Міністерство освіти і науки України
Чернівецький національний університет
імені Юрія Федьковича

Я.Й. Бігун

ЧИСЛОВІ МЕТОДИ

Навчальний посібник

Чернівці
Чернівецький національний університет
2018

УДК 519.6
ББК 22.193
Б 597

Друкується за ухвалою вченої ради
Чернівецького національного університету
імені Юрія Федьковича, протокол №13 від 26 грудня 2018 р.

Рецензенти:

Венгерський П.С., доктор фіз.-мат. наук, професор кафедри
інформаційних систем
(Львівський національний університет імені Івана Франка);
Руткас А.Г., доктор фіз.-мат. наук, професор кафедри
програмної інженерії
(Харківський національний університет радіоелектроніки);
Шевельова А.Є., доктор фіз.-мат. наук, професор
кафедри обчислювальної математики та математичної кібернетики
(Дніпровський національний університет імені Олеся Гончара).

Бігун Я.Й.

Б 597 **Числові методи:** навч. посібник / Я.Й. Бігун. – Чернівці: Чернівец. нац. ун-т, 2018. – 436 с.

У посібнику наведено елементи комп'ютерної арифметики, основні прямі та ітераційні методи розв'язування систем лінійних алгебраїчних рівнянь, ітераційні методи для нелінійних рівнянь і систем, звернуто увагу на метод Ньютона та його модифікації. Наведено однокрокові і багатокрокові методи розв'язування задачі Коші для звичайних диференціальних рівнянь, зокрема методи Рунге-Кутти, і на числове інтегрування жорстких систем. Розглянуто різницеві методи розв'язування лінійних і нелінійних звичайних диференціальних рівнянь із двоточковими крайовими умовами. Проілюстровано на прикладах розв'язування типових задач та наведено завдання для самостійного опрацювання матеріалу.

Для студентів, які навчаються за спеціальностями "Прикладна математика", "Комп'ютерні науки", "Системний аналіз", "Математика" та ін.

Бібліогр.: 108 найм.

УДК 519.6
ББК 22.193

© Чернівецький національний університет, 2018
© Бігун Я.Й., 2018

ISBN

Зміст

Список основних позначень і скорочених назв

$\forall$ – квантор всезагальності, $\exists$ – квантор існування

$:=$ – покласти за означенням або присвоїти

$n = \overline{1, N}$ – число n набуває послідовно значення від 1 до N включно

$[\cdot]^T$ –транспонована матриця або вектор

$A: X \to Y$ – оператор A відображає множину X на (або в) множину Y

R – множина дійсних чисел (числова пряма)

$[a,b]$, (a,b) – відрізок та інтервал з кінцями в точках $a, b \in R$

a – наближене значення величини a^*

$\Delta(a)$ – абсолютна похибки наближеного значення величини a^*

$\delta(a)$ – відносна похибка наближеного значення величини a^*

$\Delta^0(y)$, $\delta^0(y)$ – лінійні абсолютна і відносна оцінки похибки функції $y = y(x)$

$\|x\|_\infty$, $\|\mathrm{A}\|_\infty$ – кубічна норма (максимальна) вектора $x \in R^n$ і матриці $A(n,n)$

$\|x\|_1$, $\|\mathrm{A}\|_1$ – октаедрична норма (1-норма) вектора $x \in R^n$ і матриці $A(n,n)$

$\|x\|_2$, $\|\mathrm{A}\|_2$ – евклідова (сферична) норма вектора $x \in R^n$ і матриці $A(n,n)$

(x, y) – скалярний добуток елементів x та y евклідового простору R^n

$\rho(\mathrm{A})$ – спектральний радіус матриці A

$Q_n(x)$ – множина алгебраїчних многочленів степеня, що не перевищує n

$L_n(x)$ і $P_n(x)$ – інтерполяційні многочлени степеня n Лагранжа і Ньютона відповідно

$H_n(x)$ – алгебраїчний інтерполяційний многочлен Ерміта степеня n

$S_n(x)$ – інтерполяційний сплайн степеня n

$h_i = x_{i+1} - x_i$ – відстань (крок сітки) між вузлами x_{i+1} та x_i, $x_{i+1/2} = (x_i + x_{i+1})/2$

I_h – квадратурна формула для обчислення інтеграла I з кроком h

$y_{x,n}$ і $y_{\bar{x},n}$ – права і ліва різницева похідні сіткової функції y_h відповідно

$y_{\overset{\circ}{x}}$ – центральна різницева похідна сіткової функції y_h

$y_{\bar{x}x,n}$ – друга різницева похідна сіткової функції y_h

$y_n = y(x_n)$ – значення сіткової функції y_h в просторовій точці x_n

A_h – різницева апроксимація оператора A

$\psi_h^{(1)}$ – похибка апроксимації різницевої схеми

$C^n[a,b]$ – простір визначених на відрізку $[a,b]$ функцій, які мають неперервні похідні до n-го порядку включно

■ – ознака закінчення прикладу або доведення твердження

ГСП – головна складова похибки

ЗДР – звичайне диференціальне рівняння

МРКр – метод Рунге–Кутти порядку p

КФ – квадратурна формула

КФНАСТ – КФ найвищого алгебраїчного степеня точності

НМРК – неявні методи Рунге–Кутти

НСКН – найліпше середньоквадратичне наближення

СЛАР – система(и) лінійних алгебраїчних рівнянь

Передмова

Дослідження математичних моделей у різноманітних галузях і розв'язування переважної більшості прикладних задач неможливе без застосування числових методів. Тому студентам, як зі спеціальності прикладна математика, так і суміжних спеціальностей, у процесі комп'ютерного моделювання необхідно володіти достатньо повним арсеналом числових методів, уміти їх застосувати, аналізувати та модифікувати.

У багатьох закладах вищої освіти України видані підручники і посібники як із числових методів (обчислювальних, наближених), так і їх застосувань у різних галузях. Частина з них вийшла друком дещо раніше [15, 16, 39, 40], інші доступні переважно студентам ЗВО, в яких вони видавались [9, 27, 37, 43, 54, 55, 78-80, 82] та ін., і тираж їх досить обмежений.

Даний посібник ґрунтується на лекціях, які протягом багатьох років автор читає студентам різних спеціальностей факультету математики та інформатики в Чернівецькому національному університеті імені Юрія Федьковича. Складається посібник із 15 розділів, списку використаних джерел (108 найменувань), коротких відомостей про вчених, які згадуються у посібнику, та додатків.

Розділ 1 присвячений аналізу похибок й особливостям комп'ютерної арифметики, зокрема реалізації обчислювальних алгоритмів у системах із плаваючою крапкою. У наступних трьох розділах розглядаються прямі методи розв'язування систем лінійних алгебраїчних рівнянь, аналіз похибки розв'язку та застосування ітераційних методів відповідно. Звернуто увагу на особливості розв'язування СЛАР, а також на побудову розв'язків систем із прямокутними матрицями.

Ітераційні методи розв'язування нелінійних рівнянь розглянені в розділі 5. Наведено низку методів, основна увага звернена на метод Ньютона та його модифікації. Методам аналізу та наближеному обчисленню коренів алгебраїчних рівнянь присвячений розділ 6. У розділі 7 розглядаються ітераційні методи розв'язування систем нелінійних рівнянь і методи зведення їх до задач оптимізації.

Важливою і складною задачею обчислювальної математики є знаходження власних значень і власних векторів матриці. Методи знаходження спектра матриці, найбільших за модулем власних значень і відповідних власних векторів наведені в розділі 8.

У розділі 9 викладено методи наближення функцій однієї змінної інтерполяційними многочленами Лагража, Ньютона та Ерміта, лінійними та кубічними сплайнами, а також побудову найліпших середньоквадратичних наближень методом найменших квадратів.

Формули числового диференціювання (ФЧД) складають розділ 10. У наступному розділі розглянено квадратурні формули (КФ) обчислення визначених інтегралів і кубатурні формули, зокрема метод Монте-Карло, наближеного обчислення кратних інтегралів.

Розділи 12-15 присвячені числовому розв'язуванню звичайних диференціальних рівнянь (ЗДР). У розділі 12 детально розглянено явні одно-крокові методи Рунге–Кутти. Багатокрокові явні і неявні методи Адамса, Штермера та ін. наведені в розділі 13, проведено дослідження їх стійкості на модельному рівнянні.

Чисто неявні різницеві схеми і неявні методи Рунге–Кутти та огляд інших методів числового інтегрування жорстких систем ЗДР розглянено в розділі 14. В останньому розділі 15 наведено побудову та дослідження стійкості і збіжності різницевих схем для лінійних і нелінійних доточкових крайових задач. Показано побудову різницевих схем інтегро-інтерполяційним методом.

Кожен розділ містить розв'язування типових задач та низку завдань різної складності для самостійної роботи. У додатках наведено коефіцієнти явних методів Рунге–Кутти порядку 1-8 і 10 та деяких класів неявних однокрокових методів та багатокрокових РС.

Короткі відомості про вчених, які згадуються в посібнику і зробили внесок у розвиток числових методах, наведені в авторському показнику.

У даному посібнику використано матеріали з посібників, виданих у співавторстві з викладачами кафедри прикладної математики та інформаційних технологій Інессою Краснокутською (Березовською) – [7] та Лідією Сергєєвою – [8].

Автор висловлює щиру подяку завідувачу комп'ютерної лабораторії Наталії Романенко за постійну увагу до підготовки та комп'ютерну верстку посібника. Поліпшенню посібника посприяли побажання і зауваження доцентів кафедри прикладної математики та інформаційних технологій Василя Маценка й Інесси Краснокутської, за що їм автор висловлює свою вдячність. Автор вдячний Тетяні Книгніцькій і Ларисі Бігун за набір текстів і допомогу в оформленні, Анастасії Юрійчук – за виконання частини рисунків, а також рецензентам за прихильне відношення до моєї праці.

Посібник буде корисним студентам, які навчаються за спеціальностями прикладна математика, комп'ютерні науки, системний аналіз та інші, а також фахівцям, які у своїх дослідженнях використовують числові методи. Свої зауваження, пропозиції та запитання прошу надсилати за адресою електронної пошти *yaroslav.bihun@gmail.com.*

Розділ 1. Теорія похибок і комп'ютерна арифметика

Позиційні системи числення. Джерела похибок та їх класифікація. Абсолютна і відносна похибки. Похибка функції та похибки виконання арифметичних операцій. Заокруглення чисел. Особливості машинної арифметики для чисел у формі з фіксованою крапкою. Системи чисел з плаваючою крапкою, оцінка похибки заокруглення, похибка арифметичних операцій та особливості реалізації алгоритмів у цих системах.

Література [5, 6, 13, 14, 30, 32, 43, 73, 74, 83, 93]

1.1. Зображення дійсних чисел у позиційних системах

Стандартна позиційна система визначається основою $p \in N$, $p \geq 2$, і набором базисних чисел $\alpha_0, \alpha_1, ..., \alpha_{p-1}$, які задовольняють умови:

$$\alpha_k \in Z, \ |\alpha_k| \leq p-1, \ k = \overline{0, p-1}. \tag{1.1}$$

Зображення числа $x \in R$ в позиційній системі за основою p і з базисними числами $\alpha_0, \alpha_1, ..., \alpha_{p-1}$ набуває вигляду

$$x = \pm\left(b_n p^n + b_{n-1} p^{n-1} + ... + b_1 p + b_0 + b_{-1} p^{-1} + ...\right) \tag{1.2}$$

або в такому записі

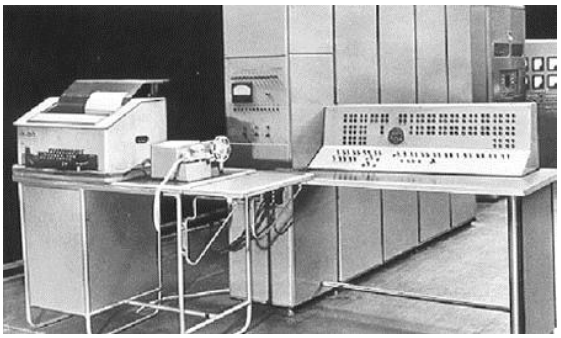
$$x = \pm\left(b_n, ..., b_1, b_0, ...\right),$$

де $b_i \in \left\{\alpha_0, ..., \alpha_{p-1}\right\}$, тобто є скінченним або нескінченним дробом за основою p. Найчастіше застосовуються системи за основою $p = 2, 8, 10$ або 16. Базисними числами двійкової системи служать цифри 0 і 1, десяткової системи – 0, 1, …, 9, наприклад $(3.8125)_{10} = 3 \cdot 10^0 + 8 \cdot 10^{-1} + 1 \cdot 10^{-2} + 2 \cdot 10^{-3} + 5 \cdot 10^{-4}$, $(11.1101)_2 = 1 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-2} + 0 \cdot 2^{-3} + 1 \cdot 2^{-4}$. Якщо $p = 16$, то базисними елементами є множина $\{0, 1, 2, ..., 9, A, B, C, D, E, F\}$.

Десяткова система числення обґрунтована Сімоном Стевіном у 1585 р. у трактаті з арифметики. Перший аналіз двійкової системи опублікований у 1670 р. в Іспанії, але загальне визнання до цієї системи прийшло після публікації Лейбніца в 1703 р. Вісімкову систему винайшов у 1717 р. шведський король Карл XII, який мав намір запровадити її у Швеції як державну. У вісімковій системі вказують право доступу для команди *chmod* в

Unix-подібних операційних системах. У шістнадцятковій системі числення кодуються помилки програмних продуктів, ця система використовується у низькорівневому програмуванні та комп'ютерній документації, при роботі з мовами високого рівня, оскільки числа в цій системі за допомогою спеціальної таблиці відповідності легко переводяться в двійкову систему.

Систему за основою $p = 3$ і базисними числами $-1,\ 0$ і 1 запропонував у 1840 р. Леон Лаллан. У 1945-1946 рр. така система розглядалась в Інституті Мура як альтернатива до двійкової системи при розробці перших ЕОМ. Скорочена система вимагає на $\ln 2 / \ln 3 \approx 0.63$, тобто на 63% цифрових позицій менше, ніж двійкова. Ця система була покладена в основу радянської ЕОМ «СЕТУНЬ» у 50-х роках минулого століття.



Детальніша інформація про системи числення як традиційні, так і з від'ємними та комплексними основами наведена в монографії [32].

Питання існування зображення (1.2) для дійсних чисел й алгоритм його побудови дається такою теоремою [14].

Теорема 1.1. *Якщо* $\{0,1,...,p-1\}$ *– множина базисних чисел, то для будь-якого дійсного числа існує зображення за основою* $p \geq 2$ *вигляду (1.2).* ■

Зауважимо, що таке зображення може бути не єдиним. Справді, 1=1.00… або 1=0.999…

Означення 1.1. *Значущими цифрами в числі (1.2) називаються всі коефіцієнти* b_i*, починаючи з першого ненульового зліва.*

Такими є підкреслені цифри в числах: 1.24600, 0.00420, 1350

1.2. Абсолютна і відносна похибки

Похибка, яка одержується при розв'язування задачі зумовлена кількома причинами.

1. Математичний опис задачі неточний, зокрема параметри математичної моделі можуть задаватися з похибками (неусувна

похибка). Наприклад у моделі коливання плоского математичного маятника

$$\ddot{u} + \omega^2 \sin u = 0,\ u(0) = u_0,\ \dot{u}(0) = \dot{u}_0,$$

такими можуть бути значення частоти ω, початкових умов u_0 і $\dot{u}_0$.

2. Похибка, викликана застосуванням наближених алгоритмів (похибка методу). Наприклад, похибка обчислення наближеного значення $\sin u$ для $|u| \le \pi/6$ згідно з формулою $u - u^3/6 + u^5/120$, не перевищує $1.5 \cdot 10^{-6}$.

3. Введення і виведення даних, виконання арифметичних операцій на комп'ютері супроводжуються заокругленням чисел (обчислювальна похибка).

Сучасне програмне забезпечення дозволяє підтримувати «точну арифметику», тобто $\sqrt{3}$ – це справді $\sqrt{3}$, а не 1.7320508... з певною кількістю цифр. Недоліком такого підходу є вельми суттєве зниження швидкості обчислень, оскільки точна арифметика апаратно не підтримується.1.7320508...

У більшості випадків замість точного значення числа x^* відомо його наближене значення x.

Означення 1.2. *Величина $\Delta(x)$, про яку відомо, що*

$$|x - x^*| \le \Delta(x), \tag{1.3}$$

називається абсолютною похибкою наближеного значення x.

Таким чином, точним значенням є $x^* = x + \varepsilon$, де $|\varepsilon| \le \Delta(x)$, а ε – точна похибка наближеного значення x. Значення похибки $\Delta(x)$ неоднозначне, тому вона вибирається найменшою з усіх можливих значень, що задовольняють нерівність (1.3).

Означення 1.3. *Величина $\delta(x)$, про яку відомо, що*

$$\left|\frac{x - x^*}{x}\right| \le \delta(x), x \ne 0,$$

називається відносною похибкою наближеного значення x.

Для відносної похибки маємо формулу

$$\delta(x) = \frac{\Delta(x)}{|x|},\ x \ne 0. \tag{1.4}$$

Використовується така форма запису:

$$x^* = x \pm \Delta(x) = x(1 \pm \delta(x)).$$

Відносна похибка відіграє головну роль у характеристиці точності числа x.

Означення 1.4. *Значуща цифра в записі числа (1.2) називається правильною, якщо абсолютна похибка числа не перевищує половини одиниці розряду, що відповідає цій цифрі.*

Для числа 0.3421 й абсолютної похибки $\Delta = 0.004$ правильними будуть цифри 3, 4 і 2. Якщо ж $\Delta = 10^{-5}$, то всі значущі цифри числа правильні.

1.3. Похибка арифметичних операцій

Знайдемо похибки операцій над двома наближеними числами x_1 і x_2. При додаванні чисел маємо

$$x_1^* + x_2^* = (x_1 + \varepsilon_1) + (x_2 + \varepsilon_2) = (x_1 + x_2) + (\varepsilon_1 + \varepsilon_2).$$

Звідси випливає, що точне значення похибки дорівнює $\varepsilon_1 + \varepsilon_2$, а для абсолютної похибка суми маємо

$$\Delta(x_1 + x_2) = \Delta(x_1) + \Delta(x_2).$$

Якщо $x_1 + x_2 \neq 0$, то відносна похибка

$$\delta(x_1 + x_2) = \frac{\Delta(x_1) + \Delta(x_2)}{|x_1 + x_2|} = \frac{|x_1|}{|x_1 + x_2|}\delta(x_1) + \frac{|x_2|}{|x_1 + x_2|}\delta(x_2).$$

Для суми $y = \alpha_1 x_1 + ... + \alpha_n x_n$, в якій коефіцієнти α_i обчислюються точно, абсолютна похибка $\Delta(y) = \sum_{i=1}^{n} |\alpha_i| \Delta_i$.

Такі ж формули правильні і для віднімання чисел. Зауважимо, що при відніманні близьких чисел x_1 і x_2 однакових знаків можливе зростання відносної похибки через малість знаменника. Наприклад, корені рівняння $x^2 - 2px + q = 0$ дорівнюють $p \pm \sqrt{p^2 - q}$. Нехай $p, q > 0$ і $q \ll p^2$. Унаслідок віднімання близьких чисел p і $\sqrt{p^2 - q}$ можлива втрата точності для кореня $p - \sqrt{p^2 - q}$. Точніший результат одержиться, якщо обчислювати корінь згідно з формулою:

$$x = \frac{q}{\sqrt{p^2 - q} + p}.$$

Похибка операції множення

$$x_1^* x_2^* - x_1 x_2 = (x_1 + \varepsilon_1)(x_2 + \varepsilon_2) - x_1 x_2 = x_1\varepsilon_2 + x_2\varepsilon_1 + \varepsilon_1\varepsilon_2.$$

Для абсолютної похибки одержується оцінка

$$\left|x_1^* x_2^* - x_1 x_2\right| \le \left|x_1\right|\Delta_1 + \left|x_2\right|\Delta_2 + \Delta_1\Delta_2.$$

Отже, для операції множення абсолютна похибка

$$\Delta = \left|x_1\right|\Delta_1 + \left|x_2\right|\Delta_2 + \Delta_1\Delta_2.$$

Відносна похибка множення

$$\delta\left(x_1 x_2\right) = \delta\left(x_1\right) + \delta\left(x_2\right) + \delta\left(x_1\right)\delta\left(x_2\right).$$

При діленні чисел x_1 і $x_2 \neq 0$ маємо

$$\frac{x_1^*}{x_2^*} - \frac{x_1}{x_2} = \frac{x_1 + \varepsilon_1}{x_2 + \varepsilon_2} - \frac{x_1}{x_2} = \frac{x_2\varepsilon_1 - x_1\varepsilon_2}{x_2\left(x_2 + \varepsilon_2\right)}.$$

Нехай $\left|x_2\right| > \Delta_2$. Тоді одержимо оцінку абсолютної похибки

$$\Delta\left(\frac{x_1}{x_2}\right) \le \frac{\left|x_1\right|\Delta_2 + \left|x_2\right|\Delta_1}{\left|x_2\right|\left(\left|x_2\right| - \Delta_2\right)}.$$

Для відносної похибки маємо

$$\delta\left(\frac{x_1}{x_2}\right) = \frac{\delta\left(x_1\right) + \delta\left(x_2\right)}{1 - \delta\left(x_2\right)}.$$

1.4. Похибка функції

Нехай функція $y = f(x_1,...,x_n)$ визначена в області $G \subset R^n$, $x := (x_1,...,x_n)$ і в цій області існують неперервні частинні похідні $\partial f / \partial x_i$, які обмежені сталими M_i. Нехай замість точних значень x_i^* відомі наближені значення $x_i \in G$ з похибками Δ_i:

$$\left|x_i - x_i^*\right| \le \Delta_i, i = \overline{1,n}.$$

Обчислимо значення $y = f(x)$ і побудуємо в області G оцінку похибки

$$\left|y - y^*\right| = \left|f(x) - f(x^*)\right|.$$

На підставі теореми про скінченні прирости маємо

$$y - y^* = f(x) - f(x^*) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x + \theta(x^* - x))(x_i - x_i^*),$$

де $\theta \in (0,1)$. Звідси одержимо

$$\left|y - y^*\right| \le \sum_{i=1}^{n} \max_{x \in G} \left|\frac{\partial f(x)}{\partial x_i}\right| \cdot \left|x_i - x_i^*\right| \le \sum_{i=1}^{n} M_i \Delta_i.$$

Отже, $\left|y-y^*\right| \le \Delta_0, \quad \Delta_0 = \sum_{i=1}^{n} M_i \Delta_i \equiv \Delta_0.$

Означення 1.5. *Величина* $A(y) = \max_{z \in G} \left|y - f(z)\right|$ *називається абсолютною граничною похибкою функції* $y = f(x)$ *в* G, *а* $A(y)/\left|y\right|$ *– відносною граничною похибкою, якщо* $y \neq 0$.

Зрозуміло, що $A(y) \le \Delta_0$. Із неперервності частинних похідних $\frac{\partial f}{\partial x_i}(x)$ випливає, що

$$\frac{\partial f}{\partial x_i}(x + \theta(x^* - x)) = \frac{\partial f}{\partial x_i}(x) + \rho_i(x - x^*),$$

де $\rho_i(z) \to 0$, при $z \to 0$. Отже,

$$y - y^* = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x)(x_i - x_i^*) + \sum_{i=1}^{n} (x_i - x_i^*)\rho_i(x - x^*).$$

Для малих $\Delta_i = \left|x_i - x_i^*\right|$ можна знехтувати другою сумою в правій частині й одержати наближену оцінку похибки:

$$\left|y - y^*\right| \le \sum_{i=1}^{n} \left|\frac{\partial f(x)}{\partial x_i}\right| \Delta_i . \qquad (1.5)$$

Величина

$$\Delta^0(y) = \sum_{i=1}^{n} \left|\frac{\partial f(x)}{\partial x_i}\right| \Delta_i \qquad (1.6)$$

називається лінійною оцінкою абсолютної похибки функції. Перевагою оцінки (1.6) в тому, що досить обчислити значення похідних у точці *x*, а не оцінювати ці похідні. Розглянемо деякі приклади.

Приклад 1.1. Нехай $y = x_1^2 + x_2^2, \ x_1 = x_2 = 1, \ \Delta_2 = 2\Delta_1 = 0.2$. Обчислимо для наближеного значення $y = 2$ оцінки похибки $\Delta_0(y), A(y)$ і $\Delta^0(y)$ в прямокутнику $G = [0.9;1.1] \times [0.8;1.2]$. Маємо

$$\frac{\partial f}{\partial x_1} = 2x_1 \le 2.2, \quad \frac{\partial f}{\partial x_2} = 2x_2 \le 2.4.$$

Тому оцінка (1.5) $\Delta_0 = 2.2 \cdot 0.1 + 2.4 \cdot 0.2 = 0.7$. Лінійна оцінка похибки $\Delta^0(y) = 2x_1\Delta_1 + 2x_2\Delta_2 = 0.6$. Гранична абсолютна похибка $A(y) = \max_{x \in G} \left|2 - \left(x_1^2 + x_2^2\right)\right|$ досягається на границі *P* у точці

із координатами (0.9;0.8) і $A(y)=0.65$. Отже, маємо оцінки $\Delta^0(y)<A(y)<\Delta_0(y)=0.7$. Нехай $x^*=(1;0.8)$, тоді $\left|y-y^*\right|=$ $=\left|2-2.65\right|=0.65$ і найкращою є гранична абсолютна похибка, а лінійна похибка – заниженою.

Приклад 1.2. Знайдемо лінійну похибку для функції

$$y=x_1^{p_1}\cdot\ldots\cdot x_n^{p_n},\ p_i\in R,\quad i=\overline{1,n},$$

визначеної в деякій області *G*, причомуy $x_i\neq 0$. Нехай відомі абсолютні похибки Δ_i, з якими обчислюються аргументи x_i або їх відносні похибки δ_i. Нагадаємо, що $\delta_i=\Delta_i/\left|x_i\right|$. Тоді

$$\Delta^0(y)=\sum_{i=1}^{n}\left|p_i\right|\left|x_1^{p_1}\cdot\ldots\cdot x_i^{p_{i-1}}\cdot\ldots\cdot x_n^{p_n}\right|\Delta_i=\sum_{i=1}^{n}\left|p_i\right|\cdot\left|y\right|\cdot\frac{\Delta_i}{\left|x_i\right|}=\left|y\right|\sum_{i=1}^{n}\left|p_i\right|\delta_i.$$

Звідси одержимо лінійну оцінку відносної похибки

$$\delta^0(y)=\sum_{i=1}^{n}\left|p_i\right|\delta_i\,. \qquad (1.7)$$

Наприклад, для функції $y=x_1x_2^{-2}x_3^{-3}$ маємо $\delta^0(y)=\delta_1+2\delta_2+3\delta_3$.

Приклад 1.3. Для лінійної відносної похибки множення і ділення на підставі оцінки (1.7) маємо

$$\delta^0\left(x\right)=\delta\left(x_1\right)+\delta\left(x_2\right).$$

Лінійна абсолютна похибка множення

$$\Delta^0\left(x_1x_2\right)=\delta^0\left(x_1x_2\right)\left|x_1x_2\right|=\left|x_2\right|\Delta\left(x_1\right)+\left|x_1\right|\Delta\left(x_2\right).$$

Для операції ділення

$$\Delta^0\left(\frac{x_1}{x_2}\right)=\left|\frac{x_1}{x_2}\right|\left(\delta\left(x_1\right)+\delta\left(x_2\right)\right)=\frac{\Delta\left(x_1\right)}{\left|x_2\right|}+\frac{\left|x_1\right|}{x_2^2}\Delta\left(x_2\right).$$

Приклад 1.4. Обчислимо лінійні абсолютну і відносну похибки функції $y=\dfrac{x_1\sqrt{x_2}}{x_1^2+x_2^2}$, якщо $x_1=x_2=1,\ \Delta_1=\Delta_2=0.1.$

Маємо:

$$\Delta^0=\left|y\right|(\delta_{чис.}+\delta_{знам.})\approx\frac{1}{2}(\delta_1+\frac{1}{2}\delta_2+\delta_{знам.})=\frac{1}{2}(\frac{\Delta_1}{\left|x_1\right|}+\frac{1}{2}\frac{\Delta_2}{\left|x_2\right|}+\frac{\Delta_{знам.}}{x_1^2+x_2^2})\leq$$

$$\leq\frac{1}{2}(0.1+\frac{1}{2}\cdot 0.1+\frac{\Delta_{x_1^2}+\Delta_{x_2^2}}{2})\leq\frac{1}{2}(0.15+\frac{1}{2}(x_1^2\cdot 2\delta_1+x_2^2\cdot 2\delta_2))=0.175.$$

Лінійна відносна похибка $\delta^0=\Delta^0/\left|y\right|\approx 0.35$ або 35%.

1.5. Обернена задача в теорії похибок

Задача полягає в знаходженні оцінок похибок аргументів, таких, що похибка обчислення значення функції $y = f(x_1,...,x_n)$ не перевищує ε.

Для функції від одного аргументу $(n=1)$ похибка $\Delta_0(y) = M\Delta(x) \le \varepsilon,$ де $|f'(x)| \le M,$ а оцінка для похибки $\Delta(x) \le \varepsilon/M$. Нехай $n>1$. Тоді можливі різні підходи до розв'язування задачі.

1. Припустимо, що в похибці функції доля кожного аргументу однакова, тобто $M_i\Delta_i = const$. Тоді $\Delta(y) \le \sum_{i=1}^{n} M_i\Delta_i = nM_i\Delta_i < \varepsilon$. Звідси випливає, що $\Delta_i < \varepsilon / nM_i,\ i=\overline{1,n}$.

2. Нехай похибки аргументів функції обмежені сталою $\bar{\Delta}$. Тоді $\Delta(y) \le \bar{\Delta}\sum_{i=1}^{n} M_i \le \varepsilon.$ Звідси маємо оцінку для похибок аргументів

$$\Delta_i \le \Delta(y) / \sum_{i=1}^{n} M_i \le \varepsilon / \sum_{i=1}^{n} M_i .$$

3. Нехай для похибки функції відома функція „вартості" $F(\Delta_1,...,\Delta_n)$, яка враховує вагу кожної із похибок $\Delta_1,...,\Delta_n$. Одержимо задачу знаходження умовного мінімуму

$$F(\Delta_1,...,\Delta_n) \to \min,$$

коли $M_1\Delta_1 + ... + M_n\Delta_n \le \varepsilon$ і $\Delta_i > 0$. Застосуємо метод множників Лагранжа. Для цього розглянемо функцію

$$\Phi(\Delta_1,...,\Delta_n,\lambda) = F(\Delta_1,...,\Delta_n) + \lambda(\varepsilon - M_1\Delta_1 - ... - M_n\Delta_n).$$

З необхідних умов екстремуму функції Φ маємо:

$$\begin{cases} \dfrac{\partial\Phi}{\partial\Delta_i} = 0, i=\overline{1,n}, \\ \dfrac{\partial\Phi}{\partial\lambda} = 0. \end{cases}$$

Звідси одержимо систему рівнянь для знаходження Δ_i

$$\begin{cases} \dfrac{\partial F}{\partial\Delta_i} - \lambda M_i = 0,\ i=\overline{1,n}; \\ \varepsilon - A_1\Delta_1 - ... - A_n\Delta_n = 0,\ \Delta_i > 0. \end{cases} \qquad (1.10)$$

Приклад 1.5. Нехай $F(\Delta_1, \Delta_2) = 0.2\Delta_1^2 + 0.8\Delta_2^2$, $2A_1 = A_2 = 4$, $\varepsilon = 0.1$. Система рівнянь (1.10) набуває вигляду:

$$\begin{cases} 0.4\Delta_1 - 2\lambda = 0, \\ 1.6\Delta_2 - 4\lambda = 0, \\ 2\Delta_1 + 4\Delta_2 = 0.1. \end{cases}$$

Звідси $\Delta_1 = 5\lambda$, $\Delta_2 = 2.5\lambda$. Із третього рівняння знаходимо $\lambda = 0.003$. Тоді $\Delta_1 = 0.025$, $\Delta_2 = 0.0125$, відповідно до ваги похибок аргументів функції $F(\Delta_1, \Delta_2)$.

1.6. Заокруглення чисел

Процес обчислень на комп'ютері супроводжується заокругленням чисел з деякою похибкою. Якими б малими не були похибки заокруглення, що виникають при виконанні арифметичних операцій, їх поява змінює математичні властивості самих операцій. Точні операції множення і додавання є комутативними, асоціативними і пов'язані між собою властивістю дистрибутивності. При виконанні обчислень на комп'ютері втрачаються властивості асоціативності та дистрибутивності арифметичних операцій. Мови програмування своєю математичною символікою створюють подібність між математичними і комп'ютерними операціями, але результати їх виконання можуть відрізнятись.

Крім виконання арифметичних операцій, джерелом похибок при обчисленнях на комп'ютері є перевід чисел із однієї системи в іншу. Розглянемо приклад. У числових методах часто використовуються числа $10^{-k}, k \geq 1$, як крок сітки. У двійковій системі число 0.1 не зображується скінченним дробом, оскільки $(0.1)_{10} = (0.00011001100...)_2$. Нехай число розрядів у мантисі 40, тоді

$$(0.1)_2 = 2^{-3} * 0.\underbrace{11001100...001100}_{40}|11... \approx 0.\underbrace{11001100...001100}_{40} * 2^{-3},$$

якщо заокруглення відбувається відкиданням розрядів. Помноживши одержане наближене значення на 10 і віднявши від 1 одержимо 2^{-40}, а не 0.

1.6.1. Способи заокруглення. Для числа $x = b_n ... b_s b_{s-1} ...$ заокруглене число $x_s = b_n ... b_s 0...$ способом «відкидання» одержується шляхом заміни $b_{s-1}, b_{s-2} ...$ нулями. Перевага цього способу – проста реалізація. Недоліком те, що похибка заокруглення

$|\varepsilon_s| \le p^s$, де $\varepsilon_s = x_s - x$. Крім того, для $x > 0$ маємо $\varepsilon_s \le 0$, а для $x < 0$ похибка $\varepsilon_s \ge 0$. Отже, обчислення над числами одного знаку може привести до нагромадження обчислювальної похибки. Рівність $|\varepsilon_s| = p^s$ досягається, коли x нескінченний дріб вигляду $b_n ... b_s (p-1)(p-1)...$

Інший спосіб, який має назву "правильне заокруглення", полягає ось у чомуу:

$$x_s^* = \begin{cases} x_s, \textit{ якщо } |x_s - x| < 0.5p^s; \\ x_s + p^s, \textit{ якщо } |x_s - x| > 0.5p^s; \\ x_s \textit{ або } x_s + p^s, \textit{ якщо } |x_s - x| = 0.5p^s. \end{cases}$$

На мові Сі $2/3 = 0.6...67$. Перевагою такого способу заокруглення є удвічі менша оцінка для похибки, а також те, що знак числа x і похибки ε_S не залежать один від одного. Недолік полягає в тому, що приблизно в половині випадків виконується операція додавання і кожен раз порівняння. До того ж потрібно з'совувати питання із невизначеністю $|x_s - x| = 0.5p^s$, наприклад, випадково вибирати той чи інший варіант.

1.6.2. Заокруглення в скорочених системах. Задача полягає в об'єднанні переваг кожного із способів. Тобто, заокругливши шляхом заміни значень у розрядах $s,\ s-1,\ldots$ нулями, ґарантувати для похибки оцінку

$$|x_s - x| \le \frac{1}{2} p^s. \qquad (1.11)$$

Цього можна досягти вибором основи p і базисних чисел $\alpha_0, ..., \alpha_{p-1}$. Для виконання нерівності (1.11) необхідно і досить, щоб всі числа $0...0 b_{s-1} b_{s-2} ...$ по модулю не перевищували $0.5p^s$. Оскільки $|b_i| \le p-1$, то необхідно і досить, щоб

$$\max_{0 \le k \le p-1} |\alpha_k| \le \frac{p-1}{2}. \qquad (1.12)$$

Наприклад, для $p = 10$ умова (1.12) не виконується, оскільки $|\alpha_k| > 9/2$ для $\alpha_k = \overline{5,9}$.

Нехай $p = 2m$. Тоді $(p-1)/2 = m - 1/2$. Базисних чисел, які задовольняють умову (1.12) є тільки $2(m-1)+1 = 2m-1 < 2m$. Отже, основа може бути хіба що непарним числом $p = 2m+1$.

Тоді $(p-1)/2=m$. Множина базисних чисел складається із $2m+1$ чисел $0,\pm1,...,\pm m$. Системи з такою основою і множиною базисних чисел називаються *скороченими* [14, 32].

Приклад 1.6. Найменшою основою у скорочених системах є $p=3$, базисними числами: $0,\bar{1},+1$, де $\bar{1}:=-1$, наприклад, число $(7/27)_{10}=1\cdot3^{-1}-1\cdot3^{-2}+1\cdot3^{-3}=(0.1\bar{1}1)_3$. У десятковій системі $7/27\approx0.259259...=0.(259)$. Якщо заокруглювати до двох значущих цифр, то $x_{-2}=0.25$ і $x-x_{-2}=0.0092...>0.2*10^{-2}$.

У скороченій системі $x_{-2}=0.1\bar{1}$ і $x-x_{-2}=0.001=1/27<1/2\cdot3^{-2}=$

$=1/18$. Зауважимо, що $-7/27=-3^{-1}+3^{-2}-3^{-3}=0.\bar{1}1\bar{1}$. Тобто знак числа у скороченій системі визначається знаком першої значущої цифри, а зміна знаку числа рівносильна зміні знаків усіх значущих цифр.

1.7. Заокруглення в системах з фіксованою крапкою

В пам'яті комп'ютера кожне число розміщується в машинному слові зі своїм знаком і фіксованою кількістю цифр (розрядів). В одному з таких способів під дробову частину виділяється строго фіксована кількість цифр. Це зображення із фіксованою крапкою характеризується трьома параметрами: p – основа системи числення, t – кількість розрядів для зображення числа і f – кількість розрядів, відведених для дробової частини. Позначимо таку систему через

$$F(p,t,f). \qquad (1.13)$$

Наприклад, $F(10,6,3)$. Тут $131.404\in F$, а $11.40438\notin P$. Числа, розміщені рівномірно із кроком 10^{-3}. Найбільше число $999.999=10^{6-3}-10^{-3}$, найменше додатне число $0.001=10^{-3}$. У загальному випадку системи (1.13) $x_{\max}=p^{t-f}-p^{-f}=p^{-f}(p^t-1)$, $0<x_{\min}=p^{-f}$ – крок. Кількість чисел у системі (1.13) дорівнює $2(p^t-1)+1=2p^t-1$. Якщо $x\in[-M_\infty,M_\infty]$, то його зображення позначимо через $fix(x)$. Для $x\in P$ маємо $fix(x)=x$, інакше

$$|x-fix(x)|\le p^{-f}/2. \qquad (1.14)$$

Отже, абсолютна похибка заокруглення в системі (1.13) не залежить від x. Відносна похибка

$$\frac{|x - fix(x)|}{|x|} \le \frac{1}{2}\frac{p^{-f}}{|x|},\ x \ne 0$$

нерівномірна по відношенню до x.

Операції “+” і “–” в системі (1.16) виконуються точно, якщо не виникає переповнення. Операції “$*$” і “/” в загальному випадку виконуються із похибкою, причому

$$fix(x \circ y) = x \circ y + \upsilon,\ \ |\upsilon| \le 0.5p^{-f}.$$

1.8. Системи з плаваючою крапкою

В системі чисел $P(p,t,L,U)$ із плаваючою крапкою p – основа системи числення, t – кількість цифр (розрядів), відведених для мантиси; L і U – межі зміни значень показника степеня чисел. Для довільного $x \in P$, $x \ne 0$, маємо зображення

$$x = \pm\left(\frac{b_1}{p} + \frac{b_2}{p^2} + ... + \frac{b_t}{p^t}\right) \times p^l,\ L \le l \le U,$$

або $x = \pm .b_1 ... b_t \times p^l$, де $1 \le b_1 \le p-1,\ 0 \le b_i \le p-1,\ i = \overline{2,t}$. Число $0 \in P$ зображується у вигляді $0 = +.0...0 \times p^L$.

Числа в системі P розміщені нерівномірно. У системі $P(10,2,-1,1)$, для якої $M_\infty = 0.99 * 10^1$, $0 < x_{\min} = 0.10 * 10^{-1}$. Додатні числа розміщені рівномірно на відрізках: $[0.010; 0.099]$ із кроком 0.001, $[0.1; 0.99]$ із кроком 0.01 і $[1; 9.9]$ із кроком 0.1. У загальному випадку маємо

$$M_\infty = +.(p-1)...(p-1) \times p^U,\ 0 < x_{\min} = +.10...0 \times 10^{-1}.$$

Для кожного l числа розміщені рівномірно з кроком $h = p^{l-t}$. Число $\omega := x_{\min} > 0$ називається машинним нулем, а множина $(-\omega, 0) \cup (0, \omega)$ – областю машинного нуля.

Зображення дійсного числа x у системі P позначимо через $fl(x)$. Кількість чисел у системі з плаваючою крапкою дорівнює $2(U-L+1)(p-1)p^{t-1}+1$. Відстань ε між 1 і наступним числом системи P називається „машинним епсилон” (англ. *Machine epsilon*). Наступне після одиниці число дорівнює $1 + p^{-t} \cdot p^1 = p^{1-t} + 1$. Тому

$$\varepsilon = p^{1-t}. \qquad (1.15)$$

Перше зліва від одиниці число дорівнює $1-p^{-t}$. Отже, відстань між ним й одиницею складає $p^{-t}=\varepsilon/p$ (рис. 1.1).

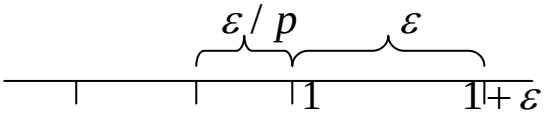


Рис. 1.1. Машинне епсилон

Зображення двох чисел збігається, якщо вони відрізняються на величину, яка менша по модулю, ніж машинне епсилон (1.15). На мові С існують граничні сталі FLT_EPSILON для типу *float* і дорівнює 1.2E-7 та DBL_EPSILON типу *double* і має значення 2.2E-16.

Нехай $x\in P$. Відстань між x і сусіднім числом задовольняє нерівність

$$\frac{\varepsilon|x|}{p}\le|x-y|\le\varepsilon|x|.$$

Справді, $|x-y|=p^{l-t}=\varepsilon\cdot p^{e-1}\le\varepsilon|x|$ і $|x-y|=\varepsilon\cdot p^{l-1}\ge\frac{\varepsilon}{p}p^{l}\ge\frac{\varepsilon}{p}|x|$.

На сьогодні загальноприйнятий IEEE-стандарт двійкової арифметики. Він реалізований на робочих станціях Sun, DEC, HP, IBM а також на всіх персональних комп'ютерах. IEEE-арифметика передбачає два типи чисел із плаваючою крапкою: числа зі звичайною точністю (32-бітове зображення) і числа подвійної точності (64-бітове зображення). Формат чисел показаний на рис. 1.2.

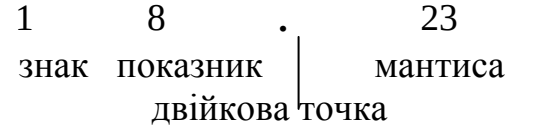


Рис. 1.2а. IEEE-число звичайної точності

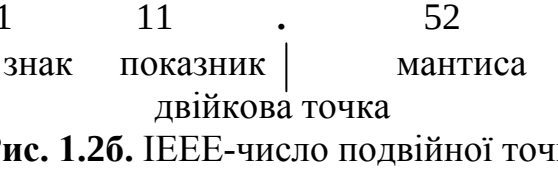


Рис. 1.2б. IEEE-число подвійної точності

Нехай в зображенні IEEE-числа звичайної точності s – однобітовий знак, f – 8-бітовий показник і f – мантиса, $f<1$. Тоді число дорівнює $(-1)^s\cdot 2^{l-127}(1+f)$. а область додатних нормалізованих чисел знаходиться в діапазоні від 2^{-126} (поріг

машинного нуля) до $2^{127}(2-2^{-23}) \approx 2^{128}$ (поріг переповнення), або, наближено, від 10^{-38} до 10^{38}.

Аналогічно для числа із подвійною точністю маємо $(-1)^s \cdot 2^{l-1023}(1+f)$. Максимальна відносна похибка зображення дорівнює $2^{-53} \approx 10^{-16}$, а границі області нормалізованих чисел складають 2^{-1022} (поріг машинного нуля) і $2^{1023}(2-2^{-52}) \approx 2^{1024}$ (поріг переповнення), наближено 10^{-308} і 10^{308}. Зауважимо, що уникнути аварійних ситуацій попадання в область машинного нуля можна, певним чином конструюючи обчислювальний алгоритм.

1.9. Оцінка похибки в системі з плаваючою крапкою

Якщо $x \in P$, то $fl(x) = x$. Якщо ж $\omega \le |x| \le M_\infty$ і $x \notin P$, то число зображується з деякою похибкою, одержаною внаслідок заокруглення мантиси до *t*-го розряду. Похибка $fl(x) - x$ нерівномірна відносно x. Нехай

$$\delta(x) = \frac{fl(x) - x}{x},\ x \neq 0. \qquad (1.16)$$

Тоді

$$fl(x) = x(1+\delta(x)).$$

Оцінимо похибку $\delta(x)$ для випадку правильного заокруглення (похибка не перевищує $p^{-t}/2$). Якщо $x = \pm .b_1 \ldots b_t * p^l$, то $.b_1 \ldots b_t \ge p^{-1}$, оскільки $b_1 \ge 1$. Тому $|x| \ge p^{-1} \cdot p^l = p^{l-1}$. Звідси маємо $|\delta(x)| \le (p^{-t} \cdot p^l)/(2p^{l-1}) = p^{l-1}/2$. Отже, для $\delta(x)$ справджується оцінка

$$|\delta(x)| \le p^{1-t}/2 \qquad (1.17)$$

або $|\delta(x)| \le \varepsilon/2$. Для заокруглення способом „відкидання” маємо $|\delta(x)| \le \varepsilon$.

Якщо ж $x = 0$, то покладемо $\delta = 0$. Для чисел з області “машинного нуля”, коли $0 < |x| < \omega$, маємо $fl(x) = 0$ і в цьому випадку $\delta(x) = -1$.

Таким чином, відносна похибка заокруглення не залежить від самого числа, а визначається параметрами p і t (або ε) системи з плаваючою крапкою P:

$$|\delta(x)| \le \begin{cases} \varepsilon / 2, \textit{ для правильного заокруглення}; \\ \varepsilon, \textit{ заокруглення "відкиданням"}. \end{cases}$$

Нехай символ $\circ$ означає одну із бінарних операцій +, –, * або / над числами x і y. Якщо точне значення $x \circ y \notin P$, то перш, ніж записати його в пам'ять або регістр, результат апроксимується певним числом $fl(x \circ y)$ із P. Різниця $fl(x \circ y) - (x \circ y)$ називається похибкою заокруглення. В стандарті IEEE $fl(x \circ y)$ − найближче до $x \circ y$ число, тобто заокруглення правильне. Якщо число $x \circ y$ знаходиться точно посередині між двома сусідніми числами з плаваючою крапкою, то з двох можливих значень для $fl(x \circ y)$ IEEE-арифметика вибирає число з нульовим останнім розрядом мантиси (заокруглення до найближчого парного). IEEE-арифметика забезпечує також те, що $fl(\sqrt{x}) = \sqrt{x}(1+\delta),\ \delta \le 0.5\varepsilon$.

IEEE-арифметика охоплює також субнормальні числа, тобто ненормалізовані числа з плаваючою крапкою, які мають найменший можливий показник. Ці числа знаходяться між нулем і найменшим нормалізованим числом як показано в [23, с. 21] на рис.1.4, де для мантис прийнято 3-бітове зображення.

IEEE-арифметика передбачає також символи $\pm\infty$ і NaN (Not a Number – не число). Символи $\pm\infty$ генеруються при переповненні і надалі задовольняють правила: $x / \pm\infty = 0 \ \forall x \in P$; $x / 0 = \pm\infty \ \forall x \in P \setminus \{0\}$; $+\infty + \infty = +\infty$ тощо. Будь-яка операція, результат якої скінченний або нескінченний, але не визначений коректно, генерує символ NaN, наприклад, $\infty - \infty$, $\frac{\infty}{\infty}$, $NaN \circ x$ та ін.

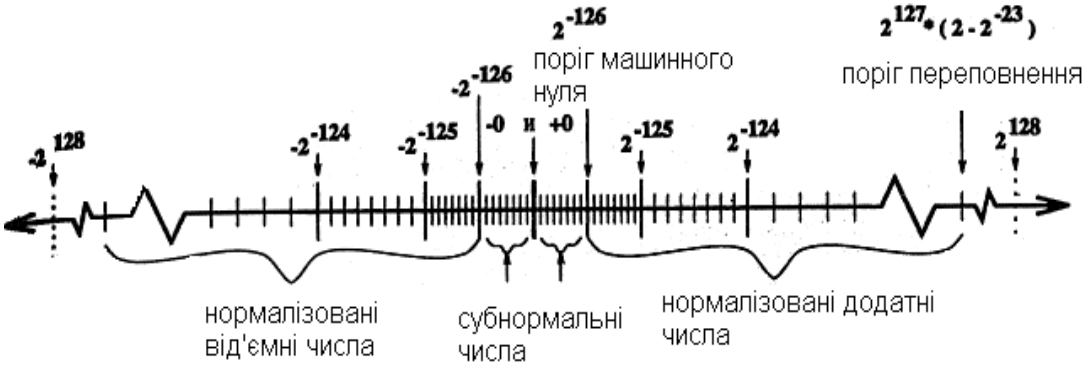


Рис. 1.4. Дійсна числова пряма, на якій числа з плаваючою крапкою помічені суцільними штрихами

Рис. 1.3. Ракета Ariane 5 (фото з Вікіпедії)

У кожному із наступних випадків арифметична операція некоректна і породжує NaN: відбулось переповнення; зустрілось ділення на нуль, внаслідок чого генерується $\pm\infty$; одержаний машинний нуль, – виставляється індикатор особливого випадку (exception flag). Надалі стан індикатора може перевірятись програмою користувача. Ця особливість арифметики дозволяє розробляти надійніші програми, оскільки програма самостійно може виявити і виправити особливі випадки, замість того, щоб припинити своє виконання. Також програма виконується швидше, оскільки можна уникнути численних можливих, але малоймовірних особливих випадків. Неправильна обробка особливого випадку в арифметиці з плаваючою крапкою привела до катастрофи 5 червня 1996 р. ракети Ariane 5 Європейського космічного агентства, оскільки сталася помилка при перетворенні 64-розрядного числа з плаваючою крапкою на 16-розрядне ціле. Надто велике значення дійсного числа не вмістилось у 16 розрядах, що спричинило переповнення.

У мовах програмування машинним числам звичайної і подвійної точності відповідають свої типи зображення даних. На С – *float* і *double*. Крім цих стандартів комп'ютерного зображення дійсних чисел, які підтримуються апаратно, мови програмування дозволяють оперувати з машинними числами підвищеної точності в розширеному діапазоні. Наприклад, під числа типу *long double* в С виділяється 10 байтів пам'яті, що відповідає 18–20-розрядній мантисі, нижня границя діапазону додатних чисел якої складає приблизно $1.2\cdot10^{-4932}$.

1.10. Особливості комп'ютерної арифметики у системі з плаваючою крапкою

Нехай „$\circ$" – знак деякої арифметичної операції і $x\circ y\notin(-\omega,0)\cup(0,\omega)$. Тоді

$$fl(x \circ y) = (x \circ y)(1+\delta),\ |\delta| \le 0.5p^{1-t}.$$

Розглянемо на прикладах деякі особливості виконання арифметичних операцій, для яких можуть порушуватися властивості асоціативності і дистрибутивності.

Приклад 1.7. Нехай $S = x + y + z$. Маємо:

а) $((x+y)+z) \Rightarrow (x+y)(1+\delta_1)+z \to ((x+y)(1+\delta_1)+z)(1+\delta_2) =$

$$= (x+y+z+(x+y)\delta_1)(1+\delta_2) = fl((x+y)+z);$$

б) $x+(y+z) \to x+(y+z)(1+\bar{\delta}_1) \to (x+(y+z)(1+\bar{\delta}_1)(1+\bar{\delta}_2) =$

$$= ((x+y+z+(y+z)\bar{\delta}_1)(1+\bar{\delta}_2) = fl(x+(y+z)).$$

У загальному випадку результати можуть відрізнятись.

Ненадійність обчислювального алгоритму деколи зумовлюється заокругленням окремих чисел з відповідною втратою критично важливої інформації. Одна з можливих причин – віднімання близьких чисел.

Приклад 1.8. Ще одним прикладом формування похибки залежно від вибору обчислювального алгоритму служить обчислення за розкладом у ряд Тейлора функції

$$e^{-x} \approx 1 - x + \frac{x^2}{2!} - \ldots + (-1)^n \frac{x^n}{n!}.$$

Якщо $x > 0$, то точніше обчислювати e^x за формулою $1/e^x$, оскільки тоді ряд для e^x не буде знакопереміжним і при відніманні чисел не втрачаються значення в молодших розрядах. Наприклад, для *n=5, x=1.5* і типу даних *real* у Паскалі похибка обчислень складає 0.01297, тоді як для $1/e^x$ маємо похибку 0.00998.

Приклад 1.9. При додаванні чисел у системі з плаваючою крапкою не виконується властивість асоціативності, тому залежно від порядку сумування породжується сім'я обчислювальних алгоритмів. Наприклад, для змінних типу *double* на мові С для $n = 10^9$

$$S_n^{(1)} = \left(\left(\ldots\left(\frac{1}{1}+\frac{1}{2}\right)+\frac{1}{3}\right)+\ldots+\frac{1}{n-1}\right)+\frac{1}{n} \approx 21.3004815023485,$$

$$S_n^{(2)} = \left(\left(\ldots\left(\frac{1}{n}+\frac{1}{n-1}\right)+\frac{1}{n-2}\right)+\ldots+\frac{1}{2}\right)+1 \approx 21.3004815023461.$$

Результати відрізняються на $2.4\cdot10^{-12}$. Для $n=10^6$ похибка менша і дорівнює $7.8\cdot10^{-13}$, а зі збільшенням n до $n=10^{10}$ зростає до $4.4\cdot10^{-12}$. Виникають питання: чому порушується властивість асоціативності та як проводити сумування, щоб похибка була якомога меншою?

Проаналізуємо похибку обчислення суми чисел

$$S_n = a_1 + a_2 + ... + a_n, \quad a_i > 0. \tag{1.18}$$

Для довільних a_k і a_e маємо, $fl(a_k + a_l) = (a_k + a_l)(1+\delta)$, де $|\delta| \le 0.5p^{1-t}$. Оскільки $a_i > 0$, то $\delta \ne -1$. Нехай сумування відбувається зі збільшенням індексу. Тоді

$$fl(S_n) = (...(a_1 + a_2)(1+\delta_2) + a_3)(1+\delta_3) + ... + a_n)(1+\delta_n) =$$

$$= (a_1 + a_2)\prod_{i=2}^{n}(1+\delta_i) + a_3\prod_{i=3}^{n}(1+\delta_i) + ... + a_n(1+\delta_n) = \sum_{j=1}^{n} a_j(1+E_j). \tag{1.19}$$

Тут $E_1 = E_2, 1+E_1 = 1+\delta_2 + ... + \delta_n + \delta_2\delta_3 + ... + \delta_{n-1}\delta_n + ... + \delta_2...\delta_n$. Оскільки $|\delta_i| \le 0.5\rho^{1-t} \ll 1$, то залишивши тільки лінійні доданки δ_i, одержимо $E_1 \approx \delta_2 + ... + \delta_n$, $E_j \approx \delta_j + ... + \delta_n$ для $j = \overline{3,n}$. Отже,

$$|E_1| = |E_2| \le |\delta_2| + ... + |\delta_n| \le 0.5(n-1)p^{1-t};$$

$$|E_j| \le 0.5(n-j+1)p^{1-t}, \quad j = \overline{3,n}.$$

Тепер із (1.18) і (1.19) маємо $fl(S_n) - S_n = \sum_{j=1}^{n} a_j E_j$ і

$$|fl(S_n) - S_n| \le \sum_{j=1}^{n} a_j |E_j| \le 0.5p^{1-t}\left[(a_1 + a_2)(n-1) + a_3(n-2) + ... + 2a_{n-1} + a_n\right].$$

Із одержаної нерівності випливає, що похибка залежить від порядку додавання чисел. Оцінка відносних збурень E_j має найбільші множники для перших доданків у сумі і найменші для останніх. Тому для зменшення обчислювальної похибки потрібно починати сумувати з найменших за абсолютною величиною доданків, а закінчувати найбільшими. При формуванні сумарної похибки кожен доданок бере участь у сумуванні один раз, а в утворенні похибки результату стільки разів, скільки разів додаються частинні суми, залежні від цього доданка.

Нехай $n = 2^k$. Щоб уникнути нерівноправності доданків можна спочатку обчислити $a_{i,i+1} = a_i + a_{i+1}$, $i = \overline{1,n-1}$, потім обчислюються $a_{i,i+1} + a_{i+2,i+3}$ і т.д. Кожен доданок бере участь в утворенні похибки рівно k разів, тому в кінцевому результаті

$$fl(S_n) = \sum_{j=1}^{n} a_j(1+E_j),\ |E_j| \le 0.5p^{1-t}k.$$

Отже, один і той же метод можна реалізувати на комп'ютері різними обчислювальними алгоритмами, які відрізняються один від одного порядком виконання арифметичних операцій, але результати їх реалізації в одному й тому ж обчислювальному середовищі можуть відрізнятись [14, 32, 56, 97, 100].

Приклади розв'язування типових задач

Задача 1. Знайти оцінку абсолютної похибки функцій $u = \sin x$ і $u = \ln x$.

Розв'язування. Маємо $|u| = |\cos x| \le 1 = M$. Тому $|u - u^*| \le \Delta_x = \Delta_u$, тобто похибка обчислення функції не перевищує похибки аргумента. Для функції $u = \ln x$ лінійна абсолютна похибка дорівнює відносній похибці аргументу, оскільки $\Delta^0(u) = \Delta_x / x = \delta_x$.

Задача 2. Обчислити лінійну абсолютну і відносну похибки функції $u = x^2\sqrt{y}(z+4)$ у точці $(-2,4,-8)$, якщо $\Delta_x = 2\Delta_y = 0{,}02$, а $\delta_z = 0{,}01$.

Розв'язування. На підставі формули (1.4) $\Delta_z = |z|\delta_2 = 2\cdot 0{,}01 = 0{,}02$, $\delta_x = 0{,}02/2 = 0{,}01$, $\delta_y = 0{,}01/4 = 0{,}0025$, $\delta_{z+4} = \Delta_{z+4} / |z+4| = 0{,}005$. Застосувавши формулу (1.7), одержимо

$$\delta^0(u) = 2|x|\delta_x + \frac{1}{2}|y|\delta_y + |z+4|\delta_{z+4} = 4\cdot 0{,}01 + 2\cdot 0{,}0025 + 4\cdot 0{,}005 = 0{,}075;$$

$$\Delta^0(u) = |u|\delta^0(u) = \frac{4\cdot 2}{4}\cdot 0{,}075 = 0{,}15.$$

Задача 3. Обчислити відносну лінійну похибку значення функції $u = x_1x_2^2x_3^3$ у точці $(37.100, 9.870, 6.052)$ з похибками аргументів $\Delta(x_1) = 0.3$, $\Delta(x_2) = 0.11$ і $\delta(x_3) = 0.0011$.

Розв'язування. Обчислимо відносні похибки $\delta(x_1) = 0.3/37.1 \approx$ ≈ 0.0081, $\delta(x_2) = 0.11/9.87 \approx 0.0112$. Тоді $\delta^0(u) = \delta(x_1) + 2\delta(x_2) +$ $+3\delta(x_3) = 0.0226$ або 2.26%.

Задача 4. Нехай додатне наближене число x має m правильних цифр. Показати, що для його відносної похибки δ виконується оцінка $\delta \le p^{1-m}/(2b_n)$, де b_n – перша значуща цифра числа.

Розв'язування. Оскільки $x = b_n p^n + b_{n-1} p^{n-1} + ... + b_{n-m+1} p^{n-m+1}$, то його абсолютна похибка $\Delta_x \le 0{,}5\, p^{n-m+1}$. Тому відносна похибка

$$\delta_x = \frac{\Delta_x}{x} \le \frac{0{,}5\, p^{n-m+1}}{b_n p^n + ... + b_{n-m+1} p^{n-m+1}} \le \frac{0{,}5\, p^{n-m+1}}{b_n p^n} = \frac{p^{1-m}}{2b_n}.$$

Зауважимо, що для $p = 2$ відносна похибка $\delta_x \approx 2^{-m}$.

Задача 5. Знайти лінійну оцінку похибки при обчисленні неявної функції $\varphi(u, x_1, x_2, ..., x_n) = 0$, якщо відома точка наближення $(x_1, x_2, ..., x_n)$, значення функції u в цій точці і похибки аргументів $\Delta_1, \Delta_2, ..., \Delta_n$.

Розв'язування. Диференціюючи функцію φ по x_i, одержимо $\frac{\partial \varphi}{\partial u} \cdot \frac{\partial u}{\partial x_i} + \frac{\partial \varphi}{\partial x_i} = 0$, звідки $\frac{\partial u}{\partial x_i} = -\frac{\partial \varphi}{\partial x_i}\left(\frac{\partial \varphi}{\partial u}\right)^{-1}$. Значення u можна знайти, розв'язавши рівняння $\varphi(u; x_1, x_2, ..., x_n) = 0$ для заданих $x_1, x_2, ..., x_n$. Обчисливши значення похідних $\partial u(u; x_1, x_2, ..., x_n)/\partial x_i$, одержимо лінійну похибку $\Delta^0(u)$ згідно з формулою (1.6).

Задача 6. Обчислити похибку коренів квадратного рівняння $\varphi(u;\, p,\, q) := u^2 + pu + q = 0$, якщо задано коефіцієнти p, q, та їх похибки $\Delta(p)$ і $\Delta(q)$.

Розв'язування. Обчислимо похідні

$$\frac{\partial u}{\partial p} = -\frac{\partial \varphi}{\partial p}\left(\frac{\partial \varphi}{\partial u}\right)^{-1} = -u(2u + p)^{-1},\ \frac{\partial u}{\partial q} = -\frac{\partial \varphi}{\partial q}\left(\frac{\partial \varphi}{\partial u}\right)^{-1} = -(2u + p)^{-1}.$$

На підставі формули (1.6) відносна лінійна похибка коренів

$$\delta^0(u) = \frac{|u|\Delta(p) + \Delta(q)}{|2u + p|}.$$

Нехай $p = 5,\ q = -6,\ \Delta(p) = \Delta(q) = 0.001$. Для розв'язку рівняння $u = -6$ маємо $\delta^0(-6) = \frac{6 \cdot 0.001 + 0.001}{2 \cdot 6 + 5} \approx 0.00041$ або 0.041%.

Задача 7. Для обчислення евклідової норми $\|x\| = \sqrt{x_1^2 + ... + x_n^2}$, $x_i \in (\omega, \sqrt{\omega})$, $\omega \ll 1$, вектора x побудувати алгоритм обчислення $\|x\|$ так, щоб $fl(\|x\|) \neq 0$.

Розв'язування. Оскільки $x_i^2 \in (\omega^2, \omega)$, то $fl(\|x\|) = 0$, хоча $\|x\| > \omega$.

Покладемо $\alpha = \max|x_i|$ і обчислимо $\|x\| = \alpha\left[\left(\frac{x_i}{\alpha}\right)^2 + ... + \left(\frac{x_n}{\alpha}\right)^2\right]^{1/2}$.

Тоді $\frac{|x_i|}{\alpha} > \frac{\omega}{\sqrt{\omega}} = \sqrt{\omega}$, тому $\left(\frac{x_n}{\alpha}\right)^2 > \omega$. Отже, $\|x\| > \omega$ і $fl(\|x\|) \neq 0$.

Можна показати [29], що $fl(\|x\|) = \|x\|(1+E)$, $|E| \le \frac{n+6}{4} p^{1-t}$.

Задача 8. Показати, що у форматі чисел з плаваючою крапкою з основою 10, чотирирозрядною мантисою, додатковими чотирма бітами при виконанні обчислень і правильним заокругленням результат обчислення виразу $0.1456 \cdot 10^3 - 0.1455 \cdot 10^3 + 0.1231 \cdot 10^4$ залежить від порядку виконання операцій.

Розв'язування. Обчисливши $a - b$, а потім додавши c, одержимо

$$a - b = 0.0001 \cdot 10^3 \to 0.1000 \cdot 10^0 := d_1;\ d_1 + c =$$
$$= (0.00001000 + 0.12310000) \cdot 10^4$$
$$= 0.12311000 \cdot 10^4 \to 0.1231 \cdot 10^4 := S_1 = c.$$

Тепер обчислимо спочатку $c - b$, а потім додамо a. Одержимо

$$(0.123110000 - 0.0145500) \cdot 10^4 = 0.10855000 \cdot 10^4 \to 0.1085 \cdot 10^4.$$

Наступна операція $(0.01456000 + 0.10850000) \cdot 10^4 =$

$$= 0.12301000 \cdot 10^4 \to 0.1230 \cdot 10^4 \neq c.$$

Задача 9. Радіус кулі $R = 10$ і виміряний з похибкою $\Delta_r = 10^{-3}$. Знайти оцінку абсолютної і відносної похибки обчислення об'єму кулі. В значенні $\pi \approx 3.1416$ всі цифри є правильними.

Розв'язування. Оскільки $V = 4\pi R^3/3$, то лінійна відносна похибка $\delta_V^0 = 4(\delta_\pi + 3\delta_R)/3$. Відносні похибки набувають значень $\delta_\pi = \Delta_\pi/\pi = 10^{-3}/(2 \cdot 3.1416) \approx 0.0002$, $\delta_R = \Delta_R/R = 10^{-4}$. Тому

$\delta_V^0 = 0.0007$ або 0.07%. Лінійна абсолютна похибка $\Delta_V^0 = \delta_V^0 \cdot V = 4 \cdot 3.1416 \cdot 1000 \cdot 0.0002/3 = 0.8378$.

Завдання та запитання для самостійної роботи

1. Як класифікуються похибки при розв'язуванні задач? Навести приклади.
2. Які є способи розв'язування оберненої задачі в теорії похибок для функції n змінних, коли $n \geq 2$?
3. Як зображаються числа у системі з фіксованою крапкою та які основні характеристики заокруглення в ній?
4. Пояснити зображення чисел у системі з плаваючою крапкою, основні характеристики цієї системи, заокруглення й особливості виконання арифметичних операцій у ній.
5. Визначити значущі та правильні цифри чисел
   $a_1 = 0{,}0443$, $a_3 = 0.04432100$, якщо $\Delta(a_1) = 0{,}002 \cdot 10^{-k}$, $k = \overline{0,3}$.
6. Для системи з фіксованою крапкою $P(10,2,-1,1)$ обчислити параметри: ω, ε, M_∞, кількість чисел у P та оцінку відносної похибки $\delta(x)$.
7. Для системи з фіксованою крапкою $F(10,6,4)$ обчислити параметри: $x_{\min} > 0$, M_∞, кількість чисел, оцінку абсолютної похибки.
8. Показати, що для будь-якого способу заокруглення в системі з плаваючою крапкою операції додавання і множення не є асоціативними і не зв'язані між собою законом дистрибутивності.
9. Проаналізувати відносну похибку заокруглення для дійсних типів даних в одній з мов програмування.
10. Сформулювати властивості скорочених систем та заокруглення в них. Навести приклади скорочених систем з основою $p = 3$ і 5.
11. Скласти на мові C, C++ або Python програму обчислення машинного епсилон.
12. Одержати оцінку відносної похибки для x_1 / x_2. Порівняти результат з лінійною відносною похибкою.
13. Обчислюється добуток n чисел $P = x_1 \cdot x_2 \cdot \ldots \cdot x_n$, серед яких є досить малі і досить великі числа. Побудувати алгоритм множення чисел, який не приводить до зникнення порядку чи переповнення, якщо значення P – допустиме число.
14. Використовуючи різні способи розв'язування оберненої задачі в теорії похибок знайти оцінки похибок аргументів функцій $y = \cos(x_1^2 + x_2^2 - 2)$ і

$y = \log_2(x_1^2 + x_2^2)$, які обчислюються при $x_1 = x_2 = 1$, $\Delta_y = 0.001$ і значення x_1 та x_2 змінюються на проміжку $[0.5; 1.5]$.

15. Обчислити відносну похибку значення функції $u(x, y, z) = xy^2 / z^3$, якщо задано $x = 1{,}2$, $y = 3{,}4$, $z = 4{,}5$, $\Delta_x = 0{,}1$, $\delta_y = 0{,}01$, $\delta_z = 0{,}02$.

16. Оцінити абсолютну та відносну похибки обчислення функції:

1) $f(x, y, z) = \ln \dfrac{xy}{z}$ при $x = 2{,}34 \pm 0{,}01$, $y = 1{,}25 \pm 0{,}02$, $z = 3{,}05 \pm 0{,}02$;

2) $f(x, y, z) = \sqrt{\dfrac{xy}{z}}$ при $x = 0{,}757 \pm 0{,}001$, $y = 21{,}7 \pm 0{,}05$, $z = 1{,}84 \pm 0{,}05$;

3) $f(x, y, z) = \dfrac{\sqrt{x} + y}{\sqrt[3]{z}}$ при $x = 4 \pm 0{,}1$, $y = 3 \pm 0{,}05$, $z = 1 \pm 0{,}08$.

17. Нехай a, b і c – дійсні числа, $a \neq b$, $c \neq 0$ і для них відомі відносні похибки δ_a, δ_b і δ_c. Для якої з функцій $u = \dfrac{a - b}{c}$ і $v = \dfrac{a}{c} - \dfrac{b}{c}$ відносна похибка наближеного обчислення її значення може бути більшою?

18. Нехай $|x| < 1$. У якому порядку ліпше обчислювати суму $\sum_{k=0}^{n} x^k$, щоб зменшити обчислювальну похибку? Відповідь обґрунтувати.

19. З яким числом правильних цифр потрібно взяти $\lg 2$, щоб обчислити корені рівняння $x^2 - 2x + \lg 2 = 0$ із чотирма правильними цифрами?

20. Показати, що при відсутності переповнень і машинних нулів,

$$fl\left(\sum_{i=1}^{n} x_i y_i\right) = \sum_{i=1}^{n} x_i y_i (1 + \delta_i), \text{ де } |\delta_i| \le n\varepsilon,\ \varepsilon = p^{1-t}.$$

21. Відстань, пройдена вільно падаючим тілом у вакуумі, обчислюється за формулою $S = gt^2 / 2$, де g – при скорення вільного падіння, де t – час з початку падіння. Нехай $g = 9.81$ м/сек^2 точне, а t вимірюється з точністю до 0,05 сек. Показати, що з ростом t абсолютна похибка обчислення S збільшується, а відносна похибка зменшується.

