад, візьмемо нескінченно диференційовну на $[-1,\ 1]$ функцію

$$f(x) = \begin{cases} 0, & x \in [-1,\ 0], \\ e^{-1/x}, & x \in (0,\ 1] \end{cases}$$

і розмістимо всі вузли інтерполяції лівіше від точки $x = 0$. Очевидно, що тоді $p_n(x;\ f) \equiv 0$ і про збіжність не може бути мови. Правда, в цьому прикладі вузли були розміщені «грубо» нерівномірно. Проте, як виявляється, рівномірне розміщення вузлів також не завжди забезпечує збіжність. С. Н. Бернштейн в 1916 р. довів, що для функції $f(x) = |x|$ на відрізку $[-1,\ 1]$, покритому рівномірною сіткою, значення інтерполяційного многочлена $p_n(x;\ f)$ в точках між вузлами інтерполяції необмежено зростають при $n \to \infty$, тобто збіжності немає. Швидкий ріст сталої Лебега у випадку рівновіддалених вузлів дає змогу припустити, що інтерполяційний процес у цьому разі збігається

лише для вузького класу функцій. Наприклад, показано, що навіть для таких гладких функцій, як

$$f_r(x) = \begin{cases} 0, & x \leqslant 0, \\ x^r, & x \geqslant 0, \end{cases}$$

де r — будь-яке натуральне число, $f_r \in C^{r-1}$, інтерполяційний процес за вузлами $x_k = -1 + \frac{2k}{n}$, $k = \overline{0,\ n}$, розбігається в усіх точках проміжку $[-1,\ 1]$, за виключенням $-1,\ 0,\ 1$.

Якщо розглядати $p_n(x;\ f)$ як оператор $P_n : C[a,\ b] \to \pi_n$, де π_n — множина многочленів степеня не вищого n, $\pi_n \subset C[a,\ b]$, то збіжність інтерполяційного процесу означає сильну збіжність послідовності операторів P_n до одиничного оператора I. Нагадаємо, що теорема Банаха — Штейнгауза стверджує: для того щоб послідовність обмежених лінійних операторів $A_n : N \to M$ $(N \subset B,\ M \subset B,\ B$ — банахів простір) сильно збігалась до деякого обмеженого лінійного оператора A (тобто $\| A_n f - Af \| \to 0$ при $n \to \infty$ $\forall f \subset N$), необхідно і достатньо, щоб: 1) норми всіх A_n були обмежені в сукупності; 2) на всіх елементах всюди щільної множини в N послідовність $\{A_n x\}$ була збіжною.

Друга умова цієї теореми для операторів P_n виконана, бо за теоремою Вейєрштрасса множина многочленів щільна в $C[a,\ b]$ і для будь-якого многочлена $z(x)$ при досить великих n буде $p_n(x;\ z) \equiv z(x)$. Однак перша умова, як випливає з оцінок (п. 2.4.7) для сталих Лебега, не виконується для жодного набору вузлів. Тому з теореми Банаха — Штейнгауза випливає таке твердження.

Теорема 2. *Для будь-якого набору вузлів $x_i^{(n)} \in [a,\ b]$, $i = \overline{0,\ n}$, існує неперервна функція $f(x) \in C[a,\ b]$, для якої інтерполяційний процес розбігається.*

Для конкретних наборів вузлів і функцій, які належать більш гладким підмножинам простору $C[a,\ b]$, оцінки сталої Лебега та величин $E_n(f)$ і нерівність Лебега дають змогу формулювати теореми про збіжність інтерполяційного процесу. Друга теорема Джексона, наприклад, стверджує, що для функцій з класу $C^{k,\lambda}[a,\ b]$ (клас функцій, які мають неперервні похідні до k-го порядку, причому похідна порядку k задовольняє умову Ліпшиця з показником $0 \leqslant \lambda \leqslant 1$) справедлива нерівність $E_n(f) \leqslant \frac{C}{n^{k+\lambda}}$. Цей результат і оцінки для Λ_n (п.2.4.7) приводять, зокрема, до наступного твердження.

Теорема 3. *Нехай $f(x) \in C^{k,\lambda}[-1,\ 1]$, $k \geqslant 0$. Тоді, якщо $p_n(x;\ f)$ — інтерполяційний многочлен по вузлах Чебишева (тобто по нулях многочлена Чебишева $T_{n+1}(x)$), то*

$$\| p_n(x;\ f) - f \|_{C[a,b]} \leqslant \frac{C \ln n}{n^{k+\lambda}},$$

і тому $\| p_n(x;\ f) - f \|_{C[a,b]} \to 0$ при $n \to \infty$.

З цієї теореми випливає, що коли для функцій $f(x) = |x|$, $x \in [-1, 1]$ або $f(x) = f_r(x)$, $r \geqslant 0$, побудувати інтерполяційний многочлен по вузлах Чебишева, то він рівномірно збігатиметься до $f(x)$ при $n \to \infty$. Дійсно, $|x| \in C^{0,1}[-1, 1]$, $f_1(x) \in C^{0,1}[-1, 1]$, $f_r(x) \in C^{r-1}[-1, 1]$, $r \geqslant 2$, тому

$$\| |x| - p_n(x; |x|) \|_C \leqslant \frac{C \ln n}{n},$$

$$\| f_1(x) - p_n(x; f_1) \|_C \leqslant \frac{C \ln n}{n},$$

$$\| f_r(x) - p_n(x; f_r) \|_C \leqslant \frac{C \ln n}{n^{r-1}}, \quad r \geqslant 2,$$

де C — стала, що не залежить від n.

Досі ми фактично розглядали збіжність $p_n(x; f)$ до $f(x)$ в чебишевський нормі. Збіжності в середньому для будь-якої функції $f(x) \in C[a, b]$ можна досягти вибором вузлів.

Дійсно, нехай $f(x) \in C[a, b]$. Тоді за першою теоремою Джексона

$$E_n(f) \leqslant \omega\left(f; \frac{(b-a)\pi}{2(n+2)}\right), \tag{2}$$

де $\omega(f; t) = \sup\limits_{|x_1 - x_2| \leqslant t} |f(x_1) - f(x_2)|$ — модуль неперервності функції $f(x)$, причому для $f \in C[a, b]$ $\omega(f; t) \to 0$ при $t \to 0$.

Виберемо за вузли інтерполювання точки $x_k^{(n+1)}$, $k = \overline{0, n}$, які є нулями многочлена $p_{n+1}(x)$ із системи ортогональних у просторі $\tilde{L}_{2,\rho}[a, b]$ многочленів, і побудуємо інтерполяційний многочлен $q_n(x; f)$. Тоді з теореми 2.4.8 і нерівності (2) дістаємо збіжність $q_n(x; f)$ до $f(x) \in C[a, b]$ в середньому (або в просторі $\tilde{L}_{2,\rho}[a, b]$), тобто

$$\int_a^b \rho(x) [q_n(x; f) - f(x)]^2 \, dx \to 0$$

при $n \to \infty$. Теорема 7 (п. 2.4.8) разом з оцінками для $E_n(f)$ дає і швидкість збіжності для більш гладких функцій $f(x)$.

Наведемо один з наслідків збіжності інтерполяційних многочленів з кратними вузлами. Якщо $f(x) \in C[-1, 1]$ і за вузли інтерполювання вибрані нулі многочленів Чебишева першого роду $T_n(x)$, для будь-яких чисел C_{in}, що задовольняють умову

$$\lim_{n \to \infty} \max_i |C_{in} \ln n / n| = 0,$$

многочлен $\mathscr{H}_{2n-1}(x) = \mathscr{H}_{2n-1}(x; f)$ побудований за допомогою інтерполювання з кратними вузлами за числами $f(x_i)$, $i = \overline{0, n}$, C_{in}, $i = \overline{0, n}$, тобто $\mathscr{H}_{2n-1}(x_i) = f(x_i)$, $\mathscr{H}'_{2n-1}(x_i) = C_{in}$ рівномірно збігається до $f(x)$ при $n \to \infty$. Очевидно, якщо $f(x)$ має обмежену похідну, то за C_{in} можна брати значення похідної у вузлах, тобто $C_{in} = f'(x_i)$.

2.8. Застосування інтерполювання. Обернене інтерполювання

Інтерполювання застосовується не тільки для обчислення значень табульованої функції при будь-яких значеннях аргументу. За допомогою інтерполяційних многочленів розв'язуються задачі оберненого інтерполювання, будуються формули чисельного диференціювання, чисельного інтегрування, чисельного розв'язування задачі Коші, крайових задач тощо.

Задачею оберного інтерполювання називають задачу про знаходження x для довільного заданого y, якщо задана таблиця $y_i = f(x_i)$, $i = \overline{0, n}$. Якщо функція $f(x)$ монотонна, то шукане значення x єдине. Вважатимемо змінну y незалежною, а x — функцією від y. Тоді, записавши за даними (y_i, x_i) інтерполяційний многочлен, наприклад у формі Лагранжа

$$x(y) = \sum_{i=0}^{n} x_i \frac{(y - y_0) \ldots (y - y_{i-1})(y - y_{i+1}) \ldots (y - y_n)}{(y_i - y_0) \ldots (y_i - y_{i-1})(y_i - y_{i+1}) \ldots (y_i - y_n)},$$

за заданим y знайдемо x. Залишковий член у цьому разі утворюється із залишкового члена формули Ньютона, якщо в останньому поміняти місцями x і y, а похідні від функції $f(x)$ замінити похідними від оберненої функції.

Якщо функція $f(x)$ не монотонна, то записуємо інтерполяційний многочлен $p_n(x; f)$ і розв'язуємо рівняння $p_n(x; f) = y$ відносно x. Оцінимо похибку цього методу. Маємо

$$f(x) - p_n(x; f) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0) \ldots (x - x_n).$$

Нехай $\bar{x}$ — розв'язок рівняння $p_n(x; f) = y$, тоді

$$f(\bar{x}) - p_n(\bar{x}; f) = f(\bar{x}) - y = f(\bar{x}) - f(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega_n(\bar{x}).$$

Застосовуючи теорему Лагранжа, маємо

$$(\bar{x} - x) f'(\eta) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega_n(\bar{x}),$$

де η лежить між x і $\bar{x}$. Якщо $x, \bar{x} \in [a, b]$ і $\min\limits_{x \in [a,b]} |f'(x)| = m_1 \neq 0$, то з останньої рівності випливає оцінка

$$|\bar{x} - x| \leqslant \frac{M_{n+1}}{m_1 (n+1)!} |\omega_n(\bar{x})|.$$

Приклад 1. Розв'язати рівняння

$$f(x) \equiv (1 + x) e^{0,5x} - 2,5 = 0,$$

застосувавши обернене інтерполювання.

Розв'язання. Складемо таблицю значень функцій $f(x)$ і, оскільки вона монотонна, застосуємо перший із розглянутих методів:

y_i	x_i	$x(y_i;\ y_{i+1})$	$x(y_0;\ y_1;\ y_2)$
$-1{,}5$	0		
		0,540	
$-0{,}574$	0,5		$-0{,}076$
		0,365	
0,797	1,0		

Далі знаходимо

$$x(0) \approx x_0 + (0 - y_0)\, x(y_0;\ y_1) + (0 - y_0)(0 - y_1)\, x(y_0;\ y_1;\ y_2) = 0{,}744.$$

Точним є розв'язок $x(0) = 0{,}732$. Для підвищення точності доцільно взяти нові вузли, розміщені близько до грубо знайденого кореня, а не збільшувати кількість вузлів.

Приклад 2. Нехай функція $f(x)$ задана таблицею

x	-1	0	2
$f(x)$	-4	-5	-1

Знайти x, при якому $f(x) = -2$.

Розв'язання. Як видно з таблиці, функція не монотонна, тому застосуємо другий спосіб. Знаходимо

$$p_2(x;\ f) = -5\frac{(x+1)(x-2)}{1(-2)} - 4\frac{x(x-2)}{(-1)(-3)} - 1\frac{x(x+1)}{2\cdot 3} = x^2 - 5.$$

Розв'язуючи рівняння $x^2 - 5 = -2$, знаходимо $x = \pm\sqrt{3}$.

2.9. Чисельне диференціювання

Задача чисельного диференціювання формулюється так: за заданими значеннями функції $f(t)$ в точках x_i, $i = \overline{0, n}$, і заданими x, k знайти наближено $f^{(k)}(x)$, $k \geqslant 1$, і оцінити похибку. Одна з ідей побудови формул для наближеного обчислення похідних від функції $f(x)$ така. Якщо функція $\varphi(x)$ наближує функцію $f(x)$ у певному розумінні (це може бути, наприклад, інтерполяційний многочлен, інтерполяційний чи згладжуючий сплайн, многочлен найкращого середньоквадратичного наближення, див. далі), то покладають $f^{(k)}(x) \approx \varphi^{(k)}(x)$ у заданій точці x.

2.9.1. Побудова формул чисельного диференціювання. Оцінки похибок. Найпростіші формули чисельного диференціювання дістають за допомогою диференціювання інтерполяційних многочленів. Якщо

$$p_n(x;\ f) = \sum_{i=0}^{n} f(x_i)\, l_{i,n}(x) -$$

інтерполяційний многочлен для функції $f(x)$, $l_{i,n}(x)$ — фундаментальні інтерполяційні многочлени, то $f^{(k)}(x)$ наближається за допомогою формули

$$f^{(k)}(x) \approx \sum_{i=0}^{n} c_i f(x_i),$$

де $c_i = l_{i,n}^{(k)}(x)$.

Зауважимо, що задача диференціювання не є коректною в $C[a, b]$, бо немає неперервної залежності норми похідної від норми функції. Про це говорить такий приклад.

Нехай $f(x) \in C[a, b]$ та $\tilde{f}(x) \in C[a, b]$ зв'язані співвідношенням

$$\tilde{f}(x) = f(x) + n^{-1} \sin[n^2(x - a)],$$

тоді

$$\|f - \tilde{f}\|_C = \max_{x\in[a,b]} \left|\frac{1}{n}\sin[n^2(x-a)]\right| \leqslant \frac{1}{n} \underset{n\to\infty}{\to} 0,$$

але

$$\|f' - \tilde{f}'\|_C = \max_{x\in[a,b]} |n\cos[n^2(x-a)]| = n \underset{n\to\infty}{\to} \infty.$$

Така сама властивість притаманна і операції чисельного диференціювання. Зобразимо функцію $f(x)$ через інтерполяційний многочлен Ньютона

$$f(x) = f(x_0) + (x - x_0) f[x_0,\ x_1] + (x - x_0)(x - x_1) f[x_0,\ x_1,\ x_2] + \cdots + (x - x_0)(x - x_1)\ldots(x - x_{n-1}) f[x_0,\ x_1,\ \ldots,\ x_n] + (x - x_0)(x - x_1)\ldots(x - x_n) f[x,\ x_0,\ x_1,\ \ldots,\ x_n] \equiv \equiv p_n(x;\ f) + R_n(x;\ f).$$

Звідси дістаємо таке співвідношення для похідної k-го порядку:

$$f^{(k)}(x) = p_n^{(k)}(x;\ f) + R_n^{(k)}(x;\ f),$$

де похідна від залишкового члена за допомогою формули Лейбніца зображується так:

$$R_n^{(k)}(x;\ f) = \sum_{i=0}^{k} C_k^i f^{(i)}[x,\ x_0,\ \ldots,\ x_n]\, \omega_{n+1}^{(k-i)}(x),$$

$$C_k^i = \frac{k(k-1)\ldots(k-i+1)}{i!} = \frac{k!}{i!\,(k-i)!}. \qquad (1)$$

Нехай функція $g(x)$ неперервно диференційовна q разів. В 2.3 ми

вивели формулу

$$g[x,\ x+\varepsilon,\ \ldots,\ x+q\varepsilon] = g^{(q)}(x_\varepsilon)/q!,$$

де $x \leqslant x_\varepsilon \leqslant x+q\varepsilon$. Звідси маємо

$$g^{(q)}(x) = q!\lim_{\varepsilon\to 0} g[x,\ x+\varepsilon,\ \ldots,\ x+q\varepsilon]q!,$$

отже, за означенням розділеної різниці за кратними вузлами

$$[f[x,\ x_0,\ \ldots,\ x_n]]^{(q)} = q![\lim_{\varepsilon\to 0} f[x,\ x+\varepsilon,\ \ldots,\ x+q\varepsilon,\ x_0,\ \ldots,\ x_n]] =$$

$$= f[\underbrace{x,\ \ldots,\ x}_{q+1 \text{ раз}},\ x_0,\ \ldots,\ x_n]q!$$

Таким чином, співвідношення (1) можна записати у вигляді

$$f^{(k)}(x) - p_n^{(k)}(x;\ f) =$$

$$= \sum_{i=0}^{k} \frac{k!}{(k-i)!\,i!}\, i!\, f[x,\ \ldots,\ x,\ x_0,\ \ldots,\ x_n]\,\omega_{n+1}^{(k-i)}(x). \qquad (2)$$

Виражаючи розділену різницю через похідну, дістаємо оцінку

$$|R_n^{(k)}(x;\ f)| = |f^{(k)}(x) - p_n^{(k)}(x;\ f)| \leqslant \sum_{i=0}^{k} \frac{k!}{(k-i)!\,(n+i+1)!} \times$$

$$\times \max_{\xi\in[y_1,y_2]} |f^{(n+i+1)}(\xi)|\,|\omega_{n+1}^{(k-i)}(x)|, \qquad (3)$$

де $y_1 = \min(x,\ x_0,\ \ldots,\ x_n)$, $y_2 = \max(x,\ x_0,\ \ldots,\ x_n)$.

Розглянемо таке розміщення вузлів, при якому $x_i - x_{i-1} = O(h)$, $i = \overline{1,\ n}$, де h — деякий параметр, сітка вузлів згущається при $h \to 0$. При фіксованому n величина $\omega_{n+1}^{(j)}(x)$ є сумою добутків, у кожному з яких $n+1-j$ співмножників порядку $O(h)$ кожен, а тому $\omega_{n+1}^{(j)}(x) = O(h^{n+1-j})$. Отже, рівність (1) можна записати у вигляді

$$R_n^{(k)}(x;\ f) = f[x,\ x_0,\ \ldots,\ x_n]\,\omega_{n+1}^{(k)}(x) + O(h^{n+2-k}) = O(h^{n+1-k}). \qquad (4)$$

Якщо точка x така, що $\omega_{n+1}^{(k)}(x) = 0$, то порядок точності формули чисельного диференціювання $f^{(k)}(x) \approx p_n^{(k)}(x;\ f)$ збільшується на одиницю. Тому точки, в яких $\omega_{n+1}^{(k)}(x) = 0$, називаються *точками підвищеної точності*.

Розглянемо деякі найпростіші формули чисельного диференціювання. Ввівши позначення $\xi_i = x - x_i$, запишемо інтерполяційний многочлен Ньютона у вигляді

$$p_n(x;\ f) = f(x_0) + \xi_0 f[x_0,\ x_1] + \xi_0\xi_1 f[x_0,\ x_1,\ x_2] + \cdots +$$

$$+ \xi_0\xi_1\ldots\xi_{n-1} f[x_0,\ x_1,\ \ldots,\ x_n].$$

Диференціюючи цей вираз, дістанемо

$$p_n'(x;\ f) = f[x_0,\ x_1] + (\xi_0+\xi_1) f[x_0;\ x_1;\ x_2] + (\xi_0\xi_1 + \xi_0\xi_2 +$$

$$+ \xi_1\xi_2) f[x_0,\ x_1,\ x_2,\ x_3] + \cdots \qquad (5)$$

$$p_n''(x;\ f) = 2f[x_0,\ x_1,\ x_2] + 2(\xi_0+\xi_1+\xi_2) f[x_0,\ x_1,\ x_2,\ x_3] + \cdots \qquad (6)$$

. . . . . .

Для k-ї похідної маємо формулу

$$p_n^{(k)}(x;\ f) = k!\,\{[f[x_0,\ x_1,\ \ldots,\ x_k] + \left(\sum_{i=0}^{k} \xi_i\right) f[x_0,\ x_1,\ \ldots,\ x_k,\ x_{k+1}] +$$

$$+ \left(\sum_{i>j\geqslant 0}^{i=k+1} \xi_i\xi_j\right) f[x_0,\ x_1,\ \ldots,\ x_{k+1},\ x_{k+2}] + \qquad (7)$$

$$+ \left(\sum_{i>j>l\geqslant 0}^{i=k+2} \xi_i\xi_j\xi_l\right) f[x_0,\ x_1,\ \ldots,\ x_{k+3}] + \cdots\},$$

в якій $n+1-k$ доданків. Тому формулу (7) називають ще $(n+1-k)$-членною. Якщо в (7) залишити тільки перший доданок, то дістанемо одночленну формулу $f^{(k)}(x) \approx p_n^{(k)}(x,\ f) \approx k!\,f[x_0,\ \ldots,\ x_k]$. Залишковий член цієї формули за умови $x_i - x_{i-1} = O(h)$, $i = \overline{1,\ n}$ має вигляд

$$f^{(k)}(x) - k!\,f[x_0,\ \ldots,\ x_k] = R_n^{(k)}(x;\ f) + k!\left(\sum_{i=0}^{k} \xi_i\right) \times$$

$$\times f[x_0,\ \ldots,\ x_{k+1}] + \cdots = k!\left(\sum_{i=0}^{k} \xi_i\right) f[x_0,\ \ldots,\ x_{k+1}] + O(h^2). \qquad (8)$$

Двочленна формула для k-ї похідної $f^{(k)}(x) \approx p_n^{(k)}(x;\ f) \approx k!\,[f[x_0,\ \ldots,\ x_k] + \left(\sum_{i=0}^{k} \xi_i\right) f[x_0,\ \ldots,\ x_{k+1}]]$ має залишковий член

$$\left(\sum_{i>j\geqslant 0}^{i=k+1} \xi_i\xi_j\right) f[x_0,\ \ldots,\ x_{k+2}] + O(h^3) \qquad (9)$$

і так далі.

Найчастіше вживають такі одночленні формули:

$$f'(x) \approx f[x_0,\ x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}, \qquad (10)$$

$$f''(x) \approx 2f[x_0,\ x_1,\ x_2] = \frac{2}{x_1 - x_0}\left(\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}\right), \qquad (11)$$

. . . . . .

$$f^{(k)}(x) \approx k!\,f[x_0,\ \ldots,\ x_k] = k!\sum_{i=0}^{k} f(x_i) \prod_{\substack{j=0 \\ j\neq i}}^{k} (x_i - x_j)^{-1}. \qquad (12)$$

Неважко помітити, що мінімальна кількість вузлів, необхідна для обчислення k-ї похідної, є $k+1$. Оскільки залишковий член $R_n^{(k)}(x; f)$ формули (7) — це многочлен виду $\Sigma\Pi(x-x_i)$ степеня $n+1-k$ відносно x, то кількість точок підвищеної точності для $(n+1-k)$-членної формули (7) дорівнює $n+1-k$. Позначимо ці точки через $x_i^{(p,k)}$, $i=\overline{1,p}$, $p=n+1-k$. В одночленній формулі для k-ї похідної, як випливає з (8), точка підвищеної точності визначається з умови

$$\sum_{i=0}^{k}\xi_i=\sum_{i=0}^{k}(x-x_i)=0,$$

звідки

$$x_i^{(1,k)}=(x_0+\cdots+x_k)/(k+1).$$

У цій точці на рівномірній сітці з кроком h або на нерівномірній сітці такій, що $x_i-x_{i-1}=O(h)$, одночленна формула має похибку порядку $O(h^2)$ замість $O(h)$.

Вправа 1. Довести, що у двочленній формулі для k-ї похідної є дві точки підвищеної точності:

$$x_i^{(2,k)}=\left[\sqrt{k+1}\sum_{i=1}^{k+1}x_i\pm\sqrt{\sum_{i>j\geqslant 0}^{i=k+1}(x_i-x_j)^2}\right]/[(k+2)\sqrt{k+1}],$$

в яких досягається третій порядок точності за h.

Вказівка. Слід розв'язати рівняння $\sum\limits_{i>j\geqslant 0}^{i=k+1}\xi_i\xi_j=0$, яке має вигляд

$$\frac{(k+2)(k+1)}{2}x^2-x\sum_{i>j\geqslant 0}^{i=k+1}(x_i+x_j)+\sum_{i>j\geqslant 0}^{i=k+1}x_ix_j=0.$$

Якщо $p>2$, то знайти точки підвищеної точності складно, за винятком окремого випадку, про який йдеться у такій теоремі.

Теорема 1. *Нехай $p=n+1-k$ — парне, а вузли в формулі (7) вибрано так, що вони розміщені симетрично відносно точки x. Тоді x є однією з точок підвищеної точності.*

Доведення. Якщо p непарне, то головний член залишкового члена має вигляд $W\equiv\Sigma\Pi\xi_i=\Sigma\Pi(x-x_i)$ і кожний доданок є многочленом p-го, тобто непарного степеня. В силу симетричного розміщення вузлів відносно точки x величини $\xi_i=x-x_i$ мають попарно рівні абсолютні величини, але протилежні знаки. Якщо змінити нумерацію вузлів, тобто x_0 позначити через x_n, x_1 — через x_{n-1} і т. д., то величина W не повинна змінитися. Але при цьому величини ξ_i поміняють знаки і в силу непарності p знак W мав би змінитись на протилежний. Виконання останніх двох умов на W можливе лише при $W=0$, а це і треба було довести.

Очевидно, що симетричне розміщення вузлів відносно точки x означає, що при непарній кількості вузлів точка x збігається з центральним вузлом, а при парній — лежить між середніми вузлами. Умова симетрії, зокрема, легко реалізується на рівномірній сітці, що дає змогу знаходити прості формули підвищеної точності. Наприклад, одночленні формули (10), (11) для трьох рівновіддалених вузлів x_0, x_1, x_2, $x_2-x_1=x_1-x_0=h$ з урахуванням теореми 1 можна записати так:

$$f'(x_1)=\frac{f(x_2)-f(x_0)}{2h}+O(h^2),\tag{13}$$

$$f''(x_1)=\frac{f(x_2)-2f(x_1)+f(x_0)}{h^2}+O(h^2).\tag{14}$$

Формулу (13) часто записують в іншому вигляді, зручному для визначення похідної в середній точці інтервалу сітки

$$f'_{i+1/2}\equiv f'(x_{i+1/2})=\frac{f_{i+1}-f_i}{h}+O(h^2),$$

$$x_{i+1/2}=x_i+1/2h.\tag{15}$$

Аналогічно можна дістати формули більш високого порядку точності. Наприклад, тричленна формула (5) для першої похідної в середині інтервалу рівномірної сітки за чотирма сусідніми вузлами дає

$$f'_{3/2}=(-f_3+27f_2-27f_1+f_0)/(24h)+O(h^4),\tag{16}$$

а для другої похідної в центральному вузлі за п'ятьма вузлами маємо

$$f''_2=(-f_4+16f_3-30f_2+16f_1-f_0)/(12h^2)+O(h^4).\tag{17}$$

Зауважимо, що формули (13) — (17) справедливі для рівномірної сітки: якщо ж застосувати, наприклад, формулу (11) на нерівномірній сітці, то дістанемо похибку $O(h)$, $h=\max(x_2-x_1, x_1-x_0)$, а застосування формули (14) дає похибку $O(1)$, тобто приводить до грубої помилки.

Для апріорної оцінки точності формул диференціювання часто застосовують розвинення за формулою Тейлора — Маклорена. Нехай, наприклад, $f(x)$ має неперервну четверту похідну. Тоді

$$f(x_{i\pm1})=f(x_i\pm h)=f_i\pm hf'_i+\frac{h^2}{2}f''_i\pm\frac{h^3}{6}f'''_i+\frac{h^4}{24}f^{IV}(\eta_\pm),$$

$$f(x_i)=f_i,\tag{18}$$

де η_+ є деякою точкою інтервалу (x_i, x_{i+1}), а η_- — точкою з інтервалу (x_{i-1}, x_i). Використовуючи (18), маємо

$$f_{\bar{x}x,i}\equiv h^{-2}(f_{i+1}-2f_i+f_{i-1})=f''_i+\frac{h^2}{24}[f^{IV}(\eta_+)+f^{IV}(\eta_-)]=$$

$$=f''_i+O(h^2)=f''_i+\frac{h^2}{12}f^{IV}(\eta),\quad \eta\in(x_{i-1}, x_{i+1}).\tag{19}$$

Такий спосіб уточнює величину залишкового члена, який виявився рівним $h^2 f^{IV}(\eta)/12$.

Вправа 2. За допомогою розвинення за формулою Тейлора — Маклорена довести, що для односторонніх різницевих похідних $f_{x,i} = (f_{i+1} - f_i)/h$, $f_{\bar{x},i} = (f_i - f_{i-1})/h$ у випадку $f \in C^2(\bar{\Omega})$ мають місце співвідношення

$$f_{x,i} = f_i' + \frac{h}{2} f_i''(\eta), \quad \eta \in (x_i, x_{i+1}),$$

$$f_{\bar{x},i} = f_i' - \frac{h}{2} f_i''(\eta), \quad \eta \in (x_{i-1}, x_i). \tag{20}$$

Розглянемо оцінки залишкових членів формул чисельного диференціювання у випадку $f(x) \in W_2^k(\Omega)$.

Теорема 2. *Для лінійних функціоналів*

$$R^{(1)}(f) \equiv [f(x_2) - f(x_0)]/(2h) - f'(x_1),$$
$$R^{(2)}(f) \equiv [f(x_2) - 2f(x_1) + f(x_0)]/h^2 - f''(x_1),$$

заданих у просторах $W_2^3(e)$ і $W_2^4(e)$, $e = (x_1 - h, x_1 + h)$, $x_0 = x_1 - h$, $x_2 = x_1 + h$, *відповідно мають місце оцінки*

$$|R^{(1)}(f)| \leqslant Mh^{1,5} |f|_{W_2^3(e)}, \tag{21}$$

$$|R^{(2)}(f)| \leqslant Mh^{1,5} |f|_{W_2^4(e)}, \tag{22}$$

де сталі M *не залежать від* h *і* f.

Доведення. За допомогою лінійної заміни $\xi = 2hs + x_1$ відобразимо e в $E = (-0{,}5;\ 0{,}5)$, $\xi \in e$, $s \in E$, позначимо $\tilde{f}(s) = f(\xi(s))$. При цьому функціонали $R^{(1)}$ і $R^{(2)}$ перетворяться таким чином:

$$R^{(1)}(f) = R^{(1)}(\tilde{f}) = [\tilde{f}(0{,}5) - \tilde{f}(-0{,}5)]/(2h) - \tilde{f}'(0)/(2h),$$

$$R^{(2)}(f) = R^{(2)}(\tilde{f}) = [\tilde{f}(0{,}5) - 2\tilde{f}(0) + \tilde{f}(-0{,}5)]/h^2 - \tilde{f}''(0)/(4h^2).$$

В силу теореми вкладення маємо

$$|R^{(1)}(\tilde{f})| \leqslant \frac{1}{2h} \|\tilde{f}\|_{C^1(\bar{E})} \leqslant Mh^{-1} \|\tilde{f}\|_{W_2^2(E)} \leqslant Mh^{-1} \|\tilde{f}\|_{W_2^3(E)},$$

$$|R^{(2)}(\tilde{f})| \leqslant Mh^{-2} \|\tilde{f}\|_{C^2(\bar{E})} \leqslant Mh^{-2} \|\tilde{f}\|_{W_2^3(E)} \leqslant Mh^{-2} \|\tilde{f}\|_{W_2^4(E)},$$

що означає обмеженість лінійних функціоналів $R^{(1)}$ і $R^{(2)}$ у просторах W_2^3 і W_2^4 відповідно.

Покажемо, що функціонал $R^{(1)}$ перетворюється на нуль у многочленах до другого степеня включно. Для цього досить переконатися, що він перетворюється на нуль у многочленах 1, s, s^2. Маємо

$$R^{(1)}(1) = [1 - 1]/(2h) - 0/(2h) = 0, \quad R^{(1)}(s) = [0{,}5 - (-0{,}5)]/(2h) -$$
$$- 1/(2h) = 0, \quad R^{(1)}(s^{(2)}) = [(0{,}5)^2 - (-0{,}5)^2]/(2h) - 2 \cdot 0/(2h) = 0.$$

Аналогічно переконуємося, що лінійний функціонал $R^{(2)}$ перетворюється на нуль у многочленах до третього степеня включно. Тому в силу леми Бремблa — Гільберта маємо

$$|R^{(1)}(f)| \leqslant M\bar{M}h^{-1} |\tilde{f}|_{W_2^3(E)},$$

$$|R^{(2)}(f)| \leqslant M\bar{M}h^{-2} |\tilde{f}|_{W_2^4(E)},$$

де сталі M, $\bar{M}$ не залежать від h і f. Переходячи до змінної ξ, з цих оцінок дістаємо твердження теореми. На практиці конкретні формули чисельного диференціювання можна будувати також іншим способом. Нехай, наприклад, потрібно обчислити похідну $f'(x)$ в точці x_i, знаючи значення функції $f(x)$ у точках x_i, $x_{i+1} = x_i + h$, $x_{i+2} = x_i + 2h = x_{i+1} + h$. Для цього будуємо інтерполяційний многочлен за точками x_i, x_{i+1}, x_{i+2} (запишемо його в формі Лагранжа):

$$p_2(x;\ f) = f(x_i) \frac{(x - x_{i+1})(x - x_{i+2})}{(x_i - x_{i+1})(x_i - x_{i+2})} + f(x_{i+1}) \times$$
$$\times \frac{(x - x_i)(x - x_{i+2})}{(x_{i+1} - x_i)(x_{i+1} - x_{i+2})} + f(x_{i+2}) \frac{(x - x_i)(x - x_{i+1})}{(x_{i+2} - x_i)(x_{i+2} - x_{i+1})} =$$
$$= f(x_i) \frac{(x - x_{i+1})(x - x_{i+2})}{2h^2} -$$
$$- f(x_{i+1}) \frac{(x - x_i)(x - x_{i+2})}{h^2} + f(x_{i+2}) \frac{(x - x_i)(x - x_{i+1})}{2h^2}. \tag{23}$$

Функцію $f(x)$ можна подати у вигляді

$$f(x) = p_2(x;\ f) + R(x;\ f), \tag{24}$$

де

$$R(x;\ f) = \frac{f'''(\xi)}{3!}(x - x_i)(x - x_{i+1})(x - x_{i+2}),$$
$$\xi = \xi(x) \in [y_1, y_2], \quad y_1 = \min(x, x_i, x_{i+1}, x_{i+2}), \tag{25}$$
$$y_2 = \max(x;\ x_i;\ x_{i+1};\ x_{i+2}).$$

Продиференціюємо рівність (24) і покладемо $x = x_i$:

$$f'(x_i) = p_2'(x_i;\ f) + R'(x_i;\ f), \tag{26}$$

де з урахуванням (23), (25) маємо

$$p_2'(x_i;\ f) = \left[f(x_i)\left(\frac{x-x_{i+2}}{2h^2}+\frac{x-x_{i+1}}{2h^2}\right) - f(x_{i+1})\left(\frac{x-x_i}{h^2}+\frac{x-x_{i+2}}{h^2}\right) + f(x_{i+2})\left(\frac{x-x_i}{2h^2}+\frac{x-x_{i+1}}{2h^2}\right)\right]_{x=x_i} = \frac{-3f(x_i)+4f(x_{i+1})-f(x_{i+2})}{2h},$$

$$R'(x_i;\ f) = \left[\frac{f^{IV}(\xi)}{3!}\xi_k'(x-x_i)(x-x_{i+1})(x-x_{i+2}) + \frac{f'''(\xi)}{3!}((x-x_{i+1})(x-x_{i+2})+(x-x_i)(x-x_{i+2})+(x-x_i)(x-x_{i+1}))\right]_{x=x_i} = 2\frac{f'''(\xi)}{3!}h^2.$$

Таким чином,

$$f'(x_i) = \frac{-3f(x_i)+4f(x_{i+1})-f(x_{i+2})}{2h} + O(h^2). \qquad (27)$$

Аналогічно за допомогою інтерполяційного многочлена за вузлами x_i, $x_{i-1} = x_i - h$, $x_{i-2} = x_i - 2h$ можна знайти формулу

$$f'(x_i) = \frac{3f(x_i)-4f(x_{i-1})+f(x_{i-2})}{2h} + O(h^2). \qquad (28)$$

Так само можна будувати формули різного порядку точності для інших похідних за різними конфігураціями вузлів. Як показує доведення теореми 2, зменшення гладкості функції веде до зменшення порядку точності формул чисельного диференціювання.

Вправа 3. За допомогою леми Брембла — Гільберта оцінити похибку односторонніх формул для обчислення першої похідної в припущенні, що $f(x) \in W_2^2(\Omega)$.

2.9.2. Обчислювальна похибка формул чисельного диференціювання. При використанні формул чисельного диференціювання доводиться віднімати близькі значення функції, що приводить до знищення перших значущих цифр і, як наслідок, до втрати частини достовірних знаків результату. Розглянемо, як веде себе похибка чисельного диференціювання, коли значення функції відомі не точні, а з похибкою $\delta f(x)$.

Як випливає з попереднього, при використанні інтерполяційного многочлена для знаходження k-ї похідної функції $f(x)$ дістаємо формулу виду (на рівномірній сітці з кроком h)

$$f^{(k)}(x) = h^{-k}\sum_q c_q(x) f(x_q) + R^{(k)}(x),$$

причому $c_q(x) = O(1)$. Якщо формула має порядок точності p, то

$$R^{(k)}(x) \approx c(x)\,h^p,$$

і цей залишковий член визначає похибку методу. Очевидно, що мажоранта неусувної похибки $r^{(k)}(x)$ за рахунок наближеного задання функції $f(x)$ має вигляд

$$r^{(k)} = \delta h^{-k}\sum_q |c_q|$$

і необмежено зростає при $h \to 0$ (фактично ж неусувна похибка $r^{(k)}(x)$ нерегулярно залежить від величини кроку, осцилюючи в межах, що визначаються мажорантою, рис. 16). Повна похибка мажорується сумою $R^{(k)} + r^{(k)}$ і оптимальний крок відповідає її мінімуму. Неважко знайти, що

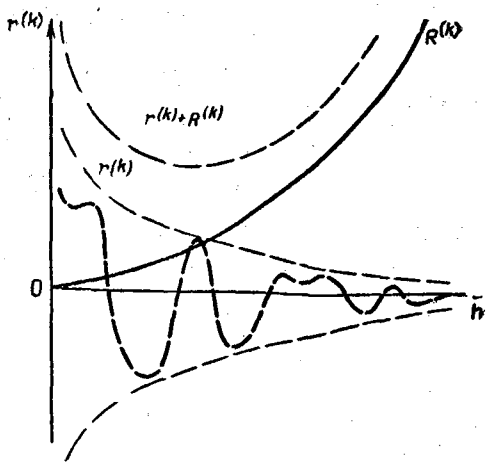


Рис. 16

$$(r^{(k)} + R^{(k)})_h' = -k\delta h^{-k-1}\sum_q |c_q| + ph^{p-1}c = h^{p-1}\times\left(-k\delta h^{-(k+1)-(p-1)}\sum_q |c_q| + pc\right) = 0,$$

$$h_0(\delta) = \left(k\delta\sum_q |c_q|/(pc)\right)^{1/(p+k)} = O(\delta^{1/(p+k)}),$$

$$\min(R^{(k)} + r^{(k)}) = ch_0^p\left[1 + \left(\delta\sum_q c_q\right)h_0^{-(k+p)}c^{-1}\right] = ch_0^p\left[1 + \frac{p}{k}\right] = O(\delta^{p/(p+k)}).$$

Як бачимо, мінімальна похибка тим менша, чим менші похибка вхідних даних δ, порядок похідної k і чим більший порядок точності формули p. Із наведеного аналізу випливає також важливий практичний висновок: при $\delta f \to 0$ можна дістати як завгодно високу точність результату, якщо крок сітки h прямує до нуля, залишаючись не меншим за $h_0(\delta)$; якщо ж припустити $h < h_0(\delta)$, то результат граничного переходу може бути неправильним. Вказаний спосіб вибору кроку $h \geqslant h_0(\delta)$ є деякою регуляцією диференціювання, яка називається *регуляцією за кроком сітки.*

2.10. Апостеріорні оцінки похибки

2.10.1. Метод Рунге — Ромберга. Загальна ідея методу така: маємо деяку наближену формулу $\zeta(x, h)$ для обчислення величини $z(x)$ за її значеннями на рівномірній сітці з кроком h, а залишковий член цієї формули

$$z(x) - \zeta(x, h) = \psi(x) h^p + O(h^{p+1}). \tag{1}$$

Наприклад, $z(x) = f'(x)$, $f(x)$ — задана функція

$$f \in C^5[a, b], \quad \zeta(x, h) = (f(x+h) - f(x-h))/(2h) \equiv f_{\circ x}(x),$$

$$z(x) - \zeta(x, h) = f'(x) - f_{\circ x}(x) = -\frac{h^2}{6} f^{(3)}(x) - \frac{h^4}{60} f^{(5)}(\xi),$$

$\xi \in [x-h, x+h]$. Тут $p = 2$. Якщо скористатися тією самою наближеною формулою для обчислення значення z в точці x, але використовуючи сітку з кроком rh, дістанемо

$$z(x) - \zeta(x, rh) = \psi(x)(rh)^p + O((rh)^{p+1}) = \psi(x)(rh)^p + O(h^{p+1}). \tag{2}$$

Віднявши (1) від (2), дістанемо першу формулу Рунге для оцінки похибки

$$R \approx \psi(h) h^p = \frac{\zeta(x, h) - \zeta(x, rh)}{r^p - 1} + O(h^{p+1}). \tag{3}$$

Перший доданок у (3) є головним членом похибки, тобто розрахунок на другій сітці дає змогу оцінити похибки на першій сітці з точністю до членів вищого порядку. Виключаючи за допомогою (3) величину $\psi(x) h^p$ з (1) дістанемо другу формулу Рунге

$$z(x) = \zeta(x, h) + \frac{\zeta(x, h) - \zeta(x, rh)}{r^p - 1} + O(h^{p+1}), \tag{4}$$

яка дає результат з вищим порядком точності, ніж (1). Іноді уточнення результату за формулою (4) називають *уточненням за Річардсоном*. Розглянемо приклади застосування описаного вище процесу для підвищення точності в задачі диференціювання.

Приклад 1. Нехай функція $y(x) = \lg x$ задана таблицею

x	$y = \lg x$
1	0,000
2	0,301
3	0,478
4	0,602
5	0,699

Обчислити $y'(3)$.

Р о з в ' я з а н н я. Скориставшись формулою (13) (п. 2.9) при $h = 1$ дістанемо

$$y'(3) = [y(4) - y(2)]/2 \approx 0{,}151.$$

Збільшуючи крок вдвічі ($r = 2$), дістанемо

$$y'(3) = [y(5) - y(1)]/2 \cdot 2 \approx 0{,}175.$$

За формулою (3) при $p = 2$

$$y'(3) \approx 0{,}143,$$

що лише на 2 % відрізняється від шуканого значення $y'(3) = 0{,}145$.

Приклад 2. За допомогою методу Рунге вивести формулу чисельного диференціювання (2.9.16) порядку $O(h^4)$ з формули (13) (п. 2.9) більш низького порядку $O(h^3)$.

Р о з в ' я з а н н я. Маємо

$$f'_{3/2}(h) \approx (f_2 - f_1)/h, \quad f'_{3/2}(3h) \approx (f_3 - f_0)/(3h).$$

Порядок точності цих формул $p = 2$, а коефіцієнт збільшення кроку $r = 3$, тому уточнення за методом Рунге дає формулу (2.9.16)

$$f'_{3/2} \approx f'_{3/2}(h) + \frac{1}{8}[f'_{3/2}(h) - f'_{3/2}(3h)] = \frac{1}{24h}(f_0 - 27f_1 + 27f_2 - f_3).$$

Як бачимо, для обчислення результату більш високого порядку точності не обов'язково використовувати безпосередньо формули високого порядку точності; можна виконати обчислення за простими формулами низької точності на різних сітках і потім уточнити результат за методом Рунге. Такий спосіб має перевагу ще й тому, що величина поправки (3) дає апостеріорну оцінку точності.

Метод Рунге узагальнюється на довільну кількість сіток.