22. Висота зрізаного конуса виміряна з відносною похибкою $\delta = 0.001$ і дорівнює 10 см. Радіуси основ $R = 20\,см$ і $r = 5\,см$ відомі з абсолютною похибкою $\Delta = 0.02$. Знайти оцінки абсолютної і відносної похибок обчислення об'єму конуса, якщо $\pi \approx 3.1416$.

23. Нехай числа a і b задані точно, $a > b > 0$. Показати, що внаслідок похибок заокруглення обчислення лівої і правої частини тотожності $a + b = \left(a^2 - b^2\right) / \left(a - b\right)$ можуть відрізнятись. Розглянути випадок, коли

$a = 3.14159$ і $b = 3.14158$, використовуючи десяткову арифметику з шістьма значущими цифрами.

24. Порівняти результати наближеного обчислення e^x за формулами $e^x \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!}$ і $e^x \approx 1 + x + \left(1 + \frac{x}{2}\left(1 + \frac{x}{3}\left(1 + \frac{x}{4}\left(1 + \frac{x}{5}\right)\right)\right)\right)$, для $x = 1,\ -1,\ 0.1$ і -0.1. Зробити висновок щодо точності обчислення.

25. Обчислити з простою і подвійною точністю інтеграл

$$E_n = \int_0^1 x^n e^{x-1} dx,\ n = \overline{1,10},$$

згідно з рекурентними формулами $E_n = 1 - nE_{n-1},\ n = \overline{2,10},\ E_1 = 1/e$ і $E_{n-1} = (1 - E_n)/n,\ n = \overline{20,2},\ E_{20} \approx 0$. Порівняти одержані результати щодо їх стійкості до обчислювальних похибок. У значенні $E_{10} \approx 0.0838771$ всі значущі цифри правильні.

26. Розв'язати аналогічну задачу з обчислення інтеграла

$$I_n = \int_0^1 \frac{x^n}{x + 10} dx,\ n = \overline{1,20},$$

застосувавши рекурентну формулу

$$I_k = -10I_{k-1} + 1.1,\ k = \overline{1,n},\ I_0 = \ln 1.1 = 0{,}0\ 953\ 101\ 798...,$$

$n = 5,10,15,20$ та іншу формулу

$$I_{k-1} = (1/10 - I_k)/10,\ k = m, m-1, ..., n+1,\ I_m \approx 0.$$

27. Проаналізувати результати обчислень із простою і подвійною точністю значень функцій $f(x) := (x + 1/3) - (x - 1/3)$ для значень $x = 1, 10^k,\ k = 3,6,9,10$ і 11 та функції $g(x) := ((3 + x^2/3) - (3 - x^2/3))/x^2$ для $x = 10^k,\quad k = \overline{-6,0}$. Точні значення $f(x) = 2/3$ для всіх $x \in R$, $g(x) = 2/3$ для $x \neq 0$.

28. Перетворити функції $1 - \cos x$ і $e^x - 1$ для $x \approx 0$ так, щоб обчислення їх значень було стійким до похибок. Наприклад, для великих $\sqrt{x+1} - \sqrt{x} = 1/(\sqrt{x+1} + \sqrt{x})$.

29. Нехай $f(x) = (n+1)x - 1$. Розглянемо ітерації

$$x_k = f(x_{k-1}),\ k = \overline{1,K};\ x_0 = 1/n.$$

Яким буде результат обчислень із простою й подвоєною точністю для $n = \overline{1,5},\ K = 10,20$ і 40?

30. У системі MathCad або Mathematica проаналізувати точність обчислення функцій на різних множинах значень:

а) $f(x) := \dfrac{1}{1+2x} - \dfrac{1-x}{1+x}$ для $-1 < x \le 1$ і $|x| \le 10^{-15}$;

б) $f(x) := \sqrt{x+1/x} - \sqrt{x-1/x}$ для $1 \le x \le 10$ і $2 \cdot 10^7 \le x \le 2 \cdot 10^8$.

31. Показати, що множина комплексних чисел, породжена системою за основою $2-i$ та базисними числами 0 і 1, є замкнена множина, яка містить деякий окіл початку координат й утворює фрактальну множину, яка називається "Двоголовий дракон" (рис. 1.5) [32].

Чи будь-яке комплексне число має "двійкове зображення" (базисні числа 0 і 1) в позиційній системі за основою $2+i$?

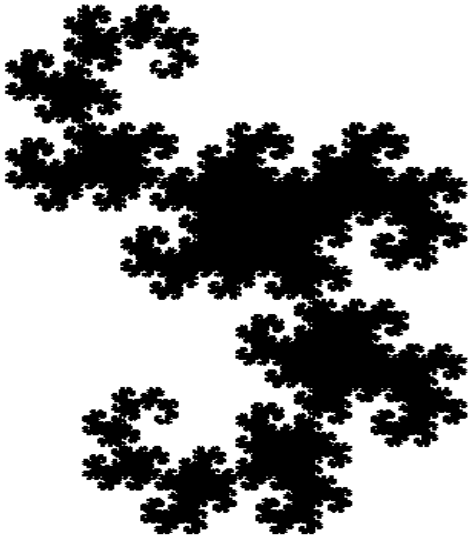


Рис. 1.5. Фрактал "Двоголовий дракон" (дракон Хартера–Хейтуея)

32. Сформувати нескінченно багато цілих чисел, трійкове зображення яких використовує тільки нулі й одиниці, а зображення за основою 4 – тільки нулі, одиниці та двійки [32].

33. Нехай x, y і z – числа в машині зі 32-розрядною мантисою. Оцінити похибку при обчисленні значення $(x+y)z$.

34. Нехай $|x| < 1, n$ – досить велике. В якому порядку обчислювати суму $\sum_{k=1}^{n} x^k$ з метою зменшення обчислювальної похибки?

35. Чи завжди правильно, що

$$fl\left(\frac{\alpha+\beta}{2}\right) \in [\alpha, \beta]?$$

36. Побудувати системи чисел вигляду $\pm(0.b_1b_2)\times 2^k$, де

а) $k \in \{-1, 0\}$; б) $k \in \{-1, 0, 1\}$; в) $k \in \{-2, -1, 0, 1, 2\}$.

Розділ 2. Прямі методи розв'язування СЛАР

Метод Ґауса та його модифікації. Схема з вибором головного елемента. Обґрунтування методу Ґауса, аналіз похибок методу та особливості його реалізації. Обчислення оберненої матриці та визначника. Метод квадратного кореня для СЛАР із симетричною матрицею. QR-метод та метод ортогоналізації. Метод прогонки для СЛАР із тридіагональною матрицею.

Література [5, 13, 14, 20–23, 36, 43, 56, 59, 73, 79, 80, 100]
Електронні джерела [1032–107]

2.1. Приклади СЛАР

Розглянемо СЛАР з квадратною невиродженою матрицею

$$Ax = b \tag{2.1}$$

або в розгорненому вигляді

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ &a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n. \end{aligned} \tag{2.2}$$

У прямих методах розв'язок системи рівнянь одержується за скінченну кількість арифметичних операцій. До прямих методів належать метод Ґауса та його модифікації, метод квадратного кореня для СЛАР із симетричними або ермітовими матрицями, метод ортогоналізації, QR-метод, метод прогонки для систем з тридіагональними матрицями та інші.

Прямі методи застосовуються для розв'язування СЛАР не дуже високого порядку, оскільки при їх розв'язуванні потрібно зберігати матрицю і результати проміжних обчислень в оперативній пам'яті комп'ютера. Також може відбутись нагромадження похибки у процесі розв'язування, оскільки в обчисленнях на кожному етапі використовуються результати попередніх операцій, виконаних із заокругленням, зокрема для СЛАР з погано обумовленою матрицею.

До СЛАР часто зводиться наближене розв'язування диференціальних задач шляхом різницевої апроксимації похідних, при побудові ітераційних наближень розв'язку нелінійних систем. Такі системи служать математичними моделями різноманітних

процесів, наприклад, модель Леонтьєва лінійного міжгалузевого балансу в матричній формі набуває вигляду [42]

$$x = Px + y,$$

де $P-$ матриця $n \times n$ коефіцієнтів прямих затрат, шуканий вектор $x-$обсяг виробленої продукції (вектор валового випуску), $n-$вектор y визначає обсяг продукції кінцевого споживача.

Для ланцюга, зображеного на рис. 2.1, ґрунтуючись на законі

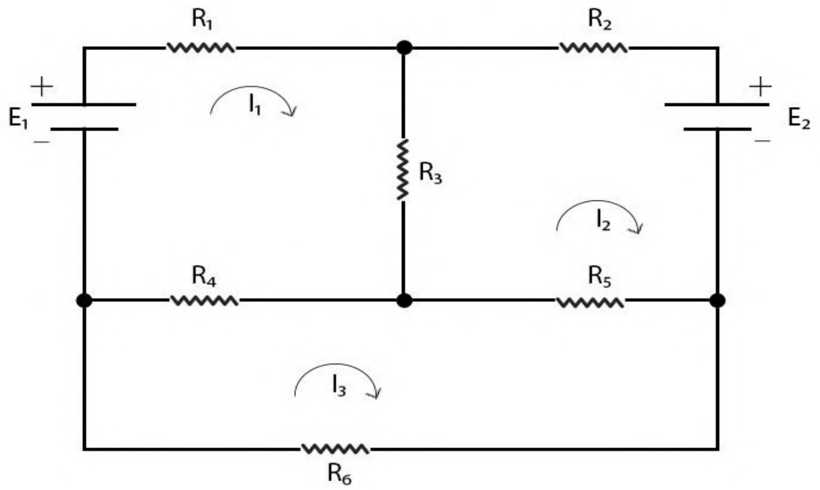


Рис. 2.1

Кірхгофа, одержується така СЛАР для знаходження значень струмів I_ν :

$$\begin{aligned}(R_1 + R_3 + R_4)I_1 + \quad R_3 I_2 + \quad I_3 &= E_1,\\ R_3 I_1 + (R_2 + R_3 + R_5)I_2 - \quad I_3 &= E_2,\\ R_4 I_1 - R_5 I_2 + (R_4 + R_5 + R_6)I_3 &= 0.\end{aligned}$$

2.2. Метод Ґауса

2.2.1. Схема методу. Метод полягає в послідовному вилученні невідомих. Розглянемо *схему єдиного ділення*. Припустимо, що $\det A \neq 0$. Нехай $a_{11} \neq 0$, інакше можна поміняти місцями перше рівняння з *i*-м, де $a_{i1} \neq 0$. Поділимо перше рівняння на a_{11} (називатимемо його *ведучим елементом*), а з решти рівнянь вилучимо x_1, помноживши перетворене перше рівняння на a_{i1}, $i = \overline{2,n}$, та віднявши його від *i*-го рівняння. Одержимо систему рівнянь

$$\begin{aligned}x_1 + u_{12}x_2 + \ldots + u_{1n}x_n &= y_1,\\ a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n &= b_2^{(1)},\\ &\ldots\ldots\ldots\ldots\ldots\ldots\\ a_{n2}^{(1)}x_2 + \ldots + a_{nn}^{(1)}x_n &= b_n^{(1)},\end{aligned} \tag{2.3}$$

де

$$u_{1j} = a_{1j} / a_{11},\ y_1 = b_1 / a_{11}; \qquad j = \overline{2,n}; \tag{2.4}$$

$$a_{ij}^{(1)} = a_{ij} - a_{i1} \cdot u_{1j},\ f_i^{(1)} = b_i - a_{i1} \cdot y_1; \qquad i, j = \overline{2,n}. \tag{2.5}$$

Нехай $a_{22}^{(1)}$, тоді аналогічні перетворення можна виконати для підсистеми (2.3), яка містить невідомі $x_2, x_3, ..., x_n$. Якщо такі перетворення можна продовжити до *n*-го рівняння, то в підсумку одержимо систему з верхньою трикутною матрицею вигляду

$$\begin{aligned} x_1 + u_{12}x_2 + u_{13}x_3 + ... + u_{1n}x_n &= y_1, \\ x_2 + u_{23}x_3 + ... + u_{2n}x_n &= y_2, \\ ..... & \\ x_{n-1} + u_{n-1n}x_n &= y_{n-1}, \\ x_n &= y_n. \end{aligned} \tag{2.6}$$

Нехай $a_{kj}^{(0)} = a_{kj};\ k, j = \overline{1,n}.$ На *k*-тому кроці, $k = \overline{1,n-1}$ коефіцієнти перетвореної системи обчислюються за формулами:

$$u_{kj} = a_{kj}^{(k-1)} / a_{kk}^{(k-1)},\ j = \overline{k+1,n}; \tag{2.7}$$

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - u_{kj} \cdot a_{ik}^{(k-1)};\ i, j = \overline{k+1,n}. \tag{2.8}$$

Праві частини:

$$y_k = b_k^{(k-1)} / a_{kk}^{(k-1)}; \tag{2.9}$$

$$b_i^{(k)} = b_i^{(k-1)} - a_{ik}^{(k-1)} \cdot y_k, \qquad i = \overline{k+1,n}. \tag{2.10}$$

На останньому кроці знаходимо

$$y_n = b_n^{(n-1)} / a_{nn}^{(n-1)}. \tag{2.11}$$

Описаний процес можливий, якщо всі ведучі елементи $a_{kk}^{(k-1)} \neq 0$, оскільки в ході процесу відбувається ділення на ці елементи. Із системи (2.6) послідовно знаходимо невідомі:

$$x_n = y_n, \qquad x_i = y_i - u_{ii+1}x_{i+1} - ... - u_{in}x_n, \qquad i = \overline{n-1,1}.$$

Процес знаходження коефіцієнтів трикутної системи називається *прямим*, а процес отримання її розв'язків – *оберненим ходом* методу Ґауса.

Послідовне вилучення невідомих, що перетворює дану систему у систему з трикутною матрицею, можна проводити й за іншими обчислювальними схемами. У схемі *ділення та віднімання* на кожному кроці діляться всі рівняння на коефіцієнт при змінній, яка вилучається, а потім саме вилучення

здійснюється відніманням одного рівняння від решти рівнянь. У схемі *множення та віднімання* на першому кроці невідома x_1 вилучається з i-го рівняння за допомогою множення цього рівняння на a_{11} і віднімання 1-го рівняння, помноженого на a_{i1}. На наступних кроках застосовується цей же прийом, так що коефіцієнти допоміжних рівнянь $\tilde{a}_{ij}^{(k)}$ на k-ому кроці обчислюються за формулами

$$\tilde{a}_{ij}^{(k)} = \tilde{a}_{kk}^{(k-1)}\tilde{a}_{ij}^{(k-1)} - \tilde{a}_{ik}^{(k-1)}\tilde{a}_{kj}^{(k-1)}, \quad \tilde{b}_{ik} = \tilde{a}_{kk}^{(k-1)}b_i^{(k-1)} - \tilde{a}_{ik}^{(k-1)}\tilde{b}_k^{(k-1)}.$$

2.2.2. Оцінка складності методу Ґауса. Оцінимо складність методу Ґауса за кількістю арифметичних операцій, зокрема, множення і ділення, які необхідні для реалізації методу. Для зведенні системи (2.1) до системи з трикутною матрицею та її розв'язання необхідно виконати

$S_r = \frac{n}{6}\left(2n^2 + 9n + 1\right) = \frac{n^3}{3} + \frac{3n^2}{2} + \frac{n}{6} = O\left(n^3\right)$ операцій множення або

$$S_r \approx \frac{n^3}{3}. \qquad (2.12)$$

Для порівняння оцінимо складність розв'язування СЛАР за правилом Крамера:

$$x_i = \frac{\Delta_i}{\Delta}, \quad i = \overline{1,n},$$

де Δ_i – допоміжні визначники. Обчислення визначника згідно з означенням вимагає $n!\left(n-1\right)$ множень. Тому обчислення за формулами Крамера вимагатиме $S_k = \left(n+1\right)n!\left(n-1\right) = \left(n^2-1\right)n!$ операцій множення.

Якщо на комп'ютері виконується мільйон операцій множень або ділень за секунду, то для знаходження розв'язку за правилом Крамера потрібно 0.0003 секунди, за методом Ґауса – $8\cdot10^{-6}$ секунд, якщо $n=5$, для $n=10$ відповідно 36 і $5\cdot10^{-4}$ секунд, а для $n=15$ вже 170 діб і $1.5\cdot10^{-3}$ секунд відповідно.

2.3. Схема Йордана

На першому кроці прямого ходу вилучається x_1 з усіх рівнянь, крім першого. На другому кроці x_2 вилучається також з усіх рівнянь, крім другого, і система набуває вигляду

$$x_1 + a_{13}^{(2)} x_3 + \ldots + a_{1n}^{(2)} x_n = b_1^{(2)},$$
$$x_2 + a_{23}^{(2)} x_3 + \ldots + a_{2n}^{(2)} x_n = b_2^{(2)},$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$a_{n3}^{(2)} x_3 + \ldots + a_{nn}^{(2)} x_n = b_n^{(2)}.$$

На останньому кроці x_n вилучається з перших $n-1$ рівнянь. У результаті одержуємо розв'язок $x_v = b_v^{(n)}$, $v = \overline{1,n}$, системи рівнянь.

У схемі Йордана матриця A у системі (2.1) еквівалентними перетвореннями зводиться до одиничної матриці I, що рівносильно множенню матриці A на обернену матрицю. Тому, виконавши ці ж перетворення над матрицею I, одержимо обернену матрицю A^{-1}.

2.4. Метод Ґауса з вибором головного елемента

Кожну із описаних схем можна застосовувати для вилучення невідомих як з наперед визначеним порядком, так і вибираючи порядок вилучення в ході процесу. Зокрема, у схемі єдиного ділення потрібно вимагати, щоб усі ведучі елементи були відмінними від нуля. Близькість ведучих елементів до нуля може бути причиною нагромадження обчислювальної похибки, зменшити яку можна за допомогою методом Ґауса з вибором *головного елемента*. При виборі головного елемента в стовпці на k-му кроці прямого ходу за ведучий береться найбільший за абсолютною величиною елемент у перетворюваній частині k-го стовпця, тобто

$$|a_{kk}| = \max_i |a_{ik}|, \quad i = \overline{k+1,n}.$$

Згідно з іншим способом на k-му кроці ведучим вибирається найбільший за абсолютною величиною елемент у перетворюваній частині матриці, тобто

$$|a_{kk}| = \max_{i,j} |a_{ij}|, \quad i, j = \overline{k,n}.$$

У першому способі кількість порівнянь чисел та обчислень абсолютних величин складає $(2n-1)+(2n-3)+\ldots+3 = n^2-1$, що на порядок менше, ніж у методі Ґауса. Якщо вибір ведучого елемента здійснювати в матриці, то число таких операцій складе

$(2n^2-1)+(2(n-1)^2-1)+...+2\cdot 2^2-1=2(1^2+2^2+...+n^2)-n-1=$
$=(2n^3+3n^2-2n-3)/3\approx 2n^3/3$, тобто такий же порядок як в методі Ґауса.

2.5. Обґрунтування методу Ґауса

Якщо прямий хід методу Ґауса завершений, то система (2.1) зводиться до системи рівнянь вигляду

$$Ux=y \tag{2.13}$$

з верхньою трикутною матрицею

$$U=\begin{bmatrix} 1 & u_{12} & ... & u_{1n} \\ 0 & 1 & ... & u_{2n} \\ .... & .... & ... & ... \\ 0 & 0 & ... & 1 \end{bmatrix}.$$

Елементи правої частини (2.1) перетворюються за формулами:

$b_1=a_{11}y_1,\ \mathrm{b}_2=a_{21}y_1+a_{22}^{(1)}y_2,...,\ b_k=l_{k1}y_1+...+l_{k,k-1}y_{k-1}+l_{kk}y_k$,

де $l_{k1},...,l_{k,k-1}$ – деякі коефіцієнти, причому

$$l_{11}=a_{11}\neq 0;\ l_{22}=a_{22}^{(1)}\neq 0;\ \ l_{kk}\neq 0,\ k=\overline{3,n}.$$

Тому

$$Ly=b, \tag{2.14}$$

де L – нижня трикутна матриця з діагональними елементами, які не дорівнюють нулю, і має вигляд

$$L=\begin{bmatrix} l_{11} & 0 & ... & 0 \\ l_{21} & l_{22} & ... & 0 \\ ... & ... & ... & ... \\ l_{n1} & l_{n2} & ... & l_{nn} \end{bmatrix},\ l_{ii}\neq 0,\ \ i=\overline{1,n}.$$

Підставивши y з (2.13) в (2.14), одержимо

$$LUx=b. \tag{2.15}$$

Із порівняння систем рівнянь (2.1) і (2.15) випливає, що реалізація методу Ґауса рівносильна розкладу матриці A на добуток матриць L і U:

$$A=LU, \tag{2.16}$$

де матриці L і U задовольняють зазначені вище умови, відтак до розв’язування двох систем: (2.14) з нижньою трикутною

матрицею для знаходження вектора y і системи (2.13) з верхньою трикутною матрицею U, з якої просто одержати розв'язок x.

Елементи матриць L і U можна знайти з матричного рівняння

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

Звідси маємо: $l_{i1} = a_{i1},\ l = \overline{1,n};\ l_{11}u_{1j} = a_{1j}$, отже, $u_{1_j} = a_{1_j} / l_{11},\ \ j = \overline{2,n};$

$$a_{ij} = l_{ii} + \sum_{j=1}^{i-1} l_{ij}u_{ji}, \quad i = \overline{2,n}.$$

Враховуючи, що $u_{kj} = 0$ для $k > j$ і $l_{ik} = 0$ для $k > i$ одержимо:

$$u_{ij} = \left(a_{ii} - \sum_{k=1}^{i-1} l_{ik}u_{ik}, \right) / l_{ii}, \quad i = \overline{2,n-1}, \quad j = \overline{i+1,n};$$

$$l_{ij} = \left(a_{ji} - \sum_{k=1}^{i-1} l_{jk}u_{kj}, \right) / l_{ii}, \quad i = \overline{2,n-1}, \quad j = \overline{i+1,n}.$$

Результат факторизації матриці A у вигляді добутку LU зручно зберігати в одній матриці вигляду

$$\begin{bmatrix} l_{11} & u_{12} & u_{13} \cdots & u_{1n} \\ l_{21} & l_{22} & u_{23} \cdots & u_{2n} \\ \cdots & \cdots & \cdots \cdots & \cdots \\ l_{n1} & l_{n2} & l_{n3} \cdots & l_{nn} \end{bmatrix}.$$

Позначимо через Δ_i – головний мінор порядку i матриці A:

$$\Delta_1 = a_{11},\ \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \dots,\ \Delta_n = \det A.$$

Теорема 2.1 *(теорема про LU – розклад)* [59]. *Якщо всі головні мінори матриці A не дорівнюють нулю, то її можна зобразити у вигляді (2.16), де L – нижня трикутна матриця з ненульовими діагональними елементами, U – верхня трикутна матриця з одиницями на головній діагоналі. Такий розклад єдиний.*

Доведення. Доведення проведемо методом математичної індукції. Для $n = 2$ маємо

$$\begin{bmatrix} l_{11} & 0 \\ l_{12} & l_{22} \end{bmatrix}\begin{bmatrix} 1 & u_{12} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Для знаходження l_{11}, l_{12}, l_{22} і u_{12} одержимо систему лінійних рівнянь

$$l_{11} = a_{11}, \quad l_{11}u_{12} = a_{12},$$
$$l_{21} = a_{21}, \quad l_{21}u_{12} + l_{22} = a_{22}.$$

Звідси знаходимо елементи матриць L і U:

$$l_{11} = a_{11}, \quad l_{21} = a_{21}, \quad u_{12} = a_{12}/a_{11},$$
$$l_{22} = a_{22} - l_{21}u_{12} = (a_{11}a_{22} - a_{21}a_{12})/a_{11} = (\det A)/a_{11} \neq 0.$$

Нехай твердження теореми справджується для матриці порядку $k-1 < n$. Доведемо, що воно правильне для матриці порядку k. Запишемо матрицю A у вигляді

$$A = \left[\begin{array}{c|c} A_{k-1} & a_{k-1} \\ \hline f_{k-1} & a_{kk} \end{array}\right], \; A_{k-1} = \begin{bmatrix} a_{11} & \dots & a_{1,k-1} \\ \dots & \dots & \dots \\ a_{k-1,1} & \dots & a_{k-1,k-1} \end{bmatrix},$$

$$a_{k-1} = \begin{bmatrix} a_{1k} \\ \vdots \\ a_{k-1,k} \end{bmatrix}, \; f_{k-1}^T = \begin{bmatrix} a_{k1} \\ \vdots \\ a_{k,k-1} \end{bmatrix}.$$

Згідно з припущенням

$$A_{k-1} = L_{k-1}U_{k-1}$$

і матриці L_{k-1}, U_{k-1} задовольняють умови, такі ж як для матриць L і U. Нехай

$$A = L_k U_k = \begin{bmatrix} L_{k-1} & 0 \\ l_{k-1} & l_{kk} \end{bmatrix}\begin{bmatrix} U_{k-1} & u_{k-1} \\ 0 & 1 \end{bmatrix},$$

де $l_{k-1} = (l_{k1}, \dots, l_{kk-1}), u_{k-1} = (u_{1k}, \dots, u_{k-1,k})^T$ – невідомі поки що вектори. Також потрібно знайти число l_{kk}. Маємо

$$l_{k-1}u_{k-1} + l_{kk} = a_{kk}, \; l_{k-1}U_{k-1} = f_{k-1}, \; L_{k-1}u_{k-1} = a_{k-1}.$$

Матриці L_{k-1}^{-1} і U_{k-1}^{-1} існують, тому

$$u_{k-1} = L_{k-1}^{-1}a_{k-1}, \quad l_{k-1} = f_{k-1}U_{k-1}^{-1}, \quad l_{kk} = a_{kk} - l_{k-1}u_{k-1}.$$

Перевіримо, що $l_{kk} \neq 0$. Справді,

$$\det A = \det L_{k-1} \cdot \det U_{k-1} = \det L_k \cdot \det U_k = (\det L_{k-1})l_{kk}.$$

Оскільки $\det L_{k-1} \neq 0$ і $\det A \neq 0$, то $l_{kk} \neq 0$.

Доведемо єдиність розкладу (2.16). Нехай матрицю A можна розкласти на два добутки:

$$A = L_1U_1 \text{ і } A = L_2U_2.$$

Тоді $L_1U_1 = L_2U_2$, звідки $U_1U_2^{-1} = L_1^{-1}L_2$. Неважко перевірити, що матриця в лівій частині – верхня трикутна, а в правій частині – нижня трикутна. Така рівність можлива, якщо ліва і права частини одержаної рівності – діагональні матриці. Але діагональні елементи $U_1U_2^{-1} = L_1^{-1}L_2 = I_n$, де I_n – одинична матриця. Тому $U_1 = U_2$, $L_1 = L_2$ тобто розклад єдиний. ■

Приклад 2.1. Для матриці $A = \begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}$ побудуємо розклад (2.17). Маємо: $l_{11} = 2,\ u_{12} = 1/2,\ l_{21} = 3,\ l_{22} = 5/2.$ Отже,

$$\begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 3 & 5/2 \end{bmatrix}\begin{bmatrix} 1 & 1/2 \\ 0 & 1 \end{bmatrix}.$$

Приклад 2.2. Нехай $A = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 3 & 6 \\ 4 & 5 & 2 \end{bmatrix}$. Тут мінори $\Delta_1 = \Delta_2 = 1 > 0,\quad \Delta_3 = \det A = 20 > 0$, отже, умови теореми 2.1 виконані. Із розкладу

$$\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix}\begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 3 & 6 \\ 4 & 5 & 2 \end{bmatrix}$$

одержимо $l_{11} = l_{21} = 1,\quad l_{31} = 4,\quad u_{12} = 2,\quad u_{13} = 0;\quad l_{21}u_{12} + l_{22} = 3,$ $l_{21}u_{13} + l_{22}u_{23} = 6,\quad l_{31}u_{12} + l_{32} = 5,\quad l_{31}u_{13} + l_{32}u_{23} + l_{33} = 2.$ Розв'язавши систему, дістанемо $l_{22} = 1,\quad u_{23} = 6,\quad l_{32} = -3,\quad l_{33} = 20.$ Отже,

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 4 & -3 & 20 \end{bmatrix},\quad U = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 6 \\ 0 & 0 & 1 \end{bmatrix}$$

Розглянемо тепер метод Ґауса з вибором головного елемента у стовпці. Якщо на i-му кроці, $1 \le i \le n-1$, вибирається головний елемент з k-го рядка, $i+1 \le k \le n$, то потрібно поміняти місцями

рядки з номерами k та i. Це виконується множенням перетворюваної матриці на матрицю елементарних перестановок

$$P_{ik} = \begin{bmatrix} 1 & & | & & & | & & \\ & 1 & | & & & | & & \\ - & - & 0 & - & - & 1 & - & - \\ & & | & 1 & & | & & \\ & & | & & 1 & | & & \\ - & - & 1 & - & - & 0 & - & - \\ & & | & & & | & 1 & \\ & & | & & & | & & 1 \end{bmatrix} \begin{matrix} \\ \\ i \\ \\ \\ k \\ \\ \\ \end{matrix}$$

$$\begin{matrix} & & \mathrm{i} & & & k & & \end{matrix}$$

Матриця P_{ik} одержується із одиничної матриці I перестановкою рядків з номерами k та i. Наприклад, щоб поміняти місцями 1-й і 3-й рядки:

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 4 & -1 & -1 \\ 1 & 2 & -2 \\ 2 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -1 & 1 \\ 1 & 2 & -2 \\ 4 & -1 & -1 \end{bmatrix}.$$

Можна показати [10], що метод Ґауса з вибором головного елемента рівносильний цьому ж методу, який застосовується до системи $PAx = Pb$, де P – деяка матриця перестановок, одержана з одиничної матриці. Схема з вибором головного елемента обґрунтовується наступною теоремою.

Теорема 2.2 [59, с. 65-67]. *Якщо* $\det A \neq 0$*, то існує матриця перестановок P така, що матриця PA має головні мінори, які не дорівнюють нулю, і справджується розклад* $PA = LU$*, де L – нижня трикутна матриця з відмінними від нуля діагональними елементами, U – верхня трикутна матриця з одиничною головною діагоналлю.* ■

2.6. Аналіз похибок у методі Ґауса

Ґрунтуючись на формулах (2.18) і, як показано в [29, 35], можна одержати оцінки похибок вилучень невідомих за схемою Ґауса. Нехай в арифметиці *P(p, t, L, U)* не відбувається переповнень і машинних нулів, $fl(x_i \circ y_i) = (x_i \circ y_i)(1+\delta)$, де $|\delta| \le p^{1-t} = 2\varepsilon$. Тоді

$$fl(\sum_{i=1}^{n} x_i y_i) = \sum_{i=1}^{n} x_i y_i (1+\delta_i), \text{ де } |\delta_i| \le n\varepsilon / 2.$$

наслідок похибок заокруглень [29, 33] матриця $A = LU + \Delta$, причому $\|\Delta\| \le n\varepsilon \|L\| \cdot \|U\|$, де $\|\cdot\|$ – деяка норма матриці, наприклад $\|A\| = (\sum_{i,j=1}^{n} a_{ij}^2)^{1/2}$. Розглянемо, як це впливає на розв'язування задачі $LUx = b$, якщо розв'язуються окремо системи рівнянь $Ly = b$ і $Ux = y$. Розв'язавши систему (2.14), одержується наближений розв'язок $\tilde{y}$, який задовольняє систему $(L + \delta L)\tilde{y} = b$, де $\|\delta L\| \le n\varepsilon \|L\|$. Аналогічно, розв'язуючи систему $Ux = \tilde{y}$, одержимо вектор $\tilde{x}$, який задовольняє систему $(U + \delta U)\tilde{x} = \tilde{y}$ і $\|\delta U\| \le n\varepsilon \|U\|$.

Об'єднуючи ці співвідношення, одержимо

$$b = (L + \delta L)\tilde{y} = (L + \delta L)(U + \delta U)\tilde{x} \equiv (A + \delta A)\tilde{x},$$

де $\delta A = -E + L \cdot \delta U + \delta L \cdot U + \delta L \cdot \delta U$.

На підставі оцінок для $\delta U, \delta L, E$ і нерівності трикутника оцінимо норму δA:

$$\|\delta A\| \le \|E\| + \|L \cdot \delta U\| + \|\delta L \cdot U\| + \|\delta L \cdot \delta U\| \le$$
$$\le 3n\varepsilon \|L\| \cdot \|U\| + n^2 \varepsilon^2 \|L\| \cdot \|U\| \approx 3n\varepsilon \|L\| \cdot \|U\|.$$

Отже, вилучення невідомих згідно з методом Ґауса є стійким алгоритмом, якщо правильне співвідношення

$$3n\varepsilon \|L\| \cdot \|U\| = O(\varepsilon)\|A\|.$$

Саме тоді величина $\|\delta A\| / \|A\| = O(\varepsilon)$.

Як підтверджується обчислювальною практикою, метод Ґауса з вибором головного елемента у стовпці майже завжди забезпечує виконання співвідношення $\|L\| \cdot \|U\| \approx \|A\|$. Метод гарантує, що $|l_{ij}| \le 1$. У [33, 35] визначений коефіцієнт росту $g_{pp} = \|U\|_{\max} / \|A\|_{\max}$, де $\|A\|_{\max} = \max_{i,j} |a_{ij}|$. Тобто стійкість методу рівносильна тому, що число g_{pp} мале або повільно росте із зростанням *n*. На практиці g_{pp} майже завжди не перевищує *n*.

Можна припустити, що в середньому поведінка g_{pp} описується функцією $\sqrt[3]{n^2}$ або й $\sqrt{n}$. Але існують приклади, коли $g_{pp} = 2^{n-1}$ [20].

Використавши оцінки $\|L\|_\infty \le n$ і $\|U\|_\infty \le n g_{pp} \|A\|_\infty$, одержуємо

$$\|\delta A\|_\infty \le 3 g_{pp} n^3 \varepsilon \|A\|_\infty .$$

Зробимо деякі висновки.

1. Доведено [56], що повний вибір ведучого елемента є надійною стратегією, оскільки похибка зростає не дуже сильно і коефіцієнт зростання не перевищує

$$f(n) = \left(n \cdot 2 \cdot 3^{1/2} \cdot 4^{1/3} \cdot \ldots \cdot n^{1/(n-1)} \right)^{1/2} .$$

Похибка в коефіцієнтах системи чи похибка в процесі обчислень, яка впливає на розв'язок, множиться на f_n. Наведемо значення $f(n)$ для деяких n:

n	2	10	15	20	30	50	80	100
f_n	5,7	18,3	39,1	69,8	155,5	536,2	1915,5	3552,4

2. В обчислювальній практиці ймовірність значного зростання похибки при виборі ведучого елемента в стовпці досить мала і не перевищує в 2-4 рази значення $f(n)$.

3. Побудована система рівнянь, для якої вибір ведучого елемента в стовпці дає коефіцієнт росту похибки, що дорівнює 2^n. Для $n = 10$, це 1024 раз, $n = 20$ – більше 10^6, а для $n = 100$ вже складає $\approx 10^{30}$.

4. Без певного вибору ведучого елемента похибка заокруглення може значно зрости. Зокрема показано, що для системи порядку 20, можна очікувати зростання похибки в 10^{12} і більше разів. При виконанні прямого ходу методу Ґауса зростання модулів елементів ілюструє перетворення матриці:

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{bmatrix}.$$

5. В алгоритмі методу Ґауса здійснюється перевірка на дорівнювання нулю ведучого елемента. Але в арифметиці з плаваючою крапкою числа, які мають бути нулями у звичайній арифметиці, майже завжди не дорівнюють нулю внаслідок похибок заокруглення. Тому потрібно враховувати як зображення чисел у системі рівнянь, так і чисел при реалізації програми. Якщо на комп'ютері виконуються дії з $t = 10$ десятковими

знаками, то ведучий елемент $1.5*10^{-10}$ можна вважати результатом заокруглення і присвоїти йому значення 0. Якщо ж елементи матриці мають порядок 10^{-9}, то цього стверджувати вже не можна. Тому доцільно масштабувати систему $Ax=b$, наприклад, помножити елементи рівнянь на деякий множник так, щоб нові значення мали порядок 1. Можна помножити стовпці матриці, що відповідає зміні одиниць виміру невідомих, наприклад, перехід від метрів до міліметрів. Наведемо такий приклад:

$$\begin{bmatrix} 10^{-7} & 10 & -10^{-10} \\ 10^{-6} & 15 & 10^{-4} \\ 10^{-8} & 20 & 10^{-9} \end{bmatrix} \to \begin{bmatrix} 1 & 1.0 & -1 \\ 10 & 1.5 & 0.1 \\ 0.1 & 2.0 & 104 \end{bmatrix}$$

де $y_1 = 10^7 x_1$, $y_3 = 10^{10} x_3$, $y_2 = 10^{-1} x_2$.

Доцільним способом масштабування є множення на відповідний множник так, щоб найбільший елемент рядка дорівнював 1. Перевірка на нуль ведучого елемента здійснюється порівнянням з числом $c \cdot p^{-t}$, де p – основа системи числення, t – кількість знаків після коми при реалізації алгоритму на комп'ютері, c – деяке невелике число, яке зростає зі збільшенням порядку n. Оскільки над кожним елементом a_{ij} виконується найбільше $2n^2$ арифметичних операцій, то найбільше значення c, яке варто розглядати, дорівнює n^2. Похибки заокруглення майже взаємно знищуються, тому $c = n$ або довільне невелике число.

2.7. Обчислення визначника й оберненої матриці

Схему методу Ґауса можна застосувати для обчислення визначників. Зупинимось на схемі єдиного ділення. Нехай $a_{11} \neq 0$, тоді

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} = a_{11} \begin{vmatrix} 1 & b_{12} & \dots & b_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}.$$

Далі, від кожного рядка віднімемо перший рядок, помножений відповідно на перший елемент цього рядка. Отримаємо

$$\Delta = a_{11}\begin{vmatrix} 1 & b_{12} \ldots b_{1n} \\ 0 & a_{22}^{(1)} \ldots a_{2n}^{(1)} \\ \ldots\ldots\ldots\ldots\ldots \\ 0 & a_{n2}^{(1)} \ldots a_{nn}^{(1)} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22}^{(1)} \ldots a_{2n}^{(1)} \\ \ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots \\ a_{n2}^{(1)} \ldots a_{nn}^{(1)} \end{vmatrix}.$$

У визначнику $(n-1)$-го порядку виконаємо аналогічні перетворення, якщо тільки $a_{22}^{(1)} \neq 0$. Продовжуючи цей процес, отримаємо

$$\Delta = a_{11}^{(0)}{}_{11}a_{22}^{(1)}\ldots a_{nn}^{(n-1)},\ a_{11}^{(0)} = a_{11}.$$

Якщо на деякому кроці $a_{ii}^{(i-1)} = 0$ і $a_{ki}^{(i-1)} = 0, \quad k < i$, то $\Delta = 0$. Оптимальний результат одержується при застосуванні схеми Ґауса з вибором головного елемента. Кількість множень і ділень, необхідних для обчислення визначника *n*-го порядку, дорівнює

$$n^2 + (n-1)^2 + \ldots + 2^2 + n - 1 = (n-1)(2n^2 + 5n + 12)/6.$$

Зауважимо, що в цьому випадку потрібно врахувати знак визначника від перестановки рядків.

Задача розв'язування СЛАР і задача обчислення оберненої матриці пов'язані між собою. Справді, якщо для матриці *A* відома її обернена матриця, то отримаємо

$$x = A^{-1}b.$$

Навпаки, визначення елементів оберненої матриці можна звести до розв'язування *n* систем рівнянь вигляду

$$A\bar{x}_i = \bar{e}_i, \quad i = \overline{1,n},$$

де $\bar{x}_i$ – *i*-й стовпець матриці $X = A^{-1}$, $\bar{e}_i$ – *i*-й стовпець одиничної матриці *I*. Останнє випливає з означення оберненої матриці $AA^{-1} = I$ і правила множення матриць. При цьому зводити матрицю *A* до трикутного вигляду потрібно тільки один раз.

Числове розв'язування *n* систем рівнянь, що дають елементи оберненої матриці, можна здійснити, наприклад, за схемою єдиного ділення для декількох систем зі спільною матрицею коефіцієнтів. У результаті отримаємо матрицю, яка складається із рядків оберненої матриці, розміщених у протилежному порядку. Для прямого ходу потрібно виконати $n^3/3$ дій множення і ділення. Обернений хід вимагає здійснення $n\left(\frac{n(n+1)}{2} + \frac{n(n-1)}{2}\right) = n^3$ операцій множення і ділення. Отже,

загальна кількість операцій становить приблизно $4n^3/3$. Для контролю обчислення і оцінки точності результату можна здійснити множення A на A^{-1}.

Зауваження 2.1. *Існують ітераційні методи знаходження оберненої матриці. Наведемо один з них, відомий як метод Шульца*[1]. *Нехай матриці* U_0 *і* Y_0 *такі, що існує* U_0^{-1}, $\Psi_0 = I - AU_0$ *і* $\|\Psi_0\| < 1$. *Тоді існує матриця* A^{-1} *і до неї збігається послідовність матриць* U_k, *визначених ітераційним процесом*

$$\Psi_k = I - AU_0,\ U_{k+1} = U_k(I + \Psi_k + ... + \Psi_k^m),\ m \geq 1,\ k = 0, 1, ...$$

Найчастіше використовується метод другого порядку $(m = 1)$, *при цьому*

$$\|A^{-1} - U_k\| \leq \frac{\|U_0\|}{1 - \|\Psi_0\|} \|\Psi_0\|^{2^k}.$$

Детальніше метод наведено в [13].

2.8. Метод квадратного кореня

Цей метод є модифікацією методу Ґауса на випадок систем лінійних рівнянь із симетричною (у більш загальному випадку ермітовою) матрицею, тобто коли $A^T = A$, $a_{ij} \in \mathbf{R}$. Симетричність матриці дозволяє одержати розклад матриці у вигляді

$$A = S^T DS, \tag{2.17}$$

де S – верхня трикутна матриця з додатними діагональними елементами, D – діагональна матриця з елементами ± 1. Матриця D потрібна для того, щоб елементи s_{ii} були дійсними числами. Розклад (2.17) є наслідком теореми про LU-розклад для симетричної матриці A, коли всі головні мінори матриці A, які не дорівнюють нулю [25, 73].

Систему (2.1), врахувавши (2.17), можна записати у вигляді двох систем:

$$S^T Dy = b, \tag{2.18}$$

$$Sx = y. \tag{2.19}$$

[1] Schulz G. Iterative berechnung der reziproken matrix // ZAMM. – 1933, №13. – P. 57-59.

Перша із них – система із нижньою трикутною, а друга – з верхньою трикутною матрицею. Розв'язування систем (2.18) і (2.19) вимагає $2(1+2+...n) = n^2 + n$ операцій множення і ділення.

Розглянемо реалізацію методу для системи другого порядку. Маємо

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} = \begin{bmatrix} s_{11} & 0 \\ s_{12} & s_{22} \end{bmatrix} \cdot \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix} \cdot \begin{bmatrix} s_{11} & s_{12} \\ 0 & s_{22} \end{bmatrix}.$$

Покажемо, що при виконанні умов $a_{11} \neq 0$ і $\det A \neq 0$ можна підібрати $s_{ii} > 0$ і $d_i = \pm 1,\ i = 1,2$, так, щоб виконувалась рівність (2.17). Виконавши множення в правій частині, одержимо

$$\begin{bmatrix} s_{11}^2 & s_{11}s_{12}d_1 \\ s_{11}s_{12}d_1 & s_{12}^2 d_1 + s_{22}^2 d_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}.$$

Прирівнявши відповідні елементи матриць, отримаємо систему рівнянь для знаходження матриць *S* і *D*:

$$s_{11}^2 d_1 = a_{11}, \quad s_{11}s_{12}d_1 = a_{12}, \quad s_{12}^2 d_1 + s_{22}^2 d_2 = a_{22}.$$

Нехай $d_1 = sign(a_{11})$. Тоді $s_{11}^2 = a_{11}d_1 > 0$ і

$$s_{11} = \sqrt{d_1 a_{11}} > 0. \qquad (2.20)$$

Далі,

$$s_{12} = a_{12} / s_{11}d_1,\ \ s_{22}^2 d_2 = a_{22} - s_{12}^2 d_1\ .$$

Візьмемо $d_2 = sign(a_{22} - s_{12}^2 d_1)$. Тоді

$$s_{22} = \sqrt{(a_{22} - s_{12}^2 d_1)d_2}\ .$$

Зауважимо, що $a_{22} - s_{12}^2 d_1 = a_{22} - \dfrac{a_{12}^2}{a_{11}} = (a_{11}a_{22} - a_{12}^2)/a_{11} = a_{11}^{-1} \det A \neq 0$,

отже, $s_{22} > 0$.

Розглянемо систему порядку $n \geq 2$. Нехай головні мінори $\Delta_i \neq 0, i = \overline{1,n}$. Для знаходження елементів s_{ij} одержується система рівнянь

$$\sum_{l=1}^{i-1} s_{li}s_{lj}d_l + s_{ii}s_{ij}d_i + \sum_{l=i+1}^{n} s_{li}s_{lj}d_l = a_{ij},\ \ i, j = \overline{1,n}.$$

Враховуючи, що $s_{li} = 0$ для $l > i$, одержимо

$$s_{ii}s_{ij}d_i + \sum_{l=1}^{i-1} s_{li}s_{lj}d_l = a_{ij},\ \ i \leq j\ . \qquad (2.21)$$

Для $i=1$ значення d_1 і s_{11} визначаються згідно з (2.20), а

$$s_{1j} = a_{1j} / s_{11} d_1, \quad j = \overline{2,n}.$$

Якщо $i = j \geq 2$, то $s_{ii}^2 d_i = a_{ii} - \sum_{l=1}^{i-1} s_{li}^2 d_l$, тобто

$$d_i = sign(a_{ii} - \sum_{l=1}^{i-1} s_{li}^2 d_l), \quad s_{ii} = \sqrt{a_{ii} - \sum_{l=1}^{i-1} s_{li}^2 d_l}, \quad i = \overline{2,n}.$$

При $i < j$ із (2.21) випливає

$$s_{ij} = (a_{ij} - \sum_{l=1}^{i-1} s_{li} s_{lj} d_l) / (s_{ii} d_i), \quad j = \overline{i+1,n}.$$

Елементи вектора *y* визначаються за рекурентними формулами:

$$y_1 = \frac{b_1}{s_{11}}; \quad y_i = \frac{b_i - \sum_{\nu=1}^{i-1} s_{\nu i} y_\nu}{s_{ii}}, i = \overline{2,n}.$$

Розв'язок системи рівнянь $Ax = b$ знаходиться за формулами

$$x_n = \frac{y_n}{s_{nn}}, \quad x_i = \frac{y_i - \sum_{\nu=i+1}^{n} s_{i\nu} x_\nu}{s_{ii}}, i = \overline{n-1,1}.$$

Обернений хід, який полягає в розв'язуванні систем (2.18) і (2.19), вимагає n^2+n операцій множення і ділення. Побудова розкладу (2.17) – $n(n-1)(n+4)/6$ таких операцій і n операцій обчислення квадратного кореня.Загальна кількість зазначених операцій складає

$$n(n+1) + \frac{n(n-1)(n+4)}{6} + n = \frac{n(n^2+9n+2)}{6}.$$

Для великих *n* це приблизно $n^3/6$, що майже вдвічі менше, ніж у методі Ґауса.

2.9. QR-метод

Серед методів, для реалізації яких потрібно приблизно $4n^3/3$ операцій, QR-метод на даний час найбільш стійкий до нагромадження обчислювальної похибки.

Метод ґрунтується на розкладі матриці СЛАР на добуток ортогональної матриці Q і верхньої трикутної матриці R:

$$A = QR. \qquad (2.22)$$

Ортогональною називається матриця, для якої виконується умова

$$Q^T Q = I,$$

або $Q^T = Q^{-1}$. Наприклад, матриця, що задає на площині перетворення повороту на кут φ:

$$\begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix}.$$

Теорема 2.2 [19, 23]. *Будь-яку дійсну невироджену матрицю можна розкласти на добуток ортогональної матриці на праву трикутну матрицю.* ■

Важливою властивістю ортогонального перетворення є те, що воно не змінює евклідову норму перетворюваного вектора. Справді, якщо Q – ортогональна матриця, то

$$\| Qx \|_2^2 = (Qx)^T (Qx) = x^T Q^T Qx = x^T x = \| x \|_2^2 .$$

Звідси випливає, що зведення матриці до трикутного вигляду послідовністю ортогональних перетворень вектор-стовпців, не буде „підсилювати” похибки задання елементів системи та обчислювальні похибки. Тобто такі алгоритми стійкіші порівняно з тими, що ґрунтуються на LU – розкладі, хоч і вимагають більшої кількості операцій.

У цьому методі матриця A системи (2.1) зводиться до вигляду

$$Rx = Q^T b, \qquad (2.23)$$

де R – верхня трикутна, а Q – ортогональна матриця.

Одним із можливих способів побудови QR – розкладу матриці A є використання перетворення Хаусхолдера, яке дозволяє перетворити в нуль групу піддіагональних елементів стовпця матриці. Перетворення здійснюється з використанням матриці Хаусхолдера[2]

$$H = I - \frac{2}{\|w\|_2^2} ww^T \qquad (2.24)$$

де w – ненульовий вектор-стовпець, ww^T – квадратна матриця порядку n. Матриця H – симетрична, оскільки

$$(ww^T)^T = (w^T)^T w^T = ww^T,$$

а також ортогональна.

[2] Householder A.S. Unitary Triangularization of a Nonsymmetric Matrix // Journal ACM. – 1958. – **5** (4). – P. 339–342.

Нехай для вектора $y=(y_1,\ldots,y_n)^T$, $y_1\neq 0$, перетворенням $\bar{y}=Hy$ потрібно одержати вектор $\bar{y}=(\tilde{y}_1,0,\ldots,0)^T$. Результат досягається, якщо вибрати

$$w=y+sign(y_1)\|y\|_2\, e_1,\ e_1=(1,0,\ldots,0)^T.$$

Застосуємо таке перетворення для факторизації матриці A до вигляду (2.22). Нехай $A_0=A$, $A_1=H_1A_0$, де матриця H_1 в (2.24) така, що всі елементи першого стовпця матриці A_1, крім першого, дорівнюють нулю. Тоді $y=(a_{11}^{(0)},a_{21}^{(0)},\ldots,a_{n1}^{(0)})$, а компоненти вектора w обчислюються так:

$$\begin{aligned} w_1^{(1)} &= a_{11}^{(0)}+sign(a_{11}^{(0)})\|a_{11}^{(0)}\|,\\ w_i^{(1)} &= a_{i1}^{(0)},\quad i=\overline{2,n}, \end{aligned} \tag{2.25}$$

де $a_{ij}^{(0)}:=a_{ij}$, $a_1^{(0)}:=(a_{11},a_{21},\ldots,a_{n1})^T$.

На другому кроці матрицю H_2 потрібно задати так, щоб отримати нулі в другому стовпці нижче головної діагоналі. Нехай

$$y:=a_2^{(1)},\quad a_2^{(1)}=(0,a_{22}^{(1)},\ldots,a_{2n}^{(1)})^T.$$

Тоді

$$w_1^{(2)}=0,\quad w_2^{(2)}=a_{22}^{(1)}+sign(a_{22}^{(1)})\|a_2^{(1)}\|_2,\quad w_i^{(2)}=a_{2i}^{(1)},\ i=\overline{3,n}.$$

Якщо перетворення вдається виконати $n-1$ раз, то одержимо шуканий розклад (2.22). Розв'язок знаходимо із СЛАР (2.23) із трикутною матрицею $R=A_{n-1}$, матриця

$$Q=(H_{n-1}H_{n-2}\ldots H_1)^T=H_1H_2\ldots H_{n-1}.$$

Зауважимо, що перетворення матриці A в $QR-$методі відбувається аналогічно як у методі Гауса, але з допомогою ортогонального перетворення. На відміну від методу Якобі, на одній ітерації якого утворюється два елементи, що дорівнюють нулю, але в наступних ітераціях вони можуть стати ненульовими, в методі Хаусхолдера на кожній ітерації утворюється кілька нулів, які не змінюють наступні ітерації.

Модифікації QR-розкладу матриці наведені в [80, розділ 4.2.1; 20, гл. 5].

2.10. Метод ортогоналізації

Цим методом можна розв'язувати як СЛАР загального вигляду, так і значно ефективніше системи з симетричними додатно визначеними матрицями. Розв'язок системи (2.1) будується у вигляді

$$x = \sum_{k=1}^{n} C_k x^{(k)}, \tag{2.22}$$

де $C_1,...,C_n$ – невідомі коефіцієнти, $x^{(1)},...,x^{(n)}$ – n-вимірні вектори, що задовольняють умови

$$\left(Ax^{(k)}, x^{(l)}\right) = 0, \text{ коли } k > l. \tag{2.23}$$

Нехай ненульові вектори $x^{(1)}, x^{(2)},..., x^{(n)}$ вже знайдені. Покажемо, як обчислити коефіцієнти C_k. Із (2.1) і (2.22) одержимо

$$\left(Ax, x^{(l)}\right) = \sum_{k=1}^{n} C_k \left(Ax^{(k)}, x^{(l)}\right) = \left(b, x^{(l)}\right), \quad l = \overline{1,n}.$$

Врахувавши умови (2.23), коли $l = 1,...,n$, одержимо для знаходження C_k систему лінійних рівнянь із трикутною матрицею

$$\begin{aligned}
&\left(Ax^{(1)}, x^{(1)}\right)C_1 = \left(b, x^{(1)}\right),\\
&\left(Ax^{(1)}, x^{(2)}\right)C_1 + \left(Ax^{(2)}, x^{(2)}\right)C_2 = \left(b, x^{(2)}\right),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\left(Ax^{(1)}, x^{(n)}\right)C_1 + \left(Ax^{(2)}, x^{(n)}\right)C_2 + \ldots + \left(Ax^{(n)}, x^{(n)}\right)C_n = \left(b, x^{(n)}\right).
\end{aligned} \tag{2.24}$$

Якщо $\left(Ax^{(k)}, x^{(k)}\right) \neq 0,\ k = 1,...,n$, то система має єдиний розв'язок.

Нехай A – симетрична матриця. Тоді

$$\left(Ax^{(k)}, x^{(l)}\right) = \left(x^{(k)}, Ax^{(l)}\right) = \left(Ax^{(l)}, x^{(k)}\right).$$

Тому з умов (2.23) випливає, що $\left(Ax^{(k)}, x^{(l)}\right) = 0, l \neq k, k, l = \overline{1, n},$ і система (2.24) набуде вигляду

$$C_k \left(Ax^{(k)}, x^{(k)}\right) = \left(b, x^{(k)}\right), k = \overline{1, n}.$$

Звідси одержимо

$$C_k = \frac{\left(b, x^{(k)}\right)}{\left(Ax^{(k)}, x^{(k)}\right)}, k = \overline{1, n}.$$

Для побудови системи векторів $x^{(1)},..., x^{(n)}$ візьмемо деяку систему лінійно незалежних векторів, наприклад,

$$e^{(1)} = (1,0,...,0),\ e^{(2)} = (0,1,...,0),...,\ e^{(n)} = (0,0,...,1).$$

Розглянемо спочатку випадок, коли A – додатно визначена матриця $(A > 0)$, тобто $(Ax, x) \geq 0 \quad \forall x \in R^n; \quad (Ax, x) = 0 \Leftrightarrow x = 0$. Нехай $x^{(1)} = e^{(1)}$, а $x^{(2)} = e^{(2)} + \alpha_1^{(2)} x^{(1)}$, $\alpha_1^{(2)}$ – деякий коефіцієнт. Із умови $(x^{(2)}, Ax^{(1)}) = 0$ маємо

$$\left(Ax^{(1)}, e^{(2)}\right) + \alpha_1^{(2)} \left(Ax^{(1)}, x^{(1)}\right) = 0,$$

звідки знаходимо

$$\alpha_1^{(2)} = -\frac{\left(Ax^{(1)}, e^{(2)}\right)}{\left(Ax^{(1)}, x^{(1)}\right)}.$$

Зауважимо, що $x^{(2)} \neq 0$, оскільки $e^{(2)}$ і $e^{(1)}$ - ортогональні.

Далі, шукаємо $x^{(3)}$ у вигляді

$$x^{(3)} = e^{(3)} + \alpha_1^{(3)} x^{(1)} + \alpha_2^{(3)} x^{(2)}.$$

Оскільки $(x^{(3)}, Ax^{(1)}) = 0$ і $(x^{(3)}, Ax^{(2)}) = 0$, то

$$\alpha_1^{(3)} = -\frac{\left(Ax^{(1)}, e^{(3)}\right)}{\left(Ax^{(1)}, x^{(1)}\right)}, \quad \alpha_2^{(3)} = -\frac{\left(Ax^{(2)}, e^{(3)}\right)}{\left(Ax^{(2)}, x^{(2)}\right)}.$$

У загальному випадку маємо

$$x^{(k)} = e^{(k)} + \alpha_1^{(k)} x^{(1)} + ... + \alpha_{k-1}^{(k)} x^{(k-1)},$$

$$\alpha_\nu^{(k)} = -\frac{\left(Ax^{(\nu)}, e^{(k)}\right)}{\left(Ax^{(\nu)}, x^{(\nu)}\right)},\ \nu = \overline{1, k-1}; \qquad k = \overline{1, n}.$$

Зауважимо, що $(Ax^{(\nu)}, x^{(\nu)}) \neq 0$, оскільки $A > 0$ і $x^{(\nu)} \neq 0$.

Нехай тепер A – матриця загального вигляду порядку n, $x^{(1)} = e^{(1)}$ і вектори $x^{(1)},...,x^{(k)}$ вже побудовані. Шукаємо $x^{(k+1)}$ у вигляді

$$x^{(k+1)} = e^{(k+1)} + \sum_{i=1}^{k} \gamma_i^{(k+1)} x^{(i)}.$$

Для знаходження коефіцієнтів $\gamma_i^{(k+1)}$, $i = \overline{1, n}$, одержимо систему рівнянь

$$\sum_{i=1}^{k} \gamma_i^{(k+1)} \left(Ax^{(i)}, x^{(j)}\right) = -\left(Ae^{(k+1)}, x^{(j)}\right),\ j = \overline{1, k}. \qquad (2.25)$$

На підставі умови (2.23) матриця системи (2.25) трикутна і набуває вигляду $(Ax^{(1)}, x^{(1)})\gamma_1^{(k+1)} = -(Ae^{(k+1)}, x^{(1)})$,

$$(Ax^{(1)}, x^{(2)})\gamma_1^{(k+1)} + (Ax^{(2)}, x^{(2)})\gamma_2^{(k+1)} = -(Ae^{(k+1)}, x^{(2)}),$$

$$(Ax^{(1)}, x^{(k)})\gamma_1^{(k+1)} + (Ax^{(2)}, x^{(k)})\gamma_2^{(k+1)} + ... + (Ax^{(k)}, x^{(k)})\gamma_k^{(k+1)} =$$
$$= -(Ae^{(k+1)}, x^{(k)}),\ \text{k=}\overline{1, n-1}\,.$$

2.11. Метод прогонки для систем рівнянь із тридіагональними матрицями

2.11.1. Постановка задачі. Розглянемо СЛАР вигляду

$$\begin{aligned} &-c_0 y_0 + b_0 y_1 = -f_0, \\ &a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \quad i = \overline{1, N-1}; \\ &a_N y_{N-1} - c_N y_N = -f_N. \end{aligned} \tag{2.28}$$

Матриця системи тридіагональна, порядку $N+1$ і набуває вигляду

$$A = \begin{bmatrix} -c_0 & b_0 & 0 & 0 & ... & 0 & 0 & 0 \\ a_1 & -c_1 & b_1 & 0 & ... & 0 & 0 & 0 \\ 0 & a_2 & -c_2 & b_2 & ... & 0 & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & a_{N-1} & -c_{N+1} & b_{N-1} \\ 0 & 0 & 0 & 0 & ... & 0 & a_N & -c_N \end{bmatrix}.$$

Матриці, для яких елементи головної і кількох побічних діагоналей не дорівнюють нулю, називаються *стрічковими***.**

Система (2.28) одержується, зокрема при числовому розв'язуванні крайової задачі

$$u''(x) - q(x)u(x) = -f(x), \quad 0 < x < l;$$
$$u(0) = \mu_0,\ u(l) = \mu_1,$$

якою моделюється стаціонарний розподіл тепла у тонкому стержні довжиною l. На відрізку $[0, l]$ побудуємо рівномірну сітку $x_i = ih,\ i = \overline{0, N}$, з кроком $h = l/N$. У кожному вузлі x_i, $i = \overline{1, N}$, апроксимуємо похідну $u_i'' = u''(x_i)$ різницевою похідною $u_i'' \approx (u_{i-1} - 2u_i + u_{i+1})/h^2$. Одержимо різницеву задачу

$$(y_{i-1} - 2y_i + y_{i+1})/h^2 - q_i y_i = -f_i,\ i = \overline{1, N-1};\ y_0 = \mu_0,\ y_N = \mu_1,$$

або, після простих перетворень,

$$\begin{aligned} &y_{i-1} - (2 + h^2 q_i) y_i + y_{i+1} = -h^2 f_i, \quad i = \overline{1, N-1}, \\ &y_0 = \mu_0,\ y_N = \mu_1. \end{aligned}$$

Отже, маємо СЛАР із тридіагональною матрицею, де

$$b_0 = a_N = 1, \quad a_i = b_i = 1, \quad i = \overline{1, N-1}; \quad c_i = 2 + h^2 q_i, \quad i = \overline{0, N}.$$

Якщо на кінцях стержня задано крайові умови

$$u'(0) = -\nu_0, -u'(l) = -\nu_1,$$

то, апроксимувавши похідні

$$u'(0) \approx (u_1 - u_0)/h, \quad u'(l) = (u_N - u_{N-1})/h,$$

доповнимо систему (2.53) ще двома рівняннями

$$-y_0 + y_1 = -h\nu_0, \quad y_{N-1} - y_N = -h\nu_1.$$

Ефективним методом розв'язування системи (2.28) служить модифікація методу Ґауса, яка має назву *метод про*гонки. Кількість арифметичних операцій у методі прогонки пропорційна порядку системи. Крім того, із $(N+1)^2$ елементів матриці A зберігати потрібно тільки $3N+1$. Якщо $N = 1000$, то ненульових елементів не більше 0.29%.

2.11.2. Метод правої прогонки. Нехай $c_0 \neq 0$. Застосуємо схему Ґауса, взявши за ведучий елемент c_0. З першого рівняння знаходимо $y_0 = \dfrac{b_0}{c_0} y_1 + \dfrac{f_0}{c_0}$. Нехай $\alpha_1 = \dfrac{b_0}{c_0}, \quad \beta_1 = \dfrac{f_0}{c_0}$. Тоді

$$y_0 = \alpha_1 y_1 + \beta_1.$$

Підставимо y_0 у друге рівняння системи (2.28). Одержимо $(a_1\alpha_1 - c_1)y_1 + b_1 y_2 = -a_1\beta_1 - f_1$. Якщо $\gamma_1 = c_1 - a_1\alpha_1 \neq 0$, то

$$y_1 = \alpha_2 y_2 + \beta_2, \text{ де } \alpha_2 = \frac{b_1}{\gamma_1}, \quad \beta_2 = \frac{a_1\beta_1 + f_1}{\gamma_1}.$$

Нехай виконано i кроків, $2 \le i \le N-1$. У системі рівнянь

$$y_{i-1} = \alpha_i y_i + \beta_i,$$

$$a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i$$

вилучимо з другого рівняння (2.55) y_{i-1}. Для цього введемо позначення: $\gamma_i = c_i - a_i\alpha_i$ і $\alpha_{i+1} = \dfrac{b_i}{\gamma_i}, \ \beta_{i+1} = \dfrac{f_i + a_i\beta_i}{\gamma_i}$, якщо $\gamma_i \neq 0$. Тоді

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, \ i = \overline{1, N-1}.$$

Залишається знайти y_N із системи рівнянь

$$y_{N-1} = \alpha_N y_N + \beta_N,$$

$$a_N y_{N-1} - c_N y_N = -f_N.$$

Для y_N одержимо

$$y_N = (f_N + a_N\beta_N)/(c_N - a_N\alpha_N).$$

Отже, прямий хід методу прогонки полягає в обчисленні коефіцієнтів

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

$$\gamma_i = c_i - a_i\alpha_i,\ \ \alpha_{i+1} = \frac{b_i}{\gamma_i},\ \ \beta_{i+1} = \frac{f_i + a_i\beta_i}{\gamma_i},\ i = \overline{1, N-1}. \qquad (2.29)$$

В оберненому ході знаходимо розв'язок системи:

$$y_N = (f_N + a_N\beta_N)/(c_N - a_N\alpha_N),$$

$$y_i = \alpha_{i+1}y_{i+1} + \beta_{i+1},\ i = \overline{N-1, 0}. \qquad (2.30)$$

Обчислення за формулами методу прогонки вимагає $8N+1$ арифметичних операцій, із них $5N+1$ ділень і множень.

Зауваження 2.2. *Коефіцієнти α_i не залежать від правої частини системи рівнянь (2.28) і визначаються через значення a_i, b_i, c_i. Тому, якщо потрібно розв'язати кілька СЛАР із різними правими частинами, але з однією і тією ж тридіагональною матрицею, то коефіцієнти α_i обчислюються тільки один раз, а коефіцієнти β_i і значення розв'язку y_i обчислюється для кожної із систем з відповідною правою частиною.*

2.11.3. Обґрунтування методу правої прогонки. Формули (2.29) мають зміст, якщо $c_0 \neq 0$ і $\gamma_i \neq 0,\ \ i = \overline{1, N}$. Значення розв'язку обчислюється за рекурентною формулою (2.30), при цьому можуть нагромаджуватись похибки заокруглення. Справді, припустимо, що α_{i+1} і β_{i+1} знаходяться точно, а в обчисленні y_{i+1} допущена похибка ε_{i+1}. Тоді замість y_i одержимо значення $\tilde{y}_i = \alpha_{i+1}(y_{i+1} + \varepsilon_{i+1}) + \beta_i$. Похибка $\varepsilon_i = \tilde{y}_i - y_i$ задовольняє рівняння

$$\varepsilon_i = \alpha_{i+1}\varepsilon_{i+1},\quad i = \overline{N-1, 0}.$$

Якщо $|\alpha_{i+1}| > 1$, то $|\varepsilon_i| \geq |\varepsilon_{i+1}|$ і для великих i початкова похибка може сильно зрости. Якщо ж $|\alpha_{i+1}| \leq 1$, то $|\varepsilon_i| \leq |\varepsilon_{i+1}|$, тому нагромадження похибки не відбудеться. У цьому випадку кажуть, що метод правої прогонки *стійкий*.

Наступна теорема дає достатні умови коректності методу правої прогонки, тобто, що існує єдиний розв'язок системи (2.28), який обчислюється згідно з формулами (2.29), (2.30) і метод прогонки стійкий до похибок обчислення коефіцієнтів.

Теорема 2.3 [59, 60]. *Нехай коефіцієнти системи (2.28) задовольняють умови:*

$$a_i \neq 0,\ i=\overline{1,N};\ b_i \neq 0,\ i=\overline{0,N-1}; \tag{2.31}$$

$$|c_i| \geq |a_i| + |b_i|, \quad i=\overline{1,N-1}; \tag{2.32}$$

$$|c_0| \geq |b_0|, \quad |c_N| \geq |a_N|, \tag{2.33}$$

причому хоч б одна з нерівностей (2.32) або (2.30) виконується строго (A −матриця із діагональною перевагою).

Тоді алгоритм методу правої прогонки коректний, тобто $\gamma_i \neq 0$ і $|\alpha_i| \leq 1, \quad i=\overline{1,N}$.

Доведення. Оскільки $|\alpha_1| = |b_0|/|c_0|$, то з першої з умов (2.31) випливає, що $|\alpha_1| \leq 1$. Нехай $|\alpha_i| \leq 1, \quad i \leq N-1$. Тоді

$$|\gamma_i| = |c_i - a_i\alpha_i| \geq |c_i| - |a_i||\alpha_i| \geq |b_i| + |a_i|(1-|\alpha_i|) \geq |b_i| > 0.$$

Отже, $\gamma_i \neq 0$ і $|\alpha_{i+1}| = |b_i|/|\gamma_i| \leq |b_i|/|b_i| = 1, \quad i=\overline{1,N-1}$.

Залишається довести, що $\gamma_N = c_N - a_N\alpha_N \neq 0$. Скористаємось умовою, що виконується строго хоча б одна з нерівностей (2.32) або (2.33). Нехай $|c_N| > |a_N|$. Тоді $|\gamma_N| > |a_N|(1-|\alpha_N|) \geq 0$. Якщо ж для деякого k, $1 \leq k \leq N-1$ маємо $|c_k| > |a_k| + |b_k|$, то $|\gamma_k| = |c_k - a_k\alpha_k| > |b_k| > 0$. Тому $|\alpha_{k+1}| = |b_k|/|\gamma_k| < 1$. Методом математичної індукції просто перевірити, що й $|\alpha_i| < 1$ для $i \geq k+1$. Отже, $|\alpha_N| < 1$ і тому $\gamma_N \neq 0$. Якщо ж $|c_0| > |b_0|$, то $|\alpha_1| < 1$ і тому $|\alpha_i| < 1$ для $i \geq 2$, а також $|\alpha_N| < 1$ і $\gamma_N \neq 0$. ■

Зауваження 2.3. *Умови (2.31)-(2.33) є тільки достатніми. Якщо деякий коефіцієнт* a_k *чи* b_k, $1 < k < N$, *дорівнює нулю, то система (2.28) розпадається на дві системи такого ж вигляду. До кожної з них можна застосувати формули вигляду (2.29), (2.30).*

Зауваження 2.4. *При виконанні умов попередньої теореми система (2.28) має єдиний розв'язок. Через L і U позначимо матриці порядку* $N+1$ *вигляду:*

$$L=\begin{bmatrix} -c_0 & 0 & 0 & \dots & 0 & 0 & 0 \\ a_1 & -\gamma_1 & 0 & \dots & 0 & 0 & 0 \\ 0 & a_2 & -\gamma_2 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & a_{N-1} & -\gamma_{N-1} & 0 \\ 0 & 0 & 0 & \dots & 0 & a_N & -\gamma_N \end{bmatrix},$$

$$U=\begin{bmatrix} 1 & -\alpha_1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -\alpha_2 & \dots & 0 & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 & -\alpha_N \\ 0 & 0 & 0 & \dots & 0 & 0 & 1 \end{bmatrix}.$$

Якщо помножити матрицю L *на* U, *то одержимо розклад* $LU=A$. *Оскільки* $c_0 \neq 0$ *і* $\gamma_i \neq 0, \quad i=\overline{1,N}$, *то* $\det A = \det L \cdot \det U =$

$= (-1)^{N+1} c_0 \prod_{i=1}^{n} \gamma_i \neq 0$. *Отже, розв'язок системи рівнянь (2.8) існує і єдиний.*

2.11.4. Метод лівої та зустрічної прогонок. У методі *лівої прогонки* значення розв'язку знаходяться за формулами

$$y_0=\eta_0, \quad y_{i+1}=\xi_{i+1}y_i+\eta_{i+1}, \quad i=\overline{1,N-1},$$

після обчислення коефіцієнтів

$$\xi_N = a_N / c_N, \quad \delta_I = c_I - b_I\xi_{I+1}, \quad \xi_I = a_I / \delta_I, \quad i=\overline{N-1,1};$$

$$\eta_N = f_N / c_N, \quad \eta_i = (f_i + b_i\eta_{i+1}) / \delta_i, \quad i=\overline{N-1,0}.$$

Якщо потрібно знайти одне невідоме y_m, $1 \le m \le N-1$, або групу невідомих, які розміщені підряд, то зручно використати метод *зустрічних* прогонок [59, 60]. Цей метод одержується комбінацією правої і лівої прогонок. Формули методу зустрічних прогонок набувають вигляду:

$$\alpha_1 = \frac{b_0}{c_0}; \ \gamma_i = c_i - a_i\alpha_i, \ \alpha_{i+1} = \frac{b_i}{\gamma_i}, \ i=\overline{1,m-1};$$

$$\beta_1 = \frac{f_0}{c_0}; \ \beta_{i+1} = \frac{f_i + a_i\beta_i}{\gamma_i}, \ i=\overline{1,m-1};$$

$$\xi_N = \frac{a_N}{c_N}; \quad \delta_i = c_i - b_i\xi_{i+1i}, \quad \xi_1 = \frac{a_i}{\delta_i}, \quad i = \overline{N-1, m},$$

$$y_i = \alpha_{i+1}y_{i+1} + \beta_{i+1}, \quad i = \overline{m-1, 0};$$

$$y_{i+1} = \beta_{i+1}y_i + \eta_{i+1}, \quad i = \overline{m, N-1}; \quad y_m = \frac{\eta_m + \xi_m\beta_m}{1-\xi_m\alpha_m}.$$

2.11.5. Метод циклічної прогонки. Розглянемо таку систему:

$$a_i y_{i-1} + c_i y_i - b_i y_i = f_i,$$

$$a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \qquad i = 0, \pm 1, \pm 2, ...$$

коефіцієнти і права частина якої періодичні з періодом N:

$$a_i = a_{i+N}, \ b_i = b_{i+N}, \ c_i = c_{i+N}, \ f_i = f_{i+N}. \qquad (2.34)$$

Системи такого вигляду одержуються, наприклад, при розгляді триточкових різницевих схем, які виникають при знаходженні періодичних розв'язків звичайних диференціальних рівнянь другого порядку, а також при наближеному розв'язуванні рівнянь з частинними похідними в циліндричних і сферичних координатах.

При виконанні умов (2.34) розв'язок системи рівнянь, якщо він існує, також буде періодичним з періодом N. Тому досить знайти розв'язок y_i наприклад при $i = \overline{0, N-1}$. У цьому випадку задачу можна записати так:

$$a_0 y_{N-1} + c_0 y_0 - b_0 y_1 = f_0,$$

$$a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \qquad i = \overline{1, N-1},$$

$$y_N = y_0.$$

Алгоритм розв'язування цієї задачі, який носить назву *методу циклічної прогонки,* такий:

$$\alpha_2 = b_1 / c_1, \ \beta_2 = f_1 / c_1, \ \gamma_2 = a_1 / c_1; \ A_k^+, k = \overline{1, N}$$

$$\alpha_{i+1} = b_i / \delta_i, \quad \beta_{i+1} = (f_i + a_i\beta_i) / \delta_i, \ \gamma_{i+1} = a_i\gamma_i / \delta_i, \qquad i = \overline{2, N}.$$

$$u_{N-1} = \beta_N, \quad v_{N-1} = \alpha_N + \gamma_N;$$

$$u_i = \alpha_{i+1}u_{i+1} + \beta_{i+1}, \qquad v_i = \alpha_{i+1}v_{i+1} + \gamma_{i+1}, \qquad i = \overline{N-2, 1};$$

$$y_0 = \frac{\beta_{N+1} + \alpha_{N+1}u_1}{1 - \gamma_{N+1} - \alpha_{N+1}v_1}, \quad y_i = u_i + y_0 v_i, \qquad i = \overline{1, N-1}.$$

Для реалізації алгоритму циклічної прогонки потрібно виконати $14N - 8$ арифметичних операцій.

Якщо не виконуються умови переваги головної діагоналі в матриці A, то можуть з'явитися знаменники, що дорівнюють нулю або близькі до нуля. У таких випадках використовується метод немонотонної прогонки. Цей метод ґрунтується на виборі головного елемента по рядках, як у методі вилучення невідомих Ґауса. У такому алгоритмі монотонний порядок визначення невідомих y_i може порушуватись, тому він і має назву немонотонної прогонки.

Метод прогонки модифікований для систем з п'ятидіагональними матрицями та інших систем лінійних рівнянь спеціальної структури із розрідженими матрицями. Названі методи наведені в [24, 53, 60, 81].

2.12. Розв'язування лінійних систем із прямокутними матрицями

2.12.1. Псевдорозв'язок СЛАР. Розглянемо СЛАР

$$Ax = b \tag{2.35}$$

де A – $(m \times n)$-матриця, b і x відповідно m і x-вимірні вектор-стовпці, тобто

$$A - \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, \; b = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_m \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}.$$

При $m > n$ система перевизначена. Такі системи зустрічаються, зокрема при обробці даних, обсяг яких перевищує кількість невідомих коефіцієнтів функції, за допомогою якої здійснюється апроксимація.

Означення 2.1. *Псевдорозв'язком системи (2.35) називається n-вектор* $x = x^0$ *такий, що евклідова норма*

$$\left\| Ax - b \right\|_2^2 = \sum_{i=1}^{m} \left(\sum_{k=1}^{n} a_{ik} x_k - b_i \right)^2$$

має найменше значення і серед усіх векторів x, для яких це значення досягається, вектор x^0 *має найменшу довжину, тобто*

$$\left\| x^0 \right\|_2^2 := \sum_{i=}^{n} (x_i^0)^2 = \min_x \sum_{i=1}^{n} x_i^2 .$$

Доведено [25, 80], що система (2.35) завжди має тільки один псевдорозв'язок, який обчислюється згідно з формулою

$$x^0 = A^+ b, \qquad (2.36)$$

де A^+ – $(n \times m)$-матриця, яка називається *псевдооберненою* до матриці A.

Зауважимо, що для системи (2.1) із квадратною невиродженою матрицею псевдорозв'язок x^0 збігається із єдиним розв'язком системи (2.35). У цьому випадку $x = A^{-1}b$ і $A^+ = A^{-1}$.

Покажемо, що $\|Ax - b\|_2^2$ досягає найменшого значення тоді, коли матриця A^T ортогональна до нев'язки $r(x) = b - Ax$ для розв'язку $x = x^0$, тобто для $x = x^0$

$$A^T(b - Ax^0) = A^T r(x^0) = 0.$$

Нехай x – деякий розв'язок системи (2.35), $x \neq x^0$. Тоді

$$r(x) = (b - Ax^0) + (Ax^0 - Ax) = r(x^0) + A(x^0 - x).$$

На підставі умови ортогональності одержимо

$$r^T(x)r(x) = ((A(x - x^0))^T + r^T(x^0))(A(x - x^0) + r(x^0)) =$$

$$= (A(x - x^0))^T (A(x - x^0)) + r^T(x^0)r(x^0) = \|A(x - x^0)\|_2^2 + \|r(x^0)\|_2^2 > \|r(x^0)\|_2^2.$$

Отже, $\|r(x)\|_2^2 > \|r(x^0)\|_2^2$ для будь-якого розв'язку $x \neq x^0$.

Із умови ортогональності при $n \leq m$ одержимо так звану *нормальну систему рівнянь*

$$(A^T A)x = A^T b. \qquad (2.37)$$

Матриця $A^T A$ – симетрична, її порядок n і ранг дорівнює n, якщо ранг матриці $rang(A) = n$. Тобто, якщо стовпці матриці A лінійно незалежні, то існує $(A^T A)^{-1}$ і псевдорозв'язок

$$x = (A^T A)^{-1} A^T b. \qquad (2.38)$$

Звідси випливає, що псевдообернена матриця

$$A^+ = (A^T A)^{-1} A^T. \qquad (2.39)$$

2.12.2. Обчислення псевдооберненої матриці

Означення 2.2. *Псевдообернена матриця A^+ – це нетотожне перетворення матриці A, яке задовольняє такі умови:*

1. $AA^+A = A$; *2.* $A^+AA^+ = A^+$; *3.* $(A^+A)^T = A^+A$;

4. $(AA^+)^T = AA^+$.

Якщо рядки матриці A лінійно незалежні, то матриця AA^T має обернену. В цьому разі псевдообернена матриця обчислюється за формулою

$$A^+ = A^T (AA^T)^{-1}. \qquad (2.40)$$

Якщо стовпці матриці A лінійно незалежні, то існує обернена матриця до матриці $A^T A$ і в цьому випадку

$$A^+ = (A^T A)^{-1} A^T. \qquad (2.41)$$

Якщо і стовпці і рядки матриці A лінійно незалежні (що справджується для квадратних невироджених матриць), тоді

$$A^+ = A^{-1}.$$

Нехай k – ранг матриці A. Тоді A можна зобразити як добуток BC, де B – матриця розміру $m \times k$, а C – $k \times n$. У цьому випадку

$$A^+ = C^T (CC^T)^{-1} (B^T B)^{-1} B^T.$$

Наведемо приклад застосування формули (2.41).

Приклад 2.3. Розглянемо систему рівнянь [32]

$$\begin{aligned} x_1 - x_2 + 2x_3 &= 3, \\ -x_1 + 2x_2 - 3x_3 + x_4 &= 0, \\ x_2 - x_3 + x_4 &= 0, \end{aligned}$$

із матрицею

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

Оскільки третій рядок матриці A є сумою перших двох рядків і $\begin{vmatrix} 1 & -1 \\ -1 & 2 \end{vmatrix} = 3 \neq 0,$ то ранг матриці 2. За B візьмемо перші два стовпці матриці A. Тоді

$$A = BC = \begin{bmatrix} 1 & -1 \\ -1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 1 \end{bmatrix},$$

$$B^T B = \begin{bmatrix} 2 & -3 \\ -3 & 6 \end{bmatrix}, \quad (B^T B)^{-1} = \begin{bmatrix} 2 & 1 \\ 1 & 2/3 \end{bmatrix},$$

$$CC^T = \begin{bmatrix} 3 & 0 \\ 0 & 63 \end{bmatrix}, \quad (CC^T)^{-1} = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \end{bmatrix} = \frac{1}{3} I.$$

Тому, згідно з формулою (2.41) , псевдообернена матриця

$$A^{+}=\frac{1}{2}\begin{bmatrix}1 & 0\\ 0 & 1\\ 1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}2 & 1\\ 1 & 2/3\end{bmatrix}\cdot\begin{bmatrix}1 & -1 & 0\\ -1 & 2 & 1\end{bmatrix}=\begin{bmatrix}1/3 & 0 & 1/3\\ 1/9 & 1/9 & 2/9\\ 2/9 & -1/9 & 1/9\\ 4/9 & 1/9 & 5/9\end{bmatrix}.$$

Псевдорозв'язок системи рівнянь

$$x^{0}=A^{+}\begin{bmatrix}3\\ 6\\ 0\end{bmatrix}=\begin{bmatrix}1\\ 1\\ 2\end{bmatrix}.$$

Отже, для квадратної невиродженої матриці ($k=m=n$) псевдообернена матриця $A^{+}=A^{-1}$. Якщо стовпці матриці A лінійно незалежні, то A^{+} обчислюється згідно з (2.39), якщо рядки матриці A лінійно незалежні, то маємо формулу (2.40). Інакше, коли $rang(A)=k<\min(m,n)$, то можна застосувати формулу (2.41).

2.12.3. Методи Келлі–Гамільтона і Гревілля. У формулах (2.39)–(2.41) потрібно знаходити обернену матрицю. Оригінальні формули для обчислення псевдообернених матриць одержали А.Ф. Турбін[3] та В.І. Кублановська[4] Розроблено ряд алгоритмів, у яких обернення матриці не використовується [19], наведемо два з них.

Алгоритм методу Келлі–Гамільтона

1. $A_1 := A^T A$;
2. $k := 1$
3. *do* 3.1. $\gamma_k := tr(A_k / k)$ (обчислення сліду матриці A_k / k)
   3.2. *if* $\gamma_k = 0$ *then* $r := k$
   3.3. $B_k = A_k - \gamma_k I$
   3.4. $A_{k+1} = B_k A_k$
   3.5. *if* $A_k B_k = 0$ *then* $k := n$

   until $k < n$ *end do*
4. $A^{+} := B_r A^T / \gamma_{r-1}$ (псевдообернена матриця)

[3] Турбин А.Ф. Формулы для вычисления полуобратной и псевдообратной матриц / ЖВМиМФ. – 1974. – Т. 14, № 3. – с. 772–776.

[4] Кублановская В.И. О вычислении обобщенной обратной матрицы и проектора / ЖВМиМФ. – 1966. – Т. 6, № 2. – с. 326–332.

Метод Гревілля послідовного знаходження псевдооберненої матриці. Нехай a_k – k-й стовпець у $m\times n$ – матриці A, $A_k=(a_1,...,a_k)$ - матриця утворена першими k стовпцями матриці A, b_k останній рядок у матриці $A_k^+, k=\overline{1,n}$, $A_1=a_1, A_n=A$. Якщо $A_1=a_1=0$ то $A_1^+=0$. Тоді

$$A_1^+=a_1^+=\frac{a_1^+}{a_1^+a_1}$$

і для $k>1$ виконуються рекурентні формули

$$A_k^+=\begin{pmatrix}B_k\\ b_k\end{pmatrix},\quad B_k=A_{k-1}^+-d_kb_k,\quad d_k=A_{k-1}^+a_k.$$

Якщо $c_k=a_k-A_{k-1}d_k\neq 0$, то $b_k=c_k^+=(a_k-A_{k-1}d_k)^+$;
якщо ж $c_k=0$, тобто $a_k=A_{k-1}d_k$, то $b_k=(1+d_k^+d_k)^{-1}d_k^+A_{k-1}^+$.

Даний метод може бути використаний і для знаходження оберненої матриці A^{-1}.

Приклади розв'язування типових задач

Задача 1. Застосувавши LU -розклад матриці, розв'язати СЛАР

$$\begin{aligned}4x_1+\ x_2+3x_3&=8,\\ 3x_1+5x_2-6x_3&=2,\\ 5x_2-\ x_3&=4.\end{aligned}$$

Розв'язування. Елементи матриць L і U в розкладі $A=LU$ знаходимо з матричного рівняння

$$\begin{bmatrix}l_{11}&0&0\\ l_{21}&l_{22}&0\\ l_{31}&l_{32}&l_{33}\end{bmatrix}\begin{bmatrix}1&s_{12}&s_{13}\\ 0&1&s_{23}\\ 0&0&1\end{bmatrix}=\begin{bmatrix}4&1&3\\ 3&5&-1\\ 0&5&-1\end{bmatrix}.$$

Послідовно визначаємо:

$l_{11}=4,\quad s_{12}=1/l_{11}=0.25,\quad s_{13}=3/l_{11}=0.75;$

$l_{21}=3,\quad l_{22}=5-l_{21}s_{12}=4.25,\quad s_{13}=(-6-l_{21}s_{13})/l_{22}=-33/17;$

$l_{31}=0,\quad l_{32}=5,\quad l_{33}=-1-l_{31}s_{13}-l_{32}s_{23}=148/17.$

Із СЛАР з трикутною матрицею

$$\begin{aligned}4y_1&=8,\\ 3y_1+17y_2/4&=2,\\ 5y_2+148y_3/17&=4\end{aligned}$$

знаходимо $y_1 = 2, \quad y_2 - -16/17, \quad y_3 = 1.$

Розв'яжемо СЛАР із верхньою трикутною матрицею

$$x_1 + x_2/4 + 3x_3/4 = 2,$$
$$x_2 - 33x_3/17 = -16/17,$$
$$x_3 = 1,$$

звідки знаходимо $x_3 = x_2 = x_1 = 1.$

Задача 2. Розв'язати систему лінійних рівнянь із задачі 1, застосувавши QR-метод.

Розв'язування. Оскільки $a_1^{(0)} = [4,3,0]^T$, $\left\|a_1^{(0)}\right\|_2 = 5,$ то

$$w_1 = [4+5,3,0]^T = [9,3,0]^T, \|w_1\|_2^2 = 90.$$

Ортогональна матриця

$$H_1 = I - \frac{2}{\|w_1\|_2^2} w_1 w_1^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \frac{2}{90}\begin{bmatrix} 9 \\ 3 \\ 0 \end{bmatrix}\begin{bmatrix} 9 & 3 & 0 \end{bmatrix} =$$

$$= \begin{bmatrix} -0.8 & -0.6 & 0 \\ -0.6 & 0.8 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Далі маємо,

$$A_1 = H_1 A_0 = \begin{bmatrix} -0.8 & -0.6 & 0 \\ -0.6 & 0.8 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 4 & 1 & 3 \\ 3 & 5 & -6 \\ 0 & 5 & -1 \end{bmatrix} = \begin{bmatrix} -5 & = 3.8 & 1.2 \\ 0 & 3.4 & -6.6 \\ 0 & 5 & -1 \end{bmatrix}.$$

Утворимо нуль на місці елемента $a_{32}^{(1)}$. Згідно з QR-алгоритмом

$$a_1^{(0)} = [0, 3.4, 5], \left\|a_1^{(1)}\right\|_2 \approx 6.046487,$$

$$w_2 = \begin{bmatrix} 0 \\ 3.4 + 6.046487 \\ 5 \end{bmatrix} = \begin{bmatrix} 0 \\ 9.446487 \\ 5 \end{bmatrix}, \|w_2\|_2^2 = 114.236109.$$

Наступна ортогональна матриця

$$H_2 = I - \frac{2}{\|w_2\|_2^2}\begin{bmatrix} 0 \\ 9.446487 \\ 5 \end{bmatrix}[0, 9.446487, 5] =$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.562346 & -0.826945 \\ 0 & 0.826945 & 0.562300 \end{bmatrix}.$$

Верхня трикутна матриця з похибкою 0.000113 має вигляд

$$R = A_2 = H_2 A_1 = \begin{bmatrix} -5 & -3.8 & 1.2 \\ 0 & -6.046701 & 4.538429 \\ 0 & 0.000113 & 4.895537 \end{bmatrix}.$$

Права частина СЛАР із матрицею R набуває вигляду

$$Q^T b = (H_1 H_2)^T b =$$

$$= \begin{bmatrix} -0.8 & -0.6 & 0 \\ 0.377408 & -0.449877 & -0.826945 \\ 0.496167 & -0.661556 & 0.562300 \end{bmatrix} \begin{bmatrix} 8 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} -7.600000 \\ 1.508273 \\ 4.895424 \end{bmatrix}.$$

Із системи рівнянь

$$-5.000000x_1 - 3.800000x_2 + 1.200000x_3 = -7.600000,$$
$$-6.046701x_2 + 4.538429x_3 = -1.508273,$$
$$4.895537x_3 = 4.895424$$

знаходимо $x_1 = 1.004367, \quad x_2 = 0.994246, \quad x_3 = 0.9999769.$

Похибка одержаного розв'язку $\max\limits_{i=1,n} \left| x_i - x_i^* \right| = 0.005754.$

Далі, скориставшись формулами (2.23), знайдемо розв'язок системи $A_2 x = Q^T b$, яка набуває вигляду

$$5.000x_1 + 3.800x_2 - 1.200x_3 = 7.600,$$
$$6.046x_2 - 4.538x_3 = 1.512,$$
$$4.895x_3 = 4.892.$$

Отже, наближений розв'язок системи має вигляд:

$x_1 \approx 0.99939, \quad x_2 \approx 1.00017, \quad x_3 \approx 0.99972.$

Задача 3. Методом квадратного кореня розв'язати систему ліній них рівнянь

$$5x_1 + 2x_2 + x_3 = 8,$$
$$2x_1 + 6x_2 + 3x_3 = 11,$$
$$x_1 + 3x_2 + 7x_3 = 11.$$

Розв'язування. Побудуємо розклад матриці A вигляду $A = S^T DS =$

$$= \begin{bmatrix} s_{11} & 0 & 0 \\ s_{12} & s_{22} & 0 \\ s_{13} & s_{23} & s_{33} \end{bmatrix} \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ 0 & s_{22} & s_{23} \\ 0 & 0 & s_{33} \end{bmatrix} =$$

$$= \begin{bmatrix} s_{11}^2 d_1 & s_{11}s_{12}d_1 & s_{11}s_{13}d_1 \\ * & s_{12}^2 d_1 + s_{22}^2 d_2 & s_{12}s_{13}d_1 + s_{22}s_{23}d_2 \\ * & * & s_{13}^2 d_1 + s_{23}^2 d_2 + s_{33}^2 d_3 \end{bmatrix},$$

де $s_{ii} > 0, \quad |d_i| = 1,$ символом «*» позначені елементи матриці, симетричні відносно головної діагоналі. Тоді, маємо

$$s_{11}^2 d_1 = 5, d_1 = sign(5) = 1, s_{11} = \sqrt{5}, s_{12} = 2/\sqrt{5}, s_{13} = 1/\sqrt{5}.$$

Для елементів другого рядка

$$\frac{4}{5} \cdot 1 + s_{22}^2 d_2 = 6 \Rightarrow d_2 = 1, \quad s_{22} = \sqrt{26}/\sqrt{5}.$$

Із рівняння $\frac{2}{\sqrt{5}} \cdot \frac{1}{\sqrt{5}} + \frac{\sqrt{26}}{\sqrt{5}} s_{23} = 3$ знаходимо $s_{23} = \sqrt{13}/\sqrt{10}.$

Залишається знайти d_3 і s_{33} із рівняння $\frac{1}{5} + \frac{13}{10} + s_{33}^2 d_3 = 7.$ Звідки маємо, що $d_3 = 1, \quad s_{33} = \sqrt{55}/10.$ Отже,

$$S = \begin{bmatrix} \sqrt{5} & 2/\sqrt{5} & 1/\sqrt{5} \\ 0 & \sqrt{26}/\sqrt{5} & \sqrt{13}/\sqrt{10} \\ 0 & 0 & \sqrt{55}/\sqrt{10} \end{bmatrix}$$

і зі СЛАР

$$\begin{aligned} \sqrt{5} y_1 &= 8, \\ \frac{2}{\sqrt{5}} y_1 + \frac{\sqrt{26}}{\sqrt{5}} y_2 &= 11, \\ \frac{1}{\sqrt{5}} y_1 + \frac{\sqrt{13}}{\sqrt{10}} y_2 + \frac{\sqrt{55}}{\sqrt{10}} y_3 &= 11 \end{aligned}$$

знаходимо $y_1 = 8/\sqrt{5}, \quad y_2 = 3\sqrt{13}/\sqrt{10}, \quad y_3 = \sqrt{55}/\sqrt{10}.$

Із СЛАР з верхньою трикутною матрицею

$$\sqrt{5}x_1 + \frac{2}{\sqrt{5}}x_2 + \frac{1}{\sqrt{5}}x_3 = \frac{8}{\sqrt{5}},$$
$$\frac{\sqrt{26}}{\sqrt{5}}x_2 + \frac{\sqrt{13}}{\sqrt{10}}x_3 = \frac{3\sqrt{13}}{\sqrt{10}},$$
$$\frac{\sqrt{55}}{\sqrt{10}}x_3 = \frac{\sqrt{55}}{\sqrt{10}}$$

знаходимо $x_3 = x_2 = x_1 = 1.$

Задача 4. Методом ортогоналізації розв'язати систему рівнянь із симетричною матрицею

$$x_1 + 2x_2 + 3x_3 = 6,$$
$$2x_1 + 5x_2 + 6x_3 = 13,$$
$$3x_1 + 6x_2 + 10x_3 = 19.$$

Розв'язування. Матриця A – симетрична і $A > 0$, оскільки головні мінори $\Delta_1 = \Delta_2 = \Delta_3 = 1 > 0$. Розв'язок системи будуємо у вигляді

$$x = C_1 x^{(1)} + C_2 x^{(2)} + C_3 x^{(3)},$$

де $x^{(1)} = (1,0,0)^T$, $x^{(2)} = e^{(2)} + \alpha_1^{(2)} x^{(1)}$, $x^{(3)} = e^{(3)} + \alpha_1^{(3)} x^{(1)} + \alpha_2^{(3)} x^{(2)} + \alpha_3^{(3)} x^{(3)}$. Знайдемо $x^{(2)}$, обчисливши

$$\alpha_1^{(2)} = -\frac{\left(Ax^{(1)}, e^{(2)}\right)}{\left(Ax^{(1)}, x^{(1)}\right)} = \frac{-2}{1} = -2.$$

Тому $x^{(2)} = e^{(2)} - 2e^{(1)} = (-2,1,0)^T$. На підставі формул (2.47)

$$\alpha_1^{(3)} = -\frac{(1,2,3)(0,0,1)^T}{(1,2,3)(1,0,0)^T} = -3, \quad \alpha_2^{(3)} = -\frac{(0,1,0)(0,0,1)^T}{(0,1,0)(-2,1,0)^T} = 0.$$

Отже, $x^{(3)} = e^{(3)} - 3x^{(1)} = (-3,0,1)^T$. Тепер, згідно з (2.47), маємо

$$C_1 = \frac{\left(b, x^{(1)}\right)}{\left(Ax^{(1)}, x^{(1)}\right)} = \frac{6}{1} = 6, C_2 = \frac{\left(b, x^{(2)}\right)}{\left(Ax^{(2)}, x^{(2)}\right)} = \frac{1}{1} = 1, \quad C_3 = \frac{\left(b, x^{(3)}\right)}{\left(Ax^{(3)}, x^{(3)}\right)} = \frac{1}{1} = 1.$$

Задача 5. Методом ортогоналізації розв'язати систему рівнянь із несиметричною матрицею

$$x_1 + 2x_2 + 3x_3 = 6,$$
$$3x_1 + 2x_2 + 8x_3 = 13,$$
$$2x_1 + 8x_2 + 9x_3 = 19.$$

Розв'язування. Тут матриця A вже не симетрична і не додатна, оскільки $\Delta_2 = -4 < 0$. Покладемо
$x^{(1)} = e^{(1)} = (1,0,0)^T$, $x^{(2)} = e^{(2)} + \gamma_1^{(2)} x^{(1)}$. Тоді

$$\gamma_1^{(2)} = -\frac{(Ae^{(2)}, x^{(1)})}{(Ax^{(1)}, x^{(1)})} = -\frac{(2,2,8)(1,0,0)^T}{(1,3,2)(1,0,0)^T} = -2.$$

Тому $x^{(2)} = (0,1,0)^T - 2(1,0,0)^T = (-2,1,0)^T$.

Коефіцієнти $\gamma_1^{(3)}$ і $\gamma_2^{(3)}$ в розкладі $x^{(3)} = e^{(3)} + \gamma_1^{(3)} x^{(1)} + \gamma_2^{(3)} x^{(2)}$ знаходяться із системи рівнянь

$$\gamma_1^{(3)}(Ax^{(1)}, x^{(1)}) = -(Ae^{(3)}, x^{(1)}),$$
$$\gamma_1^{(3)}(Ax^{(1)}, x^{(2)}) + \gamma_2^{(3)}(Ax^{(2)}, x^{(2)}) = -(Ae^{(3)}, x^{(2)}).$$

Після обчислення скалярних добутків одержимо $\gamma_1^{(3)} = -3$,
$\gamma_2^{(3)} = -1/4$. Отже, $x^{(3)} = e^{(3)} - 3x^{(1)} - \frac{x^{(2)}}{4} = \left(-\frac{5}{2}, -\frac{1}{4}, 1\right)^T$.

Запишемо систему (2.48):

$$C_1 = 6,$$
$$C_1 - 4C_2 = 1,$$
$$-\frac{5}{4}C_1 + 5C_2 + 2C_3 = \frac{3}{4},$$

звідки знаходимо $C_1 = 6, C_2 = \frac{5}{4}, C_3 = 1$, тоді

$x = 6x^{(1)} + \frac{5}{4}x^{(2)} + x^{(3)} = (1,1,1)^T$, тобто $x_1 = x_2 = x_2 = 1$.

Задача 6. Методом Келлі–Гамільтона побудувати псевдообернену матрицю A^+ до матриці

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

Розв'язування. Для для $k = 1$ маємо:

$$A_1 = A^T A = \begin{bmatrix} 1 & 1 & 2 & 3 \\ 1 & 0 & 1 & 1 \\ 1 & -1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \\ 2 & 1 & 0 \\ 3 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 15 & 6 & -3 \\ 6 & 3 & 0 \\ -3 & 0 & 3 \end{bmatrix},$$

$$tr(A_1) = 21,\ B_1 = A_1 - 21I = \begin{bmatrix} -6 & 6 & -3 \\ 6 & -18 & 0 \\ -3 & 6 & -18 \end{bmatrix},$$

$$A_2 = A_1 B_1 = 9 \begin{bmatrix} -5 & -2 & 1 \\ -2 & -2 & -2 \\ 1 & -2 & -5 \end{bmatrix}.$$

Для $k = 2$ одержимо

$$\gamma_2 = tr(A_2 / 2) = -54, \quad B_2 = A_2 - \gamma_2 I = 9 \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix},$$

$$A_3 = A_2 B_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$ Отже, $M = r = 2$ і

$$A^+ = \frac{1}{\gamma_2} B_1 A^T = -\frac{1}{18} \begin{bmatrix} -1 & -1 & -2 & -3 \\ -4 & 2 & -2 & 0 \\ -7 & 5 & -2 & 3 \end{bmatrix}.$$

Розглянемо тепер систему лінійних рівнянь з матрицею A

$$x_1 + x_2 + x_3 = 3,$$
$$x_1 - x_3 = 0,$$
$$2x_1 + x_2 = 3,$$
$$3x_1 + x_2 - x_3 = 4.$$

Система несумісна, оскільки, віднявши від четвертого рівняння третє, одержимо $x_1 - x_3 = 1$, тоді як $x_1 - x_3 = 0$.

Псевдорозв'язок $x^0 = -\frac{1}{18} \begin{bmatrix} -1 & -1 & -2 & -3 \\ -4 & 2 & -2 & 0 \\ -7 & 5 & -2 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 7/6 \\ 1 \\ 5/6 \end{bmatrix},$

нев'язка $\left\|b - Ax^0\right\|_2 = \left\|(0, -\frac{2}{6}, -\frac{2}{6}, \frac{2}{6})^T\right\|_2 = \frac{\sqrt{12}}{6}$.

Задача 7. Методом Ґревіля побудувати псевдообернену матрицю A^+ до матриці $A = \begin{Vmatrix} 1 & -1 & 0 \\ -1 & 2 & 1 \\ 2 & -3 & -1 \\ 0 & 1 & 1 \end{Vmatrix}$.

Розв'язування. Маємо $A_1^+ = (A_1'A_1)^{-1}A_1' = \frac{1}{6}A_1' = \left(\frac{1}{6} \quad -\frac{1}{6} \quad \frac{1}{3} \quad 0\right)$,

$d_2 = A_1^+ a_2 = -\frac{3}{2}, \quad c_2 - a_2 - A_1 d_2 = \left(\frac{1}{2} \quad \frac{1}{2} \quad 0 \quad 1\right)^T$,

$b_2 = c_2^{-1}(c_2'c_2)^{-1}c_2' = \frac{2}{3}c_2' = \left(\frac{1}{3} \quad \frac{1}{3} \quad 0 \quad \frac{2}{3}\right)$,

$B_2 = A_1^+ - d_2 b_2 = \left(\frac{2}{3} \quad \frac{1}{3} \quad \frac{1}{3} \quad 1\right)$.

Отже,

$A_2^+ = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & \frac{1}{3} & 1 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{2}{3} \end{pmatrix}$. Далі, $d_3 = A_2^+ a_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $c_3 = a_3 - A_2 d_3 = 0$.

Тому $b_3 = (1 + d_3'd_3)^{-1}d_3'A_2^+ = \left(\frac{1}{3} \quad \frac{1}{3}\right)$, $A_2^+ = \left(\frac{1}{3} \quad \frac{2}{9} \quad \frac{1}{9} \quad \frac{5}{9}\right)$,

$B_3 = A_3^+ - d_3 b_3 = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & \frac{1}{3} & 1 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{2}{3} \end{pmatrix} - \begin{pmatrix} \frac{1}{3} & \frac{2}{9} & \frac{1}{9} & \frac{5}{9} \\ \frac{1}{3} & \frac{2}{9} & \frac{1}{9} & \frac{5}{9} \end{pmatrix} =$

$= \begin{Vmatrix} \frac{1}{3} & \frac{1}{9} & \frac{2}{9} & \frac{4}{9} \\ 0 & \frac{1}{9} & -\frac{1}{9} & \frac{1}{9} \end{Vmatrix}$. Отже, $A^+ = A_3^+ = \begin{Vmatrix} \frac{1}{3} & \frac{1}{9} & \frac{2}{9} & \frac{4}{9} \\ 0 & \frac{1}{9} & -\frac{1}{9} & \frac{1}{9} \\ \frac{1}{3} & \frac{2}{9} & \frac{1}{9} & \frac{5}{9} \end{Vmatrix}$

Завдання та запитання для самостійної роботи

1. Як побудувати обернену матрицю, використовуючи метод Ґауса?
2. У чому полягає ідея методу LU -розкладу?
3. До яких СЛАР можна застосовувати метод LU – розкладу?
4. Яка норма матриці називається узгодженою з нормою вектора?
5. Як користуватися ітераційною формулою в методі простої ітерації?
6. При виконанні яких умов метод простої ітерації збігається?
7. Чим ітераційна формула методу Зейделя відрізняється від ітераційної формули методу простої ітерації?
8. Чим відрізняється збіжність методу простої ітерації від методу Зейделя?
9. На що впливає вибір початкового наближення $x^{(0)}$ в ітераційних методах?
10. Який критерій зупинки ітераційного процесу в методі простої ітерації та в методі Зейделя?
11. Як СЛАР звести до вигляду зручного для застосування ітерацій? Чи впливає спосіб зведення на швидкість збіжності ітераційного методу?
12. Показати, що матриця Хаусхолдера є ортогональною.
13. Використовуючи метод Ґауса, побудувати обернену матрицю A^{-1} до матриці

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & -5 & 9 \\ 3 & 9 & 24 \end{bmatrix}.$$

14. Методом LU– розкладу розв'язати СЛАР

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 7, \\ 2x_1 - 5x_2 + 9x_3 = 11, \\ 3x_1 + 9x_2 + 24x_3 = 54. \end{cases}$$

Попередньо перевірити умови застосування методу.

15. Побудувати LU-розклад матриць Вандермонда і розв'язати відповідні системи рівнянь.

$$1)\ A_1 = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 4 & 9 \\ 1 & 8 & 27 \end{bmatrix}, \quad b_1 = \begin{bmatrix} 2 \\ 10 \\ 190 \end{bmatrix};$$

$$2)\ A_2=\begin{bmatrix}1&2&3&4\\1&4&9&16\\1&8&27&64\\1&16&81&256\end{bmatrix},\qquad b_1=\begin{bmatrix}2\\10\\44\\190\end{bmatrix}.$$

16. Методом квадратного кореня побудувати розклад матриці Гільберта 3-го і 4-го порядку та розв'язати відповідні системи рівнянь:

$$1)\ A_1=\begin{bmatrix}1&\frac{1}{2}&\frac{1}{3}\\\frac{1}{2}&\frac{1}{3}&\frac{1}{4}\\\frac{1}{3}&\frac{1}{4}&\frac{1}{5}\end{bmatrix},\qquad b_1=\begin{bmatrix}1\\0\\0\end{bmatrix};\ 2)\ A_2=\begin{bmatrix}1&\frac{1}{2}&\frac{1}{3}&\frac{1}{4}\\\frac{1}{2}&\frac{1}{3}&\frac{1}{4}&\frac{1}{5}\\\frac{1}{3}&\frac{1}{4}&\frac{1}{5}&\frac{1}{6}\\\frac{1}{4}&\frac{1}{5}&\frac{1}{6}&\frac{1}{7}\end{bmatrix},\qquad b_2=\begin{bmatrix}1\\0\\0\\1\end{bmatrix}.$$

17. Проілюструвати схему Ґауса і схему з вибором головного елемента у стовпці для СЛАР

$$3.241\cdot10^{-k}x_1+1.000x_2=1.000+3.241\cdot10^{-k},$$
$$1.020\cdot10^{4}x_1+2.000x_2=2.000+1.020\cdot10^{4},\ k=1,4,6,8.$$

Проалізувати одержані розв'язки, порівнявши їх між собою і з точним розв'язком $x_1=x_2=1.$

18. Розв'язати СЛАР

$$\begin{aligned}(R_1+R_3+R_4)I_1+\quad R_3I_2+\quad\ I_3&=E_1,\\ R_3I_1+(R_2+R_3+R_5)I_2-\ I_3&=E_2,\\ R_4I_1-R_5I_2+(R_4+R_5+R_6)I_3&=0.\end{aligned}$$

методом Ґауса, методом Ґауса з вибором головного елемента та ортогоналізації СЛАР для знаходження значень струму I_i в електричному ланцюгу, якщо:

а) $2R_1=R_2=R_3=2,\quad R_4=R_5=3,\quad E_1=24,\quad E_2=28;$

б) $3R_1=4R_2=R_3=6,\quad R_4=2R_5=3,\quad E_1=39,\quad E_2=42.$

19. Знайти точний розв'язок СЛАР порядку 4 з матрицею Гільберта H і правою частиною b:

$$H=\begin{bmatrix}1&1/2&1/3&1/4\\1/2&1/3&1/4&1/5\\1/3&1/4&1/5&1/6\\1/4&1/5&1/6&1/7\end{bmatrix},\quad b=\begin{bmatrix}1\\0\\0\\0\end{bmatrix}.$$

Скласти програму і розв'язати СЛАР методом Ґауса і методом квадратного кореня для даних із простою та подвійною точністю. Порівняти одержані розв'язки.

20. Застосувавши схему Ґауса, знайти матрицю, обернену до матриці A:

$$A=\begin{bmatrix} 1 & 1 & 0 & 4 \\ 2 & -1 & 5 & 0 \\ 5 & 2 & 1 & 2 \\ -3 & 0 & 2 & 6 \end{bmatrix}.$$

21. Розв'язати СЛАР із матрицею з попередньої задачі і вектором правої частини $b=[6,6,9,5]^T$, здійснивши LU –розклад матриці A.

22. Методом правої або лівої прогонки розв'язати СЛАР із тридіагональною матрицею із підрозділу 2.11.1, якою апроксимується крайова задача для ЗДР на сітці з кроком $h=N^{-1}$, якщо $q(x)=x+1, r(x)=\sin x,\ l=1,\quad N=100.$

23. Задана система рівнянь

$$\begin{aligned} 2x_1 + x_2 &= 1, \\ x_1 + 3x_2 + x_3 &= 1, \\ -2x_2 + 3x_3 + x_4 &= 6, \\ x_3 - 2x_4 &= 1. \end{aligned}$$

1) Перевірити виконання умов теореми 2.4. Якщо не всі умови виконуються, то перетворити систему;

2) Знайти розв'язки методом правої та лівої прогонок.

3) Знайти x_3 методом зустрічних прогонок.

24. Методом Келлі–Гамільтона знайти псевдообернену матрицю для СЛАР із прямокутними матрицями та відповідний псевдорозв'язок цих систем рівнянь:

$$1)\ \begin{aligned} 2x_1 - 3x_2 + x_3 - x_4 &= 1, \\ 3x_1 + x_2 + 3x_3 + x_4 &= 2, \\ 4x_1 - x_2 - 4x_3 + 2x_4 &= 3. \end{aligned} \qquad 2)\ \begin{aligned} x_1 - 5x_2 + 4x_3 - 3x_4 &= 2, \\ 2x_1 - 4x_2 + 3x_3 - 2x_4 &= 3, \\ 3x_1 - 3x_2 + 2x_3 - x_4 &= 4. \end{aligned}$$

25. Знайти нормальний псевдорозв'язок СЛАР з прямокутними матрицями:

$$1)\ \begin{aligned} 2x_1 - 3x_2 + x_3 - x_4 &= 1, \\ 3x_1 + x_2 + 3x_3 + x_4 &= 2 \end{aligned} \qquad 2)\ \begin{aligned} x_1 - 5x_2 &= 2, \\ 2x_1 - 4x_2 &= 3, \\ 3x_1 - 3x_2 &= 4. \end{aligned}$$

Розділ 3. Оцінки похибки розв'язку систем лінійних алгебраїчних рівнянь

Норми векторів і матриць. Числа обумовленості матриці, їх порівняння та обчислення. Оцінка відносної похибки розв'язку при збуренні правої частини та матриці системи. Регуляризація СЛАР.

Література [5, 13, 14, 20, 23, 59, 80]

3.1. Норми векторів і матриць

3.1.1. Векторні норми. Нехай дійсному або комплексному n-вектору $x=(x_1,...,x_n)^T$ покладено у відповідність невід'ємне число $\|x\|$ таке, що для довільного числа α і n-вектора y виконується умови:

1) $\|x\|>0$, якщо $x\neq 0$, $\|0\|=0$ (додатна визначеність);

2) $\|\alpha x\|=|\alpha|\|x\|$ (однорідність);

3) $\|x+y\|\le\|x\|+\|y\|$ (нерівність трикутника).

Тоді число $\|x\|$ називається нормою вектора x. Найчастіше в обчислювальній практиці використовуються такі векторні норми:

$\|x\|_1=\sum_{i=1}^{n}|x_i|$ − октаедрична (рис. 3.1);

$\|x\|_\infty=\max\limits_{i=1...n}|x_i|$ − кубічна (рис. 3.2);

$\|x\|_2=|(x,x)|=|x^Tx|=(\sum_{i=1}^{n}x_i^2)^{1/2}$ − евклідова, де через $(\cdot,\cdot)$ позначено скалярний добуток векторів (рис. 3.3).

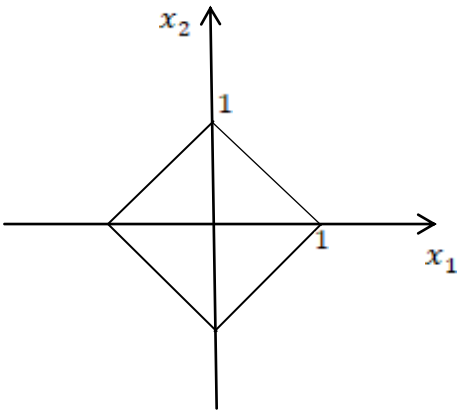


Рис. 3.1. $\|x\|_1=1$

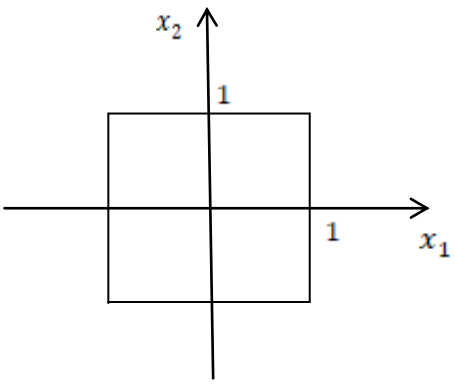


Рис. 3.2. $\|x\|_\infty=1$

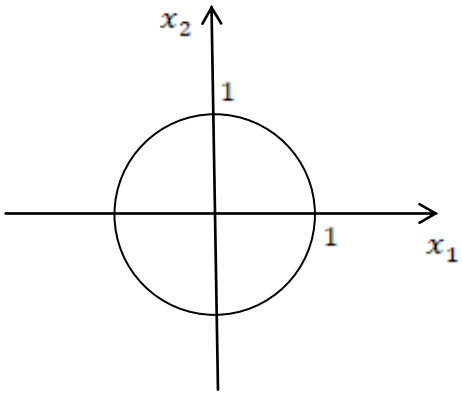


Рис. 3.3. $\|x\|_2 = 1$

Дві векторні норми $\|\cdot\|_I$ і $\|\cdot\|_{II}$ називаються еквівалентними, якщо існують такі додатні числа c_1 і c_2, що для довільного вектора x виконується умова (відношення еквівалентності) $c_1\|x\|_I \le \|x\|_{II} \le c_2\|x\|_I$, причому сталі c_1 і c_2 не залежать від вибору x. Розглянуті вище норми є еквівалентними.

3.1.2. Матричні норми. Для дійсної або комплексної матриці A порядку n визначена норма $\|A\|$, якщо для довільного числа α і довільної матриці B порядку n виконуються такі умови:

а) $\|A\| > 0$, якщо $A \neq 0$, $\|0\| = 0$;

б) $\|\alpha A\| = |\alpha|\|A\|$;

в) $\|A + B\| \le \|A\| + \|B\|$;

г) $\|AB\| \le \|A\|\|B\|$ (умова мультиплікативності).

Якщо норма $\|A\|$ задовольняє тільки перші три умови, то її називають матричною нормою, якщо всі чотири умови, то – мультиплікативною матричною нормою. Надалі під матричною нормою будемо розуміти мультиплікативну матричну норму.

Найчастіше використовуються такі матричні норми:

− октаедрична норма

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|;$$

− кубічна норма $\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$;

− спектральна норма (друга норма або норма Гільберта)

$$\|A\|_2 = \left(\max_{1\le i\le n} \lambda_i(A^T A)\right)^{1/2} = (\max_{1\le i\le n} \lambda_i(AA^T))^{1/2},$$

де $\lambda_i(A^T A)$ – власні значення матриці $A^T A$,

– сферична норма (норма Фробеніуса або евклідова норма)

$$\|A\|_E = (\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2)^{1/2}.$$

Наведені матричні норми є еквівалентними. Зауважимо, що $\lambda(A^T A) \ge 0$. Справді, із рівності $A^T Ax = \lambda x$ випливає, що $x^T A^T Ax = \lambda x^T x$ або $\|Ax\|^2 = \lambda\|x\|^2 \ge 0$, тому $\lambda \ge 0$.

Число $\rho(A) = \max\limits_{1\le i\le n}|\lambda_i|$ називається *спектральним радіусом матриці*.

Матрична норма $\|\cdot\|_M$ називається узгодженою з векторною нормою $\|\cdot\|_V$, якщо

$$\|Ax\|_V \le \|A\|_M \|x\|_V$$

для довільної матриці A порядку n і будь-якого n-вектора x.

Нехай задана деяка векторна норма $\|\cdot\|_V$. Тоді числова функція

$$\|A\| = \sup_{\|x\|_V \ne 0} \frac{\|Ax\|_V}{\|x\|_V} = \sup_{\|x\|_V = 1} \|Ax\|_V$$

називається нормою матриці, підпорядкованою векторній нормі $\|\cdot\|_V$. Серед усіх матричних норм, узгоджених із заданою векторною нормою, підпорядкована норма є мінімальною в тому сенсі, що в нерівності $\|Ax\|_V \le \|A\|_M \|x\|_V$ число $\|A\|_M$ неможливо зменшити. Спектральна, 1-а і ∞-норми матриць підпорядковані по відношенню, відповідно до евклідової, 1- і ∞-норми векторів, отже, й узгоджені з ними.

3.2. Обумовленість матриці

3.2.1. Число обумовленості матриці. Нехай A – невироджена матриця, порядку n.

Означення 3.1. *Величина* $\kappa(A) = \|A\|\|A^{-1}\|$ *називається числом обумовленості матриці* A.

Значення $\kappa(A)$ залежить від вибору матричної норми, однак в силу їх еквівалентності цією різницею можна знехтувати при

оцінках збурення розв'язку. Нехай для вироджених матриць $\kappa(A)=\infty$.

Оскільки будь-яка норма матриці не менша від її найбільшого по модулю власного значення, то $\|A\|\geq\max|\lambda_A|$. Власні значення матриць A і A^{-1} взаємно обернені, тому $\|A^{-1}\|\geq\max\dfrac{1}{|\lambda_A|}=\dfrac{1}{\min|\lambda_A|}$. Отже, $\kappa(A)\geq\max|\lambda_A|/\min|\lambda_A|\geq 1$. Зокрема, для симетричної матриці маємо $\|A\|_2=\max|\lambda_A|$ і $\|A^{-1}\|_2=\max\dfrac{1}{|\lambda_A|}=\dfrac{1}{\min|\lambda_A|}$. Тобто

$$\kappa(A)=\max|\lambda_A|/\min|\lambda_A|$$

для норми $\|\cdot\|_2$. Матриці, для яких число $\kappa(A)$ досить велике, називаються *погано обумовленими*. Межа, з якої матрицю можна класифікувати як погано обумовлену, визначається конкретною задачею і може мати порядок 10^3 або значно більший.

Погано обумовленими є матриці Гільберта, елементи яких

$$a_{ij}=(i+j-1)^{-1},\ i,j=\overline{1,n}.$$

Якщо для матриці Гільберта другого порядку $\kappa(A)=19.28$, то для $n=5$ маємо $4.77\cdot 10^5$, $\kappa(A)=1.60\cdot 10^{13}$ для $n=10$ і $\kappa(A)=1.42\cdot 10^{74}$ для $n=50$

Приклад 3.1. Нехай $A=\begin{bmatrix}1.005 & 1\\ 2 & 2\end{bmatrix}$. Тоді $\|A\|_\infty=4$. Обернена матриця $A^{-1}=\dfrac{1}{0.01}\begin{bmatrix}2 & -1\\ -2 & 1.005\end{bmatrix}=\begin{bmatrix}200 & -100\\ -200 & 100.5\end{bmatrix}$.

Отже, $\kappa(A)=\|A\|_\infty\cdot\|A^{-1}\|_\infty=4\cdot 300.5=1202\gg 1$ і можна вважати матрицю A погано обумовленою.

Зауваження 3.1. *Ознакою поганої обумовленості лінійної системи є поява малих ведучих елементів у прямому ході методу Гауса, хоча є погано обумовлені матриці і з не малими ведучими елементами. Ще одною ознакою може служити поява великого за нормою розв'язку. Нехай* $\|A\|=\|b\|=1$, *тоді* $\|x\|\leq\|A^{-1}\|\cdot\|b\|=\kappa(A)$. *Якщо значення* $\|x\|$ *– велике, то й* $\kappa(A)$*– велике. Але в цьому випадку можуть бути розв'язки з малою нев'язкою* $r=b-Ax$.

Зауваження 3.2. *Для обчислення* $\kappa(A)$ *потрібно знати матрицю* A^{-1}*, знаходження якої за схемою Ґауса вимагає* $\approx n^3$ *арифметичних операцій. Можна оцінити* $\|A^{-1}\|$ *на підставі виразу* [56, с.104]*:*

$$\|A^{-1}\| \approx \frac{3}{2} \max_{1 \le \upsilon \le k} \frac{\|z_\upsilon\|}{\|y_\upsilon\|},$$

де z_υ *– розв'язок задачі* $Az_\upsilon = y_\upsilon$, y_υ *– псевдовипадкові вектори. Досить добре можна оцінити* $\|A^{-1}\|$ *вже при* $k = 3$. *Такий підхід вимагає додатково* $\approx kn^2$ *операцій, що є невеликою частиною затрат при розв'язуванні системи.*

3.2.2. Інші критерії обумовленості. Розглянутий критерій оцінки обумовленості матриці задовільний для обчислень з невисокою точністю і не характеризує втрату значущих цифр. Наприклад для СЛАР з діагональною матрицею, діагональні елементи якої 1.000 і 0.000001 маємо $\kappa(A) = \|A\|_\infty \cdot \|A^{-1}\|_\infty = 10^6$, хоча прямі методи дають розв'язок з максимальною точністю.

Дж. Ортега [49] для обчислень із плаваючою крапкою запропонував критерій обумовленості, виходячи з геометричних міркувань. Нехай $a_i = \left(a_{i_1}, a_{i_2}, ..., a_{i_n}\right)$ – i-ий рядок матриці A. Критерій Ортеги:

$$\kappa_v = V_0 / V,$$

де V – об'єм паралелепіпеда, побудованого на векторах $a_1, ..., a_n$, $V = |\det A|$; $V_0 = \|a_1\| \cdot ... \cdot \|a_n\|$ – об'єм прямокутного паралелепіпеда, $\|a_i\|^2 = a_{i1}^2 + ... + a_{in}^2$.

Для діагональної матриці $\chi_v = 1$. Якщо яке-небудь рівняння СЛАР помножити на сталу $c \neq 0$, то число χ_v не зміниться. Для тридіагональної матриці з елементами $a_{ii} = 2$, $a_{i,i\pm1} = -1$, $i = \overline{1,n}$; $a_{ij} = 0$, $|i - j| > 1$, $n \ge 2$, число обумовленості

$$\kappa_v = 5 \cdot 6^{(n/2)-1} / (n+1).$$

Для $n \approx 1000$ маємо $\chi_v \approx 10^{400}$, що означає дуже погану обумовленість, хоча прямі алгоритми для СЛАР з такими матрицями працюють добре.

Результатом розв'язування СЛАР є точка перетину гіперплощин, які відповідають рівнянням системи. До втрати точності приводить наявність малих кутів між гіперплощинами. Кут між $(n-1)$-вимірною гранню і ребром a_i позначимо через α_i. За кутовий критерій береться число $\kappa_a = 1/\min_i \sin\alpha_i$. Показано[1], що

$$\kappa_a = \max\left(\|a_i\| \cdot \|c_i\|\right),$$

де c_i – i-ий стовпець матриці A^{-1}.

Для тридіагональної матриці з елементами порядку n

$$\kappa_a \approx \left(\frac{n+1}{2}\right)^{3/2}.$$

Число κ_a повільно зростає при збільшенні n, тобто для СЛАР з такими матрицями має спостерігатись добра обумовленість.

Із властивостей числа обумовленості κ_a відзначимо такі:

1) $\kappa_a \le \kappa_v$ при $n > 2$;
2) $\kappa_a = 1$ для діагональної матриці;
3) для довільної квадратної невиродженої матриці $\kappa_a \le \kappa$;
4) для ермітових матриць $\kappa_a(A) = \kappa_a\left(A^{-1}\right)$;
5) при множенні будь-якого рівняння СЛАР на сталу $c \neq 0$ число κ_a не змінюється.

Зауваження 3.3. *Оскільки для обчислень доступна збурена матриця $A + \Delta A$, то замість числа обумовленості матриці A точної системи можна обчислити це число для збуреної матриці. Ця особливість непринциповою. Нехай $\varepsilon = \|\Delta A\|/\|A\|$ і близьке до точності обчислення на комп'ютері. Якщо, крім того, $\varepsilon \le 0.1$, тобто розв'язок одержуєть хоча б з одним правильним десятковим знаком, то справджується нерівність* [31]

$$\frac{8}{9}(1-\varepsilon) \le \frac{\kappa(A+\Delta A)}{\kappa(A)} \le \frac{10}{9}(1+\varepsilon).$$

Це означає, що числа обумовленості точної і збуреної матриць приблизно рівні.

---

[1] *Калиткин Н.Н., Юхно Л.Ф., Кузьмина Л.В. Количественный критерий обусловленности систем линейных алгебраических уравнений. – Матем. моделирование. – 2011. – Т. 23, № 2. – С. 3–26.*

3.3. Оцінка відносної похибки розв'язку при збуренні правої частини

Розглянемо систему лінійних алгебраїчних рівнянь

$$Ax = b \tag{3.1}$$

з квадратною невиродженою матрицею A. Унаслідок обчислювальної похибки або похибок заокруглення, які виникають у процесі введення чисел та при розв'язуванні лінійної системи, отримуємо наближений розв'язок $\tilde{x}$, який можна розглядати як точний розв'язок *збуреної* системи

$$(A + \Delta A)(x + \Delta x) = \tilde{b} \equiv b + \Delta b, \tag{3.2}$$

де матриця збурень ΔA і вектор Δb малі в деякій нормі, $\Delta x = \tilde{x} - x$.

Нехай задано деяку векторну норму. Тоді число $\|x - \tilde{x}\|$ називається *абсолютною похибкою* для $\tilde{x}$. Якщо $x \neq 0$, то $\|x - \tilde{x}\| / \|x\|$ – *відносна* похибка для $\tilde{x}$. Відносна похибка в кубічній нормі може розглядатися як оцінка кількості правильних цифр в числі $\tilde{x}$. Якщо $\dfrac{\|x - \tilde{x}\|_\infty}{\|x\|_\infty} \approx 10^{-p}$, то найбільший за модулем з компонент у $\tilde{x}$ має приблизно p правильних цифр. Як правило, при оцінці відхилення розв'язку $\tilde{x} = x + \Delta x$ системи (3.2) від його точного розв'язку x застосовується обернений аналіз похибок, коли $\tilde{x}$ розглядається як точний розв'язок збуреної системи (3.2).

Нехай у системі (3.1) збурюється тільки вектор b, унаслідок чого розв'язується збурена система

$$A(x + \Delta x) = b + \Delta b. \tag{3.3}$$

Теорема 3.1. *Якщо $\tilde{x}$ – точний розв'язок збуреної системи (3.3), то для відносної похибки розв'язку x правильна оцінка*

$$\frac{\|\Delta x\|}{\|x\|} \le \kappa(A) \frac{\|\Delta b\|}{\|b\|}. \tag{3.4}$$

Доведення. Справді, із систем (3.1) і (3.3) маємо $A\delta x = \delta b$ і $\Delta x = A^{-1}\Delta b$. Тому $\|\Delta x\| = \|A^{-1}\Delta b\|$, отже, $\|\Delta x\| \le \|A^{-1}\| \|\Delta b\|$.

Оскільки $b = Ax$, то $\|b\| \le \|A\| \|x\|$ і $\|b\| \|\Delta x\| \le \|A\| \|x\| \|A^{-1}\| \|\Delta b\|$.

Поділивши обидві частини нерівності на $\|b\|$ і $\|x\|$, одержимо (3.4).

Зауваження 3.4. *Якщо* $\tilde{x}$ *– розв'язок збуреної системи (3.3), а* $r = b - A\tilde{x}$ *– його нев'язка, тоді* $\Delta x = \tilde{x} - x = \tilde{x} - A^{-1}b = -A^{-1}r$. *Звідси одержується оцінка*

$$\|\Delta x\| \le \|A^{-1}\| \cdot \|r\|.$$

Із нерівності (3.4) випливає, що навіть для вектора нев'язки $r = b - A\tilde{x}$ з малою нормою відносні збурення в розв'язку можуть бути великими, якщо число $\kappa(A)$ велике. Число обумовленості може розглядатися як міра *чутливості* розв'язку СЛАР до збурення системи. Із (3.4) одержується нерівність

$$\frac{\|\Delta x\|}{\|x\|} : \frac{\|\Delta b\|}{\|b\|} \le \kappa(A).$$

Отже, в найгіршому випадку відносна похибка розв'язку може перевищувати відносну похибку правої частини в κ разів. Якщо число обумовленості велике, то й похибка $\|\Delta x\| / \|x\|$ може також бути великою.

Розглянемо два приклади.

Приклад 3.2.

$$x_1 + x_2 = 2,$$
$$(1 - 10^{-10})x_1 + x_2 = 2 - 10^{-10}.$$

Точний розв'язок системи: $x_1 = x_2 = 1$. Внесемо у праву частину збурення $\Delta b = \left(0, 10^{-10}\right)^T$. Збурена система

$$y_1 + y_2 = 2,$$
$$(1 - 10^{-10})y_1 + y_2 = 2.$$

вже має розв'язок $y_1=0$, $y_2=2$ і $\Delta x = y - x = \left(-1, 1\right)^T$. Абсолютна похибка $\|\Delta x\|_\infty = 1$ і відносна похибка $\|\Delta x\|_\infty / \|x\|_\infty = 1/1 = 1$ значно перевищують відповідні похибки правої частини $\|\delta b\| = 10^{-10}$ і $\|\Delta b\| / \|b\| = 0.5 \cdot 10^{-10}$. Далі маємо $\|A\|_\infty = \max(2,\ 2 - 10^{-10}) = 2$, $A^{-1} = 10^{-10}\begin{bmatrix} 1 & -1 \\ 10^{-10} - 1 & 1 \end{bmatrix}$, $\|A^{-1}\|_\infty = 2 \cdot 10^{10}$. Отже, число обумовленості $\kappa(A) = 4 \cdot 10^{10} \gg 1$ і матриця A – погано обумовлена. Згідно з (3.4) одержимо оцінку похибки

$$\frac{\|\Delta x\|}{\|x\|} \le 4 \cdot 10^{10} \cdot 0.5 \cdot 10^{-10} = 2,$$

що вдвічі більше від точного значення відносної похибки і перевищує відносну похибку правої частини у $4\cdot 10^{10}$ разів.

Приклад 3.3. Для системи рівнянь

$$x_1 + 2x_2 = 3,$$
$$3x_1 + 4x_2 = 7$$

$A^{-1} = \begin{bmatrix} 2 & 1 \\ 1.5 & -0.5 \end{bmatrix}$ і $\kappa(A) = 7\cdot 3 = 21.$ Отже, матриця A – добре обумовлена. Похибка $\Delta b = \left(0, 10^{-10}\right)^T$ породжує похибку розв'язку $\delta x = y - x = \left(1+10^{-10},\ 1-0.5\cdot 10^{-10}\right) - (1,1) = \left(10^{-10},\ 0.5\cdot 10^{-10}\right)$ з абсолютною похибкою $\|\Delta x\|_\infty = 10^{-10} = \|\Delta b\|_\infty$ і $\|\Delta x\|_\infty / \|x\|_\infty = 1.$ Відносна оцінка похибки складає $21(10^{-10}/7) = 3\cdot 10^{-10}$, що втричі перевищує точне значення похибки і в 21 раз похибку правої частини.

3.4. Оцінка відносної похибки розв'язку при збуренні матриці системи

Якщо в системі (3.1) збурено матрицю A і вектор b, то розв'язується відповідна збурена система

$$(A + \Delta A)(x + \Delta x) = b + \Delta b. \qquad (3.5)$$

Теорема 3.2. *Нехай* $\left\|A^{-1}\Delta A\right\| \le 1$, *тоді*

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

Доведення. Виконання умови теореми забезпечує існування матриці, оберненої до матриці $A + \Delta A$. Із систем (3.1) і (3.3) маємо $A\Delta x = \Delta b - \Delta A\cdot x - \Delta A\cdot \Delta x$. Помноживши обидві частини рівності на A^{-1}, отримаємо $\Delta x = A^{-1}\Delta b - A^{-1}\Delta A x - A^{-1}\Delta A\Delta x$. Звідси маємо

$$\|\delta x\| = \left\|A^{-1}\Delta b - \Delta A A^{-1}x - A^{-1}\Delta A\Delta x\right\| \le$$
$$\le \left\|A^{-1}\right\|\cdot\|\Delta b\| + \left\|A^{-1}\right\|\cdot\|\Delta A\|\cdot\|x\| + \left\|A^{-1}\Delta A\right\|\cdot\|\Delta x\|.$$

Оскільки $(1 - \left\|A^{-1}\Delta A\right\|)\|\Delta x\| \le \left\|A^{-1}\right\|(\|\Delta b\| + \|\Delta A\|\cdot\|x\|)$ і $1 - \left\|A^{-1}\Delta A\right\| > 0,$ то

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

Нарешті маємо

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

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

Якщо ж $\|A^{-1}\|\|\Delta A\| \le 1$, то має місце менш точна оцінка

$$\frac{\|x-\tilde{x}\|}{\|x\|} \le \frac{\kappa(A)}{1-\kappa(A)\frac{\|\Delta A\|}{\|A\|}}\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right). \quad \blacksquare$$

3.5. Похибка розв'язку СЛАР унаслідок заокруглення у правій частині

Розглянемо питання про похибку розв'язку внаслідок заокруглення у правій частині системи (3.1). Нехай t – кількість двійкових розрядів в арифметиці з плаваючою крапкою. Кожен елемент b_i у правій частині заокруглюється з відносною похибкою, яка не перевищує $2^{1-t}/2 = 2^{-t}$ і з абсолютною похибкою $2^{-t}|b_i|$. Тому $\|\Delta b\| = \|b\|2^{-t}$ і $\|\Delta b\|/\|b\| \le 2^{-t}$. Отже,

$$\frac{\|\Delta x\|}{\|x\|} \le 2^{-t}\,\kappa(A)\ \cdot \qquad (3.6)$$

В обчислювальній практиці питання про строгу оцінку похибки наближеного розв'язку СЛАР за допомогою одержаних нерівностей розглядається рідко. Але інформація про порядок похибки розв'язку корисна для якісних висновків про те, з якою точністю доцільно розв'язувати задачу. Оцінка (3.6) дає змогу оцінити похибку розв'язку зверху, що є наслідком похибки вхідних даних. Ця оцінка досить точна, тому здебільше нема потреби одержувати розв'язок з похибкою, значно меншою значення $\kappa(A)\cdot 2^{-t}$.

3.6. Регуляризація систем лінійних алгебраїчних рівнянь

Нехай матриця A системи рівнянь (3.1) погано обумовлена, тобто число $\kappa(A)$ досить велике. Це означає, що похибки елементів матриці і правої частини, або похибки заокруглення при обчисленні, можуть досить сильно вплинути на розв'язок системи. Для зменшення похибки можна провести обчислення з подвійною точністю, але за наявності похибок коефіцієнтів це не може дати бажаного результату. Можна змінити саму задачу, замінивши її іншою лінійною системою, яка ліпше обумовлена, з розв'язком, близьким до розв'язку системи (3.1). Інші способи уточнення розв'язки для системи із погано обумовленою матрицею наведені в [23, с. 307-315; 69].

Систему (3.1) можна замінити задачею мінімізації

$$f_0(x) := (Ax-b, Ax-b) = \|Ax-b\|^2 \to \min\,, \qquad (3.7)$$

де $\|\cdot\|$ – евклідова норма. Покажемо, що задачі (3.1) і (3.7) рівносильні. Справді, якщо x_0 – розв'язок (3.1), то $\|Ax_0-b\|=0$ і найменше значення 0 функції $f_0(x)$ досягається. Нехай тепер x – розв'язок задачі (3.7). Необхідна умова екстремуму набуває вигляду

$$\frac{\partial}{\partial x} f_o(x) = \frac{\partial}{\partial x}(Ax-b_1, Ax-b) = \frac{\partial}{\partial x}((Ax-b)^T(Ax-b)) = 0.$$

Обчислимо спочатку похідну $\frac{\partial}{\partial x}(x^T x) = \frac{\partial}{\partial x}(x_1^2+\ldots+x_n^2) = 2x^T$.

Отже,

$$\frac{\partial f_0(x)}{\partial x} = 2(Ax-b)^T \frac{\partial}{\partial x}(Ax-b) = 2(Ax-b)^T A = 0$$

тоді і тільки тоді, коли $Ax-b=0$, тобто x є розв'язком (3.1).

Якщо елементи матриці A або вектора b задаються з похибками, то й розв'язок також наближений. Тому замість (3.7) виконується наближена рівність

$$(Ax-b, Ax-b) \approx 0.$$

Щоб позбутися невизначеності, накладемо додаткові умови, наприклад, будемо вибирати x так, щоб мінімізувати функцію

$$f_\alpha(x) = (Ax-b, Ax-b) + \alpha(x,x),$$

де α – малий додатний параметр.

Із необхідної умови екстремуму одержимо

$$\frac{\partial f_\alpha(x)}{\partial x} = 2(Ax-b)^T \frac{\partial}{\partial x}(Ax-b) + 2\alpha x^T = 2(Ax-b)^T A + 2\alpha x^T = 0.$$

Виконавши операцію транспонування, одержимо систему рівнянь вигляду

$$(A^T A + \alpha I)x = A^T b. \qquad (3.8)$$

Зауважимо, що матриця системи симетрична, а сама система (3.8) називається *регуляризованою*. Розв'язуючи систему (3.8) із симетричною матрицею одним із методів, одержимо регуляризоване значення $x = x(\alpha)$.

Якщо $\alpha = 0$, то маємо погано обумовлену систему (3.1). Якщо ж α не мале, то матриця системи (3.8) буде добре обумовленою, завдяки доданку αI, але значення $x(\alpha)$ може не бути близьким до розв'язку системи (3.1). Оптимальним буде найменше значення α, при якому обумовленість матриці системи (3.1) ще буде задовільною. Значення α вибирається експериментально, розв'язуючи систему (3.8) із значеннями $\alpha_\nu, \alpha_1 > \alpha_2 > \ldots$. Зокрема, для знаходження α обчислюють нев'язку $r(x_\alpha) = b - Ax_\alpha$ і порівнюють її за нормою з відомою похибкою Δb та впливом похибки коефіцієнтів матриці $\Delta A \cdot x$. Якщо α досить велике, то нев'язка помітно більша від цих похибок, якщо ж α надто мале, то відчутно менша. Оптимальним вважається те значення α, для якого $\|r(x_\alpha)\| \approx \|\Delta b\| + \|\Delta A \cdot x_\alpha\|$.

Приклади розв'язування типових задач

Задача 1. Знайти оцінку відносної похибки розв'язку системи лінійних рівнянь

$$1.005x_1 + x_2 = 2.005,$$
$$2x_1 + 2x_2 = 4,$$

розв'язок якої $x_1 = x_2 = 1$, а збурення $\Delta A = \begin{bmatrix} 0.001 & 0 \\ 0 & 0 \end{bmatrix}$.

Розв'язування. Для норми $\|\cdot\|_\infty$ маємо $\kappa(A) = 1020.$ Число обумовленості досить велике, тому відносна похибка розв'язку може бути досить великою. Оскільки $\|\Delta A\| / \|A\| = 0.00025,$ то

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{1202}{1-1202\cdot 0.00025}0.00025 = 0.43,$$

що в 2.6 разу перевищує точне значення похибки $\frac{\|\Delta x\|}{\|x\|} = \|\Delta x\| =$ $= \max(|0.8333-1|, 1.1667-1) = 0.1667$. При цьому відношення похибок $\frac{\|\Delta x\|}{\|x\|} : \frac{\|\Delta A\|}{\|A\|} \approx 467.$

Задача 2. Знайти норму $\|\cdot\|_2$ оцінки відносної похибки розв'язку СЛАР

$$y_1 + 0.99y_2 = 1.99,$$
$$0.99y_1 + 0.98y_2 = 1.97,$$

точний розв'язок якої $x_1 = x_2 = 1$, якщо збурення правої частини $\delta b = (-0.000097, 0.000106)^T$.

Розв'язування. Збурена система рівнянь має вигляд

$$y_1 + 0.99y_2 = 1.989903,$$
$$0.99y_1 + 0.98y_2 = 1.970106.$$

Збурення правої частини породжує збурення розв'язку, евклідова норма якого

$\|\Delta x\| = \|y - x\| = \|[3.000693; -1.0210000]^T - [1; 1]\| = \sqrt{(\Delta x_1)^2 + (\Delta x_2)^2} =$ $= 2.843803, \ \|\Delta x\|/\|x\| = 2.843803/\sqrt{2} = 2.010872.$

Відносна похибка правої частини $\|\Delta b\|/\|b\| = \frac{0.000144}{2.800179} = 0.000051$. Власні значення матриці A знаходяться із рівняння $\begin{bmatrix} 1-\lambda & 0.99 \\ 0.99 & 0.98-\lambda \end{bmatrix} = = 0$ або $\lambda^2 + 1.98\lambda - 0.001 = 0$. Одержимо $\lambda_1 = -1.98005, \ \lambda_2 = 0.000051.$ Далі, $\kappa(A) = |\lambda_1|/\lambda_2 = 39202 \gg 1,$ тому матриця погано обумовлена. Відношення $\frac{\|\Delta x\|}{\|x\|} : \frac{\|\Delta b\|}{\|b\|} = 39128,$ що близьке до $\kappa(A)$.