Приклад 3. За допомогою розвинення в ряд Тейлора для функції $f(x) \in C^{2m+3}[x-a, x+a]$ і $|h| \leqslant a$ дістаємо

$$z(x) - \zeta(x, h) \equiv f'(x) - f_{\circ x}(x) = \psi_1(x) h^2 + \psi_2(x) h^4 + \cdots + \psi_m(x) h^{2m} + h^{2m+2} \tilde{\psi}_{m+1}(x, h), \quad \psi_k(x) = f^{(2k+1)}(x)/(2k+1)!$$

$$\tilde{\psi}_{m+1}(x, h) = \psi_{m+1}(x) + O(1), \quad k = \overline{1, m+1}. \tag{5}$$

Приклад 4. Для односторонньої різницевої похідної $f_x(x) = (f(x+h) - f(x))/h \equiv \zeta(x, h)$ при $f(x) \in C^{m+3}[x, x+a]^x$, $|h| \leqslant a$ маємо

$$z(x) - \zeta(x, h) \equiv f'(x) - f_x(x) = \psi_1(x) h + \psi_2(x) h^2 + \cdots + \psi_m(x) h^m + h^{m+1}(\psi_{m+1}(x) + O(1)), \quad \psi_k(x) = -f^{(k+1)}(x)/(k+1)!,$$

$$k = \overline{0, m+1}. \tag{5'}$$

Нехай розрахунки виконано на q різних сітках h_j, $1 \leqslant j \leqslant q$. Тоді із залишкового члена (5) можна вилучити $q - 1$ складових. Для цього перепишемо (5) у вигляді

$$z(x) - \sum_{m=p}^{p+q-2} \psi_m(x) h_j^m = \zeta(x, h_j) + O(h^{p+q-1}), \quad h_j \leqslant h, \quad 1 \leqslant j \leqslant q. \tag{5''}$$

Це система лінійних рівнянь відносно величин $z(x)$ і $\psi_m(x)$, $m = p, p+1, \ldots, p+q-2$. Використавши формули Крамера, дістанемо уточнений розв'язок за формулою Ромберга:

$$z(x) = \Delta^{-1}\begin{vmatrix} \zeta(x, h_1) & h_1^p & h_1^{p+1} & \ldots & h_1^{p+q-2} \\ \zeta(x, h_2) & h_2^p & h_2^{p+1} & \ldots & h_2^{p+q-2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \zeta(x, h_q) & h_q^p & h_q^{p+1} & \ldots & h_q^{p+q-2} \end{vmatrix} + O(h^{p+q-1}), \qquad (6)$$

де

$$\Delta = \begin{vmatrix} 1 & h_1^p & h_1^{p+1} & \ldots & h_1^{p+q-2} \\ 1 & h_2^p & h_2^{p+1} & \ldots & h_2^{p+q-2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & h_q^p & h_q^{p+1} & \ldots & h_q^{p+q-2} \end{vmatrix}$$

(вважаємо, що $\Delta \neq 0$). Ця формула виражає $z(x)$ через обчислені з точністю до $O(h^p)$ величини $\zeta(x, h_i)$ з більш високою точністю $O(h^{p+q-1})$ (тобто розрахунок на кожній новій сітці дає змогу підвищити порядок точності на одиницю). Розкладаючи визначник за першим стовпчиком, формулу для $z(x)$ можна записати також у вигляді

$$z(x) = \sum_{i=1}^{q} \zeta(x, h_i) L_{i,q}(0) + O(h^{p+q-1}),$$

де

$$L_{i,q}(h) = \Delta^{-1}\begin{vmatrix} 1 & h_1^p & h_1^{p+1} & \ldots & h_1^{p+q-2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & h_{i-1}^p & h_{i-1}^{p+1} & \ldots & h_{i-1}^{p+q-2} \\ 1 & h^p & h^{p+1} & \ldots & h^{p+q-2} \\ 1 & h_{i+1}^p & h_{i+1}^{p+1} & \ldots & h_{i+1}^{p+q-2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & h_q^p & h_q^{p+1} & \ldots & h_q^{p+q-2} \end{vmatrix}.$$

Функції $L_{i,q}(h)$ мають, очевидно, такі дві властивості:

а) $L_{i,q}(h_j) = \delta_{ij}$, δ_{ij} — символ Кронекера;

б) $L_{i,q}(h) = \alpha_0^{(i)} + \alpha_1^{(i)} h^p + \alpha_2^{(i)} h^{p+1} + \ldots + \alpha_{q-1}^{(i)} h^{p+q-2}$ ($\alpha_j^{(i)}$ — дійсні коефіцієнти), тобто $L_{i,q}(h)$ є многочленами від h. Тому функція

$$P(h) \equiv P(h; \zeta) = \sum_{i=1}^{q} \zeta(x, h_i) L_{i,q}(h) \equiv$$
$$\equiv \beta_0 + \beta_1 h^p + \cdots + \beta_{q-1} h^{p+q-2} \qquad (7)$$

є інтерполюючою функцією для $\zeta(x, h)$ (β_i — дійсні коефіцієнти), а величина

$$\sum_{i=1}^{q} \zeta(x, h_i) L_{i,q}(0) \approx z(x)$$

є значенням цієї функції при $h = 0$, причому $h = 0$ не належить найменшому інтервалу $l[h_1, h_2, \ldots, h_q]$, що охоплює всі точки $h_1, h_2, \ldots, h_q$. З цієї причини у випадку методу Рунге — Ромберга говорять також про екстраполяцію. Вживають також терміни «екстраполяція за Річардсоном», «екстраполяція до нуля», «екстраполяція до кроку нуль».

Оскільки система функцій $\varphi_0(h) = 1$, $\varphi_1(h) = h^p, \ldots, \varphi_{q-1}(h) = h^{p+q-2}$ не при всіх p і не на довільному інтервалі буде системою Чебишева, то інтерполяційна функція (7) існує не для будь-якої послідовності $h_1, \ldots, h_q$. Але для послідовностей, які найчастіше зустрічаються на практиці, а саме: а) h_1, $h_2 = \alpha h_1$, $h_3 = \alpha h_2 = \alpha^2 h_1$, ... (послідовність Рунге — Ромберга); б) h_1, $h_2 = h_1/2$, $h_3 = h_1/3$, ... можна довести, що $\Delta \neq 0$, і тим самим існування многочлена $P(x)$ гарантується.

Зауваження. 1. Формула Рунге — Ромберга має ту перевагу, що вона може бути застосована для довільних кроків h_j та числа сіток q (за умови $\Delta \neq 0$). Недоліком її є те, що потрібно розв'язувати систему лінійних алгебраїчних рівнянь і в проміжних розрахунках не контролюється точність.

2. Метод Ромберга можна застосувати не лише для розвинення вигляду

$$z(x) - \zeta(x, h) = \sum_{m \geq 0} \psi_m(x) \varphi_m(h)$$

з функціями $\varphi_m(h) = h^{p+m}$, як в (5), але й для довільних функцій $\varphi_m(h)$, $\varphi_m(0) = 0$. Коли $\varphi_m(h) = h^{\gamma(m+1)}$ ($\gamma = \text{const}$), що часто зустрічається на практиці, тоді інтерполяційний многочлен $P(h)$ має вигляд

$$P(h) \equiv P(s) = \beta_0 + \beta_1 s + \beta_2 s^2 + \cdots + \beta_{q-1} s^{q-1},$$
$$s = h^{\gamma}$$

і може бути обчислений в точці $s = 0$ ($h = 0$) за допомогою, наприклад, алгоритму Ньютона без розв'язування системи (5″).

3. Якщо сітки такі, що $h_j = r h_{j-1} = \ldots = r^{j-1} h_1$, тобто згущення їх відбувається в одну і ту саму кількість разів, то зручніше застосувати рекурентно метод Рунге. Це робиться таким чином. Спочатку на кожній парі сіток (h_1, h_2), (h_2, h_3), ..., (h_{q-1}, h_q) методом Рунге вилучають головний член похибки $\psi_q(x) h^p$. Уточнені значення таким чином групуються в пари і далі вилучається похибка наступного

порядку $O(h^{p+1})$. Всього можна виконати $q-1$ уточнень. При кожному уточненні за формулою (3) обчислюється апостеріорна оцінка точності.

4. Якщо формула для обчислення $\zeta(x, h)$ має симетричний вигляд, то на рівномірній сітці часто всі непарні члени ряду (5) перетворюються на нуль (див. приклад 3). В такій ситуації користуватися формулою (6) невигідно. Слід залишити в сумі (5) члени $\psi_p h^p$, $\psi_{p+2}h^{p+2}$, $\psi_{p+4}h^{p+4}$, ... і відповідно змінити формулу Ромберга. Те саме стосується і рекурентної процедури Рунге — при черговому вилученні порядок точності підвищується на 2, а не на 1.

5. Число членів суми (5) зв'язане з кількістю неперервних похідних у функції, для якої обчислюються $z(x)$ і $\zeta(x, h)$ (див. приклади 4, 5). Для не досить гладких функцій недоцільно брати велике число сіток для уточнення. Практично навіть для гладких функцій використовують не більше 3—5 сіток, причому, як правило, беруть відношення r кроків сіток рівним 2.

6. Метод Рунге — Ромберга можна застосувати лише тоді, коли вірно (5), причому коефіцієнти $\psi_m(x)$ однакові для всіх сіток. Для формул чисельного диференціювання ці коефіцієнти залежать від положення вузлів сітки. Але якщо вибрані конфігурації вузлів на всіх сітках подібні відносно точки x, то залежність від вузлів однакова. У такому разі метод Рунге — Ромберга можна застосувати, в інших же випадках його застосувати неможливо. Тому при чисельному диференціюванні метод Рунге — Ромберга застосовується лише для знаходження похідних у вузлах або в середніх точках інтервалів рівномірних і на деяких «близьких» до них сіток. Це так звані квазірівномірні сітки, які добирають так, щоб «найкращим чином» передати поведінку конкретної функції. Сітка (в змінних x) називається *квазірівномірною,* якщо існує двічі неперервно диференційовна функція $x = \xi(t)$, яка переводить відрізок $0 \leqslant t \leqslant 1$ у відрізок $a \leqslant x \leqslant b$ так, що кожній сітці $x_i^{(N)}$ відповідає рівномірна сітка $t_i^{(N)} = i/N$, причому на цьому відрізку $\xi'(t) \geqslant \varepsilon > 0$, а $\xi''(t)$ обмежена. Якщо ці умови виконано, то крок сітки $h_i \approx \xi'(t_i)/N$, а різниця двох сусідніх кроків $h_i - h_{i-1} \approx \xi''(t_i)/N^2$, тобто при значній кількості вузлів різниця сусідніх кроків є величина порядку $O(h^2)$ і сусідні інтервали майже рівні (хоча відношення довжин далеких один від одного інтервалів $h_i/h_j \approx \xi'(t_i)/\xi'(t_j)$ може бути великим).

2.10.2. Процес Ейткена. Метод розрахунків на декількох сітках застосовується для підвищення порядку точності і в тому випадку, коли невідомий порядок головного члена похибки, і носить назву *процесу Ейткена.* Нехай

$$z(x) - \zeta(x, h) = \psi_p(x)h^p + \psi_q(x)h^q + \cdots, \quad q > p,$$

але p — невідоме.

Проводяться обчислення на трьох сітках з кроками $h_1 = h$, $h_2 = \rho h$, $h_3 = \rho^2 h$ $(0 < \rho < 1)$. Нехтуючи членами порядку $O(h^q)$, дістаємо

$$A = \frac{\zeta(x, h_1) - \zeta(x, h_2)}{\zeta(x, h_2) - \zeta(x, h_3)} \approx \frac{h_1^p - h_2^p}{h_2^p - h_3^p} = \frac{1 - \rho^p}{\rho^p(1 - \rho^p)} = \left(\frac{1}{\rho}\right)^p.$$

Звідси знаходимо

$$p \approx \frac{\ln A}{\ln(1/\rho)}.$$

Далі можна скористатися вже відомим методом Рунге, який можна трактувати таким чином. Утворимо комбінацію $\tilde{\zeta}(x, h) = \sigma\zeta(x, h_1) + (1 - \sigma)\xi(x, h_2)$ і виберемо σ так, щоб $\sigma h_1^p + (1 - \sigma)h_2^p = (\sigma + (1 - \sigma)\rho^p)h^p = 0$. Дістанемо $\sigma = \rho^p/(\rho^p - 1) = 1/(1 - A)$, причому $\tilde{\zeta}(x, h) - z(x) = O(h^q)$.

ГЛАВА 3
НАБЛИЖЕННЯ ФУНКЦІЙ В ЛІНІЙНИХ НОРМОВАНИХ ПРОСТОРАХ

3.1. Класифікація методів наближення функцій

Задача наближення функцій часто зустрічається на практиці. При цьому вимоги до якості наближення і відповідно методи наближення функцій можуть бути дуже різними. За принципами наближень

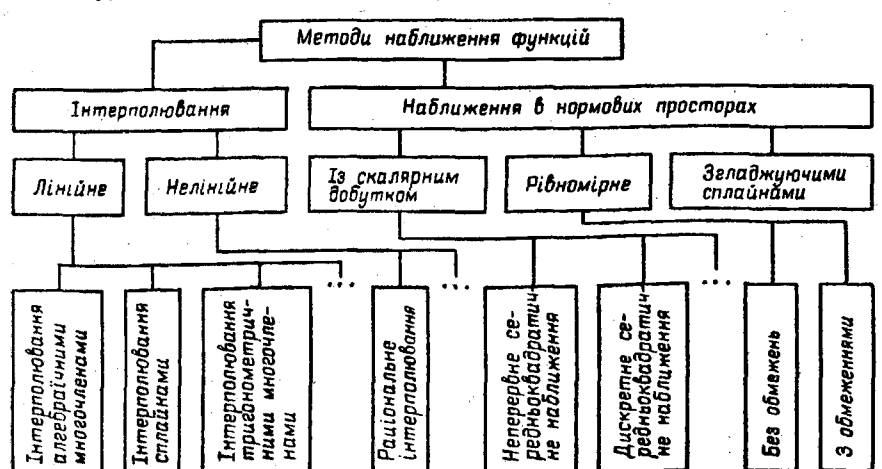


Рис. 17

і методами побудови цих наближень можна дати таку класифікацію методів, яка не може претендувати на повноту і носить частково суб'єктивний характер (рис. 17).

Зауважимо, що, як показує нерівність Лебега, інтерполяційний многочлен можна розглядати як деяке наближення (при виборі вузлів у нулях многочлена Чебишева), близьке до найкращого для заданої функції в просторі $C[a, b]$. Теорема 8 (п. 2.4) показує, що інтерполяційний многочлен можна розглядати як наближення до заданої функції $f(x)$ і в просторі $L_{2,\rho}[a, b]$. Як ми побачимо далі, інтерполяційний многочлен за певних умов можна розглядати і як многочлен найкращого дискретного середньоквадратичного наближення. Але історично задача інтерполювання не ставилась спочатку як задача наближення в нормованому просторі. Крім того, певна специфіка інтерполювання дозволяє виділити цей спосіб наближення окремо.

3.2. Постановка задачі. Наближення функцій в лінійному нормованому просторі. Умови існування і єдиності елемента найкращого наближення

Нехай R — лінійний нормований простір функцій і $f \in R$. Нехай $\varphi_i \in R$, $i = \overline{0, n}$,— лінійно незалежні функції, а M — лінійний підпростір узагальнених многочленів

$$\Phi = \sum_{i=0}^{n} c_i \varphi_i \tag{1}$$

з дійсними коефіцієнтами c_i.

Числова множина

$$\Delta(f, \Phi) = \|f - \Phi\|, \tag{2}$$

де f — фіксований елемент з R, $\Phi \in M$, обмежена знизу нулем. Тому існує число $\Delta(f)$ таке, що

$$\Delta(f) = \inf_{\Phi \in M} \Delta(f, \Phi). \tag{3}$$

Виникає запитання: чи існує елемент $\Phi_0 \in M$ такий, що

$$\Delta(f) = \|f - \Phi_0\|? \tag{4}$$

Будь-який елемент Φ_0, для якого виконується (4), називатимемо *елементом найкращого наближення для* f у підпросторі M нормованого простору R або проекцією f на M.

3.2.1. Теорема існування. Відповідь на поставлене вище запитання дає така теорема.

Теорема 1. *Для будь-якого елемента $f \in R$ в M існує елемент найкращого наближення, причому множина таких елементів опукла.*

Доведення. З нерівностей

$$\left\| \left\| f - \sum_{i=0}^{n} c_i^{(1)} \varphi_i \right\| - \left\| f - \sum_{i=0}^{n} c_i^{(2)} \varphi_i \right\| \right\| \leqslant$$

$$\leqslant \left\| \sum_{i=0}^{n} [c_i^{(1)} - c_i^{(2)}] \varphi_i \right\| \leqslant \sum_{i=0}^{n} |c_i^{(1)} - c_i^{(2)}| \, \|\varphi_i\|$$

випливає, що

$$F(c_0, \ldots, c_n) = \left\| f - \sum_{i=0}^{n} c_i \varphi_i \right\|$$

є неперервною функцією аргументів c_i при будь-якому $f \in R$. Нехай $\|\vec{c}\| = \left(\sum_{i=0}^{n} c_i^2 \right)^{1/2}$ — евклідова норма вектора $\vec{c} = (c_0, \ldots, c_n)$. Функція

$$F_0(c_0, \ldots, c_n) = \left\| \sum_{i=0}^{n} c_i \varphi_i \right\|$$

також неперервна на одиничній сфері $\|\vec{c}\| = 1$ і тому в деякій її точці $(\tilde{c}_0, \tilde{c}_1, \ldots, \tilde{c}_n)$ досягає своєї нижньої грані $\tilde{F}$ (по сфері). Оскільки рівність $\tilde{F} = \|\tilde{c}_0 \varphi_0 + \ldots + \tilde{c}_n \varphi_n\| = 0$ суперечить лінійній незалежності елементів φ_i, то $\tilde{F} \neq 0$. Для будь-якого $\vec{c} = (c_0, \ldots, c_n) \neq (0, 0, \ldots, 0)$ справедлива оцінка

$$\|c_0 \varphi_0 + \cdots + c_n \varphi_n\| = F_0(c_0, \ldots, c_n) =$$

$$= \|\vec{c}\| F_0\left(\frac{c_0}{\|\vec{c}\|}, \ldots, \frac{c_n}{\|\vec{c}\|} \right) \geqslant \|\vec{c}\| \tilde{F}. \tag{5}$$

Нехай $\gamma > \dfrac{2\|f\|}{\tilde{F}}$. Функція $F(c_0, \ldots, c_n)$ неперервна в замкненій кулі $\|\vec{c}\| \leqslant \gamma$ і тому в деякій її точці $(c_0^{(0)}, \ldots, c_n^{(0)})$ досягає своєї нижньої грані F_*, причому $F_* \leqslant F(0, 0, \ldots, 0) = \|f\|$. Зовні цієї кулі в силу (5)

$$F(c_0, \ldots, c_n) \geqslant \|c_0 \varphi_0 + \cdots + c_n \varphi_n\| - \|f\| =$$

$$= \|\vec{c}\| \left\| \frac{c_0}{\|\vec{c}\|} \varphi_0 + \cdots + \frac{c_n}{\|\vec{c}\|} \varphi_n \right\| - \|f\| > \frac{2\|f\|}{\tilde{F}} \tilde{F} - \|f\| = \|f\| \geqslant F_*.$$

Таким чином,

$$F(c_0, \ldots, c_n) \geqslant F_* = F(c_0^{(0)}, \ldots, c_n^{(0)}) \; \forall \vec{c},$$

і першу частину теореми доведено.

Якщо

$$\Phi_0 = \sum_{i=0}^{n} c_i \varphi_i, \quad \tilde{\Phi}_0 = \sum_{i=0}^{n} \tilde{c}_i \varphi_i$$

є елементами найкращого наближення, то

$$\|f - \Phi_0\| = \|f - \tilde{\Phi}_0\| = \Delta(f)$$

і у випадку $\Delta(f) = 0$ маємо $\Phi_0 = \tilde{\Phi}_0 = f$. Якщо ж $\Delta(f) > 0$, то нехай m — точка відрізка, що сполучає елементи Φ_0 та $\tilde{\Phi}_0$, тобто

$$m = a\Phi_0 + b\tilde{\Phi}_0, \quad a, b \geqslant 0, \quad a + b = 1,$$

тоді

$$\Delta(f) \leqslant \|f - m\| = \|a(f - \Phi_0) + b(f - \tilde{\Phi}_0)\| \leqslant$$

$$\leqslant a\|f - \Phi_0\| + b\|f - \tilde{\Phi}_0\| = \Delta(f),$$

отже, $\|f - m\| = \Delta(f)$, тобто m є елементом найкращого наближення. Це і означає опуклість множини многочленів найкращого наближення. Теорему доведено.

Наведемо приклад, який показує, що елемент найкращого наближення в нормованому просторі може бути не єдиним (навіть у скінченновимірному випадку).

Приклад. У просторі $l_1^{(2)}$ двовимірних векторів $x = (\xi_1, \xi_2)$ з нормою $\|x\| = |\xi_1| + |\xi_2|$ візьмемо елемент $x_0 = (1, -1)$ і одновимірний підпростір M_1 з базисним вектором $e = (1, 1)$, тобто $M_1 = \{\alpha e : \alpha \in R^1\}$. Неважко помітити, що

$$\Delta(x_0) = \inf_{\alpha \in R^1} \|x_0 - \alpha e\| = \inf_{\alpha \in R^1} (|1 - \alpha| + |1 + \alpha|) = 2,$$

і це значення досягається при $\alpha \in [-1, 1]$. Отже, в даному випадку маємо нескінченну множину елементів найкращого наближення $\Phi_0 = \alpha e$, $\alpha \in [-1, 1]$, для елемента $x_0 = (1, -1)$.

3.2.2. Достатня умова єдиності елемента найкращого наближення. Розглянемо питання про умови єдиності елемента найкращого наближення. Одну з достатніх умов дає наступна теорема.

Теорема 2. *Якщо простір R строго нормований, то елемент найкращого наближення єдиний.*

Доведення. Припустимо від супротивного, що існують два елементи f_1, f_2 такі, що

$$\|f - f_1\| = \|f - f_2\| = \Delta(f), \quad f_1 \neq f_2.$$

Очевидно, $\Delta \neq 0$, бо інакше $f_1 = f = f_2$. В силу того що множина елементів найкращого наближення опукла, елемент $(f_1 + f_2)/2$ також є елементом найкращого наближення і

$$\left\|f - \frac{f_1 + f_2}{2}\right\| = \left\|\frac{f - f_1}{2} + \frac{f - f_2}{2}\right\| = \Delta(f) =$$

$$= \frac{\Delta(f)}{2} + \frac{\Delta(f)}{2} = \left\|\frac{f - f_1}{2}\right\| + \left\|\frac{f - f_2}{2}\right\|.$$

Оскільки простір R строго нормований, то $\frac{f - f_1}{2} = \alpha \frac{f - f_2}{2}$. Якщо $\alpha = 1$, то $f_1 = f_2$ і дістанемо суперечність.

Якщо ж $\alpha \neq 1$, то $f = \frac{1}{1 - \alpha} f_1 - \frac{\alpha}{1 - \alpha} f_2 \in M$, тому $\Delta(f) = 0$, що знову приводить до суперечності $f_1 = f_2 = f$.
Теорему доведено.

Вправа 1. Довести, що будь-який простір із скалярним добутком є строго нормованим.

Вправа 2. Довести, що простір $L_p[0, 1]$ з нормою $\|f\|_{L_p[0,1]} = \left(\int_0^1 |f|^p dx\right)^{1/p}$ строго нормований при $1 < p < \infty$.

Вказівка. Розглянути умови, за яких досягається знак рівності в нерівності Мінковського (див. також п. 3.10).

Вправа 3. Довести, що простори $L_1[0, 1]$ і $C[0, 1]$ не є строго нормованими.

Вказівка. У просторі $C[a, b]$ розглянути функції $f_1(x) = (x - a)(b - x)$ і $f_2(x) = \frac{(b - a)^2}{4} \sin \pi \frac{x - a}{b - a}$. У просторі $L_1[0, 1]$ розглянути функції $f_1(x) = 1$ і $f_2(x) = 2x$, для яких $\|f_1 + f_2\|_{L_1} = \|f_1\|_{L_1} + \|f_2\|_{L_1}$, але f_1, f_2 лінійно незалежні.

3.2.3. Характеристика елемента найкращого наближення в просторі із скалярним добутком. Побудова елементів найкращого наближення грунтується на теоремах, які встановлюють ті властивості цих елементів, якими вони відрізняються від усіх інших. Такі властивості і алгоритми, які на них грунтуються, є особливо простими у випадку просторів із скалярним добутком.

Нехай M — лінійний підпростір простору із скалярним добутком H, $f \in H$. Розглянемо властивості, які має елемент $h_0 \in M$ такий, що

$$\|f - h_0\| = \inf_{h \in M} \|f - h\|,$$

де $\|f\| = (f, f)^{1/2}$ (символ $(,)$ позначає скалярний добуток в H).

Теорема 3. *Нехай в M існує елемент h_0 найкращого наближення для f. Тоді різниця $f - h_0$ ортогональна до всіх елементів підпростору M.*

Доведення. Припустимо, що $h_1 \in M$, для якого $(f - h_0, h_1) = \alpha \neq 0$. Оскільки $h_1/\|h_1\| \in M$, то можна вважати $\|h_1\| = 1$. Розглянемо елемент $h_2 = h_0 + \alpha h_1 \in M$. Маємо

$$\|f - h_2\|^2 = (f - h_0 - \alpha h_1, f - h_0 - \alpha h_1) = (f - h_0, f - h_0) -$$

$$- \alpha(h_1, f - h_0) - \alpha(f - h_0, h_1) + \alpha^2(h_1, h_1).$$

Оскільки $(h_1, f - h_0) = (f - h_0, h_1) = \alpha$, то з попередньої рівності дістаємо

$$\|f - h_2\|^2 = \|f - h_0\|^2 - \alpha^2 - \alpha^2 + \alpha^2 = \|f - h_0\|^2 - \alpha^2 < \|f - h_0\|^2,$$

а це суперечить твердженню про те, що h_0 — елемент найкращого наближення.

Теорема 4. *Якщо $h_0 \in M$, $(f - h_0, h) = 0$ для будь-якого $h \in M$, то h_0 — елемент найкращого наближення.*

Доведення. Для довільного $h \in M$ маємо

$$\|f - h\|^2 = (f - h_0 + h_0 - h, f - h_0 + h_0 - h) = (f - h_0, f - h_0) + (h_0 - h, f - h_0) + (f - h_0, h_0 - h) + (h_0 - h, h_0 - h).$$

Оскільки $h_0 - h \in M$, то другий і третій доданки в силу умов теореми перетворюються на нуль і дістаємо

$$\|f - h\|^2 = \|f - h_0\|^2 + \|h - h_0\|^2. \tag{6}$$

Отже, $\|f - h\|^2 > \|f - h_0\|^2$ при $h \neq h_0$ і теорему доведено.

Оскільки простір із скалярним добутком строго нормований, то існування і єдиність елемента найкращого наближення вигляду $\Phi = \sum_{i=0}^{n} c_i \varphi_i$ випливає з теорем 1 і 2. Більш коротке доведення єдиності для просторів із скалярним добутком можна дістати і з теорем 3 і 4: якщо h_0 — елемент найкращого наближення, то $(f - h_0, h) = 0$ для будь-якого $h \in M$ і тоді відповідно до (6), будь-який елемент $h \neq h_0$, $h \in M$ не є елементом найкращого наближення.

3.3. Побудова елемента найкращого наближення в просторі із скалярним добутком

Нехай φ_i, $i = \overline{0, n}$,— система лінійно незалежних елементів. Лінійні комбінації $\sum_{i=0}^{n} c_i \varphi_i$ з дійсними коефіцієнтами утворюють лінійний підпростір M, причому якщо елемент $\varphi^{(n)} \equiv \varphi = \sum_{i=0}^{n} a_i \varphi_i$ є елементом найкращого наближення, то в силу теореми 3 (п. 2.3) попереднього параграфа

$$\left(f - \sum_{i=1}^{n} a_i \varphi_i, \varphi_j\right) = 0, \quad j = \overline{0, n}. \tag{1}$$

Оскільки елемент найкращого наближення існує, то система лінійних алгебраїчних рівнянь (1) відносно a_i, $i = \overline{0, n}$, має розв'язок. Покажемо тепер, що коли a_i, $i = \overline{0, n}$,— розв'язок системи (1), то елемент $\varphi = \sum_{i=0}^{n} a_i \varphi_i$ є єдиним елементом найкращого наближення.

Дійсно, нехай систему (1) розв'язано. Тоді

$$\left(f - \sum_{i=0}^{n} a_i \varphi_i, \sum_{i=0}^{n} c_i \varphi_i\right) = 0 \quad \forall c_i, \quad i = \overline{0, n},$$

або, що те саме,

$$\left(f - \sum_{i=0}^{n} a_i \varphi_i, h\right) = 0 \quad \forall h \in M.$$

В силу теореми 4 (п. 3.2) такий елемент $\varphi = \sum_{i=0}^{n} a_i \varphi_i$ є елементом найкращого наближення. Припустимо, що (1) має розв'язок $(b_0, \ldots, b_n)$, відмінний від $(a_0, a_1, \ldots, a_n)$. Тоді елемент $q = \sum_{i=0}^{n} b_i \varphi_i$ також є елементом найкращого наближення і в силу єдиності останнього $\sum a_i \varphi_i = \sum b_i \varphi_i$. Припущення, що $a_i \neq b_i$ при деякому i суперечить лінійній незалежності елементів φ_i і ми дійдемо суперечності.

Таким чином, ми показали, що у випадку простору із скалярним добутком задача побудови елемента найкращого наближення $\varphi = \sum_{i=0}^{n} a_i \varphi_i$ зводиться до розв'язування системи лінійних алгебраїчних рівнянь (1), яку можна переписати у вигляді

$$\sum_{i=0}^{n} a_i (\varphi_i, \varphi_j) = (f, \varphi_j), \quad j = \overline{0, n}. \tag{2}$$

Як ми вже довели, ця система невироджена, тобто має єдиний розв'язок. Матриця системи (2) $G(\varphi_0, \ldots, \varphi_n) = [(\varphi_i, \varphi_j)]_{i,j=1}^{n}$ називається *матрицею Грама* системи елементів φ_i, $i = \overline{0, n}$.

Вправа. Довести, що $G(\varphi_0, \ldots, \varphi_n) \geqslant 0$, причому $G(\varphi_0, \varphi_1, \ldots, \varphi_n) = 0$ тоді і лише тоді, коли $\varphi_0, \ldots, \varphi_n$ лінійно залежні. Довести нерівність

$$G(\varphi_0, \ldots, \varphi_{n+1}) \leqslant G(\varphi_0, \ldots, \varphi_n)(\varphi_{n+1}, \varphi_{n+1}).$$

Якщо елементи φ_i утворюють ортонормовану систему, тобто

$$(\varphi_i, \varphi_j) = \delta_{i,j}, \quad \delta_{i,j} = \begin{cases} 0, & i \neq j, \\ 1, & i = j \end{cases}$$

— символ Кронекера, то система (2) спрощується:

$$a_j = (f, \varphi_j), \quad j = \overline{0, n}. \tag{3}$$

Тоді найкраще наближене записується у вигляді

$$\varphi = \sum_{i=0}^{n} (f, \varphi_i) \varphi_i. \tag{4}$$

Нехай φ — найкраще наближення для f. Тоді за теоремою 3 (п. 3.2) маємо $(f-\varphi, h)=0\ \forall h\in M$ і, підставляючи в формулу (6) $h_0=\varphi$, $h=0$, дістаємо

$$\|f\|^2=\|f-\varphi\|^2+\|\varphi\|^2.$$

Для ортонормованої системи φ_i, $i=\overline{0,n}$ далі маємо

$$\Delta^2(f)=\|f-\varphi\|^2=\|f\|^2-\|\varphi\|^2=(f,f)-$$

$$-\left(\sum_{i=0}^{n}a_i\varphi_i,\ \sum_{i=0}^{n}a_i\varphi_i\right)=(f,f)-\sum_{i,j=0}^{n}a_ia_j\delta_{ij}=$$

$$=(f,f)-\sum_{i=0}^{n}a_i^2=(f,f)-\sum_{i=0}^{n}|(f,\varphi_i)|^2. \qquad (5)$$

Оскільки $\|f-\varphi\|^2\geqslant 0$, то звідси дістаємо нерівність Бесселя

$$(f,f)\geqslant\sum_{i=0}^{n}|(f,\varphi_i)|^2. \qquad (6)$$

Зауважимо, що коли вихідні елементи не утворюють ортонормовану систему, то їх можна ортогоналізувати за допомогою відомого в лінійній алгебрі процесу Шмідта.

Хоча матриця системи (2) невироджена, вона може мати дуже погані обчислювальні властивості. Розглянемо як приклад випадок простору із скалярним добутком $L_2[0,1]$ і систему функцій $\varphi_i(x)=x^i$, $i=\overline{0,n}$. Тоді

$$a_{ij}=(\varphi_i,\varphi_j)=\int_0^1 x^{i+j}dx=\frac{x^{i+j+1}}{i+j+1}\Big|_0^1=\frac{1}{i+j+1}$$

і матриця системи (2) має вигляд

$$H_{n+1}=\left(\frac{1}{i+j+1}\right)_{i,j=0}^{n} \qquad (7)$$

і називається *матрицею Гільберта*, про її погану обумовленість ми вже зазначали у п. 1.2 6. Наприклад, $\mathrm{cond}_\infty H_6\sim 1{,}5\cdot 10^7$. Це означає, що, вибравши базис $\varphi_i(x)=x^i$, зіткнемося із значними труднощами при розв'язуванні системи (2) на ЕОМ при великих n. Погані «обчислювальні» властивості такої системи функцій пов'язані з її «переповненістю», про яку говорить така *теорема Мюнца.*

Теорема. *Для того щоб система функцій* $1, \{x^{n_j}\}$ $(0<n_1<n_2<\ldots)$ *була повною в* $C[0,1]$, *необхідно і достатньо, щоб розбігався ряд* $\sum_{j=1}^{\infty}n_j^{-1}$.

З теореми Мюнца випливає, що видаливши з системи $\{x^i\}_{i=0}^{\infty}$ нескінченну кількість членів, дістаємо повну в $C[0,1]$ систему (нагадаємо, що система функцій повна в нормованому просторі, якщо її замикання збігається з усім простором).

Наприклад, можна залишити лише функції $1, x^2, \ldots, x^p, \ldots$, в яких показник — просте число або нуль. Оскільки ряд Σp^{-1} розбігається, то ця система буде повною. У той самий час в ній залишилася лише мізерна кількість членів вихідної послідовності, бо на відрізку $[1, n]$ містяться прості числа порядку $n/\ln n$. Грубо кажучи, в системі $\{x^i\}_{i=0}^{\infty}$ є багато «майже лінійно залежних» елементів (до речі, це добре видно на графіках функцій x^i, $i=0, 1, \ldots$), що і призводить до поганих обчислювальних властивостей.

3.4. Приклад найкращого наближення в просторі із скалярним добутком — середньоквадратичне наближення функцій алгебраїчними многочленами (неперервний випадок)

Розглянемо простір неперервних дійсних функцій $\tilde{L}_{2,\rho}[a,b]$ із скалярним добутком

$$(f,g)=\int_a^b\rho(t)f(t)g(t)\,dt, \qquad (1)$$

де $\rho(x)\geqslant 0$ на $[a,b]$, причому $\rho(x)=0$ не більш ніж на множині міри нуль (як зазначалося раніше, простір $\tilde{L}_{2,\rho}[a,b]$ не є гільбертовим, бо він неповний).

Зауважимо, що результати п. 3.3 про найкращі наближення правильні для будь-яких просторів із скалярним добутком. Але для завершеності теорії в тому розумінні, щоб існувала границя $\lim_{n\to\infty}\varphi^{(n)}=f$, де $\varphi^{(n)}$ — найкраще наближення для підпростору M розмірності n, вимога, щоб простір був гільбертовим, є вирішальною (див. 3.8).

Простір $\tilde{L}_{2,\rho}[a,b]$ можна поповнити, якщо за його елементи брати не окремі неперервні функції, а класи функцій, які відрізняються на множині міри нуль, або ж класи фундаментальних (в нормі $\tilde{L}_{2,\rho}[a,b]$) еквівалентних ($\{x_n\}\sim\{x_n'\}$, якщо $\|x_n-x_n'\|\to 0$, $n\to\infty$) послідовностей неперервних функцій (див. п. 1.3.2). Якщо класи, які містять послідовності $\{x_n\}$ і $\{y_n\}$, позначити через $\hat{x}$ і $\hat{y}$ відповідно, то

$$(\hat{x},\hat{y})=\int_a^b\rho(t)\hat{x}(t)\hat{y}(t)\,dt\overset{\text{def}}{=}\lim_{n\to\infty}\int_a^b\rho(t)x_n(t)y_n(t)\,dt. \qquad (2)$$

Такий простір із скалярним добутком (2) є повним, тобто гільбертовим, і позначається $L_{2,\rho}(a,b)$.

Найкращі наближення Φ_0 у просторах $\tilde{L}_{2,\rho}[a, b]$ або $L_{2,\rho}(a, b)$ називаються найкращими середньоквадратичними наближеннями, або наближеннями за методом найменших квадратів. Величина $\Delta(f, \Phi_0)$ в цьому разі називається середньоквадратичним відхиленням Φ_0 від f.

Виберемо за M лінійну оболонку, натягнуту на лінійно незалежну систему функцій $1, x, x^2, \ldots, x^n$. Відповідно з загальною теорією 3.3 коефіцієнти многочлена найкращого середньоквадратичного наближення $p_n^{(2)}(x) \equiv p^{(2)}(x; f) = \sum_{i=0}^{n} a_i x^i$ є розв'язком системи лінійних алгебраїчних рівнянь

$$\sum_{k=0}^{n} (x^m, x^k) a_k = (f, x^m), \quad m = \overline{0, n}. \tag{3}$$

Однак, як відмічалося в попередньому параграфі, при великих n матриця системи (3) має погані обчислювальні властивості (погано обумовлена), тому краще користуватися не системою функцій $1, x, x^2, \ldots, x^n$, а ортогональною в $\tilde{L}_{2,\rho}[a, b]$ системою многочленів. Такі системи многочленів для деяких вагових функцій $\rho(x)$ розглядалися в п. 1.6.

3.5. Приклад найкращого наближення в просторі із скалярним добутком — дискретне середньоквадратичне наближення функцій. Многочлени Чебишева дискретного аргументу

Нехай відомі значення деякої функції $f(x)$ в дискретному наборі точок відрізка $[a, b]: x_0, x_1, \ldots, x_n$. Якщо в лінійному просторі таких функцій дискретного аргументу, який позначимо H_{n+1}, ввести скалярний добуток за формулою

$$(f, g) = \sum_{i=0}^{n} \rho_i f(x_i) g(x_i), \quad \rho_i > 0, \tag{1}$$

то він стане гільбертовим (як відомо з математичного аналізу, такий простір, як простір $(n+1)$-вимірних векторів, є повним). Тому для нього справедливі всі результати п. 3.2, 3.3.

Нехай функції неперервного аргументу $\varphi_0(x), \varphi_1(x), \ldots, \varphi_m(x)$ утворюють систему Чебишева. Очевидно, що вектори $\{\varphi_0(x_i)\}_{i=0}^n, \ldots, \{\varphi_m(x_i)\}_{i=0}^n$, $m \leqslant n$, є лінійно незалежними в H_{n+1} і, таким чином, визначають в H_{n+1} підпростір H_{m+1} розмірності $m+1$, який можна також розглядати як множину векторів вигляду $\{p^h(x_i)\}_{i=0}^n$, де $p^h(x) \equiv p_m^h(x) = \sum_{i=0}^{m} a_i \varphi_i(x)$ — узагальнений многочлен. Нехай вектор $\Phi = (\Phi_0, \ldots, \Phi_n)$ є найкращим наближенням для вектора

$\{f(x_i)\}_{i=0}^n$ в H_{m+1}. За цим вектором Φ можна побудувати єдиний узагальнений многочлен $p^h(x) \equiv p_m^h(x) \equiv p_m^h(x, f)$ такий, що $p^h(x_i) = \Phi_i$, $i = \overline{0, n}$. Дійсно, якщо таких многочленів було б два, а саме $p^h(x)$ і $\tilde{p}^h(x)$, то їхня різниця $p^h - \tilde{p}^h$ перетворювалася б на нуль в $n+1$ точці, і оскільки $m \leqslant n$, то $p^h(x) \equiv \tilde{p}^h(x)$. При $m = n$ многочлен $p^h(x)$ є інтерполяційним многочленом. Дійсно, при $m = n$ підпростір H_{m+1} збігатиметься з усім простором H_{n+1}, і тому елементом найкращого наближення для вектора $\{f(x_i)\}_{i=\overline{0,n}} \in H_{n+1}$ в підпросторі H_{m+1} буде той самий вектор $\{f(x_i)\}_{i=\overline{0,n}} \equiv \Phi$, а величина дискретного середньоквадратичного наближення дорівнює нулю. Тому узагальнений многочлен $p_m^h(x; f)$ такий, що $p_m^h(x_i; f) = \Phi_i$ і буде інтерполяційним для функції $f(x)$. Якщо $m < n$, то говоритимемо, що узагальнений многочлен $p^h(x)$ здійснює найкраще наближення функції $f(x)$ за дискретним методом найменших квадратів. На практиці найчастіше потреба в такому наближенні виникає тоді, коли вигляд функції $f(x)$ невідомий, але її значення в деяких точках можна обчислити (виміряти, спостерігати) і потрібно для неї знайти простий аналітичний вираз. Часто точки x_i, $i = \overline{0, n}$, (точки спостережень) перебувають у розпорядженні дослідника і, крім того, для досягнення мети він може розпоряджатися також вибором базисних функцій $\varphi_i(x)$ та їхнім числом m, прагнучи, зрозуміло, до мінімізації m. Коефіцієнти a_i такого многочлена $p^h(x) = \sum_{i=0}^{m} a_i \varphi_i(x)$ знаходять відповідно до загальної теорії (п. 3.2, 3.3), а саме, для коефіцієнтів a_i маємо систему лінійних алгебраїчних рівнянь

$$\sum_{i=0}^{m} a_i (\varphi_j, \varphi_k) = (f, \varphi_k), \quad k = \overline{0, m}, \tag{2}$$

яка за допомогою позначень

$$s_{jk} = \sum_{i=0}^{n} \rho_i \varphi_j(x_i) \varphi_k(x_i), \quad r_k = \sum_{i=0}^{n} \rho_i f(x_i) \varphi_k(x_i) \tag{3}$$

записується у вигляді

$$\sum_{j=0}^{m} a_j s_{jk} = r_k, \quad k = \overline{0, m}. \tag{4}$$

Оскільки вектори $\{\varphi_j(x_i)\}_{i=0}^n$, $j = \overline{0, m}$, є лінійно незалежними (бо $\varphi_j(x)$ — система Чебишева), то визначник системи (4) є визначником Грама і не дорівнює нулю. Тому (4) має єдиний розв'язок.

Якщо за базисну систему $\{\varphi_i(x)\}$, $i = \overline{0, m}$ вибрати систему функцій $\{x^i\}$, $i = \overline{0, m}$, то викладений вище метод має два недоліки: 1) для

знаходження коефіцієнтів узагальненого многочлена доводиться розв'язувати систему $m+1$ рівнянь із заповненою матрицею, яка має погані обчислювальні властивості (погано обумовлена); 2) якщо для деякого m многочлен побудований, то, збільшивши m, ми не зможемо використати результати виконаних розрахунків при визначенні $p_m^h(x)$ і змушені повторити всі обчислення.

Цих недоліків можна позбутися, якщо знайти систему ортогональних, у розумінні скалярного добутку (1), многочленів. Природно, що ця система залежатиме від розміщення вузлів і від вагової сіткової функції ρ_i. Якщо $x_i = i$, $i = \overline{0, n}$, $\rho_i = 1$, то така система многочленів відома — це *многочлени Чебишева дискретного аргументу*, які визначаються формулою

$$P_{m,n}(x) = \frac{n!}{(n-m)!} \sum_{j=0}^{m} (-1)^j \frac{C_m^j C_{m+j}^j}{n^{(j)}} x^{(j)}, \qquad (5)$$

де $m = 0, 1, \ldots, n$, $x^{(j)} = x(x-1)\ldots(x-(j-1))$ — так званий факторіальний многочлен степеня j, $x^{(0)} = 1$. Співвідношення ортогональності має вигляд

$$(P_{m,n}, P_{j,n}) = \sum_{i=0}^{n} P_{m,n}(i) P_{j,n}(i) = \begin{cases} 0, & j \neq m, \\ \dfrac{(n+m+1)!}{(2m+1)(n-m)!}, & j = m. \end{cases} \qquad (6)$$

Многочлени Чебишева дискретного аргументу на сітці $\omega_n = \{x_i = i : i = \overline{0, n}\}$ задовольняють різницеве рівняння

$$[x(n+1-x)(P_{j,n}(x))_{\bar{x}}]_x + j(j+1) P_{j,n}(x) = 0. \qquad (7)$$

Ці многочлени пов'язані також рекурентним співвідношенням

$$\left(x - \frac{n}{2}\right) P_{m,n}(x) + \frac{m+1}{2(2m+1)} P_{m+1,n}(x) + \frac{m((n+1)^2 - m^2)}{2(2m+1)} P_{m-1,n} = 0, \qquad (8)$$

причому

$$P_{0,n}(x) = 1, \quad P_{1,n}(x) = n - 2x, \quad P_{2,n}(x) = n(n-1) - 6nx + 6x^2,$$
$$P_{3,n}(x) = n(n-1)(n-2) - 2(6n^2 - 3n + 2)x + 30nx^2 - 20x^3. \qquad (9)$$

Зауваження. Іноді використовують ортогональні многочлени Чебишева дискретного аргументу

$$\tilde{P}_{m,n}(x) = \frac{(n-m)!}{n!} P_{m,n}(x),$$

які лише нормуючим множником відрізняються від многочленів (5). Для многочленів $\tilde{P}_{m,n}(x)$ очевидно змінюються і співвідношення (6) — (9).

Якщо вузли рівновіддалені, тобто $x_{i+1} - x_i = h$, то після заміни $x' = \frac{x - x_0}{h}$ точки $x_0, x_1, \ldots, x_n$ перейдуть у точки $0, 1, 2, \ldots, n$, тобто можна користуватися системою многочленів (5).