Задача 3. Регуляризувати СЛАР із задачі 1.

Розв'язування. Для збуреної системи

$$1.005y_1 + y_2 = 2,$$
$$2y_1 + 2y_2 = 4,$$

маємо $A^TA = \begin{bmatrix} 5.010025 & 5.005 \\ 5.005 & 5 \end{bmatrix}$, $A^TA = \begin{bmatrix} 10.01 \\ 10 \end{bmatrix}$. Система (3.8) набуває вигляду

$$(5.010025 + \alpha)z_1 + 5.005z_2 = 10.01,$$
$$5.005z_1 + (5 + \alpha)2z_2 = 10.$$

Для $\alpha = 0.1$ маємо систему рівнянь

$$5.110025z_1 + 5.005z_2 = 10.01,$$
$$5.005z_1 + 5.1z_2 = 10.$$

Її розв'язок $z_1 = 1.99761,\ z_2 = 0.00039.$ Міра обумовленості $\kappa(A) \ll 1202.$ Для $\alpha = 0.05$ вже $z_1 = 0.994830,\ z_2 = 0.994830.$ Ще ліпше наближення $z_1 = 0.99711,\ z_2 = 0.99691$ одержується для $\alpha = 0.025$, причому $\kappa(A) = 401.6.$ Подальше зменшення α до 0.001 вже погіршує результат.

Завдання та запитання для самостійної роботи

1. Знайти сталі еквівалентності, що зв'язують норми $\|x\|_\infty, \|x\|_1, \|x\|_2$, а також вектори, на яких досягаються рівності.

   Відповідь: $\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty, \|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty,\ \frac{1}{\sqrt{n}}\|x_1\| \le \|x_2\| \le \|x_3\|.$

2. Знайти матричні норми, підпорядковані векторним нормам $\|x\|_\infty$ і $\|x\|_1$.

3. Для системи $Ax = b$ із квадратною матрицею, $\det A \neq 0$, довести оцінку

$$\|x - \tilde{x}\| \le \|A^{-1}\| \cdot \|r\|,$$

   де $r = b - A\tilde{x}$, $\tilde{x}$ – наближений розв'язок.

4. Нехай $\|\Delta b\| \le \delta\|b\|, \delta \cdot \kappa(A) = r < 1.$ Чи правильна оцінка

$$\frac{\|\Delta x\|}{\|x\|} \le \kappa(A)\frac{2\delta}{1-r}?$$

Інші оцінки похибки розв'язку СЛАР одержані в [20, с. 83-84].

5. Для матриці $A=\begin{bmatrix}1.000 & 1.001\\ 1.000 & 1.000\end{bmatrix}$ обчислити число $\kappa(A)$ для норм $\|\cdot\|_1$, $\|\cdot\|_2$ і $\|\cdot\|_\infty$. Чи можна вважати матрицю погано обумовленою?

6. Знайти число обумовленості в нормі $\|\cdot\|_1$ для матриці

$$A=\begin{bmatrix}1 & 1\\ 0 & 10^{-7}\end{bmatrix}$$

та оцінити це число за допомогою алгоритму з розділу 3.2.

7. Нехай A – невироджена матриця і $\|A^{-1}\Delta A\|<1$. Показати, що матриця $A+\Delta A$ не вироджена і виконується оцінка

$$\|(A+\Delta A)^{-1}\|\le\frac{\|A^{-1}\|}{1-\|A^{-1}\Delta A\|}.$$

8. Показати, що $\kappa(A)\ge 1$ для довільної невиродженої матриці і $\kappa(A)=1$ для довільної ортогональної матриці і норми $\|\cdot\|_2$. Матриця A – ортогональна, якщо $AA^T=I$.

9. Чи можна стверджувати, що якщо визначник матриці малий, то матриця погано обумовлена? Перевірити це для матриці $D=\varepsilon I$, де $0<\varepsilon\ll 1$.

10. Показати , що матриця порядку *n*

$$A=\begin{bmatrix}1 & -1 & -1 & \dots & -1\\ 0 & 1 & -1 & \dots & -1\\ \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & \dots & 1\end{bmatrix}$$

погано обумовлена. *Вказівка*. Обчислити $\kappa(A)$, яке для норми $\|\cdot\|_\infty$ дорівнює $n\cdot 2^{n-1}$. Як змінюється число обумовленості матриці з ростом її порядку? Розглянути випадки, коли порядок дорівнює 10, 50, 100, 1000.

11. Здійснити *LU*-розклад матриці $A=\begin{pmatrix}\varepsilon & 1\\ 1 & 1\end{pmatrix}$, $0<\varepsilon\ll 1$. Обчислити числа обумовленості матриць *A, L* і *U* в нормі $\|\cdot\|_\infty$, порівняти їх та зробити висновки. Виконати ті ж дії за схемою Ґауса з вибором головного елемента.

12. Обчислити $\kappa(A)$ в нормі $\|\cdot\|_1$ і $\|\cdot\|_\infty$ матриці системи

$$\begin{cases}2x_1+3x_2-4x=1,\\ 5x_1-6x_2+2x_3=1,\\ 7x_1-3x_2-(2-10^{-k})x_3=2+10^{-k}.\end{cases}$$

і зробити висновок про міру її обумовленості для $k = 1, 5, 10.$

13. Для СЛАР

$$\begin{aligned} 1.000x_1 + 1.001x_2 &= 2.001 \\ 1.000x_1 + 1.000x_2 &= 2.001 \end{aligned} \quad \text{і} \quad \begin{aligned} x_1 + 10x_2 &= 11, \\ 100x_1 + 10001x_2 &= 1101 \end{aligned}$$

знайти оцінку відносної похибки $\|\Delta x\| / \|x\|$ розв'язку, якщо похибка правої частини СЛАР $\Delta b = (-0.001; 0.000)^T$ і $\Delta b = (0.01; 0)$ відповідно. Порівняти одержану оцінку з точним значенням і з відносною похибкою $\|\Delta b\| / \|b\|$.

14. Нехай L – нижня трикутна матриця, $l_{ii} \neq 0, i = \overline{1, n}.$ Система рівнянь $Lx = b$ розв'язується послідовним визначенням x_1, x_2 і т.д. Показати, що при відсутності переповнень і машинних нулів знайдений розв'язок $\tilde{x}$ задовольняє систему $(L + \Delta L)\tilde{x} = b$, де $|\Delta l_{ij}| \le n\varepsilon |l_{ij}|$ і ε – "машинне епсилон".

15. На прикладі матриць A

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \text{ і } \begin{bmatrix} 1 & 2 \\ 1 & 2.01 \end{bmatrix}$$

перевірити, що

$$\kappa(A) \approx (\max_j \|a_j\|) \frac{\|z\|}{\|y\|},$$

де y і z – розв'язки СЛАР $A^T y = l;\quad Az = y,$ вектор l з компонентами ± 1 вибирається так, щоб максимізувати зростання у процесі оберненої підстановки для вектора y [78].

16. Дослідити вплив збурення елементів головної діагоналі тридіагональних матриць A і B на роз'язок СЛАР $Ax = b$ з ростом їх порядку, якщо

$$A = \begin{bmatrix} 3 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 3 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 3 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 3 \end{bmatrix},$$

$$B=\begin{bmatrix} 3 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 2 & 2 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 2 & 2 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 3 \end{bmatrix},\ b=\begin{bmatrix} 4 \\ 5 \\ 5 \\ \vdots \\ 4 \end{bmatrix}.$$

17. Показати, що в методі Ґауса $\dfrac{\|\Delta x\|}{\|x\|}=\kappa(A)O(n2^{-t})$, де $n-$ порядок СЛАР, $t-$кількість розрядів мантиси і $\|\Delta A\|/\|A\|=O(n2^{-t})$.
18. Для СЛАР
$$\begin{aligned} x_1+2x_2+3x_3+5.00001x_4&=11.00001,\\ 2x_1+3x_2+4x_3+7.00002x_4&=16.00002,\\ 4x_1+4x_2+5x_3+9.00003x_4&=22.00003,\\ 5x_1+5x_2+5x_3+10.00004x_4&=25.00004, \end{aligned}$$
розв'язок якої $x_1=x_2=x_3=x_4=1$, знайти оцінку нормі $\|.\|_\infty$ або $\|.\|_1$ похибки розв'язку і порівняти її з точним значенням похибки, якщо збурення правої частини $\Delta b=(-0.00001,-0.00002,-0.00003,-0.00004)^T$.
19. Для матриці Гільберта, елементи якої $a_{ij}=(i+j-1)^{-1}, i,j=\overline{1,n}$, обчислити число обумовленості $\kappa(A)$, число критерію Ортеги κ_ν і кутового критерію κ_a для $n-2,5,10$. Зробити висновок про обумовленість матриць за цими критеріями.
20. Для матриці $A=\begin{bmatrix} 1 & 2 \\ 1 & 2+10^{-6} \end{bmatrix}$
обчислити числа обумовленості $\kappa(A)$, число критерію Ортеги κ_ν і кутового критерію κ_a. Порівняти їх між собою та зробити висновок про обумовленість матриць.
21. Показати, що для ермітової матриці в нормі $\|\ \|_2$ виконується рівність $\kappa(A^2)=(\kappa(A))^2$. Зробити висновок про ріст числа обумовленості з зростання степеня матриці.
22. Навести приклад несиметричної матриці, для якої $\kappa(A^2)=(\kappa(A))^2$.
23. Застосувати процес регуляризації для розв'язку СЛАР із задачі 12.
24. Довести, що $\kappa(AB)\le\kappa(A)\kappa(B)$.

Розділ 4. Ітераційні методи розв'язування СЛАР

Класифікація ітераційних методів. Метод простої ітерації (МПІ) і Зейделя. Метод Якобі і Ґауса–Зейделя. Необхідна і достатня та достатня умови збіжності МПІ. Оцінка похибки МПІ. Метод Річардсона з найменшою похибкою на $k-$й ітерації. Метод релаксації. Методи варіаційного типу: найшвидшого спуску, мінімальних нев'язок, спряжених градієнтів.

Література [5, 6, 13, 20, 23, 31, 59, 73, 76, 79, 80, 83, 93]
Електронні джерела [103 – 107]

4.1. Канонічна форма однокрокових ітераційних методів

Розглянемо СЛАР з квадратною матрицею порядку *n*

$$Ax = b. \qquad (4.1)$$

В однокрокових ітераційних методах задається початкове наближення $x^{(0)}$ і будується послідовність наближень $x^{(0)}, x^{(1)}, \ldots,$ а розв'язок x^* системи (4.1) одержується як границя цієї послідовності, тобто $x^* = \lim\limits_{k\to\infty} x^{(k)}$. Наближеним розв'язком служить вектор $x^{(k)}$ для деякого значення $k = k_0$, обумовленого точністю обчислень. В *однокрокових* ітераційних методах $x^{(k+1)}$ обчислюється за відомим значенням наближення $x^{(k)}$, у *багатокрокових* – за значеннями $x^{(k)}, \ldots, x^{(k-m)}$, $m > 1$.

Канонічною формою однокрокових ітераційних методів розв'язування системи (4.1) називається їх запис у вигляді [59]

$$B_{k+1}\frac{x^{(k+1)} - x^{(k)}}{\tau_{k+1}} + Ax^{(k)} = b, \quad k = 0,1,..., \qquad (4.2)$$

де τ_{k+1} – ітераційний параметр, B_{k+1} – невироджена матриця, яка вибирається так, щоб обернена матриця B_{k+1}^{-1} обчислювалась досить просто, наприклад діагональною або трикутною.

Задамо початкове наближення $x^{(0)}$, наприклад, $x^{(0)} = b$. Із (4.2) знаходимо на $(k+1)-$й ітерації

$$x^{(k+1)} = (I - \tau_{k+1}B_{k+1}^{-1}A)x^{(k)} + \tau_{k+1}B_{k+1}^{-1}b, \ k = 0,1,...$$

Матриця $P_k = (I - \tau_{k+1}B_{k+1}^{-1}A)$ називається *ітераційною*. Ітераційний метод *явний*, якщо $B_{k+1} \equiv I$, інакше – неявний. Ітераційний метод

стаціонарний, якщо $B_{k+1} = B$ і $\tau_{k+1} = \tau$, тобто не залежать від номера ітерації, і нестаціонарний – в інших випадках.

4.2. Метод простої ітерації

Нехай система (4.1) зведена до рівносильного вигляду

$$x = Px + f, \tag{4.3}$$

де P – квадратна матриця порядку n, f – n-вектор, і задано початкове наближення $x^{(0)}$, за яке можна взяти $x^{(0)} = f$. В МПІ наближення $x^{(k+1)}$ обчислюється за рекурентною формулою

$$x^{(k+1)} = Px^{(k)} + f, \quad k = 0,1,... \tag{4.4}$$

або в координатній формі

$$x_i^{(k+1)} = \sum_{j=1}^{n} p_{ij} x_j^{(k)} + f_i .$$

Цей метод набуває вигляду (4.2), якщо $B_{k+1} = B$, де B – довільна невироджена матриця, а $\tau_{k+1} = 1$. Тоді ітераційною є матриця $P = -B^{-1}A$. Якщо ітераційний процес збіжний, то x^* – розв'язок системи (4.3), отже, і системи (4.1). Справді, при $k \to \infty$ з (4.4) випливає, що

$$x^* = Px^* + f .$$

Зображення системи рівнянь у вигляді (4.4) можна одержати, припустивши, що в (4.1) $A = A_1 + A_2$, $\det A_1 \neq 0$. Тоді

$$x = -A_1^{-1}A_2 x + A_1^{-1}b.$$

Тут $P = -A_1^{-1}A_2$, $f = A^{-1}b$. Матрицю A_1 вибирають так, щоб виконувались умови збіжності методу, а обернена матриця A_1^{-1} нескладно обчислювалась. Інший спосіб полягає в зображенні системи (4.1) у вигляді

$$x = x + F(b - Ax),$$

де F – матриця, така, що $\det F \neq 0$, або деяке число, $F \neq 0$. Тоді

$$x = (I - FA)x + Fb, \; P = I - FA.$$

На кожній з ітерацій в (4.4) виконується n^2 операцій множень і n^2 додавань. За n ітерацій кількість операцій досягне $2n^3$ (метод Ґауса вимагає приблизно $n^3/3$ множень і ділень).

Розглянемо одну з реалізацій методу простої ітерації, відому як *метод Якобі* [5, 20, 22, 23]. Нехай коефіцієнт $a_{ii} \neq 0$, інакше

можна поміняти i-те рівняння з іншим або утворити лінійну комбінацію з рівнянь так, щоб коефіцієнт при x_i не дорівнював нулю. З i-го рівняння знаходимо

$$x_i = \left(b_i - \sum_{j \neq i} a_{ij} x_j \right) / a_{ii}, \quad i = \overline{1, n}. \tag{4.5}$$

Починаючи з початкового наближення $x^{(0)}$, обчислимо $x^{(k+1)}$ за допомогою рекурентних формул:

$$x_i^{(k+1)} = \left(b_i - \sum_{j \neq i} a_{ij} x^{(k)} \right) / a_{ii}, \quad i = \overline{1, n}; \quad k = 0, 1, \ldots \tag{4.6}$$

Запишемо метод Якобі (4.6) у матричній формі. Для цього введемо такі матриці:

$$H = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ a_{21} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & 0 \end{bmatrix}, \; D = \begin{bmatrix} a_{11} & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & a_{nn} \end{bmatrix},$$

$$G = \begin{bmatrix} 0 & a_{12} & \ldots & a_{1n} \\ 0 & 0 & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 \end{bmatrix}.$$

Тоді $A = H + D + G$ і система (4.1) набуває вигляду

$$Dx = -Hx - Gx + b \text{ або } x = -D^{-1}(H + G)x + D^{-1}b.$$

Ітерації в методі Якобі набувають вигляду:

$$x^{(k+1)} = -D^{-1}(H + G)x^{(k)} + D^{-1}b.$$

Метод набуває канонічного вигляду (3.2), якщо покласти $B_{k+1} = D$ і $\tau_{k+1} = 1$, ітераційна матриця $P = -D^{-1}(H + G)$.

Приклад 4.1. Для системи рівнянь

$$3x_1 + x_2 + x_3 = 5,$$
$$x_1 - 5x_2 + x_3 = -3,$$
$$2x_1 - 3x_2 + 6x_3 = 5$$

метод Якобі записується у вигляді

$$x_1^{(k+1)} = (-x_2^{(k)} - x_3^{(k)} + 5)/3,$$
$$x_2^{(k+1)} = (x_1^{(k)} + x_3^{(k)} + 3)/5,$$
$$x_3^{(k+1)} = (-x_1^{(k)} + 3x_2^{(k)} + 5)/6.$$

4.3. Метод Зейделя

У формулах (4.4) при обчисленні наближеного значення $x_i^{(k+1)}$, $i \geq 2$, доцільно використати вже знайдені значення $(k+1)$-го наближення $x_1^{(k+1)},..., x_{i-1}^{(k+1)}$, що може прискорити збіжність методу. Така модифікація методу простої ітерації називається *методом Зейделя* і записується у вигляді

$$x_1^{(k+1)} = \sum_{j=1}^{n} p_{1j} x_j^{(k)} + f_1, \tag{4.7}$$

$$x_i^{(k+1)} = \sum_{j=1}^{i-1} p_{ij} x_j^{(k+1)} + \sum_{j=i}^{n} p_{ij} x_j^{(k)} + f_i, \quad i = \overline{2,n}.$$

У матричній формі ітерації (4.7) запишуться так

$$x^{(k+1)} = H_1 x^{(k+1)} + (D_1 + G_1) x^{(k)} + f,$$

де $P = H_1 + D_1 + G_1$, а матриці H_1, D_1 і G_1 мають такий же вигляд, як H, D і G із заміною a_{ij} на p_{ij}. Якщо існує матриця $(I - H_1)^{-1}$, то

$$x^{(k+1)} = (I - H_1)^{-1}(D_1 + G_1)x^{(k)} + (I - H_1)^{-1} f, \quad k = 0,1,...$$

Застосування процедури Зейделя до системи рівнянь (4.5) приводить до таких формул:

$$x_1^{(k+1)} = (b_1 - \sum_{j=2}^{n} a_{1j} x_j^{(k)}) / a_{11},$$

$$x_1^{(k+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)}) / a_{ii}, \quad i = \overline{2,n-1}, \tag{4.8}$$

$$x_n^{(k+1)} = (b_n - \sum_{j=1}^{n-1} a_{nj} x_j^{(k+1)}) / a_{nn}.$$

У матричній формі одержимо

$$x^{(k+1)} = -D^{-1} H x^{(k+1)} - D^{-1} G x^{(k)} + D^{-1} b.$$

Цей варіант методу Зейделя називають *методом Ґауса-Зейделя* й одержується з (4.3), коли $B_{k+1} = D + H$ і $\tau_{k+1} = 1$.

Для системи з прикладу 4.1 метод Ґауса-Зейделя набуває вигляду

$$x_1^{(k+1)} = (-x_2^{(k)} - x_3^{(k)} + 5)/3,$$
$$x_2^{(k+1)} = (x_1^{(k+1)} + x_3^{(k)} + 3)/5,$$
$$x_3^{(k+1)} = (-x_1^{(k+1)} + 3x_2^{(k+1)} + 5)/6, \quad k = 0,1,...$$

Зауважимо, що вигляд формул (4.8) складніший, порівняно з (4.4), але метод Зейделя програмується простіше. Справді, для

методу простої ітерації потрібно два масиви для $x^{(k)}$ і $x^{(k+1)}$. Після обчислення $x^{(k+1)}$ здійснюється присвоєння $x^{(k)} := x^{(k+1)}$. У методі Зейделя досить одного масиву, *i*-й компоненті якого одразу присвоюється значення $x_i^{(k+1)}$.

Як правило, метод Зейделя збігається щвидше, ніж метод простої ітерації, оскільки використовуються вже знайдені у попередніх рівняннях наближення. Метод Зейделя може збігатись, якщо розбігається метод простої ітерації, і навпаки. Перевагою методу Зейделя є його збіжність для нормальних систем лінійних рівнянь, тобто таких в яких матриця A симетрична і додатньо визначена. Цього можна досягти для невиродженої матриці A, якщо помножити її зліва на A^T і розв'язати СЛАР $A^T Ax = A^T b$.

4.4. Умови збіжності методу простої ітерації

4.4.1. Означення та допоміжні леми. Нехай $\lambda_1, \lambda_2, ..., \lambda_n$ – влас-ні значення матриці P, тобто корені характеристичного рівняння

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

Лема 4.1. *Для довільної норми матриці P справджується нерівність*

$$\rho(P) \le \|P\|. \qquad (4.9)$$

Доведення. Нехай спектральний радіус матриці $\rho(P) = |\lambda|$, x – відповідний власному значенню λ власний вектор, тобто $Px = \lambda x, \quad x \ne 0$. Утворимо квадратну матрицю $X = [x, x, ..., x]$. Тоді $\lambda X = PX$. Перейшовши до оцінки норм, одержимо

$$\rho(P)\|X\| = \|PX\| \le \|P\| \cdot \|X\|,$$

звідки й випливає оцінка (4.9). ■

Нехай P^k − добуток *k* матриць *P*, $\bar{0}$ − нуль-матриця.

Означення 4.1. *Послідовність* $P^k \to \bar{0}$ *при* $k \to \infty$, *якщо* $\|P^k\| \to 0$ *при* $k \to \infty$.

Лема 4.2 [36, с. 19-20]. $\|P^k\| \to 0$ *при* $k \to \infty$ *тоді і тільки тоді, коли спектральний радіус задовольняє умову*

$$\rho(P) < 1. \qquad (4.10)$$

Лема 4.3. *Якщо для деякої норми виконується нерівність* $\|P\| < 1$, *то* $P^k \to \bar{0}$ *при* $k \to \infty$.

Справді, $\|P^k\| \le \|P\|^k$, тому $\|P^k\| \to 0$, коли $k \to \infty$. ■

Розглянемо матричний ряд

$$I + P + P^2 + ... + P^k + ... \qquad (4.11)$$

Означення 4.2. *Матричний ряд збіжний і має границею матрицю* S, *якщо*

$$\lim_{k\to\infty} \|I + P + P^2 + ... + P^k - S\| = 0.$$

Лема 4.4. *Для збіжності матричного ряду (4.10) необхідно і досить виконання умови* (4.10). *Матриця* $I - P$ *має обернену і*

$$I + P + P^2 + ... = (I - P)^{-1}. \qquad (4.12)$$

Необхідність. Нехай ряд (4.11) – збіжний. Тоді $P^k \to \bar{0}$ при $k \to \infty$ і з леми 4.2 маємо $\rho(P) < 1$. Оскільки

$$(I - P)(I + P + ... + P^{k-1}) = I - P^k,$$

То, перейшовши до границі при $k \to \infty$, одержимо

$$(I - P)(I + P + P^2 + ...) = I, \qquad (4.13)$$

звідки випливає рівність (4.12).

Достатність. На підставі умови (4.10) $P^k \to \bar{O}$ при $k \to \infty$. Тому виконується рівність (4.13), отже, і (4.12). ■

Лема 4.5. *Якщо* $\| P \| < 1$ *для деякої норми, то матричний ряд (4.11) збіжний і виконується рівність (4.12).* ■

Досягнути виконання достатньої умови $\| P \| < 1$ можна, зобразивши СЛАР у вигляді

$$x = (I - \tau A)x + \tau b,$$

де значення τ потрібно вибрати так, щоб $\|I - \tau A\| < 1$. Наприклад, для матриці A і параметра τ

$$A = \begin{bmatrix} 1 & -2 & -3 \\ 4 & 5 & 6 \\ -7 & 8 & 3 \end{bmatrix}, \ \tau = 0.05$$

маємо $I - \tau A = \begin{bmatrix} 0.95 & 0.10 & 0.15 \\ -0.20 & 0.80 & -0.30 \\ 0.35 & 0.40 & 0.85 \end{bmatrix}$, яка є матрицею з перевагою головної діагоналі.

4.4.2. Необхідна і достатня умова збіжності

Теорема 4.1. *Метод простої ітерації збігається до єдиного розв'язку системи (4.4) для довільного початкового наближення* $x^{(0)} \in R^n$ *тоді і тільки тоді, коли всі власні значення матриці* P *лежать усередині одиничного круга.*

Доведення. Запишемо ітераційні процес (4.4) у вигляді

$$x^{(k+1)} = P(Px^{(k-1)} + f) + f = P^2 x^{(k-1)} + (I + P) f = \ldots$$
$$= P^{k+1} x^{(0)} + (I + P + P^2 + \ldots + P^k) f. \qquad (4.14)$$

На підставі лем 4.2 і 4.4 умова (4.10) необхідна і достатня для того, щоб $P^{k+1} \to \bar{0}$ при $k \to \infty$ і ряд (4.11) збігався. Тому $\lim\limits_{k\to\infty} x^{(k+1)} = x^*$ для довільного $x^{(0)} \in R^n$ і з (4.14) випливає

$$x^* = (I - P)^{-1} f.$$

Звідси маємо, що $(I - P)x^* = f$ і $x^* = Px^* + f$.

Отже, x^* – розв'язок системи (4.4). Оскільки $x - x^* = P(x - x^*)$ і $(I - P)(x - x^*) = 0,$ то з невиродженості матриці $I - P$ випливає, що $x = x^*,$ тобто розв'язок єдиний. ■

Наслідок 4.1. *Необхідною і достатньою умовою збіжності методу Якобі (4.6) є умова того, що всі корені алгебраїчного рівняння*

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22}\lambda & \ldots & a_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn}\lambda \end{vmatrix} = 0$$

лежать усередині одиничного круга.

Справді, метод Якобі є методом простої ітерації з матрицею $P = -D^{-1}(H + G).$ Для матриці P маємо

$$\det(-D^{-1}(H + G) - \lambda I) = -\det D^{-1} \cdot \det(H + G + \lambda D) = 0.$$

Отже, $\det(H + G + \lambda D) = 0.$

Наслідок 4.2. *Необхідною і достатньою умовою збіжності методу Ґауса–Зейделя є умова того, що корені рівняння*

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \ldots & a_{1n} \\ a_{21}\lambda & a_{22}\lambda & \ldots & a_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}\lambda & a_{22}\lambda & \ldots & a_{2n}\lambda \end{vmatrix} = 0$$

лежать усередині одиничного круга $|\lambda|<1$. Справді, для методу Ґауса–Зейделя $P=-(I+D^{-1}H)^{-1}D^{-1}G$. Тому

$$\det(P-\lambda I)=-\det(I+D^{-1}H)^{-1}\cdot\det(D^{-1}G+\lambda(I+D^{-1}H))=$$
$$=-\det(I+d^{-1}H)^{-1}\cdot\det D^{-1}\cdot\det(G+\lambda D+\lambda H).$$

Завершення доведення одержується з теореми 4.1.

4.4.3. Достатня умова збіжності

Теорема 4.2. *Якщо матриця* P *така, що* $\|P\|<1$ *для деякої матричної норми, тоді метод простої ітерації збіжний до єдиного розв'язку системи рівнянь (4.4) для довільного початкового наближення* $x^{(0)}\in R^n$.

Доведення випливає з наслідку леми 4.5, оскільки при $\|P\|<1$ в рівності (4.7) матричний ряд збіжний і $P^k\to\overline{O}$ при $k\to\infty$.

Наслідок 4.3. *Якщо виконується нерівність*

$$\|D^{-1}(H+G)\|<1,$$

то метод Якобі збіжний. ■

Для норми $\|\cdot\|_1$ нерівність (4.17) набуває вигляду

$$\max_{i\le j\le n}\sum_{i\ne j}\left|\frac{a_{ij}}{a_{ii}}\right|<1$$

або

$$\sum_{i\ne j}|a_{ij}|<|a_{jj}|,\ j=\overline{1,n}. \qquad (4.15)$$

Аналогічно для норми $\|\cdot\|_\infty$ маємо

$$\max_{i\le i\le n}\sum_{j\ne i}\left|\frac{a_{ij}}{a_{ii}}\right|<1,$$

інакше

$$\sum_{j\ne i}|a_{ij}|<|a_{ii}|,\ i=\overline{1,n}. \qquad (4.16)$$

Умови (4.15) і (4.16) забезпечують перевагу головної діагоналі в матриці A відповідно по стовпцях і рядках. У прикладі 4.1 матриця A є матрицею з переважаючою головною діагоналлю по рядках і стовпцях, тому обидві умови виконується. У деяких випадках перевага головної діагоналі досягається, якщо поміняти місцями деякі рівняння або виконати елементарні перетворення над рівняннями системи.

Означення 4.3. *Матриця* $A>0$ *називається додатно визначеною* $(A>0)$*, якщо*

$$x^T Ax=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_i x_j>0\ \forall x\in R^n, x\neq 0.$$

Необхідною і достатньою умовою того, що $A>0$, є додатність всіх головних міноpів матриці A [19]. Наприклад, матриця

$$A=\begin{bmatrix}1 & 2 & 3\\ 2 & 5 & 4\\ 3 & 6 & 28\end{bmatrix}>0,$$

оскільки $\Delta_1=1>0,\ \Delta_2=\begin{vmatrix}1 & 2\\ 2 & 5\end{vmatrix}=1>0$ і $\det A=1>0$. Зауважимо, що власні значення матриці $A>0$ дійсні і додатні. Справді, із рівності $Ax=\lambda x,\ x\neq 0$ одержимо $x^{\tau}Ax=\lambda x^{\tau}x=\lambda\|x\|_2^2$. Оскільки $x^T Ax>0$, то $\lambda>0$.

Нехай $A>0$. Візьмемо $\tau=q/\|A\|,\ 0<q<2$. Тоді

$$x=(I-\tau A)x+\tau b=(I-\frac{q}{\|A\|}A)x+\tau b.$$

Для власних значень μ матриці $P=I-\frac{q}{\|A\|}A$ маємо

$$P-\mu I=-\frac{q}{\|A\|}\left[A-\frac{(1-\mu)\|A\|}{q}I\right].$$

Оскільки $\lambda(A)\le\|A\|$, то $0<\frac{(1-\mu)\|A\|}{q}\le\|A\|$, звідси $0<1-\mu\le q$ або $-1<1-q<\mu<1$. Тобто виконується необхідна і достатня умова збіжності методу простої ітерації.

Зауваження 4.1. *Виконання достатньої умови збіжності можна досягнути, перетворивши відповідно матрицю* A*. Нехай* $\det A\neq 0$*. Введемо матрицю* $D=A^{-1}-F$*, де* F *– матриця з досить малими елементами. Домноживши систему рівнянь (4.1) зліва на* D*, одержимо*

$$x=FAx+Db.$$

Для досить малих елементів F_{ij} *маємо* $\|FA\| \le \|F\| \cdot \|A\| < 1$. *Зокрема, якщо* $F = \sigma E$, *то при* $\sigma \le \left(2\|A\|\right)^{-1}$ *метод збіжний, оскільки* $\|FA\| \le 1/2$. *Зауважимо, що при цьому потрібно обчислити* A^{-1}, *що рівносильне знаходженню розв'язку СЛАР.*

4.5. Точність методу простої ітерації

Нехай виконується достатня умова $\| P \| < 1$ збіжності МПІ. Із (4.5) і (4.4) для єдиного розв'язку $x = x^*$ СЛАР одержимо

$$x^{(k+1)} - x^* = P(x^{(x)} - x^*) = P(x^{(k+1)} - x^*) + P(x^{(k)} - x^{(k+1)}).$$

Тому $\| x^{(k+1)} - x^* \| \le \| P \| \cdot \| x^{(k+1)} - x^* \| + \| P \| \cdot \| x^{(k+1)} - x^{(k)} \|$.

Звідси маємо:

$$\| x^{(k+1)} - x^* \| \le \frac{\| P \|}{1 - \| P \|} \| x^{(k+1)} - x^{(k)} \|. \qquad (4.17)$$

Із (4.4) і (4.5) випливає нерівність $\| x^{(k+1)} - x^* \| \le \|P\| \cdot \left\|x^{(k)} - x^*\right\|$, що означає лінійну швидкість збіжності МПІ.

Маючи $x^{(k)}$ і $x^{(k+1)}$, можна можна оцінити точність наближеного розв'язку $x^{(k+1)}$. Як випливає з (4.17), для заданої точності $\varepsilon > 0$ оцінка $\| x^{(k+1)} - x^* \| \le \varepsilon$ виконується, якщо

$$\| x^{(k+1)} - x^{(k)} \| < \frac{\varepsilon(1 - \| P \|)}{\| P \|}.$$

Оскільки

$$\| x^{(k+1)} - x^{(k)} \| \le \| P \| \cdot \| x^{(k)} - x^{(k-1)} \| \le \ldots \le \| P \|^k \cdot \| x^{(1)} - x^{(0)} \|,$$

то на підставі нерівності

$$\| x^{(k+1)} - x \| \le \frac{\| P \|^{k+1}}{1 - \| P \|} \| x^{(1)} - x^{(0)} \| \le \varepsilon \qquad (4.18)$$

можна оцінити кількість ітерацій, необхідних для досягнення заданої точності ε. Із (4.18) маємо

$$\| P \|^{k+1} \le \frac{\varepsilon(1 - \| P \|)}{\| x^{(1)} - x^{(0)} \|},$$

звідки

$$k \ge \left(\lg \frac{\varepsilon(1 - \| P \|)}{\| x^{(1)} - x^{(0)} \|} \right) / \lg \| P \|.$$

Вектор

$$r_k = b - Ax^{(k)}$$

називається нев'язкою й характеризує наскільки точно наближений розв'язок $x^{(k)}$ задовольняє систему (4.1). Зауважимо, що малість норми $\| r_k \|$ ще не гарантує високої точності розв'язку. Наприклад, для системи рівнянь

$$2.001x_1 + 1.000x_2 = 3.001,$$

$$2.000x_1 + 1.000x_2 = 3.000$$

наближений розв'язок $z_1 = 1.2,\ z_2 = 0.6$ відрізняється від точного розв'язку $x_1 = x_2 = 1$ на $\| x - x^* \|_\infty = 0.4$, а нев'язка $\| \delta \|_\infty = 0.0002$ мала.

Зауваження 4.2. *Інколи точність вважається досягненою, якщо виконується нерівність*

$$\| x^{(k+1)} - x^{(k)} \| < \varepsilon. \qquad (4.19)$$

Це так, якщо $\| P \| /(1 - \| P \|) \le 1$, *тоді* $\| P \| \le 0.5$. *Якщо ж* $0.5 < \| P \| < 1$, *то оцінка (4.19) може бути заниженою.*

Наприклад, коли $\| P \| = 0.9$, то, згідно з (4.17), маємо $\| x^{(k+1)} - x \| < 9 \| x^{(k+1)} - x^{(k)} \|$. Якщо ж $\| P \| = 0.1$, то оцінка (4.19) завищена, оскільки

$$\| x^{(k+1)} - x \| < (\| P \| /(1 - \| P \|)) \| x^{(k+1)} - x^{(k)} \| = \| x^{(k+1)} - x^{(k)} \| /9.$$

Якість ітераційного процесу зручно характеризувати також швидкістю спадання норми відношення похибки $z_k = x^{(k)} - x$ до початкової похибки $z_0 = x^{(0)} - x.$ Нехай $x^{(0)} = f$. Оскільки $z_k = Pz_{k-1} = P^k z_0$, то після k ітерацій маємо:

$$\sigma_k = \sup_{x^{(0)} \neq 0} \frac{\| z_k \|}{\| z_0 \|} = \sup_{z_0 \neq 0} \frac{\| P^k z_0 \|}{\| z_0 \|} = \| P^k \|.$$

Можна гарантувати, що для ρ, $0 < \rho < 1$, виконується оцінка $\sigma_k \le \rho$, якщо $\| P^k \| \le \| P^k \| \le q$, тобто для

$$k \ge k_q = (\lg q) / \lg | P \|.$$

4.6. Метод Річардсона

4.6.1. Ідея методу. Метод Річардсона – це явний нестаціонарний ітераційний метод, який одержується із (4.3), коли $B_{k+1} = I$, і набуває вигляду

$$\frac{1}{\tau_{k+1}}\left(x^{(k+1)}-x^{(k)}\right)+Ax^{(k)}=b,\ k=0,1,... \qquad (4.20)$$

Ітераційні параметри τ_{k+1} вибираються таким чином, щоб для заданого числа ітерацій m похибка на m-ій ітерації була найменшою, тобто, щоб на наборі параметрів $\tau_1,\tau_2,\ldots,\tau_m$ досягнути $\min\left\|x^{(m)}-x\right\|$. Ці параметри є коренями алгебраїчного многочлена, введеного П.Л. Чебишевим.

4.6.2. Многочлени Чебишева. Такі многочлени відіграють важливу роль у теорії та практиці застосування числових методів. За їх допомогою можна розв'язувати задачі оптимізації обчислювальних алгоритмів. На проміжку [–1,1] многочлени Чебишева $T_n(x)$ степеня n задаються так [68]:

$$T_0(x)=1,\ T_1(x)=x,\ T_n(x)=2xT_{n-1}(x)-T_{n-2}(x), n=2,3...$$

Для $n=2, 3, 4$ і 5 маємо:

$$T_2(x)=2x^2-1,$$
$$T_3(x)=4x^3-3x,$$
$$T_4(x)=8x^4-8x^2+1,$$
$$T_5(x)=16x^5-20x^3+5x.$$

Графіки функцій, які визначаються многочленами Чебишева, показані на рис. 4.1.

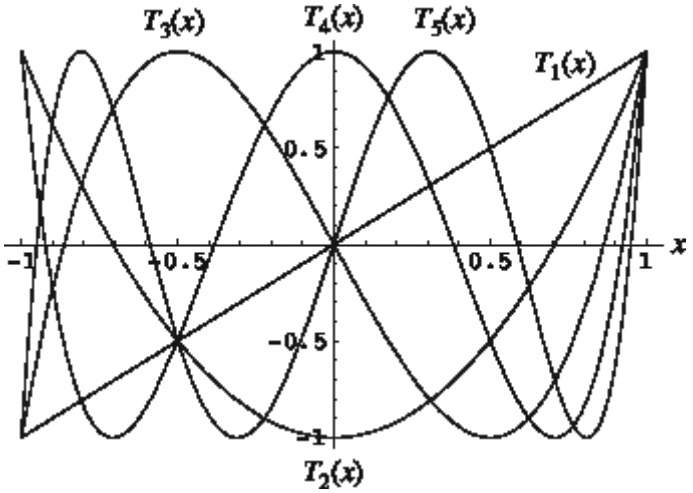


Рис. 4.1. Графіки многочленів Чебишева $T_n(x)$, $n=\overline{1,5}$

Наведемо деякі властивості многочленів Чебишева [59, 68].

1. Якщо $n=2m$, то $T_{2m}(x)$ – парна функція. Для $n=2m+1$ функція $T_{2m+!}$ – непарна. Справді, $T_1(x)$ – непарна, а $T_2(x)$ – парна функції, тому $T_3(x)$ – непарна функція і т. д.

2. Многочлени Чебишева можна записати у вигляді

$$T_n(x) = \cos(n \arccos(x)),\ -1 \le x \le 1.$$

4. Із тригонометричного рівняння $\cos(n \arccos x) = 0$ і періодичності функції $\cos x$ випливає, що нулями многочлена $T_n(x)$ є

$$x_k = \cos \frac{\pi(2k-1)}{2n},\ k = \overline{1,n}.$$

5. Екстремуми функції $T_n(x)$ досягається в точках

$$x_k = \cos \frac{k\pi}{n},\ k = \overline{1,n-1}.$$

де $|T_n(x)| = 1$, причому $T_n(x_k) = \cos k\pi = (-1)^k$.

4.6.3. Алгоритм методу. Даний метод обґрунтований для систем із додатно визначеною матрицею A [1]. Параметри $\tau_{k+1},\ k = \overline{0,m-1}$, у методі Річардсона вибираються так:

$$\tau_0 = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)},\ \rho_0 = \frac{1-\xi}{1+\xi},\ \xi = \frac{\lambda_{\min}}{\lambda_{\max}};$$

$$\tau_k = \frac{\tau_0}{1+\rho_0 t_{k+1}},\quad t_{k+1} = \cos \frac{(2k+1)\pi}{2m},\ k = \overline{0,m-1} \qquad (4.21)$$

Обґрунтування методу Річардсона дається наступною теоремою.

Теорема 4.3 [59, с. 110-111]**.** *Нехай A – симетрична і додатно визначена матриця. Тоді серед методів вигляду (4.20) найменшу похибку на m-ій ітерації має метод, для якого параметри τ_{k+1} визначаються формулами (4.21). Для похибки на m-ій ітерації виконується нерівність*

$$\left\| x^m - x \right\| \le q_m \left\| x^{(0)} - x \right\|, \qquad (4.22)$$

де

$$q_m = \frac{2\rho_1^m}{1+\rho_1^{2m}},\quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}.$$ ■

Зауважимо, що оцінка (4.22) правильна для набору параметрів (4.21) у будь-якому порядку. Але в обчислювальній практиці спостерігається, що порядок вибору параметрів суттєво впливає на числову стійкість методу й може привести до недопустимо сильного зростання обчислювальної похибки. Справа в тому, що метод не гарантує монотонного спадання похибки від ітерації до ітерації. Запишемо рівняння для похибки

$z^{(k)} = x^{(k)} - x$. Для цього $x^{(k)} = z^{(k)} + x$ підставимо в (4.20). Одержимо

$$\frac{1}{\tau_{k+1}}\left(z_{k+1} - z_k\right) + Az^{(k)} + Ax = b.$$

Враховуючи, що $Ax = b$, одержимо $z_{k+1} = (I - \tau_{k+1}A)z_k$. Норма оператора ξ може набути значень, більших від одиниці для декількох сусідніх ітерацій, що й веде до зростання похибки. У [1] наведений алгоритм вибору параметрів τ_{k+1}, для якого ітераційний метод (4.20) стійкий.

4.7. Метод релаксації

Прискорити швидкість збіжності ітераційного процесу можна, ввівши в методі Зейделя деякий параметр ω. Цей метод одержується із (4.3), якщо $B_k = (D + \omega G)$, $\tau_{k+1} = \omega$. Тоді

$$(D + \omega H)\frac{x^{(k+1)} - x^{(k)}}{\omega} + Ax^{(k)} = b, \qquad k = 0,1,... \tag{4.24}$$

Якщо ω є (0,2) і матриця A – симетрична і додатно визначена, то при цьому досягається збіжність методу релаксації [20, 23, 59]. Для ω є (0,1) маємо *метод нижньої релаксації*, а для ω є (1,2) – *верхньої релаксації*. Якщо параметр релаксації ω=1, то одержимо *метод Гауса–Зейделя*.

Щоб вивести формули для обчислення $x_i^{(k+1)}$, запишемо (4.24) у вигляді

$$(I + \omega D^{-1}H)x^{(k+1)} = ((1-\omega)I - \omega D^{-1}G)x^{(k)} + \omega D^{-1}b.$$

Тоді

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

Тут $\sum_{n+1}^{n} = 0$, $\sum_{1}^{0} = 0$. Починаючи із $i=1$, послідовно знаходимо $x_i^{(k+1)}$:

$$x_1^{(k+1)} = (1-\omega)x_1^{(k)} + \frac{\omega}{a_{11}}(b_1 - \sum_{i=2}^{n}a_{1j}x_j^{(k)}),$$

$$x_2^{(k+1)} = -\omega\frac{a_{21}}{a_{11}}x_1^{(k+1)} + (1-\omega)x_2^{(k)} + \frac{\omega}{a_{22}}(b_2 - \sum_{j=3}^{n}a_{2j}x_j^{(k)}),$$

$$x_3^{(k+1)} = -\omega\frac{a_{31}}{a_{33}}x_1^{(k+1)} - \omega\frac{a_{32}}{a_{33}}x_j^{(k+1)} + (1-\omega)x_3^{(k)} + \frac{\omega}{a_{33}}(b_3 - \sum_{j=4}^{n} a_{3j}x_j^{(k)}),...$$

Вибором параметра ω можна прискорити швидкість збіжності методу релаксації. Порівняємо метод Якобі та метод релаксації при різних значеннях параметра ω для деяких систем.

Приклад 4.2. У табл. 4.1 наведено похибки різних варіантів методу релаксації для системи рівнянь із симетричною і додатно визначеною матрицею та перевагою головної діагоналі

$$5x_1 + 2x_2 + x_3 = 8,$$
$$2x_1 + 6x_2 + 3x_3 = 11,$$
$$x_1 + 3x_2 + 7x_3 = 11.$$

Задамо початкові умови $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$. Точний розв'язок цієї системи $x_1 = x_2 = x_3 = 1$. Найліпший результат одержується методом релаксації при $\omega = 1$ (метод Ґауса–Зейделя).

Таблиця 4.1.

Похибки методу релаксації для різних значень параметра ω

$\|x^{(k)} - x\|_\infty$	*Метод Якобі*	*Метод релаксації* ($\omega = 0.1$)	*Метод релаксації* ($\omega = 0.5$)	*Метод Гауса-Зейделя* ($\omega = 1$)	*Метод релаксації* ($\omega = 1.5$)	*Метод релаксації* ($\omega = 1.9$)
k=5	0.16718	0.45994	0.04695	0.0032899	0.06576	0.73146
k=15	0.00337	0.11892	0.0020103	$8.34119*10^{-10}$	0.0001555	0.35622
k=25	0.0000696	0.04313	0.0000732	0	$3.17787*10^{-7}$	0.14785
k=35	0.0000014	0.02192	0.0000026	0	$6.22769*10^{-10}$	0.05088

Приклад 4.3. Для СЛАР

$$5x_1 + 2x_2 + x_3 = 8,$$
$$2x_1 + 6x_2 + 3x_3 = 11,$$
$$x_1 + 3x_2 - 7x_3 = -3$$

матриця A – не додатно визначена, оскільки $\det A = -221$, але головна діагональ переважаюча. Метод релаксації при $\omega > 1$ – розбіжний, швидша збіжність спостерігається для $\omega \in (0,1)$. Найвищу точність дає метод Якобі та Ґауса–Зейделя. Для початкових умов $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$ результати наведені в табл. 4.2:

Таблиця 4.2

Похибка $\left\|x^{(k)} - x\right\|_\infty$	*Метод Якобі*	*Метод релаксації* $(\omega = 0.1)$	*Метод релаксації* $(\omega = 0.5)$	*Метод Гаусса-Зейделя* $(\omega = 1)$	*Метод релаксації* $(\omega = 1.5)$	*Метод релаксації* $(\omega = 1.9)$
k =5	0.02471	0.718	0.0465	0.00144	1.12472	8.50322
k =15	$8.67996*10^{-6}$	0.27071	$4.36073*10^{-4}$	$3.0066*10^{-10}$	1.18692	432.54701
k =25	$2.83705*10^{-10}$	0.07872	$3.34329*10^{-6}$	0	1.14274	$2.31007*10^{4}$
k=35	$8.66862*10^{-13}$	0.04348	$2.66507*10^{-8}$	0	1.10021	$1.23388*10^{6}$

4.8. Методи варіаційного типу

У цих методах розв'язування СЛАР замінюється деякою рівносильною оптимізаційною задачею [5, 12, 59, 77].

4.8.1. Метод найшвидшого спуску. Цей метод породжений задачею мінімізації функціонала, який є квадратичною формою вигляду

$$F(x) = \frac{1}{2}(Ax, x) - (b, x). \qquad (4.25)$$

Нехай x^* – розв'язок системи (4.1) з симетричною додатною матрицею A. Запишемо функцію (4.25) у такому вигляді:

$$F(x) = F(x^*) + \frac{1}{2}(A(x - x^*), x - x^*).$$

Оскільки $(Au,u) \geq 0 \quad \forall u \in R^n$, то функція $F(x)$ досягає найменшого значення в точці x^*. Навпаки, якщо для $x = \bar{x}$ досягається найменше значення $F(x)$, то з необхідної умови екстремуму F випливає, що $gradF(x) = 0$. Оскільки

$$gradF(x) = Ax - b,$$

то $\bar{x} = x^*$ є розв'язком системи (4.1).

У методі найшвидшого спуску рух із деякої точки $x^{(k)}$ здійснюється в напрямі антиградієнта. Наступне наближення

$$x^{(k+1)} = x^{(k)} - \alpha_k gradF(x^{(k)}), \qquad k = 0,1,..., \qquad (4.26)$$

де стала α_k визначається з умови мінімуму функції

$$g(\alpha) = F(x^{(k)} - \alpha \cdot gradF(x^{(k)})).$$

З необхідної умови екстремуму маємо

$$\frac{dg}{d\alpha} = (b - Ax^{(k)}, b - Ax^{(k)}) - \alpha(A(b - Ax^{(k)}), b - Ax^{(k)}) = 0$$

маємо:

$$\alpha_k = \frac{(r_k, r_k)}{(Ar_k, r_k)}, \qquad (4.27)$$

де $r_k = b - Ax^{(k)}$ − нев'язка.

Алгоритм методу найшвидшого спуску:

1. Задати початкове наближення $x^{(0)}$, $k := 0$.
2. Обчислити α_k за формулою (4.27).
3. Знайти наступне наближення згідно з (4.26).
4. Якщо $\alpha_k \|r_k\| \le \varepsilon$, де ε характеризує точність, то ітерації завершені і $x^* \approx x^{(k+1)}$. Якщо число ітерацій k не перевищує допустимого значення $k_{\max}$, то $k := k+1$ і повернутись до пункту 2.

Якщо відомі границі для власних значень матриці A: $\mu_1 \le \lambda(A) \le \mu_2$, то, як показано в [59, с. 242], для похибки розв'язку виконується нерівність

$$\left\|x^{(k)} - x^*\right\|_2 \le \left(\frac{\mu_2 - \mu_1}{\mu_2 + \mu_1}\right)^k \sqrt{\frac{\mu_2}{\mu_1}} \left\|x^{(0)} - x^*\right\|_2 .$$

Метод найшвидшого спуску нелінійним, оскільки на кожній ітерації параметр α_k залежить від одержаного наближення. Недоліком методу є те, що на кожній ітерації потрібно двічі виконувати операції множення матриці на вектор (знаходження r_k і α_k). Економніше записати ітераційний алгоритм у вигляді (4.26), (4.27), але обчислювати r_{k+1} за рекурентною формулою

$$r^{(k+1)} = r^{(k)} - \alpha_k A \cdot gradF(x^{(k)}).$$

4.8.2. Метод мінімальних нев'язок. *У* методі Річардсона потрібно знати власні значення $\lambda_{\min}(A)$ і $\lambda_{\max}(A)$ матриці A. В методі мінімальних нев'язок

$$\frac{x^{(k+1)} - x^{(k)}}{\tau_{k+1}} + Ax^{(k)} = b, \; k = 0,1,...,$$

де параметр τ_{k+1} на кожній ітерації вибирається так, щоб мінімізувати евклідову норму нев'язки $r^{(k+1)} = b - Ax^{(k+1)}$ наближення $x^{(k+1)}$. Якщо параметр τ_{k+1} знайдено, то

$$x^{(k+1)} = x^{(k)} + \tau_{k+1} r^{(k)}.$$

Далі маємо

$$r^{(k+1)} - r^{(k)} = -Ax^{(k+1)} + Ax^{(k)} = b - Ax^{(k+1)} - b + Ax^{(k)} =$$
$$= Ax^{(k)} - Ax^{(k+1)} = A(x^{(k)} - x^{(k)} - \tau_{k+1} r^{(k)}) = -\tau_{k+1} Ar^{(k)}.$$

Звідси випливає, що $r^{(k+1)} = r^{(k)} - \tau_{k+1} Ar^{(k)}$.

Розглянемо скалярний добуток

$$(r^{(k+1)}, r^{(k+1)}) = (r^{(k)}, r^{(k)}) - 2\tau_{k+1}(Ar^{(k)}, r^{(k)}) + \tau_{k+1}^2 (Ar^{(k)}, Ar^{(k)}),$$

або $\left\| r^{(k+1)} \right\|^2 = \left\| r^{(k)} \right\|^2 - 2\tau_{k+1}(Ar^{(k)}, r^{(k)}) + \tau_{k+1}^2 \left\| Ar^{(k)} \right\|^2$.

Оскільки A – додатно визначена матриця, то $(Ar^{(k)}, r^{(k)}) \geq 0$, і квадратична функція змінної τ_{k+1} набуває найменшого значення, коли

$$\tau_{k+1} = \frac{(Ar^{(k)}, r^{(k)})}{\left\| Ar^{(k)} \right\|^2}.$$

Теорема 4.4 [59]. *Нехай* $A = A^T > 0$. *Тоді метод мінімальних нев'язок збіжний і для похибки справджується оцінка*

$$\left\| A(x^{(k)} - x^*) \right\| \leq \rho_0^k \left\| A(x^{(0)} - x^*) \right\|, \qquad k = 0,1,...,$$

де $\rho_0 = \dfrac{1-\xi}{1+\xi}, \qquad \xi = \dfrac{\lambda_{\min}(A)}{\lambda_{\max}(A)}.$ ■

Метод мінімальних ітерацій має ту ж швидкість збіжності, що й метод простої ітерації з параметром $\tau = 2/(\lambda_{\max}(A) + \lambda_{\min}(A))$.

4.8.3. Метод спряжених градієнтів. Цей метод застосовується для розв'язування СЛАР із симетричною додатно визначеною матрицею A. Оптимізується градієнтний метод вибором параметра так, щоб на наступному кроці нев'язка була ортогональною до всіх попередніх нев'язок. На першому кроці нев'язка $\xi_0 = r_0$, як і в МНС. Одержані нев'язки утворюють ортогональний базис. Теоретично на останньому кроці нев'язка дорівнює нулю й ітераційний процес завершується. Але внаслідок похибок в обчисленнях, особливо при поганій обумовленості матриці, норму нев'язки потрібно порівнювати з допустимим рівнем абсолютної похибки $\varepsilon > 0$.

Алгоритм методу спряжених градієнтів

1. Задати початкове наближення $x^{(0)}$ і число $\varepsilon > 0$ (допустимий рівень абсолютної похибки нев'язки).
2. Обчислити нев'язку $r^{(0)} = b - Ax^{(0)}$ початкового наближення $x^{(0)}$.
3. k=0 (номер ітерації), $p^{(0)} = r^{(0)}$.

Обчислити: $q^{(k)} = Ap^{(k)}$; множник на кроці $\alpha_k = \dfrac{(r^{(k)}, p^{(k)})}{(q^{(k)}, p^{(k)})}$;

наступне наближення $x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$;

нев'язку наступного наближення $r^{(k+1)} = r^{(k)} - \alpha_k q^{(k)}$.

4. **Якщо** $\left\| r^{(k+1)} \right\| \le \varepsilon$, **то**:

4.1. $x := x^{(k+1)}$ (розв'язок задовольняє СЛАР з точністю ε).

інакше:

4.2. Обчислити коефіцієнт $\beta_k = \dfrac{(r^{(k+1)}, q^{(k)})}{(p^{(k)}, q^{(k)})}$,

новий напрям мінімізації $p^{(k+1)} = r^{(k+1)} - \beta_k p^{(k)}$,

$k := k + 1$ і перейти до кроку 4.

Отже, в цьому методі система векторів, яка підлягає процесу ортогоналізації, не задається заздалегідь, а будується паралельно з побудовою спряжених напрямків і відповідних їм послідовних наближень. Зауважимо, що назва методу пов'язана з тією обставиною, що нев'язка r_i є градієнтом функції похибок, обчисленим у точці *x*. Процес побудови послідовних наближень обірветься, як тільки деяке наближення збігається з точним розв'язком системи.

Обчислювальна складність алгоритму після виконання *n* ітерацій складає $S_{МСГ} = 2n^3 + 13n^2$.

4.9. Висновки

1. Ефективне розв'язання СЛАР значно залежить від вибору ітераційного методу. Проте, щоб одержати оптимальніший результат, потрібно брати до уваги обчислювальні аспекти методу. В прямих методах вихідну матрицю *A* часто потрібно зображати у вигляді добутку матриць, алгоритми ж ітераційних методів не вимагають цього. До того ж, основні дії в ітераційних методах часто використовують непряму адресацію, залежно від структури даних. Такі операції мають відносно низьку ефективність виконання. Проте це не впливає на загальний час розв'язання системи. До того ж, ітераційні методи зазвичай простіше реалізувати, ніж прямі методи. Крім того, їх можна застосовувати для ширшого класу систем.

При виборі методу потрібно взяти до уваги властивості матриці *A*. Не кожним методом можна реалізувати розв'язування

будь-якої СЛАР, тому знання властивостей матриці – головний критерій для вибору ітераційного методу. Також потрібно врахувати потужність комп'ютера, можливість розпаралелювання алгоритму та реалізацію методу на багатопроцесорній системі [51].

Зробимо висновки про переваги і недоліки деяких ітераційних методів.

1. *Метод простих ітерацій.* Простий для реалізації, володіє лінійною швидкістю збіжності, якщо норма ітераційної матриці менша одиниці або головна діагональ матриці переважаюча.

2. *Метод Зейделя.* Збіжність може бути швидшою, ніж у методі простих ітерацій. Збіжний для симетричних додатно визначених матриць або матриць із перевагою головної діагоналію і є частковим випадком методу верхньої релаксації ($\omega = 1$). Якщо A – матриця зі строгою перевагою по рядках, то методи Якобі і Ґауса-Зейделя збіжні, причому другий метод збігається швидше.

3. *Метод релаксації.* Дозволяє прискорити збіжність методу Зейделя (ω>1, верхня релаксація). Метод верхньої релаксації може бути збіжним, тоді як метод Зейделя або метод нижньої релаксації застосувати не вдається. Швидкість збіжності залежить від ω. При певних умовах оптимальне значення для ω може бути оцінене спектральним радіусом ітераційної матриці Якобі. У цьому методі нерівність $0 < \omega < 2$ необхідна для збіжності. Якщо ж $A^T = A > 0$, то ця нерівність і достатня для збіжності.

4. *Метод спряжених градієнтів.* Цей метод найліпший для систем з матрицею $A^T = A > 0$. При його реалізації передбачено одночасне зберігання в пам'яті тільки чотирьох векторів. Крім того, в його внутрішньому циклі використовується тільки матрично-векторний добуток, два скалярних добутки, три операції додавання одного вектора до іншого і невелика кількість скалярних операцій. Тобто і пам'ять, і обчислення в методі дуже незначні. Якщо власні значення матриці A такі, що $\lambda_{\max} / \lambda_{\min}$ досить велике, то швидкість збіжності може бути надлінійною. Варіанти методу спряжених градієнтів та їх аналіз наведено в [20, 23, 97, 100].

5. *Метод мінімальних нев'язок.* Застосовується і до систем з несиметричними матрицями. У цьому методі найменша нев'язка

для фіксованої кількості ітерацій, але їх громіздкість зростає з кожною ітерацією.

6. *Вплив похибки заокруглення на ітераційний процес.* Сумарну похибку заокруглення при виконанні однієї ітерації можна розглядати як збурення $\varepsilon^{(k)}$ у правій частині ітераційного процесу (4.4), тобто

$$\bar{x}^{(k+1)} = P\bar{x}^{(k)} + f + \varepsilon^{(k)}. \qquad (4.28)$$

Враховуючи, що $q := \|P\| < 1$, із (4.4) і (4.28) одержимо

$$\left\|\bar{x}^{(k+1)} - x^{(k+1)}\right\| \le q\left\|\bar{x}^{(k)} - x^{(k)}\right\| + \varepsilon^{(k)} \le q^2\left\|\bar{x}^{(k-1)} - x^{(k-1)}\right\| + q\left\|\varepsilon^{(k-1)}\right\| + \left\|\varepsilon^{(k)}\right\|.$$

Нехай $\left\|\varepsilon^i\right\| \le \varepsilon \ \ \forall i$. Оскільки $\bar{x}^{(0)} - x^{(0)} = 0$, то

$$\left\|\bar{x}^{(k+1)} - x^{(k+1)}\right\| \le \frac{\varepsilon\left(q^{k+1} - 1\right)}{q-1} \le \frac{\varepsilon}{q-1}.$$

Отже, $\varepsilon/(q-1)$ – похибка заокруглення внаслідок скінченної розрядності зображення чисел з плаваючою крапкою. Похибка не залежить від кількості ітерацій і характеризує стійкість ітераційного процесу по відношенню до похибок заокруглення.

Приклади розв'язування типових задач

Задача 1. Проілюструємо методи Якобі та Ґауса-Зейделя на прикладі системи лінійних рівнянь

$$4x_1 + x_2 + 9x_3 = 1,$$
$$3x_1 + 8x_2 - 7x_3 = 2,$$
$$x_1 + x_2 - 8x_3 = 3.$$

Розв'язування. Перетворимо систему рівнянь так, щоб виконувалася достатня умова збіжності. Для цього до першого рівняння додамо третє, а потім від другого віднімемо третє. Одержимо систему, в якій досягається перевага діагональних елементів

$$5x_1 + 2x_2 + x_3 = 4,$$
$$2x_1 + 7x_2 + x_3 = -1,$$
$$x_1 + x_2 - 8x_3 = 3.$$

Метод Якобі і Ґауса-Зейделя набувають відповідно вигляду

$$x_1^{(k+1)} = (2x_2^{(k)} + x_3^{(k)} + 4)/5, \qquad x_1^{(k+1)} = (2x_2^{(k)} + x_3^{(k)} + 4)/5,$$
$$x_2^{(k+1)} = (2x_1^{(k)} + x_3^{(k)} - 1)/7, \quad \text{і} \qquad x_2^{(k+1)} = (2x_1^{(k+1)} + x_3^{(k)} - 1)/7,$$
$$x_3^{(k+1)} = (x_1^{(k)} + x_2^{(k)} + 3)/(-8). \qquad x_3^{(k+1)} = (x_1^{(k+1)} + x_2^{(k+1)} + 3)/(-8).$$

Задача 2. Знайти три наближення методом Річардсона для системи лінійних рівнянь

$$2x_1 + x_2 = 3,$$
$$x_1 + 4x_2 = 5.$$

Розв'язування. Задамо початкові умови $x_1^{(0)} = x_2^{(0)} = 0$. Точний розв'язок СЛАР $x_1 = x_2 = 1$, власні значення матриці $\lambda_{\min}(A) = 3 - \sqrt{2}$, $\lambda_{\max}(A) = 3 + \sqrt{2}$, $\tau_0 = 1/3$. Матриця $A > 0$, оскільки головні мінори додатні. На третій ітерації ($m = 3$) знаходимо нулі $t_1 = \dfrac{\sqrt{3}}{2}$ $t_2 = 0$, $t_3 = -\dfrac{\sqrt{3}}{2}$ многочлена $4x^3 - 3x$. Тоді параметри $\tau_1 = \dfrac{6 - \sqrt{6}}{15}$, $\tau_2 = \dfrac{1}{3}$, $\tau_3 = \dfrac{6 + \sqrt{6}}{15}$. Результати обчислень згідно з методом Річардсона, а також методами Якобі та простої ітерації наведено в табл. 4.3.

Таблиця 4.3

Номер ітерації	*Метод Якобі*	*Метод Річардсона*
1	1.5000 1.1250	0.7101 1.1835
2	0.9375 0.7500	0.8422 1.0055
3	1.1250 0.8906	1.0000 1.0444
Похибка $\left\|x^{(3)} - x\right\|_\infty$	0.1250	0.0444

Похибка в методі Річардсона тут найменша і дорівнює 0.044.

Задача 3. Нехай у методі простої ітерації

$$\frac{x^{(k+1)} - x^{(k)}}{\tau} + Ax^{(k)} = b$$

з параметром $\tau = \left\|A\right\|^{-1}$ власні значення матриці A дійсні і $\lambda(A) > 0$. Довести, що метод збіжний.

Розв'язування. Маємо $\lambda^{(k+1)} = (I - \tau A)x^{(k)}$. Покажемо, що власні значення матриці $P = I - \tau A$ належать $[0, 1)$, тобто виконується умова збіжності теореми 4.1. Оскільки $P - \lambda(P) = -\tau(A - (\lambda(B) - 1)\tau^{-1} I)$, то $\lambda(P) = 1 - \tau\lambda(A)$. Із умови $\lambda(A) > 0$ і нерівності $\lambda(A) \le \|A\|$ випливає, що $\lambda(P) \ge 1 - \|A\|^{-1}\|A\| = 0$ і $\lambda(P) = 1 - \tau\lambda(A) < 1$, тобто $0 \le \lambda(P) < 1$ і метод збіжний.

Завдання та запитання для самостійної роботи

1. Записати в канонічній формі ітераційні методи Зейделя, Ґауса–Зейделя та релаксації.
2. Що таке нев'язка розв'язку СЛАР і як вона характеризує точність ітераційного методу?
3. Проілюструвати геометричну інтерпретацію методу Зейделя для системи двох лінійних рівнянь.
4. Навести алгоритм методу релаксації та сформулювати умови збіжності.
5. Розкрити суть методів варіаційного типу, мінімальних напрямків та спряжених градієнтів.
6. Які умови збіжності методу простої ітерації? Оцінка похибки розв'язку та оцінка знизу кількості ітерацій для досягнення заданої точності.
7. У чому суть методу найшвидшого спуску? Умови застосування, збіжність, геометрична інтерпретація.
8. Довести, що для СЛАР другого порядку методи Якобі і Ґауса–Зейделя збіжні та розбіжні одночасно.
9. Проілюструвати в системі координат *xOy* збіжність і розбіжність методів Якобі і Ґауса–Зейделя вдповідно для СЛАР
$$\begin{aligned} 2x_1 + x_2 &= 2, \\ x_1 - 2x_2 &= -2 \end{aligned} \quad \text{і} \quad \begin{aligned} x_1 - 2x_2 &= -2, \\ 2x_1 + x_2 &= 2. \end{aligned}$$
10. Перевірити, чи виконуються необхідні і достатні умови збіжності методів Якобі і Ґауса–Зейделя для системи рівнянь
$$\begin{aligned} x_1 + x_2 &= 2, \\ x_1 + 2x_2 + x_3 &= 4, \\ x_2 + 2x_3 &= 3. \end{aligned}$$
11. Виписати формули та дослідити умови збіжності ітераційних методів Якобі, Зейделя, верхньої релаксації для системи рівнянь

$$2x_1 + x_2 = 1,$$
$$x_1 + 2x_2 = -1.$$

12. Знайти необхідні і достатні умови збіжності ітераційних методів Якобі та Зейделя для системи рівнянь із матрицею A вигляду

$$A = \begin{pmatrix} a & b & 0 \\ b & a & b \\ 0 & b & a \end{pmatrix}.$$

13. Нехай $\sum_{j \neq i} |a_{ij}| \leq q|a_{ii}|$, $0 < q < 1$ для всіх i. Одержати оцінку

$$\left\| x^{(k)} - x^* \right\| \leq q^k \left\| x^{(0)} - x^* \right\|.$$

14. Записати алгоритми методів простої ітерації та Зейделя для системи лінійних рівнянь, звівши їх спочатку до вигляду, що забезпечує збіжність:

$$1)\ A_1 = \begin{bmatrix} 2.8 & 10.2 & 3.4 \\ 3.6 & 1.9 & 3.1 \\ 4.2 & 5.9 & 1.7 \end{bmatrix}, \quad b_1 = \begin{bmatrix} 2.5 \\ 3.6 \\ 4.5 \end{bmatrix};$$

$$2)\ A_2 = \begin{bmatrix} 1.6 & -2.9 & 1.8 \\ 2.2 & 3.5 & 1.9 \\ 4.5 & -1.5 & 1.1 \end{bmatrix}, \quad b_2 = \begin{bmatrix} 3.6 \\ 4.7 \\ 5.5 \end{bmatrix}.$$

15. Для систем лінійних рівнянь:

$$1)\ \begin{aligned} x_1 + 2x_2 &= 3, \\ 3x_1 + 4x_2 &= 7; \end{aligned} \qquad 2)\ \begin{aligned} 2x_1 + x_2 &= 4, \\ 3x_1 - 2x_2 &= 1 \end{aligned}$$

виконати першу і другу симетризацію Ґауса.

16. Для систем рівнянь:

$$1)\ \begin{aligned} x_1 + 2x_2 &= 3, \\ 2x_1 + 5x_2 &= -3; \end{aligned} \qquad 2)\ \begin{aligned} 2x_1 - x_2 &= 1, \\ -x_1 + x_2 &= 0 \end{aligned}$$

застосувати метод спряжених градієнтів, метод найшвидшого спуску і метод мінімальних ітерацій. Проаналізувати результати.

17. При яких α і β методи Якобі і Ґауса–Зейделя для систем із матрицею $A = \begin{bmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{bmatrix}$ збігаються?

18. Для яких значень параметра τ метод $x^{(k+1)} = (I - \tau A)x^{(k)} + \tau b$ для системи рівнянь $Ax = b$ з матрицею

$$1)\ A = \begin{bmatrix} 5 & 0.8 & 4 \\ 2.5 & 2 & 0 \\ 2 & 0.8 & 4 \end{bmatrix};\ 2)\ A = \begin{bmatrix} 2 & 1 & 0.5 \\ 3 & 5 & 1 \\ 1 & 3 & 3 \end{bmatrix}$$

збігається для довільного початкового наближення?

19. Одержати оцінку швидкості збіжності методу найшвидшого спуску вигляду $\left\|x^{(k)} - x^*\right\|_2 \le (1 - \frac{\lambda_{\min}}{\lambda_{\max}})\left\|x^{(0)} - x^*\right\|_2$, де $\lambda_{\min}$ і $\lambda_{\max}$ – найменше і найбільше власні значення матриці $A > 0$.

20. Для лінійної системи рівнянь

$$\begin{aligned} x_1 + \quad & \quad x_3 = 2, \\ -x_1 + x_2 \quad & \quad = 0, \\ x_1 + 2x_2 - 3x_3 &= 0 \end{aligned}$$

і кількох початкових значень показати, що ітерації Якобі збіжний, а Гауса–Зейделя – розбіжними.

21. Задати СЛАР порядку n з матрицею A, діагональні елементи якої дорівнюють $2n$, а інші елементи в кожному рядку – випадкові числа $a_{ij} \in (0,1), i \ne j.$ Елементи вектора правої частини $b_i = \sum_{j=1}^{n} a_{ij}, i = \overline{1,n}.$ Випробувати для такої СЛАР методи Якобі та Ґауса–Зейделя з початковим значенням $x^{(0)} = b$ і різними, досить великим, значенням n.

22. Проілюструвати ітераційні методи Якобі, Ґауса–Зейделя і релаксації зі значеннями параметра $\omega = 0.90$ і 1.10 для СЛАР

$$\begin{bmatrix} 10 & 2 & -1 & 3 & 1 \\ 2 & 10 & 2 & -1 & 3 \\ -1 & 2 & 10 & 2 & -1 \\ 3 & -1 & 2 & 10 & 2 \\ 1 & 3 & -1 & 2 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 5 \\ -2 \\ 4 \\ -2 \\ 5 \end{bmatrix}.$$

23. Потрібно розв’язати методом Гауса–Зейделя СЛАР [56]

$$\begin{aligned} 2x_1 + x_2 - 0.5x_3 + 8x_4 &= 4, \\ 6x_1 + 2x_2 + 8x_3 + 14x_4 &= 12, \\ x_1 - 6x_2 + 10x_3 + 9x_4 &= -1, \\ 2x_1 + 11x_2 + 3x_3 - 4x_4 &= 4. \end{aligned}$$

Які дії потрібно виконати спочатку? Узяти за початкове наближення нульовий вектор й оцінити скільки ітерацій потрібно зробити, щоб отримати п'ять правильних цифр у розв'язку.

24. Побудувати алгоритм розв'язування СЛАР

$$\begin{bmatrix} A_1 & B_2 \\ B_1 & A_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} c \\ d \end{bmatrix},$$

блочним методом Якобі. Тут A_1, A_2 – квадратні матриці порядку n_1 і n_2 відповідно, x, c і y, d – вектори розміру n і m.

25. Проаналізувати збіжність методів Якобі і Гауса–Зейделя для СЛАР із матрицею Гільберта $a_{ij} = (i + j - 1)^{-1}, i, j = \overline{1,n},$ порядку 3, 4 і 5.

26. Побудувати СЛАР, для якої метод Гауса–Зейделя збіжний, а метод Якобі розбіжний.

27. Показати, що метод Гауса–Зейделя збіжний для СЛАР із матрицею без переваги головної діагоналі [56]

$$A = \begin{bmatrix} 8 & 2 & 1 \\ 10 & 4 & 1 \\ 50 & 25 & 2 \end{bmatrix}.$$

Чи буде збіжним метод Якобі?

28. Узагальнити метод Зейделя, приєднавши головну діагональ. Як зміниться метод, якщо приєднати одну або більше діагоналей над головною?

29. У методі Річардсона вибрати параметри $\tau_1, ..., \tau_m$, так, щоб похибка методу монотонно спадала з ростом номера ітерації[1].

30. Застосувати метод Гауса та метод релаксації з різними значеннями параметра ω для знаходження розв'язку СЛАР

$$10x_1 + 0.01x_2 = 20.05,$$
$$0.01x_1 + 0.001x_2 + 0.1x_1 = 0{,}125,$$
$$0.1x_1 + 1000x_3 = 1000.5.$$

$\lambda_1 = 0.000979999, \lambda_2 = 10.00001, \lambda_3 = 1000.00001,$ точний розв'язок $x = [2,5,1].$

[1] Самарский А.А. Введение в теорию разностных схем. – М.: Наука, 1971. – 552 с.

Розділ 5. Ітераційні методи розв'язування нелінійних рівнянь

Відокремлення коренів нелінійних рівнянь. Швидкість збіжності ітераційного методу. Метод половинного поділу та лінійної інтерполяції. Методи простої ітерації. Метод Ньютона: алгоритм, збіжність, оцінка похибки, графічна ілюстрація. Випадок кратних коренів. Модифікації методу Ньютона. Метод січних та швидкість його збіжності. Комбіновані ітераційні методи.

Література [5, 13, 25, 28, 43, 45, 49, 50, 70, 73, 83]
Електронні джерела [103–108]

5.1. Приклади нелінійних рівнянь

Одним із перших трансцендентних рівнянь, для якого будувались ітераційні наближення, було рівняння Кеплера (1619 р.) із визначення ексцентричної аномалії E еліптичної орбіти планети

$$E - \varepsilon \sin E = M,$$

де ε – ексцентриситет, $0 < \varepsilon < 1$, M – середня аномалія (рис. 5.1).

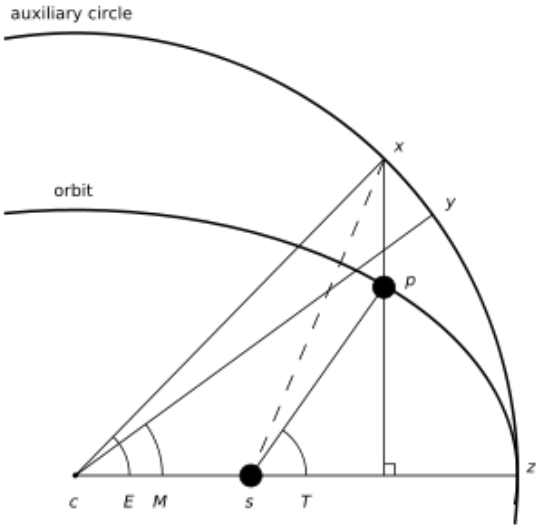


Рис. 5.1. Ексцентрична та середня аномалії еліптичної орбіти (uk.wikipedia.org/wiki/Елементи_орбіти)

Модель динаміки популяції з чисельністю $N(t)$ у момент часу t і міграцією υ описується диференціальним рівнянням

$$\frac{dN}{dt} = \lambda N(t) + \upsilon, \quad t > 0,$$

де λ – коефіцієнт зростання. Розв'язок рівняння з початковою умовою $N(0) = N_0$ і постійною міграцією ν має вигляд

$$N(t) = N_0 e^{\lambda t} + \frac{\upsilon}{\lambda}(e^{\lambda t} - 1).$$

Нехай для деякої популяції $N_0 = 10$, $\upsilon = 7$ і $N(1) = 15$. Тоді для знаходження коефіцієнта зростання λ одержимо нелінійне рівняння

$$15 = 10\lambda e^{\lambda} + 7(e^{\lambda} - 1).$$

При моделюванні епідемій важливо знати долю $z \in (0,1)$ схильного до інфікування населення. Для моделі Кермака–МакКендрика поширення епідемії [42]

$$\frac{dS}{dt} = -\beta SI, \frac{dI}{dt} = \beta SI - \gamma I, \frac{dR}{dt} = \gamma I,$$
$$S(t) + I(t) + R(t) = N,$$

де $S(t)$ – кількість схильного до інфікування населення із постійної кількості N, $I(t)$ – інфікованого, а $R(t)$ – тих, хто одужали або померли, величина z є розв'язком нелінійного рівняння

$$z = \exp(R_0(z-1)).$$

Розв'язок $z \in (0,1)$ існує, якщо основне репродуктивне число $R_0 > 1$, тобто коли має місце епідемія.

До розв'язування нелінійних рівнянь зводяться також важливі математичні задачі. Наприклад, абсциси точок екстремуму диференційовної функції $y = g(x)$ знаходяться з рівняння

$$\frac{dg}{dx} = 0.$$

Ще одним прикладом є задача про знаходження стаціонарних розв'язків автономного диференціального рівняння $\frac{du}{dt} = f(x)$, які у скалярному випадку знаходяться із рівняння $f(x) = 0$. Для системи диференціальних рівнянь маємо відповідну систему нелінійних рівнянь.

Математичні моделі процесів, швидкість зміни яких залежить від стану процесу як у момент часу t, так і в попередній момент $t - \Delta$, $\Delta > 0$, у лінійному випадку описуються диференціальним рівнянням із запізненням аргументу

$$\dot{u}(t) = au(t) + bu(t - \Delta),$$

де a і $b \neq 0$ – деякі сталі. Якщо шукати розв'язок рівняння у вигляді $u(t) = e^{\lambda t}$, то для параметра λ одержимо нелінійне рівняння $\lambda = a + be^{-\lambda\Delta}$, яке має нескінченну множину комплексних коренів.

5.2. Відокремлення коренів

Розглянемо скалярне рівняння

$$f(x) = 0, \tag{5.1}$$

де f – неперервна функція, визначена на скінченному або нескінченному інтервалі.

Означення 5.1. *Корінь* x^* *рівняння (5.1) або нуль функції* f *має кратність* $m > 0$*, якщо*

$$f(x) = (x - x^*)^m g(x),$$

де $\lim\limits_{x \to x^*} g(x) \neq 0$. *Якщо* $m = 1$*, то корінь називається простим.*

Прості та кратні корені проілюстровано на рис. 5.1. Наприклад, для рівняння $(x-1)\sqrt[3]{(x-2)^2} = 0$ корінь $x_1 = 1$ – простий, а $x_2 = 2$ має кратність $2/3$. Для рівняння $1 - \cos 4x = 0$ корінь $x = 0$ має кратність 2. Справді,

$$1 - \cos 4x = x^2 \frac{1 - \cos 4x}{x^2}, \quad \lim_{x \to 0} \frac{1 - \cos 4x}{x^2} = \lim_{x \to 0} \frac{2\sin^2 2x}{x^2} = 8.$$

Нехай f $f \in C^m(a,b)$, $x^* \in (a,b)$ і

$$f(x^*) = f'(x^*) = \ldots = f^{(m-1)}(x^*) = 0, \; f^{(m)}(x^*) \neq 0.$$

Тоді з розкладу функції f за формулою Тейлора у точці x з малого околу x^* маємо:

$$f(x) = f(x^*) + (x - x^*) f'(x^*) + \frac{(x - x^*)^2}{2} f(x^*) + \ldots \tag{5.2}$$

$$+ \frac{(x - x^*)^{m-1}}{(m-1)!} f^{(m-1)}(x^*) + \frac{(x - x^*)^m}{m!} f^{(m)}(\xi) = \frac{(x - x^*)^m}{m!} f^{(m)}(\xi), \; \xi \in (a, b).$$

Із (5.2) випливає, що корінь x^* має кратність m, оскільки $f^{(m)}(\xi(x)) \neq 0$ для близьких до x^* значень аргументу.

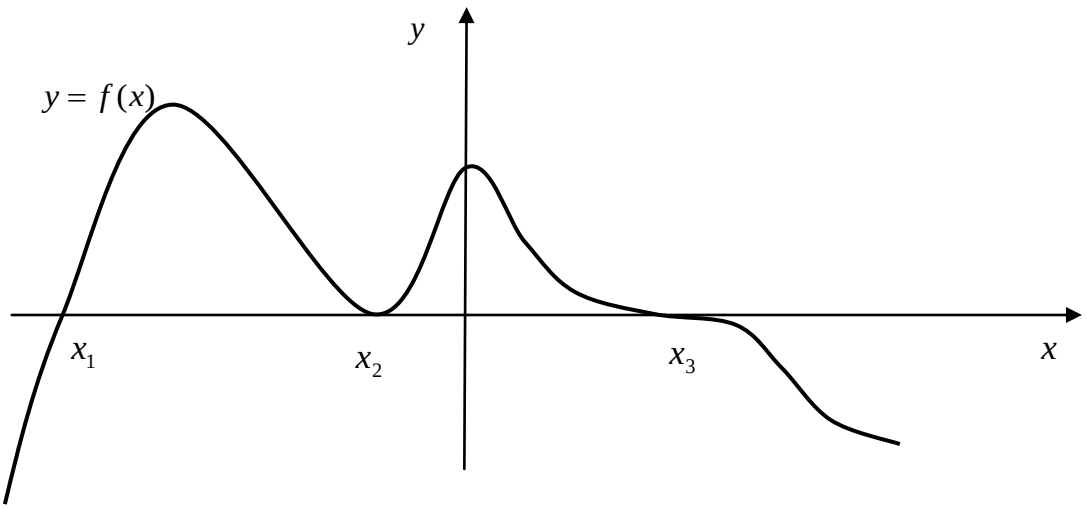


Рис. 5.2. Корінь x_1 – простий, x_2 –парної, а x_3 – непарної кратності

Наближене обчислення коренів рівняння (5.1) передбачає такі етапи: відокремлення коренів, вибір початкового (нульового) наближення для відокремленого кореня та уточнення кореня деяким ітераційним методом. Відокремлення коренів полягає у знаходженні таких інтервалів (a_v, b_v), на кожному з яких існує єдиний корінь рівняння (5.1) і ці інтервали попарно не перетинаються.

Якщо нескладно побудувати графік функції $y = f(x)$, то можна одержати інформацію про кількість і розміщення коренів. Деколи задачу можна спростити, записавши рівняння (5.1) у вигляді

$$f_1(x) = f_2(x),$$

де f_1 і f_2 такі функції, графіки яких просто побудувати. Тоді задача відокремлення коренів зводиться до знаходження кількості точок перетину цих графіків і виділенні на осі абсцис тих проміжків, яким належать проекції точок перетину.

Перевірити, що на проміжку $[a, b]$ є корінь рівняння (5.1), де f – неперервна функція, можна на підставі теореми Больцано–Коші. Якщо

$$f(a) f(b) < 0, \tag{5.3}$$

то на інтервалі (a, b) є хоча б один корінь рівняння (5.1). Ця теорема не дає відповіді на запитання про кількість коренів рівняння на цьому інтервалі. Крім того, для коренів з парною кратністю умова (1.3) не виконується (корінь x_2 на рис. 5.1).

Якщо на цьому інтервалі функція f монотонна і виконується умова (5.3), то на (a, b) існує єдиний корінь.

Приклад 5.1. Розглянемо рівняння $x^2 - \cos x - 1 = 0$, яке запишемо у вигляді $x^2 - 1 = \cos x$. Графіки лівих і правих частин цього рівняння показано на рис. 5.3. Корені локалізовані на інтервалах $(-1.5, -1.5)$ $(1.0, -1.5)$

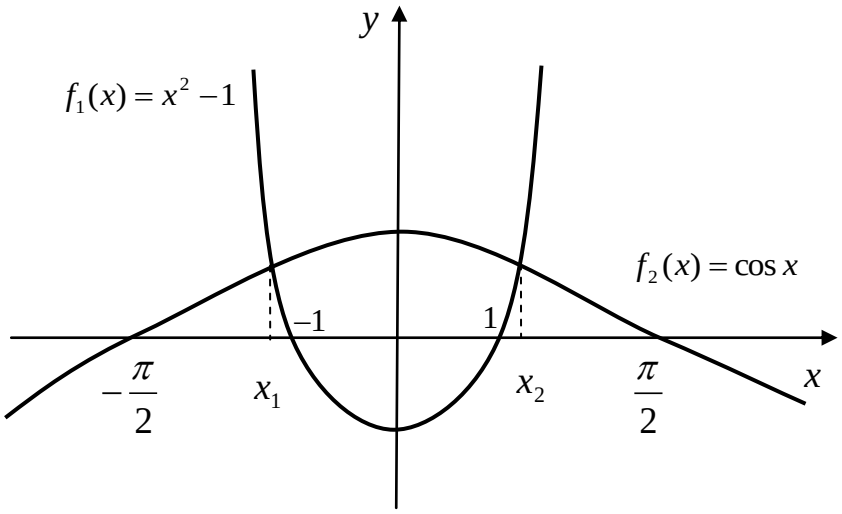


Рис. 5.3. Локалізація коренів рівняння $x^2 - \cos x - 1 = 0$

Ще одним способом відокремлення коренів є розбиття відрізка $[a, b]$, на якому знаходяться корені, точками $x_i = a + ih$, $i = \overline{1, n}$, $x_0 = a$, $x_n = b$, де $h = (b - a)/n$ – досить мале. На кожному з інтервалів $(x_\nu, x_{\nu+1})$ перевіряється умова $f(x_\nu) f(x_{\nu+1}) < 0$. Якщо відома кількість коренів і вони прості, то, зменшуючи h, наприклад, у два рази, можна локалізувати всі корені. Недоліком цього способу є те, що для близьких коренів, відстань між якими має порядок точності обчислень, для малих h прийдемо до таких проміжків, кожна з точок яких може бути прийнята за корінь рівняння (5.1).

5.3. Швидкість збіжності ітераційного методу

Нелінійні рівняння наближено розв'язуються ітераційними методами. Для оцінки швидкості збіжності та порівняння ітераційних методів між собою наведемо деякі поняття [13, 28, 49, 73]. Нехай послідовність значень $x_0, x_1, \ldots$, яка має своєю границею x^* при $k \to \infty$, є результатом виконання ітераційного процесу

$$x_{k+1} = \varphi(x_k), \quad k = 0, 1, ...$$

Означення 5.2. *Збіжність послідовності* $\{x_k\}_{k=0}^{\infty}$ *називається лінійною (ітераційний процес збігається лінійно), якщо для деякої сталої* $C \in (0,1)$ *і деякого номера* k_0 *виконується нерівність*

$$| x_{k+1} - x^* | \le C | x_k - x^* | \qquad \forall k \ge k_0 . \tag{5.4}$$

Лінійну збіжність також називають збіжністю зі швидкістю геометричної прогресії. Справді, якщо ввести позначення $\varepsilon_k = | x_k - x^* |$, то $\varepsilon_{k+1} = | x_{k+1} - x^* | \le C\varepsilon_k$ і мажоранти ε_k, $k = 0,1...$, утворюють геометричну прогресію, збіжність якої досягається при $C < 1$.

Означення 5.3. *Послідовність* $\{x_k\}_{k=0}^{\infty}$ *збігається з надлінійною швидкістю, якщо існує така додатна послідовність* $\{C_k\}_{k=0}^{\infty}$, *що* $C_k \to 0$ *при* $k \to \infty$ *і*

$$| x_{k+1} - x^* | \le C_k | x_k - x^* | .$$

Означення 5.4. *Послідовність* $\{x_k\}_{k=0}^{\infty}$ *збігається із порядком* p *(ітераційний процес має хоча б* p*-й порядок), якщо знайдуться такі сталі* $C > 0$, $p > 1$, $k_0 \ge 0$, *що*

$$| x_{k+1} - x^* | \le C | x_k - x^* |^p, \quad k \ge k_0 . \tag{5.5}$$

Для $p = 2$ збіжність називається квадратичною. Якщо можна вказати послідовність $\{C_k\}$ таку, що $| x_{k+1} - x^* | \le C_k | x_k - x^* |^p$ і $C_k \to C$ при $k \to \infty$, то маємо асимптотичну збіжність із порядком p. Аналогічно, маємо асимптотично лінійну збіжність, якщо $p = 1$ і $C_k \to C \in (0,1)$.

Якщо ітераційний метод має порядок p, то зі зростанням k можна спостерігати стабілізацію відношення

$$\frac{| x_{k+1} - x^* |}{| x_k - x^* |^p}$$

відносно сталої . Якщо ж швидкість збіжності не дорівнює p, то такої стабілізації вже не буде (див. приклад 5.4).

Означення 5.5. *Якщо ітераційний процес збігається для всіх початкових значень* x_0 *із деякої області, то збіжність називається глобальною. Якщо ж тільки для початкових значень*

x_0 із деякого досить малого околу шуканого розв'язку x^*, то ітераційний метод збігається локально.

При дослідженні збіжності ітераційних процесів використовуються *апріорні оцінки* похибки вигляду

$$|x_{k+1} - x^*| \le C\nu^{p^k},$$

де $C > 0, \nu \in (0,1), p -$ порядок методу, або *апостеріорні оцінки*

$$|x_{k+1} - x^*| \le C|x_{k+1} - x_k|^p.$$

Контроль точності ітераційного наближення x_{k+1} здійснюється згідно з апостеріорною оцінкою

$$C|x_{k+1} - x_k|^p \le \varepsilon,$$

де $\varepsilon > 0$ характеризує точність обчислення кореня, або малістю відносної похибки. Інші поняття збіжності та їх деталізація наведена у працях [43, 70, 77].

5.4. Метод половинного поділу

Нехай для рівняння $f(x) = 0$, де $f \in C[a,b]$, виконується нерівність (5.3) і корінь $x^* \in (a,b)$ – єдиний. Запишемо алгоритм методу половинного поділу у такому вигляді.

1. Увести: a, b, абсолютну або відносну похибку розв'язку і/або точність обчислення функції f.
2. $k := 1;$
3. ***do*** $c := (a+b)/2;$
   if $f(a)f(c) \le 0$ ***then***
   if $f(c) = 0$ ***then*** $x := c$
   else $b := c; \quad k := k+1;$
   else $a := c; \quad k := k+1;$
   until $abs(a-b) \le \varepsilon \ \vee \ abs(f(c)) \le \delta$
4. $x := c;$ $x - k$-те наближення розв'язку.

Геометрична ілюстрація методу половинного поділу показана на рис. 5.4. Метод може збігатись досить повільно, якщо корінь знаходиться близько до одного з кінців проміжку $[a,b]$ (див. приклади у підрозділі 5.11).

З алгоритму методу половинного поділу випливає, що

$$|x_{k+1} - x^*| \le \frac{1}{2}|x_k - x^*|, \quad k = 1,2,...$$

тобто $C = 0.5,\ \ p = 1$ і метод має лінійну збіжність. Оскільки на кожній ітерації довжина попереднього відрізка зменшується удвічі, то оцінка похибки набуває вигляду

$$| x_k - x^* | \le \frac{1}{2^k}(b-a), \quad k = 1, 2, \ldots$$

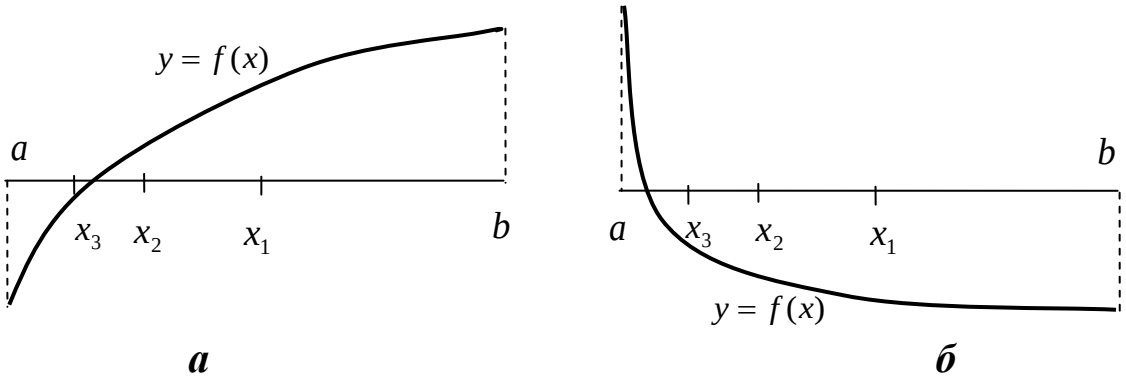


Рис. 5.4. Геометрична ілюстрація методу половинного поділу

5.5. Метод лінійної інтерполяції

На відміну від методу половинного поділу, у цьому методі точкою поділу відрізок $[a,b]$ ділиться на частини пропорційно ординатам $\left|f(a)\right|$ і $\left|f(b)\right|$ (рис. 5.5.)

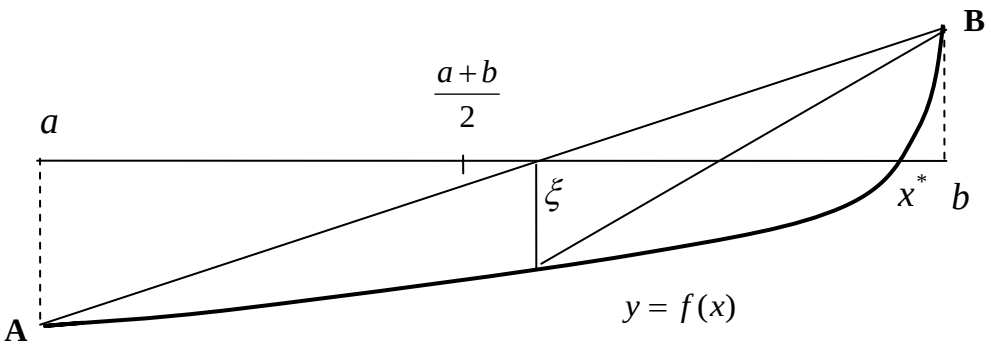


Рис. 5.5. Ілюстрація методу лінійної інтерполяції

Запишемо рівняння прямої, яка проходить через точки $A(a, f(a))$ і $B(b, f(b))$

$$\frac{y - f(b)}{f(a) - f(b)} = \frac{x - b}{a - b}.$$

Звідси знаходимо точку $\xi = b - \dfrac{f(b)(b-a)}{f(b) - f(a)}$ перетину хорди з віссю абсцис Ox. Якщо ввести позначення $x_0 = a,\ x_1 = b$, то метод

лінійної інтерполяції (його ще називають методом хорд, або *regula falsi* [50], набуває вигляду

$$x_{k+1} = x_k - \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})}, \quad k = 1, 2, \ldots \tag{5.6}$$

Метод лінійної інтерполяції детально вивчений у [23, 36, 50]. Розглянемо випадок, коли $f \in C^1[a, b]$ і

$$0 < m_1 \le \left| f'(x) \right| \ge M_1, \; x \in [a, b]. \tag{5.7}$$

Згідно з формулою Тейлора маємо

$$f(x_k) = f(x^*) + f'(\theta_1)(x_k - x^*) = f'(\theta_1)(x_k - x^*),$$

де точка θ_1 знаходиться між x_k і x^*. Згідно з теоремою про скінченні прирости $f(x_k) - f(x_{k-1}) = f'(\theta_2)(x_k - x_{k-1})$, точка θ_2 належить інтервалу з кінцями x_{k-1} і x_k. Тоді з рівності (5.6) випливає, що

$$x_{k+1} - x^* = x_k - x^* - \frac{f(\theta_1)}{f'(\theta_2)}(x_k - x^*) = \frac{f'(\theta_2) - f(\theta_1)}{f'(\theta_2)}(x_k - x^*).$$

Враховуючи нерівності (5.7), одержимо

$$| x_{k+1} - x^* | \le \frac{M_1 - m_1}{m_1} | x_k - x^* |.$$

Якщо метод лінійної інтерполяції збіжний і $M_1 < 2m_1$, то $C = (M_1 - m_1) / m_1 < 1$ і збіжність лінійна.

Повніше проаналізувати метод лінійної інтерполяції вдається для випадку, коли зберігаються знаки першої та другої похідної на $[a, b]$. Нехай

$$f'(x) > 0, \; f''(x) > 0, \; x \in [a, b]. \tag{5.8}$$

Інші випадки розглядаються аналогічно. Уведемо позначення: $x_0 = b, \; x_1 = a$ (рис. 5.5а), тоді

$$f(x_0) f''(x_0) > 0. \tag{5.9}$$

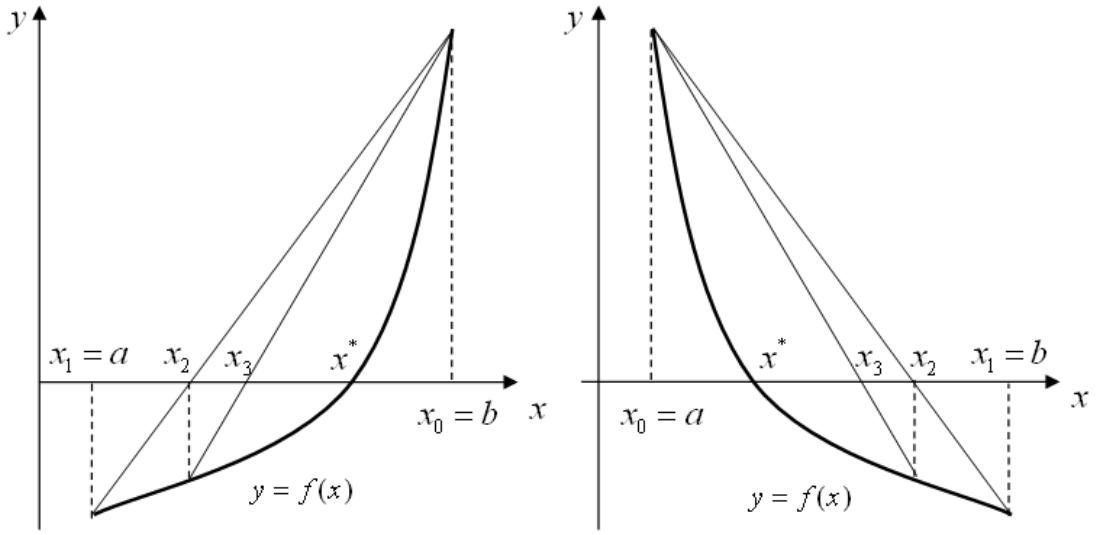
Нерівність (5.9) *називається умовою Фур'є* [23, 50]. Точка x_0 у цьому випадку залишається нерухомою і формула (5.6) набуває вигляду

$$x_{k+1} = x_k - \frac{f(x_k)(x_k - b)}{f(x_k) - f(b)}, \; k = 1, 2, \ldots \tag{5.10}$$

На рис. 5.6а видно, що $x_1 < x_2 < x_3 < \ldots$ Справді, із (5.10) випливає, що на першій ітерації ($k = 1$)

$$x_2 = a - \frac{f(a)(a-b)}{f(a) - f(b)} > a,$$

оскільки $f(a) < 0$ і $f(b) > f(a)$. Далі залишається застосувати метод математичної індукції.



а) $f''(x) > 0,\ f'(x) > 0$ *б)* $f''(x) < 0,\ f'(x) < 0$

Рис. 5.6. Метод лінійної інтерполяції для знакопостійних f' і f'' на $[a, b]$

Послідовність $\{x_n\}_{n=1}^{\infty}$ монотонно зростає і обмежена зверху числом x^*, тому існує границя $\lim\limits_{k\to\infty} x_k = \xi$. Перейшовши в (5.10) до границі при $k \to \infty$, одержимо, що $f(\xi) = 0$. Оскільки корінь єдиний на $[a, b]$, то $\xi = x^*$.

Якщо $f \in C^2[a, b]$, похідні f' і f'' зберігають знак на $[a, b]$, m_1 і M_1 визначені згідно з (5.7) і x_0 задовольняє умову Фур'є (5.9), то, як показано в [5, 59], правильна апостеріорна оцінка похибки методу лінійної інтерполяції:

$$| x_{k+1} - x^* | \le \frac{M_1 - m_1}{m_1} | x_{k+1} - x_k |. \qquad (5.11)$$

У загальному випадку довжина проміжку локалізації кореня може не прямувати до нуля. Без додаткових умов, наприклад $|f'(x)| \ge m_1$, метод лінійної інтерполяції може збігатись повільніше, ніж метод половинного поділу.

Приклад 5.2. Результати розв'язування рівнянь методом лінійної інтерполяції та методом половинного поділу для різної кількості ітерацій *k* наведені в табл. 5.1. Усі цифри результатів ітерацій правильні.

Таблиця 5.1

Результати розв'язання рівнянь $f(x)=0$ за допомогою методу лінійної інтерполяції (МЛІ) і методу половинного поділу (МПП)

	a	b	$k=5$		$k=10$		$k=20$	
			МЛІ	МПП	МЛІ	МПП	МЛІ	МПП
$x-\exp(-0.5x)$	0	1	0,703468	0,71875	0,7034674	0,704102	0,703467	0,703468
$x-1/2018$	0	1	0,000498	0,03125	0,0004975	0,000977	0,000498	0,000497
x^2-2	1	2	1,414141	1,40625	1,4142135	1,415039	1,414214	1,414214
x^4+2x^3-x-1	0	1	0,866085	0,84375	0,86621093	0,865303	0,866760	0,866761
$x\cos x-2x^2+$ $+3x-1$	0,2	0,3	0,298500	0,296875	0,2975305	0,297559	0,297530	0,297530

5.6. Метод простої ітерації

Запишемо рівняння (5.1) у рівносильному вигляді

$$x=\varphi(x), \tag{5.12}$$

де функція $\varphi:[a,b]\to[a,b]$. Для довільного $x_0\in[a,b]$ розглянемо ітераційну послідовність

$$x_{k+1}=\varphi(x_k),\ k=0,1,... \tag{5.13}$$

Доведення збіжності ітерацій (5.13) випливає із теореми про стискаючі відображення [34 , с.73].

Теорема 5.1. *Нехай $\varphi: B\to B$, де B – повний метричний простір з метрикою ρ, а відображення φ – стискаюче, тобто існує таке $q\in(0,1)$, що для будь-яких $x_1,x_2\in B$ виконується нерівність $\rho(\varphi(x_1),\varphi(x_2))\le q\rho(x_1,x_2)$.*

Тоді відображення φ має єдину нерухому точку x^* (рівняння (5.12) має єдиний розв'язок x^*), яка є границею послідовних наближень $x^*=\lim\limits_{k\to\infty}x_k$ для довільного початкового значення $x_0\in B$. ■

Нагадаємо, що простір B – повний, якщо в ньому збігається будь-яка фундаментальна послідовність. Повним метричним простором є R^n із метрикою $\rho(x,y)=\|x-y\|$, породженою, наприклад, нормами $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$, замкнена куля в R^n, простір $C[a,b]$ з нормою $\|\varphi\|=\max\limits_{a\le x\le b}|\varphi(x)|$. Із теореми 5.1 випливає

обґрунтування збіжності послідовності (5.13) до розв'язку рівняння (5.12).

Теорема 5.2. *Нехай функція*

$$\varphi:[a,b]\to[a,b]$$

задає стискаюче відображення, тобто існує таке $q\in(0,1)$*, що*

$$\left|\varphi(x_2)-\varphi(x_1)\right|\le q\left|x_2-x_1\right| \quad \forall x_1,x_2\in[a,b]. \tag{5.14}$$

Тоді рівняння (5.12) має єдиний розв'язок x^**, який для довільного* $x_0\in[a,b]$ *є границею послідовності (5.13), причому*

$$\left|x_{k+1}-x^*\right|\le\frac{q}{1-q}\left|x_k-x_{k-1}\right|\le\frac{cq^k}{1-q},\quad c=\left|x_1-x_0\right|. \quad \blacksquare \tag{5.15}$$

Оцінки (5.15) випливають з таких нерівностей:

$$\left|x_{k+1}-x^*\right|=\left|\varphi(x_k)-\varphi(x^*)\right|\le q\left|x_k-x^*\right|\le q\left|x_{k+1}-x^*\right|+q\left|x_{k+1}-x_k\right|,$$

$$\left|x_{k+1}-x_k\right|=\left|\varphi(x_k)-\varphi(x_{k-1})\right|\le q\left|x_k-x_{k-1}\right|\le\ldots\le q^k\left|x_1-x_0\right|.$$

Отже, метод простої ітерації є глобально збіжним лінійним ітераційним методом.

Зауваження 5.1. *Нехай* $\varphi\in C^1[a,b]$*, тоді замість умови стиску (5.14) простіше перевірити умову*

$$\left|\varphi'(x)\right|\le q<1,\ x\in[a,b]. \tag{5.16}$$

Справді, для довільних $x_1,x_2\in[a,b]$ *маємо:*

$$\left|\varphi(x_2)-\varphi(x_1)\right|=\left|\varphi'(\theta)\right|\cdot\left|x_2-x_1\right|\le\max_{[a,b]}\left|\varphi'(x)\right|\cdot\left|x_2-x_1\right|\le q\left|x_2-x_1\right|.$$

Тут θ *– точка між* x_1 *і* x_2*. Тепер твердження теореми 1.3 одержується із теореми 5.2.* ■

Геометрична ілюстрація методу простої ітерації для рівняння (5.12), коли похідна φ' зберігає знак на (a,b) показана на рис. 5.7.

Нехай $B=[x_0-r,x_0+r]$. При виконанні ітерацій досить перевірити, що всі $x_k\in B$. Тоді умову $\varphi:B\to B$ можна замінити перевіркою умови

$$\left|\varphi(x_0)-x_0\right|\le(1-q)r, \tag{5.17}$$

де $1>q$ – стала Ліпшіця.

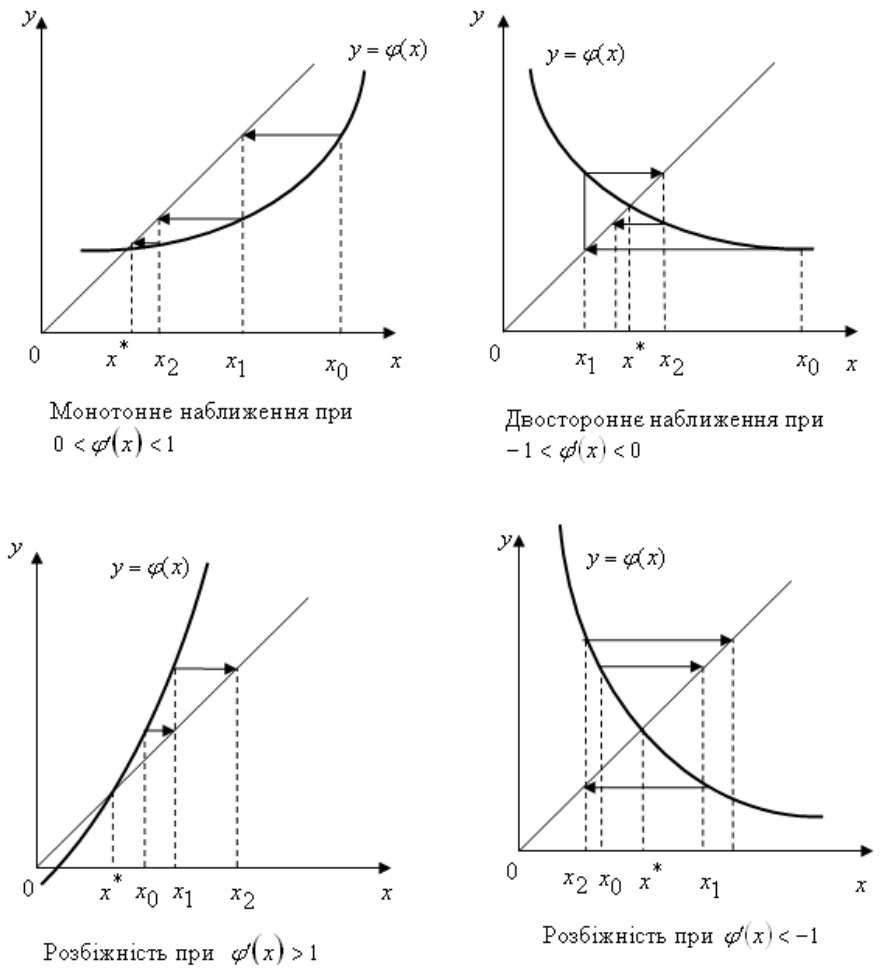


Рис. 5.7. Геометрична ілюстрація методу простої ітерації

Теорема 5.3 *(Ознака існування розв'язку). Якщо на проміжку* B *функція* φ *задовольняє умову Ліпшіца зі сталою* q, $0<q<1$, *і виконується нерівність (5.17), то рівняння (5.12) має єдиний розв'язок* x^*, *який є границею послідовності (5.13).*

Доведення. Досить довести, що $\varphi: B \to B$. Справді, для будь-якого $x \in B$ на підставі (5.14) і (5.17) маємо

$$\left|\varphi(x)-x_0\right| \le \left|\varphi(x)-\varphi(x_0)\right| + \left|\varphi(x_0)-x_0\right| \le$$

$$\le q\left|x-x_0\right| + (1-q)r \le qr + (1-q)r = r.$$

Якщо припустити існування кореня $x = x^*$ рівняння (5.12), то маємо наступну ознаку збіжності методу простої ітерації.

Теорема 5.4. *Нехай функція* φ *на проміжку* B *задовольняє умову Ліпшіця зі сталою* $q<1$. *Тоді послідовність (5.13) збігається до розв'язку* x^* *при будь-якому* $x_0 \in B$ *і має лінійну швидкість збіжності, тобто*

$$\left|x_{k+1}-x^*\right| \le q\left|x_k-x^*\right|. \qquad (5.18)$$

Доведення. Умова стиску випливає з умови Ліпшіця. Для довільного $x \in B$ маємо

$$\left|\varphi(x) - x^*\right| \le \left|\varphi(x) - \varphi\left(x^*\right)\right| \le q\left|x - x^*\right| \le qr < r,$$

тобто $\varphi : B \to B$ є *стискаючим відображенням*. ■

В обчислювальній практиці популярний емпіричний підхід до завершення ітераційного процесу (5.13) й оцінки похибки наближення x_{k+1}. А саме: обчислення ведуться до тих пір, поки не виконається нерівність $|x_{k+1} - x_k| < \varepsilon$, де ε – належно вибране досить мале число. Якщо порядок x_k наперед невідомий, то правильніше орієнтуватись на виконання нерівності $|x_{k+1} - x_k| \le \delta |x_k|$ або $|x_{k+1} - x_k| \le \delta(|x_k| + \sigma)$, де δ – відносна похибка, σ – характеристика точності виконання обчислень.

У загальному випадку перехід від рівняння (5.1) до (5.12) можна здійснити, записавши рівняння (5.1) у вигляді $x = x - \lambda f(x)$.

Параметр $\lambda \ne 0$ вибирається так, щоб похідна $\varphi'(x) = 1 - \lambda f'(x)$ в деякій області задовольняла нерівність (5.16). Наприклад, якщо $0 < \alpha \le f'(x) \le \gamma$, то для функції $\varphi(x) = x - \lambda f(x)$ одержимо

$$1 - \gamma\lambda \le \varphi'(x) \le 1 - \alpha\lambda.$$

Отже, $\left|\varphi'(\tilde{o})\right| \le q(\lambda) := \max\left\{\left|1 - \lambda\alpha\right|, \left|1 - \lambda\gamma\right|\right\}$. Аналізуючи одержані нерівності, можна побачити, що при будь-яких $\lambda \in (0, 2/\gamma)$ виконується оцінка $q(\lambda) < 1$. Зокрема, при $\lambda = 1/\gamma$ маємо:

$$0 \le \varphi'(x) \le 1 - \frac{\alpha}{\gamma} < 1,$$

що забезпечує монотонну збіжність ітераційного процесу із лінійною швидкістю з коефіцієнтом $q = 1 - \frac{\alpha}{\gamma}$. Оптимальним є значення параметра $\lambda = \lambda_0 := \frac{2}{\alpha + \gamma}$. При цьому значенні параметра $1 - \lambda\gamma = \frac{\alpha - \gamma}{\alpha + \gamma}$ і $1 - \lambda\alpha = \frac{\gamma - \alpha}{\alpha + \gamma}$, тобто максимум похідної, що дорівнює $q(\lambda_0) = \frac{\gamma - \alpha}{\alpha + \gamma}$, досягається на кожному з елементів

множини $\{|1-\lambda\alpha|,\ |1-\lambda\gamma|\}$. У прикладі 5.3 таким значенням є $\lambda_0 = 2/(2+4) = 1/3$.

Зауваження 5.2. *Якщо знайдено три послідовних наближення* x_n, x_{n+1}, x_{n+2} *розв'язку рівняння (5.1), то наступне значення* x_{n+1} *можна знайти згідно з* Δ^2 *– процесом Ейткена. Цей ітераційний процес має квадратичну швидкість збіжності і полягає ось у чому. Нехай* x^* *– точний розв'язок. Тоді*

$$x_{n+1} - x^* = \varphi(x_n) - \varphi(x^*) \approx \varphi'(x^*)(x_n - x^*),$$
$$x_{n+2} - x^* \approx \varphi'(x^*)(x_{n+1} - x^*).$$

Звідси маємо

$$\frac{x_{n+1} - x^*}{x_{n+2} - x^*} \approx \frac{x_n - x^*}{x_{n+1} - x^*}.$$

Розв'язавши це рівняння відносно x^*, *знайдемо*

$$x^* \approx x_{n+2} - \frac{(x_{n+2} - x_{n+1})^2}{x_{n+2} - 2x_{n+1} + x_n}.$$

Ітераційний процес набуває вигляду

$$x_{n+3} = x_{n+2} - \frac{(x_{n+2} - x_{n+1})^2}{x_{n+2} - 2x_{n+1} + x_n},\ n = 0,1,...$$

5.7. Побудова ітерацій у методі Ньютона

Розглянемо рівняння (5.1), для якого x^* – єдиний корінь на $[a,b]$. Одним із основних методів наближеного розв'язування нелінійних рівнянь і систем є метод Ньютона[1], що пов'язано з його простою реалізацією та квадратичною збіжністю[2]. Метод Ньютона узагальнений у 1948 р. для розв'язування нелінійних

[1] Ypma T.J. Historical development of the Newton–Raphson method // SIAM Review. – 1995. – 37 (4). – P. 531–551.

[2] Метод описаний І. Ньютоном у 1669 р., але застосований він був для многочленів. Обчислювалась не послідовність наближень, а послідовність многочленів, наслідком чого був наближений розв'язок. Уперше метод був опублікований у трактаті «Алгебра» Джона Валліса в 1685 р. У 1690 р. Джозеф Рафсон розглянув метод Ньютона вже як загальний ітераційний алгоритм, але також обмежився многочленами, тому метод і одержав назву «Метод Ньютона-Рафсона». Тільки в 1740 р. метод Ньютона сформульований Томасом Сімпсоном як ітераційний метод розв'язування нелінійного рівняння з використанням похідної в тому вигляді, як він використовується тепер. Він також узагальнив метод Ньютона для системи двох нелінійних рівнянь.

операторних рівнянь у функціональних просторах і відомий як *метод Ньютона-Канторовича* [29, 33].

Нехай x_k – деяке наближення розв'язку, де $f \in C^2[a,b]$. За формулою Тейлора для довільної точки $t \in [a,b]$ маємо

$$f(t) = f(x_k) + f'(x_k)(t - x_k) + f''(\theta_k)(t - x_k)^2 / 2,$$

де θ_k – деяка точка між t і x_k. Оскільки $f(x^*) = 0$, то вважаючи, що $t = x^*$, одержимо

$$f(x_k) + f'(x_k)(x^* - x_k) + f''(\theta_k)(x^* - x_k)^2 / 2 = 0. \qquad (5.19)$$

Якщо значення x_k близьке до x^*, то $(x^* - x_k)^2$ досить мале, порівняно з $|x^* - x_k|$, тому, нехтуючи останнім доданком у правій частині (5.19), одержимо $f(x_k) + f'(x_k)(x^* - x_k) \approx 0$. Запишемо точну рівність, замінивши x^* наближеним значенням x_{k+1}. Одержимо ітераційний ***метод Ньютона***, який визначається лінійним рівнянням

$$f(x_k) + f'(x_k)(x_{k+1} - x_k) = 0, \qquad (5.20)$$

звідки

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, k = 0,1,... \qquad (5.21)$$

Формула (5.21) має зміст, якщо $f'(x_k) \neq 0$. Метод Ньютона має простий геометричний зміст, тому він ще має назву ***методу дотичних*** (рис. 5.8). Запишемо рівняння дотичної до графіка функції $y = f(x)$ у точці $(x_k, f(x_k))$

$$y - f(x_k) = f'(x_k)(x - x_k)$$

і позначимо абсцису точки перетину дотичної через x_{k+1}. Тоді при $y = 0$ одержимо формулу (5.21).

Вивести формулу (5.21) можна також з методу ітерацій, якщо записати рівняння (5.1) у вигляді $x = x - \lambda f(x)$, а параметр λ на k-й ітерації вибирати так, щоб $\varphi'(x_k) = 1 - \lambda f'(x_k) = 0$. Тоді $\lambda_k = 1/f'(x_k)$ (метод нестаціонарний) і

$$x_{k+1} = x_k - \lambda_k f(x_k) = x_k - \frac{f(x_k)}{f'(x_k)}, k = 0,1,...$$

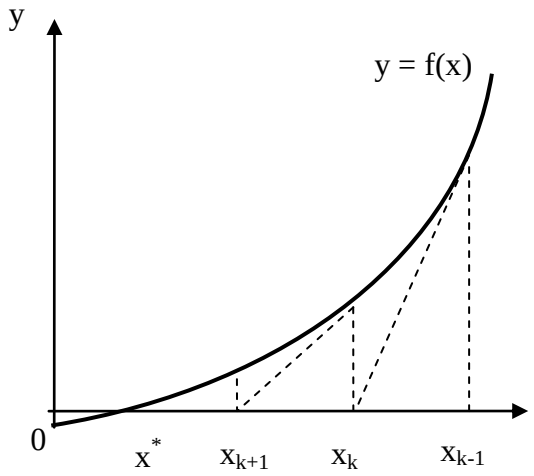


Рис. 5.8. Ілюстрація методу Ньютона

5.8. Оцінка похибки та збіжність методу Ньютона

Теорема 5.5. *Нехай функція* $f \in C^2[a,b]$,

$$|f'(x)| \geq m_1 > 0,\ |f''(x)| \leq M_2$$

, *(5.22)*

ітераційні наближення $x_k \in [a,b]\ \forall k$ *і ця послідовність збіжна до кореня* x^* *рівняння (5.1), тоді для довільного цілого* $k \geq 0$

$$|x_{k+1} - x^*| \leq \frac{M_2}{2m_1}(x_k - x^*)^2, \qquad (5.23)$$

$$|x_{k+1} - x^*| \leq \frac{M_2}{2m_1}(x_{k+1} - x_k)^2. \qquad (5.24)$$

Перша з нерівностей означає, що метод має другий порядок, а друга служить для оцінки похибки методу.

Доведення. Прирівнявши ліві частини рівностей (5.19) і (5.20), одержимо,

$$f(x_k) + f'(x_k)(x^* - x_k) + f''(\theta_k)(x_k - x^*)^2/2 = f(x_k) + f'(x_k)(x_{k+1} - x_k),$$

або

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

На підставі нерівностей (5.22) одержуємо оцінку (5.23).

Для доведення нерівності (5.24) запишемо формулу Тейлора для функції $f(x)$ у точці x_{k+1}

$$f(x_{k+1}) = f(x_k) + f'(x_k)(x_{k+1} - x_k) +$$

$$+ f''(\overline{\theta_k})(x_{k+1} - x_k)^2)/2 = f''(\overline{\theta_k})(x_{k+1} - x_k)^2/2.$$

Згідно з формулою Лагранжа

$$f(x_{k+1}) - f(x^*) = f'(\xi_{k+1})(x_{k+1} - x^*), \qquad (5.25)$$

де ξ_{k+1} – точка між x_{k+1} і x^*. Врахувавши, що $f(x^*)=0$ і рівність (5.20), одержимо

$$f'(\xi_{k+1})(x_{k+1}-x^*)=f''(\overline{\theta_k})(x_{k+1}-x_k)^2/2.$$

Звідси випливає шукана оцінка

$$\left|x_{k+1}-x^*\right|=\frac{\left|f(x_{k+1})\right|}{\left|f'(\xi_{k+1})\right|}\le\frac{M_2}{2m_1}(x_{k+1}-x_k)^2. \qquad \blacksquare$$

Зауваження 5.3. *Із (5.25) для n=k+1 одержимо*

$$\left|x_n-x^*\right|=\frac{\left|f(x_n)\right|}{\left|f'(\xi_n)\right|}\le\frac{\left|f(x_n)\right|}{m_1},$$

тобто маємо можливість контролювати точність обчислення кореня x_n за величиною нев'язки $f(x_n)$. А саме: точність ε досягнена, якщо $\left|x_n-x^*\right|\le\dfrac{\left|f(x_n)\right|}{m_1}<\varepsilon$ *або*

$$\left|f(x_n)\right|<m_1\varepsilon. \qquad (5.26)$$

Якщо $m_1\ge 1$, то $\left|x_n-x^*\right|\le\left|f(x_n)\right|$ і $x^*:=x_n$ із точністю ε.. Якщо ж $m_1<1$, то потрібно домагатися виконання нерівності (5.26). ■

Доведемо збіжність методу Ньютона, коли похідні $f'(x)$ і $f''(x)$ знакопостійні на $[a,b]$, тобто $f'(x)f''(x)\ne 0\ \ \forall x\in[a,b]$.

Теорема 5.6. Якщо $f\in C^2[a,b]$, $f(a)f(b)<0$ і похідні f' і f'' зберігають знак на $[a,b]$, то, починаючи з початкового наближення $x_0\in[a,b]$, для якого виконується умова Фур'є

$$f(x_0)f''(x_0)>0, \qquad (5.27)$$

метод Ньютона збіжний до єдиного розв'язку x^* рівняння (5.1), де x_0 – це a або b, для якого виконується умова (5.27).

Доведення. Нехай похідні $f'(x)>0$ і $f''(x)>0$ для $x\in[a,b]$ (рис. 5.1а). Тоді можна взяти $x_0\in(x^*,b]$, зокрема $x_0=b$. Умова $f(a)f(b)<0$ і знакопостійність $f'(x)$ гарантують єдиність розв'язку рівняння (5.1). Для $x_0>x^*$ згідно з формулою Тейлора

$$f(x_0)+f'(x_0)(x^*-x_0)+f''(\theta_0)(x^*-x_0)^2/2=0.$$

Оскільки $f''(\theta_0)>0$, то рівність виконується, якщо $f(x_0)+$ $+f'(x_0)(x^*-x_0)<0$, звідки випливає, що $x^*<x_0-f(x_0)/f'(x_0)=$ $=x_1<x_0$. За індукцією просто довести, що $x^*<x_k<x_{k+1}\ \ \forall k$.

Оскільки послідовність $\{x_k\}$ монотонно спадна й обмежена знизу x^*, то $\lim\limits_{k\to 0} x_k = \bar{x} \geq x^*$. Рівність $\bar{x} = x^*$ випливає з (5.21) при $k \to \infty$. Інші 3 випадки розглядаються аналогічно (рис. 5.9, б-в).

Зауваження 5.4. *Якщо за початкове наближення взяти точку x_0, для якої не виконується умова Фур'є, то x_1 може не належати проміжку $[a,b]$.*

Існування кореня рівняння (5.1) і збіжність методу Ньютона обґрунтовано й при слабкіших обмеженнях на функцію f. Сформулюємо одну з таких теорем.

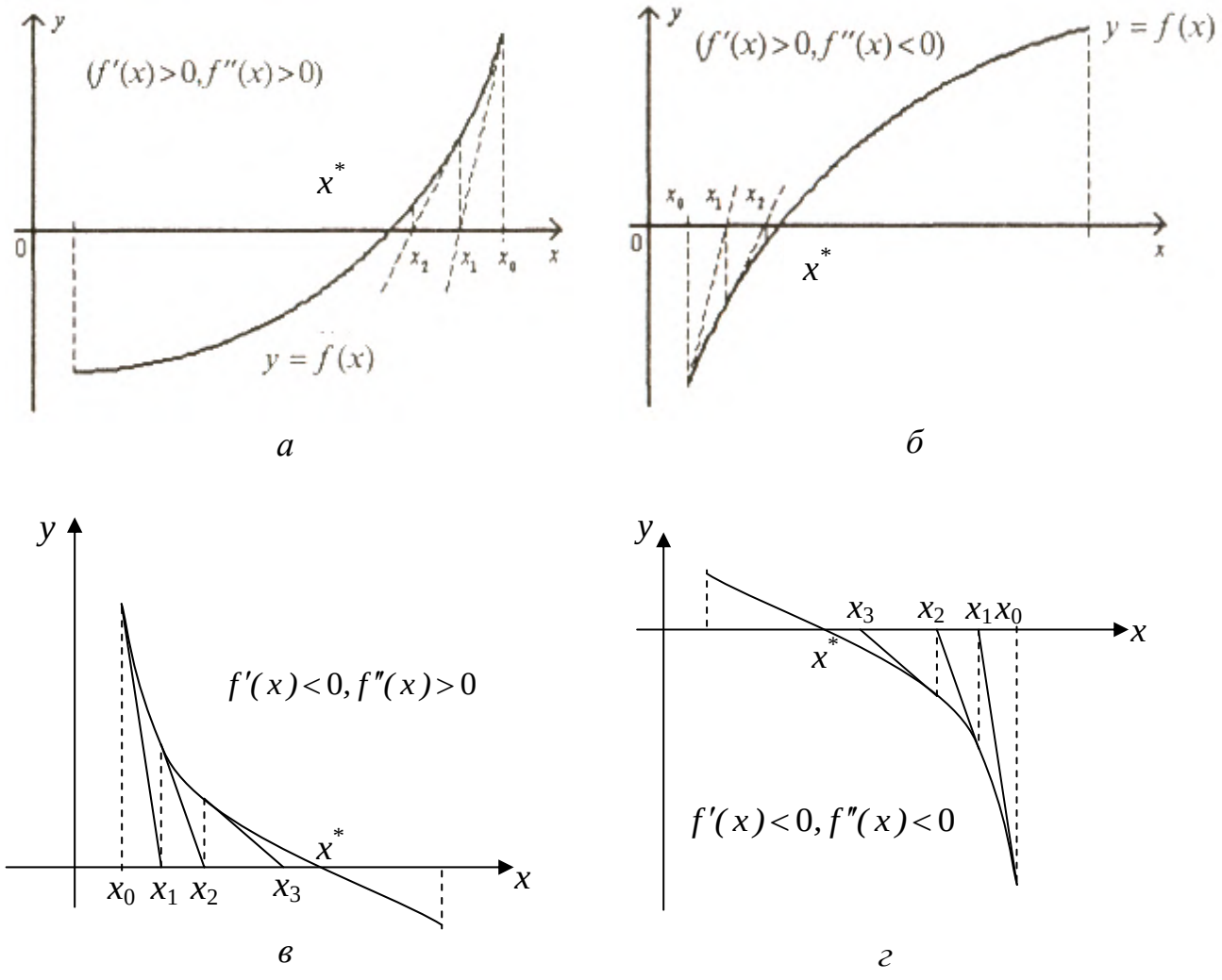


Рис. 5.9. Метод Ньютона для знакопостійних f' і f''

Теорема 5.7 [22, 28, 36, 50]. Нехай виконуються наступні умови:

1) $f \in C^2[x_0, x_0 + 2h_0]$, де $f(x_0) f'(x_0) \neq 0$, $h_0 = -f(x_0)/f'(x_0)$;

2) $2|h_0| \max\limits_{[x_0, x_0+2h_0]} |f''(x)| \leq |f'(x_0)|$.

Тоді всі ітерації $x_k \in [x_0, x_0 + 2h_0]$, де x_k визначається згідно з (5.21), $\lim\limits_{k\to\infty} x_k = x^*$, де x^* – єдиний корінь рівняння *f(x)=0*, і виконується оцінка

$$|x_{k+1} - x^*| \leq \frac{M_2}{2|f'(x_k)|}(x_k - x_{k-1})^2, k = 1,2,... \qquad (5.28)$$

Приклад 5.3. Застосуємо метод Ньютона для обчислення $\sqrt{a}$, що рівносильне знаходженню додатного кореня рівняння $f(x) = x^2 - a = 0$. Згідно з (5.21) ітерації набувають вигляду

$$x_{k+1} = x_k - \frac{x_k^2 - a}{2x_k} = \frac{x_k}{2} + \frac{a}{x_k} = \frac{x_k^2 + a}{2x_k}, k = 0,1,... \qquad (5.29)$$

Похибка на (k+1)-й ітерації

$$r_{k+1} = x_{k+1} - \sqrt{a} = \frac{x_k^2 + a}{2x_k} - \sqrt{a} = \frac{\left(x_k - \sqrt{a}\right)^2}{2x_k} = \frac{r_k^2}{2x_k}.$$

Покажемо, що абсолютна величина похибки зменшується з кожною ітерацією принаймні вдвічі. Справді,

$$r_{k+1} = \frac{r_k}{2x_k} \cdot r_k = \frac{x_k - \sqrt{a}}{2x_k} r_k = \left(\frac{1}{2} - \frac{\sqrt{a}}{2x_k}\right) r_k. \qquad (5.30)$$

Оскільки $x_k > 0$, то $r_{k+1} \geq 0$ для $k = 0,1,...$ Тому $x_k - \sqrt{a} \geq 0$, тобто $x_k \geq \sqrt{a}$, $k = 0,1,...$ Унаслідок цього $0 \leq \frac{1}{2} - \frac{\sqrt{a}}{2x_k} \leq \frac{1}{2}, k = 1,...$

Отже, $|r_{k+1}| \leq |r_k|/2, k = 0,1,....$

Розглянемо відносну похибку $\delta_k = \frac{|r_k|}{\sqrt{a}} = \left|1 - \frac{x_k}{\sqrt{a}}\right|$. Тоді

$$\delta_{k+1} = \frac{|r_{k+1}|}{\sqrt{a}} = \frac{r_k^2}{2x_k\sqrt{a}} < \frac{1}{2}\frac{r_k^2}{(\sqrt{a})^2} \leq \frac{1}{2}\delta_k^2.$$

Це означає, що для достатньо близького наближення до $\sqrt{a}$ кожне наступне наближення наближеного подвоює число правильних цифр.

Таблиця 5.2

Порівняльна таблиця обчислення $\sqrt{2} = 1.41421356...$ різними ітераційними методами

k	Метод поділу відрізка пополам		Метод хорд		Метод Ньютона	
	x_k	r_k	x_k	r_k	x_{k+1}	r_k
0	1,0	–	2.0	–	2	–
1	2,0	–	1.0	–	1,5	0,0857864
2	1,5	0,275255	1,3333333	-0,0808801	1,4166666	0,0024530
3	1,25	0,131966	1,3999999	-0,0142135	1,4142156	2,1215639E-6
4	1,375	0,202396	1,4117647	-0,0024488	1,41421356 5	-2,4203139E-8
5	1,4375	0,238542	1,4137930	-0,0004204		
6	1,406250	0,220396	1,4141414	-7,2145820E-5		
7	1,421875	0,229450	1,4142011	-1,2421969E-5		
8	1,417969	0,227184	1,4142113	-2,1699704E-6		
9	1,416016	0,226051	1,4142131	-3,8183100E-7		
10	1,415039	0,225485	1,4142135	-2,4203139E-8		

Метод січних		Метод ітерацій	
x_k	r_k	x_{k+1}	r_k
1.0	–	2	-
2.0	–	1,5	0,0857864
1,3333333	-0,0808801	1,4375	0,0232864
1,3999999	-0,0142135	1,4208984	0,0066848
1,4146341	0,0004205	1,4161603	0,0019467
1,4142113	-2,1699704E-6	1,4147827	0,0005692
1,4142135	-2,4203139E-8	1,4143801	0,0001666
		1,4142624	4,8851605E-5
		1,4142278	1,4280911E-5
		1,4142177	4,1481221E-6
		1,4142147	1,1678897E-6

Процес (5.30) знаходження квадратних коренів за методом Ньютона відомий як ***процес Горнера***. Оскільки $f'(x) = 2x > 0$ і $f''(x) = 2,$ то умови монотонності виконуються. Залишається задовольнити умову Фур'є (5.28). Для цього потрібно вибрати x_0 так, щоб $x_0^2 > a.$ Наприклад, якщо взяти $x_0 = 2^{0.5m}$, якщо m – парне, і $x_0 = 2^{0.5(m+1)}$; якщо m непарне, де m таке, що $a = 2^m q,\ q \in [0.5; 1].$ Тоді $2^{m-1} \le a < 2^m.$

Якщо потрібно обчислити $\sqrt[n]{a}$ для $a > 0,$ то процес наближень для $\sqrt[n]{a}$ будується методом Ньютона, як розв'язок рівняння $x^n - a = 0.$ Ітераційний процес набуває вигляду

5.9. Випадок кратних коренів

Припустимо, що відома точна кратність m кореня x^* рівняння (5.1). Тоді $f^{(m)}(x^*) \neq 0$. Оскільки значення $f'(x)$ мале для x, близьких до x^*, то можна показати, що метод Ньютона збігається лінійно і знаменник відповідної геометричної прогресії наближено дорівнює $1-m^{-1}$, де m – кратність кореня. Нехай для парних m значення $f(x) > 0$, коли $x \in [a,b] / \{x^*\}$, інакше розглянемо рівняння $-f(x) = 0$. Тоді для рівняння

$$g(x) = \sqrt[m]{f(x)} = 0$$

x^* – простий корінь і метод Ньютона для нього набуває вигляду

$$x_{k+1} = x_k - \frac{g(x_k)}{\sqrt{g'(x_k)}} = x_k - \frac{\sqrt[m]{f(x_k)}}{m^{-1}\sqrt{(f(x_k))^{m-1}} \cdot f(x_k)} =$$

$$= x_k - m\frac{f(x_k)}{f'(x_k)}.$$

Отже, якщо відомо точка кратність m, то маємо таку модифікацію методу Ньютона:

$$x_{k+1} = x_k - m\frac{f(x_k)}{f'(x_k)}, k = 0,1,... \qquad (5.31)$$

Приклад 5.4. Розглянемо рівняння $f(x) := x^3 - 3x + 2 = 0$. Тут $x^* = 1$ – корінь кратності $m = 2$. Результати порівняння наближених значень $x_{k+1} = x_k - \frac{x_k^3 - 3x_k + 2}{3\left(x_k^2 - 1\right)} = \frac{2x_k^3 - 2}{3\left(x_k^2 - 1\right)}$ і $\bar{x}_{k+1} = \bar{x}_k - 2\frac{\bar{x}_k^3 - 3\bar{x}_k + 2}{3\left(\bar{x}_k^2 - 1\right)} = \frac{\bar{x}_k^3 + 3\bar{x}_k - 4}{3\left(\bar{x}_k^2 - 1\right)}$, обчислених згідно з (5.21) і (5.31) відповідно, з однаковими початковими значеннями $x_0 = \bar{x}_0 = 2$. наведено в таблиці 5.3. Швидкість збіжності порівнюється обчисленням відношення $\frac{|r_{k+1}|}{|r_k|}$ і $\frac{|r_{k+1}|}{r_k^2}$, де $r_k = x_k - x^*$, та відповідних відношень для $\bar{x}_k$, $\bar{r}_k = \bar{x}_k - x^*$.

Ще однією модифікацією методу Ньютона на випадок кратних коренів є ***метод Хейлі***. Якщо x^* – корінь кратності $p \geq 2$, то для функції $g(x) = f(x)/f'(x)$ цей корінь уже простий.

Таблиця 5.3

Наближення для кратного кореня рівняння $x^3 - 3x + 2 = 0$

k	Метод Ньютона				Модифікований метод Ньютона (p=2)			
	x_k	r_k	$\frac{\lvert r_{k+1}\rvert}{\lvert r_k\rvert}$	$\frac{\lvert r_{k+1}\rvert}{\lvert r_k\rvert^2}$	x_k	$\bar{r}_k$	$\frac{\lvert \bar{r}_{k+1}\rvert}{\lvert \bar{r}_k\rvert}$	$\frac{\lvert \bar{r}_{k+1}\rvert}{\lvert \bar{r}_k\rvert^2}$
0	2	–	–	–	2	–	–	–
1	1,5555555	0,55556	0,55556	0,55556	1,1111111	0,1111	0,1111	0,11111
2	1,2979066	0,29791	0,53623	0,96522	1,0019493	0,0019	0,0019	0,15789
3	1,1553901	0,15539	0,52161	1,75091	1,0000006	6,3269E-7	6,3269E-7	0,16650
4	1,0795622	0,07956	0,51202	3,29503	1,0000000	6,7502E-14	6,7502E-14	0,16863
5	1,0402884	0,04029	0,50638	6,36454				
6	1,0202768	0,02028	0,50330	12,49221				
7	1,0101723	0,01017	0,50168	24,74121				
8	1,0050947	0,00510	0,50084	49,23589				
9	1,0025495	0,00255	0,50042	98,22354				
10	1,0012753	0,00128	0,50021	196,1979				

Застосувавши метод Ньютона до рівняння $g(x) = 0$, одержимо метод Хейлі

$$x_{k+1} = x_k - \frac{f(x_k) f'(x_k)}{(f'(x_k))^2 - f(x_k) f''(x_k)}, \quad k = 0, 1, \ldots \tag{5.32}$$

На відміну від методу (5.31), тут не потрібно знати точної кратності кореня. Наприклад, для рівняння $x^3 - 3x + 2 = 0$, де $x^* = 1$ – корінь кратності $m = 2$, маємо

$$x_{k+1} = \frac{2(2x_k + 1)}{x_k^2 + 2x_k + 3}, k = 0,1,\ldots$$

Ітераційний процес знаходження $\sqrt{a}$ методом Хейлі набуває вигляду

$$x_{k+1} = x_k - \frac{(x_k^2 - a) \cdot 2x_k}{4x_k^2 - (x_k^2 - a) \cdot 2} = \frac{2ax_k}{x_k^2 + a}.$$

Для простого кореня метод Хейлі має кубічну збіжність.

5.10. Огляд інших ітераційних методів

5.10.1. Спрощений метод Ньютона. На першому кроці обчислюється $f'(x_0)$ і надалі $f'(x_k) := f'(x_0)$ або значення похідної змінюється з деяким періодом $r \geq 2$. Ітераційний процес набуває вигляду

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)}, k = 0,1,...$$

Геометрично це означає, що дотичні паралельні дотичній, проведеній у точці $(x_0, f(x_0))$ (рис. 5.9). Метод має тільки лінійну швидкість збіжності, що випливає з аналізу збіжності методу простої ітерації з функцією $\varphi(x) = x - f(x)/f'(x_0)$.

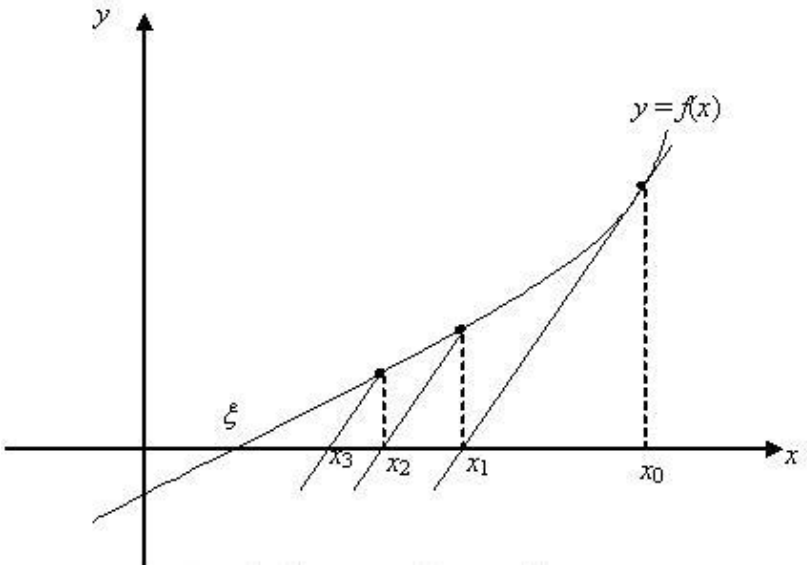


Рис. 5.10. Спрощений метод Ньютона

5.10.2. Різницевий метод Ньютона. Надлінійну швидкість збіжності можна досягнути тоді, коли похідну $f'(x_k)$ замінити на кожній ітерації деяким близьким значенням, яке обчислюється через значення функції $f(x)$. Наприклад, похідну можна апроксимувати різницевим відношенням

$$f'(x_k) \approx \frac{f(x_k + h_k) - f(x_k)}{h_k},$$

де h_k – малий параметр, який вибирається у певний спосіб. Тоді метод Ньютона набуває вигляду

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

Що стосується вибору h_k. Зі зростанням k значення h_k має зменшуватися з метою точнішої апроксимації $f'(x_k)$. Наприклад, $h_k = \lambda h_0$, де $\lambda < 1$. Можна взяти $\lambda = 0.5$ або 0.1. Недоліком такого вибору є відсутність зв'язку між $h_k \to 0$ і $x_k \to x^*$. Може виявитися, що x_k ще не достатньо близьке до x^*, а $|h_k|$ досить

мале і $f(x_k+h)$ і $f(x_k)$ реально не відрізняється. Протилежна ситуація може привести до втрати швидкості збіжності.