Якщо $x_i = i$, $i = \overline{0, n}$, $\rho_i = 1$ і скалярний добуток в H_{n+1} визначається формулою (1), то за многочленами $P_{k,n}(x)$, $k > n$, легко побудувати в H_{m+1} многочлен, який дає найкраще наближення. Цей многочлен шукаємо у вигляді

$$P^h(x; f) \equiv P^h(x) \equiv P_m^h(x) = a_0 P_{0,n}(x) + \cdots + a_m P_{m,n}(x).$$

Відповідно до загальної теорії (див. п. 3.3) для знаходження його коефіцієнтів дістаємо систему

$$a_i (P_{i,n}, P_{i,n}) = (f, P_{i,n}), \quad i = \overline{0, n}.$$

З урахуванням (6) маємо

$$a_i = \frac{\sum_{k=0}^{n} f(k) P_{i,n}(k)}{\sum_{k=0}^{n} P_{i,n}^2(k)} = \frac{(2i+1)(n-i)!}{(n+i+1)!} \sum_{k=0}^{n} f(k) P_{i,n}(k).$$

Середньоквадратичне відхилення знаходять за формулою

$$\Delta^2(f, P^h) = \Delta^2(f) = \sum_{k=0}^{n} [f(k) - P_m^h(k)]^2 =$$
$$= \sum_{k=0}^{n} f^2(k) - \sum_{k=0}^{m} \frac{(n+k+1)!}{(2k+1)(n-k)!} a_k^2.$$

3.6. Приклад найкращого наближення в просторі із скалярним добутком — середньоквадратичне наближення тригонометричними многочленами

3.6.1. Неперервний випадок. Функція

$$\Phi_m(x) = a_0 + \sum_{p=1}^{m} (a_p \cos px + b_p \sin px),$$

де a_p, b_p — довільні числові коефіцієнти, причому $|a_m| + |b_m| \neq 0$, називається тригонометричним многочленом порядку m. Щоб побудувати тригонометричний многочлен найкращого середньоквадратичного наближення в просторі із скалярним добутком

$$(f, g) = \int_0^{2\pi} f(x) g(x)\, dx, \qquad (1)$$

можна вибрати за $\varphi_i(x)$, $i = \overline{0, m}$, функції $\frac{1}{\sqrt{2\pi}}$, $\frac{1}{\sqrt{\pi}}\cos x$,

$$\frac{1}{\sqrt{\pi}}\sin x, \ \ldots, \ \frac{1}{\sqrt{\pi}}\cos mx, \ \frac{1}{\sqrt{\pi}}\sin mx, \qquad (2)$$

які утворюють ортонормовану систему (вона є також системою Чебишева) і далі скористатися результатами п. 3.3. Як наслідок, для коефіцієнтів многочлена найкращого наближення дістанемо формули

$$a_0 = \frac{1}{\sqrt{2\pi}}\int_0^{2\pi} f(x)\,dx, \quad a_p = \frac{1}{\sqrt{\pi}}\int_0^{2\pi} f(x)\cos pxdx,$$

$$b_p = \frac{1}{\sqrt{\pi}}\int_0^{2\pi} f(x)\sin pxdx, \quad p = \overline{1, m}. \qquad (3)$$

3.6.2. Дискретний випадок. Нехай n, m — натуральні числа, $m \leqslant n/2$, а також

$$\omega = \{x_i = 2\pi i/(n+1), \quad i = \overline{0, n}\}, \qquad (4)$$

$$\varphi_0(x) = 1, \quad \varphi_p(x) = \sqrt{2}\cos px, \quad \psi_p(x) = \sqrt{2}\sin px, \quad p = \overline{1, m}. \qquad (5)$$

У $(n+1)$-вимірному просторі H_{n+1} функцій дискретного аргументу, заданих на сітці ω, введемо скалярний добуток

$$(f, g) = \frac{1}{n+1}\sum_{i=0}^{n} f(x_i)\,g(x_i) \qquad (6)$$

і поставимо задачу: знайти многочлен найкращого середньоквадратичного наближення (у смислі скалярного добутку (6)) на дискретній множині точок (4) в $(2m+1)$-вимірному просторі H_{2m+1} многочленів вигляду

$$P_m(x) = \alpha_0 + \sum_{p=1}^{m}(\alpha_p\cos px + \beta_p\sin px), \quad x \in \omega. \qquad (7)$$

Вправа 1. Довести, що розмірність підпростору многочленів вигляду (7) є $2m+1$.

Вправа 2. Показати, що система функцій дискретного аргументу (5) ортонормована в смислі скалярного добутку (6), тобто

$$(\varphi_j, \psi_k) = 0, \quad j = \overline{0, m}, \quad k = \overline{1, m};$$
$$(\varphi_r, \varphi_s) = 0, \quad (\psi_r, \psi_s) = 0, \quad r \neq s;$$
$$(\varphi_p, \varphi_p) = (\psi_p, \psi_p) = 1, \quad p = \overline{0, m}; \quad m \leqslant n/2.$$

Оскільки система (5) ортонормована, то для коефіцієнтів многочлена (7) маємо формули

$$\alpha_0 = \frac{1}{n+1}\sum_{i=0}^{n} f\left(\frac{2\pi i}{n+1}\right), \quad \alpha_p = \frac{\sqrt{2}}{n+1}\sum_{i=0}^{n} f\left(\frac{2\pi i}{n+1}\right)\cos p\,\frac{2\pi i}{n+1},$$

$$\beta_p = \frac{\sqrt{2}}{n+1}\sum_{i=0}^{n} f\left(\frac{2\pi i}{n+1}\right)\sin p\,\frac{2\pi i}{n+1}, \quad p = \overline{1, n}. \qquad (8)$$

Середньоквадратичне відхилення обчислюється за формулою

$$\Delta(f, P_m) = \Delta(f) = (f - P_m, f - P_m)^{1/2} =$$
$$= [\|f\|^2 - \alpha_0^2 - \frac{1}{2}\sum_{p=1}^{m}(\alpha_p^2 + \beta_p^2)]^{1/2},$$

де

$$\|f\|^2 = (f, f) = \frac{1}{n+1}\sum_{i=0}^{n} f^2\left(\frac{2\pi i}{n+1}\right).$$

Зауваження. При парному n і при $m = n/2$ маємо $2m + 1 = n + 1$ і система функцій (5) утворює базис в H_{n+1}. Це означає, що

$$\Delta(f) = \Delta(f, P_{\frac{n}{2}}) = \left(\frac{1}{n+1}\sum_{i=0}^{n}(f(x_i) - P_{\frac{n}{2}}(x_i))^2\right)^{1/2} = 0$$

(бо для елемента $\{f(x_i)\}_{i=0}^{n} \in H_{n+1}$ шукаємо елемент найкращого наближення в H_{n+1}). Звідси $P_m(x_i) = f(x_i)$, $i = \overline{0, n}$, тобто тригонометричний многочлен виду (7) від неперервної змінної x є інтерполяційним для функції f.

3.7. Похибка середньоквадратичних наближень

Розглянемо питання про похибку методу найменших квадратів наближення функцій в дискретному варіанті як найбільш важливому для практичних обчислень. Розглянемо окремо вплив випадкових похибок в значеннях функції і похибку методу, що виникає внаслідок заміни функції многочленом, який її наближає.

3.7.1. Вплив випадкових похибок на значення функції, яка наближається. Нехай H_{n+1} — гільбертів простір функцій, заданих на скінченній множині точок $\{x_i\}_{i=0}^{n} \subset [a, b]$, скалярний добуток задається формулою (1) (п. 3.5) $\varphi_0, \varphi_1, \ldots, \varphi_m$, $m \leqslant n$, ортогональна система функцій, тобто $(\varphi_i, \varphi_j) = 0$, якщо $i \neq j$ і $(\varphi_i, \varphi_i) \neq 0$, $i = \overline{0, m}$. Ця система функцій в H_{n+1} утворює $(m+1)$-вимірний простір, в якому для сіткової функції ($(n+1)$-вимірного вектора) $f(x) \in H_{n+1}$ знаходимо найкраще наближення у вигляді

$$P_m(x) = \sum_{i=0}^{m} a_i\varphi_i(x). \qquad (1)$$

Припустимо, що замість значень $f(x_i)$ функції f відомі їхні наближені значення

$$f_i^* = f(x_i) + \eta_i, \quad i = \overline{0, n}, \qquad (2)$$

де η_i — незалежні випадкові величини з нульовим середнім значенням і дисперсією σ, яке не залежить від i, тобто

$$M[\eta_i] = 0 \; \forall \, i = \overline{0, n}, \quad M[\eta_i, \eta_j] = \sigma^2 \delta_{ij},$$

$M[\xi]$ — математичне сподівання випадкової величини ξ, δ_{ij} — символ Кронекера. Тоді замість коефіцієнтів a_i знайдемо коефіцієнти

$$a_i^* = \frac{(f^*, \varphi_i)}{(\varphi_i, \varphi_i)} = \frac{(f, \varphi_i)}{(\varphi_i, \varphi_i)} + \frac{(\eta, \varphi_i)}{(\varphi_i, \varphi_i)} \equiv a_i + \gamma_i, \tag{3}$$

де $\gamma_i = (\eta, \varphi_i)/(\varphi_i, \varphi_i)$ — випадкова похибка i-го коефіцієнта. Звідси

$$M[\gamma_i] = M\left[\frac{(\eta, \varphi_i)}{(\varphi_i, \varphi_i)}\right] = M[(\eta, \varphi_i)]/(\varphi_i, \varphi_i) =$$

$$= M\left[\sum_{j=0}^{n} \rho_j \eta_j \varphi_i(x_j)\right]/(\varphi_i, \varphi_i) = \left(\sum_{j=0}^{n} \rho_j \varphi_i(x_j) M[\eta_j]\right)/(\varphi_i, \varphi_i) = 0;$$

$$M[\gamma_j; \gamma_k] = \frac{1}{(\varphi_j, \varphi_j)(\varphi_k, \varphi_k)} M\left[\sum_{i=0}^{n} \rho_i \eta_i \varphi_j(x_i), \sum_{q=0}^{n} \rho_q \eta_q \varphi_k(x_q)\right] =$$

$$= \frac{1}{(\varphi_j, \varphi_j)(\varphi_k, \varphi_k)} \sum_{i=0}^{n} \sum_{q=0}^{n} \rho_i \rho_q \varphi_j(x_i) \varphi_k(x_q) M[\eta_i, \eta_q] =$$

$$= \frac{1}{(\varphi_j, \varphi_j)(\varphi_k, \varphi_k)} \sum_{i=0}^{n} \rho_i^2 \varphi_j(x_i) \varphi_k(x_i) \sigma^2. \tag{4}$$

Далі для спрощення обмежимося випадком $\rho_i \equiv \text{const} = h$. Тоді

$$M[\gamma_j, \gamma_k] = \frac{(\varphi_j, \varphi_k) h\sigma^2}{(\varphi_j, \varphi_j)(\varphi_k, \varphi_k)} = \frac{h\sigma^2}{(\varphi_j, \varphi_j)} \delta_{kj}, \tag{5}$$

тобто похибки різних коефіцієнтів не корелюють між собою, а дисперсія похибки j-го коефіцієнта обчислюється за формулою

$$D[\gamma_j] = \frac{h\sigma^2}{(\varphi_j, \varphi_j)}, \tag{6}$$

де σ^2 — дисперсія похибок обчислення (спостережень) функції f. Із (3) випливає

$$P_m^*(x) = \sum_{i=0}^{m} a_i^* \varphi_i(x) = P_m(x) + \Gamma_m(x), \tag{7}$$

де $\Gamma_m(x) = \sum_{i=0}^{m} \gamma_i \varphi_i(x)$ — випадкова похибка апроксимуючого многочлена $P_m^*(x)$. У виразі (7) аргумент x може набувати значень із скінченної множини $\{x_i\}_{i=0}^{n}$, а також будь-яких інших значень, при яких визначені $\varphi_i(x)$, $i = \overline{0, m}$, зокрема, якщо за $\varphi_i(x)$ вибрати многочлени Чебишева дискретного аргументу, то $x \in (-\infty, \infty)$.

Оскільки математичне сподівання суми випадкових величин дорівнює сумі їхніх математичних сподівань, а дисперсія суми некорелюючих випадкових величин з деякими коефіцієнтами, а дорівнює сумі їхніх дисперсій, помножених на квадрати коефіцієнтів, то з виразів (4) — (7) дістаємо

$$M[\Gamma_m(x)] = \sum_{i=0}^{m} \varphi_i(x) M[\gamma_i] = 0, \tag{8}$$

$$D[\Gamma_m(x)] = h\sigma^2 \sum_{i=0}^{m} \frac{\varphi_i^2(x)}{(\varphi_i, \varphi_i)}. \tag{9}$$

Це означає, що математичне сподівання випадкової похибки многочлена найкращого дискретного середньоквадратичного наближення дорівнює нулю, а її дисперсія зростає із збільшенням m, тобто степеня многочлена.

Якщо вибрати $x_i = i$, $i = \overline{0, n}$, $\varphi_i(x) = P_{i,n}(x)$, $i = \overline{0, m}$, $m \leqslant n$, де $P_{i,n}(x)$ — многочлени Чебишева дискретного аргументу, то матимемо

$$P_m(x) = \sum_{i=0}^{m} a_i P_{i,n}(x),$$

$$a_i = \frac{(f, P_{i,n})}{(P_{i,n}, P_{i,n})}, \quad (f, g) = \frac{1}{n+1} \sum_{i=0}^{n} f(i) g(i).$$

Якщо використовуються наближені значення функції (2), то з (8), (9) маємо

$$M[\Gamma_m(x)] = 0, \quad D[\Gamma_m(x)] = \frac{\sigma^2}{n+1} \sum_{j=0}^{m} \frac{P_{jn}^2(x)}{(P_{jn}, P_{jn})}. \tag{10}$$

Зокрема, при апроксимації многочленом першого степеня, тобто $m = 1$, маємо

$$D[\Gamma_1(x)] = \frac{\sigma^2}{n+1}\left[1 + \frac{12\left(x - \frac{n}{2}\right)^2}{n(n+1)}\right]. \tag{11}$$

Звідси видно, що крім росту дисперсії з ростом m (див. (10)) має місце зміна її залежно від x: в середині інтервалу спостережень вона значно менша, ніж біля його кінців.

У випадку $[a, b] = [0, 2\pi]$, $x_i = 2\pi i/(n+1)$, $i = \overline{0, n}$, $\rho_i = \frac{1}{n+1}$, і вибору координатних функцій у вигляді (2) (п. 3.6), дістанемо многочлен вигляду (7) (п. 3.6) з коефіцієнтами (8) (п. 3.6). Якщо замість точних значень $f(x_i)$ використовуються значення (2), то при $m \leqslant n/2$

відповідно до (9) матимемо

$$D[\Gamma_m(x)] = \frac{\sigma^2}{n+1}\left[\frac{\varphi_0^2(x)}{(\varphi_0,\varphi_0)} + \sum_{p=1}^{m}\left(\frac{\varphi_p^2(x)}{(\varphi_p,\varphi_p)} + \frac{\psi_p^2(x)}{(\psi_p,\psi_p)}\right)\right] =$$

$$= \frac{\sigma^2}{n+1}\left[1 + 2\sum_{p=1}^{m}(\cos^2 px + \sin^2 px)\right] = \frac{2m+1}{n+1}\sigma^2.$$

Позначивши середньоквадратичне значення похибки $\Gamma_m(x)$ через $\sigma_m(x)$, дістанемо

$$\sigma_m(x) = \sqrt{D[\Gamma_m(x)]} = \sigma\sqrt{\frac{2m+1}{n+1}}, \quad m \leqslant n/2,$$

де σ — середньоквадратичне відхилення похибок спостережень. Таким чином, у періодичному випадку дисперсія похибки $\Gamma_m(x)$ зростає пропорційно m і не залежить від x (на відміну від неперіодичного випадку), а $M[\Gamma_m(x)] = 0$.

3.7.2. Похибка дискретного середньоквадратичного наближення у випадку $f \in \tilde{W}_2^l(\bar{\Omega})$, $l \geqslant 1$. Нехай $p_m^h(x) \equiv p_m^h(x; f) = \sum_{j=0}^{m} a_j P_{j,n}\left(\frac{x}{h}\right)$ — алгебраїчний многочлен найкращого дискретного середньоквадратичного наближення для функції $f(x) \in \tilde{W}_2^l(\bar{\Omega})$ за системою точок $\bar{\omega}_h = \{x_i = ih,\ i = \overline{0, n}\}$, $P_{j,n}(z)$ — многочлени Чебишева дискретного аргументу. Розглянемо залишковий член

$$R_m(x) \equiv R_m(x; f) = f(x) - p_m^h(x), \tag{12}$$

який при фіксованому x, очевидно, є лінійним функціоналом у просторі $\tilde{W}_2^m(\bar{\Omega})$, $\Omega = (0,\ nh)$ (n вважатимемо фіксованим цілим числом). Це випливає з того, що коефіцієнти многочлена

$$a_j = \frac{(f, P_{j,n})}{\|P_{j,n}\|^2} = \frac{\sum_{i=0}^{n} f(x_i) P_{j,n}(i)}{\sum_{i=0}^{n} P_{j,n}^2(i)}, \quad j = \overline{0, m},$$

лінійно виражаються через функцію $f(x)$.

Оцінимо коефіцієнти a_j за допомогою нерівності Коші — Буняковського:

$$|a_j| \leqslant \|f\|_{C(\bar{\Omega})}\sqrt{n+1}\left(\sum_{i=0}^{n} P_{j,n}^2(i)\right)^{-1/2} =$$

$$= \|f\|_{C(\bar{\Omega})}\sqrt{n+1}\left[\frac{(2j+1)(n-j)!}{(n+j+1)!}\right]^{\frac{1}{2}}. \tag{13}$$

Теорема 1. *Нехай* $f(x) \in \tilde{W}_2^l(\bar{\Omega})$, $l \geqslant 1$, $\Omega = (0, nh)$, n *— фіксоване ціле число незалежне від* h, $nh \leqslant 1$, i $p_m^h(x) \equiv p_m^h(x; f)$ *— алгебраїчний многочлен степеня* m *найкращого дискретного середньоквадратичного наближення для функції* $f(x)$ *за системою точок* $\bar{\omega}_h = \{x_i = ih : i = \overline{0, n}\}$. *Тоді для залишкового члена* $R_m(x) \equiv R_m(x; f) = f(x) - p_m^h(x)$ *в точці* $x \in \Omega$ *справедлива оцінка*

$$|R_m(x)| \leqslant M(m, n, x, h)\,\overline{M}(l, m)\,(nh)^{k-1/2}\,|f|_{\tilde{W}_2^k(\bar{\Omega})}, \tag{14}$$

де

$$M(m, n, x, h) = 1 + \sqrt{n+1}\sum_{j=0}^{n}\left\{\left[\frac{(2j+1)(n-j)!}{(n+j+1)!}\right]^{1/2}\left|P_{j,n}\left(\frac{x}{n}\right)\right|\right\} \leqslant M^*,$$

$$M^* = 1 + \sqrt{n+1}\sum_{j=0}^{m}\left[\frac{(2j+1)(n-j)!}{(n+j+1)!}\right]^{1/2} M_j,$$

$$M_j = \max_{x\in[0,n]} |P_{j,n}(x)|,$$

$$\overline{M}(l, m) = \left(1 + \sum_{j=0}^{k-1}\frac{1}{(2k-2j-1)\,[(k-j-1)!]^2}\right)^{1/2},$$

$$k = \begin{cases} m+1, & \text{якщо } l > m, \\ l, & \text{якщо } l \leqslant m, \end{cases}$$

$P_{j,n}$ *— многочлен Чебишева дискретного аргументу.*

Доведення. За допомогою лінійної заміни $\xi = snh$ відобразимо відрізок $\bar{\Omega}$ в одиничний відрізок $\bar{E} = [0, 1]$, $\xi \in \bar{\Omega}$, $s \in \bar{E}$ і позначимо $\tilde{f}(s) = f(\xi(s))$. Очевидно, що $\tilde{f}(s) \in \tilde{W}_2^k(\bar{E})$. Покажемо, що лінійний функціонал (при фіксованому x) $R_m(x; f) \equiv R_m(\tilde{snh}; \tilde{f})$, $x = \tilde{snh}$, обмежений в $\tilde{W}_2^k(\bar{E})$. Використовуючи оцінку (13) і теорему вкладення (див. п. 1.5), дістаємо

$$|R_m(x; f)| = |f(x) - p_m^h(x; f)| \leqslant |f(x)| + |p_m^h(x; f)| \leqslant$$

$$\leqslant \|\tilde{f}\|_{C(\bar{E})} + \sum_{j=0}^{m}|a_j|\left|P_{j,n}\left(\frac{x}{h}\right)\right| \leqslant$$

$$\leqslant \|\tilde{f}\|_{C(\bar{E})}\left(1 + \sqrt{n+1}\sum_{j=0}^{m}\left\{\left[\frac{(2j+1)(n-j)!}{(n+j+1)!}\right]^{1/2}\left|P_{j,n}\left(\frac{x}{h}\right)\right|\right\}\right) \leqslant$$

$$\leqslant 2\|\tilde{f}\|_{\tilde{W}_2^l[0,1]}\left(1 + \sqrt{n+1}\sum_{j=0}^{m}\left\{\left[\frac{(2j+1)(n-j)!}{(n+j+1)!}\right]^{1/2}\left|P_{j,n}\left(\frac{x}{h}\right)\right|\right\}\right) \leqslant$$

$$\leqslant M\|\tilde{f}\|_{\tilde{W}_2^l[0,1]}, \tag{15}$$

де

$$M \equiv M(m, n, x, h) = 1 + \sqrt{n+1} \sum_{j=0}^{m} \left\{ \left[\frac{(2j+1)(n-j)!}{(n+j+1)!} \right]^{1/2} \left| P_{j,n}\left(\frac{x}{h}\right) \right| \right\}. \tag{16}$$

Оскільки $x \in [0, nh]$, то $\frac{x}{h} \in [0, n]$, а тому

$$M \leqslant M^* \equiv 1 + \sqrt{n+1} \sum_{j=0}^{m} \left[\frac{(2j+1)(n-j)!}{(n+j+1)!} \right]^{1/2} M_j, \tag{17}$$

$$M_j = \max_{x \in [0,n]} |P_{j,n}(x)|.$$

Нерівності (15) — (17) означають обмеженість лінійного функціонала $R_m(x; f)$ в просторі $\tilde{W}_2^l[0, 1]$. Очевидно, що $R_m(x; p) \equiv 0$ для довільного многочлена p степеня, не вищого ніж m. Тому в силу леми Брембла — Гільберта маємо

$$|R_m(x; f)| = |R_m(x; \tilde{f})| \leqslant M\overline{M} |\tilde{f}|_{\tilde{W}_2^k(\bar{E})}, \tag{18}$$

де

$$k = \begin{cases} m+1, & \text{якщо } l > m, \\ l, & \text{якщо } l \leqslant m, \end{cases}$$

$$\overline{M} = \overline{M}(l, m) = \left(1 + \sum_{j=0}^{k-1} \frac{1}{(2k-2j-1)[(k-j-1)!]^2} \right)^{1/2}.$$

Переходячи знову до змінної x, дістаємо

$$|\tilde{f}|_{\tilde{W}_2^k(\bar{E})} = \left(\int_0^1 [\tilde{f}^{(k)}(s)]^2 \, ds \right)^{1/2} = \left((n^{-1}h^{-1}(nh)^{2k} \times \right.$$

$$\left. \times \int_0^{nh} [f^{(k)}(x)]^2 \, dx \right)^{1/2} = (nh)^{k-1/2} |f|_{\tilde{W}_2^k(\bar{\Omega})}. \tag{19}$$

Звідси і з виразу (18) випливає твердження теореми 1.

Теорема 1 стверджує, що при фіксованих m, n, x і при $h \to 0$ матимемо

$$|R_m(x)| = |f(x) - p_m^h(x)| = O(h^{k-1/2}). \tag{20}$$

Якщо ж $f(x) \in C^l[0, nh]$, то замість (19), очевидно, матиме місце нерівність

$$|\tilde{f}|_{\tilde{W}_2^k(\bar{E})} = \left((nh)^{2k-1} \int_0^{nh} [f^{(k)}(x)]^2 \, dx \right)^{1/2} \leqslant (nh)^k \|f\|_{C^k(\bar{\Omega})}$$

і тому при фіксованих m, n, x

$$|R_m(x)| = |f(x) - p_m^h(x)| \leqslant M(m, n, x, h)\, \overline{M}(l, m)(nh)^k \|f\|_{C^k(\bar{\Omega})}, \tag{21}$$

тобто при $h \to 0$

$$|R_m(x)| = |f(x) - p_m^h(x)| = O(h^k). \tag{22}$$

З виразів (14), (21) видно, що величина залишкового члена залежить від положення точки x. Щоб з'ясувати характер цієї залежності, припустимо, що $m \ll n$ (саме цей випадок найбільш цікавий для практики). Тоді можна довести, що при $n \to \infty$

$$(n+1)^{-j} P_{j,n}\left(\frac{n+1}{2}(1+s) - \frac{1}{2} \right) = P_j(s) + O(n^{-2}), \tag{23}$$

де $P_j(s)$ — многочлен Лежандра степеня j. В свою чергу можна показати, що для многочленів Лежандра має місце нерівність

$$|P_j(x)| < \frac{2}{\pi j \sqrt[4]{1-x^2}}. \tag{24}$$

Вигляд величини $M(m, n, x, h)$ та співвідношення (23), (24) показують, що похибка методу дискретного середньоквадратичного наближення алгебраїчними многочленами в середині проміжку менша ніж на його краях.

Вправа. Розглянути характер залежності величини $M(m, n, x, h)$ від x при $m = 1$, записавши явний вигляд многочленів $P_{0,n}(x)$ та $P_{1,n}(x)$.

В цілому з розгляду випадкової похибки і похибки методу в дискретному неперіодичному випадку приходимо до наступних висновків: а) точки обчислення функції f (моменти спостережень) доцільно вибирати так, щоб найцікавіший відрізок виявився ближче до середини проміжку спостережень, де менша як середньоквадратична випадкова похибка апроксимуючого многочлена (див. (10), (23), (24)), так і похибка методу; б) при постановці спостережень і виборі степеня m апроксимуючого многочлена слід дотримуватися «розумного» співвідношення між випадковою похибкою, яка зростає із збільшенням m ($m \leqslant n$), але спадає з ростом n, і похибкою методу, яка зростає із збільшенням nh, але спадає із збільшенням m і зменшенням h (при фіксованому n і $l > m$). На практиці, як правило, поступають таким чином. Вибравши $m = 1$, коли середньоквадратичне відхилення $\Delta(f, p_1^h) \gg \varepsilon$, де ε — відома похибка вимірювання $f(x)$, збільшують степінь многочлена доти, доки стане $\Delta(f, p_m^h) \approx \varepsilon$. Якщо при цьому $m \ll n$, то вигляд апроксимуючої функції (алгебраїчний многочлен) вибрано вдало. Якщо ж $m \sim n$, то слід підшукати відповідну систему функцій $\varphi_i(x)$.

3.8. Збіжність найкращих наближень у гільбертовому просторі

Збіжність у нормі гільбертового простору. Нехай $P_m(x) \equiv P_m(x; f)$ — елемент найкращого наближення для заданого $f \in H$ у підпросторі M_m (розмірності $m+1$) гільбертового простору H. Що розуміти під збіжністю $P_m(x)$ до f? Коли простір H нескінченновимірний (див., наприклад, п. 3.4), тоді природно вивчати збіжність $P_m(x)$ до елемента f при $m \to \infty$ у нормі простору H. Така постановка питання у випадку, коли простір $H = H_n$ скінченновимірний (див., наприклад, п. 3.5), не має сенсу, бо тоді $m \leqslant n$ і при $m = n$ маємо $P_m = f$. Однак в умовах п. 3.5 многочлен найкращого дискретного наближення $p_m^h(x) = \sum_{j=0}^{m} a_j \varphi_j(x)$ можна розглядати як функцію неперервного аргументу x, коефіцієнти якої a_j визначають з умови найкращого наближення елемента $\{f(x_i)\}$, $i = \overline{0, n}$, елементом $\{p_m^h(x_i)\}_{i=\overline{0,n}}$ підпростору $M_m \subset H_n$ розмірності $m+1$ ($m < n$). У цьому випадку, який ми назвали дискретним випадком, можна розглядати два види збіжності: а) якщо $f(x)$ задана на відрізку $[a, b]$ і $x_i \in [a, b]$, $i = \overline{0, n}$, то можна вивчати питання про збіжність послідовності функцій неперервного аргументу $p_m^h(x)$ в нормі простору $C[a, b]$ при $n \to \infty$; б) можна розглядати збіжність $p_m^h(x)$ до $f(x)$ у нормі простору $C[a, b]$ при $m \to \infty$, $n \to \infty$, $m \leqslant n$; в) якщо точки x_i — рівновіддалені з кроком h і точка x фіксована, то можна розглядати збіжність $p_m^h(x)$ при $h \to 0$ і фіксованих m, n. Однак зрозуміло, що у випадку а) $p_m^h(x)$ не може збігатися до $f(x)$, бо розмірність простору H_n прямуватиме до нескінченності, а елемент найкращого наближення залишатиметься в підпросторі M_m фіксованої розмірності m. Отже, немає сенсу розглядати цей випадок. Випадок б) найскладніший і розглядається в спеціальній літературі. Щодо випадку в), то з теореми 3.7.1 випливає, що при $h \to 0$, $f(x) \in \tilde{W}_2^l(\overline{\Omega})$ і фіксованих m, n, x справедлива збіжність $p_m^h(x) \to f(x)$.

Перш ніж перейти до випадку нескінченновимірного гільбертового простору H, дамо наступне означення.

Означення 1. Ортогональна система $\{\varphi_k\}_{k=0}^{\infty} \subset H$ називається *повною*, якщо з умови $(\varphi^*, \varphi_k) = 0 \; \forall k = \overline{0, \infty}$ випливає, що $\varphi^* \equiv 0$.

Нехай $\{\varphi_k\}_{k=0}^{\infty}$ — ортонормована система елементів з гільбертового простору H і $P_m = \sum_{j=0}^{m} a_j \varphi_i$ — елемент найкращого наближення для f в H. Тоді справедливе таке твердження.

Теорема 1. В гільбертовому просторі H послідовність найкращих наближень P_m для елемента $f \in H$, побудованих за повною ортонормованою системою, збігається до цього елемента.

Доведення. Покажемо спочатку, що послідовність P_m фундаментальна в H.

Дійсно, в силу ортонормованості φ_i маємо

$$\|P_n - P_m\|^2 = \left\| \sum_{i=m+1}^{n} a_i \varphi_i \right\|^2 = \sum_{i=m+1}^{n} |a_i|^2. \qquad (1)$$

З нерівності Бесселя (3.6) випливає, що числовий ряд $\sum_{i=0}^{\infty} |a_i|^2 = \sum_{i=0}^{\infty} |(f, \varphi_i)|^2$ збігається, тому права частина в (1) прямує до нуля при $m, n \to \infty$ (за критерієм Коші). В силу повноти H існує $\lim_{m\to\infty} P_m = \tilde{f}$. Розглянемо скалярний добуток

$$(f - \tilde{f}, \varphi_k) = \left(f - \sum_{i=0}^{\infty} a_i \varphi_i, \varphi_k\right) = \left(f - \sum_{i=0}^{m} a_i \varphi_i, \varphi_k\right) - \left(\sum_{i=m+1}^{\infty} a_i \varphi_i, \varphi_k\right) = 0 \; \forall k = 0, 1, \ldots, m,$$

який дорівнює нулю за умови ортогональності φ_i і того, що $\sum_{i=0}^{m} a_i \varphi_i$ є елементом найкращого наближення. Оскільки система $\{\varphi_i\}$ — повна, то маємо $f = \tilde{f}$, що і треба було довести.

З умов теореми 1 маємо

$$\|f - P_m\|^2 = \|f\|^2 - \sum_{i=0}^{m} |a_i|^2,$$

звідки після граничного переходу при $m \to \infty$ дістаємо *рівність Парсеваля*

$$\sum_{i=0}^{\infty} |a_i|^2 = \|f\|^2.$$

Зауважимо, що виконання рівності Парсеваля є необхідною і достатньою умовою того, щоб система $\{\varphi_i\}$ була повною. Іншим критерієм повноти ортогональної системи є такий: ортогональна система $\{\varphi_k\}$ з гільбертового простору H повна в тому і лише в тому випадку, коли її лінійна оболонка L щільна в H, тобто замикання L збігається з H. У цьому разі система $\{\varphi_k\}$ називається *замкнутою*.

Як зазначалося в п. 1.3, всякий нормований простір можна поповнити, перетворити на банаховий, а простір із скалярним добутком — на

гільбертовий. Зокрема, так визначається простір $L_{2,\rho}(a, b)$, який є простором класів фундаментальних еквівалентних у нормі, пов'язаній із скалярним добутком (1) (п. 3.4), послідовностей неперервних функцій. Які ж системи функцій є повними в $L_{2,\rho}(a, b)$? Відповідь дає наступна теорема, що узагальнює відому теорему Вейерштрасса.

Теорема 2. *Нехай* $f(x) \in L_{2,\rho}(a, b)$, $|a|, |b| < +\infty$. *Тоді для будь-якого* $\varepsilon > 0$ *існує такий алгебраїчний многочлен* $P_n(x)$, *що*

$$\|f - P_n\|_{L_{2,\rho}(a,b)} \equiv \left\{\int_a^b \rho(x)\,[f(x) - P_n(x)]^2\,dx\right\}^{1/2} < \varepsilon,$$

тобто система ортогональних в $L_{2,\rho}(a, b)$ *многочленів є замкнутою і, отже, повною в* $L_{2,\rho}(a, b)$.

Доведення. В силу побудови простору $L_{2,\rho}(a, b)$ будь-яку функцію (клас) $f(x) \in L_{2,\rho}(a, b)$ можна як завгодно точно наблизити в нормі простору $L_{2,\rho}(a, b)$ послідовністю неперервних функцій $f_n(x)$. За теоремою Вейерштрасса для всіх $f_n(x)$ і $\varepsilon_n > 0$ ($\varepsilon_n \to 0$ при $n \to \infty$) існує многочлен $P_n(x)$ з раціональними коефіцієнтами такий, що $\|f_n - P_n\|_{C[a,b]} \leqslant \varepsilon_n$. Звідси

$$\|f - P_n\|_{L_{2,\rho}(a,b)} \leqslant \|f - f_n\|_{L_{2,\rho}(a,b)} + \|f_n - P_n\|_{L_{2,\rho}(a,b)} \leqslant$$

$$\leqslant \|f - f_n\|_{L_{2,\rho}(a,b)} + K\|f_n - P_n\|_{C[a,b]} \to 0, \quad n \to \infty,$$

де $K = \left[\int_a^b \rho(x)\,dx\right]^{1/2}$, тобто множина многочленів з раціональними коефіцієнтами щільна в $L_{2,\rho}(a, b)$, що і треба було довести.

Якщо інтервал (a, b) — необмежений, то має місце теорема, яку наводимо без доведення.

Теорема 3. *Система* $\{e^{-x/2}x^{\alpha/2}L_n^{\alpha}(x)\}_{n=0}^{\infty}$, *де* $L_n^{\alpha}(x)$ — *многочлен Лагерра, повна в* $L_{2,1}(0, \infty)$, *а система* $\{\exp(-x^2/2)\,H_n(x)\}_{n=0}^{\infty}$, *де* $H_n(x)$ — *многочлени Ерміта, повна в* $L_{2,1}(-\infty, \infty)$.

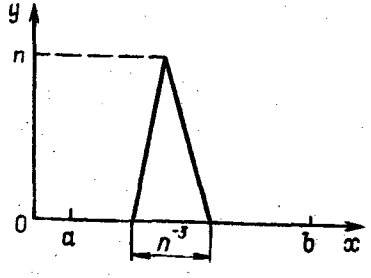


Рис. 18

З теорем 1—3 випливає така.

Теорема 4. *Для будь-якої функції* $f(x) \in L_{2,\rho}(a, b)$, *де* $\rho(x)$ *і* (a, b) *відповідають класичним ортогональним многочленам, елемент* $\varphi_m(x)$ *найкращого наближення в* $L_{2,\rho}(a, b)$ *побудований за відповідною системою ортонормованих ортогональних многочленів* $\{p_i(x)\}_{i=0}^{m}$, *збігається в* $L_{2,\rho}(a, b)$ *до* $f(x)$ *або, іншими словами, ряд Фур'є, побудований за відповідною системою нормованих ортогональних многочленів, збігається до* $f(x)$ *в* $L_{2,\rho}(a, b)$.

Зауваження 1. Близькість двох неперервних функцій в нормі простору $L_{2,\rho}(a, b)$ не гарантує малості їхнього максимального відхилення однієї від другої, тобто не гарантує близькості їх в нормі $C[a, b]$ (хоча саме близькість в нормі $C[a, b]$ часто важлива на практиці). Прикладом цього є задача, яку пропонуємо розв'язати.

Вправа. Показати, що для функції $g(x) = 0$, $x \in [a, b]$ і функції $f(x)$, зображеної на рис. 18, виконуються нерівності

$$\|f - g\|_{L_{2,1}(a,b)} = \sqrt{\frac{1}{b-a}\int_a^b (f(x) - g(x))^2\,dx} \leqslant \frac{1}{\sqrt{(b-a)\,n}},$$

$$\|f - g\|_{C[a,b]} = n,$$

тобто вибором n можна величину $\|f - g\|_{L_{2,1}(a,b)}$ зробити як завгодно малою, а $\|f - g\|_{C[a,b]}$ — як завгодно великою.

3.9. Застосування ідей методу найменших квадратів у суміжних питаннях

3.9.1. Згладжування результатів спостережень. Нехай внаслідок спостережень для значень аргументу $x_0, x_1, \ldots, x_k$ побудовано таблицю значень функції $f(x)$: $f(x_0), f(x_1), \ldots, f(x_k)$. На практиці, як правило, $x_0, x_1, \ldots, x_k$ визначаються точно, або принаймні, значно точніше ніж значення $f(x_0), f(x_1), \ldots, f(x_k)$, які мають похибки спостережень. Як випливає з результатів п. 3.7.1, коли значення, що спостерігаються, мають вигляд $f_i^* = f(x_i) + \eta_i$, де η_i — незалежні випадкові величини з нульовим середнім значенням і дисперсією, яка не залежить від i, математичне сподівання похибки многочлена найкращого середньоквадратичного наближення по дискретній системі точок дорівнює нулю. Цей результат і є основою так званого згладжування наслідків спостережень, тобто визначення деяких нових $\overline{f}(x_i)$.

Нехай $x_i - x_{i-1} = h$ і функція $f(x)$ на кожному відрізку, що охоплює N вузлів, може бути добре наближена многочленом m-го степеня, $m \leqslant N - 1$. Щоб знайти $\overline{f}(x_i)$, вибирають N парним і множину вузлів $x_{i-\frac{N}{2}+j}$, $j = \overline{0, N}$, для якої точка x_i є середньою. За відомими внаслідок спостережень значеннями $f(x_{i-\frac{N}{2}+j})$, $j = \overline{0, N}$, наприклад за допомогою многочленів Чебишева дискретного аргументу, будують многочлен найкращого дискретного середньоквадратичного

наближення $p_m(x)$ в просторі із скалярним добутком

$$(u, v) = \frac{1}{N+1} \sum_{i=0}^{N} u(x_i)\, v(x_i)$$

(див. 3.5) і покладають $\bar{f}(x_i) = p_m(x_i)$. Наведемо деякі з формул згладжування (тут для зручності позначено $f(x_i) = f_i$):

$$m = 1, \quad N = 4, \quad \bar{f}(x_i) = \frac{1}{5}(f_{i-2} + f_{i-1} + f_i + f_{i+1} + f_{i+2});$$

$$m = 3, \quad N = 4, \quad \bar{f}(x_i) = \frac{1}{35}(-3f_{i-2} + 12f_{i-1} + 17f_i + 12f_{i+1} - 3f_{i+2}),$$

$$N = 6, \quad \bar{f}(x_i) = \frac{1}{21}(-2f_{i-3} + 3f_{i-2} + 6f_{i-1} + 7f_i + 6f_{i+1} + 3f_{i+2} - 2f_{i+3});$$

$$m = 5, \quad N = 6, \quad \bar{f}(x_i) = \frac{1}{231}(5f_{i-3} - 30f_{i-2} + 75f_{i-1} + 131f_i + 75f_{i+1} - 30f_{i+2} + 5f_{i+3}),$$

$$N = 8, \quad \bar{f}(x_i) = \frac{1}{429}(15f_{i-4} - 55f_{i-3} + 30f_{i-2} + 135f_{i-1} + 179f_i + 135f_{i+1} + 30f_{i+2} - 55f_{i+3} + 15f_{i+4}).$$

Іноді згладжування проводиться декілька разів, але слід мати на увазі, що багаторазове згладжування може сильно змінити (замаскувати) реальну поведінку функції $f(x)$.

3.9.2. Розв'язування несумісних систем лінійних алгебраїчних рівнянь методом найменших квадратів. Нехай треба розв'язати систему лінійних алгебраїчних рівнянь

$$B\alpha = l, \tag{1}$$

де $B = \{b_{ij}\}_{i=\overline{1,n}}^{j=\overline{1,m}}$ — прямокутна матриця, $\alpha = (\alpha_1, \ldots, \alpha_m)^T$ — невідомий m-вимірний вектор, $l = (l_1, \ldots, l_n)^T$ — заданий n-вимірний вектор, причому $m < n$. Ця система, в якій число рівнянь більше, ніж число невідомих, взагалі кажучи, несумісна, тому неможливо знайти числа $\alpha_1, \ldots, \alpha_m$, які перетворили б (1) на тотожність. На практиці в цьому і немає потреби, бо матриця B та вектор l визначаються наближено в результаті спостережень або обчислень.

Метод найменших квадратів розв'язування таких систем полягає в тому, що визначаються невідомі, які мінімізують суму квадратів нев'язок, тобто суму вигляду

$$S = \sum_{k=1}^{n} \left[l_k - \sum_{i=1}^{m} b_{ki}\alpha_i \right]^2. \tag{2}$$

З умови мінімуму величини S як функції від $\alpha_1, \ldots, \alpha_m$ дістаємо систему m рівнянь з m невідомими виду

$$\frac{\partial S}{\partial \alpha_i} = 0, \quad i = \overline{1, m}. \tag{3}$$

Рівняння системи (1) називаються умовними, а система (1) *системою умовних рівнянь*. Рівняння (3) називаються нормальними, а вся система (3) — *системою нормальних рівнянь*.

Запишемо нормальні рівняння в явному вигляді. Маємо

$$\frac{\partial S}{\partial \alpha_j} = -2 \sum_{k=1}^{n} \left[l_k - \sum_{i=1}^{m} b_{ki}\alpha_i \right] b_{kj} = 0, \quad j = \overline{1, m},$$

або

$$\sum_{i=1}^{m} \left[\sum_{k=1}^{n} b_{ki} b_{kj} \right] \alpha_i = \sum_{k=1}^{n} b_{kj} l_k, \quad j = \overline{1, m}. \tag{4}$$

Розв'язок системи m лінійних алгебраїчних рівнянь (4) з m невідомими вважаємо наближеним розв'язком системи (1).

Приклад 1. Розв'язати систему лінійних алгебраїчних рівнянь

$$B\alpha = l, \tag{5}$$

$$B = \begin{pmatrix} 1 & 3 \\ 2 & 4 \\ 5 & 6 \end{pmatrix}, \quad l = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

Відповідь.

$$\alpha = \begin{bmatrix} 17/149 \\ 13/149 \end{bmatrix}.$$

3.9.3. Побудова емпіричних формул і розв'язування систем нелінійних алгебраїчних рівнянь. Нехай в результаті вимірювань функції $y(x)$ при $x = x_1, x = x_2, \ldots, x = x_n$, $x_i \in [a, b]$, $i = \overline{1, n}$ дістаємо таблицю значень y_i, $i = \overline{1, n}$. За даними цієї таблиці треба побудувати аналітичну формулу

$$\bar{y}(x) = f(x, a_1, \ldots, a_m), \tag{6}$$

яка залежить від m ($m < n$) параметрів a_i, $i = \overline{1, m}$, причому функція $\bar{y}(x)$ має «досить добре» наближувати функцію $y(x)$ на всьому проміжку $[a, b]$. Вигляд функції f і число параметрів у деяких випадках відомі на основі додаткових міркувань. В інших випадках вони визначаються за графіком, побудованим за відомими значеннями $y(x_i)$ так, щоб залежність (6) була досить простою і добре відображала результати спостережень.

Якщо система рівнянь

$$y_1 = f(x_1, a_1, \ldots, a_m),$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$y_n = f(x_n, a_1, \ldots, a_m) \qquad (7)$$

має єдиний розв'язок, то він може бути знайдений з яких-небудь m рівнянь системи (7). Однак у загальному випадку значення y_i, x_i, $i = \overline{1, n}$, є наближеними і точний вигляд залежності $y(x)$ невідомий, і через це система (7) переважно є несумісною. Тому визначимо параметри $a_1, \ldots, a_m$ так, щоб у деякому розумінні всі рівняння системи (7) задовольнялись з найменшою похибкою, точніше, щоб мінімізувати функцію

$$S(a_1, \ldots, a_m) = \sum_{i=1}^{n} [y_i - f(x_i, a_1, \ldots, a_m)]^2. \qquad (8)$$

Такий метод розв'язування системи (7) називають *методом найменших квадратів*.

Якщо функція $S(a_1, \ldots, a_m)$ досягає абсолютного мінімуму в області зміни параметрів $a_1, \ldots, a_m$, то, розв'язуючи систему

$$\frac{\partial S}{\partial a_k} = -2 \sum_{i=1}^{n} [y_i - f(x_i, a_1, \ldots, a_m)] \frac{\partial f(x_i, a_1, \ldots, a_m)}{\partial a_k} = 0,$$
$$k = \overline{1, m}, \qquad (9)$$

знаходимо точки, в яких може бути екстремум. Вибравши той розв'язок, який належить області зміни параметрів $a_1, \ldots, a_m$ і в якому функція $S(a_1, \ldots, a_m)$ має абсолютний мінімум, знаходимо шукані значення $a_1, \ldots, a_m$.

Якщо $f(x, a_1, \ldots, a_m)$ лінійно залежить від параметрів $a_1, \ldots, a_m$, тобто

$$f(x, a_1, \ldots, a_m) = \sum_{i=1}^{m} f_i(x) a_i,$$

то система (7) набирає вигляду

$$y_j = \sum_{i=1}^{m} f_i(x_j) a_i, \quad j = \overline{1, n}, \qquad (10)$$

і, позначаючи $\alpha = (a_1, \ldots, a_m)^T$, $l = (y_1, \ldots, y_n)^T$, $b_{ij} = f_j(x_i)$, дістаємо систему умовних рівнянь виду (1). Зауважимо, що систему нормальних рівнянь можна дістати, помноживши систему умовних рівнянь (1) зліва на транспоновану до B матрицю B^*, тобто система нормальних рівнянь має вигляд

$$B^* B\vec{\alpha} = B^* \vec{l}. \qquad (11)$$

У випадку нелінійної залежності $f(x, a_1, \ldots, a_m)$ від параметрів для розв'язування системи (9) іноді можна скористатися таким прийомом.

Нехай відомі наближені значення параметрів $a_1^0, \ldots, a_m$, які відрізняються від шуканих $a_1, \ldots, a_m$ малими поправками $\alpha_1, \ldots, \alpha_m$, і функція $f(x, a_1, \ldots, a_m)$ диференційовна по $a_1, a_2, \ldots, a_m$. Тоді наближено можна покласти

$$f(x_i, a_1, \ldots, a_m) \approx f(x_i, a_1^0, \ldots, a_m^0) + \sum_{k=1}^{m} f'_{a_k}(x_i, a_1^0, \ldots, a_m^0)\alpha_k.$$

Ввівши позначення

$$y_i^* = y_i - f(x_i, a_1^0, \ldots, a_m^0), \quad i = \overline{1, n}, \quad l = (y_i^*)_{i=\overline{1,n}},$$
$$B = \{b_{ij}\} = \{f'_{a_j}(x_i, a_1^0, \ldots, \alpha_m^0)\}_{i=\overline{1,n}}^{j=\overline{1,m}}$$
$$\alpha = (\alpha_1, \ldots, \alpha_m),$$

із (7) дістаємо умовну систему виду (1). Позначивши її розв'язок через $\alpha_1^0, \ldots, \alpha_m^0$, знаходимо такі наближення для параметрів: $a_1^1 = a_1^0 + \alpha_1^0, \ldots, a_m^1 = a_m^0 + \alpha_m^0$. Приймаючи їх за нове початкове наближення, можна продовжити процес уточнення доти, доки, наприклад, для заданого $\varepsilon > 0$ не буде виконано нерівність $\| \alpha^k \| < \varepsilon$, де через $\| \alpha^k \|$ позначено деяку норму вектора $\alpha^k = (\alpha_1^k, \ldots, \alpha_m^k)^T$.

Цей метод легко поширюється на випадок визначення параметрів $a_1, \ldots, a_m$ із системи r емпіричних формул з n незалежними змінними, тобто

$$y_i = f_i(x_1, \ldots, x_n, a_1, \ldots, a_m), \quad i = \overline{1, r},$$

якщо для точок $(x_1^{(j)}, \ldots, x_n^{(j)})$, $j = \overline{1, N}$, відомі наближені значення $y_i^{(j)} = f_i(x_1^{(j)}, \ldots, x_n^{(j)}, a_1, \ldots, a_m)$, причому $N > m$.

Зазначимо, що на задачу наближення функцій, заданих таблицею значень, за допомогою узагальненого многочлена степеня m можна дивитися як на задачу побудови емпіричної формули, що має вигляд цього узагальненого многочлена, причому роль параметрів у цьому випадку грають коефіцієнти многочлена.

3.10. Про елементи найкращого наближення в нормованих просторах $L_p(a, b)$ та $C[a, b]$

Побудова елемента найкращого наближення в банаховому або нормованому просторі пов'язана із значними труднощами. Це викликано тим, що для банахових просторів невідомі теореми характеризації елемента найкращого наближення, які б давали конструктивну основу алгоритму для визначення елемента, як у просторах із скалярним добутком.

Розглянемо важливі для практики нормовані простори $L_p(a, b)$ з нормою

$$\|f\|_{L_p(a,b)} = \left\{\int_a^b |f(x)|^p dx\right\}^{1/p}, \quad p \geqslant 1, \quad p \neq 2, \tag{1}$$

і простір $C[a, b]$ з нормою

$$\|f\|_{C[a,b]} = \max_{x\in[a,b]} |f(x)|. \tag{2}$$

Випадок $p = 2$ дає нам гільбертовий простір $L_2(a, b)$, для якого мають місце результати 3.3, 3.4.

3.10.1. Деякі властивості просторів $L_p(a, b)$ і елемента найкращого наближення в ньому. Насамперед доведемо таке твердження про одну властивість норми в $L_p(a, b)$ при $p \to \infty$.

Теорема 1. *Нехай функція $f(x)$ вимірна (за Лебегом) і обмежена на відрізку $[a, b]$. Тоді*

$$\lim_{p\to\infty} \|f\|_{L_p(a,b)} = \sup_{x\in[a,b]} \operatorname{vrai} |f(x)| = M, \tag{3}$$

права частина якого є так званий суттєвий максимум $|f(x)|$ на відрізку $[a, b]$, тобто найменше з чисел M таких, що множина всіх $x \in [a, b]$, для яких $|f(x)| > M$, має міру нуль.

Доведення. З нерівності

$$\|f\|_{L_p(a,b)} \leqslant M(b-a)^{1/p}$$

маємо

$$\overline{\lim_{p\to\infty}} \|f\|_{L_p(a,b)} \leqslant M. \tag{4}$$

З іншого боку, якщо E означає множину відрізка $[a, b]$, на якому

$$|f(x)| > M - \varepsilon$$

(тут $\varepsilon > 0$ — як завгодно мале число), а $\operatorname{mes} E$ — міру E, то

$$\|f\|_{L_p(a,b)} \geqslant \left(\int_E |M-\varepsilon|^p dx\right)^{1/p} = (M-\varepsilon)(\operatorname{mes} E)^{1/p},$$

звідки

$$\underline{\lim_{p\to\infty}} \|f\|_{L_p(a,b)} \geqslant M - \varepsilon,$$

і в силу довільності ε

$$\underline{\lim_{p\to\infty}} \|f\|_{L_p(a,b)} \geqslant M. \tag{5}$$

Нерівності (4) і (5) доводять теорему.

Зауважимо, що коли $f \in C[a, b]$, то з виразу (3)

$$\lim_{p\to\infty} \|f\|_{L_p(a,b)} = \|f\|_{C[a,b]}. \tag{6}$$

Ця рівність дає підстави розглядати простір $C[a, b]$ як граничний випадок простору $L_p(a, b)$ при $p \to \infty$.

Теорема 2. *При будь-якому $p > 1$ простір $L_p(a, b)$ є строго нормованим.*

Доведення. Для доведення досить з'ясувати, за яких умов досягається знак рівності у відомій нерівності Гьольдера

$$\left|\int_a^b f(x)g(x)\,dx\right| \leqslant \left[\int_a^b |f(x)|^p\,dx\right]^{1/p}\left[\int_a^b |g(x)|^q dx\right]^{1/q}, \tag{7}$$

$$\frac{1}{p} + \frac{1}{q} = 1, \quad p > 1, \quad q > 1,$$

і, як наслідок, в нерівності Мінковського

$$\left\{\int_a^b [f(x)+g(x)]^p\,dx\right\}^{1/p} \leqslant \left[\int_a^b |f(x)|^p\,dx\right]^{1/p} + \left[\int_a^b |g(x)|^p\,dx\right]^{1/p}. \tag{8}$$

Наслідком теорем 2 і 3.2.1, 3.2.2 є таке твердження.

Теорема 3. *Елемент найкращого наближення в просторі $L_p(a, b)$ існує і єдиний.*

Доведення. Нехай $f(x) \in L_p(a, b)$ і елемент найкращого наближення визначають у підпросторі M_{n+1}, натягнутому на систему лінійно незалежних функцій $\{\varphi_i(x)\}_{i=0}^n$, $\varphi_i(x) \in L_p(a, b)$. Тоді задача знаходження цього елемента зводиться до визначення глобального мінімуму функціонала

$$J(a^{(p)}) \equiv J(a_0^{(p)}, \ldots, a_n^{(p)}) = \left\{\int_a^b \left|f(x) - \sum_{i=0}^n a_i^{(p)}\varphi_i(x)\right|^p dx\right\}^{1/p}, \tag{9}$$

який в силу теореми 3 існує і єдиний. Неважко помітити, що функціонал (9) строго опуклий в M_{n+1}. Дійсно, в силу нерівності Мінковського

$$J(\alpha a^{(p)} + (1-\alpha)b^{(p)}) = \left\|\alpha f + (1-\alpha)f - \alpha\sum_{i=0}^n a_i^{(p)}\varphi_i(x) - \right.$$

$$\left. -(1-\alpha)\sum_{i=0}^n b_i^{(p)}\varphi_i(x)\right\|_{L_p(a,b)} \leqslant$$

$$\leqslant \alpha\left\|f - \sum_{i=0}^n a_i^{(p)}\varphi_i(x)\right\|_{L_p(a,b)} + (1-\alpha)\left\|f - \sum_{i=0}^n b_i^{(p)}\varphi_i(x)\right\|_{L_p(a,b)} = \tag{10}$$

$$= \alpha J(a^{(p)}) + (1-\alpha)J(b^{(p)}), \quad 0 \leqslant \alpha \leqslant 1,$$

$$\forall\, a^{(p)},\ b^{(p)} \in M_{n+1},$$

причому знак рівності можливий тоді і лише тоді, коли для деякого $\lambda > 0$

$$\alpha\left[f - \sum_{i=0}^{n} a_i^{(p)}\varphi_i(x)\right] = \lambda(1-\alpha)\left[f - \sum_{i=0}^{n} b_i^{(p)}\varphi_i(x)\right] \quad \forall\, a^{(p)},\ b^{(p)}.$$

Остання рівність, очевидно, неможлива для жодних $\alpha \neq 0$, $\alpha - 1 \neq 0$; тобто рівність в (10) можлива лише при $\alpha = 0$ або $\alpha = 1$, що і означає строгу опуклість функціонала J. Тому в силу відомої теореми про екстремум опуклої функції на опуклих множинах функціонал J має лише одну точку екстремуму, яка є точкою глобального мінімуму, і для її знаходження можна застосувати один з методів мінімізації.

Теорема 4. *Нехай v — опукла множина, а функція $J(u)$ визначена і опукла на v. Тоді всяка точка локального мінімуму $J(u)$ одночасно являється точкою її глобального мінімуму на v, причому множина $v_* = \{u : u \in v,\ J(u) = J_* = \inf_v J(v)\}$ опукла. Якщо $J(u)$ строго опукла на v, то v_* містить не більше однієї точки.*

Зауважимо, що в силу того, що функція $y(x) = x^{1/p}$, $x \geqslant 0$, $p > 1$, є монотонною, то замість (9) можна мінімізувати функціонал

$$J_1(a^{(p)}) = J_1(a_0^{(p)}, \ldots, a_n^{(p)}) = \int_a^b \left| f(x) - \sum_{i=0}^{n} a_i^{(p)}\varphi_i(x) \right|^p dx.$$

Цю задачу при парному p, в свою чергу, можна замінити задачею розв'язування системи лінійних алгебраїчних рівнянь

$$\frac{\partial J_1}{\partial a_j^{(p)}} = -p\int_a^b \left[f(x) - \sum_{i=0}^{n} a_i^{(p)}\varphi_i(x) \right]^{p-1} \varphi_j(x)\,dx = 0$$

(необхідна умова мінімуму функцій багатьох змінних).

3.10.2. Деякі властивості простору $C[a, b]$ і елемента найкращого наближення у ньому. У випадку простору $C[a, b]$ теорема 1 (п. 3.2) гарантує існування елемента найкращого наближення для $f \in C[a, b]$ в підпросторі M розмірності $n + 1$. Однак на відміну від простору $L_p(a, b)$, $p > 1$, тут не можна скористатися теоремою 2 (п. 3.2) про єдиність елемента найкращого наближення, бо нормований простір $C[a, b]$ не є строго нормованим (див. п. 3.2). Розглянемо умови, за яких елемент найкращого наближення в $C(\bar{s})$, де $\bar{s}$ — деякий компакт, єдиний. Їх дає наступна теорема Хаара, яку ми приведемо без доведення.

Теорема 5. *Для того щоб для будь-якої функції $f(x) \in C[a, b]$ існував єдиний многочлен найкращого рівномірного наближення вигляду* (1) (п. 3.2), *необхідно і достатньо, щоб система $\{\varphi_i(x)\}$ була системою Чебишева на $\bar{s}$.*

Теореми характеризації елемента найкращого наближення в просторі $C(\bar{s})$ виходять за рамки цієї книги, а тому наведемо одну з них без доведення і проілюструємо деякими прикладами. Оскільки для випадку алгебраїчних многочленів найкращого рівномірного наближення вона вперше була доведена П. Л. Чебишевим, то будемо її називати *узагальненою теоремою Чебишева.*

Теорема 6. *Нехай $\{\varphi_i(x)\}_{i=0}^{n} \subset C(\bar{s})$ — система Чебишева на $\bar{s}$ і M — її лінійна оболонка. Тоді для того щоб елемент $\Phi_0(x) \in M$ був найкращим наближенням для $f(x) \in C(\bar{s})$ $(f \notin M)$ в M необхідно і достатньо, щоб існували $x_i \in \bar{s}$, $p_i > 0$, $i = \overline{1, n+2}$, які задовольняють умови*

$$\delta(x_i)\operatorname{sign}\delta(x_i) = \Delta(f), \quad i = \overline{1, n+2},$$

$$\sum_{i=1}^{n+2} p_i \operatorname{sign}\delta(x_i)\,\Phi(x_i) = 0 \quad \forall\, \Phi \in M,$$

де $\delta(x) = \Phi(x) - f(x)$.

У випадку $\bar{s} = [a, b]$ ця теорема спрощується і формулюється таким чином.

Теорема 7. *Нехай виконано умови теореми 6 для $s = [a, b]$. Елемент $\Phi_0(x) \in M \subset C[a, b]$ буде елементом найкращого наближення до $f(x) \in C[a, b]$ тоді і лише тоді, коли існують точки x_i, які задовольняють умови*

$$a \leqslant x_1 < x_2 < \cdots < x_{n+2} \leqslant b, \tag{11}$$

$$|\delta(x_i)| = \Delta(f), \quad i = \overline{1, n+2}, \tag{12}$$

$$\delta(x_i) = -\delta(x_{i+1}), \quad i = \overline{1, n+1}. \tag{13}$$

Точки x_i з (11) — (13) називаються *чебишевським альтернансом.*

Наведемо деякі приклади, в яких за допомогою теорем 5 і 6 можна побудувати алгебраїчний многочлен найкращого наближення в просторі $C[a, b]$.

Приклад 1. Нехай функція $f(x)$, $x \in [a, b]$, неперервна і треба знайти многочлен найкращого рівномірного наближення нульового степеня. Неважко помітити, що многочлен

$$\varphi_0(x) = (M + m)/2,$$

де $M = \max_{[a,b]} f(x) = f(x_1)$, $m = \min_{[a,b]} f(x) = f(x_2)$ є шуканим, а точки x_1, x_2 є точками чебишевського альтернансу. Це випливає з теореми 7.

Приклад 2. Знайдемо многочлен найкращого рівномірного наближення першого степеня для опуклої на $[a, b]$ функції $f(x) \in C[a, b]$. Шукатимемо його у вигляді $\varphi_1(x) = a_0 + a_1x$. В силу опуклості $f(x)$ різниця $f(x) - \varphi_1(x)$ може мати тільки внутрішню точку екстремуму, тому точки a, b є точками чебишевського альтернансу. Нехай d — третя точка. Тоді з теореми 7

$$f(a) - (a_0 + a_1a) = \alpha\Delta(f),$$
$$f(d) - (a_0 + a_1d) = -\alpha\Delta(f),$$
$$f(b) - (a_0 + a_1b) = \alpha\Delta(f),$$

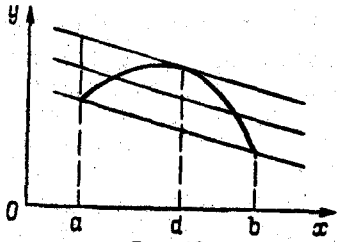


Рис. 19.

де $\alpha = +1$ або $\alpha = -1$. Віднімаючи перше рівняння від третього, маємо

$$f(b) - f(a) = a_1 (b - a),$$

звідки $a_1 = \dfrac{f(b) - f(a)}{b - a}$. Оскільки d — точка екстремуму $f(x) - \varphi_1(x)$, то вона визначається з рівняння $f'(d) - a_1 = 0$. Складаючи перше і друге рівняння, знайдемо

$$2a_0 = f(a) + f(d) - a_1 (a + d).$$

Геометрично ця процедура виглядає таким чином (рис. 19):

а) проводимо січну через точки $(a, f(a))$, $(b, f(b))$; тангенс кута її нахилу до осі OX дорівнює a_1; б) проводимо паралельно їй дотичну до кривої $y = f(x)$; в) проводимо посередині цих двох прямих нову, котра і буде шуканим елементом найкращого наближення.

Приклад 3. Нехай функція $f(x)$ задовольняє умову $f^{(n+1)}(x) \geqslant 0$ і треба оцінити величину відхилення $\Delta(f) = \inf\limits_{p \in \pi_n} \| f - p \| \equiv \Delta_n(f)$ многочлена її найкращого рівномірного наближення, де π_n — множина алгебраїчних многочленів не вище n-го степеня.

У п. 3.4.2 була визначена така *оцінка похибки інтерполяції за вузлами* $x_k = \dfrac{1}{2}\left[(a + b) + (b - a)\cos\left(\dfrac{\pi(2k+1)}{2(n+1)}\right)\right]$, $k = \overline{0, n}$, які є нулями модифікованого многочлена Чебишева $T_n(x)$, що найменше відхиляються від нуля на $[a, b]$:

$$|f(x) - p_n(x)| \leqslant \max_{x \in [a,b]} |f^{(n+1)}(x)| \frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}. \qquad (14)$$

Звідси випливає

$$\Delta_n(f) \leqslant \max_{x \in [a,b]} |f^{(n+1)}(x)| \frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}. \qquad (15)$$

Нехай $\varphi_n(x)$ — многочлен найкращого рівномірного наближення. В силу теореми 7 різниця $f(x) - \varphi_n(x)$ перетворюється на нуль в $(n+1)$-й точці $y_1, \ldots, y_{n+1}$. Тому многочлен можна розглядати як інтерполяційний з вузлами інтерполяції $y_1, \ldots, y_{n+1}$, причому

$$f(x) - \varphi_n(x) = f^{(n+1)}(\xi) \frac{\omega_{n+1}(x)}{(n+1)!},$$

$$\omega_{n+1}(x) = (x - y_1) \ldots (x - y_{n+1}), \quad \xi \in [a, b].$$

Нехай $\max\limits_{[a,b]} |\omega_{n+1}(x)| = |\omega_{n+1}(x_0)|$, тоді

$$\Delta_n(f) = \|f - \varphi_n\|_{C[a,b]} \geqslant |f(x_0) - \varphi_n(x_0)| =$$

$$= |f^{(n+1)}(\xi(x_0))| \frac{|\omega_{n+1}(x_0)|}{(n+1)!} \geqslant \min_{x \in [a,b]} |f^{(n+1)}(x)| \max_{x \in [a,b]} \frac{|\omega_{n+1}(x)|}{(n+1)!}.$$

В силу результатів 2.4.2

$$\max_{[a,b]} |\omega_{n+1}(x)| \geqslant \frac{(b-a)^{n+1}}{2^{2n+1}},$$

тому

$$\Delta_n(f) \geqslant (\min_{[a,b]} |f^{(n+1)}(x)|) \frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}. \qquad (16)$$

Оцінки (15) і (16) показують, що коли $f^{(n+1)}(x)$ зберігає знак і мало змінюється на $[a, b]$, то різниця між похибкою многочлена найкращого наближення і інтерполяційного многочлена неістотна.

Приклад 4. Розглянемо задачу знаходження многочлена найкращого наближення степеня n у випадку, коли

$$f(x) = p_{n+1}(x) = a_0 + \cdots + a_{n+1}x^{n+1}, \quad a_{n+1} \neq 0.$$

Тоді $f^{(n+1)}(x) = a_{n+1}(n+1)!$ і в силу (15), (16)

$$\Delta_n(f) = |a_{n+1}| (b-a)^{n+1} 2^{-2n-1},$$

тобто многочленом найкращого наближення є інтерполяційний многочлен *за вузлами* $x_k = \dfrac{1}{2}\left[(a + b) + (b - a)\cos\left(\dfrac{\pi(2k+1)}{2(n+1)}\right)\right]$, $k = \overline{0, n}$. Легко помітити, що цей многочлен степеня n має вигляд

$$\varphi_n(x) = p_{n+1}(x) - a_{n+1}T_{n+1}\left(\frac{2x - (a+b)}{b-a}\right)\frac{(b-a)^{n+1}}{2^{2n+1}},$$

де $T_{n+1}(x)$ — многочлен Чебишева $(n+1)$-го степеня першого роду.

Приклад 5. Нехай функція $f(x)$ парна на відрізку $[-1, 1]$, тобто $f(-x) = f(x)$. Покажемо, що многочлен найкращого наближення $\varphi_n(x)$ довільного степеня n буде парним. Дійсно,

$$|f(x) - \varphi_n(x)| \leqslant \Delta_n(f).$$

Замінюючи x на $-x$, маємо

$$|f(-x) - \varphi_n(-x)| = |f(x) - \varphi_n(-x)| \leqslant \Delta_n(f),$$

тобто многочлен $\varphi_n(-x)$ також є многочленом найкращого рівномірного наближення.

В силу єдиності $\varphi_n(x) = \varphi_n(-x)$. Якщо тепер потрібно наблизити, наприклад, функцію $f(x) = \exp x^2$ на $[-1, 1]$ многочленом третього степеня, то з вищевикладеного випливає, що цей многочлен має вигляд $a_0 + a_2x^2$. Задача стає еквівалентною задачі найкращого наближення функції $f_1(y) = \exp y$ на $[0, 1]$ многочленом першого степеня $a_0 + a_2y$ і розв'язується так, як у прикладі 2.

Аналогічно можна показати, що многочлен найкращого наближення непарної функції $f(x)$ також є непарним, тобто являє собою лінійну комбінацію непарних степенів x. Існують і загальні алгоритми визначення елементів найкращого наближення в банахових і нормованих просторах. Однак істотною відміною їх від алгоритму визначення елемента найкращого наближення в просторі із скалярним добутком, де цей елемент визначається за скінченне число дій, є те, що вони ітераційні і мають досить складну логіку.

3.11. Зв'язок між елементами найкращого наближення в просторах $L_p(a, b)$ та $C[a, b]$

На початку 3.10 ми вже розглядали «зв'язок» між просторами $L_p(a, b)$ і $C[a, b]$, а саме те, що коли $f \in C[a, b]$, то $\lim\limits_{p \to \infty} \|f\|_{L_p(a,b)} = \|f\|_{C[a,b]}$. Покажемо, що подібний зв'язок має місце і між многочленами найкращого наближення в просторах $L_p(a, b)$ і $C[a, b]$.

Нехай $\Phi_0^{(p)}(x) \in M \subset C[a, b]$ — многочлен найкращого наближення до $f(x) \in C[a, b]$ у смислі норми простору $L_p(a, b)$, тобто

$$\Delta^{(p)}(f) \equiv \Delta_n^{(p)}(f) = \inf_{\Phi \in M} \| f - \Phi \|_{L_p(a,b)} = \Delta_n^{(p)}(f, \Phi_0^{(p)}). \tag{1}$$

Насамперед доведемо таке допоміжне твердження.

Лема 1. *Нехай задана послідовність узагальнених многочленів*

$$\Phi^{(k)}(x) = a_0^{(k)}\varphi_0(x) + a_1^{(k)}\varphi_1(x) + \cdots + a_n^{(k)}\varphi_n(x), \quad k = 1, 2, \ldots,$$

заданого степеня n за системою Чебишева $\{\varphi_i(x)\}_{i=0}^n$, причому $\varphi_i(x) \in C[0, 1]$ $\forall\, i = \overline{0, n}$ і

$$\| \Phi^{(k)}(x) \|_{L_{p_k}(0,1)} \leqslant M, \quad 1 \leqslant p_k \leqslant +\infty, \tag{2}$$

де стала M не залежить від k. Тоді з цієї послідовності можна виділити підпослідовність $\{\Phi^{(k_l)}(x)\}$, що рівномірно збігається на відрізку $[0, 1]$, або що те саме, підпослідовність, для якої існує границя

$$\lim_{k_l \to \infty} a_s^{(k_l)} = a_s, \quad s = \overline{0, n}, \tag{3}$$

де a_s — деякі числа.

Доведення. Нехай $M_i = \sup_{x \in [0,1]} |\varphi_i(x)|$, $M^* = \max_{i=\overline{0,n}} M_i$.

В силу нерівності Гельдера маємо

$$\left| \sum_{s=0}^n a_s^{(k)} (\varphi_s, \varphi_l) \right| \equiv \left| \sum_{s=0}^n a_s^{(k)} \int_0^1 \varphi_s(x)\varphi_l(x)\,dx \right| =$$

$$= \left| \int_0^1 \Phi^{(k)}(x)\varphi_l(x)\,dx \right| \leqslant M^* \int_0^1 |\Phi^{(k)}(x)|\,dx \leqslant$$

$$\leqslant M^* \left(\int_0^1 |\Phi^{(k)}(x)|^{p_k} dx \right)^{1/p_k} \left(\int_0^1 1^{q_k} dx \right)^{1/q_k} \leqslant M^* M,$$

$$k = 1, 2, \ldots, \quad \frac{1}{p_k} + \frac{1}{q_k} = 1.$$

Звідси випливає, що для кожного k система лінійних рівнянь

$$\sum_{s=0}^n a_s^{(k)} (\varphi_s, \varphi_l) = \lambda_l^{(k)}, \quad l = \overline{0, n},$$

$$(\varphi_s, \varphi_l) \equiv \int_0^1 \varphi_s(x)\varphi_l(x)\,dx, \tag{4}$$

$$\lambda_l^{(k)} = \int_0^1 \Phi^{(k)}(x)\varphi_l(x)\,dx = (\Phi^{(k)}, \varphi_l)$$

з матрицею $A = \{(\varphi_s, \varphi_l)\}_{s,l=\overline{0,n}}$ і невідомими $a_s^{(k)}$, $s = \overline{0, n}$, має праві частини, які задовольняють нерівність

$$|\lambda_l^{(k)}| \leqslant M^* M. \tag{5}$$

Визначник системи (4) як визначник Грама системи лінійно незалежних функцій не дорівнює нулю і не залежить від k, тобто існує матриця $A^{-1} = \{a_{ij}^{(-1)}\}$ і $\| A^{-1} \|_\infty = M^{**}$, де M^{**} не залежить від k, $\| A^{-1} \|_\infty = \max_i \sum_{j=0}^n |a_{ij}^{(-1)}|$. В силу узгодженості матричної норми $\| A^{-1} \|_\infty$ і векторної норми $\| a^{(k)} \|_\infty = \max_{s=\overline{0,n}} |a_s^{(k)}|$, $a^{(k)} = (a_0^{(k)}, \ldots, a_n^{(k)})^T$, з (4) і (5) маємо

$$\max_s |a_s^{(k)}| \leqslant \| A^{-1}\lambda^{(k)} \|_\infty \leqslant \| A^{-1} \|_\infty \| \lambda^{(k)} \|_\infty \leqslant M M^* M^{**}, \tag{6}$$

де $\lambda^{(k)} = (\lambda_0^{(k)}, \ldots, \lambda_n^{(k)})^T$.

Розглянемо тепер послідовність

$$a_0^{(1)}, a_0^{(2)}, \ldots.$$

В силу (6) вона обмежена і тому з неї можна виділити підпослідовність

$$a_0^{(n_1^{(0)})}, a_0^{(n_2^{(0)})}, \ldots, \tag{7}$$

що збігається до деякого числа a_0. В силу (6) послідовність

$$a_1^{(n_1^{(0)})}, a_1^{(n_2^{(0)})}, \ldots$$

також обмежена і із неї можна виділити підпослідовність

$$a_1^{(n_1^{(1)})}, a_1^{(n_2^{(1)})}, \ldots,$$

яка збігається до деякого a_1. Продовжуючи цей процес $n + 1$ разів, ми, нарешті, дістаємо підпослідовність натуральних чисел $k_1 = n_1^{(n)}$, $k_2 = n_2^{(n)}$, ..., для якої будуть одночасно мати місце всі $n + 1$ рівностей (3), а це все одно, що має місце рівність $\lim_{l \to \infty} \Phi^{(k_l)}(x) = \Phi(x)$ рівномірно на $[0, 1]$, де $\Phi(x) = a_0\varphi_0(x) + \ldots + a_n\varphi_n(x)$ — деякий узагальнений многочлен. Лема доведена.

Теорема 1. *Послідовність $\Phi^{(p)}(x)$ узагальнених многочленів степеня n найкращого наближення для $f(x) \in C[a, b]$ в сенсі норми простору $L_p(a, b)$ при $p \to \infty$ рівномірно збігається до узагальненого многочлена найкращого рівномірного наближення $\Phi_0(x)$, тобто рівномірно по x*

$$\lim_{p \to \infty} \Phi^{(p)}(x) = \Phi_0(x).$$

Доведення. Не обмежуючи загальності, вважатимемо, що $[a, b] = [0, 1]$. Нехай $\{m_k\}_{k=0}^{\infty}$ — послідовність, $m_k > 1\ \forall\ k$, причому $\lim\limits_{k\to\infty} m_k = +\infty$ (цю послідовність будемо називати вихідною). Оскільки $\Phi^{(m_k)}(x)$ — многочлен найкращого наближення в $L_{m_k}(0, 1)$, то

$$\|\Phi^{(m_k)} - f\|_{L_{m_k}(0,1)} \leqslant \|\Phi_0(x) - f(x)\|_{L_{m_k}(0,1)} \leqslant$$

$$\leqslant \|\Phi_0(x) - f(x)\|_{C[0,1]} = \Delta(f),$$

де $\Delta(f) = \Delta(f, \Phi_0) = \inf\limits_{\Phi\in M} \|\Phi - f\|_{C[0,1]}$, $\Phi_0 = \sum\limits_{i=0}^{n} a_i^0 \varphi_i(x)$.

В силу леми 1 з послідовності $\{m_k\}$ можна виділити підпослідовність $\{m_{k_s}\}$ таку, що

$$\lim_{s\to\infty} \Phi^{(m_{k_s})}(x) = \Phi^{(\infty)}(x), \tag{8}$$

причому збіжність рівномірна на $[0, 1]$, де $\Phi^{(\infty)}(x)$ — узагальнений многочлен степеня n. Іншими словами, для будь-якого $\varepsilon > 0$ існує $s = s(\varepsilon)$ таке, що для всіх $s > s(\varepsilon)$

$$|\Phi^{(m_{k_s})}(x) - \Phi^{(\infty)}(x)| < \varepsilon, \quad \forall\ x \in [0, 1].$$

Звідси і з того, що $\Phi^{(m_{k_s})}(x)$ є многочленом найкращого наближення в $L_{m_{k_s}}(0, 1)$, випливає такий ланцюжок нерівностей:

$$\|\Phi^{(\infty)} - f\|_{L_{m_{k_s}}(0,1)} - \varepsilon \leqslant \|\Phi^{(\infty)} - f\|_{L_{m_{k_s}}(0,1)} -$$

$$- \|\Phi^{(\infty)} - \Phi^{(m_{k_s})}\|_{L_{m_{k_s}}(0,1)} \leqslant \|\Phi^{(m_{k_s})} - f\|_{L_{m_{k_s}}(0,1)} \leqslant$$

$$\leqslant \|\Phi^{(\infty)} - f\|_{L_{m_{k_s}}(0,1)} \leqslant \|\Phi^{(\infty)} - f\|_{C[0,1]}. \tag{9}$$

З огляду на формулу (6) з п. 3.10 маємо

$$\lim_{s\to\infty} \|\Phi^{(\infty)} - f\|_{L_{m_{k_s}}(0,1)} = \|\Phi^{(\infty)} - f\|_{C[0,1]},$$

тому при достатньо великих $s > s(\varepsilon)\ \forall\ \varepsilon > 0$

$$\|\Phi^{(\infty)} - f\|_{L_{m_{k_s}}(0,1)} \geqslant \|\Phi^{(\infty)} - f\|_{C[0,1]} - \varepsilon$$

і з (9) маємо

$$\|\Phi^{(\infty)} - f\|_{C[0,1]} - 2\varepsilon \leqslant \|\Phi^{(m_{k_s})} - f\|_{L_{m_{k_s}}(0,1)} \leqslant \|\Phi^{(\infty)} - f\|_{C[0,1]},$$

що внаслідок довільності ε означає

$$\lim_{s\to\infty} \|\Phi^{(m_{k_s})} - f\|_{L_{m_{k_s}}(0,1)} = \|\Phi^{(\infty)} - f\|_{C[0,1]}. \tag{10}$$

Переходячи тепер до границі в нерівності

$$\|\Phi^{(m_{k_s})} - f\|_{L_{m_{k_s}}(0,1)} \leqslant \|\Phi_0 - f\|_{L_{m_{k_s}}(0,1)}$$

і використовуючи в лівій частині співвідношення (10), а в правій — (6) з п. 3.10, маємо

$$\|\Phi^{(\infty)} - f\|_{C[0,1]} \leqslant \|\Phi_0 - f\|_{C[0,1]}.$$

Внаслідок єдиності многочлена Φ_0 найкращого наближення в $C[0, 1]$ випливає рівність $\Phi^{(\infty)} \equiv \Phi_0$. Таким чином, ми показали, що рівномірно по x

$$\lim_{s\to\infty} \Phi^{(m_{k_s})}(x) = \Phi_0(x), \tag{11}$$

звідки

$$\lim_{s\to\infty} a_j^{(m_{k_s})} = a_j^{(0)}, \tag{11'}$$

де $a_j^{(0)}$ — j-й коефіцієнт многочлена $\Phi_0(x)$.

Покажемо тепер, що рівномірно по x

$$\lim_{k\to\infty} \Phi^{(m_k)}(x) = \Phi_0(x). \tag{12}$$

Для цього, очевидно, достатньо показати, що для будь-якого фіксованого $j \in [0, n]$

$$\lim_{k\to\infty} a_j^{(m_k)} = a_j^{(0)}, \tag{13}$$

де $a_j^{(0)}$ — коефіцієнт многочлена $\Phi_0(x)$. Тоді буде

$$\|\Phi^{(m_k)} - \Phi_0\|_{C[0,1]} = \left\| \sum_{j=0}^{n} (a_j^{(m_k)} - a_j^{(0)}) \varphi_j(x) \right\|_{C[0,1]} \leqslant$$

$$\leqslant \max_{j=\overline{0,n}} \|\varphi_j(x)\|_{C[0,1]} \sum_{j=0}^{n} |a_j^{(m_k)} - a_j^{(0)}|,$$

звідки випливає рівність

$$\lim_{k\to\infty} \|\Phi^{(m_k)} - \Phi_0\|_{C[0,1]} = 0, \tag{14}$$

що еквівалентно (12).

Припустимо, що (13) не справджується. Тоді $\forall\ \varepsilon > 0$ зовні ε-околу числа $a_j^{(0)}$ лежить нескінченне число членів $a_j^{(m_{k_t})}$ $(t = \overline{0, \infty})$ послідовності $\{a_j^{m_k}\}_{k=0}^{\infty}$. Вибравши послідовність $\{m_{k_t}\}_{t=0}^{\infty}$ як вихідну, ми не змогли б дістати (11'), що суперечить доведеному вище. Таким чином, наше припущення про те, що (13) не справедливе, не вірне і разом з (13) доведено і (12).

Оскільки за вихідну послідовність $\{m_k\}_{k=0}$ ми брали довільну послідовність, то твердження теореми 1 доведено.

Зауваження. Якщо немає обмеження на степінь n многочлена найкращого рівномірного наближення, то можна скористатися теоремою вкладення з 1.5, в силу якої

$$\|f\|_{C[a,b]} \leqslant 2\max\left(\sqrt{b-a}, \frac{1}{\sqrt{b-a}}\right)\|f\|_{W_2^1(a,b)}. \tag{15}$$

Нехай $M \subset C^1[a, b]$ (тобто $\varphi_i(x) \in C^1[a, b]$, $i = \overline{0, n}$), $f \in C^1[a, b]$ і $\Phi_1(x)$ — многочлен найкращого наближення в просторі $W_2^1(a, b)$. Оскільки $W_2^1(a, b)$ — простір із скалярним добутком, то цей елемент визначається за скінченне число дій за допомогою алгоритму з п. 3.3. З нерівності (15) випливає, що при достатньо малій нормі $\|f - \Phi_1\|_{W_2^1(a,b)}$ (цього можна досягти збільшенням степеня n многочлена Φ_1), елемент Φ_1 добре наближатиме $f(x)$ і в просторі $C[a, b]$.

ГЛАВА 4

СПЛАЙНИ

У наш час теорія наближення функцій збагатилася новими методами, які називаються *сплайн-апроксимаціями*. Зазначимо, що *сплайном* називається функція, для якої існує подрібнення її області визначення D на підобласті таке, що в середині кожної підобласті функція є многочленом деякого степеня m. Крім цього, ця функція, як правило, неперервна в D разом з похідними до $(m-1)$-го порядку і має інтегровну з квадратом похідну m-го порядку. Найпоширеніші в інженерних розрахунках сплайни складені з многочленів третього степеня (кубічні сплайни).

4.1. Інтерполювання функцій однієї змінної за допомогою кубічних сплайнів

Нехай на відрізку $[a, b]$ дійсної осі задано сітку $\omega = \{x_i : x_0 = a < x_1 < x_2 < \ldots < x_n = b\}$, у вузлах якої задано значення $\{f_k\}_{k=0}^n$ функції $f(x)$, визначеної на $[a, b]$. Задача кусково-кубічної інтерполяції ставиться таким чином: знайти функцію $g(x)$, яку називатимемо *кубічним сплайном*, визначену на $[a, b]$ і таку, що

1) $g(x) \in C^{(2)}[a, b]$; (1)

2) на кожному з відрізків $[x_{k-1}, x_k]$ функція $g(x)$ являється кубічним многочленом виду

$$g(x) \equiv g_k(x) = \sum_{i=0}^{3} a_i^{(k)} (x_n - x)^i, \quad k = 1, \ldots, n; \tag{2}$$

3) у вузлах сітки ω виконуються рівності

$$g(x_k) = f_k, \quad k = \overline{0, n}; \tag{3}$$

4) $g''(x)$ задовольняє граничні умови

$$g''(a) = g''(b) = 0. \tag{4}$$

Покажемо, що поставлена задача має єдиний розв'язок і вкажемо алгоритм його обчислення.

В силу (2) для $x \in [x_{i-1}, x_i]$, $i = \overline{1, n}$, маємо

$$g''(x) = m_{i-1}\frac{x_i - x}{h_i} + m_i\frac{x - x_{i-1}}{h_i}, \tag{5}$$

де $h_i = x_i - x_{i-1}$, $m_i = g''(x_i)$. Інтегруючи двічі обидві частини рівності (5), дістаємо

$$g(x) = m_{i-1}\frac{(x_i - x)^3}{6h_i} + m_i\frac{(x - x_{i-1})^3}{6h_i} + A_i\frac{x_i - x}{h_i} + B_i\frac{x - x_{i-1}}{h_i}, \tag{6}$$

де A_i, B_i — деякі сталі інтегрування. Знайдемо їх з умов $g(x_{i-1}) = f_{i-1}$, $g(x_i) = f_i$ (див. (3)). Підставляючи в (6) $x = x_i$ і $x = x_{i-1}$, маємо

$$m_i\frac{h_i^2}{6} + B_i = f_i, \quad m_{i-1}\frac{h_i^2}{6} + A_i = f_{i-1}, \tag{7}$$

звідки $B_i = f_i - \frac{m_i h_i^2}{6}$, $A_i = f_{i-1} - \frac{m_{i-1}h_i^2}{6}$. Підставивши ці значення в (6), знайдемо для $x \in [x_{i-1}, x_i]$

$$g(x) = m_{i-1}\frac{(x_i - x)^3}{6h_i} + m_i\frac{(x - x_{i-1})^3}{6h_i} + \left(f_{i-1} - \frac{m_{i-1}h_i^2}{6}\right)\frac{x_i - x}{h_i} + \left(f_i - \frac{m_i h_i^2}{6}\right)\frac{x - x_{i-1}}{h_i}, \tag{8}$$

$$g'(x) = -m_{i-1}\frac{(x_i - x)^2}{2h_i} + m_i\frac{(x - x_{i-1})^2}{2h_i} + \frac{f_i - f_{i-1}}{h_i} - \frac{m_i - m_{i-1}}{6}h_i.$$

Із виразу (8)

$$g'(x_i - 0) = \frac{h_i}{6}m_{i-1} + \frac{h_i}{3}m_i + \frac{f_i - f_{i-1}}{h_i}. \tag{9}$$

Записавши (8) для відрізка $[x_i, x_{i+1}]$, тобто

$$g'(x) = -m_i\frac{(x_{i+1} - x)^2}{2h_{i+1}} + m_{i+1}\frac{(x - x_i)^2}{2h_{i+1}} + \frac{f_{i+1} - f_i}{h_{i+1}} - \frac{m_{i+1} - m_i}{6}h_{i+1},$$

знайдемо

$$g'(x_i+0) = -\frac{h_{i+1}}{3}m_i - \frac{h_{i+1}}{6}m_{i+1} + \frac{f_{i+1}-f_i}{h_{i+1}}. \qquad (10)$$

За умовою (1) функції $g'(x)$ і $g''(x)$ мають бути неперервними на $[a, b]$. Враховуючи (9), (10), з умови неперервності $g'(x)$ в точках x_i $i = \overline{1, n-1}$, дістаємо $n-1$ рівнянь

$$\frac{h_i}{6}m_{i-1} + \frac{h_i+h_{i+1}}{3}m_i + \frac{h_{i+1}}{6}m_{i+1} = \frac{f_{i+1}-f_i}{h_{i+1}} - \frac{f_i-f_{i-1}}{h_i}. \qquad (11)$$

Доповнивши ці рівняння рівностями $m_0 = m_n = 0$, які випливають з (4), запишемо систему лінійних алгебраїчних рівнянь для знаходження невідомих $m_1, \ldots, m_{n-1}$:

$$Am = Hf. \qquad (12)$$

Тут квадратна матриця A має вигляд

$$A = \begin{bmatrix} \frac{h_1+h_2}{3} & \frac{h_2}{6} & 0 & 0 & \ldots & 0 & 0 & 0 \\ \frac{h_2}{6} & \frac{h_2+h_3}{3} & \frac{h_3}{6} & 0 & \ldots & 0 & 0 & 0 \\ 0 & \frac{h_3}{6} & \frac{h_3+h_4}{3} & \frac{h_4}{6} & \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & \frac{h_{n-2}}{6} & \frac{h_{n-2}+h_{n-1}}{3} & \frac{h_{n-1}}{6} \\ 0 & 0 & 0 & 0 & \ldots & 0 & \frac{h_{n-1}}{6} & \frac{h_n+h_{n-1}}{3} \end{bmatrix}, \qquad (13)$$

а вектори m, f і прямокутна матриця H, яка має $n+1$ стовпчиків і $n-1$ рядків, такі:

$$m = \begin{pmatrix} m_1 \\ m_2 \\ \ldots \\ m_{n-1} \end{pmatrix}, \quad f = \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ \ldots \\ f_{n-1} \\ f_n \end{pmatrix}, \qquad (14)$$

$$H = \begin{bmatrix} \frac{1}{h_1} & \left(-\frac{1}{h_1}-\frac{1}{h_2}\right) & \frac{1}{h_2} & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{h_2} & \left(-\frac{1}{h_2}-\frac{1}{h_3}\right) & \frac{1}{h_3} & \ldots & 0 & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & \frac{1}{h_{n-2}} & \left(-\frac{1}{h_{n-2}}-\frac{1}{h_{n-1}}\right) & \frac{1}{h_{n-1}} & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & \frac{1}{h_{n-1}} & \left(-\frac{1}{h_{n-1}}-\frac{1}{h_n}\right) & \frac{1}{h_n} \end{bmatrix}.$$

4.1.1. Властивості матриці A. Матриця A симетрична і для її елементів a_{ij}, $i, j = \overline{1, n-1}$, виконується співвідношення

$$\min_i \left(|a_{ii}| - \sum_{j \neq i} |a_{ij}|\right) = q > 0, \qquad (15)$$

де

$$q = \min\left(\frac{h_1}{3} + \frac{h_2}{6}, \min_{i=2,\ldots,n-2} \frac{h_i+h_{i+1}}{6}, \frac{h_n}{3} + \frac{h_{n-1}}{6}\right).$$

Означення. Квадратна матриця A називається *матрицею з строгою діагональною перевагою*, якщо для неї виконується умова (15).

Лема 1. *Матриця зі строгою діагональною перевагою невироджена, причому*

$$\|A^{-1}\|_\infty \equiv \max_i \sum_j |a_{ij}^{(-1)}| \leqslant \left\{\min_i \left(|a_{ii}| - \sum_{i \neq j} |a_{ij}|\right)\right\}^{-1} = q^{-1}.$$

Доведення. Припустивши супротивне, дістаємо, що однорідна система лінійних алгебраїчних рівнянь $Ax = 0$ має нетривіальний розв'язок. Запишемо цю систему в координатній формі:

$$\sum_{j=1}^{n} a_{ij}x_j = 0, \quad i = \overline{1, n-1}.$$

Нехай індекс k такий, що $|x_k| = \|x\| \equiv \max\limits_i |x_i| \geqslant x_i$, $i = \overline{1, n-1}$. Тоді з k-го рівняння системи дістанемо

$$|a_{kk}|\|x\| \leqslant \sum_{j\neq k} |a_{kj}||x_j| \leqslant \|x\| \sum_{j\neq k} |a_{kj}|,$$

звідки, враховуючи, що $x = (x_1, \ldots, x_{n-1})$ — нетривіальний розв'язок $\|x\| \neq 0$, знаходимо $|a_{kk}| \leqslant \sum\limits_{j\neq k} |a_{kj}|$. Остання нерівність суперечить тому, що матриця A має строгу діагональну перевагу. Ця суперечність доводить першу частину леми. Норма матриці $\|A^{-1}\|_\infty$ породжується першою нормою вектора $\|x\|_\infty = \max\limits_i |x_i|$, тобто

$$\|A^{-1}\|_\infty = \sup_{y\neq 0} \frac{\|A^{-1}y\|_\infty}{\|y\|_\infty} = \sup_{Ax\neq 0} \frac{\|x\|_\infty}{\|Ax\|_\infty}, \tag{16}$$

і тому є узгодженою з нею, тобто $\|A^{-1}y\|_\infty \leqslant \|A^{-1}\|_\infty \|y\|_\infty$. Нехай $Ax = y$, $x = A^{-1}y$, і $\|x\|_\infty = |x_k|$, $1 \leqslant k \leqslant n$, тоді

$$\|y\|_\infty = \|Ax\|_\infty = \max_i \left|\sum_j a_{ij}x_j\right| \geqslant \left|\sum_j a_{kj}x_j\right| =$$

$$= \left|a_{kk}x_k + \sum_{j\neq k} a_{kj}x_j\right| \geqslant |x_k||a_{kk}| - \left|\sum_{j\neq k} a_{kj}x_j\right| \geqslant$$

$$\geqslant |x_k||a_{kk}| - |x_k| \sum_{j\neq k} |a_{kj}| \geqslant \|x\|_\infty \min\left(|a_{ii}| - \sum_{j\neq i} |a_{ij}|\right) = \|x\|_\infty q.$$

Тому з (16) маємо

$$\|A^{-1}\|_\infty = \sup_{Ax\neq 0} \frac{\|x\|_\infty}{\|Ax\|_\infty} = \sup \frac{\|x\|_\infty}{\|y\|_\infty} \leqslant q^{-1}.$$

Лему доведено.