5.10.3. Метод січних. У цьому випадку $h_k = x_k - x_{k-1}$ і маємо ітераційний процес

$$x_{k+1} = x_k - \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})}, k = 0, 1, ... \qquad (5.33)$$

значення x_0 і x_1 потрібно задати. Формула (5.33) визначає двокроковий метод, оскільки значення x_{k+1} вимагає обчислення двох попередніх наближень x_{k-1} і x_k. На кожній ітерації обчислюється тільки одне значення функції $f(x_k)$, а $f(x_{k-1})$ − відоме з попередньої ітерації. Формулою (5.33) задається і метод хорд, але породжуються ці методи різними підходами. Геометрична ілюстрація методу січних показана на рис. 5.11. Наближення x_{k+1} – абсциса точки перетину січної, проведеної через точки $(x_{k-1}, f(x_{k-1}))$, $(x_k, f(x_k))$ із віссю x.

З'ясуємо швидкість збіжності методу січних при виконанні

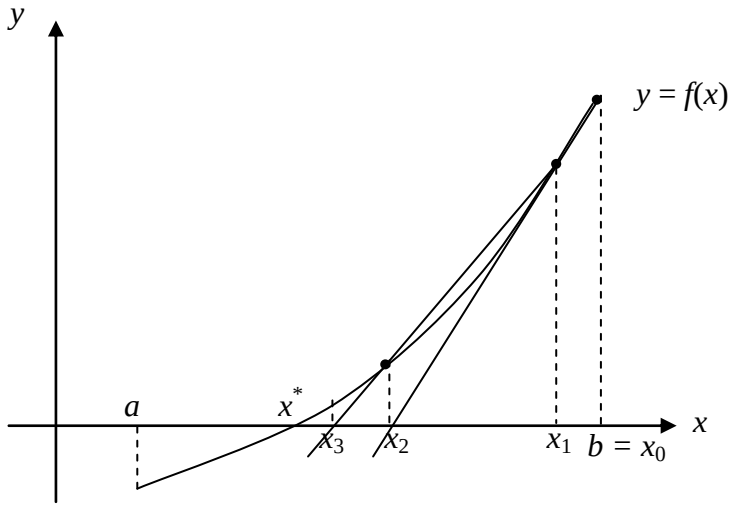


Рис. 5.11. Метод січних

умов теореми 5.2. Для цього проаналізуємо спадання похибки наближення. Із (5.33) маємо

$$x_{k+1} - x^* = x_k - x^* - \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})} = x_k - x^* - \frac{f(x_k)}{f(\theta_k)},$$

де θ_k − точка між x_k і x^*. За формулою Тейлора

$$f(x_k) = f(x^*) + f'(x^*)(x_k - x^*) + \frac{1}{2} f''(\overline{\theta_k})(x_k - x^*)^2.$$

Підставимо значення $f(x_k)$ у праву частину (5.33), врахувавши, що $f(x^*)=0$. Одержимо

$$x_{k+1}-x^*=\frac{x_k-x^*}{f'(\theta_k)}\left[\,f'(\theta_k)-f'(x^*)-\frac{1}{2}f''(\overline{\theta_k})(x_k-x^*)\right]=$$

$$=\frac{x_k-x^*}{f'(\theta_k)}\left[\,f''(\tilde{\theta}_k)(\theta_k-x^*)-\frac{1}{2}f''(\overline{\theta_k})(x_k-x^*)\right].$$

Якщо метод січних збіжний, то

$$\left|\theta_k-x^*\right|<\left|x_{k-1}-x^*\right|,\quad \left|x_k-x^*\right|<\left|x_{k-1}-x^*\right|,$$

і в підсумку маємо

$$\left|x_{k+1}-x^*\right|\le\frac{\left|x_k-x^*\right|}{\left|f'(\theta_k)\right|}\left[\left|f''(\tilde{\theta}_k)\right|\cdot\left|x_{k-1}-x^*\right|+\frac{1}{2}\left|f''(\overline{\theta_k})\right|\cdot\left|x_{k-1}-x^*\right|\right].$$

Звідси, одержимо

$$\left|x_{k+1}-x^*\right|\le\frac{3M_2}{2m_1}\left|x_k-x^*\right|\cdot\left|x_{k-1}-x^*\right|.$$

Уведемо позначення: $\varepsilon_k=\left|x_k-x^*\right|$, $C=3M_2(2m_1)^{-1}$. Тоді остання нерівність набуде вигляду $\varepsilon_{k+1}<C\varepsilon_k\varepsilon_{k-1}, k=1,2,...$ Для $k=\overline{1,4}$ маємо

$$\varepsilon_2<C\varepsilon_1\varepsilon_0<C^{-1}(C\varepsilon_0)^2,$$
$$\varepsilon_3<C\varepsilon_2\varepsilon_1<C^{-1}(C\varepsilon_0)^2\cdot C\varepsilon_0=C^{-1}(C\varepsilon_0)^3,$$
$$\varepsilon_4<C\varepsilon_3\varepsilon_2<C^{-1}(C\varepsilon_0)^5,$$
$$\varepsilon_5<C\varepsilon_4\varepsilon_3<C^{-1}(C\varepsilon_0)^8.$$

Пам’ятник Леонардо з Пізи (Фібоначчі)

Показники степенів утворюють послідовність чисел Фібоначчі: 1, 1, 2, 3, 5, 8, 13, 21,… Нагадаємо, що такі числа задаються рекурентними співвідношеннями [59]: $F_{k+1}=F_k+F_{k-1},\ k=1,2,...;\quad F_0=F_1=1.$

Тоді оцінка для ε_k набуває вигляду

$$\varepsilon_k<C^{-1}(C\varepsilon_0)^{\Phi_k},\ k=2,3,... \qquad (5.34)$$

Згідно з формулою Біне

$$F_k=\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{k+1}-\left(\frac{-1+\sqrt{5}}{2}\right)^{k+1}\right].$$

Для великих k, маємо $F_k \approx \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^{k+1}$, оскільки при $k \to \infty$ $\left(\frac{-1+\sqrt{5}}{2}\right)^{k+1} \to 0$, тому з (5.34) випливає

$$\varepsilon_k < \frac{1}{C}\left(\left(C\varepsilon_0\right)^{\frac{1+\sqrt{5}}{2\sqrt{5}}}\right)^{\left(\frac{1+\sqrt{5}}{2}\right)^k} = C_1 \upsilon^{\left(\frac{1+\sqrt{5}}{2}\right)^k},$$

де $C_1 = C^{-1}$, $\upsilon = \left(C\varepsilon_0\right)^{\frac{1+\sqrt{5}}{2\sqrt{5}}} < 1$, якщо початкове наближення x_0 близьке до кореня рівняння x^*. Нерівність дає підставу стверджувати, що збіжність методу січних надлінійна і має порядок, не менший

$$p = (\sqrt{5}+1)/2 \approx 1.618.$$

Зауваження 5.5. *Метод січних, як і метод лінійної інтерполяції належать класу двокрокових ітераційних методів. З оглядом більше 200 дво- і багатокрокових ітераційних методів можна ознайомитися в монографії*[3].

5.10.4. Метод Стеффенсена[4]**.** Оскільки $f\left(x_k\right) \to 0$ при $x_k \to x^*$, то в (5.32) можна покласти $h_k = f\left(x_k\right)$, якщо x_k близьке до x^*. Тоді ітерації обчислюється згідно з формулою

$$x_{k+1} = x_k - \frac{f^2\left(x_k\right)}{f\left(x_k + f\left(x_k\right)\right) - f\left(x_k\right)}, \; k = 0,1,...$$

У цьому методі, запропонованому Стеффенсеном**,** який по суті збігається з методом Ейткена, запропонованим у 1931 р., не потрібно обчислювати значення похідної але, як і в методі Ньютона, збіжність залишається квадратичною. У швидкості збіжності метод виграє у методу січних, в якому на кожній ітерації обчислюється одне значення функції f, а в даному методі два.

5.10.5. Гібридні методи. Якщо виконуються умови теореми 5.2, то вибір початкового наближення x_0 в методах хорд і Ньютона збігається і задовольняє умову (5.9). Крім того, обчислення ітерацій за методом хорд

[3] Petkovic M. S. Multipoint Methods for Solving Nonlinear Equations / M. S. Petkovic, B. Neta, L. D. Petkovic, J. Dzunic. – Amsterdam : Elsevier, 2013. – 344 p.

[4] Steffensen J.P. Remarks on iteration // Skand. Aktuar. Tidskr., 1933. – 16. – P. 64–72.

$$\bar{x}_{k+1} = \bar{x}_k - \frac{f(\bar{x}_k)(\bar{x}_k - \bar{x}_{k-1})}{f(\bar{x}_k) - f(\bar{x}_{k-1})},\ k = 1,2,...,\ \bar{x}_0 = x_0,$$

і за методом Ньютона

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},\ k = 1,2,...,x_1 = x_0$$

забезпечує двостороннє наближення до розв'язку x^* (рис. 5.11). Отже, виконання оцінки $|x_k - \bar{x}_k| < \varepsilon$ гарантує задану точність обчислення кореня рівняння.

Ще одним підходом є поєднання глобально збіжного методу, наприклад методу половинного поділу, з локально збіжним методом Ньютона. Можна стартувати зі швидко збіжного методу Ньютона і підключити „повільний" глобально збіжний метод або навпаки. Розглянемо перший випадок і запишемо алгоритм методу Ньютона–половинного поділу.

1. Задати початкове значення x_0; $k := 0$.

2. Обчислити $\overline{x_k} = x_k - \frac{f(x_k)}{f'(x_k)}$.

3. Якщо $\left|f\left(\overline{x_k}\right)\right| < \left|f(x_k)\right|$, то $x_{k+1} := \bar{x}_k$. інакше $\bar{x}_k := \left(x_k + \bar{x}_k\right)/2$ і повернутись на початок кроку 2.

4. Перевірити алгоритм на завершення. Робота алгоритму або припиняється із $x^* :\approx x_{k+1}$, або продовжується з переходом на крок 1 і $k := k+1$.

Такий алгоритм враховує, що метод Ньютона задає локально правильний напрям, але щоб запобігти надмірному переміщенню, відбувається корекція за допомогою поділу відрізка пополам, якщо не виконується умова релаксації $\left|f\left(\overline{x_k}\right)\right| < \left|f(x_k)\right|$ на кроці 2. Алгоритм не завжди глобально збіжний, але дозволяє розширити границі застосування методу Ньютона і вести пошук коренів, якщо знаки похідних f' і f'' не знаковизначені.

Розглянемо можливі умови завершення алгоритму.

1. За апостеріорною нестрогою оцінкою $|x_{k+1} - x_k| < \varepsilon$, де ε – задана точність, надійніше взяти $\sigma\varepsilon$, $\sigma < 1$.

2. За величиною відносної похибки

$$\delta_k = \frac{|x_{k+1} - x_k|}{|x_k|} \text{ або } \delta_k = \frac{|x_{k+1} - x_k|}{|x_k| + p^{-m}},$$

якщо значення x_k близьке до нуля, $p = 2$ або 10, m характеризує точність обчислень або обчислювальну похибку, наприклад 10^{-6}.

3. За величиною нев'язки $f(x_{k+1})$: $x^* :\approx x_{k+1}$, якщо $|f(x_{k+1})| \le \varepsilon$, надійніше, $|f(x_{k+1})| < m_1\varepsilon$, де $m_1 = \min\limits_{x\in[a,b]} |f'(x)|$.

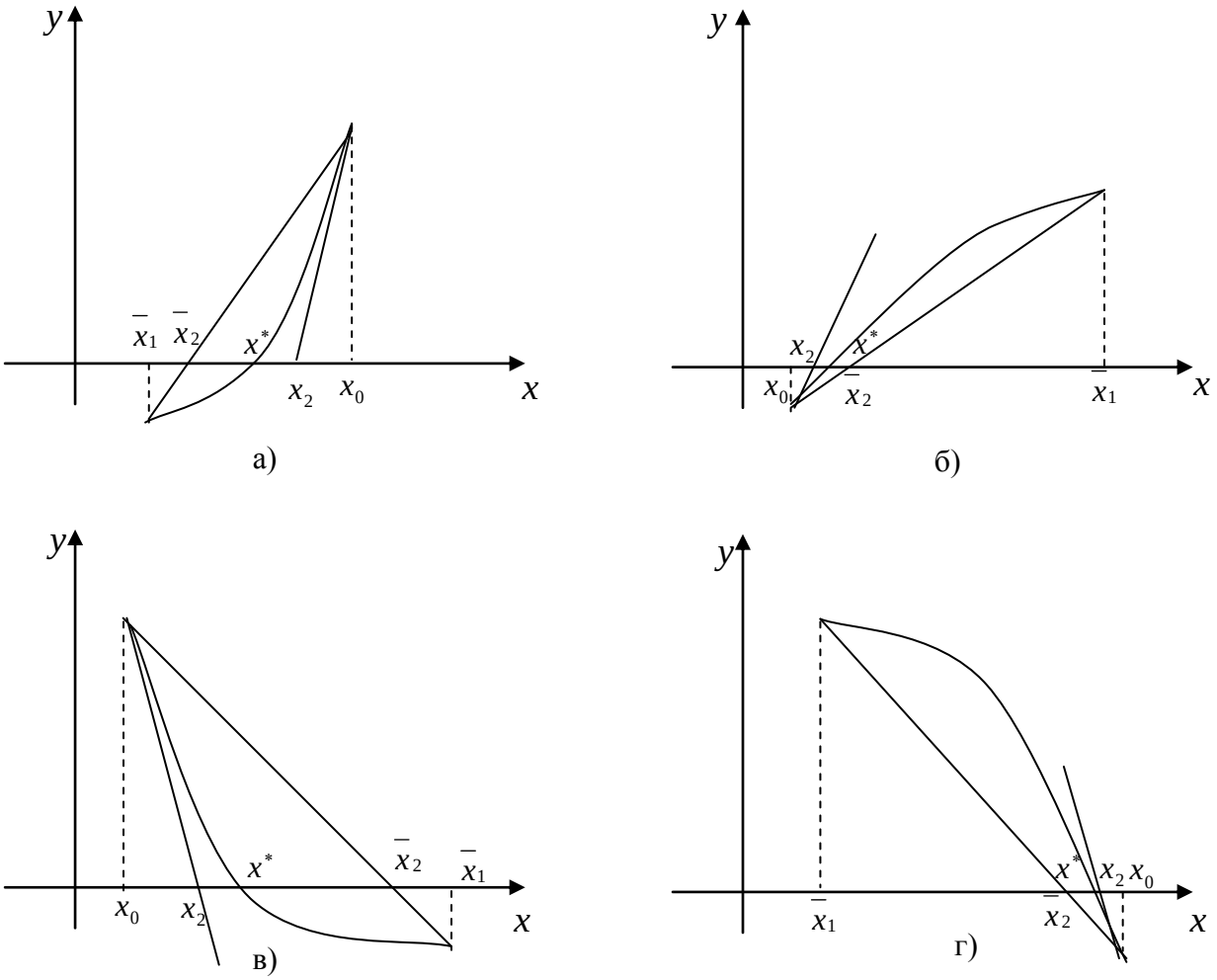


Рис. 5.12. Комбінація методу хорд і Ньютона

5.10.6. Ітераційні методи вищих порядків. Існують методи, порядок збіжності яких вищий від другого. Прикладом є метод з кубічною збіжністю

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f''(x_k) f^2(x_k)}{2\left(f'(x_k)\right)^2}, k = 0,1,...$$

Ітераційний методо, порядок збіжності якого $1+\sqrt{2}$ запропоновано у праці С.М. Шахно[5]. Побудовано ітераційні методи вищих порядків, зокрема методи до 16-го порядку точності[6], методи 4, 6 і 12-го порядку розглянуто у праці[7] .

5.11. Особливості реалізації методу Ньютона

1. Ділення на нуль. Помилку ділення на нуль $f'(x_k)=0$ легко попередити перевіркою $|f'(x_k)|>\delta>0.$ Тут δ — досить мале число, узгоджене із точністю обчислень і типом даних.

2. Відсутність коренів. Якщо значення x_k – коливаються, не наближаючись до деякого значення, то це може служити ознакою того, що рівняння $f(x)=0$ не має коренів. Наприклад, для рівняння $\sin 3x-2=0,$ яке не має дійсних коренів.

3. Випадок хибних коренів. Нехай функція f – додатна і монотонно спадна на $[a,\infty)$. Якщо $x_0>a,$ то $x_k\to\infty$ при $k\to\infty.$ Але, якщо контролювати точність за величиною нев'язки $f(x_k),$ то $f(x_k)\to 0$ при $k\to\infty.$ Тому для деякого k, значення x_k помилково можна взяти за корінь. Наприклад, для рівняння $f(x)=x^2e^{-2x}=0$ і початкового наближення $x_0=2,$ маємо в методі Ньютона

$$x_{k+1}=\frac{x_k(1-2x_k)}{2(1-x_k)}.$$

Тобто $x_1=4.0,\ \ x_2=5.333,\ \ x_3=19.724,\ldots,$ але $f(x_k)\to 0$. Тому доцільно використовувати критерій зупинки, що ґрунтується на оцінці відносної похибки $\delta_k=|x_{k+1}-x_k|/(|x_k|+10^{-m}),$ де $m>0$ таке, щоб значення 10^{-m} відповідало абсолютній похибці.

---

[5] Shakhno S.M. On an iterative algorithm with superquadratic convergence for solving non-linear operator equations // J. of Computational and Applied Mathematics**.** – V. 231, Is. 1, 1 September 2009, P. 222-235.

[6] Sharifi S., Salimib M., Siegmundb S., Lotfi T A new class of optimal four-point methods with convergence order 16 for solving nonlinear equations // Mathematics and Computers in Simulation. – 2016. – V. 231, Is. 1. − P. 69–90.

[7] Жмурко М.А., Красношлик Н.О. Порівняння багатокрокових ітераційних методів вищих порядків розв'язування нелінійних рівнянь та систем // Вісник Черкаського університету. – 2015. – № 18 (351) . − С. 1–11.

4. Зациклювання. Випадок, коли x_k повторюються або майже повторюються. Наприклад, якщо $f(x) = x^3 - x - 3$ і початкове значення $x_0 = 0$, то одержимо послідовність наближень: $x_1 = -3.000000$, $x_2 = \quad = -1.961538$, $x_3 = -1.147176$, $x_4 = -0.006579$, $x_5 = -3.000389$, $x_6 = \quad = -1.961818$, $x_7 = -1.147430,...$ Отже, $x_{k+4} \approx x_k$, $k = 0,1,...$(рис. 5.13).

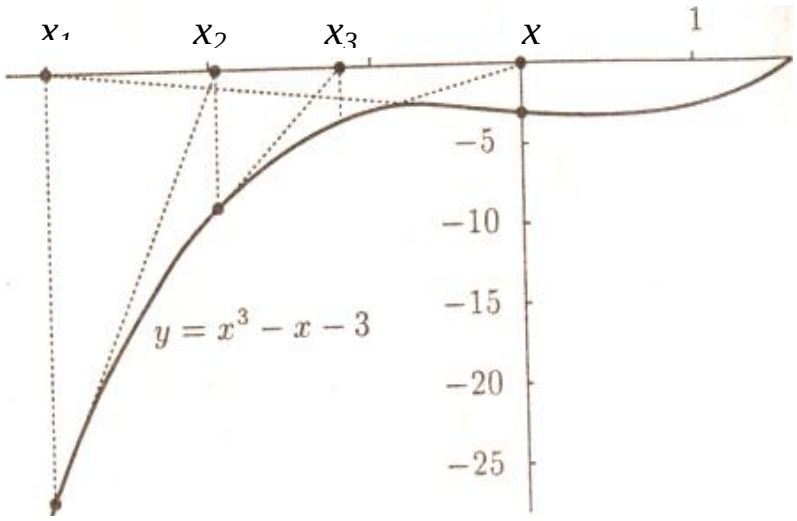


Рис. 5.13. Циклічне повторення ітерації

Приклади розв'язування типових задач

Задача 1. Відокремити корені рівняння $x^2 \ln x = 1$.

Розв'язування. Запишемо рівняння $\ln x = 1/x^2$. Графіки лівих і правих частин цих рівняння показані відповідно на рис. 5.14.

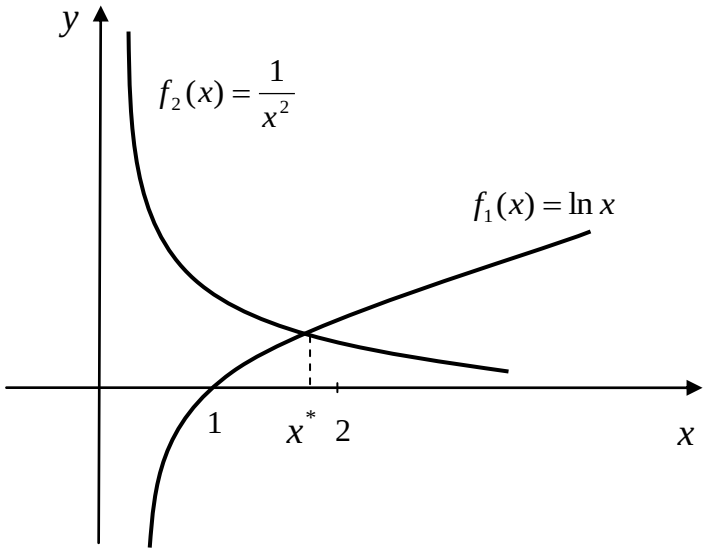


Рис. 5.14. Локалізація коренів рівняння $x^2 \ln x = 1$, $x^* \in (1, 2)$

Оскільки функція $f = x^2 \ln x - 1$ монотонна на інтервалах $(0, \frac{1}{\sqrt{e}})$ і $(\frac{1}{\sqrt{e}}, \infty)$ і $f(1) = -1 < 0$, $f(2) \approx 1.77 > 0$, то на інтервалі $(1, 2)$ функція зростає і корінь єдиний.

Задача 2. Застосувати для знаходження додатного кореня рівняння $x^2 - 2 = 0$, метод простої ітерації.

Розв'язування. Запишемо рівняння у вигляді

$$x = \varphi(x), \quad \varphi(x) = x - \lambda(x^2 - 2).$$

Параметр λ виберемо так, щоб виконувались умови теореми 5.2. Нехай $\lambda = 0.25$. Тоді для $x \in [1,2]$ $|\varphi'(x)| = |1 - 0.5x| \le q = 0.5$. Оскіль-ки похідна $\varphi'(x) = 1 - 0.5x > 0$ для $x \in (1, 2)$, то функція $\varphi(x) = = -0.25x^2 + x + 0.5$ зростає на $(1,2)$. Крім того, $\varphi(1) = 1.25$, $\varphi(2) = 1.5$. Отже, $\varphi : [1,2] \to [1.25, 1.5] \subset [1,2]$.

Для початкового значення $x_0 = 2$ наближення $x_{10} = \underline{1.4142147}$ одержується з похибкою, яка не перевищує $0.5 \cdot 10^{-6}$.

Задача 3. Проілюструвати квадратичну збіжність методу Ньютона на прикладі рівняння $x^3 = \cos x$.

Розв'язування. Графіки функцій $y = x^3$ і $y = \cos x$ перетинаються тільки в одній точці з абсцисою $x \in (0,1)$. Візьмемо $x_n = 0.5$ й обчислимо наближені значення розв'язку за формулою

$$x_{k+1} = x_k - \frac{x_k^3 - \cos x_k}{3x_k^2 + \sin x_k}.$$

Одержимо такі значення:

$$x_1 = 1.112141637097$$
$$x_2 = 0.909672693735$$
$$x_3 = 0.\underline{86}7263818208$$
$$x_4 = 0.\underline{86547}7135298$$
$$x_5 = 0.\underline{8654740331}1$$
$$x_6 = 0.\underline{865474033102}$$

Підкреслені цифри правильні, їх кількість, починаючи з четвертої ітерації, зростає, щонайменше, вдвічі.

Задача 4. Процесор не виконує операцію ділення. Застосувати метод Ньютона для обчислення $1/a$. Проілюструвати алгоритм для обчислення 1/7 із 6 десятковими правильними цифрами.

Розв'язування. Нехай $a>0$, інакше візьмемо $-a$. Розглянемо функцію $f(x)=a-1/x$. Тоді $x=1/a$ Згідно з методом Ньютона,

$$x_{k+1}=x_k-\frac{a-x_k^{-1}}{x_k^{-2}}=x_k(2-ax_k).$$

Оскільки $f''(x)=-2x^{-2}<0$, то для x_0 умова Фур'є виконується, якщо $a-x_0^{-1}<0$ або $x_0<1/a$.

Для $a=7$ візьмемо $x_0=0.125$. Ітерації набувають значень:

$$x_1=0.\underline{14}062500$$
$$x_2=0.\underline{1428}2226\ldots$$
$$x_3=0.\underline{142857}13\ldots$$
$$x_4=0.\underline{14285714}\ldots$$

На кожній з наведених ітерацій спостерігається подвоєння кількості правильних цифр, отже, з потрібною точністю $1/7\approx 0.142857$.

Завдання та запитання для самостійної роботи

1. Що означає корінь кратності m? Навести приклади рівнянь, корені яких мають кратність 3, 4 і 2.5.
2. У чому полягає лінійна і квадратична збіжність ітераційного процесу?
3. Що означає локальна і глобальна збіжність ітераційного процесу? Навести приклади глобально збіжних ітераційних процесів.
4. Як оцінити похибку наближеного розв'язку в методі простої ітерації та Ньютона?
5. Геометрична ілюстрація методу простої ітерації та методу Ньютона у випадку знакопостійності першої і другої похідної. Розглянути чотири випадки.
6. Дослідити збіжність методу січних.
7. Побудувати алгоритм ньютонівських ітерацій для обчислення $\sqrt[n]{a}$. Обґрунтувати вибір початкового значення x_0.
8. Застосувати метод Ньютона до рівняння $(x-\pi)^2+\cos x+1=0$. Довести, що ітерації збігаються лінійно, а не квадратично. Застосувати модифікований метод Ньютона (5.31) і метод Хейлі.

9. Із точністю $\varepsilon = 10^{-6}$ знайти розв'язки рівнянь:

1) $x\exp(x^2) - \sin^2 x + 3\cos x + 5 = 0$;
2) $\cos 1.3x + 1.1 = x^3$;
3) $1.5\ln x - 2\sin x + 1 = 0$;
4) $\exp(-x) + 0.25x - 0.88 = 0$;
5) $\exp(x) + \cos(\pi x)x^2 - 1 = 0$;
6) $4\cos(1.5x) - 2.1\exp(0.9x) - x = 0$;
7) $\exp(2x) - 0.8\sin(1.2x) = 0$;
8) $\ln(7.9x) = 9x - 3.1$;
9) $x = 2.5 + 3.1\exp(-x) + 1.4\exp(-2x) = 0$;
10) $x^3 - \sin^2(x) + 3\cos(x) + 5 = 0$;
11) $(x-1)^2\exp(1.2x) - 0.7 = 0$;
12) $0.9x - 3\sin(1.3x) - 0.25 = 0$;
13) $x + 1 - \exp(\sin x) = 0$;
14) $1 + \exp(-2x^2) = \exp(2x)$.

10. Довести, що корінь рівняння $(x-1)^2 e^x - 7 = 0$ та уточнити його з точністю 0.001 методами половинного поділу, Ньютона та січних.
11. Наближене значення $1/a$, $a > 0$, можна обчислити методом Ньютона, виконуючи операції віднімання і множення. Застосувати метод Ньютона з функцією $f(x) = x^{-1} - a$. Записати формулу методу Ньютона, дослідити збіжність і вказати правило вибору початкового значення. Обчислити наближені значення π^{-1}, e^{-1} і $(0.1)^{-1}$ з точністю 10^{-10}, якщо $\pi \approx 3/1415926535$, $e \approx 2.7182818284$.
12. Методом простої ітерації розв'язати рівняння Кеплера $E - \varepsilon \sin E = M$ відносно ексцентричної аномалії E, при $\varepsilon = 0,01671123$ і $M = 24.851090$. Дослідити збіжність ітераційного процесу.
13. Знайти із точністю 0.001 найменших 10 додатних значень x, для яких пряма $y = x$ перетинає графік функції $y = tgx$. Розв'язок цієї задачі використовується при визначенні максимального навантаження, яке може витримати стержень без зміни форми.
14. Нехай $\varphi \in C^1[a,b]$, $|\varphi'(x)| \geq 1$ при $x \in [a,b]$ і x^* – єдиний корінь рівняння $x = \varphi(x)$ на $[a,b]$. Показати, що метод простої ітерації розбіжний для довільного $x_0 \in [a,b] \setminus \{x^*\}$.
15. Для знаходження простого нуля функції $f \in C^4$ використовується ітераційний процес $x_{n+1} = (u_{n+1} + v_{n+1})/2$, де

$$u_{n+1} = x_n + \frac{f(x_n)}{f'(x_n)}, \quad v_{n+1} = x_n + \frac{g(x_n)}{g'(x_n)}, \quad g(x) = \frac{f(x)}{f'(x)}.$$

Довести, що для збіжного методу швидкість збіжності – кубічна.

16. Показати, що якщо x^* – корінь рівняння $f(x) = 0$ кратності p, то модифікація методу Ньютона (5.30) має квадратичну збіжність.

17. Показати, що метод Стеффенсена

$$x_{k+1} = x_k - \frac{f^2(x_k)}{f(x_k + f(x_k)) - f(x_k)}, \quad k = 0,1,...$$

володіє квадратичною збіжністю.

18. Застосувати метод Ньютона до рівняння $(x-\pi)^2 + \cos x + 1 = 0.$ Перевірити, що ітерації збігаються лінійно, а не квадратично. Застосувати модифіковані методи Ньютона (5.31) і (5.32).

19. Показати, що у випадку кратного кореня рівняння $f(x) = 0$ метод Ньютона, застосований для рівняння $F(x) = 0$, $F(x) = f(x)/f'(x)$, має квадратичну швидкість збіжності.

20. Довести, що послідовність $x_{k+1} = \varphi(x_k)$ збігається до розв'язку x^* рівняння $x = \varphi(x)$ з порядком m, якщо

$$\varphi'(x^*) = \varphi''(x^*) = ... = \varphi^{(m-1)}(x^*) = 0, \quad \varphi^{(m)}(x^*) \neq 0.$$

21. Рівняння $f(x) = 0$, корінь якого x^*, еквівалентне рівнянню $x = x + \lambda(x) f(x)$, де $0 < \lambda(x)$ – довільна неперервна функція. Вибрати $\lambda(x)$ так, щоб послідовність $x_{k+1} = x_k + \lambda(x_k) f(x_k)$ збігалась квадратично. Який метод при цьому одержується?

22. Рівняння $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3 e^{0.3x}}{6}$ має додатний корінь на інтервалі (2,3). Методом Ньютона знайти цей корінь з точністю 10^{-5}.

23. Двоетапний метод Ньютона має вигляд

$$y^{(k)} = x^{(k)} - \frac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)}, \quad x^{(k+1)} = y^{(k)} - \frac{f\left(y^{(k)}\right)}{f'\left(y^{(k)}\right)}.$$

Показати, що такий метод має кубічну збіжність і

$$\lim_{k\to\infty} \frac{x^{(k+1)} - x^*}{\left(x^{(k)} - x^*\right)^3} = \frac{1}{2}\left[\frac{f''\left(x^*\right)}{f'\left(x^*\right)}\right]^2.$$

24. Для рівняння $x = \cos x$:

а) довести збіжність методу простої ітерації для $x \in \left[0, \frac{\pi}{2}\right]$;

б) для початкових значень із цього ж відрізку довести збіжність методу Ньютона і знайти оцінку похибки для $x_{k+1} - x^*$.

25. Показати, що метод Ньютона для рівняння $arctgx = 0$, яке має єдиний корінь $x^* = 0$, збігається тоді, коли $|x_0| < \bar{x}$, де $0 < \bar{x}$ - корінь рівняння $2x = (1 + x^2) arctgx.$

Розділ 6. Алгебраїчні рівняння

Математичної моделі, які приводять до алгебраїчних рівнянь. Кількість і межі дійсних коренів. Метод Мюллера. Особливості розв'язування алгебраїчних рівнянь, басейни Ньютона.

Література [22, 38, 43, 45, 50, 73, 78, 79, 80]
Електронні джерела [103, 105–107]

6.1. Приклади алгебраїчних рівнянь

Математична модель популяційного спалаху комах зі щільністю популяції $u(t)$ має вигляд [11]

$$\frac{du}{dt} = ru\left(1-\frac{u}{q}\right)-\frac{u^2}{1+u^2},$$

де r і q – додатні сталі. Ненульові положення рівноваги в цій моделі знаходяться із кубічного рівняння

$$f(u) := r\left(1-\frac{u}{q}\right)-\frac{u}{1+u^2}=0.$$

Залежно від значення r рівняння може мати 1, 2 або 3 різні дійсні додатні корені (рис. 6.1).

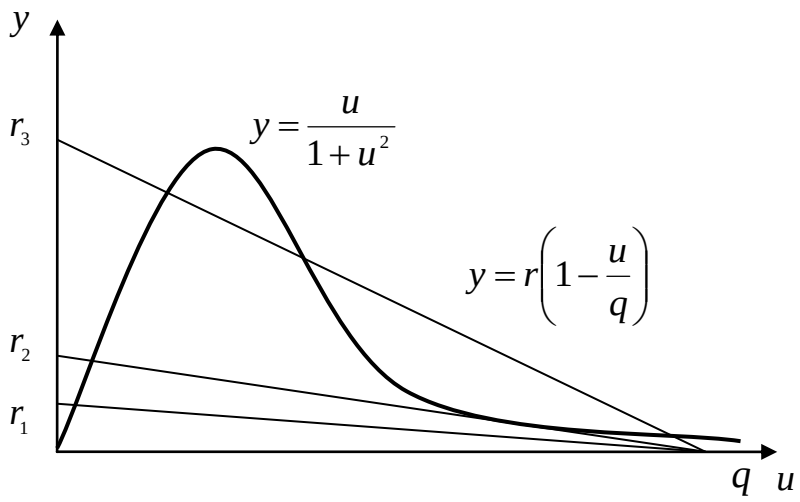


Рис. 6.1. Кількість коренів рівняння $f(u)=0$ залежно від значення r

Алгебраїчні рівняння виникають і при проведенні деяких фінансових розрахунків. Наприклад, обчислення процентної ставки при капіталовкладенні. Нехай закуповуються прилади на

10000 гривень. Припустимо, що фінанси вкладено на 12 місяців по 500 гривень/місяць, а на наступні 12 місяців – по 400. Відомо, що прилади коштуватимуть 2500 гривень у кінці цього періоду. Розв'язування полягає ось у чому: поточна ціна однієї гривні через n місяців складатиме $1/(1+s)^{-n}$, де s – невідома процентна ставка. Тому

$$10000 = \sum_{j=1}^{12} \frac{500}{(1+s)^j} + \sum_{j=13}^{24} \frac{400}{(1+s)^j} + \frac{2500}{(1+s)^{24}}.$$

Звідси отримуємо алгебраїчне рівняння порядку 24 відносно $(1+s)$

$$10000(1+s)^{24} - \sum_{j=1}^{12} 500(1+s)^{24-j} - \sum_{j=13}^{24} 400(1+s)^{24-j} - 2500 = 0.$$

Розв'язки лінійного однорідного диференціального рівняння зі сталими коефіцієнтами порядку n

$$u^{(n)} + a_1 u^{(n-1)} + \ldots + a_{n-1}\dot{u} + a_n u = 0,$$

де $u = u(x)$ – невідома функція, будуються у вигляді $u(t) = e^{\lambda t}$, де значення λ є коренями алгебраїчного рівняння степеня n

$$\lambda^n + a_1\lambda^{n-1} + \ldots + a_{n-1}\lambda + a_n = 0,$$

яке називається характеристичним і набуває вигляду

$$a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1}x + a_n,\ a_0 \neq 0.$$

При знаходженні власних значень квадратної матриці A порядку n також одержуємо алгебраїчне рівняння степеня n

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

6.2. Властивості розв'язків алгебраїчних рівнянь

6.2.1. Загальні властивості. Розглянемо задачу про знаходження нулів многочлена з дійсними коефіцієнтами вигляду

$$P_n(x) := a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1}x + a_n,\ a_0 \neq 0,$$

тобто задачу про знаходження коренів алгебраїчного рівняння

$$P_n(x) = 0. \qquad (6.1)$$

Будь-яке алгебраїчне рівняння, степінь якого не перевищує 4, розв'язується в радикалах, що можна робити в системі Mathematica [105]. Для кубічних рівнянь корені записуються за допомогою формули Кардано, але при цьому доводиться

проводити обчислення з комплексними числами. Корені рівняння (6.1) степеня $n > 4$, як довів Е. Галуа у 1830 р., у загальному випадку такого зображення не мають. У часткових випадках, наприклад, $x^n - a = 0$ або $x^{2n} + bx^n + c = 0$ розв'язок можна записати в явному вигляді.

Теорема 6.1 [38]. *Алгебраїчне рівняння степеня n має рівно n коренів, дійсних або комплексних, за умови, що кожен корінь враховується стільки разів, якою є його кратність.* ■

Нагадаємо, що корінь x^* алгебраїчного рівняння (6.1) має кратність m, $1 \le m \le n$, якщо виконуються умови

$$P_n(x^*) = P_n'(x^*) = ... = P_n^{(m-1)}(x^*) = 0,\ P_n^{(m)}(x^*) \neq 0.$$

Корені рівняння (2.1) із дійсними коефіцієнтами комплексно спряжені, тобто $x = \alpha \pm \beta i,$, де i − уявна одиниця, $\alpha, \beta \in R$. Звідси випливає, що для непарного n рівняння (6.1) має хоча б один дійсний корінь.

Для уточнення відокремлених дійсних коренів рівняння (6.1) можна застосувати метод лінійної інтерполяції (5.6), метод Ньютона (5.22), комбінований метод або інші методи, розглянені у розділі 5. Наближене значення комплексних коренів обчислюється за допомогою методу Ньютона [50] або інших спеціальних методів [22, 50, 78].

Обчислити значення многочлена $P_n(x)$ можна за схемою Горнера, виконавши n множень і n додавань. Справді, згідно з теоремою Безу для довільного c існують $b_0, b_1, ..., b_{n-1}$ такі, що

$$\begin{aligned} &a_0x^n + a_1x^{n-1} + ... + a_{n-1}x + a_n = \\ &(x-c)(b_0x^{n-1} + b_1x^{n-2} + ... + b_{n-2}x + b_{n-1}). \end{aligned} \tag{6.2}$$

Після множення многочленів у правій частині і прирівнювання коефіцієнтів при однакових степенях x одержимо:

$$b_0 = a_0, b_1 = a_1 + b_0c, ..., b_n = a_n + b_{n-1}c, \tag{6.3}$$

причому $b_n = P_n(c)$. Числа $b_0, b_1, ..., b_n$ називається коефіцієнтами Горнера. Якщо відоме значення кореня $x = c$ рівняння (6.1), то із (6.2) випливає, що $b_n = 0$ і на підставі співвідношення (6.3) можна перейти до знаходження коренів многочлена

$$b_0x^{n-1} + b_1x^{n-2} + ... + b_{n-2}x + b_{n-1},$$

виділивши лінійний множник $x - c$ згідно з (6.2).

6.2.2. Межі коренів. Розглянемо питання про межі коренів рівняння (6.1), зокрема додатних x_+ і від'ємних x_- коренів. Якщо $a_n = 0$, то $x = 0$ є коренем рівняння, і в цьому випадку переходимо до розв'язування рівняння $a_0 x^{n-1} + a_1 x^{n-2} + ... + a_{n-2} x + a_{n-1}$.

Як відомо [22, 38, 78], всі корені рівняння (6.1) лежать у кільці

$$\frac{|a_0|}{A + |a_n|} \le |x| \le 1 + \frac{B}{|a_0|}, \qquad (6.4)$$

де $A = \max(|a_0|, ..., |a_{n-1}|)$, $B = \max(|a_1|, ..., |a_n|)$.

Зауважимо, що можна обмежитися знаходженням верхньої границі R лише для додатних коренів рівняння (6.1). Для цього розглянемо допоміжні алгебраїчні рівняння:

$$x^n P_n(1/x) = 0, \quad P_n(-x) = 0, \quad x^n P_n(-1/x) = 0.$$

Нехай додатні корені рівняння

$$a_n y^n + a_{n-1} y^{n-1} + ... + a_1 y + a_0 = 0, \quad a_n \neq 0,$$

обмежені зверху числом $R_1 > 0$, тоді, згідно з (6.4), верхньою межею додатних коренів є $R_1 = 1 + \frac{A}{|a_n|}$. Отже, $y_+ = x_+^{-1} \le R_1$.

Отже, $R_1^{-1} \le x_+ \le R$.

Для знаходження меж від'ємних коренів розглянемо многочлени $P_n(-x)$ і $x^n P_n(-1/x)$. Тоді $-R_2 \le x_- \le -R_3^{-1}$, де R_2 і R_3 – нижня і верхня межі коренів рівнянь $P_n(-x) = 0$ і $P_n(-1/x) = 0$ відповідно.

Точніше, порівняно з (6.4), обчислити значення R можна згідно з теоремами Лагранжа або Ньютона.

Теорема 6.2 (Лагранжа), [22, 38]. *Нехай $a_0 > 0$ і a_k $(k \ge 1)$ – перший із від'ємних коефіцієнтів многочлена $P_n(x)$, C – найбільша з абсолютних величин від'ємних коефіцієнтів многочлена. Тоді за верхню границю додатних коренів рівняння (6.1) можна взяти число*

$$R = 1 + \sqrt[k]{\frac{C}{a_0}}. \qquad (6.5)$$

Якщо всі $a_k > 0$*, то додатних коренів немає.* ■

Теорема 6.3 (*Ньютона*), [22, 38]. *Нехай* $a_0 > 0$ *і для* $x = c > 0$ *значення* $P_n(x)$ *та похідних* $P_n'(x), ..., P_n^{(n-1)}(x)$ *невід'ємні. Тоді число* c *можна взяти за верхню межу додатних коренів.* ■

Цей результат випливає із формули Тейлора для многочлена

$$P_n(x) = P(c) + P'(c)(x-c) + \frac{P''(c)}{2}(x-c)^2 + ... + \frac{P^{(n)}(c)}{n!}(x-c)^n.$$

6.2.3. Кількість дійсних коренів многочлена. Питання полягає в знаходженні на заданому інтервалі кількості коренів рівняння (6.1) або її оцінки. Якщо $P_n(a_1)P_n(b_1) < 0$, то на інтервалі (a_1, b_1) є непарне число коренів із урахуванням їхньої кратності. Якщо ж $P(a_2)P(b_2) > 0$, то на (a_2, b_2) або немає коренів рівняння (6.1), або їх парне число. Ілюстрацію цих висновків наведено на рис. 5.1.

Деяку інформацію про кількість додатних і від'ємних коренів можна одержати з наступних теорем.

Теорема 6.4 (Декарта), [22]. *Кількість додатних коренів рівняння (6.1) із врахуванням їх кратності дорівнює числу знакозмін в системі коефіцієнтів* $a_0, a_1, ..., a_n$ *(нульові коефіцієнти не враховуються) або менше його на парне число.* ■

Інформацію про кількість від'ємних коренів одержимо, застосувавши теорему Декарта до многочлена $P_n(-x)$.

Необхідна умова того, що всі корені рівняння (6.1) дійсні, дається теоремою Гюа [22].

Теорема 6.5. *Якщо коефіцієнти рівняння (6.1) дійсні і всі його корені дійсні, то*

$$a_k^2 > a_{k-1}a_{k+1}, \quad k = \overline{1, n-1}.$$ ■

Важливий наступний наслідок з цієї теореми.

Наслідок 6.1. *Якщо для деякого* k, $1 \le k \le n-1$*, для трьох послідовних коефіцієнтів виконується нерівність* $a_k^2 \le a_{k-1}a_{k+1}$*, то існує хоча б одна пара комплексно спряжених коренів.* ■

Приклад. 6.1. Розглянемо рівняння

$$P_5(x) := x^5 - 2x^4 - 11.25x^3 + 22.5x^2 - 12.25x + 24.5 = 0, \quad (6.6)$$

коренями якого є числа $\pm i$, -3.5, 2, 3.5. Тут $A = 22.5$, $B = 24.5$, $a_0 = 1$, $a_5 = 24.5$. Згідно з (6.4), корені рівняння знаходять в кільці

$$0.02 < (22.5 + 22.5)^{-1} \le |x| \le 1 + 24.5 = 25.5.$$

На підставі теореми Лагранжа $\bar{R} = 1 + 12.25 = 13.25$, що точніше, порівняно із уже знайденим значенням $R = 25.5$. Для многочлена $\bar{P}_5(y) = 24.5y^5 - 12.25y^4 + 22.5y^3 - 11.25y^2 - 2y + 1$ маємо: $C = 12.25$, $a_0 = 24.5$, $k = 1$. Тому $R_1 = 1 + 12.25/24.5 = 1.5$. Отже, додатні корені знаходять на проміжку $(0.66,\ 13.25)$ або на $(0.66,\ 13.25)$.

Застосуємо теорему Ньютона для $C = 5$. Тоді $P_5(5) = 994.5 > 0$ і, що нескладно перевірити похідні $P_5^{(k)}(5) > 0$, $k = \overline{1,5}$. Тому за верхню межу додатних коренів можна взяти $R = 5$.

Для рівняння (6.6) система коефіцієнтів має знаки $+--+-+$. Число знакозмін 4, тому додатних коренів 4, 2 або їх немає. Для многочлена $P_5(-x)$ маємо послідовність знаків $--++++$. Отже, існує один від'ємний корінь. Оскільки $12.25^2 < 22.5 \cdot 22.5$, то з наслідку 1 випливає, що існує одна або дві пари комплексно спряжених коренів. Якщо $x = 3$, то $P_5(3) = -32.5 < 0$, тому рівняння (6.6) має два додатних корені і пару комплексно спряжених коренів.

Точніший результат можна одержати за допомогою системи Штурма [22, 38]. Для многочлена $P_n(x)$ система Штурма будується за таким правилом:

$$Q_0(x) = P_n(x), \quad Q_1(x) = P_n'(x), \quad Q_2(x), \ldots, Q_m(x),$$

де многочлен $Q_i(x)$ – остача при діленні $Q_{i-2}(x)$ на $Q_{i-1}(x)$, взята з протилежним знаком; і так доти, доки не дійдемо до $R_m(x) = const$. Зауважимо, що систему Штурма можна обчислити з точністю до додатного множника. Позначимо через $N(c)$ число змін знаків у системі Штурма при $x = c$, за умови, що нульові елементи цієї системи викреслені.

Теорема 6.6. (Штурма[1]), [22, 38, 50]. *Якщо многочлен $P_n(x)$ не має кратних коренів, $P(a) \neq 0$ і $P(b) \neq 0$, то кількість його дійсних коренів $N(a,b)$ на інтервалі $a < x < b$ дорівнює числу втрачених знакозмін у системі Штурма для $P_n(x)$ при переході від x=a до x=b, тобто*

$$N(a,b) = N(a) - N(b). \qquad \blacksquare$$

[1] Коли на лекції Штурм викладав цей результат, то здебільше говорив: " Ось теорема, ім'я якої я ношу".

Наслідок 6.2. *Якщо* $P(0) \neq 0$, *то число* N_+ *додатних і кількість* N_- *від'ємних коренів многочлена* $P_n(x)$ *відповідно дорівнюють*

$$N_+ = N(0) - N(+\infty),\ N_- = N(-\infty) - N(0).$$

Наслідок 6.3. *Для того щоб усі корені многочлена* $P_n(x)$, *який не має кратних коренів, були дійсними, необхідно і досить, щоб виконувалась умова* $N(-\infty) - N(+\infty) = n$. ■

Отже, всі корені рівняння $P_n(0) = 0,\ a_0 > 0$ будуть дійсними тоді і тільки тоді, коли:

1) система Штурма має максимальне число елементів $n+1$, тобто $m = n$;
2) виконується нерівність $Q_k(+\infty) > 0, k = \overline{1,n}$, тобто старші коефіцієнти всіх функцій Штурма $Q_k(x)$ повинні бути додатними.

6.3. Метод Мюллера

Метод Мюллера (метод парабол) призначений для знаходження всіх коренів як алгебраїчного рівняння (6.1), так і нелінійного рівняння [73]

$$f(x) = 0.$$

Цей метод нескладно реалізувати на комп'ютері і забезпечити високу швидкість збіжності та досить точно визначати прості корені многочлена. Кратні і близькі за модулем корені знаходяться дещо гірше, але точніше, порівняно з іншими методами. Усі корені многочлена визначаються послідовно один за одним.

У методі січних (5.33) за двома початковими наближеннями x_0 та x_1 і визначається наступне наближення x_2 як абсциса точки перетину осі x та прямої, що проходить через точки $(x_0, f(x_0))$ і $(x_1, f(x_1))$ (рис. 6.2). У методі Мюллера потрібно задати три початкові наближення x_0, x_1 і x_2. Наступне наближення x_3 знаходиться як абсциса точки перетину осі x з параболою, що проходить через точки $(x_0, f(x_0))$, $(x_1, f(x_1))$ та $(x_2, f(x_2))$ (рис. 6.3).

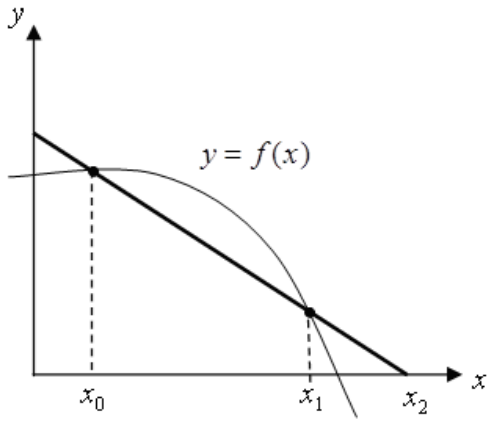


Рис. 6.2. Метод січних

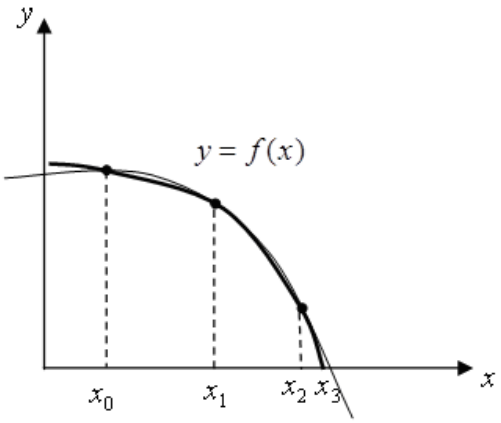


Рис. 6.3. Метод парабол

Побудуємо многочлен другого порядку

$$P_2(x) = a(x - x_2)^2 + b(x - x_2) + c,$$

графік якого проходить через точки $(x_0, f(x_0))$, $(x_1, f(x_1))$ та $(x_2, f(x_2))$. Невідомі a, b та c знаходяться з умов інтерполяції:

$$f(x_0) = a(x_0 - x_2)^2 + b(x_0 - x_2) + c,$$

$$f(x_1) = a(x_1 - x_2)^2 + b(x_1 - x_2) + c,$$

$$f(x_2) = a \cdot 0^2 + b \cdot 0 + c.$$

Із одержаної системи лінійних рівнянь маємо:

$$c = f(x_2),$$

$$b = \frac{(x_0 - x_2)^2[f(x_1) - f(x_2)] - (x_1 - x_2)^2[f(x_0) - f(x_2)]}{(x_0 - x_2)(x_1 - x_2)(x_0 - x_1)},$$

$$a = \frac{(x_1 - x_2)[f(x_0) - f(x_2)] - (x_0 - x_2)[f(x_1) - f(x_2)]}{(x_0 - x_2)(x_1 - x_2)(x_0 - x_1)}.$$

Наступне наближення x_3 знаходимо як корінь рівняння $P_2(x) = 0$ згідно з формулою

$$x_3 = x_2 - \frac{2c}{b \pm \sqrt{b^2 - 4ac}}.$$

Щоб зменшити обчислювальну похибку в методі Мюллера, яка може виникнути при відніманні досить близьких чисел, знак вибирають таким самим, як і знак коефіцієнта b. Обчислений у такий спосіб знаменник буде більшим за абсолютною величиною, тому за x_3 буде взятий той з коренів, що знаходиться ближче до x_2. Отже,

$$x_3 = x_2 - \frac{2c}{b + sign(b)\sqrt{b^2 - 4ac}}.$$

Далі процедура повторюється вже для наступної трійки чисел x_1, x_2 та x_3, щоб визначити наближення розв'язку x_4. Алгоритм завершується при досягненні заданої точності або вичерпанні максимальної кількості ітерацій.

Зауваження 6.1. *На кожній ітерації в цьому методі обчислюється корінь* $\sqrt{b^2 - 4ac}$*, тому при* $b^2 - 4ac < 0$ *маємо пару комплексних коренів.* ■

6.4. Особливості розв'язування алгебраїчних рівнянь

6.4.1. Чутливість задач до похибок. Алгебраїчне рівняння $(x-1)^{10} = 0$ має кратний корінь $x = 1$ кратності $m = 10$. Внесемо малу похибку $\varepsilon > 0$ у праву частину рівняння. Змінене рівняння $(x-1)^{10} = \varepsilon$ має корінь $\bar{x} = 1 + \sqrt[10]{\varepsilon}$. Для $\varepsilon = 10^{-10}$ похибка розв'язку складає $\Delta = 10^{-1}$, що в 10^9 разів перевищує збурення вільного члена рівняння. Для наближення $x = 1 + 10^{-4}$ маємо нев'язку $f(x) = 10^{-40}$, що не попадає в діапазон $[10^{-38}, 10^{38}]$ чисел звичайної точності.

Приклад 6.2. Розглянемо приклад Уілкінсона про нулі многочлена (1963 р., [71, 74])

$$P_{20}(x) =: (x-1)(x-2)...(x-20) = x^{20} - 210x^{19} + ... + 20!,$$

нулі якого $x = \overline{1,20}$. Внесемо похибку $2^{-23} \approx 10^{-7}$ у коефіцієнт при x^{19}. А саме, замість многочлена $P_{20}(x)$ розглянемо многочлен $\bar{P}_{20}(x) = P_{20}(x) + 2^{-23}x^{19}$. Корені рівняння $\bar{P}_{20}(x) = 0$, обчислені з усіма правильними цифрами, такі:

1.00000 0000	8.91725 0249
2.00000 0000	20.84690 8101
3.00000 0000	10.095266145 ± 0.64350 0904i
4.00000 0000	11.79363 3881 ± 1.65232 9728i
4.99999 9928	13.99235 8137 ± 2.51883 0070i
6.00000 6944	16.73073 7466 ± 2.81262 4894i
6.99969 7234	19.50243 9400 ± 1.94033 0347i
8.00726 7603	

Отже, менші за модулем корені змінилися мало, а більші значно відрізняються від коренів многочлена $P_{20}(x)$, з'явилося 5 пар комплексних коренів. Причина сильної зміни коренів не пов'язана з похибками заокруглення чи алгоритмом, а полягає в чутливості до похибок самої задачі.

6.4.2. Басейни Ньютона. Для знаходження комплексних коренів $z = x + iy$ рівняння $P_n(z) = 0$ алгоритм методу Ньютона залишається без змін:

$$z_{k+1} = z_k - \frac{P_n(z_k)}{P_n'(z_k)},\ k = 0,1,...$$

Знаходження комплексних коренів можна звести до розв'язування системи двох алгебраїчних рівнянь, якщо виділити дійсну і уявну частини функції $P_n(z)$. Тоді $P_n(x+iy) = f_1(x,y) + if_2(x,y) = 0$ й одержимо систему двох нелінійних рівнянь вигляду

$$f_1(x,y) = 0,$$
$$f_2(x,y) = 0.$$

Цікавим питанням є знаходження множини початкових значень для кожного з коренів, для яких метод Ньютона збіжний. Виявляється, що на межах цих областей початкових значень утворюються фрактали – нескінченні самоподібні структури [41]. Уперше на це звернув увагу в 1879 р. Артур Келі при узагальненні методу Ньютона на випадок уявних коренів поліномів, степінь яких перевищує два, і комплексних початкових значень[2].

Приклади фракталів показані на рис. 6.4 (заповнена множина Жюліа для відображення $f(z) = z^2 + 0.28 + 0.0113i$) і на рис. 6.5 (множина точок c на комплексній площині, для яких рекурентне співвідношення $z_n = z_n^2 + c,\ z_0 = 0,$ при всіх натуральних n задає обмежену послідовність)

Оскільки Ньютон розвинув свій метод для алгебраїчних рівнянь, то фрактали такого типу називаються фракталами, або басейнами Ньютона. Для рівнянь $z^5 - 1 = 0$ і $z^3 - 1 = 0$ басейни Ньютона показані на рис. 6.6 відповідно зліва і справа.

[2] Кроновер Р.М. Фракталы и хаос в динамических системах. Основы теории. – М.: Постмаркет, 2000. – 352 с.

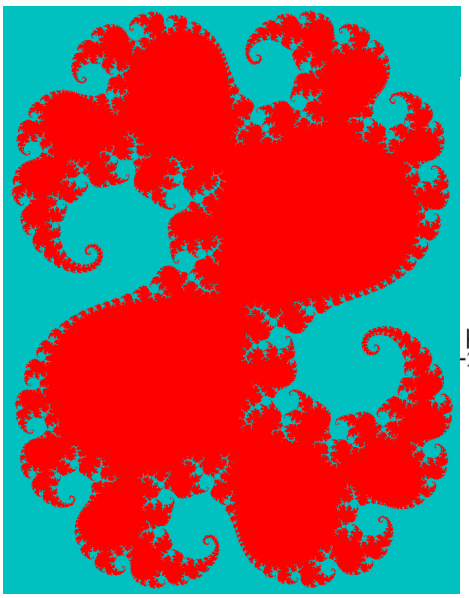


Рис. 6.4. Множина Жюліа

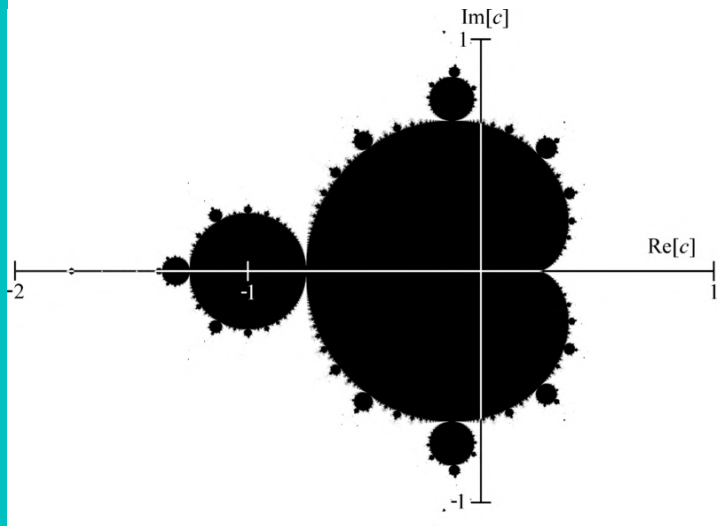
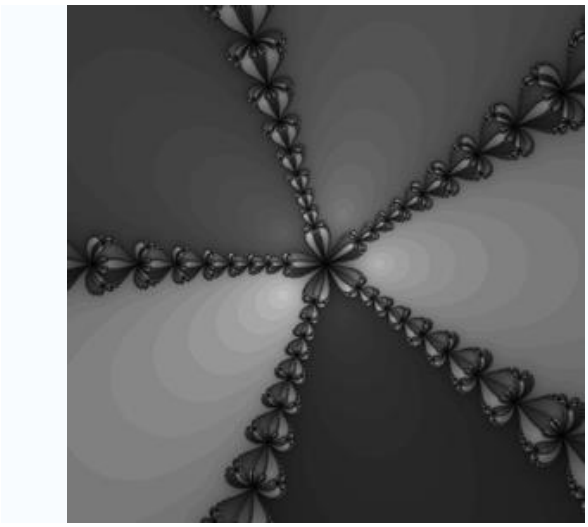


Рис. 6.5. Множина Мандельброта



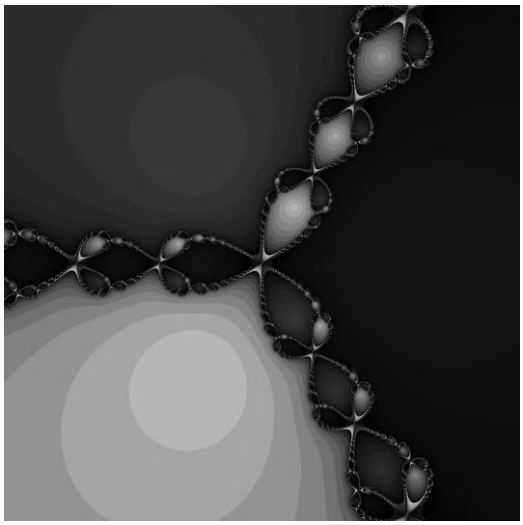


Рис. 6.6. Басейни Ньютона. Областям темнішого кольору відповідають початкові значення з більшою кількістю ітерацій

Приклади розв'язування типових задач

Задача 1. Методом Мюллера з точністю 10^{-5} знайти корені рівняння $x^4 - 3x^3 + 3x^2 - 3x + 2 = 0$ для різних початкових значень x_0, x_1, x_2, точні значення $x_{1.2} = \pm i$, $x_3 = 1$, $x_4 = 2$.

Розв'язування. Результати обчислень за формулами підрозділу 6.4 наведені в табл. 6.2.

Таблиця 6.2. Наближення коренів рівняння $x^4-3x^3+3x^2-3x+2=0$ методом Мюллера

$x_0=0, x_1=0.5, x_2=0.7$			$x_0=1.5, x_1=2.2, x_2=2.5$			$x_0=0, x_1=-0.5, x_2=-0.6$		
k	x_k	$f(x_k)$	k	k x_k	$f(x_k)$	k	x_k	$f(x_k)$
3	1.08913	-0.177492	3	2.039086	0.2094740	3	0.122+0.512i	1.219-0.887i
4	1.00192	-0.003847	4	1.994663	0.0264291	4	0.195+0.762i	0.989-0.649$\cdot i$
5	0.99998	0.000041	5	2.000118	0.0005849	5	0.124+1.059i	0.712+0.452$\cdot i$
6	0.99999	$3.842\cdot10^{-9}$	6	2.000000	$1.2341\cdot10^{-7}$	6	-0.015+0.99i	-0.089-0.051i
7	1	0	7	2	0	7	$-3.1\cdot10^{-5}$+1.0i	-0.001+0.002i
						8	$-4.9\cdot10^{-7}$+0.9i	$-1.4\cdot10^{-6}$- $-5.4\cdot10^{-6}i$
						9	i	0

Задача 2. Куля радіуса $r=10$ см занурена у воду на глибину h. Яка частина кулі буде знаходитись у воді, якщо вона виготовлена з бука, густина якого $\rho=0.620$ г/см^3?

Розв'язування. Маса витісненої води $m_в$, коли куля занурена на глибину $h, 0<h<r,$ дорівнює (рис. 6.7)

$$m_в=\int_0^h \pi(r^2-(x-r)^2\,dx=\frac{\pi h^2}{3}(3r-h).$$

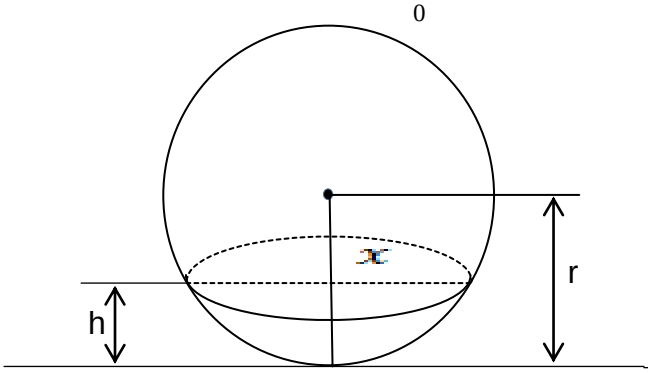


Рис. 6.7

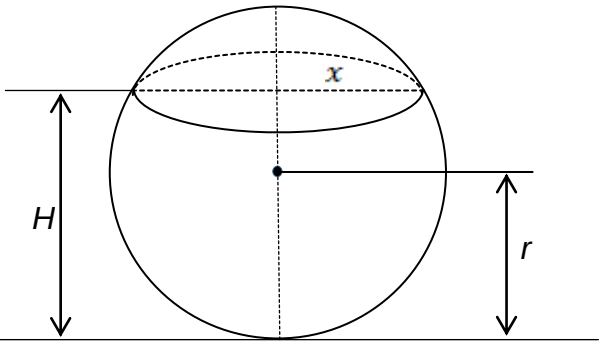


Рис. 6.8

На підставі закону Архімеда $m_в$ дорівнює масі кулі, отже,

$$\frac{\pi h^2}{3}(3r-h)=\frac{4\pi}{3}r^3\rho.$$

У підсумку одержимо рівняння

$$f(h)=h^3-30h^2+2480=0.$$

Оскільки $f(10)=480>0,\ f(12)=-9<0,\ f(14)=904>0,$ то додатні корені рівняння більші від радіуса $r=10$ і не є розв'язком задачі. Тому розглянемо випадок $H=2r-h>r, h<r$ (рис. 6.8). Тоді із закону Архімеда випливає

$$\frac{\pi h^2}{3}(r+h)=\frac{4\pi}{3}r^3\rho,$$

звідки маємо рівняння

$$g(h):=h^3+30h^2-2480=0.$$

Оскільки $g(8)=-48<0, g(10)=1520>0,$ то на відрізку [22, 50] існує єдиний розв'язок рівняння $g(h)=0.$ Єдиність випливає з того, що $g'(h)=3h(h+20)\ge 672=m_1>0,$ при $h\in[8,10].$ Друга похідна $g''(h)=6(h+10)\le 26=M_2$ і зберігає знак на [8,10]. Згідно з умовою Фур'є, за початкове значення візьмемо $h_0=10.$ Застосуємо метод Ньютона

$$h_{k+1}=h_k-\frac{g(h_k)}{g'(h_k)}=\frac{2(h_k+15)h_k^2+2480}{3h_k(h_k+20)}.$$

Результати обчислень чотирьох ітерацій наведені в таблиці

Номер ітерації	0	1	2	3	4
h_k	10	8.31	8.0755	8.071024	8.071022

Значення h_4 і h_3 збігаються з точністю до 5 десяткових цифр. Апостеріорна оцінка похибки

$$\left|x_4-x^*\right|\le\frac{M_2}{2m_1}(x_4-x_3)^2<0.2\cdot10^{-10}.$$

Отже, куля зануриться у воду на глибину $H=2r-h\approx 11.93$ см (рис. 6.6).

Задача 3. Знайти оцінку кількості та межі дійсних коренів алгебраїчного рівняння

$$P_5(x):=x^5-3.9x^4+1.37x^3+2.579x^2+0.37x+6.479=0,$$

коренями якого є числа: $\pm i, -1.1, 1.9, 3.1.$ Тут $a_0=1, a_5=6.$

Розв'язування. У коефіцієнтах многочлена $P_5(x)$ число знакозмін дорівнює 2, тому додатних коренів 2 або немає. Оскільки

$P_5(1) > 0,\ \ P_5(2) < 0,$ то додатних коренів 2. Один із них на відрізку [1,2], другий на проміжку [3,4], бо $P_5(3) < 0,$ а $P_5(4) > 0.$ У системі коефіцієнтів $P_5(-x)$ три знакозміни, але $a_4^2 = 0.37^2 < a_3 a_5 \approx 16.7,$ тому з наслідку теореми Гюа випливає, що існує пара комплексно значних коренів, а єдиний від'ємний корінь $x_- \in [-2, -1].$

Згідно з формулою (6.4), де $A = 3.9,\ B = 6.479,\ \ a_5 = 6.479,$ $a_0 = 1,$ тому корені рівняння знаходяться в кільці

$$r = 0.9 < (3.9 + 6.479)^{-1} \le |x| \le 1 + 6.479 = 7.479 = R.$$

Із теореми 6.2 випливає, що верхня межа додатних коренів $\overline{R} = 1 + 3.9 = 4.9 < R.$ Нижня межа $\overline{r} = R_1^{-1},$ де $\overline{R}_1 = 1 + \sqrt[4]{\dfrac{3.9}{6.479}} \approx 1.881$ – верхня межа додатних коренів рівняння $P_5(1/x) = 0.$ Отже $0.51 < x < 4.79.$ Аналогічно для від'ємних коренів x – на підставі рівнянь $-P_5(-x) = 0$ і $-P_5(-1/x) = 0$ одержимо –

$$-2.87 \approx 1 + \sqrt[3]{6.479} \le x_- \le -\left(1 + \sqrt{\frac{2.579}{6.479}}\right)^{-1} < -0.61.$$

Завдання та запитання для самостійної роботи

1. Які засоби розв'язування алгебраїчних рівнянь передбачено у пакетах MathCad, Mathematica і Maple?
2. Які є методи оцінки кількості додатних і від'ємних коренів алгебраїчного рівняння?
3. Навести способи знаходження меж для додатних, від'ємних і комплексних коренів алгебраїчного рівняння степеня n.
4. Навести алгоритм і дати геометричну ілюстрацію методу Мюллера.
5. У чому полягає чутливість алгебраїчного рівняння до похибок коефіцієнтів? Проілюструвати на прикладі Уілкінсона або на іншому прикладі.
6. Відокремити дійсні корені алгебраїчних рівнянь і знайти з точністю 0.001 додатний корінь:

$$1)\ x^3 - 3x - 2 = 0;$$

$$2)\ x^4 - 2x - 4 = 0;$$

3) $x^4 + 2x^3 - x - 1 = 0$;

4) $x^3 - 3x^2 + 2.9999x - 0.9999 = 0$.

7. Знайти на інтервалах $(0, \infty)$ і $(-4, 0)$ кількість коренів алгебраїчного рівняння $5x^4 - 2x^3 - 33.75x^2 + 45x - 12.25 = 0$ за допомогою системи Штурма.
8. Використовуючи один з ітераційних методів, знайти з точністю 10^{-3} додатний корінь алгебраїчного рівняння
$$x^3 + 5x^2 - 8x - 48 = 0.$$
Виділити цей корінь за схемою Горнера і знайти інші два корені.
9. Застосувати метод Ньютона для розв'язування рівняння
$$z^5 + (7 - 2i)z^4 + (20 - 12i)z^3 + (20 - 28i)z^2 + (19 - 12i)z + 13 - 26i = 0$$
з початковим значенням $z_0 = 3i$ та точністю 10^{-6}.
10. Методом Ньютона з точністю $\varepsilon = 10^{-6}$ знайти додатний корінь $x^* = 2.09455148...$ кубічного рівняння[3] $x^3 - 2x - 5 = 0$.
11. Застосувати метод Ньютона для обчислення комплексних коренів рівняння $x^3 - 2x - 5 = 0$, на якому Джозеф Рафсон у 1690 р. проілюстрував цей метод.
12. Знайти проміжок одиничної довжини, на якому є від'ємний корінь рівняння $x^3 - x^2 + 4 = 0$. За скільки кроків методу половинного поділу можна уточнити цей корінь з точністю 0.1, 0.01 і 10^{-6}?
13. З точністю $\varepsilon = 0.001$ знайти всі дійсні корені рівняння
$$x^4 - 3x^2 + 75x - 10000 = 0.$$
14. У хімічній реакції $CO + \frac{1}{2}O_2 \leftrightarrow CO_2$ процентний вміст x дисоційованого моля CO_2 визначається з рівняння $\left(P / K^2 - 1\right)x^3 + 3x - 2 = 0$, де P – тиск в атмосфері, K – стала рівноваги. Знайти x точністю $\varepsilon = 0.001$, якщо $K = 1.648$ і $P = 1$.
15. Многочлен $P_4(x) = x^4 - 12x^3 + 46x^2 - 60x + 25$ має кратний корінь $x = 1$. З точністю $\varepsilon = 10^{-3}$ обчислити цей корінь різними методами.
16. Перевірити, що метод Ньютона для рівняння $x^3 - 2x + 2 = 0$ зациклюється, якщо початкове значення $x_0 = 0$.
17. У 1225 р. Фібоначчі для рівняння $f(x) = x^3 + 2x^2 + 10x - 20 = 0$ знайшов наближене значення розв'язку $\bar{x} = 1.368808107$, усі цифри якого правильні.