З леми 1 випливає, що система (12) має єдиний розв'язок m_1, m_2, ..., m_{n-1}. Таким чином, сплайн-функція $g(x)$ також однозначно відтворюється за формулами (17) і розв'язує задачу кусково-кубічної інтерполяції.

Для розв'язування системи (12) використовується ефективний метод прогонки, який розглядали в п. 1.2.

4.1.2. Екстремальна властивість інтерполяційного кубічного сплайна. Очевидно, що $g(x) \in \tilde{W}_2^3([a, b]) \subset \tilde{W}_2^2([a, b])$. Кубічні сплайн-функції мають важливу властивість, яка встановлюється такою теоремою.

Теорема. *Мінімум функціонала*

$$\Phi(u) = \int_a^b [u''(x)]^2 dx \tag{17}$$

на класі функцій $u \in W_2^2(a, b)$ *і* $u(x_k) = f_k$, $k = \overline{0, n}$, *де* f_k *— задані, досягається тільки на кусково-кубічній сплайн-функції* $g(x)$.

Доведення. Розглянемо функціонал

$$\Phi(u - g) = \int_a^b [u'' - g'']^2 dx \geqslant 0. \tag{18}$$

Інтегруючи частинами, дістанемо

$$\Phi(u - g) = \int_a^b [u'']^2 dx - 2\int_a^b u''g''dx + \int_a^b [g'']^2 =$$

$$= \int_a^b [u'']^2 dx - \int_a^b [g'']^2 dx - 2\int_a^b (u - g)''g''dx =$$

$$= \Phi(u) - \Phi(g) - 2\left[(u' - g')g''\Big|_a^b - \int_b^b (u' - g')g'''dx\right] =$$

$$= \Phi(u) - \Phi(g) + 2\sum_{k=1}^n \int_{x_{k-1}}^{x_k} (u' - g')g'''dx.$$

Але на відрізку $[x_{k-1}, x_k]$ маємо $g''' = c_k = \text{const}$. Тому

$$\Phi(u - g) = \Phi(u) - \Phi(g) + 2\sum_{k=1}^n c_k(u - g)\Big|_{x_{k-1}}^{x_k} = \Phi(u) - \Phi(g). \tag{19}$$

Звідси і з виразу (18) випливає, що

$$\Phi(g) = \Phi(u) - \Phi(u - g) \leqslant \Phi(u),$$

$$\forall u(x) \in W_2^2(a, b), \quad u(x_k) = f_k, \quad k = \overline{0, n},$$

тобто на сплайн-функції $g(x)$ досягається мінімум функціонала $\Phi(u)$. Покажемо, що інших точок мінімуму цей функціонал не має.

Припустимо від супротивного, що існує функція $\bar{g}(x) \neq g(x)$ з $W_2^2(a, b)$, на якій досягається мінімум функціонала, тобто $\Phi(\bar{g}) = \Phi(g)$, $\bar{g}(x_k) = f_k$. Тоді з (19) дістанемо $\Phi(\bar{g} - g) = \Phi(\bar{g}) - \Phi(g) = 0$, тобто $\bar{g}'' - g'' = 0$. Остання рівність означає, що $\bar{g}(x) = g(x) + p_1(x)$, де $p_1(x) = ax + b$ — довільний многочлен першого степеня. Але в силу того, що $\bar{g}(x_k) = g(x_k)$, $k = \overline{0, n}$, маємо $p_1(x_k) = 0$, $k = \overline{0, n}$. Звідси випливає $p_1(x) \equiv 0$, тобто $\bar{g}(x) \equiv g(x)$, що суперечить припущенню. Теорему доведено.

Зауваження 1. Грунтуючись на теоремі 1, можна дати інше еквівалентне означення кусково-кубічної сплайн-функції: це функція з класу $W_2^2(a, b)$, яка набуває у вузлах сітки заданих значень і мінімізує функціонал (17).

Зауваження. 2. Враховуючи формулу (5), значення функціонала на сплайн-функції $g(x)$ можна записати у вигляді

$$0 < \Phi(g) = \int_a^b [g''(x)]^2\,dx = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \left[m_{k-1} \frac{x_k - x}{h_k} + m_k \frac{x - x_{k-1}}{h_k} \right]^2 dx = \sum_{k=1}^{n} m_k \left| m_{k-1} \frac{h_k}{6} + \frac{h_k + h_{k+1}}{3} m_k + \frac{h_{k+1}}{6} m_{k+1} \right| = (Am, m), \tag{20}$$

де матриця A визначається формулою (13), а $(\cdot, \cdot)$ визначає скалярний добуток у просторі $(n-1)$-вимірних векторів. Зазначимо, що для довільного ненульового вектора m справедлива строга нерівність $0 < \Phi(g) = (Am, m)$. Розглянемо довільний вектор $m^* = (m_1^*, \ldots, m_{n-1}^*) \neq 0$ в цьому просторі. Для цього вектора завжди можемо знайти такий вектор $f^* = (f_0^*, f_1^*, \ldots, f_n^*)$, що виконуватиметься рівність (12), тобто

$$Am^* = Hf^*. \tag{21}$$

Дійсно, розглянемо (21) як систему лінійних алгебраїчних рівнянь з матрицею H відносно невідомих $f_0^*, \ldots, f_n^*$. Ранг матриці H дорівнює $n-1$, тому що, наприклад, її головний мінор

$$\det \begin{vmatrix} \frac{1}{h_1}\left(-\frac{1}{h_1} - \frac{1}{h_2}\right) & \frac{1}{h_2} & 0 & \ldots & 0 & 0 \\ 0 & \frac{1}{h_2} & \left(-\frac{1}{h_2} - \frac{1}{h_3}\right) & \frac{1}{h_3} & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & \frac{1}{h_{n-2}}\left(-\frac{1}{h_{n-2}} - \frac{1}{h_{n-1}}\right) \\ 0 & 0 & 0 & 0 & \ldots & 0 & \frac{1}{h_{n-1}} \end{vmatrix} = \prod_{k=1}^{n-1} \frac{1}{h_k},$$

відмінний від нуля. Тому і ранг розширеної матриці дорівнює $n-1$, отже система (21) завжди має розв'язок відносно $f_0^*, \ldots, f_n^*$ для довільного заданого m^*. Звідси випливає, що довільний вектор m^* визначає деяку кубічну сплайн-фукцію g^*, яку можемо знайти за формулою (7), підставивши в неї значення m^* і f^*. Через те що вираз (20) справедливий для довільної кубічної сплайн-функції g^*, то для довільного вектора $m^* \neq 0$ маємо

$$(Am^*, m^*) > 0,$$

тобто $A > 0$, де матриця A визначається рівністю (13). Це означає, що матриця A додатно визначена. Таким чином, доведено таке твердження.

Лема 2. *Матриця A, яка визначається рівністю* (13), *додатно визначена, тобто існує стала $\delta > 0$ така, що для довільного вектора m*

$$(Am, m) \geqslant \delta \|m\|^2, \quad \|m\|^2 = (m, m) = \sum_{i=1}^{n-1} m_i^2.$$

Приклад. Побудувати інтерполяційний кубічний сплайн для функції, заданої таблицею

i	0	1	2	3	4	5	6
x_i	−15	−12	−6	0	6	12	15
$f(x_i)$	1	1	1	1	1	1	0

Розв'язання. За формулою $h_i = x_i - x_{i-1}$, $x_0 = -15$ знаходимо h_i, $i = \overline{1,6}$, $h_1 = h_6 = 3$, $h_2 = h_3 = h_4 = h_5 = 6$. Формуємо матриці (13), (14):

$$A = \begin{bmatrix} 3 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix},$$

$$H = \begin{bmatrix} \frac{1}{3} & -\frac{1}{2} & \frac{1}{6} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{6} & -\frac{1}{2} & \frac{1}{3} \end{bmatrix},$$

$$Hf = \left(0, 0, 0, 0, -\frac{1}{3}\right).$$

Далі для знаходження вектора $m = (m_1, m_2, m_3, m_4, m_5)$ треба розв'язати систему (12):

$$Am = Hf,$$

яка, як неважко помітити, збігається з системою вже розглянутого прикладу 4.1.2. І розв'язок, знайдений методом прогонки, має вигляд

$$m_1 = -\frac{1}{1254}, \quad m_2 = \frac{1}{418}, \quad m_3 = -\frac{11}{1254}, \quad m_4 = \frac{41}{1254}, \quad m_5 = -\frac{31}{418}.$$

Крім того, $m_0 = m_6 = 0$. Тепер за формулою (7) знаходимо кубічний сплайн:

$$g(x) =$$

$$= \begin{cases} -\frac{1}{1254}\frac{(x+15)^3}{18} - \frac{x+12}{3} + \frac{837}{836}\left(\frac{x+15}{3}\right), & x \in [-15, -12], \\ \frac{1}{1254}\frac{(x+6)^3}{36} + \frac{1}{418}\frac{(x+12)^3}{36} - \frac{215}{209}\frac{(x+5)}{6} + \frac{412}{418}\left(\frac{x+12}{6}\right), & x \in [-12, -6[, \\ -\frac{x^3}{416\cdot 36} - \frac{11}{1254\cdot 36}(x+6)^3 - \frac{103\cdot 206}{209\cdot 3}x + \frac{110}{209\cdot 3}(x+6), & x \in [-6, 0], \\ -\frac{11}{1254\cdot 36}(6-x)^3 + \frac{41}{1254\cdot 36}x^3 + \frac{110}{209\cdot 3}(6-x) + \frac{28}{209}x, & x \in [0, 6], \\ \frac{41}{1254\cdot 36}(12-x)^3 - \frac{17}{418\cdot 12}(x-6)^3 + \frac{28}{209}(12-x) + \frac{181}{209\cdot 3}(x-6), & x \in [6, 12], \\ -\frac{51}{418\cdot 12}(15-x) + \frac{989}{836\cdot 3}(15-x), & x \in [12, 15]. \end{cases}$$

4.2. Кубічні згладжуючі сплайн-функції. Зв'язок між згладжуючими та інтерполяційними сплайнами

Якщо в точках сітки $\omega = \{a = x_0 < x_1 < \ldots < x_n = b\}$ задано не точні значення функції $f(x_k)$, а наближені, то немає смислу наближати функцію $f(x)$ інтерполяційним сплайном. У цьому разі природно будувати наближаючу функцію, яка належить класу $W_2^2(a, b)$, за умови, що вона мінімізує функціонал:

$$\Phi_1(u) = \int_a^b [u'']^2\,dx + \sum_{k=0}^{n} p_k [u(x_k) - \tilde{f}_k]^2, \tag{1}$$

де p_k — додатні числа. Покажемо, що розв'язком цієї задачі є кубічний сплайн. Нехай $u_0 \in W_2^2(a, b)$ — розв'язок задачі. Побудуємо кубічний сплайн $\tilde{g}(x)$, такий, що $\tilde{g}(x_k) = u_0(x_k)$, $k = \overline{0, n}$. Тоді другий доданок в (1) буде однаковим для функції $\tilde{g}(x)$ і $u_0(x)$, тому

$$\int_a^b [u_0'']^2\,dx \leqslant \int_a^b [\tilde{g}'']^2\,dx. \tag{2}$$

Але, як випливає з теореми 4.1.1, $\tilde{g}(x)$ — єдина функція, яка надає мінімум функціонала (4.1.17). Тому $u_0 \equiv \tilde{g}$, тобто мінімум функціонала $\Phi_1(u)$ слід шукати в класі кубічних сплайнів. Оскільки кубічний сплайн цілком визначається множиною його значень $\{\mu_k\}_{k=0}^n$, яких він набуває у вузлах $\{x_k\}_{k=0}^n$, то мінімізація $\Phi_1(u)$ зводиться до знаходження мінімуму функції від змінних $\mu_0, \ldots, \mu_n$.

Ми бачили, що

$$\tilde{g}''(x) = m_{k-1}\frac{x_k - x}{h_k} + m_k\frac{x - x_{k-1}}{h_k}, \quad x \in [x_{k-1}, x_k],$$

$$m_k = g''(x_k), \quad h = \overline{1, n-1}, \quad m_0 = m_n = 0.$$

Тому

$$\Phi_1(\tilde{g}) = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \left[m_{k-1}\frac{x_k - x}{h_k} + m_k\frac{x - x_{k-1}}{h_k}\right]^2 dx + \sum_{k=0}^{n} p_k(\mu_k - \tilde{f}_k)^2. \tag{3}$$

Після інтегрування маємо

$$\Phi_1(\tilde{g}) = \sum_{k=1}^{n} \int_{x_{k-1}}^{x_n} \left[m_{k-1}\frac{x_k - x}{h_k} + m_k\frac{x - x_{k-1}}{h_k}\right]^2 dx + \sum_{k=0}^{n} p_k(\mu_k - \tilde{f}_k)^2 =$$

$$= \sum_{k=1}^{n} m_k\left[m_{k-1}\frac{h_k}{6} + \frac{h_k + h_{k+1}}{3}m_k + \frac{h_{k+1}}{6}m_{k+1}\right] + \sum_{k=0}^{n} p_k(\mu_k - \tilde{f}_k)^2 =$$

$$= (Am, m) + \sum_{k=0}^{n} p_k(\mu_k - \tilde{f}_k)^2, \tag{4}$$

де A — відома матриця (13) (п. 4.1). З формули (12) (п. 4.1) випливає, що вектор m лінійно виражається через вектор $\mu = (\mu_0, \mu_1, \ldots, \mu_n)$, а оскільки матриця A додатно визначена, то $\Phi_1(\tilde{g})$ — додатно визначена форма від μ. Її екстремум може бути тільки мінімумом, необхідною умовою якого є

$$\frac{\partial \Phi_1}{\partial \mu_s} \equiv \frac{\partial}{\partial \mu_s}(Am, m) + 2p_s(\mu_s - \tilde{f}_s) = 0, \quad s = \overline{0, n}.$$

Оскільки матриця A не залежить від μ, то в силу (12) (п. 4.1) маємо

$$\frac{\partial}{\partial \mu_s}(Am, m) = 2\left(\frac{\partial (Am)}{\partial \mu_s}, m\right) = 2\left(\frac{\partial (H\mu)}{\partial \mu_s}, m\right) =$$

$$= 2\left(\frac{\partial \mu}{\partial \mu_s}, H^*m\right) = 2(H^*m)_s,$$

де H визначається формулою (14) (п. 4.1). Звідси випливає, що у векторній формі умова мінімуму має вигляд

$$H^*m + P\mu = P\tilde{f}, \tag{5}$$

де $\tilde{f}=(\tilde{f}_0, \tilde{f}_1, \ldots, \tilde{f}_n)$,

$$P=\begin{pmatrix} p_0 & 0 & \ldots & 0 \\ 0 & p_1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & p_n \end{pmatrix}.$$

Помноживши (5) зліва на HP^{-1}, дістанемо

$$HP^{-1}H^*m + H\mu = H\tilde{f}$$

або з урахуванням (12) (п. 4.1)

$$(A + HP^{-1}H)\, m = H\tilde{f}. \tag{6}$$

Неважко помітити, що матриця системи (6) п'ятидіагональна, симетрична і додатно визначена, тому система має єдиний розв'язок, який можна знайти методом квадратних коренів або методом Гаусса. Після знаходження m визначаємо μ за формулою

$$\mu = \tilde{f} - P^{-1}H^*m, \tag{6'}$$

яка випливає з (5). Потім за формулою (7) (п. 4.1) відтворюємо сплайн $\tilde{g}(x)$.

Неважко помітити, що при $f_k = \tilde{f}_k$, $k = \overline{0, n}$, і при $\min\limits_k p_k \to \infty$ система (6) переходить в (12) (п. 4.1). Природно припустити, що за цих умов розв'язок системи (6) збігається до розв'язку системи (4.1.12), а це означає, що $\tilde{g}(x)$ рівномірно по x збігається до $g(x)$. Доведемо це.

Теорема 1. *Нехай $g(x)$ — згладжуючий кубічний сплайн, а $g(x)$ — інтерполяційний сплайн, побудований за значеннями $f_k = f(x_k)$, $k = \overline{0, n}$. Тоді рівномірно по x $\lim \tilde{g}(x) = g(x)$ при $\min\limits_k p_k \to \infty$.*

Доведення. Нехай $\tilde{m}$ — розв'язок системи (6), а m — розв'язок системи (12) (п. 4.1). Тоді різниця $\Delta m = \tilde{m} - m$ задовольняє рівняння

$$A\Delta m = -HP^{-1}H^*\tilde{m}. \tag{7}$$

Покажемо насамперед, що $\| HP^{-1}H^*\tilde{m} \|_\infty \to 0$ при $\min\limits_k p_k \to \infty$.

Для початку покажемо, що $\| \tilde{m} \|_\infty$ обмежена величиною, яка не залежить від p_k, $k = \overline{0, n}$. Дійсно, з (6) маємо

$$(A + HP^{-1}H^*)\,\tilde{m} = H\tilde{f}, \quad ((A + HP^{-1}H^*)\,\tilde{m}, \tilde{m}) = (H\tilde{f}, \tilde{m}), \tag{8}$$

$$(A\tilde{m}, \tilde{m}) + (HP^{-1}H^*\tilde{m}, \tilde{m}) = (H\tilde{f}, \tilde{m}),$$

де $(u, v) = \sum\limits_{i=1}^{n-1} u_i v_i$, $\| u \|^2 = (u, u)$. Оскільки $p_i > 0$, $i = \overline{0, n}$, то $(HP^{-1}H^*\tilde{m}, \tilde{m}) = (P^{-1}H^*\tilde{m}, H^*\tilde{m}) > 0$. Тому з (8) маємо

$$(A\tilde{m}, \tilde{m}) \leqslant \| H\tilde{f} \| \| \tilde{m} \|.$$

В силу леми 2 (п. 4.1)

$$\delta \| \tilde{m} \|^2 \leqslant H\tilde{f} \| \| \tilde{m} \|,$$

або

$$\| \tilde{m} \| \leqslant \delta^{-1} \| H\tilde{f} \|, \tag{9}$$

тобто $\| \tilde{m} \|$ обежена величиною, яка не залежить від $p_0, \ldots, p_n$. Неважко помітити, що $\| \tilde{m} \|_\infty = \max\limits_{i=\overline{1,n-1}} | \tilde{m}_i | \leqslant \sqrt{\sum\limits_{i=1}^{n-1} \tilde{m}_i^2} = \| \tilde{m} \|$, тому з (9) випливає

$$\| \tilde{m} \|_\infty \leqslant \delta^{-1} \| H\tilde{f} \|. \tag{10}$$

Далі з (7), (10) маємо

$$\Delta m = -A^{-1}HP^{-1}H^*\tilde{m},$$

звідки

$$\| \Delta m \|_\infty \leqslant \| A^{-1} \|_\infty \| H \|_\infty \| H^* \|_\infty \frac{1}{\min\limits_{k=\overline{0,n}} p_k} \| \tilde{m} \|_\infty \leqslant$$

$$\leqslant \| A^{-1} \|_\infty \| H \|_\infty \| H^* \|_\infty \delta^{-1} \| H\tilde{f} \| \frac{1}{\min\limits_{k=\overline{0,n}} p_k}$$

і, отже, $\| \Delta m \|_\infty \to 0$ при $\min\limits_k p_k \to \infty$, тобто $\tilde{m} \to m$. Крім того, з (6') випливає, що при $\min\limits_k p_k \to \infty$ буде $\mu \to f$ в нормі $\| \cdot \|_\infty$. З формули (8) (п. 4.1) випливає, що для $x \in [x_{i-1}, x_i]$ маємо

$$g(x) = m_{i-1}\frac{(x_i - x)^3}{6h_i} + m_i\frac{(x - x_{i-1})}{6h_i} + \left(f_{i-1} - \frac{m_{i-1}h_i^2}{6}\right)\frac{x_i - x}{6} + \left(f_i - \frac{m_i h_i^2}{6}\right)\frac{x - x_{i-1}}{h_i},$$

$$\tilde{g}(x) = \tilde{m}_{i-1}\frac{(x_i - x)^3}{6h_i} + \tilde{m}_i\frac{(x - x_{i-1})^3}{6h_i} + \left(\mu_{i-1} - \frac{\tilde{m}_{i-1}h_i^2}{6}\right)\frac{x_i - x}{6} + \left(\mu_i - \frac{\tilde{m}_i h_i^2}{6}\right)\frac{x - x_{i-1}}{h_i}$$

і далі

$$\max_{x\in[x_{i-1},x_i]}|g(x)-\tilde g(x)|\leqslant|m_{i-1}-\tilde m_{i-1}|\max_{x\in[x_{i-1},x_i]}\frac{(x_i-x)^3}{6h_i}+$$
$$+|m_i-\tilde m_i|\times$$
$$\times\max_{x\in[x_{i-1},x_i]}\frac{(x-x_{i-1})^3}{6h_i}+\left|f_{i-1}-\mu_{i-1}-\frac{m_{i-1}-\tilde m_{i-1}}{6}h_i^2\right|\times$$
$$\times\max_{x\in[x_{i-1},x_i]}\frac{x_i-x}{h_i}+\left|f_i-\mu_i-\frac{m_i-\tilde m_i}{6}h_i^2\right|\max_{x\in[x_{i-1},x_i]}\frac{x-x_{i-1}}{h_i}=$$
$$=|m_{i-1}-\tilde m_{i-1}|\frac{h_i^2}{6}+|m_i-\tilde m_i|\frac{h_i^2}{6}+|f_{i-1}-\mu_{i-1}|+$$
$$+|m_{i-1}-\tilde m_{i-1}|\frac{h_i^2}{6}+|f_i-\mu_i|+|m_i-\tilde m_i|\frac{h_i^2}{6}.$$

Звідси видно, що

$$\|g(x)-\tilde g(x)\|_{C[a,b]}=\max_{i=\overline{1,n}}\max_{x\in[x_{i-1},x_i]}|g(x)-\tilde g(x)|\to 0$$

при $\min_i p_k\to\infty$, що і треба було довести.

4.3. Деякі узагальнення

Можна розглядати сплайни, які на окремих інтервалах являються многочленами не третього степеня, як це було в попередніх параграфах, а многочленами степеня $2q-1$.

Означення 1. Простір дійсних функцій $s(x)$, визначених на відрізку $[a, b]$, що задовольняють умови

1) $s(x)\in\pi_{2q-1}\ \forall\, x\in(x_i, x_{i+1})$, $i=\overline{1, n-1}$;
2) $s(x)\in\pi_{q-1}\ \forall\, x\in[a, x_1)$, $(x_n, b]$;
3) $s(x)\in C^{(2q-2)}[a, b]$,

де $n\geqslant q$, π_k — простір многочленів степеня не вище k, називається *простором сплайн-функцій порядку q відносно точок x_i* і позначається через S.

Введемо позначення

$$(x)^+=\max(x, 0).$$

Неважко помітити, що функція

$$\delta_i(x)=[(x-x_i)^+]^{2q-1}/(2q-1)!$$

належить простору $C^{(2q-2)}[a, b]$:

$$\delta_i^{(2q-1)}(x)=\begin{cases}1, & \text{якщо } x\geqslant x_i,\\ 0, & \text{якщо } x<x_i.\end{cases}$$

Нехай $p_i(x)\in\pi_{2q-1}$ і $p_i(x)=s(x)\ \forall\, x\in(x_i, x_{i+1})$, тобто

$$p_i(x)=p_{i-1}(x)+d_i\delta_i(x),\quad x\in(x_i, x_{i+1}),$$

де

$$d_i=s^{(2q-1)}(x_i+0)-s^{(2q-1)}(x_i-0).$$

Неважко помітити, що $\forall\, s(x)\in S$ справедливе зображення

$$s(x)=p_0(t)+\sum_{i=1}^n d_i\frac{[(x-x_i)^+]^{2q-1}}{(2q-1)!},$$

де $p_0(x)\in\pi_{q-1}$. Умова 2) з означення сплайна порядку q виконуватиметься тільки тоді, коли

$$\sum_{i=1}^n d_i\frac{(x-x_i)^{2q-1}}{(2q-1)!}\equiv 0.$$

Прирівнюючи тут коефіцієнти при степенях x, дістаємо

$$\sum_{i=1}^n d_i(x_i)^k=0,\quad k=\overline{0, q-1}. \tag{1}$$

Таким чином, $s(x)\in S$ тоді і лише тоді, коли

$$s(x)=\sum_{j=0}^{q-1}\alpha_j x^j+\sum_{i=1}^n d_i\frac{[(x-x_i)^+]^{2q-1}}{(2q-1)!}, \tag{2}$$

причому d_i задовольняють умову (1).

Лема 1. *Для будь-якої функції $f(x)\in\tilde W_2^q[a, b]$ і для будь-якого сплайну $s(x)\in S$ справедливе співвідношення*

$$\int_a^b s^{(q)}(x)f^{(q)}(x)\,dx=(-1)^q\sum_{i=1}^n[s^{(2q-1)}(x_i+0)-$$
$$-s^{(2q-1)}(x_i-0)]f(x_i)\equiv(-1)^q\sum_{i=1}^n d_i f_i. \tag{2'}$$

Доведення. Інтегруючи зліва в (2') частинами і використовуючи умову (2) означення 1, маємо

$$\int_a^b s^{(q)}(x)f^{(q)}(x)\,dx=[s^{(q)}f^{(q-1)}(x)]_{x=a}^{x=b}-\int_a^b s^{(q+1)}f^{(q-1)}dx=$$
$$=\int_a^b s^{(q+1)}f^{(q-1)}dx=\cdots=(-1)^{q-1}\int_a^b s^{(2q-1)}f'\,dx=$$
$$=(-1)^{q-1}\sum_{i=1}^{n-1}\int_{x_i}^{x_{i+1}}s^{(2q-1)}(x)f'(x)\,dx=$$
$$=(-1)^{q-1}\sum_{i=1}^{n-1}\sum_{j=1}^{i}d_j[f(x_{i+1})-f(x_i)].$$

Застосовуючи тепер першу різницеву формулу Гріна (див. п. 1.4.3)

$$(u, (av_{\bar{x}})_x) = \sum_{i=1}^{N-1} hu_i (av_{\bar{x}})_{x,i} = -(u_{\bar{x}}, av_{\bar{x}}] + a_N u_N v_{\bar{x},N} - a_1 u_0 v_{\bar{x},1},$$

дістаємо (2'), що і треба було довести.

Означення 2. Нехай у точках x_i, $i = \overline{0, n}$, відомі значення деякої функції $f(x)$. Сплайн-функція $s(x) \in S$ порядку q називається *інтерполяційною*, якщо $s(x_i) = f(x_i)$.

Можна ввести фундаментальні інтерполяційні сплайн-функції $s_j(x)$ такі, що $s_j(x) \in S$ і $s_j(x_i) = \delta_{ij}$, $i, j = \overline{1, n}$. Тоді інтерполяційну сплайн-функцію можна записати у вигляді

$$s(x) = \sum_{i=1}^{n} f_i s_i(x).$$

Приклад. Фундаментальними сплайн-функціями першого порядку на відрізку [0, 1] є функції

$$s_j(t) = \begin{cases} 0, & t \in [0, x_{j-1}], \\ (t - x_{j-1})/(x_j - x_{j-1}), & t \in [x_{j-1}, x_j], \\ (t - x_{j+1})/(x_j - x_{j+1}), & t \in [x_j, x_{j+1}], \\ 0, & t \in [x_{j+1}, 1]. \end{cases}$$

Графік функції $s_j(t)$ показано на рис. 20.

Теорема 1. *Для довільних чисел f_i, $i = \overline{1, n}$, існує єдина функція s така, що $s(x_i) = f_i$, $i = \overline{1, n}$, тобто $s(x)$ — інтерполяційна сплайн-функція.*

Доведення. Візьмемо $s(x)$ у вигляді (2), де параметри d_i мають задовольняти систему (1). За умови, що $s(x_i) = f_i$ для визначення $n + q$ параметрів α_j, $j = \overline{0, q-1}$, і d_i, $i = \overline{1, n}$, маємо систему рівнянь

$$s(x_i) = f_i, \quad i = \overline{1, n},$$

$$\sum_{j=1}^{n} d_j (x_j)^k = 0, \quad k = \overline{0, q-1}.$$

Покажемо, що відповідна однорідна система має лише тривіальний розв'язок.

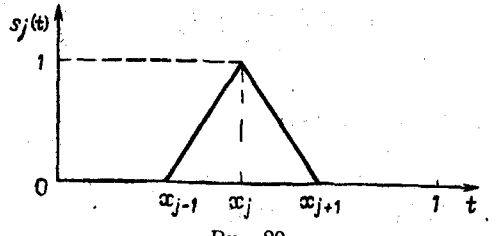


Рис. 20

Дійсно, нехай функція $\tilde{s}(x) \not\equiv 0$ задовольняє однорідну систему зазначеного виду. Тоді, поклавши в лемі 1 $f = \tilde{s}$, маємо

$$\int_a^b [\tilde{s}^{(q)}(x)]^2 dx = 0,$$

тобто $\tilde{s}^{(q)}(x) \equiv 0$. Це означає, що $\tilde{s}(x) \in \pi_{q-1}$ і $\tilde{s}(x_i) = 0$, $i = \overline{1, n}$, а оскільки $q \leqslant n$, то це можливо лише при $\tilde{s}(x) \equiv 0$. Суперечність доводить теорему.

Теорема 2. *Нехай $s(x) \in S$, $s(x_i) = f_i$, $i = \overline{1, n}$, $I_f = \{u \in \tilde{W}_2^q : u(x_i) = f_i, \ i = \overline{1, n}\}$. Тоді*

1) $\int_a^b [s^{(q)}(x) - f^{(q)}(x)]^2 dx = \min_{\tilde{s} \in S} \int_a^b [\tilde{s}^{(q)}(x) - f^{(q)}(x)]^2 dx \ \forall f \in I_f$

і всяка інша функція $s(x) \in S$, яка має цю властивість, відрізняється від S на многочлен $(q-1)$-го степеня

2) $\int_a^b [s^{(q)}(x) - \tilde{s}^{(q)}(x)]^2 dx = \min_{u \in I_f} \int_a^d [u^{(q)}(x) - \tilde{s}^{(q)}(x)]^2 dx \ \ \forall \tilde{s} \in S$

і $s(x)$ — єдина функція з I_f, яка має цю властивість.

Доведення. Для будь-якого $\tilde{s}(x) \in S$ маємо

$$\int_a^b [\tilde{s}^{(q)}(x) - f^{(q)}(x)]^2 dx = \int_a^b [s^{(q)}(x) - f^{(q)}(x) + \tilde{s}^{(q)}(x) - s^{(q)}(x)]^2 dx =$$

$$= \int_a^b [s^{(q)}(x) - f^{(q)}(x)]^2 dx + \int_a^b [\tilde{s}^{(q)}(x) - s^{(q)}(x)]^2 dx +$$

$$+ 2 \int_a^b [s^{(q)}(x) - f^{(q)}(x)] [\tilde{s}^{(q)}(x) - s^{(q)}(x)] dx.$$

Останній інтеграл дорівнює нулю в силу леми 1, якщо покласти $\tilde{s}(x) - s(x)$ замість s і $s(x) - f(x)$ замість f і врахувати, що $s(x_i) = f(x_i)$, $i = \overline{1, n}$. Отже,

$$\int_a^b [\tilde{s}^{(q)}(x) - f^{(q)}(x)]^2 dx = \int_a^b [s^{(q)}(x) - f^{(q)}(x)]^2 dx + \int_a^b [\tilde{s}^{(q)}(x) - s^{(q)}(x)]^2 dx,$$

що доводить першу частину теореми. Якщо властивість 1) має ще функція $s^*(x)$, то, поклавши в попередній рівності $\tilde{s} = s^*$, дістанемо

$$\int_a^b [s^{*(q)}(x) - s^{(q)}(x)]^2 dx = 0,$$

звідки $s^*(x) = s(x) + p(x)$, $p \in \pi_{q-1}$.

Властивість 2) доводиться аналогічно. Теорему доведено.

Теорема 3. *Нехай виконано умови теореми 2. Тоді інтерполяційна сплайн-функція $s(x)$ має таку екстремальну властивість:*

$$\int_a^b [s^{(q)}(x)]^2 dx = \min_{u \in I_f} \int_a^b [u^{(q)}(x)]^2 dx$$

і $s(x)$ єдина функція з I_f, яка має цю властивість.

Доведення. Для доведення достатньо в пункті 2) попередньої теореми покласти $\tilde{s}(x) = 0$.

Зауваження. Покажемо, що між сплайн-функціями порядку 2 і кубічними сплайнами з 4.1 існує простий зв'язок.

Дійсно, розглянемо сплайн-функцію порядку 2 на відрізку $[x_1, x_n]$. Тоді з означення 1 випливає, що на цьому відрізку маємо к бічний сплайн, причому $\tilde{s}_3''(x_1) = \tilde{s}_3''(x_n) = 0$. Зазначимо, що умова 2) означення 1 гарантує єдиність інтерполяційного сплайну.

Означення 3. Функція $\sigma(x) \in S$ називається *згладжуючою сплайн-функцією порядку q з ваговим коефіцієнтом* $\rho > 0$, якщо вона задовольняє умову

$$\sigma(x_i) + \frac{(-1)^q}{\rho}[\sigma^{(2q-1)}(x_i + 0) - \sigma^{(2q-1)}(x_i - 0)] = f_i, \quad i = \overline{1, n}, \tag{3}$$

де $f_i = f(x_i)$, $f \in \tilde{W}_2^q[a, b]$.

Теорема 4. *Згладжуюча сплайн-функція $\sigma(x)$ існує і єдина.*

Доведення. Оскільки $\sigma(x) \in S$, то функція $\sigma(x)$ має вигляд (2), причому виконується (1). Для визначення $n + q$ параметрів α_j, $j = \overline{0, q-1}$, і d_i, $i = \overline{1, n}$, маємо систему рівнянь

$$\begin{gathered}\sum_{i=1}^{n} d_i (x_i)^k = 0, \quad k = \overline{0, q-1}, \\ \sigma(x_i) + \frac{(-1)^q}{\rho} d_i = f_i, \quad i = \overline{1, n}.\end{gathered} \tag{4}$$

Припустивши, що існує не єдина згладжуюча сплайн-функція, також припускаємо, що однорідна система (4) має нетривіальний розв'язок.

За цим розв'язком за допомогою (2) визначається деякий сплайн $\tilde{\sigma}(x)$ такий, що

$$\tilde{\sigma}(x_i) + \frac{(-1)^q}{\rho} d_i = 0. \tag{5}$$

З цього співвідношення маємо

$$(-1)^q \sum_{i=1}^{n} d_i \tilde{\sigma}(x_i) + \frac{1}{\rho} \sum_{i=1}^{n} d_i^2 = 0. \tag{6}$$

Покладаючи в (2′) $s(x) = f(x) = \tilde{\sigma}(x)$, маємо

$$\int_a^b [\tilde{\sigma}^{(q)}]^2 dx = (-1)^q \sum_{i=1}^{n} d_i \tilde{\sigma}(x_i),$$

або з урахуванням (6)

$$\int_a^b [\tilde{\sigma}^{(q)}]^2 dx + \frac{1}{\rho} \sum_{i=1}^{n} d_i^2 = 0.$$

Оскільки відповідно до (5) маємо $d_i = \rho(-1)^{q-1} \tilde{\sigma}(x_i)$, то остання рівність переписується у вигляді

$$\int_a^b [\tilde{\sigma}^{(q)}(x)]^2 dx + \rho \sum_{i=1}^{n} [\tilde{\sigma}(x_i)]^2 = 0,$$

звідки випливає, що $\tilde{\sigma}^{(q)}(x) \equiv 0$, $\tilde{\sigma}(x) \in \pi_{q-1}$, $\tilde{\sigma}(x_i) = 0$, $i = \overline{1, n}$. Оскільки $n \geqslant q$, то звідси $\tilde{\sigma}(x) \equiv 0$, а це суперечить зробленому вище припущенню, що доводить існування єдиного розв'язку системи (4). Таким чином, існує єдина згладжуюча сплайн-функція $\sigma(x)$.

Якщо ввести фундаментальні згладжуючі сплайн-функції $\sigma_j(x) \in S$, які задовольняють умови

$$\sigma_j(x_i) + \frac{(-1)^q}{\rho}[\sigma_j^{(2q-1)}(x_i + 0) - \sigma_j^{(2q-1)}(x_i - 0)] = \delta_{ij}, \tag{7}$$

тоді, очевидно, згладжуюча сплайн-функція $\sigma(x)$ запишеться у вигляді

$$\sigma(x) = \sum_{i=1}^{n} f_i \sigma_i(x). \tag{8}$$

Справедливе таке твердження.

Теорема 5. *Нехай $\sigma(x)$ — єдина сплайн-функція, яка задовольняє умови* (3). *Тоді, якщо позначити*

$$\begin{aligned} R(\sigma(x) - u(x)) &= \int_a^b [\sigma^{(q)}(x) - f^{(q)}(x)]^2 dx + \\ &+ \rho \sum_{i=1}^{n} \left\{ \frac{(-1)^q}{\rho}[\sigma^{(2q-1)}(x_i + 0) - \sigma^{(2q-1)}(x_i - 0)] + u(x_i) - f_i \right\}^2 = \\ &= \int_a^b [\sigma^{(q)}(x) - u^{(q)}(x)]^2 dx + \rho \sum_{i=0}^{n} [\sigma(x_i) - u(x_i)]^2, \end{aligned} \tag{9}$$

то:

1) $$R(\sigma(x) - f(x)) = \min_{\tilde{\sigma} \in S} R(\tilde{\sigma}(x) - f(x)) \; \forall f \in \tilde{W}_2^q \tag{10}$$

і всяка функція $\tilde{\sigma}(x) \in S$, яка має цю властивість, відрізняється від $\sigma(x)$ на многочлен $(q-1)$-го степеня;

2) $$R(\tilde{\sigma}(x) - \sigma(x)) = \min_{u \in \tilde{W}_2^q} R(\tilde{\sigma}(x) - u(x)), \quad \forall \tilde{\sigma}(x) \in S, \tag{11}$$

і $\sigma(x)$ — єдина функція з $\tilde{W}_2^q$, яка має цю властивість.

Доведення. Маємо

$$R(\tilde{\sigma}(x)-f(x))=R(\sigma(x)-f(x)+\tilde{\sigma}(x)-\sigma(x))=R(\sigma(x)-f(x))+$$

$$+R(\tilde{\sigma}(x)-\sigma(x))+2\int_a^b[\sigma^{(q)}(x)-f^{(q)}(x)][\tilde{\sigma}^{(q)}(x)-\sigma^{(q)}(x)]\,dx+$$

$$+2\rho\sum_{i=1}^{n}\left\{[\sigma^{(2q-1)}(x_i+0)-\sigma^{(2q-1)}(x_i-0)]\frac{(-1)^q}{\rho}+f_i-\tilde{f}_i\right\}\times$$

$$\times\left\{\frac{(-1)^q}{\rho}[\tilde{\sigma}^{(2q-1)}(x_i+0)-\tilde{\sigma}^{(2q-1)}(x_i-0)]+\sigma(x_i)-\tilde{f}_i\right\}. \quad (12)$$

Оскільки $\sigma\in S$, то для неї справедливе зображення (2), причому

$$d_i=\sigma^{(2q-1)}(x_i+0)-\sigma^{(2q-1)}(x_i-0). \quad (13)$$

З іншого боку, оскільки $\sigma(x)$ — згладжуюча сплайн-функція, то в силу (4)

$$\sigma(x_i)-\tilde{f}_i=-\frac{(-1)^q}{\rho}d_i. \quad (14)$$

Тому з (12) маємо

$$R(\tilde{\sigma}(x)-f(x))=R(\sigma(x)-f(x))+R(\tilde{\sigma}(x)-\sigma(x))+$$

$$+2\int_a^b[\sigma^{(q)}(x)-f^{(q)}(x)][\tilde{\sigma}^{(q)}(x)-\sigma^{(q)}(x)]\,dx+$$

$$+2\frac{1}{\rho}\sum_{i=1}^{n}[\tilde{\sigma}^{(2q-1)}(x_i+0)-\tilde{\sigma}^{(2q-1)}(x_i-0)-d_i]\,d_i. \quad (15)$$

Замінивши в лемі 1 $f(x)$ на $\sigma(x)-f(x)$ і $s(x)$ на $\tilde{\sigma}(x)-\sigma(x)$, дістанемо за допомогою (13), (14)

$$\int_a^b[\tilde{\sigma}^{(q)}(x)-\sigma^{(q)}(x)][\sigma^{(q)}(x)-f^{(q)}(x)]\,dx=$$

$$=(-1)^q\sum_{i=1}^{n}[\tilde{\sigma}^{(2q-1)}(x_i+0)-\tilde{\sigma}^{(2q-1)}(x_i-0)-\sigma^{(2q-1)}(x_i+0)+$$

$$+\sigma^{(2q-1)}(x_i-0)][\sigma(x_i)-\tilde{f}(x_i)]=-(-1)^q\frac{(-1)^q}{\rho}\sum_{i=1}^{n}[\tilde{\sigma}^{(2q-1)}(x_i+0)-$$

$$-\tilde{\sigma}^{(2q-1)}(x_i-0)-d_i]\,d_i.$$

Тому з (15) випливає нерівність

$$R(\tilde{\sigma}(x)-f(x))=R(\sigma(x)-f(x))+R(\tilde{\sigma}(x)-\sigma(x))\geqslant$$
$$\geqslant R(\sigma(x)-f(x)), \quad (16)$$

а з неї — (10).

Нехай $\sigma^*(x)$ — інша сплайн-функція, яка має властивість (10), тоді в силу (16) $R(\sigma^*-\sigma)=R(\sigma^*-f)-R(\sigma-f)\equiv 0$. Враховуючи позначення (9), дістаємо $\sigma^*(x)=\sigma(x)+p(x)$, $p(x)\in\pi_{q-1}$, що завершує доведення 1). Властивість 2) доводиться аналогічно.

Як наслідок теореми 5 дістаємо таку теорему.

Теорема 6. *Нехай $\sigma(x)$ — єдина згладжуюча сплайн-функція для заданих $\rho>0$ і f_i, $i=\overline{1,n}$, тоді*

$$\int_a^b[\sigma^{(q)}(x)]^2\,dx+\rho\sum_{i=1}^{n}[s(x_i)-f_i]^2=$$

$$=\min_{u\in\tilde{W}_2^q}\left\{\int_a^b[u^{(q)}(x)]^2\,dx+\rho\sum_{i=1}^{n}[u(x_i)-f_i]^2\right\}$$

і $\sigma(x)$ — єдина функція з $\tilde{W}_2^q$, яка має цю властивість.

Для доведення достатньо в (11) покласти $\tilde{\sigma}(x)=0$.

4.4. Оцінка похибки при інтерполюванні функції кубічними сплайнами

Нехай функція $f(x)$, яка інтерполюється, належить до класу $W_2^k(\Omega)$ і x — фіксована точка проміжку $[a,b]$. Нехай $x\in[x_{i-1},x_i]$ для деякого фіксованого i, $g(x)$ — кубічна інтерполяційна функція. Розглянемо різницю

$$R(x)\equiv R(x;f)=f(x)-g(x)=$$

$$=f(x)-m_{i-1}\frac{(x_i-x)^3}{6h}-m_i\frac{(x-x_{i-1})^3}{6h}-\left(f_{i-1}-\frac{m_{i-1}h^2}{6}\right)\frac{x_i-x}{h}-$$

$$-\left(f_i-\frac{m_ih^2}{6}\right)\frac{x-x_{i-1}}{h}=R_1(x;f)+R_2(x;f), \quad (1)$$

де для спрощення прийнято $h_i\equiv h$ і введено позначення

$$R_1(x;f)=f(x)-f_{\bar{x}x}(x_{i-1})\frac{(x_i-x)^3}{6h}-f_{\bar{x}x}(x_i)\frac{(x-x_{i-1})^3}{6h}-$$

$$-\left[f(x_{i-1})-\frac{h^2}{6}f_{\bar{x}x}(x_{i-1})\right]\frac{x_i-x}{h}-\left[f(x_i)-\frac{h^2}{6}f_{\bar{x}x}(x_i)\right]\frac{x-x_{i-1}}{h}, \quad (2)$$

$$R_2(x;f)=-v_{i-1}\frac{(x_i-x)^3}{6h}-v_i\frac{(x-x_{i-1})^3}{6h}+$$

$$+\frac{h}{6}v_{i-1}\cdot(x_i-x)+\frac{h}{6}v_i\cdot(x-x_{i-1}), \quad (3)$$

$$v_i=m_i-f_{\bar{x}x}(x_i),\quad i=\overline{1,n-1},\quad f_{\bar{x}x}(x_0)=f_{\bar{x}x}(x_n)=0.$$

Зауважимо, що згідно з п. 4.1.11 v_i, $i = \overline{0, n}$, задовольняють наступну систему лінійних алгебраїчних рівнянь

$$\frac{h}{6}[v_{i-1} + 4v_i + v_{i+1}] = -\frac{h^3}{6} f_{\bar{x}x\bar{x}x}(x) \equiv hl(x; f), \quad i = \overline{1, n-1}, \quad (4)$$

$$v_0 = v_n = 0.$$

Дослідження лінійних функціоналів $R_1(x; f)$, $l(x; f)$ будемо проводити за відомою з попереднього викладу схемою: 1) перейдемо за допомогою лінійної заміни до проміжку, довжина якого не залежить від h і за допомогою теорем вкладення покажемо обмеженість функціоналів $R_1(x; f)$, $l(x; f)$ в W_2^k; 2) покажемо, що R_1 та l перетворюються на нуль у многочленах відповідного степеня і застосуємо лему Брембла — Гільберта; 3) виконавши обернену заміну, перейдемо до старої незалежної змінної і знайдемо шукану оцінку.

Отже, за допомогою заміни $s = (x - x_{i-1})/h$ відобразимо відрізок $[x_{i-2}, x_{i+1}]$ на $[-1, 2]$ і позначимо $f(x_{i-1} + sh) = \tilde{f}(s)$. Тоді

$$|R_1(x; f) \equiv |R_1(\tilde{f})| = \left|\tilde{f}(s) - \frac{1}{6}[\tilde{f}(-1) - 2\tilde{f}(0) + \tilde{f}(1)](1-s)^3 - \right.$$

$$- \frac{1}{6}[\tilde{f}(0) - 2\tilde{f}(1) + \tilde{f}(2)]s^3 - \left\{\tilde{f}(0) - \frac{1}{6}[\tilde{f}(-1) - 2\tilde{f}(0) + \tilde{f}(1)]\right\} \times$$

$$\times (1-s) - \left\{\tilde{f}(1) - \frac{1}{6}[\tilde{f}(0) - 2\tilde{f}(1) + \tilde{f}(2)]\right\} s \Big| \leqslant \tilde{M} \|\tilde{f}\|_{C[-1,2]} \leqslant$$

$$\leqslant M \|\tilde{f}\|_{W_2^k(-1,2)},$$

що означає обмеженість лінійного функціоналу $R_1(\tilde{f})$ в просторі $W_2^k(-1, 2)$, $k = \overline{1, 4}$. Крім того,

$$R_1(\tilde{f}) = 0, \quad \forall \tilde{f} \in \pi_3.$$

Тому з леми Брембла — Гільберта випливає оцінка

$$|R_1(\tilde{f})| \leqslant M\tilde{M} |\tilde{f}|_{W_2^k(-1,2)}, \quad k = \overline{1, 4},$$

яка за допомогою заміни s на стару змінну x приводить до нерівності

$$|R_1(x; f)| \leqslant M_1 h^{k-1/2} |f|_{W_2^k((x_{i-2}, x_{i+1}) \cap (a,b))}, \quad (5)$$

$$k = \overline{1, 4}, \quad x \in [x_{i-1}, x_i], \quad i = \overline{1, n}.$$

Аналогічно знаходимо оцінку для функціоналу $l(x; f)$:

$$|l(x; f)| \leqslant M_1 h^{k-5/2} |f|_{W_2^k((x_{i-2}, x_{i+2}) \cap (a,b))}, \quad (6)$$

$$k = \overline{1, 4}, \quad x \in \omega_h.$$

При відшуканні оцінки в $L_2(a, b)$-нормі $R(x; f)$ зрозуміло, що буде потрібна оцінка в цій самій нормі як $R_1(x; f)$, так і $R_2(x; f)$. Якщо наявність нерівності (5) одразу дає змогу записати, що

$$\|R_1(x; f)\|_{L_2(a,b)} \leqslant M_2 h^k |f|_{W_2^k(a,b)}, \quad k = \overline{1, 4}, \quad (7)$$

то для оцінки

$$\|R_2(x; f)\|_{L_2(a,b)} \leqslant \frac{2}{3} h^2 \|v\|_{L_2(\omega_h)} \quad (8)$$

потрібно оцінити норму $\|v\|_{L_2(\omega_h)}$.

З системи (4) маємо

$$v = hA^{-1} l(f),$$

де $v = (v_1, \ldots, v_{n-1})^T$, $l(f) = (l(x_1; f), \ldots, l(x_{n-1}; f))^T$, що приводить з врахуванням леми 1 (п. 4.1) до оцінки

$$\|v\|_{L_2(\omega_h)} \leqslant \|A^{-1}\|_3 h \|l(f)\|_{L_2(\omega_h)} \leqslant \frac{h}{\delta} \|l(f)\|_{L_2(\omega_h)}, \quad (9)$$

$$\|A^{-1}\|_3 = \sup_{y \neq 0} (A^{-1}y, y)/(y, y) = \sup_z (z, z)/(Az, z).$$

Причому сталу δ у випадку рівномірної сітки ω_h легко визначити. Дійсно,

$$(Az, z) = \frac{h}{6} \sum_{i=1}^{n-1} (z_{i-1} + 4z_i + z_{i+1}) z_i =$$

$$= \frac{h}{3} \sum_{i=1}^{n-2} z_i z_{i+1} + \frac{2}{3} h \sum_{i=1}^{n-1} z_i^2 \geqslant -\frac{h}{6} \sum_{i=1}^{n-2} (z_i^2 + z_{i+1}^2) +$$

$$+ \frac{2}{3} h \sum_{i=1}^{n-1} z_i^2 \geqslant \frac{h}{3} \sum_{i=1}^{n-1} z_i^2,$$

тобто $\delta = h/3$ і нерівність (9) з використанням (6) набуває вигляду

$$\|v\|_{L_2(\omega_h)} \leqslant 3 \|l(f)\|_{L_2(\omega_h)} \leqslant \tilde{M}_1 h^{k-2} |f|_{W_2^k(a,b)}, \quad k = \overline{1, 4}. \quad (10)$$

Нерівності (8) і (10) разом з

$$\|f(x) - g(x)\|_{L_2(a,b)} \leqslant \|R_1(x; f)\|_{L_2(a,b)} + \|R_2(x; f)\|_{L_2(a,b)}$$

дають остаточний результат

$$\|f(x) - g(x)\|_{L_2(a,b)} \leqslant M h^k |f|_{W_2^k(a,b)}, \quad k = \overline{1, 4}. \quad (11)$$

Можна дістати оцінку типу (11) і у рівномірній метриці. Для цього треба мати оцінки для $\|R_1(x; f)\|_{C[a,b]}$, $\|R_2(x; f)\|_{C[a,b]}$. З (2), (3),

(5) випливає, що

$$\|R_1(x;\ f)\|_{C[x_{i-1},x_i]} \leqslant M_1 h^{k-1/2} |f|_{W_2^k((x_{i-2},x_{i+1})\cap(a,b))}, \qquad (12)$$

$$i = \overline{1,\ n},\quad k = \overline{1,\ 4},$$

$$\|R_2(x;\ f)\|_{C[x_{i-1},x_i]} \leqslant \frac{2}{3} h^2 \|v\|_\infty,\quad i = \overline{1,\ n}. \qquad (13)$$

Скориставшись лемою 1 (п. 4.1) для розв'язку системи (4), дістанемо оцінку

$$\|v\|_\infty \leqslant h\|A^{-1}\|_\infty \|l(f)\|_\infty \leqslant \frac{h}{q}\|l(f)\|_\infty \leqslant$$

$$\leqslant 3M_1 h^{k-5/2} \max_i |f|_{W_2^k((x_{i-2},x_{i+2})\cap(a,b))},\quad k = \overline{1,\ 4}. \qquad (14)$$

З виразів (12) — (14) остаточно дістаємо

$$\|f(x) - g(x)\|_{C[a,b]} \leqslant Mh^{k-1/2} |f|_{W_2^k(a,b)},\quad k = \overline{1,\ 4}. \qquad (15)$$

Якщо $f(x) \in C^k[a,\ b]$, тоді вирази (12) — (14) приводять до більш точної оцінки

$$\|f(x) - g(x)\|_{C[a,b]} \leqslant Mh^k |f|_{C^k[a,b]},\quad k = \overline{1,\ 4}. \qquad (16)$$

Вправа 1. За допомогою вище викладеної методики показати, що для різниці похідних $f^{(j)}(x) - g^{(i)}(x)$, $j = \overline{1,\ 2}$, справедливі оцінки

$$\|f^{(j)} - g^{(j)}\|_{C[a,b]} \leqslant Mh^{k-1/2-j} |f|_{W_2^k(a,b)},\quad 1 \leqslant k \leqslant 4,\quad j = 1,\ 2;$$

$$\|f^{(j)} - g^{(j)}\|_{L_2(a,b)} \leqslant Mh^{k-j} \|f\|_{W_2^k(a,b)},\quad 1 \leqslant k \leqslant 4,\quad j = 1,\ 2;$$

якщо

$$f(x) \in W_2^k(a,\ b),\quad 1 \leqslant k \leqslant 4,$$

і

$$\|f^{(j)}(x) - g^{(j)}(x)\|_{C[a,b]} \leqslant Mh^{k-j} \|f^{(k)}\|_{C[a,b]},\quad 1 \leqslant k \leqslant 4,\quad j = 1,\ 2,$$

за умови $f \in C^k[a,\ b]$, де сталі M не залежать від f і від h,

$$h = \max_{i=\overline{1,n}} h_i.$$

Таким чином, справедливе таке твердження.

Теорема. *Нехай* $f(x) \in W_2^k(a,\ b)$, $1 \leqslant k \leqslant 4$ *і виконана умова* $h_i \equiv h$, *тоді мають місце оцінки*

$$\|f^{(j)}(x) - g^{(j)}(x)\|_{C[a,b]} \leqslant Mh^{k-1/2-j} |f|_{W_2^k(a,b)};$$

$$\|f^{(j)}(x) - g^{(j)}(x)\|_{L_2(a,b)} \leqslant Mh^{k-j} |f|_{W_2^k(a,b)},$$

де $j = 0,\ 1,\ 2$, M — *сталі що не залежать від* h і f. *Якщо* $f(x) \in C^k[a,\ b]$, $1 \leqslant k \leqslant 4$, *то*

$$\|f^{(j)}(x) - g^{(j)}(x)\|_{C[a\ b]} \leqslant Mh^{k-j} \|f^{(k)}\|_{C[a,b]},$$

де $j = 0,\ 1,\ 2$.

З теореми, зокрема, випливає, що за умови

$$f(x) \in W_2^k(a,\ b,)\quad 1 \leqslant k \leqslant 4,\quad 0 \leqslant j \leqslant k,$$

функція $g^{(j)}(x)$ збігається при $h = \max h_i \to 0$ до $f^{(j)}(x)$ зі швидкістю $O(h^{k-1/2-j})$ в нормі $C[a,\ b]$ і зі швидкістю $O(h^{k-j})$ в нормі $L_2(a,\ b)$. Якщо $f(x) \in C^k[a,\ b]$, $1 \leqslant k \leqslant 4$, $0 \leqslant j < k$, то $g^{(j)}(x)$ збігається при $h \to 0$ до $f^{(j)}(x)$ зі швидкістю $O(h^{k-j})$.

ГЛАВА 5

НАБЛИЖЕНЕ ОБЧИСЛЕННЯ ВИЗНАЧЕНИХ ІНТЕГРАЛІВ

5.1. Класифікація формул чисельного інтегрування

Перш ніж перейти до побудови та вивчення методів наближеного обчислення визначених інтегралів, запропонуємо можливу класифікацію їх (рис. 21).

Тут мова йтиме про методи обчислення визначених інтегралів без особливостей, тобто інтегралів зі скінченними границями інтегрування від функцій, які на проміжку інтегрування не перетворюються на нескінченність.

Визначений інтеграл можна розглядати як функціонал

$$J(f) = \int_a^b \rho(x) f(x)\, dx, \qquad (1)$$

заданий на деякій множині $f \in D(J)$, $\rho(x) \geqslant 0$ — задана вагова функція.

Однією із загальних ідей при побудові алгоритмів наближеного обчислення інтеграла (1) є така: для функції $f(x)$ будується деяке наближення $f^h(x)$ і наближено покладають

$$J(f) \approx J^h(f) = \int_a^b \rho(x) f^h(x)\, dx \qquad (2)$$

(звичайно, наближення $f^h(x)$ має бути таким, щоб інтеграл (2) обчислювався простіше, ніж (1)). На практиці поширені наближення $f^h(x)$, які лінійно виражаються через значення функції $f(x)$ та її похідних

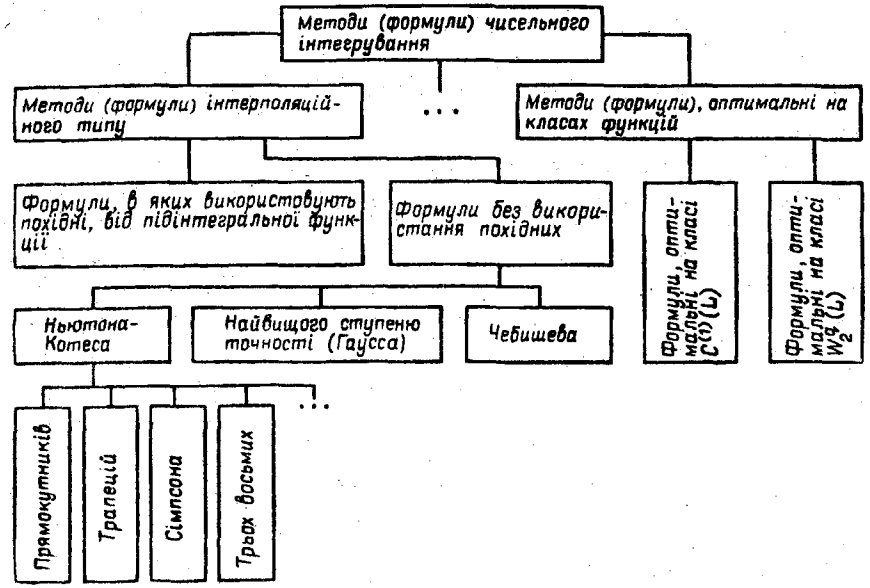


Рис. 21

в точках сітки $\omega = \{x_i \in [a, b]: i = \overline{1, n},\ x_i < x_{i+1}\}$, відповідна формула для $J^h(f)$ має вигляд

$$J^h(f) = \sum_{k=1}^{n} \sum_{i=1}^{\alpha_k} A_{k,i}^{(n)} f^{(i)}(x_k), \quad \sum_{k=1}^{n} \alpha_k = n \tag{3}$$

і називається *квадратурною формулою*. Числа x_k, $k = \overline{1, n}$, називаються *вузлами* або *абсцисами квадратурної формули*, а числа $A_{k,i}^{(n)}$ — *коефіцієнтами* або *ваговими коефіцієнтами*. Величина

$$R^h(f) = J(f) - J^h(f) \tag{4}$$

називається *залишковим членом квадратурної формули*. На практиці найчастіше вживаються квадратурні формули виду:

$$J^h(f) = \sum_{i=1}^{n} c_i^{(n)} f(x_i), \tag{5}$$

в яких використовуються лише значення функції $f(x)$ і не використовуються похідні. Залишковий член таких формул, очевидно, є лінійним функціоналом від f.

Формулу (5) можна дістати, якщо вибрати f^h у вигляді

$$f^h(x) \equiv p_{n-1}(x;\ f) = \sum_{k=1}^{n} f(x_k)\, l_{k,n-1}(x), \tag{6}$$

де $p_{n-1}(x;\ f)$ — інтерполяційний многочлен степеня $n-1$ для функції $f(x)$ за вузлами x_i, $i = \overline{1, n}$; $l_{k,n-1}$ фундаментальні інтерполяційні многочлени

$$l_{k,n-1} = \frac{\omega_n(x)}{(x - x_k)\,\omega_n'(x_k)} =$$

$$= \frac{(x - x_1) \ldots (x - x_{k-1})(x - x_{k+1}) \ldots (x - x_n)}{(x_k - x_1) \ldots (x_k - x_{k-1})(x_k - x_{k+1}) \ldots (x_k - x_n)}, \tag{7}$$

$$\omega_n(x) = (x - x_1) \ldots (x - x_n).$$

У цьому випадку

$$c_i^{(n)} = \int_a^b \rho(x)\, l_{i,n-1}(x)\, dx = \int_a^b \rho(x) \frac{\omega_n(x)}{(x - x_i)\,\omega_n'(x_i)}\, dx \tag{8}$$

формула (5) називається *формулою інтерполяційного типу*.

Означення. Якщо залишковий член квадратурної формули дорівнює нулю на множині π_n всіх алгебраїчних многочленів не вище n-го степеня, то кажуть, що *квадратурна формула має алгебраїчний ступінь точності n*.

Очевидно, що алгебраїчний степінь точності формули інтерполяційного типу (5) з ваговими коефіцієнтами (8) є щонайменше $n-1$, бо $f^h(x) \equiv p_{n-1}(x;\ f) = f(x)$, якщо $f \in \pi_{n-1}$.

Найпростішими серед формул інтерполяційного типу є формули Ньютона — Котеса, які широко використовуються в практичних обчисленнях.

5.2. Формули Ньютона—Котеса

Якщо в квадратурній формулі (5) (п. 5.1) з ваговими коефіцієнтами (8) (п. 5.1) для інтеграла (1) (п. 5.1) з вагою $\rho(x) \equiv 1$ вузли рівновіддалені, тобто $x_{i+1} - x_i = h$, $i = \overline{1, n-1}$, то така формула називається *формулою Ньютона — Котеса*. У формулах Ньютона — Котеса крок h можна вибирати за формулами $h = \frac{b-a}{n+1}$, якщо $x_1 = a + h$, або $h = \frac{b-a}{n-1}$, якщо $x_1 = a$. У першому випадку множина вузлів не містить точки a і b і проміжок інтегрування розбивається на $n+1$ рівних частин; у цьому разі квадратурна формула називається *формулою відкритого типу*. В другому випадку кінці проміжку інтегрування a і b є теж вузлами квадратурної формули, яка називається *формулою замкненого типу*; відрізок $[a, b]$ розбивається вузлами на $n-1$ рівних частин. Якщо покласти $x_1 = a + kh$, $x_n = b - kh$, то для формул відкритого типу $k = 1$, а для формул замкненого типу $k = 0$.

Покладемо $x_0 = x_1 - h = a + (k-1)h$, $x_n = b - kh = x_0 + nh$ і виконаємо в інтегралі заміну $x = x_0 + hy = a + h(y + k - 1)$. Якщо позначити $f(x_0 + hy) = F(y)$, то

$$\int_a^b f(x)\,dx = h \int_{1-k}^{n+k} F(y)\,dy. \tag{1}$$

В останньому інтегралі замінимо функцію $F(y)$ інтерполяційним многочленом Лагранжа з вузлами в точках 1, 2, ..., тобто

$$F(y) = \sum_{i=1}^{n} F(i) \frac{(y-1)\ldots(y-i+1)(y-i-1)\ldots(y-n)}{(i-1)\ldots(i-i+1)(i-i-1)\ldots(i-n)} +$$
$$+ (y-1)(y-2)\ldots(y-n)F(y;\ 1;\ 2;\ \ldots;\ n) =$$
$$= \sum_{i=1}^{n} f(x_0 + ih)(-1)^{n-i} \frac{(y-1)\ldots(y-i+1)(y-i-1)\ldots(y-n)}{(i-1)!\,(n-i)!} +$$
$$+ (y-1)\ldots(y-n)F(y;\ 1;\ 2;\ \ldots;\ n). \tag{2}$$

В результаті дістаємо

$$\int_a^b f(x)\,dx = h \int_{1-k}^{n+k} F(y)\,dy = h \sum_{i=1}^{n} \tilde{J}_{i,k}^{(n)} F(i) + R_{n,k}(f), \tag{3}$$

де $R_{n,k}(f)$ — залишковий член,

$$\tilde{J}_{i,k}^{(n)} = \frac{(-1)^{n-i}}{(i-1)!\,(n-i)!} \int_{1-k}^{n+k} \frac{(y-1)(y-2)\ldots(y-n)}{y-i}\,dy, \tag{4}$$

$$R_{n,k}(f) = h \int_{1-k}^{n+k} (y-1)\ldots(y-n)F(y;\ 1;\ \ldots;\ n)\,dy. \tag{5}$$

Враховуючи заміну, перепишемо (3) у вигляді

$$\int_a^b f(x)\,dx = (b-a) \sum_{i=1}^{n} J_{i,k}^{(n)} f(x_0 + ih) + R_{n,k}(f), \tag{6}$$

де

$$J_{i,k}^{(n)} = \frac{h\tilde{J}_{i,k}^{(n)}}{b-a} = \frac{\tilde{J}_{i,k}^{(n)}}{n-1+2k}. \tag{7}$$

Порівнюючи (6) з (5.1.5), дістаємо

$$c_i^{(n)} = (b-a)\,J_{i,k}^{(n)}. \tag{8}$$

Значення $J_{i,k}^{(n)}$ не залежить від проміжку інтегрування і можуть бути обчислені раз і назавжди. При цьому справджується рівність

$$J_{i,k}^{(n)} = J_{n-i+1,k}^{(n)},$$

яка означає рівність рівновіддалених від кінців проміжку інтегрування коефіцієнтів і яка приблизно вдвічі скорочує кількість обчислень.

Дійсно,

$$J_{n-i+1}^{(n)} = \frac{(-1)^{i-1}}{(n-1+2k)\,(n-i)!\,(i-1)!} \int_{1-k}^{n+k} \frac{(y-1)\ldots(y-n)}{y-n+i-1}\,dy$$

і, замінюючи y на $n - z + 1$, дістаємо

$$J_{n-i+1,k}^{(n)} = \frac{(-1)^{i}}{(n-1+2k)\,(n-i)!\,(i-1)!} \times$$
$$\times \int_{n+k}^{1-k} \frac{(n-z)(n-z-1)\ldots(1-z)}{i-z}\,dz =$$
$$= \frac{(-1)^{n-i}}{(n-1+2k)\,(n-i)!\,(i-1)!} \int_{1-k}^{n+k} \frac{(z-1)(z-2)\ldots(z-n)}{z-i}\,dz = J_{i,k}^{(n)}.$$

Можна довести, що величини $|J_{i,0}^{(n)}|$ при зростанні n необмежено зростають і $\lim\limits_{n\to\infty} |J_{i,0}^{(n)}| = +\infty$. Це означає, що при великих n малі похибки в значеннях функції $f(x_0 + ih)$ можуть дати велику похибку в квадратурній сумі (йдеться про неусувну похибку). Тому на практиці не використовуються формули Ньютона — Котеса з великим n, а щоб зменшити похибку при чисельному інтегруванні на великому проміжку $[a, b]$, його попередньо розбивають на достатньо велику кількість малих інтервалів і на кожному з них застосовують квадратурну формулу з невеликою кількістю вузлів. Знайдена таким чином квадратурна формула на відрізку $[a, b]$ інколи називається *ускладненою* або *складеною*. Розглянемо детальніше формули Ньютона — Котеса відкритого типу при $n = 1$ і замкненого типу при $n = 2$ і $n = 3$. Поклавши в (3) $k = 1$, $n = 1$, запишемо рівність

$$\int_a^b f(x)\,dx = (b-a) f\left(\frac{a+b}{2}\right) + R_{1,1}(f), \tag{9}$$

яка називається *формулою середніх прямокутників*. Розбиваючи відрізок $[a, b]$ на N рівних частин довжини $h = \frac{b-a}{N}$, матимемо:

$$\int_a^b f(x)\,dx = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} f(x)\,dx,$$

де $x_i = a + ih$, $i = \overline{0,\ N-1}$, $x_N = b$. Застосувавши до кожного з інтегралів у сумі формулу (9), знайдемо *ускладнену формулу середніх*

прямокутників

$$\int_a^b f(x)\,dx = h\sum_{i=1}^{N} f\left(a+\frac{2i-1}{2}h\right) + R_1(f) \equiv I_h^{\text{пр}} + R_{\text{пр}}(f), \tag{10}$$

$$R_{\text{пр}}(f) \equiv R_1(f),$$

де

$$I_h^{\text{пр}} = h\sum_{i=1}^{N} f\left(a+\frac{2i-1}{2}h\right),$$

$R_1(f) = \sum_{i=1}^{N} R_{1,1}^{(i)}(f)$, $R_{1,1}^{(i)}(f)$ — залишковий член квадратурної формули на i-му інтервалі. Часто і формулу (9), і формулу (10) називають просто формулами середніх прямокутників.

Поклавши у виразі (3) $k = 0$, $n = 2$, дістанемо *формулу трапецій*

$$\int_a^b f(x)\,dx = \frac{b-a}{2}[f(a)+f(b)] + R_{2,0}(f). \tag{11}$$

За допомогою зображення

$$\int_a^b f(x)\,dx = \sum_{i=0}^{N}\int_{a+ih}^{a+(i+1)h} f(x)\,dx, \quad h = \frac{b-a}{N+1},$$

із формули (1) знаходимо *ускладнену формулу трапецій*

$$\int_a^b f(x)\,dx = h\left[\frac{1}{2}f(a) + \sum_{i=1}^{N} f(a+ih) + \frac{1}{2}f(b)\right] + R_2(f) \equiv I_h^{\text{тр}} + R_{\text{тр}}(f), \tag{12}$$

де

$$I_h^{\text{тр}} = h\left[\sum_{i=1}^{N} f(a+ih) + \frac{1}{2}f(a) + \frac{1}{2}f(b)\right],$$

$R_{\text{тр}}(f) \equiv R_2(f) = \sum_{i=1}^{N+1} R_{2,0}^{(i)}(f)$, $R_{2,0}^{(i)}(f)$ — залишковий член на i-му інтервалі.

Якщо покладемо в (3) $k = 0$, $n = 3$, то дістанемо *формулу Сімпсона*

$$\int_a^b f(x)\,dx = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] + R_{3,0}(f). \tag{13}$$

Поклавши $h = (b-a)/(2N)$, маємо

$$\int_a^b f(x)\,dx = \sum_{i=0}^{N-1}\int_{a+2ih}^{a+2(i+1)h} f(x)\,dx,$$

і використавши (13), знайдемо *ускладнену формулу Сімпсона*

$$\int_a^b f(x)\,dx = \frac{h}{3}\left[f(a) + 4\sum_{i=1}^{N} f(a+(2i-1)h) + 2\sum_{i=1}^{N-1} f(a+2ih) + f(b)\right] + R_3(f) \equiv I_h^C + R_C(f), \tag{14}$$

де

$$I_h^C = \frac{h}{3}\left[f(a) + 4\sum_{i=1}^{N} f(a+(2i-1)h) + 2\sum_{i=1}^{N-1} f(a+2ih) + f(b)\right], \quad R_C(f) \equiv R_3(f) = \sum_{i=0}^{N-1} R_{3,0}^{(i)}(f),$$

$R_{3,0}^{(i)}(f)$ — залишковий член на інтервалі $[a+2ih,\ a+2(i+1)h]$.

Розглянемо питання про алгебраїчні степені точності наведених формул.

Лема. *Нехай в квадратурній формулі* (5.1.5), *яка є точною на многочленах до* $(n-1)$*-го степеня, вузли* x_i, $i = \overline{1, n}$, *розміщено симетрично відносно середини відрізка* $[a, b]$. *Тоді, якщо* $n = 2m+1$, $m = 1, 2, \ldots$, *то алгебраїчний ступінь точності формули* (6) *дорівнюватиме* n, *тобто підвищиться на одиницю.*

Доведення. Побудуємо для функції $f(x)$ інтерполяційний многочлен n-го степеня за вузлами $x_1, x_2, \ldots, x_n, x_i$, тобто з одним кратним вузлом. Запишемо його в формі Ньютона:

$$p_n(x; f) = f(x_1) + (x-x_1)f[x_1; x_2] + \cdots + (x-x_1)\ldots(x-x_{n-1})f[x_1; \ldots; x_n] + (x-x_1)\ldots(x-x_n)\times f[x_1; x_2; \ldots; x_n; x_i] = p_{n-1}(x; f) + (x-x_1)\ldots(x-x_n)f[x_1; \ldots; x_n; x_i]. \tag{15}$$

Оскільки n — непарне, а всі вузли x_k, $k = \overline{1, n}$, розміщено симетрично відносно точки $\frac{a+b}{2}$, то

$$\int_a^b (x-x_1)\ldots(x-x_n)\,dx = 0$$

і з (15) дістанемо $\int_a^b p_n(x; f)\,dx = \int_a^b p_{n-1}(x; f)\,dx$. Але $R(p_{n-1}) = 0$, тому і $R(p_n) = 0$, що і треба було довести.

Наслідок. *Формула середніх прямокутників має алгебраїчний ступінь точності* 1, *а формула Сімпсона* — 3.

Формули (13), (14) інколи також називають *формулами парабол*. Назви квадратурних формул (9) — (14) походять з геометричних міркувань: криволінійні трапеції, сума площ яких є значення інтегралу, замінюються відповідно прямокутниками, трапеціями і криволінійними трапеціями, верхньою стороною яких є парабола.

При $k = 0$, $n = 4$ формула (3) має вигляд

$$\int_a^b f(x)\,dx = (b-a)\left[\frac{1}{8}f(a) + \frac{3}{8}f\left(a + \frac{b-a}{3}\right) + \right.$$

$$\left. + \frac{3}{8}f\left(a + \frac{2(b-a)}{3}\right) + \frac{1}{8}f(b)\right] + R_{4,0}(f) \qquad (16)$$

(правило трьох восьмих). При виведенні (16) ми замінили підінтегральну функцію інтерполяційним многочленом третього степеня і оскільки n-парне, тобто не виконуються умови леми 1, то її алгебраїчний ступінь точності є 3 і збігається з алгебраїчним ступенем точності формули Сімпсона (у формулі Сімпсона на один вузол менше).

5.3. Квадратурні формули найвищого алгебраїчного ступеня точності (формули Гаусса)

У попередньому параграфі виведено формулу чисельного інтегрування заміною підінтегральної функції її інтерполяційним многочленом, побудованим за n вузлами. Такі формули називаються *формулами інтерполяційного типу*. Ми бачили, що алгебраїчний ступінь точності таких формул у загальному випадку є $n - 1$, а при симетричному розміщенні непарної кількості вузлів відносно середини відрізка дорівнює n. В зв'язку з цим виникає питання: чи можна за рахунок якого-небудь іншого розміщення вузлів ще підвищити алгебраїчний ступінь точності? У цьому параграфі дається відповідь на поставлене питання. Зазначимо, що алгебраїчний ступінь точності є дещо умовною характеристикою точності, оскільки можна навести приклади підінтегральних функцій, коли формула меншого алгебраїчного ступеня точності дає більш точне значення інтеграла. Однак далі видно, що формули високого алгебраїчного ступеня точності мають не тільки теоретичний, але й практичний інтерес.

Нехай квадратурна формула

$$\int_a^b \rho(x) f(x)\,dx = \sum_{k=1}^{N} c_k^{(n)} f(x_k^{(n)}) + R(f), \qquad (1)$$

де $\rho(x) \geqslant 0\ \forall\, x \in [a, b]$ — вагова функція, що задовольняє нерівності

$$\left|\int_a^b \rho(x)\, x^i dx\right| < +\infty, \quad i = 0, 1, \ldots,$$

є формулою інтерполяційного типу, тобто $R(f) = 0$, якщо $f \in \pi_{n-1}$ (тобто є многочленом не вище $(n-1)$-го степеня) і

$$c_k^{(n)} = \int_a^b l_{k,n-1}(x)\,\rho(x)\,dx = \int_a^b \rho(x)\,\frac{\omega'(x)}{(x - x_k^{(n)})\,\omega'(x_k^{(n)})}\,dx, \qquad (2)$$

$$\omega(x) = \prod_{i=1}^{n} (x - x_i^{(n)}), \quad x_i^{(n)} \in [a, b].$$

В нашому розпорядженні знаходяться вузли $x_i^{(n)}$, $i = \overline{1, n}$, які можна вибрати так, щоб формула (1) мала найвищий можливий ступінь алгебраїчної точності. Неважко помітити, що найвищий ступінь алгебраїчної точності менший або дорівнює $2n - 1$. Дійсно, покладаючи в (1)

$$f(x) = \prod_{i=1}^{n} (x - x_i^{(n)})^2 \in \pi_{2n},$$

дістаємо

$$R(f) = \int_a^b \prod_{i=1}^{n} (x - x_i^{(n)})^2\,\rho(x)\,dx > 0,$$

тобто $R(f)$ не може бути рівним нулю на всіх многочленах степеня $2n$. Формулу вигляду (1), для якої $R(f) = 0\ \forall\, f \in \pi_{2n-1}$, називатимемо *квадратурною формулою найвищого алгебраїчного ступеня точності*. Покажемо, що такі формули існують, і знайдемо необхідні і достатні умови для цього. Нехай $\{p, (x)\}_{i=0}^{\infty}$ — система ортогональних з вагою $\rho(x)$ многочленів на $[a, b]$, $p_n(x) = \sum_{i=0}^{n} k_{i,n} x^{n-i}$. Тоді справедливе таке твердження.

Теорема 1. *Для того щоб квадратурна формула* (1) *з коефіцієнтами* (2) *була квадратурною формулою найвищого алгебраїчного ступеня точності, необхідно і достатньо, щоб* $\omega(x) = k_{0,n}^{-1} p_n(x)$ *тобто вузли* $x_i^{(n)}$, $i = 1, n$, *збігались з нулями многочленів* $p_n(x)$, *причому така квадратурна формула єдина.*

Доведення. *Необхідність.* Нехай формула (1) є квадратурною формулою найвищого алгебраїчного ступеня точності, тобто

$$R(f) = 0 \quad \forall\, f \in \pi_{2n-1},$$

де через π_{2n-1} позначено множину всіх многочленів до $(2n-1)$-го степеня включно. Тоді $\forall\, Q(x) \in \pi_m$, $m \leqslant n - 1$, маємо $Q(x)\,\omega(x) \in \pi_{2n-1}$ і з (1) випливає

$$\int_a^b Q(x)\,\omega(x)\,\rho(x)\,dx = \Sigma c_i^{(n)} Q(x_i^{(n)})\,\omega(x_i^{(n)}) + R(Q\omega) = 0, \quad m = \overline{0, n-1}, \qquad (3)$$

бо $\omega(x_i^{(n)}) = 0$ і $R(f) = 0 \; \forall f \in \pi_{2n-1}$. Це означає, що многочлен $\omega(x)$ степеня n ортогональний до всіх многочленів до $(n-1)$-го степеня включно. Раніше ми бачили (див. 1.6), що такий многочлен єдиний і $\omega(x) = k_{0,n}^{-1} p_n(x)$, що і треба було довести.

Достатність. Нехай $\omega(x) = k_{0,n}^{-1} p_n(x)$. Покажемо, що $R(f) = 0 \; \forall f \in \pi_{2n-1}$. Дійсно,

$$f(x) = \omega(x) Q(x) + r(x), \tag{4}$$

де многочлени $Q(x)$ і $r(x)$ мають степені менші ніж n. Тоді

$$\int_a^b f(x)\rho(x)\,dx = \int_a^b \omega(x) Q(x)\rho(x)\,dx + \int_a^b r(x)\rho(x)\,dx. \tag{5}$$

Перший доданок у правій частині за припущенням дорівнює нулю. Оскільки з (4) випливає, що

$$f(x_i^{(n)}) = r(x_i^{(n)}) \quad \text{і} \quad r \in \pi_m, \quad m \leqslant n-1,$$

то $r(x)$ збігається з інтерполяційним многочленом функції $f(x)$. Але формула (1) є формулою інтерполяційного типу, яка точна для многочленів степеня $n-1$, таким чином,

$$\int_a^b f(x)\rho(x)\,dx = \int_a^b r(x)\rho(x)\,dx = \sum_{i=1}^n c_i^{(n)} r(x_i^{(n)}) =$$

$$= \sum_{i=1}^n c_i^{(n)} f(x_i^{(n)}) \quad \forall f \in \pi_{2n-1},$$

тобто $R(f) = 0$, а це й треба було довести.

Єдиність квадратурної формули найвищого алгебраїчного ступеня точності випливає з єдиності (з точністю до мультиплікативних сталих) системи многочленів $\{p_i(x)\}_{i=0}^{\infty}$ ортогональних з вагою $\rho(x)$ на $[a, b]$. Теорему повністю доведено.

Квадратурна формула (1) з коефіцієнтами (2), яка має найвищий алгебраїчний ступінь точності, називається ще *формулою механічних квадратур Гаусса,* або просто *формулою Гаусса.* Вагові коефіцієнти $c_k^{(n)}$ у формулі Гаусса (1), (2) часто називають *коефіцієнтами Крістоффеля.* Їхні властивості встановлює наступна теорема.

Теорема 2. *Коефіцієнти Крістоффеля* $c_k^{(n)}$, $k = \overline{1, n}$, *мають такі властивості:*

$$c_k^{(n)} > 0, \quad k = \overline{1, n}; \tag{6}$$

$$\sum_{k=1}^n c_k^{(n)} = \int_a^b \rho(x)\,dx; \tag{7}$$

$$c_k^{(n)} = \int_a^b \left[\frac{p_n(x)}{p_n'(x_k^{(n)})\,(x - x_k^{(n)})} \right]^2 \rho(x)\,dx; \tag{8}$$

$$c_k^{(n)} = \frac{k_{0,n+1}}{k_{0,n}} \frac{-\|p_n\|^2}{p_{n+1}(x_k^{(n)})\,p_n'(x_k^{(n)})} = \frac{k_{0,n}}{k_{0,n-1}} \frac{\|p_{n-1}\|^2}{p_{n-1}(x_k^{(n)})\,p_n'(x_k^{(n)})}; \tag{9}$$

$$[c_k^{(n)}]^{-1} = \sum_{j=0}^n \|p_j\|^{-2} [p_j(x_k^{(n)})]^2. \tag{10}$$

Доведення. Формули (7) і (8) випливають з того, що формула Гаусса (1) є точною для довільного многочлена степеня не вище $2n-1$ і, зокрема, для $f(x) = l$ і

$$f(x) = \left[\frac{p_n(x)}{p_n'(x_k^{(n)})\,(x - x_k^{(n)})} \right]^2 \in \pi_{2n-2},$$

а нерівність (6) являється наслідком (8).

Для доведення (9) покладемо у формулі Крістоффеля — Дарбу

$$\sum_{i=0}^n \|p_i\|^{-2} p_i(x) p_i(y) = \frac{k_{0,n}\|p_n\|^{-2}}{k_{0,n+1}} \frac{p_{n+1}(x) p_n(y) - p_n(x) p_{n+1}(y)}{x - y}, \tag{11}$$

$y = x_k^{(n)}$, помножимо обидві частини дістаної рівності на вираз

$$-\frac{k_{0,n+1}}{k_{0,n}\|p_n\|} \frac{1}{p_{n+1}(x_k^{(n)})\,p_n'(x_k^{(n)})}$$

і проінтегруємо з вагою $\rho(x)$ від a до b. Як наслідок дістанемо

$$-\frac{k_{0,n+1}}{k_{0,n}\|p_n\|} \frac{1}{p_{n+1}(x_k^{(n)})\,p_n'(x_k^{(n)})} \sum_{i=0}^n \|p_i\|^{-2} p_i(x_k^{(n)}) \int_a^b \rho(x) p_i(x)\,dx =$$

$$= -\|p_n\|^{-3} \int_a^b \rho(x) \frac{p_{n+1}(x) p_n(x_k^{(n)}) - p_n(x) p_{n+1}(x_k^{(n)})}{p_{n+1}(x_k^{(n)})\,p_n'(x_k^{(n)})\,(x - x_k^{(n)})}\,dx. \tag{12}$$

В силу ортогональності системи многочленів $p_i(x)$ маємо

$$p_i(x_k^{(n)}) \int_a^b \rho(x) p_i(x)\,dx = \begin{cases} 0, & i > 0, \\ \|p_0\|^2, & i = 0. \end{cases}$$

Враховуючи також, що

$$p_n(x_k^{(n)}) = 0, \quad \omega(x) = \frac{p_n(x)}{k_{0,n}},$$

$$\frac{-p_n(x)\,p_{n+1}(x_k^{(n)})}{p_{n+1}(x_k^{(n)})\,p_n'(x_k^{(n)})\,(x - x_k^{(n)})} = -\frac{\dfrac{p_n(x)}{k_{0,n}}}{\dfrac{p_n'(x_k^{(n)})}{k_{0,n}}\,(x - x_k^{(n)})} =$$

$$= -\frac{\omega(x)}{\omega'(x_k^{(n)})(x - x_k^{(n)})}$$

і формули (2), з (12) дістаємо першу рівність формул (9). Друга рівність у (9) випливає з першої з урахуванням рекурентного співвідношення (див. 1.6)

$$a_{n+1}p_{n+1}(x) + (b_n - x)p_n(x) + c_{n-1}p_{n-1}(x) = 0,$$

$$c_{n-1} = \frac{k_{0,n-1}}{k_{0,n}} \frac{\|p_n\|^2}{\|p_{n-1}\|^2}, \quad b_n = \frac{k_{1,n}k_{0,n+1} - k_{0,n}k_{1,n+1}}{k_{0,n+1}k_{0,n}},$$

$$a_{n+1} = \frac{k_{0,n}}{k_{0,n+1}}.$$

Нарешті, справедливість формули (10) встановлюється шляхом граничного переходу у тотожності Крістоффеля — Дарбу (11) при $x = x_k^{(n)}$, $y \to x_k^{(n)}$ із використанням формули (9).

Таким чином, для побудови квадратурної формули найвищого алгебраїчного ступеня точності необхідно побудувати відповідну систему ортогональних многочленів і знайти їхні корені. Для вагових функцій $\rho(x)$, пов'язаних з класичними ортогональними многочленами, є таблиці вагових коефіцієнтів і абсцис відповідних квадратурних формул Гаусса.

Приклад 1. Обчислити вагові коефіцієнти квадратурної формули Гаусса для інтегралів виду

$$\int_{-1}^{1} f(x)(1 - x^2)^{-1/2}\,dx, \tag{13}$$

тобто $\rho(x) = (1 - x^2)^{-1/2}$.

Розв'язання. З вагою $\rho(x) = (1 - x^2)^{-1/2}$ на відрізку $[-1, 1]$ пов'язана система ортогональних многочленів Чебишева першого роду

$$T_n(x) = P_n^{(-1/2;-1/2)}(x) = \cos(n \arccos x).$$

Тому вузли квадратурної формули, точної для многочленів до $(2n - 1)$-го степеня, знаходять з рівнянь

$$T_n(x) = \cos(n \arccos x) = 0,$$

звідки

$$x_k^{(n)} = \cos\frac{2k - 1}{2n}\pi, \quad k = \overline{1, n}. \tag{14}$$

Відповідні ваги знаходимо за формулою (9), враховуючи, що

$$\|p_n\|^2 = \int_{-1}^{1}(1 - x^2)^{-1/2}\cos^2(n \arccos x)\,dx = -\int_{\pi}^{0}\cos^2 nt\,dt = \int_{0}^{\pi}\frac{1 + \cos 2nt}{2}\,dt =$$

$$= \frac{\pi}{2} + \frac{1}{4n}\sin 2nt\,\Big|_0^{\pi} = \frac{\pi}{2},$$

$$k_{0,n} = 2^{n-1}.$$

Маємо

$$c_k^{(n)} = \frac{2^{n-1}}{2^{n-2}} \frac{\frac{\pi}{2}}{\cos\left[(n-1)\frac{2k-1}{2n}\pi\right]\left\{\sin\left(\pi\frac{2k-1}{2n}n\right)n\frac{1}{\sqrt{1 - \cos^2\frac{2k-1}{2n}\pi}}\right\}} =$$

$$= \frac{\pi}{n}. \tag{15}$$

Приклад 2. Побудувати квадратурну формулу Гаусса з двома вузлами для інтегралів вигляду

$$I(f) = \int_0^1 f(x)\,dx.$$

Розв'язання. Як випливає із загальної теорії, насамперед треба знайти нулі $x_1^{(2)}$, $x_2^{(2)}$ многочлена $p_2(x)$ другого степеня із системи ортогональних на $[0, 1]$ з вагою $\rho(x) \equiv 1$ многочленів $\{p_i(x)\}_{i=0}^{\infty}$. Ці нулі і будуть вузлами квадратурної формули. Оскільки формула Гаусса є формулою інтерполяційного типу, то її коефіцієнти за відомими вузлами обчислюються за формулами

$$c_1^{(2)} = \int_0^1 l_{1,2}(x)\,dx = \int_0^1 \frac{x - x_2^{(2)}}{x_1^{(2)} - x_2^{(2)}}\,dx,$$

$$c_2^{(2)} = \int_0^1 l_{2,2}(x)\,dx = \int_0^1 \frac{x - x_1^{(2)}}{x_2^{(2)} - x_1^{(2)}}\,dx.$$

Як відомо (див. п. 1.6), ортогональну систему на $[-1, 1]$ з вагою $\rho(x) \equiv 1$ утворюють многочлени Лежандра $\{p_i(x)\}_{i=0}^{\infty}$, які можна знайти, наприклад, з рекурентного співвідношення

$$p_{-1}(x) = 0, \quad p_0(x) = 1, \quad p_{n+1}(x) = \frac{2n+1}{n+1}xp_n(x) - \frac{n}{n+1}p_{n-1}(x),$$

$$n = 0, 1, \ldots.$$

Якщо виконати лінійну заміну змінних $x = 2t - 1$, за допомогою якої відрізок $[-1, 1]$ відображається на відрізок $[0, 1]$, то очевидно, що система многочленів $\{p_i(t)\}_{i=0}^{\infty} \equiv \{p_i(2t - 1)\}_{i=0}^{\infty}$ буде ортогональною на $[0, 1]$ з вагою 1. З рекурентного співвідношення маємо $p_1(x) = x$, $p_2(x) = \frac{3}{2}x^2 - \frac{1}{2} = \frac{3x^2 - 1}{2}$ і тому

$$p_2(t) \equiv \tilde{p}_2(t) = \frac{3(2t - 1)^2 - 1}{2} = \frac{12t^2 - 12t + 3 - 1}{2} = 6t^2 - 6t + 1.$$

Коренями цього многочлена є $x_1^{(2)} = \frac{3 - \sqrt{3}}{6}$, $x_2^{(2)} = \frac{3 + \sqrt{3}}{6}$, і ці корені є вузлами шуканої квадратурної формули. Користуючись наведеними вище формулами, дістаємо $c_1^{(2)} = c_2^{(2)} = \frac{1}{2}$ і, отже, формула Гаусса алгебраїчного ступеня точності 3 має вигляд

$$I_2(x) = \frac{1}{2}\left[f\left(\frac{3 - \sqrt{3}}{6}\right) + f\left(\frac{3 + \sqrt{3}}{6}\right)\right].$$

Подамо чисельні значення невід'ємних вузлів і коефіцієнтів $x_j^{(n)}$, $c_j^{(n)}$ формули Гаусса для вагової функції $\rho(x) = 1$, $n = 1, 2, 3, 4$, і відрізка $[-1, 1]$ з десятьма десятковими знаками після коми у таблиці:

n	x_1 C_1	n	x_1 C_1	x_2 C_2
1	0,0000000000 1,0000000000	3	0,0000000000 0,8888888888	0,7745966692 0,5555555556
2	0,5773502692 1,0000000000	4	0,3399810436 0,6521451549	0,8611363116 0,3478548451

Вузли формули Гаусса є нулями многочлена Лежандра $P_n(x)$ і розміщені симетрично відносно точки $x = 0$, коефіцієнти $c_j^{(n)}$ додатні і в симетричних вузлах збігаються при будь-якому n.