[3] На цьому рівнянні Ісаак Ньютон практикував розроблений ним метод.

а) Застосувати для розв'язування рівняння метод простої ітерації з початковим значенням $x_0 = 1$ і функцією $\varphi(x) = 20/(x^2 + 2x + 10)$ для розв'язування рівняння. На якій ітерації досягається значення $\bar{x}$?

б) З точністю $\varepsilon = 10^{-10}$ знайти розв'язок рівняння методом Ньютона.

18. Рівняння $x^3 - 1 = 0$ має корені $x_1^* = 1$, $x_{2,3}^* = \left(-\sqrt{3} \pm i\right)/2$. Проаналізувати збіжність методу Ньютона для початкових значень $x_{0,k} = \left(\sqrt{3} + 0{,}1k \pm (1 + 0{,}1k)i\right)/2$ та $x_{0,k} = 0{,}1k$ для $k = \pm 1, \pm 2, \pm 3$, використавши в кожному випадку 5 ітерацій.

19. Рівняння $x^5 + 8x^4 + 17x^3 - 8x^2 - 14x + 20 = 0$ має корені -1 і $+2$. Якщо в методі Ньютона за початкове наближення взяти $x_0 = -0.3$, то ітераційний процес збіжний до кореня $+5$. Пояснити чому.

20. Знайти з точністю до третього знаки після крапки корінь рівняння $x^3 - 0.39x^2 - 10.5x + 11.0 = 0$, який лежить на інтервалі $(2,3)$. Якщо коефіцієнти відомі з відносною похибкою 2%, то яка верхня межа можливої похибки при обчисленні кореня?

21. Методом Ньютона з точністю 0.0001 уточнити два дійсні та два комплексні корені рівняння

$$x^4 - x - 10 = 0.$$

22. Методом Ньютона з точністю $\varepsilon = 10^{-4}$ знайти додатне положення рівноваги моделі популяційного спалаху комах

$$\dot{u} = ru\left(1 - \frac{u}{q}\right) - \frac{u^2}{1 + u^2},\ r = 1,\ q = 2.55.$$

23. Методом Хейлі (5.32) з точністю $\varepsilon = 10^{-4}$ знайти кратні корені рівняння

$$x^4 - 0.3x^3 - 2.7275x^2 - 0.4125x + 1.890625 = 0.$$

24. З точністю $\varepsilon = 10^{-4}$ знайти хоча б один корінь таких рівнянь:

1) $x^4 + 7x^3 + 3x^2 + 4x - 7 = 0$;

2) $x^4 + 5x^3 + 5x^2 - 5x - 6 = 0$;

3) $x^5 + x^4 + 2x^2 - x - 2 = 0$.

25. Перевірити, чи два корені рівняння

$$x^4 - 5x^3 - 12x^2 - 76x - 79 = 0$$

є дійсними і методом Ньютона знайти їх з точністю 10^{-6}. Порівняти кількість ітерацій, виконаних для досягнення заданої точності.

Розділ 7. Системи нелінійних рівнянь

Приклади систем. Методи простої ітерації та Зейделя. Нелінійні методи Якобі та Ґауса–Зейделя. Метод Ньютона для системи двох і n рівнянь. Різницеві методи та метод Бройдена розв'язування системи нелінійних рівнянь. Комбіновані методи. Методи градієнтного та покоординатного спуску.

Література [12, 16, 22, 28, 30, 33, 45, 49, 59, 70, 73, 77, 100]
Електронні джерела [103, 105–107]

7.1. Приклади систем нелінійних рівнянь

Сучасні математичні моделі, як правило, нелінійні. Тому розв'язувати системи нелінійних рівнянь доводиться досить часто. Системи нелінійних рівнянь одержуються при дослідженні на екстремум функції $y = g(x_1,...,x_n)$ двох або більше змінних. Якщо існують неперервні частинні похідні, то необхідна умова екстремуму записується у вигляді системи рівнянь

$$\frac{\partial g}{\partial x_i}(x_1,...,x_n) = 0, \quad i = \overline{1,n}.$$

Положення рівноваги динамічної системи

$$\frac{dx_i}{dt} = \varphi_i(x_1,...,x_n), \quad i = \overline{1,n},$$

є розв'язками системи рівнянь

$$\varphi_i(x_1,...,x_n) = 0, \quad i = \overline{1,n}.$$

Системи нелінійних рівнянь виникають при побудові квадратурних формул Ґауса (підрозділ 11.8).

При розв'язуванні задачі про відбиття звукової хвилі на межі поділу двох середовищ одержується система чотирьох нелінійних рівнянь [43]

$$a_1\cos\varphi_1 - a_2\cos\varphi_2 + a = 0,$$
$$a_2\sin\varphi_2 + a_2\sin\varphi_2 = 0,$$
$$\frac{r_1}{p_1}(a_1\cos\varphi_1 - a) - \frac{r_2}{p_2}a_2\cos\varphi_2 = 0,$$
$$\frac{r_1}{p_1}a_1\sin\varphi_1 + \frac{r_2}{p_2}a_2\sin\varphi_2 = 0$$

відносно невідомих амплітуд a_1 і a_2 та фаз φ_1 і φ_2 відбитої хвилі і хвилі, яка розповсюджується у другому середовищі. Параметри a, r_1, r_2, p_1 і p_2 – вважаються відомими.

У моделі взаємодії двох популяцій, наприклад популяцій "хижак" і "жертва" [42], може виникнути питання про величини популяцій, коли швидкості їх зміни дорівнюють $\dot{v}_0$ і $\dot{u}_0$. Одержимо систему нелінійних рівнянь

$$\left(\beta_1 - \gamma_1 v - \delta_1 u\right)u = \dot{u}_0,$$
$$\left(-\beta_2 + \gamma_2 u - \delta_2 v\right)v = \dot{v}_0,$$

$\beta_i, \gamma_i, \delta_i$ – невід'ємні сталі.

7.2. Методи простої ітерації та метод Зейделя

Розглянемо систему n рівнянь з n невідомими $x_1,...,x_n$ вигляду

$$f_i(x_1,...,x_n) = 0,\ i = \overline{1,n}, \tag{7.1}$$

де f_i – задані функції, визначені у деякій області $G \subseteq R^n$. Нехай у цій області існує розв'язок, тобто такий вектор $x^* = (x_1^*,...,x_n^*)^T$, що $f(x^*) = 0$. Увівши позначення $x = (x_1,...,x_n)^T$, $f = (f_1,...,f_n)^T$, де T – символ транспонування, запишемо систему рівнянь (7.1) у векторній формі

$$f(x) = 0. \tag{7.2}$$

Нехай система рівнянь (7.2) зведена до рівносильної системи вигляду

$$x = \varphi(x), \tag{7.3}$$

де вектор-функція $\varphi : G \to G$. Наприклад, записавши її так

$$x = x - \lambda f(x),$$

де $\lambda > 0$ – деякий параметр або скалярна функція $\lambda(x) \neq 0$, або невироджена матриця.

Алгоритм методу простої ітерації, як і у випадку скалярного рівняння, полягає ось у чому. Задається початкове наближення $x^{(0)} \in G$. Наступні наближення будуються згідно з формулою

$$x^{(k+1)} = \varphi(x^{(k)}),\ k = 0,1,... \tag{7.4}$$

У координатній формі метод набуває вигляду

$$x_i^{(k+1)} = \varphi_i(x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}), \quad i = \overline{1,n};\ k = 0,1,...$$

Умови збіжності методу випливають із теореми про стискаючі відображення [34], аналогічно як для скалярного рівняння (теорема 5.2). Нехай $\|\cdot\|$ – деяка векторна норма.

Теорема 7.1. *Нехай G – замкнена випукла область у просторі R^n, відображення $\varphi: G \to G$ стискаюче, тобто існує таке $q \in (0,1)$, що*

$$\| \varphi(x_2) - \varphi(x_1) \| \le q \| x_2 - x_1 \| \quad \forall x_1, x_2 \in G. \tag{7.5}$$

Тоді система рівнянь (7.3) має єдиний розв'язок $x^* \in G$, який для довільного $x^{(0)} \in G$ є границею послідовності наближень (7.4) при $k \to \infty$. Крім того, виконується оцінка

$$\| x^{(k+1)} - x^* \| \le \frac{q}{1-q} \| x^{(k+1)} - x^{(k)} \|. \tag{7.6}$$

Замість умови стиску (7.5) на практиці можна перевірити зручнішу умову. Нехай у G існують неперервні частинні похідні $\frac{\partial \varphi_i}{\partial x_j}$, $i, j = \overline{1,n}$, і для матриці Якобі

$$J(x) = \begin{bmatrix} \frac{\partial \varphi_1(x)}{\partial x_1} & \cdots & \frac{\partial \varphi_1(x)}{\partial x_n} \\ & \cdots & \\ \frac{\partial \varphi_n(x)}{\partial x_1} & \cdots & \frac{\partial \varphi_n(x)}{\partial x_n} \end{bmatrix}$$

виконується оцінка

$$\| J(x) \| \le q < 1, \tag{7.7}$$

з якої і випливає умова стиску (7.5).

У методі Зейделя використовуються вже знайдені компоненти $(k+1)-$го наближення і алгоритм набуває вигляду:

$$\begin{aligned} x_1^{(k+1)} &= \varphi_1(x_1^{(k)}, x_2^{(k)}, ..., x_{n-1}^{(k)}, x_n^{(k)}), \\ x_2^{(k+1)} &= \varphi_2(x_1^{(k+1)}, x_2^{(k)}, ..., x_{n-1}^{(k)}, x_n^{(k)}), \\ &\cdots \\ x_n^{(k+1)} &= \varphi_n(x_1^{(k+1)}, x_2^{(k+1)}, ..., x_{n-1}^{(k+1)}, x_n^{(k)}). \end{aligned} \tag{7.8}$$

Умови теореми 7.1 забезпечують лінійну швидкість збіжності і методу Зейделя (7.8).

7.3. Нелінійні методи Якобі та Ґауса–Зейделя

У нелінійному методі Якобі знаходження розв'язку системи рівнянь (7.1) зводиться до послідовного розв'язування n нелінійних скалярних рівнянь

$$\begin{aligned} &f_1(x_1^{(k+1)}, x_2^{(k)}, ..., x_{n-1}^{(k)}, x_n^{(k)}) = 0, \\ &f_2(x_1^{(k)}, x_2^{(k+1)}, ..., x_{n-1}^{(k)}, x_n^{(k)}) = 0, \\ &\quad ... \\ &f_n(x_1^{(k)}, x_2^{(k)}, ..., x_{n-1}^{(k)}, x_n^{(k+1)}) = 0. \end{aligned} \tag{7.9}$$

Для розв'язання скалярних рівнянь можна застосувати один з ітераційних методів.

У методі Ґауса–Зейделя при розв'язуванні i-го рівняння, $i \geq 2$ використовується вже знайдені наближення $x_1^{(k+1)}, ..., x_{i-1}^{(k+1)}$ і метод набуває вигляду

$$\begin{aligned} &f_1(x_1^{(k+1)}, x_2^{(k)}, x_3^{(k)}, ..., x_{n-1}^{(k)}, x_n^{(k)}) = 0, \\ &f_2(x_1^{(k+1)}, x_2^{(k+1)}, x_3^{(k)}, ..., x_{n-1}^{(k)}, x_n^{(k)}) = 0, \\ &. . . . . \\ &f_n(x_1^{(k+1)}, x_2^{(k+1)}, x_3^{(k+1)}, ..., x_{n-1}^{(k+1)}, x_n^{(k+1)}) = 0. \end{aligned}$$

7.4. Метод Ньютона

7.4.1. Метод Ньютона для системи двох рівнянь. Розглянемо спочатку систему двох рівнянь

$$\begin{aligned} f_1(x_1, x_2) = 0, \\ f_2(x_1, x_2) = 0. \end{aligned} \tag{7.10}$$

Де функції f_1, f_2 – визначені і неперервні разом з першими частинними похідними в області $G \subseteq R^n$ і в деякому околі точки $x^{(0)} = (x_1^{(0)}, x_2^{(0)})^T \in G$ існує єдиний розв'язок системи рівнянь (7.10). Зберігши члени нульового і першого порядку в розкладі функцій f_1 та f_2 за формулою Тейлора у точці $x^{(0)}$, одержимо

$$f_1(x^{(0)}) + \frac{\partial f_1(x^{(0)})}{\partial x_1}(x_1 - x_1^{(0)}) + \frac{\partial f_1(x^{(0)})}{\partial x_2}(x_2 - x_2^{(0)}) \approx 0,$$

$$f_2(x^{(0)}) + \frac{\partial f_2(x^{(0)})}{\partial x_1}(x_1 - x_1^{(0)}) + \frac{\partial f_2(x^{(0)})}{\partial x_2}(x_2 - x_2^{(0)}) \approx 0.$$

Запишемо точні рівності, замінивши $x_i - x_i^{(0)}$ на $\Delta_i^{(0)}$, $i = 1,2$. Тоді для $\Delta_1^{(0)}$ і $\Delta_2^{(0)}$ одержимо систему лінійних рівнянь

$$\begin{aligned} \frac{\partial f_1(x^{(0)})}{\partial x_1}\Delta_1^0 + \frac{\partial f_1(x^{(0)})}{\partial x_2}\Delta_2^0 &= -f_1(x^{(0)}), \\ \frac{\partial f_2(x^{(0)})}{\partial x_1}\Delta_1^0 + \frac{\partial f_2(x^{(0)})}{\partial x_2}\Delta_2^0 &= -f_2(x^{(0)}) \end{aligned} \tag{7.11}$$

або в матричній формі

$$J(x^{(0)})\Delta^0 = -f(x^{(0)}), \tag{7.12}$$

де

$$J(x) = \begin{bmatrix} \dfrac{\partial f_1(x)}{\partial x_1} & \dfrac{\partial f_1(x)}{\partial x_2} \\ \dfrac{\partial f_2(x)}{\partial x_1} & \dfrac{\partial f_2(x)}{\partial x_2} \end{bmatrix}.$$

Якщо $\det J(x^{(0)}) \neq 0$, то існує єдиний розв'язок $\Delta_1^{(0)}, \Delta_2^{(0)}$ системи (8.14) і за перше наближення в методі Ньютона береться

$$x_1^{(1)} = x_1^{(0)} + \Delta_1^{(0)}, \quad x_2^{(1)} = x_2^{(0)} + \Delta_2^{(0)}. \tag{7.13}$$

Аналогічно будуються наступні наближення

$$x^{(k+1)} = x^{(k)} + \Delta^{(k)}, \ k = 1,2,...,$$

де $\Delta^{(k)}$ є розв'язком системи лінійних рівнянь вигляду (7.12)

$$J(x^{(k)})\Delta^{(k)} = -f(x^{(k)}).$$

7.4.2. Метод Ньютона для системи *n* рівнянь. Розглянемо систему рівнянь (7.1). Припустимо, що елементи матриці $J(x)$ неперервні в G і $\det J(x^{(0)}) \neq 0$, де $J(x)$ – матриця Якобі для вектор-функції f, $x^{(0)}$ – початкове наближення із G.

Наведемо алгоритм методу Ньютона.

1. Задати початкове наближення $x^{(0)}$, $k := 0$.
2. Обчислити $f(x^{(k)})$.

3. Обчислити матрицю Якобі $J(x^{(k)})$.

4. Якщо $\det J(x^{(k)}) \neq 0,$ то розв'язати СЛАР

$$J(x^{(k)})\Delta^{(k)} = -f(x^{(k)}),$$

інакше припинити ітерації або змінити $x^{(0)}$.

5. Обчислити наступне наближення $x^{(k+1)} = x^{(k)} + \Delta^{(k)}$.

6. Завершити ітераційний процес, якщо виконується заданий критерій точності або вичерпано вказану кількість ітерацій, $x^* := x^{(k+1)}$, інакше, $k := k+1$ і повернутись на виконання кроку 1.

Знаходження одного наближення вимагає обчислення n значень функцій $f_i(x)$ і n^2 частинних похідних і розв'язування системи лінійних рівнянь порядку n.

Приклад 7.1. Застосуємо метод Ньютона для знаходження одного з наближених розв'язків системи нелінійних рівнянь

$$\begin{aligned} x_1^2 + x_2^2 + x_3^2 &= 1, \\ x_1^2 + x_2^2 + x_3 &= 0, \\ x_1^2 + x_2 + x_3^2 &= 0. \end{aligned} \qquad (7.14)$$

Результати обчислень ітераційних наближень згідно за методом Ньютона із точністю 10^{-5} наведені в табл. 7.1. Початкове наближення $x^{(0)} = (-0.3, -0.3, -0.3),$ матрицею Якобі

$$J(x^{(k)}) = \begin{bmatrix} 2x_1^{(k)} & 2x_2^{(k)} & 2x_3^{(k)} \\ 2x_1^{(k)} & 2x_1^{(k)} & 1 \\ 2x_1^{(k)} & 1 & 2x_3^{(k)} \end{bmatrix}$$

Таблиця 7.1.

Метод Ньютона для системи рівнянь (7.14)

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$\|\Delta^{(k)}\|_\infty$	$\|f(x^{(k)})\|_\infty$
0	-0,3	-0,3	-0,3	0,45416667	0,73
1	-0,7541667	-0,68125	-0,68125	0,21833125	0,49697048
2	-0,53583542	-0,61972553	-0,61972553	0,04763502	0,05523905
3	-0,48820040	-0,61803527	-0,61803527	0,00232656	0,00227480
4	-0,48587384	-0,61803399	-0,61803399	0,00000497	0,00000541
5	-0,48586827	-0,61803399	-0,61803399		$3{,}27\cdot10^{-11}$

7.5. Метод Бройдена

Замість обчислення на кожній ітерації n^2 елементів матриці Якобі в методі Ньютона обчислимо n значень функцій $f_i(x^{(k)})$ і наближено значення частинних похідних, замінивши їх значеннями різницевих похідних вигляду

$$\frac{\partial f_i}{\partial x_j}(x) \approx \frac{f_i(x + e_j h) - f_i(x)}{h}, \qquad (7.15)$$

де h – крок, e_j – вектор, j-та компонента якого дорівнює 1, а всі інші – нулі. Розглянемо аналог методу січних (5.33), який для скалярних рівнянь має швидкість збіжності $(\sqrt{5}+1)/2$. Наслідком використання формул (7.15) є надлінійна швидкість збіжності замість квадратичної, тобто $\lim\limits_{k\to\infty} \frac{\| x^{(k+1)} - x^* \|}{\| x^{(k)} - x^* \|^p} = 0, \quad p > 1$, де x^* – точний розв'язок системи рівнянь (7.1).

Надлінійною швидкістю збіжності володіє метод, відомий як метод Бройдена [70, 77]. Кількість арифметичних операцій на кожній ітерації у цьому методі складає $O(n^2)$ і він є найбільш вдалим перенесенням методу січних на системи нелінійних рівнянь. Задамо початкове наближення $x^{(0)}$ й обчислимо наступне наближення $x^{(1)}$. У методі січних похідна $f'(x_1)$ замінюється різницевою похідною $(f(x_1) - f(x_0))/(x_1 - x_0)$. Для систем рівнянь така різниця не визначена, тому замінимо матрицю $J(x^{(1)})$ матрицею A_1, що задовольняє умову

$$A_1(x^{(1)} - x^{(0)}) = f(x^{(1)}) - f(x^{(0)}). \qquad (7.16)$$

Відомо [70], що матрицю A_1 можна у єдиний спосіб записати у вигляді

$$A_1 = J(x^{(0)}) + \frac{[f(x^{(1)}) - f(x^{(0)}) - J(x^{(0)})(x^{(1)} - x^{(0)})](x^{(1)} - x^{(0)})}{\| x^{(1)} - x^{(0)} \|_2^2}.$$

Тоді наближення $x^{(2)}$ обчислюється за формулою

$$x^{(2)} = x^{(1)} - A_1^{-1} f(x^{(1)}).$$

Аналогічно, замінивши $x^{(0)}$ на $x^{(1)}$, $x^{(1)}$ на $x^{(2)}$ і на місце $A_0 = J(x^{(0)})$, записавши A_1, побудуємо наближення $x^{(3)}$ і т. д. У загальному випадку

$$A_i = A_{i-1} + \frac{y_i - A_{i-1}s_i}{\| s_i \|_2^2} s_i^T, \qquad (7.17)$$

$$x^{(i+1)} = x^{(i)} - A_i^{-1} f(x^{(i)}),\ i = 1,2,..., \qquad (7.18)$$

де $y_i = f(x^{(i)}) - f(x^{(i-1)})$ і $s_i = x^{(i)} - x^{(i-1)}$.

Матрицю A_i^{-1} у формулі (7.18) можна не обчислювати, натомість розв'язувати систему лінійних рівнянь

$$A_i s_{i+1} = -f(x^{(i)}), \qquad (7.19)$$

після чого знайти $x^{(i+1)} = x^{(i)} + s_{i+1}$.

Розв'язування системи лінійних рівнянь (7.19), наприклад за методом Ґауса, вимагає $O(n^3)$ арифметичних операцій. Можна також скористатись формулою Шермана–Моррісона [70]

$$(A + xy^T)^{-1} = A^{-1} - \frac{A^{-1}xy^T A^{-1}}{1 + y^T A^{-1} x}, \qquad (7.20)$$

де A – невироджена матриця, вектори x і y такі, що $(A + xy^T)$– невироджена. Нехай $A = A_{i-1}$, $x = (y_i - A_{i-1}s_i)/\| s_i \|_2^2$ і $y = s_i$. Тоді з (7.19) і (7.20) одержується вираз для оберненої матриці

$$A_i^{-1} = A_{i-1}^{-1} + \frac{(s_i - A_{i-1}^{-1}y_i)s_i^T A_{i-1}^{-1}}{s_i^T A_{i-1}^{-1} y_i}, \quad i = 1,...,n, \qquad (7.21)$$

де $A_0 = J(x^{(0)})$. Знаходження A_i^{-1} за формулою (7.21) вимагає обчислення добутків і додавання відповідних матриць, тому обчислювальні затрати складають $O(n^2)$ арифметичних операцій.

7.6. Комбіновані методи

У таких методах здійснюється ітерація (зовнішня) одним із методів, а уточнення знайдених значень однією або невеликим числом ітерацій – іншим методом.

7.6.1. Ітерації Зейделя–Ньютона. Зовнішні ітерації виконуються методом Зейделя, а для знаходження внутрішніх ітерацій $x_i^{(k+1)}$ застосовується метод Ньютона. Найчастіше виконується

одна ітерація з початковим значенням $x_i^{(k)}$. Тоді для $x_i^{(k+1)}$ маємо систему рівнянь

$$\frac{\partial f_1}{\partial x_1}(x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)})(x_1^{(k+1)} - x_1^{(k)}) = -f_1(x_1^{(k)}, ..., x_n^{(k)}),$$

$$\frac{\partial f_i}{\partial x_i}(x_i^{(k+1)}, ..., x_{i-1}^{(k+1)}, x_i^{(k)}, ..., x_n^{(k)})(x_i^{(k+1)} - x_i^{(k)}) = -f_i(x_1^{(k)}, ..., x_n^{(k)}), \quad (7.22)$$

$i = 2, ..., n.$

Нехай $n = 2$. Тоді система (7.22) набуває вигляду

$$\frac{\partial f_1}{\partial x_1}(x_1^{(k)}, x_2^{(k)})(x_1^{(k+1)} - x_1^{(k)}) = -f_1(x_1^{(k)}, x_2^{(k)}),$$

$$\frac{\partial f_2}{\partial x_2}(x_1^{(k+1)}, x_2^{(k)})(x_2^{(k+1)} - x_2^{(k)}) = -f_2(x_1^{(k)}, x_2^{(k)}).$$

7.6.2. Ітерації Ньютона–Зейделя. Зовнішні ітерації будуються за методом Ньютона, а внутрішні – методом Зейделя. Запишемо метод Ньютона для системи рівнянь n рівнянь:

$$J(x^{(k)})\Delta^{(k)} = -f(x^{(k)}), \quad \Delta^{(k)} = x^{(k+1)} - x^{(k)}, \qquad (7.23)$$

де $J(x)$ – матриця Якобі, $x = (x_1, x_2)^T$.

Нехай $n = 2$. Система (7.23) набуває вигляду

$$\frac{\partial f_1}{\partial x_1}(x_1^{(k)})\Delta_1^{(k)} + \frac{\partial f_1}{\partial x_2}(x_1^{(k)})\Delta_2^{(k)} = -f_1(x_1^{(k)}),$$

$$\frac{\partial f_2}{\partial x_1}(x_1^{(k)})\Delta_1^{(k)} + \frac{\partial f_2}{\partial x_2}(x_1^{(k)})\Delta_2^{(k)} = -f_2(x_1^{(k)}).$$

Щоб розв’язати її методом Зейделя задамо в першому рівнянні $\Delta_2^{(k)} = 0$, а $(k+1)-$е наближення знайдемо із системи рівнянь

$$\frac{\partial f_1}{\partial x_1}(x^{(k)})\Delta_1^{(k)} = -f_1(x^{(k)}),$$

$$\frac{\partial f_2}{\partial x_1}(x^{(k)})\Delta_1^{(k+1)} + \frac{\partial f_2}{\partial x_2}(x^{(k)})\Delta_2^{(k+1)} = -f_2(x^{(k)}).$$

7.7. Градієнтні методи

7.7.1. Методи градієнтного та покоординатного спуску. Знаходження наближеного розв’язку систему нелінійних рівнянь (7.1) методом простої ітерації або методом Зейделя вимагає

зведення системи рівнянь до вигляду (7.3) так, щоб $\varphi: G \to G$ і виконувалась умова стиску. Досягнути цього вже при $n = 2$ буває досить складно, при цьому збіжність методів тільки лінійна. Метод Ньютона має квадратичну швидкість збіжності, але характер її локальний, тобто метод збіжний для початкового наближення, близького до точного розв'язку. Глобальної збіжності можна досягнути, якщо замінити розв'язування системи рівнянь (7.1) оптимізаційною задачею [5, 12]

$$\Phi(x) \to \min, \ \Phi\left(x\right) = \sum_{i=1}^{n} f_i^{\,2}(x). \qquad (7.24)$$

Нехай система (7.1) має розв'язок $x^* \in G$. Оскільки $\Phi(x) \geq 0$, то мінімум функції Φ досягається, коли $f_i(x) = 0$ і $x = x^*$ є розв'язком системи рівнянь (7.1). Навпаки, якщо $f_i(x^*) = 0$, $i = 1,...,n$, то $\Phi(x^*) = 0$ і x^* – точка мінімуму.

Послідовність наближень у градієнтних методах будується за формулою

$$x^{(k+1)} = x^{(k)} + \alpha_k \rho^{(k)}, \ k = 0,1,..., \qquad (7.25)$$

де вектор $\rho^{(k)}$ задає напрям мінімізації, а коефіцієнт α_k – величину кроку мінімізації. Отже, на кожній ітерації потрібно вибрати напрям $\rho^{(k)}$ і крок $\alpha_k > 0$ мінімізації. Найшвидше функція спадає в напрямі, протилежному до градієнта функції в цій точці. Тому в методі найшвидшого спуску за напрям береться вектор

$$\rho^{(k)} = -grad\, \Phi(x^{(k)}), \qquad (7.26)$$

де $grad\ f(x) = \left(\frac{\partial \Phi\left(x\right)}{\partial x_1}, ..., \frac{\partial \Phi\left(x\right)}{\partial x_n} \right)^T$. Оптимальною величиною кроку є таке α_k, при якому $\Phi(x^{(k)} + \alpha_k \rho^{(k)})$ досягає найменшого значення. Тобто

$$x^{(k+1)} = x^{(k)} + \alpha_k \rho^{(k)} \qquad (7.27)$$

є точкою мінімуму скалярної функції $F(\alpha) = \Phi(x^{(k)} - \alpha \rho^{(k)})$. Із необхідної умови екстремуму скалярної функції $F(\alpha)$ випливає, що $\alpha = \alpha_k$ є розв'язком рівняння

$$F'(\alpha) = 0. \qquad (7.28)$$

Отже, ітерації в методі найшвидшого спуску виконуються згідно з формулою (7.26), де вектор $\rho^{(k)}$ визначений рівністю (7.26), а коефіцієнт α_k є розв'язком рівняння (7.28). Геометрична ілюстрація показана на рис. 7.1.

У деяких випадках, аналогічно, як для спрощеного методу Ньютона, множник α_k береться одразу досить малим і сталим. Множник α_k можна зменшувати, наприклад удвічі, при переході до наступного кроку, якщо виконується умова релаксації $\Phi\left(x^{(k)} + \alpha_k p^{(k)}\right) < \Phi\left(x^{(k)}\right)$.

Критерієм зупинки ітераційного процесу може бути виконання нерівності $\| x^{(k+1)} - x^{(k)} \| \le \varepsilon$, де ε – задана точність, або $\| x^{(k+1)} - x^{(k)} \| / \| x^{(k)} \| \le \delta$, або малість нев'язки $|\Phi(x^{(k+1)})|$.

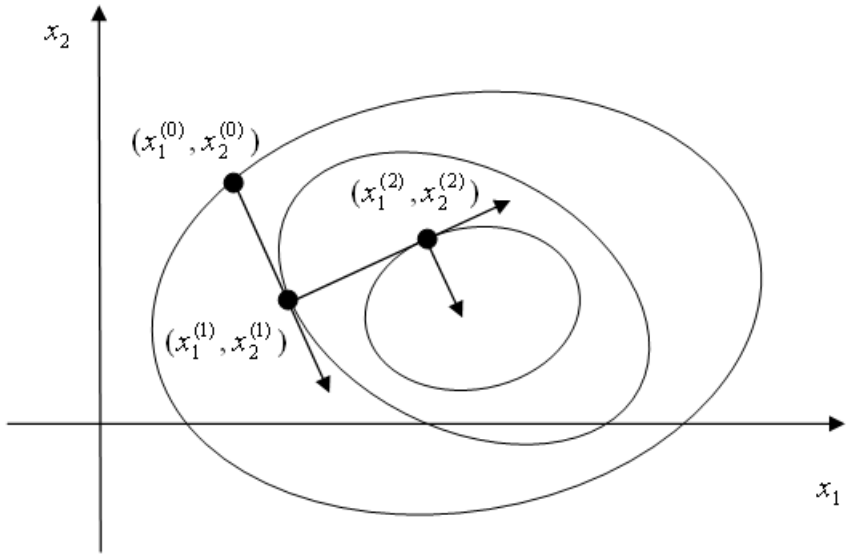


Рис. 7.1. Ілюстрація методу найшвидшого спуску

У методі покоординатного спуску задається початкове наближення $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)})$ і мінімізується функція Ф у напрямі x_1 при фіксованих інших аргументах. Відтак мінімізація відбувається в точці $(x_1^{(1)}, x_2^{(0)}, ..., x_n^{(0)})$ у напрямі x_2 і т.д. Цикл завершується мінімізацією по x_n у точці $(x_1^{(1)}, x_2^{(1)}, ..., x_{n-1}^{(1)}, x_n^{(0)})$. Наступний цикл ітерацій розпочинається з точки $(x_1^{(1)}, x_2^{(1)}, ..., x_n^{(1)})$.

Такий ітераційний процес для лінійної системи $Ax = b$ з додатно визначеною матрицею A реалізується за допомогою методу Зейделя. Процес циклічно повторюється. Послідовність

змінних, за якими здійснюється покоординатний спуск, може змінюватись у кожному з циклів. Доцільно обирати напрямок уздовж тієї осі координат, яка відповідає максимальній за абсолютною величиною частинній похідній функції $\Phi\left(x\right)$ у відповідній точці. Зокрема, номер чергової координати, за якою здійснюється спуск, може вибиратись випадково. У цьому випадку кажуть про *випадковий покоординатний спуск*.

7.7.2. Обчислення параметра спуску. У кожному з методів спуску виникає задача мінімізації функції однієї змінної. За одним зі способів за наближене значення α_k береться абсциса точки перетину дотичної до кривої

$$y = F_k(\alpha) = \Phi(x^{(k)} + \alpha \cdot grad\,\Phi(x^{(k)}))$$

у точці $(0, F_k(0))$ з віссю α. Тобто значення, одержане на першій ітерації методу Ньютона з початковим значенням $\alpha_0 = 0$. У підсумку маємо

$$\alpha_k \approx \frac{F_k(0)}{F_k'(0)} = -\frac{\Phi\left(x^{(k)}\right)}{\sum\limits_{\nu=1}^{n}\left(\dfrac{\partial\Phi\left(x^{(k)}\right)}{\partial x_\nu}\right)^2}. \qquad (7.29)$$

Тоді послідовні наближення у методі градієнтного спуску обчислюються за формулою

$$x^{(k+1)} = x^{(k)} - \frac{\Phi(x^{(k)})}{\sum\limits_{\nu=1}^{n}\left(\dfrac{\partial\Phi(x^{(k)})}{\partial x_\nu}\right)^2} grad\,\Phi(x^{(k)}).$$

Знайдене згідно з (7.29) значення α_k можна уточнити. За відомим α_k обчислимо

$$F_k\left(\frac{\alpha_k}{2}\right) = \Phi\left(x^{(k)} - \frac{\alpha_k}{2} \cdot grad\Phi\left(x^{(k)}\right)\right),$$

$$F_k\left(\alpha_k\right) = \Phi\left(x^{(k)} - \alpha_k \cdot grad\Phi\left(x^{(k)}\right)\right).$$

За значеннями функції $F_k\left(\alpha\right)$ у точках $\alpha = 0,\ \dfrac{\alpha_k}{2}$ і α_k побудуємо інтерполяційний многочлен

$$L_2(\alpha)=\frac{1}{\alpha_k^2}\big(F_k(0)(2\alpha-\alpha_k)(\alpha-\alpha_k)-$$

$$-4F_k\left(\frac{\alpha_k}{2}\right)(\alpha-\alpha_k)\alpha+F_k(\alpha_k)(2\alpha-\alpha_k)\alpha\big).$$

Тоді уточнене значення $\tilde{\alpha}_k$ знаходиться з умови $L_2'(\tilde{\alpha}_k)=0$. Оскільки

$$L_2'(\alpha)=\frac{1}{\alpha_k^2}\left(F_k(0)(4\alpha-3\alpha_k)-4F_k\left(\frac{\alpha_k}{2}\right)(2\alpha-\alpha_k)+F_k(\alpha_k)(4\alpha-\alpha_k)\right)$$

,

то одержимо

$$\tilde{\alpha}_k=\frac{3F_k(0)-4F_k\left(\frac{\alpha_k}{2}\right)+F_k(\alpha_k)}{4\left(F_k(0)-2F_k\left(\frac{\alpha_k}{2}\right)+F_k(\alpha_k)\right)}.$$

7.7.3. Оцінка градієнтних методів та їх модифікація. Перевагою градієнтних методів є глобальна збіжність. Можна довести, що процес градієнтного спуску приведе до однієї з точок мінімуму функції із довільної початкової точки. Якщо знайдений такий спосіб чином мінімум відповідає мінімуму функції $\Phi(x)$, то він буде розв'язком системи нелінійних рівнянь. Недоліком цих методів є повільна збіжність. Показано, що швидкість збіжності сповільнюється при наближенні до екстремальної точки. Градієнтний метод не дає бажаного результату, якщо $grad\Phi(x)=0$ для деякого k [25].

Доцільне є застосування гібридних алгоритмів: спочатку застосовується градієнтний метод, а в малому околі точки мінімуму розв'язок уточнити, наприклад, методом Ньютона, який володіє квадратичною збіжністю.

Розроблено ряд методів розв'язування екстремальних задач, які поєднують у собі низькі вимоги до вибору початкової точки $x^{(0)}$ і високу швидкість збіжності. До таких методів, які називаються *квазіньютонівськими*, належать, зокрема, *метод змінної метрики* (Девіда–Флетчера–Пауелла), симетричний і додатно визначений методи січних, метод спряжених градієнтів. Якщо функції $f_i(x)$ недиференційовні, то потрібно відмовитись

від використання похідних або їх апроксимації. У цьому випадку застосовуються методи прямого пошуку (циклічного покоординатного спуску, Хука і Джівса та Розенброка [25].

Приклади розв'язування типових задач

Задача 1. Застосуємо метод простої ітерації та Зейделя для наближеного розв'язування системи нелінійних рівнянь

$$\begin{aligned} x_1^2 - 2x_1 - x_2 + 0.5 = 0, \\ x_1^2 + 4x_2^2 - 4 = 0, \end{aligned} \qquad (7.30)$$

Розв'язування. Побудуємо еліпс і параболу, які визначаються першим і другим рівняннями системи відповідно. Із рис. 7.2 видно, що система (7.8) має два розв'язки (на рисунку 7.23 – координати точок M_1 і M_2). Оскільки при $x_2 = 0.5$ значення $x_1 = 0,$ то $0.5 < x_2^* < 1.0$. Для $x_1 = -0.4$ і -0.0 відповідні ординати точок на параболі $x_2 = 1.46 > 1$ і $x_2 = 0.5 < 1,$ тому точка M_1 локалізована в прямокутнику $P = [-0.4, -0] \times [0.5, 1].$

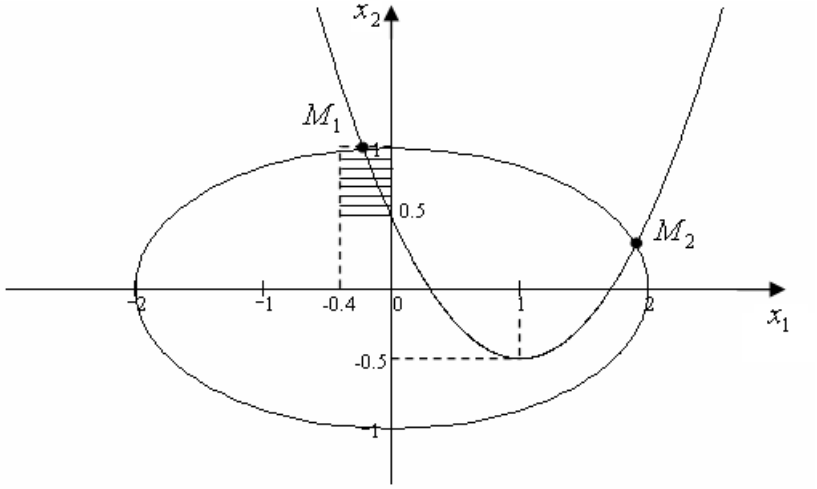


Рис. 7.2. Локалізація розв'язку в точці M_1

Для застосування методу простої ітерації запишемо систему у вигляді

$$x_1 = \frac{1}{2}\left(x_1^{\,2} - x_2 + \frac{1}{2} \right) \equiv \varphi_1(x_1, x_2),$$

$$x_2 = \frac{1}{8}\left(-x_1^{\,2} - 4x_2^{\,2} + 8x_2 + 4 \right) \equiv \varphi_2(x_1, x_2).$$

Перевіримо умову стиску відображення $\varphi = (\varphi_1, \varphi_2)^T$ у прямокутнику P. Для норми $\|\cdot\|_\infty$ маємо

$$\left|\frac{\partial \varphi_1}{\partial x_1}\right| + \left|\frac{\partial \varphi_1}{\partial x_2}\right| = |x_1| + 0.5 \le 0.4 + 0.5 = 0.9 < 1,$$

$$\left|\frac{\partial \varphi_2}{\partial x_1}\right| + \left|\frac{\partial \varphi_2}{\partial x_2}\right| = \frac{1}{4}|x_1| + |-x_2 + 1| \le 0.1 + 0.5 = 0.6 < 1.$$

Отже, $q = \max(0.9, 0.6) = 0.9$ і відображення є стискаючим. Послідовні наближення для методу простої ітерації мають вигляд

$$x_1^{(k+1)} = \frac{1}{2}\left((x_1^{(k)})^2 - x_2^{(k)} + \frac{1}{2}\right),\ x_2^{(k+1)} = \frac{1}{8}\left(-(x_1^{(k)})^2 - 4(x_2^{(k)})^2 + 8x_2^{(k)} + 4\right),$$

і методу Зейделя

$$x_1^{(k+1)} = \frac{1}{2}\left((x_1^{(k)})^2 - x_2^{(k)} + \frac{1}{2}\right),\ x_2^{(k+1)} = \frac{1}{8}\left(-(x_1^{(k+1)})^2 - 4(x_2^{(k)})^2 + 8x_2^{(k)} + 4\right).$$

Критерієм зупинки кожного з алгоритмів візьмемо виконання нерівності

$$\max\left(\left|x_1^{(k+1)} - x_1^{(k)}\right|, \left|x_2^{(k+1)} - x_2^{(k)}\right|\right) < 0.5\varepsilon,$$

де ε – задана точність. Коефіцієнт 0.5 вибраний тому, що в методі простої ітерації, згідно з (7.6), точність досягається, якщо

$$\frac{q}{1-q}\max\left(\left|x_1^{(k+1)} - x_1^{(k)}\right|, \left|x_2^{(k+1)} - x_2^{(k)}\right|\right) < \varepsilon.$$

Результати обчислень з точністю $\varepsilon = 0.001$ наведені в табл. 7.2.

Таблиця 7.2. Застосування методу простої ітерації на методу Зейделя для системи рівнянь (7.8) на мові С++

k	Метод простої ітерації			Метод Зейделя		
	$x_1^{(k)}$	$x_2^{(k)}$	$\|f(x^{(k)})\|$	$x_1^{(k)}$	$x_2^{(k)}$	$\|f(x^{(k)})\|$
0	-0,2	0,75	1,71	-0,2	0,75	1,71
1	-0,105	0,96375	0,2737186	-0,105	0,9673719	0,246347
2	-0,2263625	0,9979648	0,0349753	-0,2281734	0,9929598	0,015450
3	-0,2233624	0,9935929	0,0006805	-0,2204483	0,9939005	4.9443e-005
4	-0,2218511	0,9937431	0,0001785	-0,2226515	0,9937847	5.7066e-006
5	-0,2222626	0,9938281	0,0000206	-0,2221055	0,9938143	1.4704e-006
6	-0,2222138	0,9938059	2.0611e-005	-0,2222417	0,9938069	3.6561e-007
7	-0,2222135	0,9938084	2.6163e-007	-0.2222077	0,9938088	9.1205e-008
8	-0,2222148	0,9938085	6.0065e-007	-0.2222162	0,9938083	2.2736e-008
9	-0,2222145	0,99380849	1.2286e-007	-0.2222141	0,9938084	5.6687e-009

Задача 2. Виконати 4 ітерації в методі Ньютона для системи рівнянь (7.30).

Розв'язування. Згідно з (7.11) значення $\Delta_1^{(k)}$ та $\Delta_2^{(k)}$ визначаються із системи лінійних рівнянь

$$2(x_1^{(k)}-1)\Delta_1^{(k)} - \Delta_2^{(k)} = -((x_1^{(k)})^2 - 2x_1^{(k)} - x_2^{(k)} + 0.5),$$
$$2x_1^{(k)}\Delta_1^{(k)} - 8x_2^{(k)}\Delta_2^{(k)} = -((x_1^{(k)})^2 + 4(x_2^{(k)})^2 - 4).$$

Після чого наступне наближення обчислюється за формулами (7.13). Результати уточнення розв'язку, що відповідає точці M_1, з початковим значенням $(-0.2, 0.75)$, що є центром прямокутника, в якому локалізована точка M_1 (рис. 7.1), наведені в табл. 7.3.

Таблиця 7.3

Застосування методу Ньютона для системи (7.8)

k	$x_1^{(k)}$	$x_2^{(k)}$	$\|\Delta^{(k)}\|_\infty$	$\|f(x^{(k)})\|_\infty$
0	-0,2	0,75	-	1.71
1	-0.2385135	1.0324324	0.2824324	0.3205556
2	-0.2226170	0.9945398	0.0378926	0.0059961
3	-0.2222147	0.9938087	0.0007311	0.0000023
4	-0.2222146	0.9938084	0.0000003	$3.63\cdot10^{-12}$

Як видно, вже на четвертій ітерації збігаються шість значущих цифр наближеного розв'язку.

Задача 3. Методом найшвидшого спуску у другому квадранті знайти друге наближення для розв'язку системи нелінійних рівнянь із задачі 1.

Розв'язування. За початкове наближення візьмемо точку (-0.2, 0.75). Тоді

$$x^{(1)} = x^{(0)} + \alpha_1 \cdot grad\Phi(x^{(0)}),$$

де $\Phi(x) = (x_1^2 - 2x_1 - x_2 + 0.5)^2 + (x_1^2 + 4x_2^2 - 4)^2$,

$$grad\Phi(x) = \begin{bmatrix} 4(x_1-1)(x_1^2 - 2x_1 - x_2 + 0.5) + 4x_1(x_1^2 + 4x_2^2 - 4) \\ -2(x_1^2 - 2x_1 - x_2 + 0.5) + 8x_2(x_1^2 + 4x_2^2 - 4) \end{bmatrix}$$

Параметр α_k обчислюється за формулою

$$\alpha_k = -\frac{\Phi(x^{(k)})}{\left(\frac{\partial\Phi(x^{(k)})}{\partial x_1}\right)^2 + \left(\frac{\partial\Phi(x^{(k)})}{\partial x_2}\right)^2}.$$

Для першого наближення маємо: $\Phi(x^{(0)}) = 2.9602,\ grad\Phi(x^{(0)}) =$
$= (1.8057, -20.7651)^T$, $\alpha_0 = -2.9602 / (1.8057^2 + 20.7651^2) =$
$= -0.0068$. Отже $x^{(1)} = (-0.2, 0.75)^T - 0.0068(1.8057, -20.7651)^T =$
$(-0.2012, 0.8915)^T$. Провівши аналогічні обчислення, одержимо друге наближення
$x^{(2)} = (-0.2, 0.75)^T - 0.0068(1.8057, -20.7651)^T = (-0.2024, 0.9095)^T$.
Порівняємо з другим наближенням у методі простої ітерації – тут швидкість збіжності повільніша.

Завдання та запитання для самостійної роботи

1. Пояснити реалізацію методів простої ітерації та Зейделя для системи нелінійних рівнянь. У чому перевага програмування методу Зейделя?
2. Які достатні умови збіжності методу простої ітерації? Як можна перевірити досягнення заданої точності ε на $(k+1)$-й ітерації?
3. Пояснити алгоритми нелінійних методів Якобі і Зейделя, в яких скалярні рівняння розв'язуються методом Ньютона.
4. Пояснити реалізацію методу Ньютона для системи n нелінійних рівнянь із n невідомими. Чому зростає складність алгоритму зі зростанням n?
5. Які переваги реалізації методу Бройдена порівняно з методом Ньютона? Сформулювати кроки реалізації цього методу.
6. Порівняти складність методів Ньютона і Бройдена за кількістю обчислень значень функції на одній ітерації.
7. Чим відрізняються для системи нелінійних рівнянь методи Ньютона–Зейделя і Зейделя–Ньютона?
8. Проілюструвати методи градієнтного і покоординатного спуску при розв'язуванні системи нелінійних рівнянь.
9. Коефіцієнти A_1, A_2 і вузли x_1 і x_2 двоточкової квадратурної формули Ґауса знаходяться із системи рівнянь

$$A_1 + A_2 = 2,\ A_1x_1 + A_2x_2 = 0,$$
$$A_1x_1^2 + A_2x_2^2 = \frac{2}{3},\ A_1x_1^3 + A_2x_2^3 = 0.$$

   Розв'язати наближено систему рівнянь за допомогою методу Ньютона. Точний розв'язок такий : $x_1 = -x_2 = 1/\sqrt{3},\ \ A_1 = A_2 = 1.$

10. З точністю $\varepsilon = 10^{-3}$ методом Ньютона знайти розв'язок системи

$$\sin(2x-y)-1.2x=0.4,$$
$$0.8x^2+1.5y^2=1,$$

для якого $x>0$. *Відповідь*: $x\approx 0.491$, $y\approx -0.734$.

11. Методом простої ітерації з точністю $\varepsilon=10^{-3}$ розв'язати систему рівняь

$$\sin(x-0.6)-y=1.6,$$
$$3x-\cos y=0.9$$

Відповідь: $x\approx 0.151$, $y\approx -2.034$.

12. Методом Ньютона одержати наближені розв'язки системи рівнянь

$$x^2+y^2-2x=0,\ x^2+y^2-y=0.$$

Програма має завершувати роботу, якщо $|x_{k+1}-x_k|<\varepsilon$ і $|y_{k+1}-y_k|<\varepsilon$ або число ітерацій перевищує 20. Узяти $\varepsilon=10^{-5}$. Розглянути початкові значення (1,1), (1,-1) і (0.6,1). Проаналізувати результат і проілюструвати їх геометрично.

13. Система рівнянь

$$x^4+xy^3+y^4=1,\ x^2+xy-y^2=1$$

має 4 розв'язки. Одержати наближений розв'язок системи в околі точки (0.9,0.6) з точністю $\varepsilon=10^{-4}$ методом простої ітерації та методом Зейделя. Проаналізувати збіжність методів для таких перетворених систем

$$x=\left(1-xy^3-y^4\right)^{1/4},\ y=\left(x+\sqrt{5x^2-4}\right)/2 \text{ і}$$
$$x=\left(1-xy^3-y^4\right)^{1/4}/x^3,\ y=\left(x^2+xy-1\right)/y.$$

14. Методом Ньютона знайти наближений додатний розв'язок системи рівнянь

$$x^2+y^2+z^2=1,$$
$$2x^2+y^2-4z^2=0,$$
$$3x^2-4y^2+z^2=0.$$

15. Із точністю $\varepsilon=10^{-5}$ обчислити наближений розв'язок системи рівнянь

$$x_1=0.1x_1^2+\sin x_2,\ x_2=\cos x_1+0.1x_2^2.$$

Застосувати метод простої ітерації Зейделя і Ньютона. Розв'язок системи лежить у квадраті: $0.7\le x_1\le 0.9$, $0.7\le x_2\le 0.9$.

16. Тонка рейка довжиною 10 м. закріплена на кінцях. Рейку розрізали і приварили до неї відрізок такої ж рейки довжиною 1 м. При цьому утворена рейка набула форму дуги кола. Побудувати систему рівнянь

для наближеного обчислення Знайти значення максимального відхилення від початкового стану і радіус кола.

17. Методом найшвидшого спуску з точністю 0.01 знайти в околі початку координат наближений розв'язок системи нелінійних рівнянь

$$x + x^2 - 2yz = 0.1,$$
$$y - y^2 + 3xz = -0.2,$$
$$z + z^2 + 2xz = 0.3.$$

18. З точністю 0.001 методом Ньютона обчислити комплексні корені рівняння $x^3 - 2x - 5 = 0$, звівши задачу знаходження дійсних розв'язків системи двох рівнянь.

19. Показати, що для системи двох нелінійних рівнянь (7.10) метод Ньютона можна записати у вигляді $x = \varphi(x), \quad y = \psi(x),$ де функції φ і ψ такі:

$$\varphi = x_1 - \left(f_1 \frac{\partial f_2}{\partial x_2} - f_2 \frac{\partial f_1}{\partial x_2} \right) / \det J,$$
$$\psi = x_2 - \left(f_2 \frac{\partial f_1}{\partial x_1} - f \frac{\partial f_2}{\partial x_1} \right) / \det J,$$

$J(x)$ - матриця Якобі для системи (7.10).

20. Система нелінійних рівнянь

$$7x^3 - 10x - y = 1,$$
$$8y^3 - 11y + x = 1$$

має 9 розв'язків. За допомогою якої-небудь системи комп'ютерної математики нарисувати графіки обох ліній, вибрати початкові наближення для кожного з розв'язків і знайти їх наближені значення методом Ньютона з 9 десятковими правильними цифрами.

21. Знайти значення величини u популяції «жертва» і v популяції «хижак» за відомими швидкостями їх зміни $\dot{u} = 0.36$ і $\dot{v} = 0.1925,$ якщо динаміка взаємодії описується математичною моделлю

$$\dot{u} = (2 - v - 0.5u)u,$$
$$\dot{v} = (-0.75 + u - 0.25v)v.$$

22. Проаналізувати збіжність методу Ньютона за величинами похибки і нев'язки в нормі $\|\cdot\|_2$ для системи рівнянь

$$xy - y^3 = 1,\ x^2 y + y = 5.$$

Виконати 8 ітерацій з початковим значенням $x_0 = 2, \quad y_0 = 3$ і типом даних подвійної точності. Точний розв'язок системи $x^* = 2, \quad y^* = 1.$

Розділ 8. Наближені методи розв'язування алгебраїчної проблеми власних значень

Властивості власних значень і власних векторів матриці. LU-алгоритм розв'язування повної проблеми. QR-алгоритм розв'язування повної проблеми. Метод обертань Якобі та метод Хаусхолдера для симетричних матриць. Степеневий метод знаходження найбільшого по модулю власного значення та відповідного власного вектора для матриць загального вигляду і симетричної.

Література [5, 13, 28, 30, 43, 45, 52, 71–73, 83]
Електронні джерела [105–107]

8.1. Постановка задачі

Знаходження власних значень і власних векторів є важливою і досить складною задачею обчислювальної математики. При динамічному аналізі механічних систем власні значення відповідають власним частотам коливань, а власні вектори характеризують моди цих коливань. При розрахунку конструкцій власні значення дозволяють визначити критичні навантаження, перевищення яких веде до втрати стійкості.

Нехай A – квадратна матриця порядку n з дійсними елементами. Трактуючи A як матрицю лінійного перетворення $x \to Ax$ в просторі R^n, задача про власні значення полягає в знаходженні таких чисел λ і ненульових векторів x, для яких лінійне перетворення вектора з матрицею A змінює тільки довжину вектора x. Алгебраїчна задача полягає в знаходженні значень $\lambda \in C$, які називаються *власними значеннями*, і власних векторів $x \in R^n$.

Означення 8.1. *Число $\lambda \in C$ і ненульовий вектор x, які задовольняють рівняння*

$$Ax = \lambda x, \quad x \neq 0, \tag{8.1}$$

називаються відповідно власним значенням і власним вектором квадратної матриці A.

Означення 8.2. *Сукупність усіх власних значень, узятих із їх кратностями, називається спектром матриці та позначається через $\sigma(A)$. Число $\rho(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$ називається спектральним радіусом матриці.*

Знаходження пари (λ, x) рівносильно відшуканню нетривіальних розв'язків однорідної системи лінійних рівнянь

$$(A-\lambda I)x=0 \tag{8.2}$$

з параметром λ. Оскільки розв'язок $x\neq 0$ існує тоді і тільки тоді, коли $\det(A-\lambda I)=0$, то, обчисливши визначник

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21}-\lambda & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{vmatrix} = a_0\lambda^n + a_1\lambda^{n-1} + \dots + a_{n-1}\lambda + a_n, \quad a_0=(-1)^n,$$

для знаходження власних значень одержимо алгебраїчне рівняння степеня n

$$a_0\lambda^n + a_1\lambda^{n-1} + \dots + a_{n-1}\lambda + a_n = 0, \tag{8.3}$$

яке називається характеристичним.

Класичні методи О.М. Крилова, А.Н. Данилевського, Ж. Левер'є та ін. [22, 72, 73] якраз і полягають у побудові рівняння (8.3). Маючи його корені, із системи рівнянь (8.2) знаходимо відповідні власні вектори. При цьому потрібно обчислити коефіцієнти рівняння і знайти ненульові розв'язки n систем лінійних однорідних рівнянь (8.2). Це досить складна обчислювальна задача, тому застосовується для матриць невисокого порядку. Якщо $n\geq 4$, то розроблені спеціальні методи, орієнтовані на зведення матриці A до матриці, власні значення якої знаходяться нескладно, причому паралельно обчислюються і власні вектори. Виклад цих методів можна знайти в [20, 22, 71, 72] та ін. Розглянемо два приклади, які приводять до задачі знаходження власних значень і власних векторів.

Приклад 8.1. Нехай A – матриця порядку n. Потрібно знайти нетривіальні розв'язки системи лінійних диференціальних рівнянь

$$\frac{du}{dt} = Au, \tag{8.4}$$

де $u(t)=(u_1(t),\dots,u_n(t))^T$ – вектор-функція скалярного аргументу t. Якщо шукати розв'язок у вигляді $u(t)=e^{\lambda t}x$, де $\lambda\in C, x\in C^n$, то, підставивши його в систему рівнянь (8.4) для знаходження λ і x, прийдемо до задачі (8.1). Тобто система рівнянь (7.4) має

розв'язок заданого вигляду тоді і тільки тоді, коли λ і x є відповідно власними значеннями і власними векторами матриці A.

Приклад 8.2. Для наближеного розв'язування СЛАР $Ax = b$ порядку n із квадратною матрицею застосовується метод простої ітерації. Для цього система рівнянь зводиться до вигляду $x = Px + f$. Задавши початкове наближення $x^{(0)} \in R^n$, маємо ітераційний процес

$$x^{(x+1)} = Px^{(k)} + f, \quad k = 0,...$$

Необхідною і достатньою умовою збіжності ітераційного методу є умова того, що найбільше за модулем власне значення матриці P не перевищує 1. Отже, тут досить знати тільки одне власне значення λ, таке, що $|\lambda| = \max\limits_{i=1,...,n} |\lambda_i| < 1$.

Зауваження 8.1. Задача знаходження власних значень може бути нестійкою до похибок елементів матриці. Наприклад, для матриці A, елементи якої всі нулі крім наддіагональних $a_{i,j+1} = 1, i = \overline{1,n-1}$, усі власні значення $\lambda_i = 0, i = \overline{1,n}$. Якщо внести похибку ε в лівий нижній елемент матриці, то характеристичне рівняння $\lambda_\varepsilon^n - \varepsilon = 0$ має n коренів $\sqrt[n]{\varepsilon}$. Якщо $n = 100$, а $\varepsilon = 10^{-100} \ll 1$, то $|\lambda_\varepsilon| = 0.1 \gg \varepsilon$. ■

Задачу про власні значення можна поділити на *повну проблему*, коли потрібно знайти всі пари $\{\lambda, x\}$ з власних значень і векторів, та частинну, коли потрібно знайти найбільше або найменше по модулю власне значення і відповідний власний вектор. Зокрема, в задачах ядерної фізики, де порядок матриць перевищує 1000. У теорії коливань важливо знайти два найбільших по модулю власні значення. У прикладі 8.1 потрібно розв'язати повну, а в 8.2 – часткову проблему.

8.2. Властивості власних значень і власних векторів

1. Якщо x – власний вектор матриці A, то $\alpha x, \alpha \neq 0$, також власний вектор. Справді, $A(\alpha x) = \alpha(Ax) = \alpha\lambda x = \lambda(\alpha x)$. Зауважимо, що в деяких випадках доцільно нормувати власний вектор так, щоб його евклідова норма дорівнювала 1. Для $\alpha = 1/\|x\|$ маємо

$$\|\alpha x\| = \sqrt{\alpha^2 x_1^2 + \ldots + \alpha_n^2 x_n^2} = |\alpha| \cdot \|x\| = 1.$$

Розглянемо приклад. Для матриці

$$A = \begin{bmatrix} 2 & 2 \\ 3 & 1 \end{bmatrix}$$

власні значення $\lambda_1 = 4$ і $\lambda_2 = -1$, відповідні власні вектори $x_1 = (1,1)$ і $x_2 = (-2,-3)$. Нормовані власні вектори матриці $\bar{x}_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ і $\bar{x}_2 = (\frac{-2}{\sqrt{13}}, \frac{3}{\sqrt{13}})$.

2. Якщо $\{\lambda, x\}$ – власна пара матриці A і $\det A \neq 0$, то $\{\lambda^{-1}, x\}$ – власна пара матриці A^{-1}. Справді, матриця A невироджена, тому $x = \lambda A^{-1} x$, звідки й випливає, що $A^{-1} x = \lambda^{-1} x$. Зокрема, якщо $|\lambda| = \max |\lambda_i|$ для матриці A, то $\mu = \frac{1}{\lambda}$, $|\mu| = \min |\mu_i|$ для оберненої матриці A^{-1}.

3. Якщо A – трикутна або діагональна матриця з діагональними елементами a_{ii}, то власні значення $\lambda_i = a_{ii}, i = \overline{1,n}$.

8.3. Ортогональні матриці

Означення 8.3. *Дійсна квадратна матриця Q називається ортогональною, якщо $Q^T Q = I$.*

Прикладом ортогональної матриці є матриця обертань

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

або більш загального вигляду

$$Q_{ij} = \begin{bmatrix} 1 & & | & & & | & & & \\ & 1 & | & & & | & & & \\ - & - & c_{ij} & - & - & -s_{ij} & - & - & i \\ & & | & 1 & & | & & & \\ & & | & & 1 & | & & & \\ - & - & s_{ij} & - & - & c_{ij} & - & - & j \\ & & | & & & | & 1 & & \\ & & | & & & | & & 1 & \end{bmatrix}, \qquad (8.5)$$

i j

де $c_{ij}^2 + s_{ij}^2 = 1.$

Теорема 8.1 [19]. *Якщо* A *– дійсна симетрична матриця, то існує така ортогональна матриця* Q*, що*

$$Q^T \mathrm{A} Q = \Lambda, \tag{8.6}$$

де Λ *– діагональна матриця, діагональні елементи якої є власними значеннями матриці* A. ■

Наслідками з цієї теореми є те, що для дійсної симетричної матриці існує n лінійно незалежних власних векторів матриці A, і вони утворюють ортогональну систему, а всі її власні значні значення – дійсні числа.

При множенні матриці A зліва на матрицю Q_{ij} змінюються лише елементи *i-го* і *j-го* рядків матриці A, а при множенні на матрицю Q_{ij} справа *–i-й* та *j*-й стовпці матриці

Перетворити будь-яку дійсну невироджену матрицю у праву трикутну матрицю можна шляхом ланцюжка множень зліва на елементарні матриці обертання. При цьому всі діагональні елементи, можливо, крім останнього, додатні. У підсумку одержимо матрицю

$$B := Q_{n,n-1} \ldots Q_{21} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \ldots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \ldots & a_{2n}^{(2)} \\ \ldots & \ldots & \ldots & \ldots \\ & & \ldots & a_{nn}^{(n-1)} \end{bmatrix}.$$

5. Усі власні значення матриці A містяться у крузі $|\lambda| \le \|A\|$, де $\|\cdot\|$ – деяка норма матриці. Зокрема, для норми $\|\cdot\|_\infty$ маємо

$$|\lambda| \le \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|.$$

Точніший результат можна одержати за допомогою кругів Гершгорина, які задаються нерівностями

$$|\lambda - a_{ii}| \le \sum_{j \ne i} |a_{ij}|, \; i = \overline{1, n}. \tag{8.7}$$

Теорема 8.2. *Для того, щоб* λ *було власним значенням матриці A необхідно, щоб виконувалася б хоча б одна з нерівностей* (8.7). ■

Теорема 8.3. *Якщо* s *кругів Гершгорина,* $1 \le s \le n$, *утворюють область* G, *ізольовану від решти кругів, то в* G *є точно* s *власних значень матриці* A. ■

6. Нехай $\{\lambda, x\}$ – власна пара матриці $R = S^{-1}AS$.

Теорема 8.4. *Спектри подібних матриць збігаються*.

Доведення. Якщо матриці A і B – подібні, то існує невироджена матриця S така, що $B = S^{-1}AS$. Тоді

$$det(B - \lambda I) = det(S^{-1}AS - \lambda I) = detS^{-1} \cdot \det(A - \lambda I) \cdot detS = \det(A - \lambda I),$$

Отже, власні значення матриць A і B збігаються. ■

Із рівності $By = S^{-1}ASy$ випливає, $A(Sy) = \lambda(Sy)$. Тому, якщо y – власний вектор матриці B, то Sy – власний вектор матриці A.

8.4. LU-алгоритм розв'язування повної проблеми

У *LU*-методі не потрібно знаходити власний многочлен матриці й розв'язувати рівняння (8.3). Нехай $A = L_1U_1$, де L_1 – нижня, а U_1 – верхня трикутна матриця, причому діагональні елементи матриці L_1 відмітні від нуля, а матриці U_1 – дорівнюють 1. Таке зображення матриці A рівносильне перетворенню матриці в методі Ґауса (див. підрозділ 2.2).

Якщо ввести позначення $A_0 := U_1L_1$, то $U_1 = A_0L_1^{-1}$ і $A = L_1A_0L_1^{-1}$. Матриці A і A_0 подібні, тому їх спектри збігаються.

Нехай матриця $A_1 = L_2U_2$, де L_2 і U_2 мають такий же вигляд, як L_1 і U_1 відповідно, $A_2 := U_2L_2$. Тоді $A_1 = L_2A_2L_2^{-1}$. Звідси маємо, оержимо

$$A = L_1L_2A_2L_1^{-1}L_2^{-1} = (L_1L_2)A_2(L_1L_2)^{-1}.$$

Отже, спектри матриць A і A_2 також збігаються.

Такий ітераційний процес побудови матриць $A_1, A_2, \ldots$, подібних матриці A, називається *LU*-алгоритмом. Доведено [71, 72], що при певних обмеженнях на матрицю A, зокрема, коли всі її власні значення різні за модулем, ітераційний метод збігається до нижньої трикутної матриці. Діагональні елементи матриці $A_{k+1} = U_kL_k$ є наближеними значеннями власних значень матриці A. За наближення відповідних власних векторів можна взяти стовпці матриці

$$(L_1 L_2 ... L_k)^{-1}.$$

Недоліками LU-алгоритму є повільна збіжність і недостатня числова стійкість. Найбільш це відчутно для нестійкої несиметричної проблеми власних значень.

8.5. Метод обертань Якобі

Метод Якобі досить простий, який дозволяє знайти всі власні пари $\{\lambda_k, x_k\}$ симетричної матриці. Якщо $n \geq 10$, то цей метод конкурує зі складнішими алгоритмами.

Нехай A – симетрична матриця. На підставі теореми 8.1 існує ортогональна матриця Q така, що виконується рівність (8.6), тому $AQ = Q\Lambda$. Якщо u_i – i-ий стовпець матриці Q то $Au_i = \lambda_i u_i$, тобто цей стовпець є власним вектором матриці A, що відповідає власному значенню λ_i.

Задача полягає в побудові послідовності таких ортогональних матриць Q_k, що після деякої кількості ітерацій матриця $A_{k+1} = Q_k^T A_k Q_k \approx \Lambda$, де $A_0 = A$. Матрицю Q_k виберемо у вигляді матриці обертань (8.5), де $c_{i,j}^2 + s_{i,j}^2 = 1$, індекси i та j відповідають недіагональному елементу матриці A_k, який максимальний по модулю, тобто $\left|a_{ij}^{(k)}\right| = \max\limits_{p>l} \left|a_{pl}^{(k)}\right|$. Якщо таких елементів більше ніж один, то зафіксуємо один із них. Отже, матриця U_k одержується з одиничної матриці заміною u_{ii} і u_{jj} значенням c_k, а $u_{ij} = -s$, $u_{ji} = s_k$. Зокрема, можна взяти

$$c_k = \cos\varphi_k, \quad s_k = \sin\varphi_k, \quad \varphi_k \in \left(-\frac{\pi}{4}, \frac{\pi}{4}\right].$$

У цьому випадку матриця U_k є матрицею обертань у площині i-го та j-го рядків.

Розглянемо матрицю $A_{k+1} = U_k^T A_k U_k$. Якщо перемножити матриці у правій частині, то на місці елементів з індексами (i, j), (j, i) одержимо елемент

$$a_{ij}^{(k+1)} = (c_k^2 - s_k^2) a_{ij}^{(k)} - c_k s_k (a_{ii}^{(k)} - a_{jj}^{(k)}).$$

Виберемо $\varphi_k \in \left(\frac{\pi}{4}, \frac{\pi}{4}\right]$ так, щоб $a_{ij}^{(k+1)} = 0$. Оскільки

$$c_k = \cos\varphi_k,\ \ s_k = \sin\varphi_k,$$

то $c_k^2 - s_k^2 = \cos 2\varphi_k,\quad c_k s_k = 0.5\sin 2\varphi_k$. Тому для знаходження φ_k одержимо рівняння

$$2a_{ij}^{(s)}\cos 2\varphi - (a_{ii}^{(k)} - a_{jj}^{(k)})\sin 2\varphi = 0.$$

Якщо $a_{ii}^{(k)} = a_{jj}^{(k)}$, то $\cos 2\varphi_k = 0$ і $\varphi_k = \pi/4$.

Якщо ж $a_{ii}^{(k)} \neq a_{jj}^{(k)}$, то для знаходження φ_k одержимо рівняння,

$$tg2\varphi_k = \frac{2a_{ij}^{(k)}}{a_{ii}^{(k)} - a_{jj}^{(k)}}. \qquad (8.8)$$

Розв'язок рівняння (7.8) виберемо з інтервалу $(-\frac{\pi}{4}, \frac{\pi}{4})$. Елементи матриці $A^{(k+1)}$ обчислюються за формулами: $a_{ij}^{(k+1)} = a_{ji}^{(k+1)} = 0;$

$$a_{p_i}^{(k+1)} = c_k a_{p_i}^{(k)} + s_k a_{p_j}^{(k)},\ a_{p_j}^{(k+1)} = s_k a_{p_i}^{(k)} + c_k a_{p_j}^{(k)},\ p \neq i \text{ і } p \neq j;$$

$$a_{ii}^{(k+1)} = c_k^2 a_{ii}^{(k)} + s_k^2 a_{jj}^{(k)} + 2c_k s_k a_{ij}^{(k)};$$

$$a_{jj}^{(k+1)} = s_k^2 a_{ii}^{(k)} + c_k^2 a_{jj}^{(k)} - 2c_k s_k a_{ij}^{(k)}.$$

Доведено, що $\lim\limits_{k\to\infty} A_k = \Lambda$, тобто, виконавши деяку кількість ітерацій, можна одержати матрицю A_{k+1}, близьку до Λ [71, 72]. За похибку матриць A_{k+1} щодо діагональної матриці Λ можна взяти значення $\max\limits_{p>l}\left|a_{pl}^{(k+1)}\right| \le \varepsilon$, або $(\sum\limits_{p>l} a_{pl}^2)^{1/2} \le \varepsilon$, де ε – орієнтовна точність обчислень.

Наведемо алгоритм методу обертань Якобі.