За допомогою формули Гаусса для ваги $\rho(x) = 1$, $x \in [-1, 1]$,

$$\int_{-1}^{1} f(x)\,dx \approx \sum_{j=1}^{n} c_j^{(n)} f(x_j^{(n)}) \tag{16}$$

можна побудувати ускладнену формулу Гаусса на довільному відрізку $[a, b]$. З цією метою відрізок $[a, b]$ розбивається на N рівних відрізків $[x_k^*, x_{k+1}^*]$, де $x_k^* = a + k(b-a)/N$, $k = \overline{0, N-1}$, $x_N^* = b$, і на кожному частинному відрізку $[x_k^*, x_{k+1}^*]$ задається n вузлів

$$x_{kj} = \frac{x_k^* + x_{k+1}^*}{2} + x_j \frac{b-a}{N}, \quad j = \overline{1, n}, \tag{17}$$

де x_j — вузли канонічної формули Гаусса (16). Квадратурна формула

$$\int_{x_k^*}^{x_{k+1}^*} f(x)\,dx \approx \frac{b-a}{2N} \sum_{j=1}^{n} c_j^{(n)} f(x_{kj})$$

буде точною для многочленів степеня $2n-1$. Підсумовуючи, дістаємо ускладнену формулу Гаусса

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2N} \sum_{j=1}^{n} c_j^{(n)} \sum_{k=0}^{N-1} f(x_{kj}), \tag{18}$$

точну для многочленів степеня $2n-1$.

Квадратурна формула, вага і абсциси якої обчислені в прикладі 1, має ту особливість, що при сталих коефіцієнтах її алгебраїчний

ступінь точності є $2n-1$. Квадратурні формули з рівними коефіцієнтами називаються *формулами Чебишева*, і питання про їхню побудову для довільної вагової функції $\rho(x)$ розглянемо далі.

5.4. Квадратурні формули Чебишева

Розглянемо квадратурні формули інтерполяційного типу виду

$$\int_{-1}^{1} f(x)\rho(x)\,dx = C \sum_{k=1}^{n} f(x_k^{(n)}) + R(f), \quad \rho(x) > 0, \tag{1}$$

де ваговий множник C і вузли $x_k^{(n)} \in [-1, 1]$, $k = \overline{1, n}$, виберемо з умови, щоб

$$R(f) = 0 \quad \forall f \in \pi_m \tag{1'}$$

для максимально великого m. Оскільки формула (1) містить $n+1$ параметр $(C, x_k^{(n)}, k = \overline{1, n})$, то $m \geqslant n$.

Застосування квадратурних формул (1) найдоцільніше тоді, коли значення $f(x_k^{(n)})$ знаходяться, наприклад, вимірюванням і мають випадкові похибки. Дійсно, нехай ці похибки є незалежними випадковими величинами із сталою дисперсією σ і математичним сподіванням, що дорівнює нулю. Тоді дисперсія похибки наближеного значення інтеграла, який обчислюється за формулою

$$\int_{-1}^{1} \rho(x) f(x)\,dx \approx \sum_{i=1}^{n} C_i^{(n)} f(x_i^{(n)}), \tag{2}$$

дорівнюватиме

$$D\left(\sum_{i=1}^{n} C_i^{(n)} f(x^{(n)})\right) = \sum_{i=1}^{n} (C_i^{(n)})^2 D f(x_i^{(n)}) = \sigma \sum_{i=1}^{n} (C_i^{(n)})^2, \tag{3}$$

а її математичне сподівання дорівнює нулю. Вимагаючи, щоб формула (2) була точною для $f(x) = \text{const}$, дістаємо

$$\sum_{i=1}^{n} C_i^{(n)} = \int_{-1}^{1} \rho(x)\,dx = \mu_0. \tag{4}$$

Неважко помітити, що мінімум правої частини (3) (мінімум дисперсії) за умови (4) досягається при $C_i^{(n)} = \mu_0/n$, $i = \overline{1, n}$, тобто для квадратурних формул виду (1).

Оскільки формула (1) інтерполяційна, то

$$\frac{\mu_0}{n} = C = \int_{-1}^{1} \rho(x) \frac{\omega(x)}{(x - x_k^{(n)})\,\omega'(x_k^{(n)})}\,dx, \quad k = \overline{1, n}. \tag{5}$$

З'ясуємо, чи існують для довільного $\rho(x) > 0$ такі абсциси $x_k^{(n)} \in [-1, 1]$, $k = \overline{1, n}$, для яких формула (1) має властивості (1') і (5) (нагадаємо, що для ваги $\rho(x) = (1 - x^2)^{-1/2}$ це має місце).

Очевидно, що задача відшукання абсцис $x_k^{(n)}$ еквівалентна задачі побудови многочлена

$$\omega(x) = \prod_{k=1}^{n} (x - x_k^{(n)}) = \sum_{k=0}^{n} b_k x^{n-k}, \quad b_0 = 1. \tag{6}$$

Оскільки для формули (1) має виконуватися умова (1') при $m = n$, то дістанемо систему

$$C \sum_{k=1}^{n} [x_k^{(n)}]^i = \int_{-1}^{1} \rho(x)\, x^i dx = \mu_i, \quad i = \overline{0, n}. \tag{7}$$

Ліві частини в (7) є симетричними функціями вузлів $x_k^{(n)}$. Через ці самі симетричні функції визначаються коефіцієнти похідної від многочлена $\omega(x)$. Дійсно,

$$\omega'(x) = \sum_{i=1}^{n} \frac{\omega(x)}{x - x_i^{(n)}} = nx^{n-2} + b_1(n-1)x^{n-2} + \cdots + b_{n-1}. \tag{8}$$

За схемою Горнера

$$\frac{\omega(x)}{x - x_i^{(n)}} = x^{n-1} + (b_1 + x_i^{(n)})x^{n-2} + (b_2 + b_1 x_i^{(n)} + (x_i^{(n)})^2)x^{n-3} + \cdots +$$

$$+ (b_{n-1} + b_{n-2}x_i^{(n)} + \cdots + (x_i^{(n)})^{n-1}). \tag{9}$$

Підставляючи (9) у (8) і порівнюючи коефіцієнти при однакових степенях x, маємо

$$C^{-1} \sum_{k=0}^{i} b_{i-k}\mu_k = (n - i)\, b_i, \quad i = \overline{1, n-1}. \tag{10}$$

Оскільки $b_0 = 1$ відоме і матриця системи (10) є нижньою трикутною, то з (10) послідовно знаходимо $b_1, b_2, \ldots, b_{n-1}$. Коефіцієнт b_n визначаємо з умови

$$\sum_{k=1}^{n} \omega(x_k^{(n)}) = 0 = C^{-1} \sum_{i=0}^{n} b_i \mu_{n-i}. \tag{11}$$

Таким чином, за формулами (5), (10), (11) визначаються коефіцієнти b_i многочлена $\omega(x)$ і C. Потім, щоб побудувати формулу (1), потрібно знайти корені $x_k^{(n)}$, $k = \overline{1, n}$, многочлена $\omega(x)$. Однак при обчисленнях виявилось, що для вагової функції $\rho(x) = 1$, наприклад при $n = 8$, многочлен $\omega(x)$ має комплексні корені, тобто для цього n і вагової функції $\rho(x) = 1$ побудувати формулу Чебишева (1) неможливо. Кажуть, що вагова функція $\rho(x)$ допускає квадратуру Чебишева, якщо система рівнянь (порівняйте з (7))

$$\frac{1}{n} \sum_{i=1}^{n} (x_i^{(n)})^k = \int_{-1}^{1} x^k \rho(x)\, dx, \quad k = \overline{1, n},$$

має дійсні корені для натуральних n.

Теорема 1 (теорема Бернштейна). *При $n \geqslant 10$ в квадратурі Чебишева* (1) *для $\rho(x) \equiv 1$ серед абсцис $x_k^{(n)}$ є комплексні, тобто ця вагова функція не допускає квадратури Чебишева.*

В 1966 р. американський математик І. Л. Алмен довів, що крім вагової функції $\rho(x) \equiv (1 - x^2)^{-1/2}$ існують і інші вагові функції, які допускають квадратуру Чебишева, а саме він довів таке твердження.

Теорема 2 (теорема Алмена). *Якщо $|a| \leqslant \frac{1}{4}$, то вагова функція*

$$\rho(x) = \frac{1}{\pi\sqrt{1 - x^2}} \cdot \frac{1 + 2ax}{1 + 4a^2 + 4ax}$$

допускає квадратуру Чебишева.

Приклад. Нехай $\rho(x) \equiv 1$. Тоді маємо

$$C = \frac{2}{n}; \quad \mu_i = \int_{-1}^{1} x^i dx = \frac{1 - (-1)^{i+1}}{i + 1}, \quad i = \overline{0, n}. \tag{12}$$

Система рівнянь для визначення коефіцієнтів многочлена $\omega(x)$ має вигляд

$$\frac{n}{2} \sum_{k=0}^{i} \frac{1 - (-1)^{k+1}}{k + 1} b_{i-k} = (n - i)\, b_i, \quad i = \overline{1, n}. \tag{13}$$

Оскільки $b_0 = 1$, то з (13) маємо

$$\frac{n}{2}(2b_1 + 0 \cdot b_0) = (n - 1)\, b_1, \quad b_1 = 0,$$

$$\frac{n}{2}\left(2b_2 + 0 \cdot b_1 + \frac{2}{3} b_0\right) = (n - 2)\, b_2, \quad b_2 = -\frac{1}{3},$$

$$\frac{n}{2}\left(2b_3 + 0 \cdot b_2 + \frac{2}{3} b_1 + 0 \cdot b_0\right) = (n - 3)\, b_3, \quad b_3 = 0$$

і так далі, тобто всі b_i з непарними індексами рівні нулю. Тому

$$\omega(x) = \begin{cases} x^n + b_2 x^{n-2} + \cdots + b_{2m}x, & \text{якщо } n = 2m + 1, \\ x^n + b_2 x^{n-2} + \cdots + b_{2m}, & \text{якщо } n = 2m. \end{cases}$$

Звідси видно, що при парному n дійсні нулі многочлена $\omega(x)$, якщо вони існують

розміщені симетрично відносно початку координат. Тому

$$\sum_{k=1}^{n} (x_k^{(n)})^{n+1} = 0 = \int_{-1}^{1} x^{n+1} dx \equiv \mu_{n+1},$$

тобто формула Чебишева, яка існує для $n = \overline{1, 7}$, буде точною не тільки для многочленів степеня n, але й для степеня $n + 1$.

5.5. Оцінка залишкових членів квадратурних формул

5.5.1. Оцінки залишкових членів формул інтерполяційного типу для класів функцій $C^m[a, b]$. Як і раніше розглядатимемо квадратурні формули інтерполяційного типу. Зрозуміло, що їхній залишковий член тісно пов'язаний із залишковим членом відповідного інтерполяційного многочлена.

Розглянемо залишковий член квадратурної формули інтерполяційного типу

$$\int_a^b f(x)\rho(x)\,dx = \sum_{k=1}^{n} c_k^{(n)} f(x_k^{(n)}) + R(f), \quad \rho(x) \geqslant 0, \tag{1}$$

алгебраїчний ступінь точності якої $m \geqslant n - 1$. Замінюючи функцію $f(x)$ інтерполяційним многочленом m-го степеня $f(x) = p_m(x) + R_m(x)$ (якщо $m > n - 1$, то це буде інтерполяційний многочлен з кратними вузлами), який побудовано за вузлами $x_k^{(n)}$, $k = \overline{1, n}$, для $R(f)$ дістанемо

$$R(f) = \int_a^b \rho(x) f(x)\,dx - \sum_{k=1}^{n} c_k^{(n)} f(x_k^{(n)}) = \int_a^b \rho(x) p_m(x)\,dx +$$

$$+ \int_a^b \rho(x) R_m(x)\,dx - \sum_{k=1}^{n} c_k^{(n)} p_m(x_k^{(n)}) - \sum_{k=1}^{n} c_k^{(n)} R_m(x_k^{(n)}) =$$

$$= \int_a^b \rho(x) R_m(x)\,dx = \int_a^b \rho(x) \frac{f^{(m+1)}(\xi)}{(m+1)!} \Omega(x)\,dx, \quad \xi \in [a, b], \tag{2}$$

де $\Omega(x) = \prod_{k=1}^{n} (x - x_n^{(n)})$, якщо $m = n - 1$, і $\Omega(x) = \prod_{k=1}^{n} (x - x_k^{(n)})^{\alpha_k}$, $\sum_{k=1}^{n} \alpha_k = m$, якщо $m > n - 1$, $f(x) \in C^{(m+1)}(\bar{\Omega})$.

Формулу (2) зручніше перетворити до вигляду

$$R(f) = f^{(m+1)}(\eta) B_m, \quad \eta \in [a, b], \tag{3}$$

де B_m — стала, яка не залежить від $f(x)$. Розглянемо випадки, коли зображення (3) можливе.

По-перше, це можливо тоді, коли $\Omega(x)$ — знакостала функція. Застосовуючи теорему про середнє, з виразу (2) дістаємо (3), де

$$B_m = \frac{1}{(m+1)!} \int_a^b \rho(x) \Omega(x)\,dx. \tag{4}$$

Звідси зробимо висновок: для випадку $m > n - 1$ серед $x_k^{(n)}$, $k = \overline{1, n}$, слід вибирати кратні вузли й порядок кратності таким чином, щоб многочлен $\Omega(x)$ був знакосталим на $[a, b]$, якщо це можливо.

Другий випадок, коли справджується (3), розглядається в наступній лемі.

Лема 1. *Нехай функція*

$$k_n(y) = \frac{1}{m!} \left[\int_y^b \rho(x)(x - y)^m\,dx - \sum_{i=1}^{n} c_i^{(n)} [(x_i^{(n)} - y)^+]^m \right], \tag{4'}$$

де $(t)^+ = \max(0, t)$, *зберігає знак на* $[a, b]$. *Тоді справедлива формула* (3).

Доведення. Запишемо для функції $f(x)$ ряд Тейлора з залишковим членом в інтегральній формі

$$f(x) = \sum_{j=0}^{m} \frac{f^{(j)}(a)}{j!} (x - a)^j + \int_a^x \frac{f^{(m+1)}(y)}{m!} (x - y)^m\,dy,$$

або

$$f(x) = \sum_{j=0}^{m} \frac{f^{(j)}(a)}{j!} (x - a)^j + \int_a^b \frac{f^{(m+1)}(y)}{m!} [(x - y)^+]^m\,dy. \tag{5}$$

Застосовуючи до (5) лінійний оператор R і беручи до уваги, що формула (1) точна для многочленів степеня m, тобто $R(f) = 0\ \forall f \in \in \pi_m$, маємо

$$R(f) = R\left(\int_a^b \frac{f^{(m+1)}(y)}{m!} [(x - y)^+]^m\,dy \right) = \int_a^b \rho(x) \int_a^b \frac{f^{(m+1)}(y)}{m!} \times$$

$$\times [(x - y)^+]^m\,dy\,dx + \sum_{i=1}^{n} c_i^{(n)} \int_a^b \frac{f^{(m+1)}(y)}{m!} [(x_i^{(n)} - y)^+]^m\,dy =$$

$$= \int_a^b \frac{f^{(m+1)}(y)}{m!} \left[\int_y^b \rho(x)(x - y)^m\,dx \right] dy -$$

$$- \int_a^b \frac{f^{(m+1)}(y)}{m!} \sum_{i=1}^{n} c_i^{(n)} [(x_i^{(n)} - y)^+]^m\,dy = \int_a^b k_n(y) f^{(m+1)}(y)\,dy. \tag{5'}$$

В силу умов леми можна застосувати теорему про середнє, яка приводить до формули (3).

Зауважимо, що в формулі (3), якщо вона має місце, сталу B_m можна також подати у вигляді

$$B_m = \int_a^b \frac{(x-a)^{m+1}}{(m+1)!}\rho(x)\,dx - \sum_{k=1}^{n} c_k^{(n)} \frac{(x_k^{(n)}-a)^{m+1}}{(m+1)!}, \qquad (6)$$

тобто вона збігається із залишковим членом квадратурної формули (1) для функції

$$f(x) = \frac{(x-a)^{m+1}}{(m+1)!}.$$

Доведення цього факту очевидне.

Знайдемо залишкові члени найпростіших формул Ньютона — Котеса. Ми бачили, що формула середніх прямокутників будується за одним вузлом ($n = 1$) і має алгебраїчний ступінь точності 1 (тобто $m = 1$). Тому в силу (2), якщо $f \in C^2[a, b]$, то

$$R_{1,1}(f) = \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{f''(\xi)}{2!}\,dx = \frac{f''(\tilde{\xi})}{2!}\int_a^b \left(x - \frac{a+b}{2}\right)^2 dx =$$

$$= \frac{f''(\tilde{\xi})}{24}(b-a)^3. \qquad (7)$$

Тут $\Omega(x) = \left(x - \frac{a+b}{2}\right)^2$ і застосований інтерполяційний многочлен з одним двократним вузлом $x_1^{(2)} = x_2^{(2)} = \frac{a+b}{2}$.

Для спрощення вигляду залишкових членів ускладнених формул Ньютона — Котеса нам знадобиться елементарне твердження, яке ми сформулюємо у вигляді леми.

Лема 2. *Нехай $f \in C[a, b]$, $\xi_i \in [a, b]$ довільні точки $i = \overline{1, n}$. Тоді існує точка $\xi \in [a, b]$ така, що*

$$\frac{f(\xi_1) + \cdots + f(\xi_n)}{n} = f(\xi).$$

Твердження леми випливає з очевидних нерівностей $\min\limits_{[a,b]} f(x) \leqslant \frac{f(\xi_1) + \cdots + f(\xi_n)}{n} \leqslant \max\limits_{[a,b]} f(x)$ і теореми про проміжні значення неперервної функції. Для складеної формули середніх прямокутників (7) і леми 2 випливає, що

$$R(f) = \sum_{i=1}^{n} R_{1,1}^{(i)}(f) = \frac{h^3}{24}\sum_{i=1}^{N} f''(\xi^{(i)}) = \frac{h^2(b-a)}{24} f''(\xi), \qquad (7')$$

$$\xi \in [x, b],\quad \xi^{(i)} \in [x_{i-1}, x_i],\quad f \in C^2[a, b],\quad h = \frac{b-a}{N}.$$

У формулі трапецій $n = 2$ і її алгебраїчний ступінь точності дорівнює 2. Вважаючи $\Omega(x) = (x-a)(x-b) < 0\ \forall x \in (a, b)$, дістанемо для

$$R_{2,0}(f) = \int_a^b (x-a)(x-b)\frac{f''(\xi)}{2!}\,dx = -\frac{f''(\tilde{\xi})}{12}(b-a)^3. \qquad (8)$$

Для складеної формули трапецій маємо у випадку $f(x) \in C^2[a, b]$

$$R(f) = \sum_{i=0}^{N} R_{2,0}^{(i)}(f) = \frac{-f''(\xi)}{12}h^2(b-a),\quad \xi \in [a, b], \qquad (8')$$

де $R_{2,0}^{(i)}(f)$ — залишковий член формули трапецій для i-го інтервалу $[a + ih,\ a + (i+1)h]$, $h = (b-a)/(N+1)$.

У формулі Сімпсона число вузлів $n = 3$, а алгебраїчний ступінь точності $m = 3$. Тому для залишкового члена маємо

$$R_{3,0}(f) = \int_a^b (x-a)\left(x - \frac{a+b}{2}\right)^2 (x-b)\frac{f^{(\mathrm{IV})}(\xi)}{4!}\,dx =$$

$$= -\frac{f^{(\mathrm{IV})}(\tilde{\xi})}{2880}(b-a)^5. \qquad (9)$$

Тут ми вибрали $\Omega(x) = (x-a)\left(x - \frac{b+a}{2}\right)^2(x-b) < 0\ \ \forall x \in (a, b)$, тобто використали многочлен з кратними вузлами

$$x_1^{(n)} = a,\quad x_2^{(n)} = x_3^{(n)} = \frac{a+b}{2},\quad x_4^{(n)} = b.$$

Інший вибір кратних вузлів не забезпечив би знакосталість $\Omega(x)$ і як наслідок можливість перетворення (3).

Аналогічно попередньому знаходимо залишковий член складеної формули Сімпсона ($f \in C^4[a, b]$)

$$R(f) = \sum_{i=0}^{N-1} R_{3,0}^{(i)}(f) = -(2h)^5\frac{1}{2880}\sum_{i=0}^{N-1} f^{(\mathrm{IV})}(\xi^{(i)}) = -\frac{2h^4}{180}f^{(\mathrm{IV})}(\xi)\sum_{i=0}^{N-1} h =$$

$$= -\frac{2h^4}{180}f^{(\mathrm{IV})}(\xi)\frac{b-a}{2} = -\frac{(b-a)f^{(\mathrm{IV})}(\xi)}{180}h^4,\quad \xi \in [a, b]. \qquad (9')$$

Розглянемо правило трьох восьмих, для якого $n = 4$, алгебраїчний ступінь точності $m = 3$. Вибравши $\Omega(x) = (x-a)\left(x - \frac{2a+b}{3}\right)\left(x - \frac{a+2b}{3}\right)(x-b)$, знайдемо

$$R(f) = \int_a^b (x-a)\left(x - \frac{2a+b}{3}\right)\left(x - \frac{a+2b}{3}\right)(x-b) \times$$

$$\times \frac{f^{(\mathrm{IV})}(\xi)}{4!}\,dx = \int_a^b k_n(y) f^{(\mathrm{IV})}(y)\,dy.$$

В силу умов леми можна застосувати теорему про середнє, яка приводить до формули (3).

Зауважимо, що в формулі (3), якщо вона має місце, сталу B_m можна також подати у вигляді

$$B_m = \int_a^b \frac{(x-a)^{m+1}}{(m+1)!}\,\rho(x)\,dx - \sum_{k=1}^{n} c_k^{(n)}\,\frac{(x_k^{(n)}-a)^{m+1}}{(m+1)!}, \qquad (6)$$

тобто вона збігається із залишковим членом квадратурної формули (1) для функції

$$f(x) = \frac{(x-a)^{m+1}}{(m+1)!}.$$

Доведення цього факту очевидне.

Знайдемо залишкові члени найпростіших формул Ньютона — Котеса. Ми бачили, що формула середніх прямокутників будується за одним вузлом ($n = 1$) і має алгебраїчний ступінь точності 1 (тобто $m = 1$). Тому в силу (2), якщо $f \in C^2[a, b]$, то

$$R_{1,1}(f) = \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{f''(\xi)}{2!}\,dx = \frac{f''(\tilde{\xi})}{2!}\int_a^b \left(x - \frac{a+b}{2}\right)^2 dx =$$

$$= \frac{f''(\tilde{\xi})}{24}(b-a)^3. \qquad (7)$$

Тут $\Omega(x) = \left(x - \frac{a+b}{2}\right)^2$ і застосований інтерполяційний многочлен з одним двократним вузлом $x_1^{(2)} = x_2^{(2)} = \frac{a+b}{2}$.

Для спрощення вигляду залишкових членів ускладнених формул Ньютона — Котеса нам знадобиться елементарне твердження, яке ми сформулюємо у вигляді леми.

Лема 2. *Нехай $f \in C[a, b]$, $\xi_i \in [a, b]$ довільні точки $i = \overline{1, n}$. Тоді існує точка $\xi \in [a, b]$ така, що*

$$\frac{f(\xi_1) + \cdots + f(\xi_n)}{n} = f(\xi).$$

Твердження леми випливає з очевидних нерівностей $\min\limits_{[a,b]} f(x) \leqslant \frac{f(\xi_1) + \cdots + f(\xi_n)}{n} \leqslant \max\limits_{[a,b]} f(x)$ і теореми про проміжні значення неперервної функції. Для складеної формули середніх прямокутників (7) і леми 2 випливає, що

$$R(f) = \sum_{i=1}^{n} R_{1,1}^{(i)}(f) = \frac{h^3}{24}\sum_{i=1}^{N} f''(\xi^{(i)}) = \frac{h^2(b-a)}{24} f''(\xi), \qquad (7')$$

$$\xi \in [x, b], \quad \xi^{(i)} \in [x_{i-1}, x_i], \quad f \in C^2[a, b], \quad h = \frac{b-a}{N}.$$

У формулі трапецій $n = 2$ і її алгебраїчний ступінь точності дорівнює 2. Вважаючи $\Omega(x) = (x-a)(x-b) < 0 \ \forall x \in (a, b)$, дістанемо для

$$R_{2,0}(f) = \int_a^b (x-a)(x-b)\frac{f''(\xi)}{2!}\,dx = -\frac{f''(\tilde{\xi})}{12}(b-a)^3. \qquad (8)$$

Для складеної формули трапецій маємо у випадку $f(x) \in C^2[a, b]$

$$R(f) = \sum_{i=0}^{N} R_{2,0}^{(i)}(f) = \frac{-f''(\xi)}{12}h^2(b-a), \quad \xi \in [a, b], \qquad (8')$$

де $R_{2,0}^{(i)}(f)$ — залишковий член формули трапецій для i-го інтервалу $[a + ih, a + (i+1)h]$, $h = (b-a)/(N+1)$.

У формулі Сімпсона число вузлів $n = 3$, а алгебраїчний ступінь точності $m = 3$. Тому для залишкового члена маємо

$$R_{3,0}(f) = \int_a^b (x-a)\left(x - \frac{a+b}{2}\right)^2 (x-b)\frac{f^{(\mathrm{IV})}(\xi)}{4!}\,dx =$$

$$= -\frac{f^{(\mathrm{IV})}(\tilde{\xi})}{2880}(b-a)^5. \qquad (9)$$

Тут ми вибрали $\Omega(x) = (x-a)\left(x - \frac{b+a}{2}\right)^2 (x-b) < 0 \ \ \forall x \in (a, b)$, тобто використали многочлен з кратними вузлами

$$x_1^{(n)} = a, \quad x_2^{(n)} = x_3^{(n)} = \frac{a+b}{2}, \quad x_4^{(n)} = b.$$

Інший вибір кратних вузлів не забезпечив би знакосталість $\Omega(x)$ і як наслідок можливість перетворення (3).

Аналогічно попередньому знаходимо залишковий член складеної формули Сімпсона ($f \in C^4[a, b]$)

$$R(f) = \sum_{i=0}^{N-1} R_{3,0}^{(i)}(f) = -(2h)^5\frac{1}{2880}\sum_{i=0}^{N-1} f^{(\mathrm{IV})}(\xi^{(i)}) = -\frac{2h^4}{180} f^{(\mathrm{IV})}(\xi)\sum_{i=0}^{N-1} h =$$

$$= -\frac{2h^4}{180} f^{(\mathrm{IV})}(\xi)\frac{b-a}{2} = -\frac{(b-a) f^{(\mathrm{IV})}(\xi)}{180}h^4, \quad \xi \in [a, b]. \qquad (9')$$

Розглянемо правило трьох восьмих, для якого $n = 4$, алгебраїчний ступінь точності $m = 3$. Вибравши $\Omega(x) = (x-a)\left(x - \frac{2a+b}{3}\right)\left(x - \frac{a+2b}{3}\right)(x-b)$, знайдемо

$$R(f) = \int_a^b (x-a)\left(x - \frac{2a+b}{3}\right)\left(x - \frac{a+2b}{3}\right)(x-b) \times$$

$$\times \frac{f^{(\mathrm{IV})}(\xi)}{4!}\,dx = \int_a^b k_n(y) f^{(\mathrm{IV})}(y)\,dy.$$

Тут $\Omega(x)$ не є знакосталою, але можливо застосувати лему 1, бо

$$k_n(y) = \frac{1}{4!}(b-y)^4 - \frac{1}{48}\left\{[(a-y)^+]^3 + 3\left[\left(\frac{2a+b}{3} - y\right)^+\right]^3 + 3\left[\left(\frac{a+2b}{3} - y\right)^+\right]^3 + [(b-y)^+]^3\right\}$$

і, розбивши проміжок $[a, b]$ на три інтервали $\left[a, \frac{2a+b}{3}\right]$, $\left[\frac{2a+b}{3}, \frac{a+2b}{3}\right]$, $\left[\frac{a+2b}{3}, b\right]$, неважко довести, що в кожному з них $k_n(y) \leqslant 0$. Тому лема 1 приводить до результату

$$R(f) = -\frac{f^{(\mathrm{IV})}(\xi)}{6480}(b-a)^5, \quad \xi \in [a, b], \tag{10}$$

якщо $f \in C^4[a, b]$.

Розглянемо залишковий член квадратурних формул найвищого алгебраїчного ступеня точності.

Теорема 1. *Нехай* $\rho(x) \geqslant 0$, $\forall x \in [a, b]$, $f(x) \in C^{2n}[a, b]$. *Тоді існує точка* $\xi \in [a, b]$ *така, що для залишкового члена* $R(f)$ *квадратурної формули* (1) з (п. 5.3) *найвищого алгебраїчного ступеня точності справедлива рівність*

$$R(f) = \frac{f^{(2n)}(\xi)}{(2n)!}\int_a^b \rho(x)\,\Omega(x)\,dx, \tag{11}$$

де $\Omega(x) = \prod_{i=1}^{n}(x - x_i^{(n)})^2$, $\xi \in [a, b]$.

Доведення. Розглянемо інтерполяційний многочлен $p_{2n-1}(x)$ з кратними вузлами, який задовольняє умови

$$p_{2n-1}^{(\mu)}(x_i^{(n)}) = f^{(\mu)}(x_i^{(n)}), \quad \mu = 0, 1; \quad i = \overline{1, n}, \tag{12}$$

і підставимо в формулу (1) (п. 5.3) замість f функцію

$$f(x) = p_{2n-1}(x) + \frac{f^{(2n)}(\xi)}{(2n)!}\Omega(x), \quad \xi \in (a, b).$$

Оскільки формула (1) з п. 5.3 точна для многочленів степеня $2n-1$, дістанемо

$$R(f) = R\left(\frac{f^{(2n)}(\xi)}{(2n)!}\Omega(x)\right) = \int_a^b \rho(x)\frac{f^{(2n)}(\xi)}{(2n)!}\Omega(x)\,dx -$$

$$-\sum_{k=1}^{n} c_k^{(n)}\frac{f^{(2n)}(\xi)}{(2n)!}\Omega(x_k^{(n)}) = \int_a^b \rho(x)\frac{f^{(2n)}(\xi)}{(2n)!}\Omega(x)\,dx.$$

Звідси на основі теореми про середнє дійдемо твердження теореми.

Вправа 1. Показати, що для складеної квадратурної формули Гаусса (18) (п. 5.3) за умови $f \in C^{2n}[a, b]$ справедлива формула для залишкового члена

$$R_n(f) \equiv R(f) = \frac{(b-a)^{2n+1}}{N^{2n}}\frac{(n!)^4}{((2n)!)^3(2n+1)}f^{(2n)}(\xi), \quad \xi \in (a, b), \tag{13}$$

зокрема

$$R_2(f) = \frac{(b-a)^5}{4320N^4}f^{(4)}(\xi), \quad R_3(f) = \frac{(b-a)^7}{2\,016\,000N^6}f^{(6)}(\xi), \quad \xi \in (a, b). \tag{14}$$

Вказівка. Врахувати, що при $x = \frac{x_k^* + x_{k+1}^*}{2} + t\frac{b-a}{2N}$, $\tilde{f}(t) = f(x(t))$ маємо

$$\int_{x_k^*}^{x_{k+1}^*} f(x)\,dx = \frac{b-a}{2N}\int_{-1}^{1}\tilde{f}(t)\,dt$$

і залишковий член

$$R_n^{(k)}(f) = \int_{x_k^*}^{x_{k+1}^*} f(x)\,dx - \frac{b-a}{2N}\sum_{j=1}^{n} c_j^{(n)} f(x_{kj}) =$$

$$= \frac{b-a}{2N}\left[\int_{-1}^{1}\tilde{f}(t)\,dt - \sum_{j=1}^{n} c_j^{(n)}\tilde{f}(x_j)\right]$$

має оцінку

$$R_n^{(k)}(f) = \frac{b-a}{2N}\frac{\tilde{f}^{(2n)}(\eta_k)}{(2n)!}\int_{-1}^{1}(x-x_1)^2\ldots(x-x_n)^2\,dx =$$

$$= \frac{b-a}{2N}\frac{(b-a)^{2n}}{(2N)^{2n}}\frac{f^{2n}(\xi_k)}{(2n)!}\frac{1}{k_{0,n}^2}\int_{-1}^{1}P_n^2(x)\,dx,$$

де $\xi_k \in (x_k^*, x_{k+1}^*)$, $P_n(x)$ — многочлен Лежандра, $k_{0,n}$ — його старший коефіцієнт. Далі маємо (див. 1.6)

$$k_{0,n} = \frac{(2n)(2n-1)\ldots(n+1)}{2^n n!} = \frac{(2n)!}{2^n(n!)^2},$$

$$\|P_n\|^2 = \int_{-1}^{1}P_n^2(x)\,dx = \frac{2}{2n+1},$$

тому

$$R_n^{(k)}(f) = \frac{(b-a)^{2n+1}(n!)^4}{N^{2n+1}((2n)!)^3(2n+1)}f^{(2n)}(\xi_k), \quad \xi_k \in (x_k^*, x_{k+1}^*).$$

Порівняємо укладені квадратурні формули Сімпсона та Гаусса при $n = 2$. Оскільки в формулі Сімпсона h і N пов'язані співвідношенням $h = (b-a)/(2N)$, то її залишковий член можна записати у вигляді

$$R_C(f) = -h^4\frac{b-a}{180}f^{(4)}(\xi) = -\frac{(b-a)^5}{2880N^4}f^{(4)}(\xi).$$

Звідси видно, що залишковий член ускладненої формули Гаусса при $n = 2$ має числовий коефіцієнт в 1,5 раза менший, ніж залишковий член формули Сімпсона. Обидві формули точні для многочленів третього степеня. В формулі Сімпсона треба обчислити $2N + 1$ значень функції, а в формулі (18) (п. 5.3) при $n = 2$ використовується $2N$ значень функції. Але вказані переваги формули Гаусса слід вважати більш слабкими від тієї переваги формули Сімпсона, що її вузли розміщені рівномірно на $[a, b]$ з кроком h.

Квадратурну формулу Гаусса, в тому числі й ускладнену, слід застосувати при $n > 2$ для обчислення інтегралів від функцій, які мають відповідну високу гладкість.

5.5.2. Оцінки залишкових членів формул інтерполяційного типу для класів функцій $W_2^k(a, b)$. Нехай алгебраїчний ступінь точності формули інтерполяційного типу

$$\int_a^b f(x)\rho(x)\,dx = \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)}) + R(f) \tag{15}$$

є $m \geqslant n-1$ і

$$\sum_{k=1}^{n} |C_k^{(n)}| \leqslant M_n. \tag{16}$$

Виконаємо заміну змінних $t = \frac{x-a}{b-a}$ і перепишемо (15) у вигляді

$$R(f) = (b-a)\int_0^1 \tilde{f}(t)\tilde{\rho}(t)\,dt - \sum_{k=1}^{n} C_k^{(n)} \tilde{f}(t_k^{(n)}), \tag{17}$$

де $\tilde{f}(t) = f((b-a)t + a)$, $t_k^{(n)} = \frac{x_k^{(n)} - a}{b-a}$, $\tilde{\rho}(t) = \rho((b-a)t + a)$.

З виразу (17) за допомогою теореми вкладення (п. 1.5) дістаємо нерівність

$$|R(f)| = |R(\tilde{f})| \leqslant \|\tilde{f}\|_{C[0,1]}(b-a)\int_0^1 \tilde{\rho}(t)\,dt + M_n\|\tilde{f}\|_{C[0,1]} \leqslant M_0\|\tilde{f}\|_{W_2^l(0,1)}, \tag{18}$$

де $M_0 = \left[(b-a)\int_0^1 \tilde{\rho}(t)\,dt + M_n\right] M_l$, $l = \min(k, m+1)$, $0 < k$, M_l — стала з теореми вкладення. Нерівність (18) означає обмеженість лінійного функціонала $R(\tilde{f})$ в просторі $W_2^l(0, 1)$. За умовою цей функціонал перетворюється на нуль у многочленах m-го степеня. Тому в силу леми Брембла — Гільберта

$$|R(f)| = |R(\tilde{f})| \leqslant \overline{M} M_0 |\tilde{f}|_{W_2^l(0,1)}, \tag{19}$$

де

$$l = \min(k, m+1), \quad \overline{M} = \sqrt{1 + \sum_{j=0}^{l-2} \frac{1}{(2l-j-1)\,[(l-j-1)!]^2}}.$$

Перейшовши у виразі (19) до змінної x, дістанемо

$$|R(f)| \leqslant \overline{M} M_0 \left(\int_0^1 [\tilde{f}^{(l)}(t)]^2\,dt\right)^{1/2} = \overline{M} M_0 (b-a)^{l-1/2} |f|_{W_2^l(a,b)}. \tag{20}$$

Таким чином, ми довели наступну теорему.

Теорема 2. *Нехай $f(x) \in W_2^k(a, b)$, тоді для залишкового члена інтерполяційної квадратурної формули* (15) *алгебраїчного ступеня точності m справедлива оцінка* (20), *де* $l = \min(k, m+1)$, $M_0 = M_l\left[\int_a^b \rho(x)\,dx + \sum_{k=1}^{n} |C_k^{(n)}|\right]$, *а стала M залежить тільки від l.*

Розглянемо, наприклад, оцінку залишкового члена формули середніх прямокутників

$$\int_a^b f(x)\,dx = (b-a) f\left(\frac{a+b}{2}\right) + R_{1,1}(f),$$

для якої $n = 1$, $m = 1$. Легко знайти, що $M_0 = 4\sqrt{l}\,(b-a)$, тому

$$|R_{1,1}(f)| \leqslant \overline{M} 4\sqrt{l}\,(b-a)^{l+1/2} |f|_{W_2^l(a,b)}. \tag{21}$$

Розглянемо похибку складеної формули прямокутників

$$R_{\text{пр}}(f) = \sum_{i=1}^{N} R_{1,1}^{(i)}(f).$$

Застосовуючи до кожного $R_{1,1}^{(i)}(f)$ оцінку (21), дістаємо

$$|R_{\text{пр}}(f)| \leqslant \overline{M} 4\sqrt{l}\,h^{l+1/2} \sum_{i=1}^{N} |f|_{W_2^l(x_{i-1}, x_i)} \leqslant$$

$$\leqslant 4\sqrt{l}\,\overline{M} h^{l+1/2} \sqrt{N}\left(\sum_{i=1}^{N} |f|^2_{W_2^l(x_{i-1}, x_i)}\right)^{1/2} \leqslant$$

$$\leqslant 4\sqrt{l\,(b-a)}\,\overline{M} h^l |f|_{W_2^l(a,b)}. \tag{22}$$

Зокрема, при $k = 1$ (тобто при $f \in W_2^1(a, b)$) маємо

$$|R_{\text{пр}}(f)| \leqslant 4\sqrt{b-a}\,\overline{M}\,h\,|f|_{W_2^1(a,b)}, \tag{23}$$

при $k = 2$ (тобто, коли $f \in W_2^2(a, b)$) маємо

$$|R(f)| \leqslant 4\sqrt{2(b-a)}\,\overline{M}\,h^2 |f|_{W_2^2(a,b)}. \tag{24}$$

Вправа 2. Знайти оцінки залишкових членів формул трапеції і Сімпсона у випадку $f \in W_2^k(a, b)$, $0 < k \leqslant 2$, і $f \in W_2^k(a, b)$, $0 < k \leqslant 4$, відповідно.

Відповідь. Для канонічної формули трапеції маємо

$$|R_{2,0}(f)| \leqslant 4\sqrt{l}\,\overline{M}\,(b-a)^{l+1/2}\,|f|_{W_2^l(a,b)}, \tag{25}$$

$$l = \min(k, 2), \quad 0 < k \leqslant 2.$$

Для ускладненої формули трапецій

$$|R_{\text{тр}}(f)| = \left|\sum_{i=0}^{N} R_{2,0}^{(i)}(f)\right| \leqslant 4\sqrt{l\,(b-a)}\,\overline{M}\,h^l\,|f|_{W_2^l(a,b)}. \tag{26}$$

Для канонічної формули Сімпсона

$$|R_{3,0}(f)| \leqslant 4\sqrt{l}\,\overline{M}\,(b-a)^{l+1/2}\,|f|_{W_2^l(a,b)}, \tag{26'}$$

$$l = \min(k, 4), \quad f \in W^k(a, b).$$

Для ускладненої формули Сімпсона

$$|R_C(f)| = \left|\sum_{i=0}^{N-1} R_{3,0}^{(i)}(f)\right| \leqslant 4\sqrt{l}\,\overline{M}\,(2h)^{l+1/2}\sqrt{N}\,|f|_{W_2^2(a,b)} =$$

$$= 4\sqrt{l}\,\overline{M}\,(2h)^l\sqrt{b-a}\,|f|_{W_2^l(a,b)}, \tag{27}$$

$$l = \min(k, 4), \quad 0 < k \leqslant 4.$$

Знайдемо оцінку залишкового члена квадратурної формули найвищого алгебраїчного ступеня точності для функцій $f \in W_2^k(a, b)$, $0 < k \leqslant l$. З теореми 2 випливає, що

$$M_0 = 2\sqrt{l}\left[\int_a^b \rho(x)\,dx + \sum_{k=1}^{n} C_k^{(n)}\right] = 4\sqrt{l}\int_a^b \rho(x)\,dx.$$

Тому з (20) дістаємо

$$|R(f)| \leqslant 4\sqrt{l}\int_a^b \rho(x)\,dx\,\overline{M}\,(b-a)^{l-1/2}\,|f|_{W_2^l(a,b)}, \tag{28}$$

а у випадку $\rho(x) = 1$

$$|R(f)| \leqslant 4\sqrt{l}\,\overline{M}\,(b-a)^{l+1/2}\,|f|_{W_2^l(a,b)}, \tag{29}$$

$$f \in W_2^k(a, b), \quad 0 < k \leqslant 2n, \quad l = \min(k, 2n).$$

Для ускладненої формули Гаусса (18) з п. 5.3 аналогічно попередньому

$$|R(f)| \leqslant 4\sqrt{l}\,\overline{M}\left(\frac{b-a}{N}\right)^{l+1/2}\sqrt{N}\,|f|_{W_2^l(a,b)} =$$

$$= 4\sqrt{l(b-a)}\left(\frac{b-a}{N}\right)^{l}|f|_{W_2^l(a,b)}, \tag{30}$$

$$l = \min(k, 2n), \quad 0 < k \leqslant 2n, \quad f \in W_2^k(a, b).$$

Знайдені в даному параграфі оцінки показують, що залишкові члени ускладнених формул прямують до нуля при $h \to 0$ і вказують на порядок залишкового члена за h. Отже, зменшуючи h, за допомогою відповідних квадратурних формул при відповідній гладкості підінтегральної функції можна обчислити інтеграл як завгодно точно.

Порівнюючи оцінки (1) п. 5.5 цього пункту, бачимо, що вони однакові за порядком h для функцій з класів $C^k[a, b]$ і $W_2^k(a, b)$. Однак у випадку $f \in C^k[a, b]$ оцінки (1) (п. 5.5) кращі в тому розумінні, що у них точно відомий вигляд множника при степені h. Це дає змогу знаходити не тільки оцінки зверху для залишкових членів, але й оцінки знизу. Наприклад, для формул середніх прямокутників

$$|I - I_h^{\text{пр}}| \geqslant h^2\,\frac{(b-a)}{24}\min_{[a,b]}|f''(x)|. \tag{31}$$

Приклад. Дослідити похибки квадратурних формул прямокутників та Сімпсона для інтегралу

$$I = \int_0^{0,5} e^{-x^2}dx,$$

який не обчислюється в елементарних функціях.

Розв'язання. Маємо $f(x) = e^{-x^2}$, $f''(x) = (4x^2 - 2)\,e^{-x^2}$, $f^{(4)}(x) = 4(4x^4 - 12x^2 + 3)\,e^{-x^2}$, $e^{-1/4} \leqslant |f''(x)| \leqslant 2$, $|f^{(4)}(x)| \leqslant 12$ на $[0, 0,5]$. Звідси при $h = 0,05$ маємо $0,4 \cdot 10^{-4} \leqslant |I - I_h^{\text{пр}}| \leqslant 0,11 \cdot 10^{-3}$, $|I - I_h^c| \leqslant 0,21 \cdot 10^{-6}$. Звідси видно, що формула Сімпсона дає точніший результат, тобто справжня оцінка її залишкового члена значно менша нижньої оцінки похибки формули прямокутників.

5.5.3. Точна оцінка залишкового члена на класі функцій. Оцінки залишкових членів давалися вище для конкретної квадратурної формули і конкретної підінтегральної функції. Теоретичний і практичний інтерес мають також оцінки залишкових членів для деякого класу функцій чи деякого класу квадратурних формул. Знайдемо точну оцінку залишкового члена, якщо підінтегральна функція $f(x)$ є будь-яким представником класу $C^{(m+1)}(M; a, b)$ неперервно диференційовних до порядку m включно і таких, що мають кусково-неперервну похідну порядку $m + 1$, яка задовольняє нерівність $|f^{(m+1)}(x)| \leqslant M$.

З (1) і (5') випливає нерівність

$$\left|\int_a^b \rho(x)f(x)\,dx - \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)})\right| \leqslant M\int_a^b |K_n(x)|\,dx = MN_n, \tag{32}$$

де стала $N_n = \int_a^b |K_n(x)|\,dx$ не залежить від $f(x)$, але залежить від конкретної квадратурної формули (1). З іншого боку для функції

$g(x) \in C^{(m+1)}(M; a, b)$ такої, що $g^{(m+1)}(x) = M \operatorname{sign} K_n(x)$, нерівність (32) перетворюється в рівність, тому

$$\sup_{f \in C^{(m+1)}(M;a,b)} \left| \int_a^b \rho(x) f(x)\, dx - \sum_{k=1}^{n} C_k^{(n)} f(x_k^{(n)}) \right| = MN_n. \qquad (33)$$

Задачу, яку ми щойно розв'язали, можна інтерпретувати таким чином: задано клас функцій F і множину квадратурних формул S, для конкретної формули з S треба визначити величину $\sup_{f \in F} |R(f)|$, де $R(f)$ — залишковий член квадратурної формули. Формула (33) дає розв'язок цієї задачі для класу функцій $C^{(m+1)}(M; a, b)$ і квадратурної формули вигляду (1).

Істотно поставити також іншу задачу: знайти квадратурну формулу з S, на якій досягається

$$\inf_S \sup_F |R(f)|.$$

Розв'язок цієї задачі розглянемо в п. 5.7.

5.6. Апостеріорні оцінки похибки квадратурних формул

Кожна з формул прямокутників і трапецій окремо, як випливає з результатів попереднього параграфа, поступаються за величиною похибки перед формулою Сімпсона при інтегруванні гладких функцій. Однак якщо $f''(x)$ не змінює знака на $[a, b]$, то формули прямокутників і трапецій дають двосторонні наближення до шуканого інтеграла, бо їхні залишкові члени мають протилежні знаки. Наприклад, якщо $f'' < 0$, то $I_h^{\text{тр}} < I < I_h^{\text{пр}}$, де $I = \int_a^b f(x)\, dx$. Природно покласти $I \approx \approx (I_h^{\text{пр}} + I_h^{\text{тр}})/2$.

Тоді дістаємо таку апостеріорну оцінку похибки:

$$\left| I - \frac{I_h^{\text{пр}} + I_h^{\text{тр}}}{2} \right| \leqslant \frac{I_h^{\text{пр}} - I_h^{\text{тр}}}{2}. \qquad (1)$$

Далі помічаємо, що оцінка справедлива не тільки тоді, коли $f''(x)$ зберігає знак. Із формул (7) і (7′) з п. 5.5 випливає, що при $f \in C^2[a, b]$ формули прямокутників мають похибку порядку $O(h^2)$, причому якщо навіть функція f має більш високу гладкість, наприклад, $f \in C^4[a, b]$, то покращити ці оцінки за порядком h неможливо. Це випливає з оцінки знизу (31) (п. 5.5), яка матиме порядок $O(h^2)$, для тих $f \in C^4[a, b]$, для яких $|f''| > 0$. Однак при $f \in C^4[a, b]$ замість (7) і (7′) з п. 5.5 можна дістати в деякому смислі більш змістовні оцінки.

Нехай $F(x)$ — первісна функції $f(x)$, тобто $F(x) = \int_0^x f(t)\, dt$, $F_{\pm\frac{1}{2}} = F\left(\pm\frac{h}{2}\right)$. Тоді за умови $f \in C^4\left[-\frac{h}{2}, \frac{h}{2}\right]$ маємо

$$F_{\pm 1/2} = \pm\frac{h}{2} f_0 + \frac{h^2}{8} f_0' \pm \frac{h^3}{48} f_0''' + \frac{h^3}{384} f_0''' \pm \frac{h^5}{3840} f^{(4)}(\eta_\pm),$$

$$-\frac{h}{2} \leqslant \eta_- \leqslant \eta_+ \leqslant \frac{h}{2}.$$

Звідси

$$\int_{-h/2}^{h/2} f(x)\, dx = F_{\frac{1}{2}} - F_{-\frac{1}{2}} = hf_0 + \frac{h^3}{24} f_0'' + \frac{h^5}{1920} f^{(4)}(\eta), \qquad (2)$$

$$|\eta| \leqslant \frac{h}{2}.$$

Виконавши лінійну заміну змінних, дістаємо рівність

$$\int_{x_i}^{x_{i+1}} f(x)\, dx = hf_{i+\frac{1}{2}} + \frac{h^3}{24} f''_{i+\frac{1}{2}} + \frac{h^5}{1920} f^{(4)}(\eta_i), \qquad (3)$$

де $x_i = a + ih$, $h = (b - a)/N$, $f_{i+1/2}^{(k)} = f^{(k)}\left(a + \left(i + \frac{1}{2}\right)h\right)$, $\eta_i \in [x_i, x_{i+1}]$, $i = \overline{0, N-1}$. Підсумовуючи (3) по i від 0 до $N - 1$, маємо

$$I = \int_a^b f(x)\, dx = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)\, dx = I_h^{\text{пр}} + \frac{h^2}{24}\left[h \sum_{i=0}^{N-1} f''_{i+1/2}\right] +$$

$$+ h^4 \frac{b-a}{1920} f^{(4)}(\eta), \quad \eta \in [a, b]. \qquad (4)$$

Але за формулою середніх прямокутників

$$\int_a^b f''(x)\, dx = h \sum_{i=0}^{N-1} f''_{i+\frac{1}{2}} + h^2 \frac{b-a}{24} f^{(4)}(\xi), \quad \xi \in [a, b]. \qquad (5)$$

Із (4), (5) випливає

$$I = I_h^{\text{пр}} + Ch^2 + O(h^4), \qquad (6)$$

де

$$C = \frac{1}{24} \int_a^b f''(x)\, dx \qquad (7)$$

— стала, яка не залежить від h.

Величина Ch^2 в (6) називається *головною частиною похибки формули прямокутників.*

Вправа 1. Показати, що у випадку $f \in C^4[a, b]$

$$I = I_h^{\text{пр}} + C_1 h^2 + O(h^4), \tag{8}$$

де

$$C_1 = -\frac{1}{12}\int_a^b f''(x)\,dx. \tag{9}$$

Із співвідношень (6), (8) випливає, що у випадку $\int_a^b f''(x)\,dx \neq 0$, $f \in C^4[a, b]$, при досить малому h формули середніх прямокутників і трапецій дають двосторонні наближення інтеграла, навіть якщо f'' не зберігає знак на проміжку $[a, b]$.

Вправа 2. Показати, що у випадку $f \in C^6[a, b]$ має місце співвідношення

$$I = I_h^C + Ch^4 + O(h^6), \tag{10}$$

де C — стала, незалежна від h.

Згідно із співвідношеннями (6), (8), (10) аналогічно 2.10.1 для апостеріорної оцінки похибки чисельного інтегрування і уточнення результату (по порядку h) можна скористатися формулами Рунге (інколи кажуть правилом Рунге). Всі формули (6), (8), (10) можна записати у вигляді

$$I = I_h + Ch^k + O(h^{k+2}), \tag{11}$$

де C не залежить від h, $k = 2$ для формул середніх прямокутників і трапецій і $k = 4$ для формули Сімпсона. Вважається, що $f \in C^{k+2}[a, b]$. Тоді похибка чисельного інтегрування оцінюється за правилом Рунге таким чином

$$I - I_{h/2} \approx \frac{I_{h/2} - I_h}{2^k - 1}, \tag{12}$$

а уточнення за Річардсоном точного значення інтеграла I обчислюється за формулою

$$I_h^* = \frac{2^k I_{h/2} - I_h}{2^k - 1} \tag{13}$$

з похибкою $I - I_h^* = O(h^{k+2})$.

Зауваження 1. Формули (12), (13) можна застосовувати у випадку, коли $C \neq 0$. На практиці підтвердженням умови $C \neq 0$ є виконання нерівності

$$\left|2^k \frac{I_h - I_{h/2}}{I_{2h} - I_h} - 1\right| < 0.1. \tag{14}$$

Нерівність (14) може не виконуватися з таких причин: а) h велике настільки, що впливає складова $O(h^{k+2})$, яку ми відкидаємо; б) h занадто мале і впливають похибки заокруглень при обчисленні на реальній ЕОМ; в) $C = 0$ або близьке до нуля.

Зауваження 2. Формули трапецій і Сімпсона зручні тим, що при переході від h до $h/2$ всі обчислені раніше значення функції можуть бути використані.

Зауваження 3. За умови $f \in C^{2n+2}[a, b]$ співвідношення (11) справедливе і для ускладненої формули Гаусса, в якій $h = 1/N$, $k = 2n$.

5.7. Квадратурні формули з найкращою оцінкою залишкового члена на класах функцій

Нехай F — клас функцій, S — клас квадратурних формул для наближеного обчислення інтегралів від функцій з F. Розглянемо задачу: знайти формулу з S, на якій досягається

$$\inf_S \sup_{f \in F} |R(f)|.$$

Таку формулу, якщо вона не дуже відрізняється від інших формул класу S за обсягом обчислювальної роботи, доцільно використати при створенні математичного забезпечення ЕОМ. У цьому разі при розробці програм чисельного інтегрування доцільно орієнтуватися на досить широкий клас функцій F, які можуть поступити на вхід програми.

При цьому природно використати ту квадратурну формулу, котра навіть на «поганих» представниках класу F (для яких досягається $\sup_{f \in F} |R(f)|$) дає найменшу можливу похибку.

Розглянемо, як вирішується поставлена вище задача для деяких класів функцій.

5.7.1. Оцінки на класі функцій $C^{(1)}(L)$. Нехай $C^{(1)}(L)$ — клас неперервних функцій, які мають кусочно-неперервні похідні на відрізку $[0, 1]$ і таких, що $|f'(x)| \leqslant L$, $L > 1$. Нехай S — множина квадратурних формул виду

$$\int_0^1 f(x)\,dx \approx \sum_{k=1}^n C_k^{(n)} f(x_k^{(n)}), \quad x_k^{(n)} \in [0, 1], \tag{1}$$

в фіксованим n, які точні у випадку $f(x) = \text{const}$, тобто

$$\sum_{k=1}^n C_k^{(n)} = 1. \tag{2}$$

Для залишкового члена $R_n(f)$ квадратурної формули (1) справед-

лива оцінка

$$\sup_{f\in C^{(1)}(L)} |R_n(f)| = L\int_0^1 |K_n(y)|\,dy, \qquad (3)$$

де

$$K_n(y) = -y + \sum_{i=1}^{k} C_i^{(n)}, \quad y\in(x_k^{(n)}, x_{k+1}^{(n)}), \quad k = 0, 1, \ldots, n \qquad (4)$$

(див. 5.5). Тут $x_0^{(n)} = 0$, $x_{n+1}^{(n)} = 1$ і сума буде рівною 0 при $k = 0$.

Дійсно,

$$f(x) = f(0) + \int_0^x f'(y)\,dy = f(0) + \int_0^1 K_0(x, y) f'(y)\,dy,$$

де $K_0(x, y) = \begin{cases} 1, & y\leqslant x, \\ 0, & y > x. \end{cases}$ Оскільки квадратурна формула точна для многочленів нульового степеня, то

$$R_n(f) = \int_0^1\int_0^1 K_0(x, y) f'(y)\,dydx - \sum_{k=1}^{n} C_k^{(n)} \int_0^1 K_0(x_k^{(n)}, y) f'(y)\,dy =$$

$$= \int_0^1\int_0^1 K_0(x, y) f'(y)\,dxdy - \int_0^1 \sum_{j=1}^{n} C_j^{(n)} K_0(x_j^{(n)}, y) f'(y)\,dy =$$

$$= \int_0^1 K_n(y) f'(y)\,dy,$$

де

$$K_n(y) = \int_0^1 K_0(x, y)\,dx - \sum_{j=1}^{n} C_j^{(n)} K_0(x_j^{(n)}, y).$$

Нехай $y\in(x_j^{(n)}, x_{j+1}^{(n)})$. Тоді

$$\int_0^1 K_0(x, y)\,dx = \int_y^1 dx = 1 - y,$$

$$\sum_{i=1}^{n} C_i^{(n)} K_0(x_i^{(n)}, y) = \sum_{i=j+1}^{n} C_i^{(n)} = 1 - \sum_{i=1}^{j} C_i^{(n)}$$

і, таким чином,

$$K_n(y) = \sum_{i=1}^{j} C_i^{(n)} - y, \quad y\in(x_j^{(n)}, x_{j+1}^{(n)}).$$

Тобто

$$|R_n(f)| \leqslant L\int_0^1 |K_n(y)|\,dy.$$

Якщо припустити, що $f(x) = g(x)$, де $g'(x) = L\,\operatorname{sign} K_n(x)$, то

$$|R_n(g)| = L\int_0^1 |K_n(y)|\,dy,$$

тобто

$$\sup_{f\in C^{(1)}(L)} |R_n(f)| = L\int_0^1 |K_n(y)|\,dy.$$

Щоб побудувати найкращу квадратурну формулу виду (1) для класу $C^{(1)}(L)$, треба $x_k^{(n)}$ і $C_k^{(n)}$ вибрати так, щоб за умови (2) інтеграл в правій частині (3) мав мінімальне значення.

Припустивши, що $\xi_k = \sum_{i=1}^{k} C_i^{(n)}$, і використовуючи (4), знаходимо

$$\int_0^1 |K_n(t)|\,dt = \int_0^{x_1^{(n)}} |K_n(t)|\,dt + \sum_{k=1}^{n-1}\int_{x_k^{(n)}}^{x_{k+1}^{(n)}} |K_n(t)|\,dt +$$

$$+ \int_{x_k^{(n)}}^{1} |K_n(t)|\,dt = \int_0^{x_1^{(n)}} |t|\,dt + \sum_{k=1}^{n-1}\int_{x_k^{(n)}}^{x_{k+1}^{(n)}} |\xi_k - t|\,dt +$$

$$+ \int_{x_n^{(n)}}^{1} |t-1|\,dt = \frac{(x_1^{(n)})^2}{2} + \sum_{k=1}^{n-1}\int_{x_k^{(n)}}^{x_{k+1}^{(n)}} |\xi_k - t|\,dt + \frac{(1-x_n^{(n)})^2}{2}. \qquad (5)$$

Безпосереднім інтегруванням знаходимо

$$F(\xi_k) \equiv \int_{x_k^{(n)}}^{x_{k+1}^{(n)}} |\xi_k - t|\,dt =$$

$$= \left\{ \begin{array}{ll} \dfrac{(x_{k+1}^{(n)} - x_k^{(n)})\,(x_{k+1}^{(n)} + x_k^{(n)} - 2\xi_k)}{2}, & \xi_k < x_k^{(n)} \\ \dfrac{(x_{k+1}^{(n)} - \xi_k)^2 + (\xi_k - x_k^{(n)})^2}{2}, & x_k^{(n)} \leqslant \xi_k \leqslant x_{k+1}^{(n)}. \\ \dfrac{-(x_{k+1}^{(n)} - x_k^{(n)})\,(x_{k+1}^{(n)} + x_k^{(n)} - 2\xi_k)}{2}, & x_{k+1}^{(n)} < \xi_k \end{array} \right\}.$$

Неважко помітити, що

$$F'(\xi_k)=\begin{cases}-(x_{k+1}^{(n)}-x_k^{(n)}), & \text{якщо } \xi_k<x_k^{(n)},\\ \dfrac{4\xi_k-2(x_{k+1}^{(n)}+x_k^{(n)})}{2}, & x_k^{(n)}\leqslant\xi_k\leqslant x_{k+1}^{(n)},\\ x_{k+1}^{(n)}-x_k^{(n)}, & x_{k+1}^{(n)}<\xi_k\end{cases}$$

і $F'(\xi_k)<0$ при $\xi_k<\dfrac{x_{k+1}^{(n)}+x_k^{(n)}}{2}$, $F'(\xi_k)>0$ при $\xi_k>\dfrac{x_{k+1}^{(n)}+x_k^{(n)}}{2}$. Отже,

$$\min_{\xi_k}F(\xi_k)=F\left(\frac{x_{k+1}^{(n)}+x_k^{(n)}}{2}\right)=\frac{(x_{k+1}^{(n)}-x_k^{(n)})^2}{4}, \tag{6}$$

тому

$$F_1(x_1^{(n)},\ \dots,\ x_n^{(n)})\equiv\inf_{C_t^{(n)}}\int_0^1|K_0(t)|\,dt=\frac{(x_1^{(n)})^2}{2}+$$

$$+\sum_{k=1}^{n-1}\frac{(x_{k+1}^{(n)}-x_k^{(n)})^2}{4}+\frac{(1-x_n^{(n)})^2}{2}. \tag{7}$$

Прирівнюючи нулю похідні $\partial F_1/\partial x_i^{(n)}$, дістаємо

$$\frac{3x_1-x_2}{2}=0,\quad \frac{3x_n-2-x_{n-1}}{2}=0,\quad \frac{x_{k+1}^{(n)}-2x_k^{(n)}+x_{k-1}^{(n)}}{2}=0, \tag{8}$$

$$k=\overline{1,\ n-1}.$$

Розв'язком цієї системи є

$$x_k^{(n)}=\frac{2k-1}{2n},\quad k=\overline{1,\ n}, \tag{9}$$

і при цих значеннях

$$F_1(x_1^{(n)},\ \dots,\ x_n^{(n)})=\frac{1}{4n}. \tag{10}$$

Ми знайшли точку екстремуму функції F_1, але це не значить, що ми визначили її нижню грань, бо F_1 розглядається в області $0\leqslant x_1^{(n)}<x_2^{(n)}<\dots<x_n^{(n)}\leqslant 1$ і $\inf F_1$ може досягатися на границі області.

Перевіримо, чи значення (10) дійсно мінімальне. Для цього в нерівності Коші — Буняковського

$$\left|\sum_{j=1}^{2n}a_jb_j\right|\leqslant\left(\sum_{j=1}^{2n}a_j^2\right)^{1/2}\left(\sum_{j=1}^{2n}b_j^2\right)^{1/2}$$

покладемо $a_j\equiv 1$, $j=\overline{1,\ 2n}$, $b_1=x_1^{(n)}$, $b_2=b_3=\dfrac{x_2^{(n)}-x_1^{(n)}}{2}$, $b_4=b_5=\dfrac{x_3^{(n)}-x_2^{(n)}}{2}$, $\dots$, $b_{2n-2}=b_{2n-1}=\dfrac{x_n^{(n)}-x_{n-1}^{(n)}}{2}$, $b_{2n}=1-x_n^{(n)}$.

Очевидно, що $\sum\limits_{j=1}^{2n}a_jb_j=1$. В результаті

$$1\leqslant 2n\sum_{j=1}^{2n}b_j^2$$

або

$$\frac{1}{4n}\leqslant\frac{1}{2}\left(\sum_{j=1}^{2n}b_j^2\right)=F_1(x_1^{(n)},\ \dots,\ x_n^{(n)}),$$

а це означає, що

$$\inf_{x_j^{(n)}}F_1(x_1^{(n)},\ \dots,\ x_n^{(n)})=\frac{1}{4n}. \tag{11}$$

Залишилося обчислити $C_k^{(n)}$. Маємо

$$C_k^{(n)}=\xi_k-\xi_{k-1}=\frac{x_{k+1}^{(n)}+x_k^{(n)}-x_k^{(n)}-x_{k-1}^{(n)}}{2}=$$

$$=\frac{\frac{2k+1}{2n}-\frac{2k-3}{2n}}{2}=\frac{4}{4n}=\frac{1}{n},\quad k=\overline{1,\ n}. \tag{12}$$

Таким чином, оптимальною на класі $C^{(1)}(L)$ є квадратурна формула

$$\int_0^1 f(x)\,dx\approx\frac{1}{n}\sum_{k=1}^{n}f\left(\frac{2k-1}{n}\right), \tag{13}$$

тобто формула середніх прямокутників. Точна (тобто така, що досягається) оцінка залишкового члена цієї формули на класі $C^{(1)}(L)$ має вигляд

$$|R_n(f)|\leqslant\frac{L}{4n}. \tag{14}$$

Нерівність (14) показує, що по порядку h ($h=1/n$) оцінки з 5.5 для формули прямокутників на класі $C^{(1)}(L)$ не можуть бути покращені.

5.7.2. Оцінки на класі функцій $W_2^q(L)$. Нагадаємо, що через $W_2^q(L)$ позначається клас функцій, неперервних на відрізку $[0,\ 1]$ разом зі своїми похідними до порядку $q-1$ включно, які мають інтегровну з квадратом похідну порядку q, причому

$$\int_0^1[f^{(q)}(x)]^2\,dx\leqslant L^2. \tag{15}$$

Кожну функцію цього класу можна подати за формулою Тейлора із залишковим членом в інтегральній формі

$$f(x)=\sum_{p=0}^{q-1}\frac{f^{(p)}(0)}{p!}x^p+\int_0^1\frac{[(x-t)_+]^{q-1}}{(q-1)!}f^{(q)}(t)\,dt. \tag{16}$$

Найкращу квадратурну формулу для класу $W_2^q(L)$ шукатимемо серед формул виду

$$Q(f)\equiv\int_0^1 f(x)\,dx\approx\sum_{k=1}^n C_k^{(n)}f(x_k^{(n)})\equiv\tilde{Q}_n(f), \tag{17}$$

які є точними на многочленах до $(q-1)$-го степеня включно. Залишковий член має вигляд (5′) з ядром $K_n(t)$ виду (1.4′). За допомогою нерівності Коші — Буняковського знаходимо

$$|R_n(f)|=\left|\int_0^1 f^{(q)}(t)K_n(t)\,dt\right|\leqslant\|f^{(q)}\|_{L_2(0,1)}\|K_n\|_{L_2(0,1)}=L\|K_n\|_{L_2(0,1)}. \tag{18}$$

Розглянемо функцію $\varphi(t)$ таку, що

$$\varphi^{(q)}(t)=L\frac{K_n(t)}{\|K_n\|_{L_2(0,1)}}\in L_2(0,\ 1).$$

Оскільки очевидно, що $\|\varphi^{(q)}\|_{L_2(0,1)}=L$, то

$$\varphi\in W_2^q(L),\quad |R_n(\varphi)|=L\|K_n\|_{L_2(0,1)}.$$

Це означає, що

$$\sup_{f\in W_2^q(L)}|R_n(f)|=L\|K_n\|_{L_2(0,1)}. \tag{19}$$

Формулу виду (17), на якій досягається

$$\inf_{x_i^{(n)},C_i^{(n)}}\|K_n\|_{L_2(0,1)}, \tag{20}$$

називатимемо оптимальною квадратурною формулою на класі $W_2^q(L)$. Якщо вузли $x_i^{(n)}$, $i=\overline{1,\ n}$, задано, то можна розглядати таку задачу: знайти квадратурну формулу виду (1), тобто знайти $C_k^{(n)}$, $k=\overline{1,\ n}$, для якої досягається

$$L\inf_{C_i}\|K_n\|_{L_2(0,1)}=\inf_{C_i^{(n)}}\sup_{f\in W_2^q(L)}|R_n(f)|. \tag{21}$$

Таку квадратурну формулу називатимемо оптимальною порядку q в розумінні Сарда.

Раніше ми розглядали квадратурні формули, точні на многочленах до степеня $q-1$ включно. Якщо $n>q$, то коефіцієнти $C_k^{(n)}$ такої формули визначаються неоднозначно. Покажемо, що якщо вимагати, щоб формула (17) була точною для будь-якої сплайн-функції порядку q (множину таких функцій, побудованих на сітці $x_k^{(n)}$, $k=\overline{1,\ n}$, ми позначили S), то її коефіцієнти визначаються однозначно.

Дійсно, фундаментальні сплайн-функції $\sigma_i(x)$, $i=\overline{1,\ n}$, утворюють базис простору S. Нагадаємо, що $\sigma_i(x)$ задовольняє умови

$$\sigma_i(x_j^{(n)})=\sigma_{ij},\quad i,\ j=\overline{1,\ n}.$$

Будемо вимагати, щоб

$$\tilde{Q}_n(\sigma)=Q(\sigma)\ \forall\,\sigma\in S,\ \sigma(x)=\sum_{i=1}^n\alpha_i\sigma_i(x).$$

Тоді

$$\sum_{i=1}^n C_i^{(n)}\sigma(x_i)=\int_0^1\sigma(t)\,dt,$$

або

$$\sum_{i=1}^n C_i^{(n)}\sum_{k=1}^n\alpha_k\sigma_k(x_i)=\int_0^1\sum_{k=1}^n\alpha_k\sigma_k(t)\,dt.$$

Остання рівність перепишеться у вигляді

$$\sum_{k=1}^n\alpha_k\left(\sum_{i=1}^n C_i^{(n)}\sigma_k(x_i)\right)=\sum_{k=1}^n\alpha_k\left(\int_0^1\sigma_k(t)\,dt\right),$$

і оскільки вона має бути справедливою для будь-яких α_k, то

$$\sum_{i=1}^n C_i^{(n)}\sigma_k(x_i)=\int_0^1\sigma_k(t)\,dt,\quad k=\overline{1,\ n},$$

або

$$C_k^{(n)}=\int_0^1\sigma_k(t)\,dt,\quad k=\overline{1,\ n},$$

тобто коефіцієнти $C_k^{(n)}$ визначаються однозначно. При цьому обчислення $\tilde{Q}_n(f)$ зводиться до знаходження інтерполяційного сплайна $\sigma\in S$ такого, що $\sigma(x_i^{(n)})=f(x_i^{(n)})$. Дійсно,

$$\tilde{Q}_n(f)=\sum_{i=1}^n C_i^{(n)}f(x_i^{(n)})=\sum_{i=1}^n Q(\sigma_i)f(x_i^{(n)})=$$

$$=Q\left(\sum_{i=1}^n y_i\sigma_i(x)\right)=Q(\sigma(x)).$$

Таким чином, квадратурна формула $\tilde{Q}_n(f)$ є точною на сплайн-функціях порядку q, якщо

$$\lambda_k = Q(\sigma_k) = \int_0^1 \sigma_k(t)\,dt, \quad k = \overline{1, n}.$$

Справедливе таке твердження (теорема Шонберга).

Теорема. *Нехай функціонал $Q_n(f)$ визначений на множині $W_2^q(L)$. Тоді єдина квадратурна формула, точна на сплайн-функціях порядку q, співпадає з єдиною квадратурною формулою порядку q, оптимальною в розумінні Сарда.*

Доведення. Нехай $\tilde{Q}_n^*(f)$ — квадратурна формула вигляду (17) точна на сплайн-функціях порядку q, $\tilde{Q}_n(f)$ — будь-яка інша квадратурна формула, точна на многочленах степеня $q-1$. Нехай

$$R^*(f) = \tilde{Q}_n^*(f) - Q(f), \quad R(f) = \tilde{Q}_n(f) - Q(f),$$

тоді в силу (18)

$$R^*(f) = \int_0^1 f^{(q)}(x)\,K_n^*(x)\,dx, \quad R(f) = \int_0^1 f^{(q)}(x)\,K_n(x)\,dx,$$

де функції K_n^*, K_n називаються ядрами порядку q інтегрального представлення похибки і мають вигляд (4′) з п. 5.5. Отже,

$$\tilde{Q}_n(f) - \tilde{Q}_n^*(f) = \sum_{k=1}^n \mu_k f(x_k^{(n)}) =$$

$$= R(f) - R^*(f) = \int_0^1 M_n(x) f^{(q)}(x)\,dx, \tag{22}$$

де

$$M_n(x) = K_n(x) - K_n^*(x) = \sum_{k=1}^n \mu_k \frac{[(x_k^{(n)} - x)_+]^{q-1}}{(q-1)!}, \tag{23}$$

$$\mu_k = C_k^{(n)} - C_k^{(n)*}.$$

Неважко помітити, що

$$\sum_{k=1}^n \mu_k [x_k^{(n)}]^i = \sum_{k=1}^n C_k^{(n)} [x_k^{(n)}]^i - \sum_{k=1}^n C_k^{(n)*} [x_k^{(n)}]^i =$$

$$= \int_0^1 x^i dx - \int_0^1 x^i dx = 0, \quad i = \overline{0, q-1}, \tag{24}$$

оскільки $\pi_{q-1} \subset S$.

Покажемо, що за умов (24) функцію $M_n(x)$ можна подати у вигляді

$$M_n(x) = (-1)^q \sum_{k=1}^n \mu_k \frac{[(x - x_k^{(n)})_+]^{q-1}}{(q-1)!}. \tag{25}$$

Дійсно, нехай $x \in [x_j, x_{j+1}]$, тоді з виразу (23) маємо

$$M_n(x) = \frac{1}{(q-1)!} \sum_{k=j+1}^n \mu_k (x_k^{(n)} - x)^{q-1} =$$

$$= \frac{(-1)^{q-1}}{(q-1)!} \sum_{k=j+1}^n \mu_k (x - x_k^{(n)})^{q-1} =$$

$$= \frac{(-1)^{q-1}}{(q-1)!} \left[\sum_{k=1}^n \mu_k (x - x_k^{(n)})^{q-1} - \sum_{k=1}^j \mu_k (x - x_k^{(n)})^{q-1} \right].$$

Перший доданок у квадратних дужках на основі рівностей (24) дорівнює нулю, бо (24) означає, що $\sum_{k=1}^n \mu_k p(x_k^{(n)}) = 0, \forall\, p \in \pi_{q-1}$. Отже,

$$M_n(x) = \frac{(-1)^q}{(q-1)!} \sum_{k=1}^j \mu_k (x - x_k^{(n)})^{q-1} = \frac{(-1)^q}{(q-1)!} \sum_{k=1}^n \mu_k [(x - x_k^{(n)})_+]^{q-1}$$

і рівність (25) доведено.

Далі, відповідно до (2) з п. 4.3, маємо

$$s(x) = (-1)^q \sum_{k=1}^n \mu_k \frac{[(x - x_k^{(n)})_+]^{2q-1}}{(2q-1)!} \in S$$

і очевидно, що $S^{(q)}(x) = M_n(x)$. Оскільки квадратурна формула $\tilde{Q}_n^*(f)$ точна на всіх $s \in S$, то

$$R^*(s) = \int_0^1 s^{(q)}(x)\,K_n^*(x)\,dx = \int_0^1 M_n(x)\,K_n^*(x)\,dx =$$

$$= \int_0^1 (K_n(x) - K_n^*(x))\,K_n^*(x)\,dx = 0.$$

Звідси випливає, що

$$\int_0^1 [K_n(x)]^2 dx = \int_0^1 [K_n(x) - K_n^*(x) + K_n^*(x)]^2 dx =$$

$$= \int_0^1 M_n^2(x)\,dx + 2\int_0^1 M_n(x)\,K_n^*(x)\,dx + \int_0^1 [K_n^*(x)]^2 dx =$$

$$= \int_0^1 M_n^2(x)\,dx + \int_0^1 [K_n^*(x)]^2 dx \geqslant \int_0^1 [K_n^*(x)]^2\,dx,$$

причому рівність справджується тоді і тільки тоді, коли $M_n(x) \equiv 0$, тобто при $C_k^{(n)} = C_k^{(n)*}$, $k = \overline{1, n}$. Теорему доведено.

З теореми випливає, що квадратурну формулу, оптимальну на класі функцій $W_2^q(L)$, слід шукати серед квадратурних формул оптимальних порядку q в розумінні Сарда або, іншими словами, серед квадратурних формул, точних на сплайн-функціях порядку q_1. Для визначення явного виду її необхідно ще мінімізувати $\int_0^1 [K_n^*(x)]^2\,dx$ за всіма можливими наборами вузлів $x_k^{(n)} \in [0, 1]$, $k = \overline{1, n}$.

Приклад. Нехай $q = 1$, тоді оптимальну квадратурну формулу на класі $W_2^1(L)$ слід шукати серед формул вигляду (17), точних на сплайн-функціях першого порядку, які подаються як лінійні комбінації фундаментальних сплайн-функцій першого порядку (рис. 20). З викладеного вище випливає, що коефіцієнти $C_k^{(n)}$ оптимальної квадратурної формули задають формулами

$$C_k^{(n)} = \int_0^1 s_k(x)\,dx = \int_{x_{k-1}^{(n)}}^{x_k^{(n)}} \frac{x - x_{k-1}^{(n)}}{x_k^{(n)} - x_{k-1}^{(n)}}\,dx + \int_{x_k^{(n)}}^{x_{k+1}^{(n)}} \frac{x - x_{k+1}^{(n)}}{x_k^{(n)} - x_{k+1}^{(n)}}\,dx =$$

$$= \frac{x_{k+1}^{(n)} - x_{k-1}^{(n)}}{2}, \quad k = \overline{1, n};\ x_0^{(n)} = 0,\ x_{n+1}^{n} = 1. \tag{26}$$

Отже, якщо $x \in [x_j^{(n)}, x_{j+1}^{(n)}]$, то

$$K_n^*(x) = 1 - x - \sum_{k=1}^{n} C_k^{(n)} [(x_k^{(n)} - x)^+]^0 = 1 - x - \sum_{k=j+1}^{n} C_k^{(n)} =$$

$$= \sum_{k=1}^{j} C_k^{(n)} - x = \frac{1}{2}[x_{j+1}^{(n)} + x_j^{(n)} - x_1^{(n)}] - x$$

і

$$\int_0^1 [K_n^*(x)]^2\,dx = \sum_{j=0}^{n} \int_{x_j^{(n)}}^{x_{j+1}^{(n)}} [K_n^*(x)]^2\,dx =$$

$$= \frac{[x_1^{(n)}]^3}{3} + \frac{1}{12}\sum_{j=1}^{n-1} [x_{j+1}^{(n)} - x_j^{(n)}]^3 + \frac{[1 - x_n^{(n)}]^3}{3}. \tag{27}$$

Аналогічно тому, як це робилося в п. 5.7.1, можна визначити мінімум виразу (27), який дорівнює $\frac{1}{12n^2}$ і досягається при $x_k^{(n)} = \frac{2k-1}{n}$. Таким чином, дійдемо висновку, що оптимальною на класі $W_2^1(L)$ буде формула середніх прямокутників

$$\int_0^1 f(x)\,dx \approx \frac{1}{n}\sum_{k=1}^{n} f\left(\frac{2k-1}{2n}\right),$$

і для неї справедлива оцінка

$$\sup_{f \in W_2^1(L)} |R_n(f)| = \left| \int_0^1 f(x)\,dx - \frac{1}{n}\sum_{k=1}^{n} f\left(\frac{2k-1}{2n}\right) \right| = \frac{L}{2\sqrt{3}\,n}. \tag{28}$$

Звідси видно, що оцінка (5.5.23) за порядком h не може бути покращена.

5.8. Збіжність квадратурних формул, що не містять похідних

З оцінок, знайдених у 5.5, 5.7, випливає, що для конкретних квадратурних формул у разі належності підінтегральної функції до певного класу гладкості має місце збіжність квадратурного процесу при $h \to 0$ $(n \to \infty)$, тобто

$$\lim_{h \to 0} Q_n(f) = \lim_{n \to \infty} Q_n(f) = Q(f).$$

Розглянемо квадратурну формулу загального виду (5), п. 5.1 для функцій $f(x) \in C[a, b]$. Вона визначається нескінченними трикутними матрицями абсцис $x_k \equiv x_k^{(n)} \in [a, b]$ і коефіцієнтів $C_k^{(n)}$

$$X = \begin{bmatrix} x_0^{(0)} & & & \\ x_0^{(1)} & x_1^{(1)} & & \\ \dots & \dots & & \\ x_0^{(n)} & x_1^{(n)} & \dots & x_n^{(n)} \\ \dots & \dots & & \end{bmatrix}, \quad C = \begin{bmatrix} C_0^{(0)} & & & \\ C_0^{(1)} & C_1^{(1)} & & \\ \dots & \dots & & \\ C_0^{(n)} & C_1^{(n)} & \dots & C_n^{(n)} \\ \dots & \dots & & \end{bmatrix}. \tag{1}$$

Означення. Квадратурна формула $Q_n(f)$, яка відповідає n-м рядкам матриць X і C і яка має вигляд

$$Q_n(f) = \sum_{k=0}^{n} C_k^{(n)} f(x_k^{(n)}), \tag{2}$$

називається *збіжною*, якщо

$$\lim_{n \to \infty} Q_n(f) = \lim_{n \to \infty} \sum_{k=0}^{n} C_k^{(n)} f(x_k^{(n)}) = Q(f), \tag{3}$$

де $Q(f) = \int_a^b \rho(x) f(x)\,dx$.

Розглядатимемо квадратурну формулу $Q_n(f)$ як лінійний функціонал в банаховому просторі $C[a, b]$. Він є неперервним, і аналогічно (7) з п. 2.4 можна довести, що його норма визначається формулою

$$\|Q_n\| = \sup_{\substack{f \in C[a,b] \\ f \neq 0}} \frac{\|Q_n f\|}{\|f\|_{C[a,b]}} = \sup_{\substack{f \in C[a,b] \\ f \neq 0}} \frac{\left|\sum_{k=0}^{n} C_k^{(n)} f(x_k^{(n)})\right|}{\max_{x \in [a,b]} |f(x)|} = \sum_{k=0}^{n} |C_k^{(n)}|. \tag{4}$$

Справедлива така теорема.

Теорема 1. *Для того щоб для будь-якої функції* $f(x) \in C[a, b]$ *послідовність квадратурних формул* $Q_n(f)$, *побудована за трикутними матрицями* X *і* C, *збігалась, необхідно і достатньо, щоб*

1) $\lim\limits_{n\to\infty} Q_n(p) = Q(p)\ \forall\, p(x) \in \pi_m, \quad m = 0, 1, \ldots;$

2) *існувала стала* $M > 0$ *така, що*

$$\|Q_n\| = \sum_{k=0}^{n} |C^{(n)}| < M\ \forall\, n = 0, 1, \ldots. \tag{5}$$

Доведення. Відомо, що множина многочленів є всюди щільною підмножиною в $C[a, b]$ (теорема Вейєрштрасса). Тому твердження теореми 1 випливає з теореми Банаха — Штейнгауза.

Розглянемо два простих наслідки з теореми 1.

Теорема 2. *Для того щоб для будь-якої неперервної функції* $f(x) \in C[a, b]$ *інтерполяційна квадратурна формула збігалась, необхідне і достатнє виконання нерівностей* (5).

Доведення. Перша умова теореми 1 випливає з того, що формула інтерполяційна, а друга вимога теореми співпадає з умовою теореми, що доводиться. Таким чином, в силу твердження теореми 1 має місце і твердження теореми 2.

Теорема 3. *Якщо вагові коефіцієнти* $C_k^{(n)}$ *невід'ємні, то квадратурна формула* $Q_n(f)$, *яка є точною на сталій, збігається для будь-якої функції* $f(x) \in C[a, b]$ *тоді і лише тоді, коли вона збігається до будь-якого многочлена* $p(x) \in \pi_n$, $n = 0, 1, \ldots$.

Доведення. *Необхідність* очевидна. Доведемо *достатність*. Оскільки при $f(x) = 1$ має виконуватися граничне співвідношення

$$\lim_{n\to\infty} Q_n(1) = \lim_{n\to\infty} \sum_{k=0}^{n} C_k^{(n)} = \lim_{n\to\infty} \sum_{k=0}^{n} |C_k^{(n)}| = \lim_{n\to\infty} \|Q_n\| = 1,$$

то послідовність норм $\|Q_n\|$ має бути обмежена сталою, що не залежить від n, тобто виконується умова 2 теореми 1, що і доводить теорему 3.

Наслідок. *Квадратурна формула найвищого алгебраїчного ступеня точності збігається для будь-якої функції* $f(x) \in C[a, b]$.

Доведення. При $\rho(x) \geqslant 0$ квадратурна формула найвищого алгебраїчного ступеня точності може бути побудована для всіх n і всі вагові коефіцієнти додатні (див. 5.3). Крім того ця формула збігається для будь-якого многочлена. В силу теореми 3 вона збігатиметься для будь-якої функції $f(x) \in C[a, b]$, що і треба було довести.

Наведемо без доведення дві теореми, які вказують на вплив абсцис квадратурних формул інтерполяційного типу на збіжність.

Теорема 4. *Якщо* $x_0^{(n)} = -1$, $x_n^{(n)} = 1$, *а інші абсциси* $x_k^{(n)}$, $k = \overline{1, n-1}$, *вибираються рівновіддаленими на* $[-1, 1]$, *то коефіцієнти* $C_k^{(n)}$ *формули інтерполяційного типу є такими, що величина* $\sum\limits_{k=0}^{n} |C_k^{(n)}|$ *не обмежена по* n. *Як наслідок існує функція* $g(x) \in C[a, b]$ *така, що*

$$\lim_{n\to\infty} Q_n(g) \neq Q(g).$$

Теорема 5. *Якщо за абсциси квадратурної формули вибрати нулі многочленів Чебишева першого роду* $T_{n+1}(x)$, *тобто*

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

або нулі многочленів Чебишева другого роду $U_{n+1}(x)$, *тобто*

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

то всі коефіцієнти $C_k^{(n)}$ *формули інтерполяційного типу будуть додатні і*

$$\lim_{n\to\infty} Q_n(f) = Q(f)\ \forall\, f \in C[a, b].$$

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. *Бабенко К. И.* Основы численного анализа.— М. : Наука, 1986.— 744 с.
2. *Бахвалов Н. С.* Численные методы.— М. : Наука, 1973.— 623 с.
3. *Бахвалов Н. С., Жидков Н. П., Кобельков Г. М.* Численные методы.— М. : Наука, 1987.— 598 с.
4. *Березин И. С., Жидков Н. П.* Методы вычислений: В 2 т.— М. : ГИФМЛ, 1962.— Т. 1.— 464 с.; Т. 2.— 639 с.
5. *Васильев Ф. П.* Численные методы решения экстремальных задач.— М. : Наука, 1980.— 518 с.
6. *Воеводин В. В., Кузнецов Ю. А.* Матрицы и вычисления.— М. : Наука, 1984.— 368 с.
7. *Волков Е. А.* Численные методы.— М. : Наука, 1987.— 256 с.
8. *Даугавет И. К.* Введение в теорию приближенных функций.— Л. : Изд-во Ленингр. ун-та, 1977.— 184 с.
9. *Иванов В. В.* Методы вычислений на ЭВМ: Справ. пособие.— К. : Наук. думка, 1986.— 583 с.
10. *Калиткин Н. Н.* Численные методы.— М. : Наука, 1978.— 512 с.
11. *Канторович Л. В., Акилов Г. П.* Функциональный анализ.— М. : Наука, 1984.— 750 с.
12. *Крылов В. И., Бобков В. В., Монастырный П. И.* Вычислительные методы: В 2 т.— М. : Наука, 1976, 1977.— Т. 1—2.
13. *Кнут Д.* Искусство программирования для ЭВМ: В 2 т.— М. : Мир, 1972.— Т. 2.
14. *Ляшко И. И., Макаров В. Л., Скоробагатько А. А.* Методы вычислений.— К. Вища шк. Головное изд-во, 1977.— 406 с.
15. *Макаров В. Л., Хлобыстов В. В.* Сплайн-аппроксимация функций.— М. : Высш. шк., 1983.— 80 с
16. *Натансон И. П.* Конструктивная теория функций.— М. : Гостехиздат, 1949.
17. *Никольский С. М.* Квадратурные формулы.— М. : Наука, 1988.— 363 с.
18. *Никифоров А. Ф., Уваров В. Б.* Специальные функции математической физики.— М. : Наука, 1978 — 319 с.
19. *Рабкин Е. Л., Шапиро Е. П.* Об одном расходящемся интерполяционном процессе // Изв. вузов. Сер. мат.— 1971.— № 8.— С. 103—110.
20. *Самарский А. А.* Введение в численные методы.— М. : Наука, 1982.— 272 с.
21. *Самарский А. А., Гулин А. В.* Устойчивость разностных схем.— М. : Наука, 1973.— 415 с.
22. *Самарский А. А., Гулин А. В.* Численные методы.— М. : Наука, 1982.— 429 с.
23. *Самарский А. А., Николаев Е. С.* Методы решения сеточных уравнений.— М. : Наука, 1978.— 591 с.
24. *Треногин В. А.* Функциональный анализ.— М. : Наука, 1980.— 495 с,
25. *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений.— М. : Мир, 1980.— 280 с.
26. *Deuflhard P., Hohmann A.* Numerische Mathematik. Eine algorithmisch orientierte Einführung.— Berlin ; New York : Walter de Gruyter, 1991.
27. *Isaakson E., Keller H. B.* Analysis of Numerical Methods.— New York : Wiley, 1966.
28. *Maeß G.* Vorlesungen über numerische Matematik. Analysis.— Berlin : Akademie — Verlag, 1988.
29. *Stoer J., Bulirsch R.* Numerische Mathematik.— Berlin : Springer — Verlag, 1990.
30. *Törnig W., Spellucci P.* Numerische Mathematik für Ingeniere und Physiker. Bd. 1. Lineare Algebra.— Berlin : Springer — Verlag, 1990.
31. *Ebenda.*— Bd. 2. Numerische Methoden der Analysis. Analysis of Numerical Methods.— Berlin : Springer — Verlag, 1990.

ЗМІСТ

Навчальне видання

Гаврилюк Іван Петрович
Макаров Володимир Леонідович

МЕТОДИ ОБЧИСЛЕНЬ

У двох частинах

ЧАСТИНА 1

Оправа художника *В. Г. Самсонова*
Художній редактор *С. В. Анненков*
Технічний редактор *Т. Г. Шепновська*
Коректор *С. Г. Чиркіна*

Здано до набору 08.10.94. Підписано до друку 16.03.95. Формат 60×84¹/₁₆. Папір друк. № 2. Гарнітура літературна. Високий друк. Умовн.-друк. арк. 21,39. Умовн. фарбовідб. 21,62. Обл.-вид. арк. 21,65. Вид. № 9817. Замовлення 4—1829.

Видавництво «Вища школа». 252054, Київ-54, вул. Гоголівська, 7.

Головне підприємство республіканського виробничого об'єднання «Поліграфкнига». 252057, Київ, вул. Довженка, 3.