1. $A_1 := A,\quad k := 1.$

2. Знайти $i, j,$ для яких $\left|a_{ij}^{(k)}\right| = \max\limits_{p>l}\left|a_{pl}^{(k)}\right|.$

3. Якщо $a_{ii}^{(k)} = a_{jj}^{(k)}$, то $\varphi_k = \pi/4,\ c_k = s_k = 1/\sqrt{2},$

інакше обчислити:

3.1. $a_k := \dfrac{2a_{ij}^{(k)}}{a_{ii}^{(k)} - a_{jj}^{(k)}};$

3.2. $c_k := \cos\varphi_k = \sqrt{\dfrac{1}{2}\left(1 + \dfrac{1}{\sqrt{1+a_k^2}}\right)},$

$$s_k := \sin\varphi_k = (sign(a_k))\sqrt{\frac{1}{2}\left(1-\frac{1}{\sqrt{1+a_k^2}}\right)}..$$

4. Обчислити $A_{k+1} = U_k^T A_k U_k$

5. Якщо $\max\limits_{p>l}\left|a_{pl}^{(k+1)}\right| \le \varepsilon$ або $(\sum\limits_{p>l}(a_{pl}^{(k+1)})^2)^{1/2} \le \varepsilon$, то власні значення $\lambda_i := a_{ii}^{(k+1)}$, інакше $k := k+1$ і перейти до п. 2. Власні вектори $x_i = v_i^{(k)}, i = \overline{1,n}$, $v_i^{(k)}$ – стовпці матриці $U^{(1)}U^{(2)}\cdots U^{(k)}$. ■

Зауважимо, що контроль правильності виконання дій при кожному повороті здійснюється шляхом перевірки збереження сліду trA_{k+1} перетворюваної матриці.

8.6. QR-алгоритм

Застосовується цей підхід до матриць загального вигляду і ґрунтується на зведенні матриці ортогональноподібними перетвореннями до квазітрикутної матриці вигляду

$$B_i = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ 0 & B_{22} & \cdots & B_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & B_{mm} \end{bmatrix}$$

з $m \le n$ блоками B_{ii}, порядок яких $n_1,\cdots,n_m$, причому $n_1+\cdots+n_m = n$. Якщо матриця A – симетрична або всі її власні значення дійсні і різні, то $m_i = 1$, матриця B - трикутна і власними значеннями є діагональні елементи. Для $n_i = 2$ власні значення матриці B_{ii} є розв'язками квадратного рівняння.

Якщо $\det A \ne 0$, то згідно з теоремою 2.2

$$A = QR,$$

де R - ортогональна, а Q - верхня трикутна матриця. Тоді матриця

$$RQ = Q^T AQ$$

ортогонально подібна до A.

Алгоритм методу полягає ось у чому. Нехай $A_0 := A$.

Для $k = 0, 1, \ldots$ виконаємо розклад

$$A_k = Q_k R_k$$

й обчислимо добуток матриць

$$A_{k+1} = R_k Q_k .$$

Оскільки $A_{k+1} = Q_k^* A_k Q_k$ то спектри матриць A_k і A_{k+1} збігаються. Границею послідовності при $k \to \infty$ є квазітрикутна матриця B.

Якщо всі власні значення дійсні й різні, то їх наближеними значеннями є діагональні елементи матриці A_{k+1}, відповідними власними векторами – стовпці матриці $(Q_k \cdots Q_0)^T = Q_0 Q_1 \ldots Q_k$.

Критерієм збіжності ітераційного процесу в цьому випадку може служити виконання нерівності

$$\sum_{j=i+1}^{n} (Q_{ij}^{(k)})^2 \le \varepsilon .$$

Кожній комплексно-спряженій парі відповідає діагональний блок B_{ii} порядку 2. Елементи цього блоку змінюються від ітерації до ітерації, але власні значення $\alpha_j \pm \beta_j i$ мають тенденцію до збіжності. Критерієм закінчення ітерацій для такого блоку може бути умова $\left|\lambda^{(k)} - \lambda^{(k-1)}\right| \le \varepsilon$.

8.7. Метод Хаусхолдера

Кожне перетворення Якобі симетричної матриці дає два таких що дорівнюють нулю недіагональних елементи. Якщо порядок матриці великий, то потрібно виконати досить велике число ітерацій. Крім того, при виконанні ітерацій на місці утворених нулі можуть появитись ненульові значення. Метод Хаусхолдера дозволяє утворювати на кожній ітерації кілька нулів, які не лежать на головній діагоналі, і які не змінюються на наступних ітераціях.

Для несиметричних матриць у підсумку одержується майже трикутна матриця вигляду

$$\begin{bmatrix} * & * & * & \cdots & * & * \\ * & * & * & \cdots & * & * \\ 0 & * & * & \cdots & * & * \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & * & * \end{bmatrix},$$

яка називається *матрицею Хесенберга*. Для симетричних така матриця тридіагональна.

Основою методу є перетворення Хаусхолдера[1]

$$P = I - 2WW^T, \tag{8.9}$$

де W - вектор-стовпець, $\|W\|_2 = 1.$

Нехай A - матриця порядку $n \geq 3,$ X - довільний вектор-стовпець, $k -$ ціле число, $1 \leq k \leq n-2.$ Тоді можна побудувати вектор W_k і матрицю $P_k = I - 2W_k W_k^T$ так, що

$$P_k X = P_k \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} \\ x_{k+2} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ -s_k \\ 0 \\ \vdots \\ 0 \end{bmatrix} =: Y.$$

Значення числа s_k вибирається так, щоб $s_k^2 = x_{k+1}^2 + ... + x_n^2$ і $sign(s_k) = sign(x_{k+1}).$ Друга умова служить для зменшенням впливу похибки заокруглення. Вектор W_k виберемо так:

$$W_k = \frac{1}{\alpha_k}(X - Y) = \frac{1}{\alpha_k}[0, ..., 0, (x_{k+1} + s_k), x_{k+2}, ..., x_n]^T.$$

Із умови $\|W_k\| = 1$ випливає, що

$$\alpha_k^2 = 2x_{k+1}s_k + 2s_k^2.$$

Отже, згідно з (7.9), матриця P_k задається формулою

$$P_k = I - 2W_k W_k^T.$$

Нехай A симетрична матриця порядку $n, n \geq 3.$ Послідовністю $n-2$ перетворень вигляду PAP, де P – матриця вигляду (7.9), матриця A зводиться до симетричної тридіагональної матриці. Для матриці порядку 5 це схематично показано нижче ($*$ і $\bullet$ – деякі елементи матриць P_k і A_k відповідно) [45]:

[1] Householder Alston S. Unitary Triangularization of a Nonsymmetric Matrix // Journal ACM, 1958. – **5** (4), 1958. – P. 339–342.

$$P_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \end{bmatrix}, \quad \mathrm{A}_1 = P_1 \mathrm{A} P_1 = \begin{bmatrix} a_{11} & u_1 & 0 & 0 & 0 \\ u_1 & w_1 & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet & \bullet \end{bmatrix};$$

$$P_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{bmatrix}, \quad \mathrm{A}_2 = P_2 \mathrm{A}_1 P_2 = \begin{bmatrix} a_{11} & u_1 & 0 & 0 & 0 \\ u_1 & w_1 & u_2 & 0 & 0 \\ 0 & u_2 & w_2 & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet & \bullet \end{bmatrix};$$

$$P_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}, \quad \mathrm{A}_3 = P_3 \mathrm{A}_2 P_3 = \begin{bmatrix} a_{11} & u_1 & 0 & 0 & 0 \\ u_1 & w_1 & u_2 & 0 & 0 \\ 0 & u_2 & w_2 & u_3 & 0 \\ 0 & 0 & u_3 & \bullet & \bullet \\ 0 & 0 & 0 & \bullet & \bullet \end{bmatrix}.$$

Одне перетворення Хаусхолдера виконується так. Нехай $V = AW$. Обчислимо $c = W^T V$ і $Q = V - cW$. Тоді

$$PAP = A - 2WQ^T - 2QW^T.$$

Нехай A − симетрична матриця порядку n. Нехай $A_0 := A$. Побудуємо послідовність матриць Хаусхолдера $P_1, P_2, ..., P_{n-1}$ так, що

$$A_k = P_k A_{k-1} P_k,$$

де матриця A_k має нулі під діагоналлю в стовпцях $1, 2, \ldots, \mathrm{k}$. Тоді A_{n-2} є симетричною тридіагональною матрицею, подібною матриці A. Цей метод має назву методу Хаусхолдера. Він має більшу числову стійкість порівняно з іншими методами. Складність алгоритму на всіх $n-1$ кроках (для квадратної матриці порядку n дорівнює

$$\frac{2}{3}n^3 + n^2 + \frac{1}{3}n = O(n^3).$$

При застосуванні до матриці Хесенберга $QR-$методу одержимо квазітрикутну матрицю, клітки якої по діагоналі мають

власні значення, близькі до власних значень матриці A. при реалізації методу на комп'ютері з $t-$розрядною двійковою мантисою одержується матриця, яка ортогонально подібна деякій матриці $A_1 + \Delta A$, причому [36]

$$\|\Delta A\|_2 \le knM2^{-t}\|A_T\|_2,$$

де $k-$кількість ітерацій, $n-$порядок матриці, $M = const = O(1)$, A_1-матриця Хесенберга.

8.8. Степеневий метод знаходження найбільшого за модулем власного значення

8.8.1. Матриця загального вигляду. Нехай матриця $A-$ матриця простої структури, тобто її власні вектори лінійно незалежні. Тоді будь-який n- вектор x можна записати у вигляді $x = \alpha_1 x_1 + ... + \alpha_n x_{1n}$, де x_i- власні вектори, α_i – деякі коефіцієнти. Припустимо також, що

$$|\lambda_1| > |\lambda_2| \ge |\lambda_3| \ge ... \ge |\lambda_n|.$$

Звідси випливає, що власне значення λ_1 дійсне.

Візьмемо довільний вектор $y^{(0)} \ne 0$, наприклад вектор $y^{(0)} = (1,0,...,0)^T$, і обчислимо вектори

$$y^{(1)} = Ay^{(0)}, \quad y^{(2)} = Ay^{(1)}, ..., \quad y^{(k+1)} = Ay^{(k)}.$$

Зауважимо, що $y^{(k)} = A^2 y^{(k-2)} = A^k y^{(0)}$.

Нехай $y^{(0)} = c_1 x_1 + c_2 x_2 + ... + c_n x_n$, причому $c_1 \ne 0$, інакше можна змінити $y^{(0)}$. Оскільки $Ax_i = \lambda_i x_i$, то

$$y^{(1)} = c_1 Ax_1 + c_2 Ax_2 + ... + c_n Ax_n = c_1\lambda_1 x_1 + c_2\lambda_2 x_2 + ... + c_n\lambda_n x_n.$$

Аналогічно,

$$y^{(2)} = c_1\lambda_1^2 x_1 + c_2\lambda_2^2 x_2 + ... + c_n\lambda_n^2 x_n,$$

$$\ldots\ldots\ldots\ldots$$

$$y^{(k)} = c_1\lambda_1^k x_1 + c_2\lambda_2^k x_2 + ... + c_n\lambda_n^k x_n,$$

$$y^{(k+1)} = c_1\lambda_1^{k+1} x_1 + c_2\lambda_2^{k+1} x_2 + ... + c_n\lambda_n^{k+1} x_n.$$

Уведемо позначення: $\mu_j = \lambda_j / \lambda_1,\ j = 2,...,n, \quad \beta_{ij} = c_i x_{ij},\ i, j = 1,...,n,$ де $x_{ij} - j$-та компонента вектора x_i. Тоді

$$y_j^{(k)} = \beta_{1j}\lambda_1^k + \beta_{2j}\lambda_2^k + ... + \beta_{nj}\lambda_n^k,\ j = \overline{1,n}.$$

Із лінійної незалежності векторів $x_1,...,x_n$ випливає, що існує хоча б одна компонента $y_j^{(k)} \neq 0$. Обчислимо

$$\frac{y_j^{(k+1)}}{y_j^k} = \frac{\beta_{1j}\lambda_1^{k+1} + \beta_{2j}\lambda_2^{k+1} + ... + \beta_{nj}\lambda_n^{k+1}}{\beta_{1j}\lambda_1^k + \beta_{2j}\lambda_2^k + ... + \beta_{nj}\lambda_n^k} = \lambda_1 \frac{1 + \frac{\beta_{2j}}{\beta_{1j}}\mu_2^{k+1} + ... + \frac{\beta_{nj}}{\beta_{1j}}\mu_n^{k+1}}{1 + \frac{\beta_{2j}}{\beta_{1j}}\mu_2^k + ... + \frac{\beta_{nj}}{\beta_{1j}}\mu_n^k}$$

Оскільки $|\mu_j| < 1$ при $j \geq 2$, то $\mu_j^k \to 0$ при $k \to \infty$. Тому

$$\frac{y_j^{(k+1)}}{y_j^{(k)}} = \lambda_1(1 + \mathrm{O}(\mu_2^k)).$$

Доведення цієї рівності можна проілюструвати на такому прикладі:

$$\frac{1 + a\mu_2^{k+1}}{1 + a\mu_2^k} = 1 + \frac{a\mu_2^k(\mu_2 - 1)}{1 + a\mu_2^k} = 1 + O(\mu_2^k).$$

Отже, маємо наближення для λ_1 вигляду

$$\lambda_1 \approx \lambda_1^{(k+1)} = y_j^{(k+1)} / y_j^{(k)}. \qquad (8.9)$$

Значення λ_1 можна вважати знайденим з потрібною кількістю значущих цифр, якщо така кількість цифр збігається в $\lambda_1^{(k)}$ і $\lambda_1^{(k+1)}$. Для надійності в точності λ_1 варто обчислювати кілька або всі m відношень (8.9), для яких $y_j^{(k)} \neq 0$, й ітерації припиняти тоді, коли для всіх відношень досягається задана точність або

$$\lambda^{(k+1)} = \left(\sum_j \frac{y_j^{(k+1)}}{y_j^{(k)}} \right) / m.$$

Оскільки $y^{(k+1)} = c_1\lambda_1^{k+1}\left[x_1 + \frac{c_2}{c_1}\mu_2^{k+1} + ... + \frac{c_n}{c_1}\mu_n^{k+1} \right]$, то маємо формулу для наближеного значення відповідного λ_1 власного вектора

$$y^{(k+1)} \approx c_1\lambda_1^{k+1}x_1.$$

Вектор $y^{(k+1)}$ доцільно нормувати на кожній ітерації, щоб запобігти зростанню його компонент. Тобто обчислити $\bar{y}^{(k+1)} = y^{(k+1)} / \|y^{(k+1)}\|$, де $\|.\|$ − деяка векторна норма, наприклад, $\|y^{(k+1)}\| = \sqrt{(y_1^{(k+1)})^2 + (y_2^{(k+1)})^2 + ... + (y_n^{(k+1)})^2}$.

Зауваження 8.2. *Якщо* $|\lambda_2| > |\lambda_3|$ *і знайдено наближення для* λ_1*, то можна обчислити наближене значення* λ_2 *за формулою*

$$\lambda_2 \approx \frac{y_j^{(k+1)} - \lambda_1 y_j^{(k)}}{y_j^{(k)} - \lambda_1 y_j^{(k-1)}}. \qquad (8.10)$$

8.8.2. Випадок симетричної матриці. Нехай A − дійсна симетрична матриця, тобто $A = A^T$. У цьому випадку існує система власних векторів $\{x_1,...,x_n\}$, які утворюють базис. Можна вважати, що базис ортонормований, тобто $(x_i, x_i) = 1$ і $(x_j, x_j) = 0$, якщо $i \neq j$. Для симетричної матриці можна побудувати ітераційний процес з більш високою швидкістю збіжності.

Обчислимо для довільного вектора $y^{(0)} \neq 0$ скалярні добутки:

$$\left(y^{(k)}, y^{(k)}\right) = c_1^2 \lambda_1^{2k} + c_2^2 \lambda_2^{2k} + ... + c_n^2 \lambda_n^{2k},$$

$$\left(y^{(k+1)}, y^{(k)}\right) = c_1^2 \lambda_1^{2k+1} + c_2^2 \lambda_2^{2k+1} + ... + c_n^2 \lambda_n^{2k+1}, k = 0,1,...$$

Звідси маємо

$$\frac{\left(y^{(k+1)}, y^{(k)}\right)}{\left(y_k, y_k\right)} = \frac{c_1^2 \lambda_1^{2k+1} + c_2^2 \lambda_2^{2k+1} + ... + c_n^2 \lambda_n^{2k+1}}{c_1^2 \lambda_1^{2k} + c_2^2 \lambda_2^{2k} + ... + c_n^2 \lambda_n^{2k}} =$$

$$= \lambda_1 \frac{1 + \left(\frac{c_2}{c_1}\right)^2 \mu_2^{2k+1} + ... + \left(\frac{c_n}{c_1}\right)^2 \mu_n^{2k+1}}{1 + \left(\frac{c_2}{c_1}\right)^2 \mu_2^{2k} + ... + \left(\frac{c_n}{c_1}\right)^2 \mu_n^{2k}} \approx \lambda_1 + \mathrm{O}(\mu_2^{2k}).$$

Наближене значення λ_1 обчислюється за формулою

$$\lambda_1 \approx \lambda_1^{(k+1)} = \frac{\left(y^{(k+1)}, y^{(k)}\right)}{\left(y^{(k)}, y^{(k)}\right)}. \qquad (8.11)$$

Похибка має порядок μ_2^{2k}, тоді як у загальному випадку (8.9) порядок складає μ_2^k.

Приклади розв'язування типових задач

Задача 1. Виконати десять ітерацій в методі Якобі для матриці

$$A=\begin{bmatrix} 8 & -1 & 3 & -1 \\ -1 & 6 & 2 & 0 \\ 3 & 2 & 9 & 1 \\ -1 & 0 & 1 & 7 \end{bmatrix}.$$

Розв'язування. На першій ітерації утворимо нуль на місці елемента a_{13}. Тоді $a_1 = -6,\ \cos\varphi_1 = 0.763020,\ \sin\varphi_1 = -0.646375.$ Матриця першого повороту

$$U_1=\begin{bmatrix} 0.7631 & 0.0000 & -0.6464 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.6464 & 0.0000 & 0.7630 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{bmatrix}$$

Перетворена матриця

$$A_2 = U_1^T A_1 U_1 = \begin{bmatrix} 5.4586 & -2.0558 & 0.0000 & -1.4094 \\ -2.0558 & 6.0000 & 0.8797 & 0.0000 \\ 0.0000 & 0.8797 & 11.5414 & 0.1166 \\ -1.4094 & 0.0000 & 0.1166 & 7.0000 \end{bmatrix}.$$

На наступному кроці зануляємо елемент $a_{12}^{(2)} = -2.0558.$

На дев'ятій ітерації одержимо

$$A_{10} = U_9^T A_9 U_9 = \begin{bmatrix} 3.2959 & 0.0025 & 0.0379 & 0.0000 \\ 0.0025 & 8.4052 & -0.0050 & 0.0668 \\ 0.0379 & -0.0050 & 11.7041 & -0.0014 \\ 0.0000 & 0.0668 & -0.0014 & 6.5948 \end{bmatrix}.$$

Оскільки $\max\limits_{j>i}\left|a_{ij}^{(10)}\right| = 0.668 < 0.075,$ то процес завершено. Власні значення: $\lambda_1 \approx 3.296,\ \lambda_2 \approx 8.405,\ \lambda_3 \approx 11.704,\ \lambda_4 \approx 6.595.$

Задача 2. За допомогою методу Хаусхолдера привести до тридіагонального вигляду симетричну матрицю

$$A = \begin{bmatrix} 4 & 2 & 2 & 1 \\ 2 & -3 & 1 & 1 \\ 2 & 1 & 3 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}$$

Розв'язування. Нехай $A_0 := A$. Побудуємо матриці A_1 і A_2, де A_2 - тридіагональна. Для матриці A_1 маємо $s_1 = \sqrt{2^2 + 2^2 + 1^2} = 3$, $\alpha_1 = \sqrt{2 \cdot 2 \cdot 3 + 2 \cdot 3^2} = \sqrt{30} \approx 5.47226$. Тоді

$W_1^T = [0 \quad 5 \quad 2 \quad 1] = [0.000000 \quad 0.912871 \quad 0.365148 \quad 0.182574]$.

Далі знаходимо $V_1 = AW_1 =$

$$= \frac{1}{\sqrt{30}} [0 \quad -12 \quad 12 \quad 9] = [0.000000 \quad -2.190890 \quad 2.190890 \quad 1.643168]$$

Стала $C_1 = W_1^T V_1 = -0.9$. Після чого обчислимо $Q_1 = V_1 - C_1 W_1 =$

$$= V_1 + 0.9W_1, Q_1^T = \frac{1}{\sqrt{30}} [0.000000 \quad -7.500000 \quad 13.800000 \quad 9.900000] =$$

$$= [0.000000 \quad -1.369306 \quad 2.519524 \quad 1.807484].$$

Матриця $A_1 = A_0 - 2W_1 Q_1^T - 2Q_1 W_1^T =$

$$= \begin{bmatrix} 4.0 & -3.0 & 0.0 & 0.0 \\ -3.0 & 2.0 & -2.6 & -1.8 \\ 0.0 & -2.6 & -0.68 & -1.24 \\ 0.0 & -1.8 & -1.24 & 0.68 \end{bmatrix}.$$

На наступному кроці обчислимо сталі $s_2 = -3.162278$, $\alpha_2 = 6.036874$, $c_2 = -1.264911$ і вектори:

$W_2^T = [0.000000 \quad 0.000000 \quad -0.954514 \quad -0.298168]$,

$V_2^T = [0.000000 \quad 0.000000 \quad 1.018797 \quad 0.980843]$,

$Q_2^T = [0.000000 \quad 0.000000 \quad -0.188578 \quad 0.603687]$.

Тридіагональна матриця $A_2 = A_1 - 2W_2 Q_2^T - 2Q_2 W_2^T =$

$$= \begin{bmatrix} 4.0 & -3.0 & 0.0 & 0.0 \\ -3.0 & 2.0 & 3.162278 & 0.0 \\ 0.0 & 3.162278 & -1.4 & -0.2 \\ 0.0 & 0.0 & -0.2 & 1.4 \end{bmatrix}.$$

Задача 3. Застосувати $LU-$алгоритм для знаходження наближених власних значень матриці $A:=\begin{bmatrix}2 & 1\\ 6 & 1\end{bmatrix}$. Ітерації припинити, якщо $|a_{12}|\le 0.05$.

Розв'язування. Перший $LU-$розклад

$$A:=L_1U_1=\begin{bmatrix}2 & 0\\ 6 & -2\end{bmatrix}\begin{bmatrix}1 & 0.5\\ 0 & 1\end{bmatrix}.$$

Тоді $A_1:=U_1L_1=\begin{bmatrix}5 & -1\\ 6 & -2\end{bmatrix}$. Далі маємо

$$A_2:=L_2U_2=\begin{bmatrix}5 & 0\\ 6 & -0.8\end{bmatrix}\begin{bmatrix}1 & -0.2\\ 0 & 1\end{bmatrix},$$

звідки $A_2:=U_2L_2=\begin{bmatrix}3.8 & 0.16\\ 6 & -0.8\end{bmatrix}$.

На наступному кроці

$$A_2:=L_3U_3=\begin{bmatrix}3.8 & 0\\ 6 & -1.0526\end{bmatrix}\begin{bmatrix}1 & 0.00421\\ 0 & 1\end{bmatrix},$$

$$A_3:=U_3L_3=\begin{bmatrix}4.0526 & 0.0443\\ 6 & -1.0526\end{bmatrix}.$$

Оскільки $a_{12}<0.05$ в матриці A_3, то ітерації припиняємо. Зауважимо, що для точного значення $\lambda_1=5$ і $\lambda_2=-1$, похибка наближених значень на головній діагоналі не перевищує 0.0527.

Задача 4. Знайти наближені значення власних значень матриці $\begin{bmatrix}2 & 2\\ 3 & 1\end{bmatrix}$, виконавши чотири ітерації степеневим методом.

Розв'язування. Нехай $y^{(0)}=[1,\,0]^T$. Тоді

$$y^{(1)}=Ay^{(0)}=\begin{bmatrix}2\\ 3\end{bmatrix},\quad y^{(2)}=Ay^{(1)}=\begin{bmatrix}10\\ 9\end{bmatrix}.$$

Аналогічно

$$y^{(3)}=\begin{bmatrix}38\\ 29\end{bmatrix},\quad y^{(4)}=\begin{bmatrix}154\\ 153\end{bmatrix},\quad y^{(5)}=\begin{bmatrix}614\\ 615\end{bmatrix}.$$

За формулою (8.9) знаходимо $y_1^{(5)}/y_1^{(4)}=614/154\approx 3.987,$

$y_2^{(5)} / y_2^{(4)} = 615/153 \approx 4.020$, які відрізняються на 0.033. Уточнене значення $\lambda_{\max} \approx \approx (3.9871 + 4.020)/2 = 4/0035$, що на 0.0035 відрізняється від точного значення $\lambda_1 = 4$. Відповідний власний вектор $x \approx y^{(5)}$ або $y^{(5)} / \left\| y^{(5)} \right\|_2 = [0.707, 0.708]^T$.

Для меншого власного значення, точне значення якого -1, згідно з (8.10) $\dfrac{y_1^{(5)} - \lambda_1 y_1^{(4)}}{y_1^{(4)} - \lambda_1 y_1^{(3)}} \approx -0.735$ й аналогічне значення для другої компоненти -0.786. Уточнене значення $-(0.735 + 0.786)/2 = -0.7605$, яке менш точне порівняно з першим власним значенням.

Задача 5. Знайти наближені значення власних значень матриці

$$\begin{bmatrix} 2 & 2 \\ 3 & 1 \end{bmatrix},$$

виконавши чотири ітерації методом скалярних добутків.

Розв'язування. Нехай $y_0 = [1,0]^T$. Тоді

$$y^{(1)} = Ay^{(0)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad y^{(2)} = Ay^{(1)} = \begin{bmatrix} 5 \\ 4 \end{bmatrix}, \quad \lambda^{(2)} = \frac{(y^{(2)}, y^{(1)})}{(y^{(1)}, y^{(1)})} = \frac{14}{5} = 2.8;$$

$$y^{(3)} = Ay^{(2)} = \begin{bmatrix} 14 \\ 13 \end{bmatrix}, \quad \lambda^{(3)} = \frac{(y^{(3)}, y^{(2)})}{(y^{(2)}, y^{(2)})} = \frac{122}{41} \approx 2.9756;$$

$$y^{(4)} = Ay^{(3)} = \begin{bmatrix} 41 \\ 40 \end{bmatrix}, \quad \lambda^{(4)} = \frac{(y^{(4)}, y^{(3)})}{(y^{(3)}, y^{(3)})} = \frac{1094}{356} \approx 2.9973;$$

$$y^{(5)} = Ay^{(4)} = \begin{bmatrix} 122 \\ 121 \end{bmatrix}, \quad \lambda^{(5)} = \frac{(y^{(5)}, y^{(4)})}{(y^{(4)}, y^{(4)})} = \frac{9842}{3281} \approx 2.9997.$$

Оскільки $\left|\lambda^{(5)} - \lambda^{(4)}\right| \approx 0.0024 < 0.005$, то $\lambda \approx 2.9997$, що на 0.0003 відрізняється від точного значення.

Завдання та запитання для самостійної роботи

1. Як знаходиться степеневим методом друге найбільше по модулю власне значення? Довести правильність формули (8.10).
2. Проілюструвати методи Якобі, QL і QR методи на матриці другого порядку.
3. Як знаходяться два найбільших по модулю комплексно спряжені власні значення?

4. Скласти програму обчислення методом Якобі з допустимою похибкою $\varepsilon = 10^{-6}$ всіх пар (λ_i, y_i) матриць:

$$1)\ \mathrm{A} = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix};\ 2)\ \mathrm{A} = \begin{bmatrix} 2.25 & -0.25 & -0.75 & 1.25 \\ -0.25 & 2.25 & 1.25 & -0.75 \\ -0.75 & 1.25 & 2.25 & -0.25 \\ 1.25 & -0.75 & -0.25 & 2.25 \end{bmatrix}.$$

Порівняти результати, одержані за допомогою комп'ютерної системи Mathematica або іншої системи.

5. Виконати три ітерації методом Якобі для знаходження всіх власних значень матриць:

$$1) \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix};\ 2) \begin{bmatrix} 5 & 1 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 3 \end{bmatrix};\ 3) \begin{bmatrix} 7.25 & 0.25 & -4.25 \\ 0.25 & -5.25 & 0.75 \\ -4.25 & 0.75 & 3.25 \end{bmatrix}.$$

Застосувати цей же метод, але попередньо звівши матрицю до триді-агонального вигляду за допомогою перетворенням Хаусхолдера.

6. Застосувати цей же метод, але попередньо звести матрицю до три діагонального вигляду перетворенням Хаусхолдера.

7. Для матриці $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ виконати три ітерації:

1) LU- алгоритмом; QR- алгоритмом на підставі перетворення Гівенса;
2) QR- алгоритмом на підставі перетворення Хаусхолдера.

Порівняти одержані результати за точністю (знайшовши попередньо точні значення) та за обчислювальними затратами.

8. Довести, що для симетричної і додатно визначеної матриці A послідовність ітерацій в $LU-$методі є збіжною.

9. Показати, що для матриці

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}, a \geq c,$$

послідовність ітерацій в $LU-$методі є збіжною до діагональної матриці.

10. Задана симетрична матриця

$$\begin{bmatrix} 1.0 & 0.5 & 1.0 & 1.5 \\ 0.5 & 2.0 & 2.1 & 2.2 \\ 1.0 & 2.1 & 3.0 & 3.2 \\ 1.5 & 2.2 & 3.2 & 4.0 \end{bmatrix}$$

1) Перевірити, що матриця додатно визначена.
2) Обчислити найбільше власне значення з точністю 0.01.

11. Знайти загальний розв'язок системи лінійних диференціальних рівнянь

$$\dot{u}_1 = 4u_1 + 3u_2 + 2u_3 + u_4,$$
$$\dot{u}_2 = 3u_1 + 4u_2 + 3u_3 + 2u_4,$$
$$\dot{u}_3 = 2u_1 + 3u_2 + 4u_3 + 3u_4,$$
$$\dot{u}_4 = u_1 + 2u_2 + 3u_3 + 4u_4.$$

обчисливши з точністю 0.05 власні значення і власні вектори матриці системи.

12. Знайти всі власні пари $\{\lambda, x\}$ матриці із завдання 1 степеневим методом і методом скалярних добутків, виконавши в кожному з них чотири ітерації. Порівняти одержані результати.

13. Побудувати матрицю відображення $I - WW^T$ за вектором $W^T = [1,2,3,4]$. Показати, що одержана матриця є ортогональною.

14. Знайти методом скалярних добутків з точністю 0.001 найбільше власне значення і відповідний власний вектор симетричних матриць:

$$1)\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix},\ 2)\begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 1 \end{bmatrix}.$$

15. Виконавши 5 ітерацій степеневим методом знайти два найбільші по модулю власні значення матриць із задачі 14. Проаналізувати точність знайдених значень, якщо $\lambda_1 = 9.62347538...,\ \lambda_2 = -0.62347538...$

16. Застосувати зведення до матриці Хаусхолдера для факторизації QR-методом матриць:

$$1)\begin{bmatrix} 1 & 1 & 1 \\ 2 & -1 & -1 \\ 2 & -4 & 5 \end{bmatrix};\quad 2)\begin{bmatrix} 1 & 3 & -2 \\ -1 & -2 & 3 \\ 1 & 1 & 2 \end{bmatrix}.$$

17. Дослідити вплив похибки $0 < \varepsilon << 1$ при застосуванні LU – алгоритму до матриці порядку n

$$\begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ \varepsilon & 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Розглянути випадок $n = 10,\ \varepsilon = 10^{-10}$.

18. Для заданих матриць степеневим методом знайти наближені значення власних значень і власних векторів:

$$1)\begin{bmatrix} 5 & 30 & -48 \\ 3 & 14 & -24 \\ 3 & 15 & -25 \end{bmatrix};\quad 2)\begin{bmatrix} \frac{49}{8} & -\frac{131}{8} & -\frac{43}{4} \\ \frac{11}{8} & -\frac{17}{8} & -\frac{9}{4} \\ -\frac{1}{2} & \frac{7}{2} & 3 \end{bmatrix};\quad 3)\begin{bmatrix} \frac{15}{8} & -\frac{69}{8} & -\frac{17}{4} \\ 2 & \frac{7}{8} & -\frac{5}{4} \\ 3 & -\frac{3}{8} & \frac{17}{4} \end{bmatrix}.$$

Ітерації завершити при збіганні трьох десяткових цифр власних значень.

19. Із точністю 0.001 обчислити спектральний радіус матриці

$$\begin{bmatrix} 5 & 1 & 2 \\ 1 & 4 & 1 \\ 2 & 1 & 3 \end{bmatrix}.$$

20. Довести, що для симетричної матриці матриця подібності Q у (8.6) ортогональна ($Q^{-1} = Q^T$).

21. Довести, що для екстремальних власних значень $\lambda_{\max}$ і $\lambda_{\min}$ симетричної матриці A правильні оцінки:

$$\lambda_{\min}(A) \le \min(a_{ii}),\ \lambda_{\max}(A) \ge \max(a_{ii}).$$

22. Нехай $A = A^T > 0$. Довести, що

$$\lambda_{\max}(A) = \max R_A(x),\ \lambda_{\min}(A) = \min R_A(x),$$

де $R_A(x) = \dfrac{(Ax, x)}{(x, x)}$ – відношення Релея.

23. Нехай A– симетризовна матриця, тобто існує невироджена матриця T така, що TAT^{-1} – симетрична. Довести, що система власних векторів матриці A повна.

24. Показати, що то елементарне перетворення подібності (8.6) руйнує структуру стрічкової матриці, наприклад тридіагональної матриці.

Розділ 9. Наближення функцій

Задача наближення функцій. Інтерполяційний многочлен Лагранжа, похибка інтерполювання. Поділені різниці та їх властивості, інтерполяційний многочлен Ньютона. Інтерполювання з кратними вузлами, інтерполяційні многочлени Ерміта. Мінімізація похибки інтерполювання. Зауваження про збіжність інтерполяційного процесу. Лінійні і кубічні інтерполяційні сплайни, побудова та властивості. Середньоквадратичне наближення.

Література [1, 5, 13, 26, 28, 35, 36, 66, 73, 80, 83]
Електронні джерела [105–107]

9.1. Постановка задачі про наближення функцій

Нехай на сітці $a \le x_0 < x_1 < ... < x_n \le b$ задано значення $f_i = f(x_i)$ функції $y = f(x)$, визначеної на відрізку $[a, b]$. Задача полягає в наближенні за цими даними функції f деякою функцією $\varphi = \varphi(x, a_0, a_1, ..., a_m)$, простою для обчислень, де $a_0, a_1, ..., a_m$ – параметри, які потрібно знайти. Якщо $m > n$, то задача недовизначена. Найчастіше функція φ є лінійною за параметрами a_i:

$$\varphi(x; a_0, ..., a_m) = \sum_{i=0}^{m} a_i \varphi_i(x), \qquad (9.1)$$

де $\{\varphi_i\}_0^m$ – система лінійно незалежних на відрізку $[a, b]$ функцій, досить простих для обчислень, наприклад алгебраїчних $1, x, ..., x^m$ або тригонометричних $1, \sin x, \cos x, ..., \sin kx, \cos kx$. Коефіцієнти у формулі (9.1) вибираються згідно з деякими критеріями. Розглянемо два з них.

1) Нехай $n > m$. За міру наближення можна взяти функцію

$$F(a_0, a_1, ..., a_m) = \sum_{i=0}^{n} (y_i - \varphi(x_i; a_0, a_1, ..., a_m))^2 .$$

Елемент, для якого сума квадратів відхилень набуває найменшого значення, називається *елементом найліпшого середньоквадратичного наближення* (ЕНСН). Отже, задача тут зводиться до мінімізації функції F за змінними $a_0, a_1, ..., a_m$.

2) Якщо $n = m$, то параметрами $a_0, a_1, ..., a_n$ визначаються із системи $n+1$ рівнянь з $n+1$ невідомими

$$y_i = \varphi(x_i; a_0, a_1, ..., a_n),\ i = \overline{0,n}. \qquad (9.2)$$

Геометрично це означає, що графік функції φ проходить через задані точки (x_i, y_i), $i = \overline{0,n}$. (рис. 9.1). Якщо φ – лінійна функція за параметрами $a_0, a_1, ..., a_n$, то задача зводиться до розв'язування СЛАР

$$\sum_{i=o}^{n} a_i \varphi_i(x_j) = y_j,\ \ j = \overline{0,n}. \qquad (9.3)$$

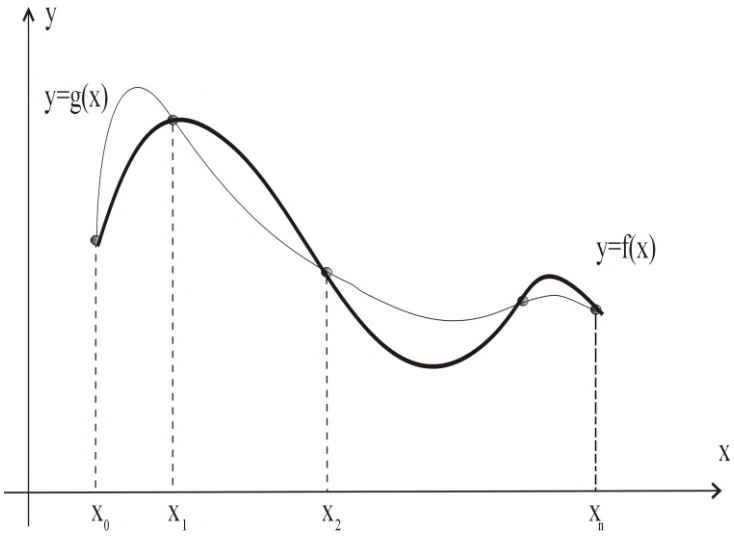


Рис. 9.1. Жирною лінією позначено графік функції, яку наближаємо

Задача *інтерполяції* полягає в заміні функції f на відрізку $[a,b]$ функцією g із деякого класу функцій такою, що функція g в точках x_0, $x_1,\ldots,$ x_n, набувала тих самих значень, що й функція f, тобто щоб виконувалась умова (9.2). Точки x_0, $x_1,\ldots,$ x_n називаються вузлами інтерполювання, Задача знаходження наближеного значення функції f в точці $x \neq x_i$, $i = \overline{0,n}$, за значенням функції $\varphi(x; a_0, ..., a_n)$, що задовольняє умову (9.2), називається *інтерполюванням.*

Лінійна за параметрами $a_0, a_1, ..., a_n$ функція g, для якої виконується умова (9.2), називається *інтерполяційним многочленом* для функції f за системою функцій $\{\varphi_i\}$.

9.2. Інтерполяційний многочлен Лагранжа

Виберемо просту для обчислень систему функцій $1, x, ..., x^n$. Ця система лінійно незалежна на довільному проміжку $[a,b]$, оскільки побудований для неї визначник Вронського

$$W[1,x,...,x^n]=\begin{vmatrix}1 & x & x^2... & x^n\\ 0 & 1 & 2x... & nx^{n-1}\\ ... & ... & ... & ...\\ 0 & 0 & 0... & n!\end{vmatrix}=\prod_{k=1}^{n}k!\neq 0.$$

Нагадаємо, що визначником Вронського $W[u_1,...,u_k]$ системи функцій $u_i\in C^{k-1}[a,b]$ називається визначник вигляду

$$W[u_1,...,u_k]=\begin{vmatrix}u_1(x) & u_2(x) & ... & u_k(x)\\ u_1'(x) & u_2'(x) & ... & u_k'(x)\\ ... & .... & ... & ...\\ u_1^{(k-1)}(x) & u_2^{(k-1)}(x) & ... & u_2^{(k-1)}(x)\end{vmatrix}.$$

Теорема 9.1 [61, с. 167]. *Якщо визначник Вронського системи функцій* $u_1(x),...,u_k(x)$ *не перетворюється в нуль хоча би в одній точці* проміжку $[a,b]$*, то така система функцій лінійно незалежна на цьому проміжку.* ■

Система рівнянь (9.3) запишеться для функцій $\varphi_i=x^i$, $i=\overline{0,n}$ так:

$$\sum_{i=0}^{n}a_i x_j^i=f_j,\ j=\overline{0,n}. \qquad (9.4)$$

Оскільки вузли x_i різні, то визначником системи є визначник Вандермонда, який відмітний від нуля. Отже, система лінійних рівнянь (9.4) має єдиний розв'язок, тому інтерполяційний многочлен існує і єдиний.

Форма запису інтерполяційного многочлена може бути різною. У формі Лагранжа він набуває вигляду

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

де $\Phi_i(x)$ − многочлени степеня n, причому $\Phi_i(x_i)=1$ і $\Phi_i(x_j)=0$, коли $j\neq i$. Тому $L_n(x_j)=f_j$. Многочлен Φ_i побудуємо у вигляді $\Phi_i(x)=C_i\prod_{j\neq i}(x-x_j)$. Оскільки $C_i\prod_{j\neq i}(x_i-x_j)=1$, то $C_i=1/\prod_{j\neq i}(x_i-x_j)$. Тепер інтерполяційний многочлен Лагранжа запишеться у вигляді

$$L_n(x)=\sum_{i=0}^{n}f(x_i)\prod_{j\neq i}\frac{x-x_j}{x_i-x_j}. \qquad (9.5)$$

Наприклад, коли $n=0$, то $L_0(x)=f_0$. Для $n=1$ маємо:

$$L_1(x)=f_0\frac{x-x_1}{x_0-x_1}+f_1\frac{x-x_0}{x_1-x_0}=\frac{(x-x_0)f_0+(x_1-x)f_1}{x_1-x_0}.$$

Для $n=2$ одержимо

$$L_2(x)=f_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}+f_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}+f_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}.$$

Оцінимо обчислювальні затрати для інтерполювання функції $y=f(x)$ у точці x згідно з формулою (9.5). Кількість арифметичних операцій складає $(n+1)(2(n-1)+2n+2)+n=4n^2+5n$. Для великих n це приблизно $4n^2$ арифметичних операцій додавання і множення.

Нехай $\omega_{n+1}(x):=(x-x_0)(x-x_1)\cdot....\cdot(x-x_n)$ – многочлен степеня $n+1$. Його похідна $\omega'_{n+1}(x)=\sum\limits_{i=0}^{n}\prod\limits_{j\neq i}(x-x_j)$. При $x=x_i$ маємо $\omega'_{n+1}(x_i)=\prod\limits_{j\neq i}(x_i-x_j)$. Тому інтерполяційний многочлен (9.5) можна записати також у вигляді

$$L_n(x)=\sum_{i=0}^{n}f(x_i)\frac{\omega_{n+1}(x)}{(x-x_j)\omega'_{n+1}(x_i)}. \qquad (9.6)$$

Зауваження 9.1. *Нехай* $f(x)=g[q(x)]$. *У деяких випадках точніший результат інтерполювання одержується, якщо скористатись узагальненим інтерполяційним многочленом Лагранжа вигляду*

$$\bar{L}_n(x)=\sum_{\nu=0}^{n}f(x_i)Q_i(x),$$

$$Q_i(x)=\prod_{j\neq 1}\frac{q(x)-q(x_j)}{q(x_i)-q(x_j)},\quad q(x_i)\neq q(x_j)\quad \textit{при } x_i\neq x_j.$$

Зауваження 9.2. *Для спрощення обчислень інтерполяційного многочлена можна застосувати схему Ейткена. Нехай* $l_{k,k+1,...,p}(x)$ *– інтерполяційний многочлен з вузлами інтерполювання* $x_k,...,x_p$. *Зокрема,* $l_k(x)=f(x_k)$. *Справджується рівність*

$$l_{k,k+1,...,p+1}(x)=\frac{l_{k+1,...,p+1}(x)(x-x_k)-l_{k,...,p}((x)(x-x_{p+1}))}{x_{p+1}-x_k}, \qquad (9.7)$$

оскільки права частина є многочленом степеня $p-k+1$ *і збігається зі значеннями функції* f *у точках* $x_k,...,x_{p+1}$. *Схема*

Ейткена для обчислення значення $L_n(x) = l_{0,1,\ldots,n}(x)$ *полягає у послідовному обчисленні за допомогою формули (9.7) елементів таблиці значень інтерполяційних многочленів*

$l_0(x)$

$l_1(x)\ l_{01}(x)$

$l_2(x)\ l_{12}(x)\ l_{012}(x)$

.....

$l_n(x)\ l_{n-1,n}(x)\ l_{n-2,n-1,n}(x)\ldots l_{012\ldots n}(x)$

9.3. Оцінка похибки інтерполювання

Припустимо, що $f \in C^{n+1}[a,b]$. Знайдемо вираз для похибки

$$R_n(x) = f(x) - L_n(x)$$

інтерполяційного многочлена у точці $x \in [a,b]$. У вузлах сітки $R_n(x_i) = 0,$ тому розглянемо випадок, коли $x \neq x_i$. Зведемо функцію $\psi(t) = f(t) - L_n(t) - K\omega_{n+1}(t),$ де сталу K виберемо так, щоб $\psi(x) = 0$. Рівність $\varphi(t) = 0$ виконується в точках $x_0, x_1, \ldots, x_n; x.$ Тому на підставі теореми Ролля похідна $\psi'(t)$ дорівнює нулю не менше, ніж в $(n+1)$-й точці, $\psi''(t)$ – у n точках і т.д. Похідна

$$\psi^{(n+1)}(t) = f^{(n+1)}(t) - L_n^{(n+1)}(t) - \omega_{n+1}^{(n+1)}(t) = f^{(n+1)}(t) - K(n+1)! = 0$$

хоча б в одній точці, позначимо її через ξ. Тоді $K = f^{(n+1)}(\xi)/(n+1)!$ Отже,

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x), \quad x \in [a,b]. \tag{9.8}$$

Зауважимо, що коли $f(x)$ – многочлен степеня $p,$ $p \le n$, то $f^{(n+1)}(t) \equiv 0$ і $R_n(x) \equiv 0$. Якщо $\max\limits_{x\in[a,b]} \left| f^{(n+1)}(x) \right| \le M_{n+1},$ то маємо для $R_n(x)$ оцінку вигляду

$$\left|R_n(x)\right| \le \frac{M_{n+1}}{(n+1)!}\left|\omega_{n+1}(x)\right|. \tag{9.9}$$

9.4. Поділені різниці

Для запису інтерполяційного многочлена у формі Ньютона, як певного аналогу формули Тейлора, потрібно ввести поняття *поділеної різниці*. Поділена різниця нульового порядку в точці x_i

збігається зі значенням $f(x_i)$. Нехай $x_i \neq x_j$, коли $i \neq j$. Різниці першого порядку задаються рівністю

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

різниці другого порядку мають вигляд

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

Різниця k-го порядку $f(x_1, x_2, ..., x_{k+1})$ визначається через різниці $(k-1)$-го порядку за формулою

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

Таблиця 9.1

Поділені різниці

x_i	$f(x_i)$	$f(x_i; x_j)$	$f(x_i; x_j; x_k)$	$f(x_i; x_j; x_k; x_l)$	…	$f(x_0; x_1; ...; _{n-1}; x_n)$
x_0	$f(x_0)$	$f(x_0; x_1)$	$f(x_0; x_1; x_2)$	$f(x_0; x_1; x_2; x_3)$	…	$f(x_0; x_1; ...; _{n-1}; x_n)$
x_1	$f(x_1)$	$f(x_1; x_2)$	$f(x_1; x_2; x_3)$	$f(x_1; x_2; x_3; x_4)$		
x_2	$f(x_2)$	$f(x_2; x_3)$	$f(x_2; x_3; x_4)$	…	…	
x_3	$f(x_3)$	$f(x_3; x_4)$	…	$f(x_{n-3}; x_{n-2}; x_{n-1}; x_n)$		
		…	$f(x_{n-2}; x_{n-1}; x_n)$			
…	…	$f(x_{n-1}; x_n)$				
x_n	$f(x_n)$					

Лема 9.1. *Справджується рівність*

$$f(x_0, x_2, ..., x_k) = \sum_{i=0}^{k} \frac{f(x_i)}{\prod_{j \neq i} (x_i - x_j)}. \qquad (9.10)$$

Для $k = 1$ маємо

$$f(x_0, x_1) = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0}. \qquad (9.11)$$

Завершується доведення леми методом математичної індукції. ■

Із рівності (9.10) випливають такі дві властивості:

1. Для фіксованих $x_1, ..., x_k$ *поділена різниця є лінійним функціоналом:*

$$(\alpha_1 f_1 + \alpha_2 f_2)(x_1; ...; x_k) = \alpha_1 f_1(x_1; ...; x_k) + \alpha_2 f_2(x_1; ...; x_k),\ \alpha_1, \alpha_2 \in R.$$

2. Поділена різниця є симетричною функцією своїх аргументів.

Зокрема, для поділеної різниці першого порядку маємо $f(x_1, x_2) = f(x_2, x_1)$. Нижче наведено таблицю поділених різниць.

9.5. Інтерполяційний многочлен Ньютона

Покажемо, що інтерполяційний многочлен $L_n(x)$ можна записати у вигляді

$$P_n(x) = f(x_0) + f(x_0; x_1)(x - x_0) + f(x_0; x_1; x_2)(x - x_0)(x - x_1) + ... \\ + f(x_0; x_1; ...; x_n)(x - x_0)(x - x_1)...(x - x_{n-1}). \qquad (9.12)$$

Запишемо інтерполяційний многочлен Лагранжа $L_n(x)$ так:

$$L_n(x) = L_0(x) + (L_1(x) - L_0(x)) + ... + (L_n(x) + L_{n-1}(x)). \qquad (9.13)$$

Зрозуміло, що $L_0(x) = f(x_0)$, а

$$L_1(x) - L_0(x) = \frac{x - x_1}{x_0 - x_1} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1) - f(x_0) =$$

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

Різниця $L_k(x) - L_{k-1}(x)$ є многочленом степеня k, корені якого $x_0, x_1, ..., x_{k-1}$. Оскільки

$$L_k(x_i) = L_{k-1}(x_i) = f(x_i),\ i = \overline{0, k-1},$$

то $L_k(x) - L_{k-1}(x) = A_k \omega_k(x)$. Нехай $x = x_k$, тоді

$$f(x_k) - L_{k-1}(x_k) = A_k \omega_k(x_k). \qquad (9.14)$$

З іншого боку,

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

$$=\prod_{i=0}^{n}(x-x_i)\left(\frac{f(x)}{\prod_{i=0}^{k}(x-x_i)}+\sum_{i=0}^{n}\frac{f(x_i)}{(x_i-x)\prod_{j\neq i}(x_i-x_j)}\right). \qquad (9.15)$$

Порівнявши рівності (9.14) із (9.15) одержимо, що

$$f(x)-L_n(x)=f(x;x_0;...;x_n)\omega_{n+1}(x). \qquad (9.16)$$

Для $n=k-1$ і $x=x_k$ із (9.13) і (9.15) маємо $A_{k-1}=f(x_0;...;x_{k-1};x_k)$. Отже, $L_k(x)-L_{k-1}(x)=f(x_0;x_1;...;x_k)\omega_k(x)$. Підставивши ці величини в (9.13), одержимо формулу (9.12), яка називається *інтерполяційним многочленом Ньютона.*

Зауваження 9.3. *Формула (9.16) служить оцінкою похибки в точці x для інтерполяційного многочлена Ньютона (9.12). Порівнявши (9.8) і (9.16), одержимо*

$$f(x;x_0;...;x_n)=\frac{f^{(n+1)}(\xi)}{(n+1)!},\ \xi\in[0,b].$$

9.6. Мінімізація похибки інтерполювання

Нехай функція $f(x)$ апроксимується на $[a,b]$ інтерполяційним многочленом Лаґранжа $L_n(x)$ за вузлами інтерполювання $x_0,x_1,...,x_n\in[a;b]$. Для заданої функції f, як випливає з оцінки похибки інтерполювання (9.8), мінімізація похибки можлива шляхом вибору вузлів $x_0,x_1,...,x_n$ у функції $w_{n+1}(x)$.

Для $n\geq 1$ коефіцієнт при x^n многочлена Чебишева $T_n(x)$ дорівнює 2^{n-1}. Многочлен $\overline{T}_n(x)=2^{1-n}T_n(x)=x^n+...$ називається многочленом, який найменше відхиляється від нуля.

Теорема 9.1. *Якщо $P_n(x)$ – многочлен степеня n зі старшим коефіцієнтом 1, то*

$$\max_{[-1;1]}|P_n(x)|\geq\max_{[-1;1]}|\overline{T}_n(x)|=2^{1-n}.$$

Доведення. Нехай висновок теореми неправильний. Оскільки степінь многочлена $\overline{T}_n(x)-P_n(x_p)$ дорівнює $n-1$, то, згідно з припущенням $|P_n(x_p)|<2^{1-n}$ для всіх p, тому

$$sign(\overline{T}_n(x_p) - P_n(x_p)) = sign((-1)^m 2^{1-n} - P_n(x_p)) = (-1)^p .$$

Отже, між кожними двома точками x_p, x_{p+1} многочлен $\overline{T}_n(x) = P_n(x)$ змінює знак. Тож многочлен $\overline{T}_n(x) - P_n(x)$ степеня $n-1$ має n різних нулів, що є суперечністю. ■

Відрізок $[-1,1]$ лінійною заміною

$$t = \frac{b+a}{2} + \frac{b-a}{2}x$$

відображається у відрізок $[a,b]$. Старший коефіцієнт многочлена $\overline{T}_n\left(\frac{2x-(b+a)}{b-a}\right)$ дорівнює $\left(\frac{2}{b-a}\right)^n$. На підставі теореми 9.1 многочлен

$$\overline{T}_n^{[a;b]}(x) = (b-a)^n 2^{1-2n} T_n\left(\frac{2x-(b+a)}{b-a}\right)$$

зі старшим коефіцієнтом 1 є многочленом, який найменше відхиляється від нуля на проміжку $[a,b]$. Це означає, що для довільного многочлена вигляду $P_n(x) = x^n + \ldots$ виконується нерівність

$$\max_{[a;b]} |P_n(x)| \geq \max_{[a;b]} |\overline{T}_n^{[a;b]}(x)| = (b-a)^n 2^{1-n}.$$

Оскільки величини $\cos\left(\frac{\pi(2k-1)}{2n}\right)$, $k = \overline{1,n}$, на $[-1,1]$ є нулями многочлена Чебишева, то вузлами інтерполювання на $[a,b]$, тобто нулями многочлена $\overline{T}_n^{[a;b]}(x)$, є точки

$$x_k = \frac{b+a}{2} + \frac{b-a}{2}\cos\left(\frac{\pi(2k-1)}{2n}\right),\ k = \overline{1,n}. \qquad (9.17)$$

Отже, можна мінімізувати похибку (9.8), якщо $\omega_{n+1}(x) = 2^{-n-1}T_{n+1}(x)$. Тоді $|\omega_{n+1}(x)| \geq (b-a)2^{1-2n}$, коли $x \in [a;b]$. Тож узявши (9.17) за вузли інтерполювання, одержимо $\omega_{n+1}(x) = \overline{T}_{n=1}^{[a;b]}(x)$. У цьому випадку $\max\limits_{x\in[a;b]} |\omega_{n+1}(x)| = (b-a)^n 2^{1-2n}$. Тобто при такому розміщенні вузлів

$$\max_{[a;b]} |f(x) - L_n(x)| \leq \frac{M_{n+1}(b-a)^{n+1}2^{-1-2n}}{(n+1)!}.$$

Одержану оцінку не вдається поліпшити, оскільки для многочлена $f(x)=a_0x^{n+1}+...+a_n,\ \ a_0\neq 0,$ степеня $n+1$ маємо $f^{(n+1)}(\theta)=a_0(n+1)!,$ тому нерівність перетворюється в рівність. Отже, для довільних вузлів інтерполювання

$$\max_{[a;b]}\left|f(x)-L_n(x)\right|\geq\max_{[a;b]}\left|f^{(n+1)}(x)\right|(b-a)^{n+1}2^{-1-2n}.$$

9.7. Інтерполювання з кратними вузлами

Нехай у точках $x_0,x_1,...,x_m,\ \ x_i\neq x_j$ коли $i\neq j,$ задано не тільки значення функції $y_i=f(x_i),$ але й значення похідних до порядку $n_i-1,\ \ n_i\geq 1,\ \ n=n_0+n_1+...+n_m-1.$ Якщо $n_i=1,$ то $f^{(0)}(x_i)=f(x_i).$ Число n_i назвемо кратністю вузла. Отже, вхідні дані набувають вигляду:

$$\begin{array}{cccc} x_0 & x_1 & ... & x_m \\ f(x_0) & f(x_1) & ... & f(x_m) \\ \multicolumn{4}{c}{.....} \\ f^{(n_0-1)}(x_0) & f^{(n_1-1)}(x_1) & ... & f^{(n_m-1)}(x_m) \end{array}$$

Задача полягає в тому, щоб побудувати многочлен $H_n(x)$ степеня n з числом коефіцієнтів $n+1$ такий, що

$$H_n^{(j)}(x_i)=f^{(j)}(x_i),\ i=\overline{0,m};\ j=\overline{0,n_i-1}. \qquad (9.18)$$

Існування та єдиність такого многочлена випливає з таких міркувань. Розглянемо відповідну (9.18) однорідну систему рівнянь

$$H_n^{(j)}(x_i)=0,\ i=\overline{0,m_j},\ j=\overline{0,n_i-1}. \qquad (9.19)$$

Корінь x_i рівняння (9.19) має кратність не мешу, ніж $n_i-1.$ Отже сумарна кратність всіх коренів не менша, ніж $n+1.$ Оскільки $H_n(x)$ многочлен степеня $n,$ то це можливо тоді і тільки тоді, коли $H_n(x)\equiv 0.$ Це означає, що система (9.19) відносно $n+1$ коефіцієнтів многочлена $H_n(x)$ має тільки нульовий, а відповідна неоднорідна система (9.18) – єдиний розв'язок.

Інтерполяційний многочлен Ерміта будується у такий же спосіб, як і многочлен Лагранжа у вигляді

$$H_n(x)=\sum_{i=0}^{n}\sum_{j=0}^{k_i-1} f^{(j)}(x_i)\Phi_{ij}(x),$$

де $\Phi_{ij}(x)$ − алгебраїчний многочлен степеня n, який задовольняє умови: $\Phi_{ij}^{k}(x_l)=1,$ коли $i=l$ і $j=k$ та $\Phi_{ij}^{k}(x_l)=0,$ в інших випадках.

9.8. Збіжність інтерполяційного процесу

На труднощі, пов'язані з інтерполюванням, указав ще у 1901 р. К. Рунге на прикладі функції $f(x)=(1+25x^2)^{-1}$, аналітичної на $x\in[-1;\ 1]$. Послідовність інтерполяційних многочленів $L_n(x)$, що інтерполює функцію $f(x)$ на рівномірній сітці, зростанням n не збігалась до неї у жодній точці. крім вузлів інтерполювання (рис. 9.2). Як показано в [10, с. 31], зростання максимальної похибки інтерполювання зі збільшенням вузлів інтерполювання набуває значень:

n	$\max\limits_{[a,b]}\lvert f(x)-L_n(x)\rvert$	n	$\max\limits_{[a,b]}\lvert f(x)-L_n(x)\rvert$
2	0.9615	12	0.5567
4	0.7070	14	1.069
6	0.4247	18	4.214
8	0.2474	20	8.573
10	0.2994		

Аналогічний приклад для функції $y=|x|,\ x\in[-1,1]$ побудований у 1916 р. С. Бернштейном.

Нехай $\{\Delta_n\}$ − послідовність сіток на відрізку $[a,b]$: $\Delta_n=$ $=\{x_i : a\le x_0^{(n)}<\ldots<x_n^{(n)}\le b\},\ \ n=0,1,\ldots$ Побудуємо послідовність інтерполяційних многочленів $L_n[f]$, які інтерполюють функцію $f\in C[a,b]$ на сітці $\Delta_n,\ \ n=0,1,\ldots$ Назвемо цю послідовність *інтерполяційним процесом.*

Означення 9.1. *Інтерполяційний процес збігається до функції* $f(x)$ *у точці* $x\in[a,b]$, *якщо* $\lim\limits_{n\to\infty}L_n[f(x)]=f(x)$. *Збіжність рівномірна на* $[a,b]$, *якщо* $\max\limits_{x\in[a,b]}\left|f(x)-L_n[f(x)]\right|=0$ *при* $n\to\infty$.

Збіжність чи розбіжність інтерполяційного процесу залежить від вибору послідовності сіток Δ_n і гладкості функції f.

Наведемо теореми про розбіжність і збіжність інтерполяційного процесу.

Теорема 9.2 (Фабер[1]). *Якою б не була послідовність сіток* Δ_n, *знайдеться неперервна на* $[a,b]$ *функція* $f(x)$ *така, що послідовність* інтерполяційних многочленів $L_n[f(x)]$ *не збігається рівномірно до* $f(x)$ *на* $[a,b]$. ■

Отже, рівномірну збіжність на класі неперервних на $[a,b]$ функцій не забезпечує жодна послідовність сіток. Але якщо зафіксувати неперервну функцію, то можна побудувати спеціальну послідовність сіток Δ_n, для яких інтерполяційний процес збігається до $f(x)$ рівномірно на $[a,b]$.

Теорема 9.3 (Марцинкевич[1]). *Якщо функція* $f(x)$ *неперервна на* $[a,b]$, *то існує послідовність сіток, для яких відповідний інтерполяційний процес збігається рівномірно на проміжку* $[a,b]$.■

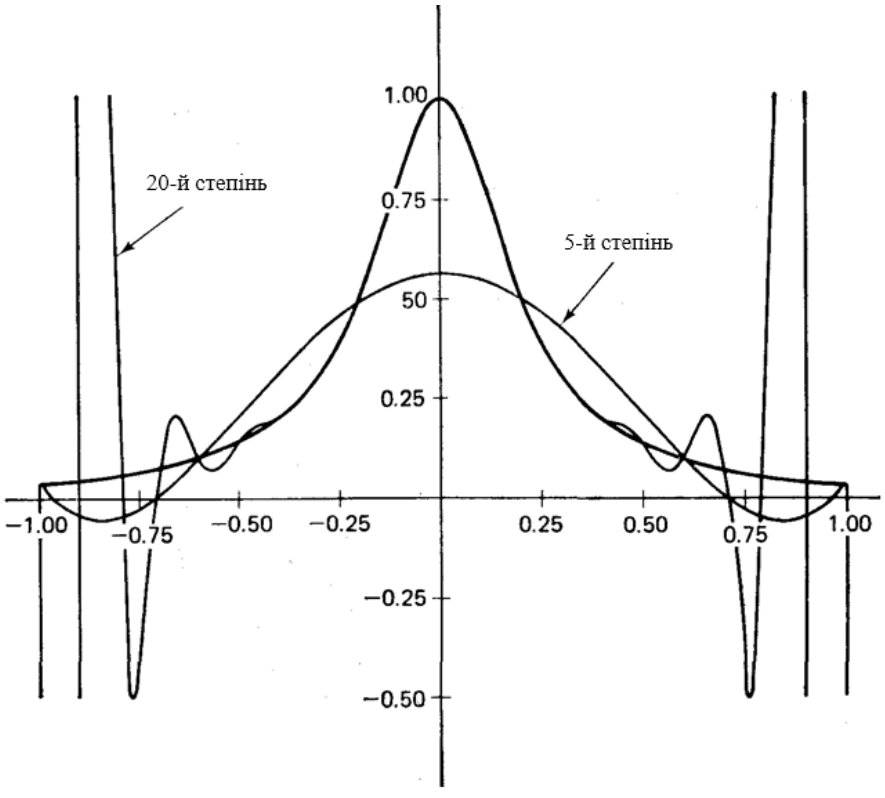


Рис. 9.2. Функція Рунге, що інтерполюється многочленами 5-го і 20-го степеня

Зауважимо, що алгоритм побудови сіток у твердженні 9.3 досить складний. Тому в обчислюваній практиці не користуються інтер-

[1] Натансон И.П. Конструктивная теория функций. – М.: Гостехиздат, 1949. – 532 с.

поляційними многочленами високих степенів. Значно ефективніший апарат апроксимації сплайнами.

9.9. Поняття сплайна

Під сплайном (від англ. spline – гнучка лінійка) розуміють агрегатну функцію, яка збігається із простішими функціями на кожному елементі розбиття відрізка $[a,b]$. Наприклад, для функції $y=f(x)$, визначеної на $[a,b]$, кубічний сплайн "склеюється" із кубічних многочленів, побудованих на елементах, якими є проміжки $[x_{i-1},x_i]$. Максимальний степінь з використаних на відповідних елементах многочленів називається *степенем сплайна*. Різниця між степенем сплайна і його гладкістю на $[a,b]$ називається *дефектом сплайна*. Наприклад, кубічний сплайн S_3 має дефект 1, якщо $S_3 \in C^2[a,b]$, і дефект 2, якщо $S_3 \in C^1[a,b]$.

Кубічний інтерполяційний сплайн дефекту 1 можна проілюструвати на прикладі гнучкої стальної лінійки, закріпленої в точках (x_0,y_0), (x_1,y_1),...,(x_n,y_n) (рис. 9.1).

При відсутності зовнішніх навантажень на кінцях лінійки профіль струни задовольняє крайові умови

$$S''(x_0)=S''(x_n)=0.$$

Згідно із законом Бернуллі-Ейлера лініарезоване диференціальне рівняння для профілю струни набуває вигляду

$$\alpha S''(x)=-M(x),$$

де $M(x)$ - момент прогину, який змінюється лінійно від однієї точки опору до іншої, $\alpha=const$ – коефіцієнт жорсткості. Після інтегрування крайової задачі одержимо, що $S(x)$ – кубічна функція, двічі неперервно диференційована на $[x_0,x_n]$.

Сплайни широко застосовуються в теорії наближень і в прикладних задачах. Зокрема, для задання поверхонь у системах комп'ютерного моделювання (наприклад криві Безьє, NURBS). Інтенсивне вивчення сплайнів розпочалося з другої половини XX століття. Термін "сплайн" запровадив Ісаак Шонберг (I. Shoenberg) у 1964 р. для поліноміальних сплайнів[2]. Детально різні питання теорії сплайнів розглянено в [1, 26, 40, 66] та ін.

[2] Schoenberg I. Contribution to the problem of approximation of equidistant data by analytic functions / I. Schoenberg // Quart. Appl. Math. Part B. − 1946, №4. – P. 112–141.

9.10. Лінійні інтерполяційні сплайни

Нехай $Q_m[a,b]$ – множина многочленів, визначена на проміжку $[a,b]$, степінь яких не перевищує m, $\Delta = \{x_i : a = x_0 < x_1 < ... < x_n = b\}$ – сітка на відрізку $[a,b]$.

Означення 9.2. *Функція* $S_1(x)$ *називається лінійним інтерполяційним сплайном, що інтерполює функцію* f *на сітці* Δ*, якщо:*

1) $S_1 \in Q_1\left[x_{i-1}, x_i\right],\ i = \overline{1,n};$

2) $S_1 \in C\left[a,b\right];$

3) $S_1(x_i) = f(x_i),\ i = \overline{0,n}.$

Отже, сплайн $S_1(x)$ – функція, "склеєна" у вузлах x_i, $i = \overline{1,n}$, із лінійних функцій (рис. 9.3). На кожному з відрізків $[x_{i-1}, x_i]$, $i = \overline{1,n}$ маємо рівняння сплайна $S_1(x) = y_i + k_i(x - x_i)$. Оскільки кутовий коефіцієнт $k_i = \dfrac{y_i - y_{i-1}}{x_i - x_{i-1}}$, то рівняння лінійного інтерполяційного сплайна на проміжку $[x_{i-1}, x_i]$ набуває вигляду

$$S_{1,i}(x) = y_i + \frac{y_i - y_{i-1}}{x_i - x_{i-1}}\left(x - x_i\right), \qquad i = \overline{1,n}. \tag{9.21}$$

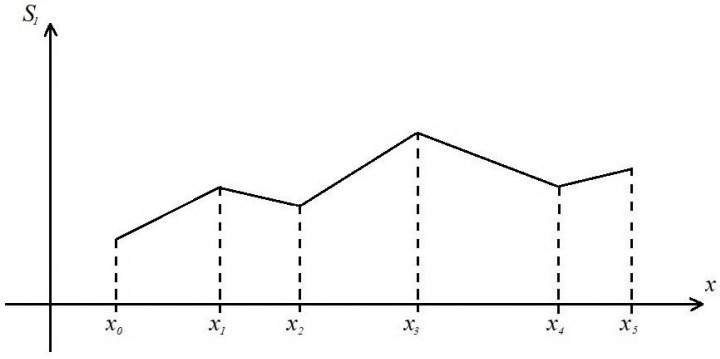


Рис. 9.3. Лінійний інтерполяційний сплайн

Для оцінки похибки інтерполювання сплайном $S_1(x)$ функції f *на сітці* Δ введемо поняття модуля неперервності функції.

Означення 9.3. *Нехай* $f \in C[a,b],\ 0 < \delta \le b-a$.

Число $\omega(\delta, f) = \max\limits_{\substack{x, x+t \in [a,b] \\ |t| \le \delta}} |f(x+t) - f(x)|$

називається модулем неперервності функції f.

Модуль неперервності показує, наскільки можуть відрізнятись значення функції f у точках $t, x \in [a, b]$, відстань між якими не перевищує δ. Зауважимо, що для неперервності функції f на відрізку $[a, b]$ необхідно і досить, щоб виконувалась умова $\lim\limits_{\delta \to 0} \omega(\delta,\ f) = 0$.

Уведемо позначення: $\|f\| = \max\limits_{x \in [a,b]} |f(x)|,\ \|\Delta\| = \max\limits_{i} (x_i - x_{i-1})$.

Теорема 9.4. *Нехай* $f \in C^1[a,b],\ S_1(x)$ *– лінійний інтерполяційний сплайн, який інтерполює функцію* f *на сітці* Δ. *Тоді справджуються оцінки:*

$$\left\| f^{(j)} - S_1^{(j)} \right\| \le K_j \|\Delta\|^{1-j} \cdot \omega(\|\Delta\|, f'),\quad j = 0, 1, \qquad (9.22)$$

де $2K_0 = K_1 = 1$, $S'(x_i)$ *– одна з похідних* $S'(x_i \pm 0)$.

Доведення. На відрізку $[x_{i-}, x_i]$ маємо

$$|f'(x) - S_1'(x)| = \left| \frac{y_i - y_{i-1}}{x_i - x_{i-1}} - f'(x) \right| = |f'(\xi_i) - f'(x)| \le \omega(\|\Delta\|, f'),$$

$\xi_i \in (x_{i-1}, x_i)$.

Нехай z – найближчий до x кінець відрізка $[x_{i-1}, x_i]$. Тоді

$S_1(z) = f(z)$ і $|S_1(x) - f(x)| = \left| \int\limits_z^x \left(S_1'(t) - f'(t) \right) dt \right| \le \int\limits_z^x \left| S_1'(t) - f'(t) \right| dt \le$

$\le 0.5 \|\Delta\| \omega\left(\|\Delta\|,\ f'\right)$.

Перейшовши до норми $\|\cdot\|$ одержимо другу з оцінок (9.22).

Наслідок 9.1. *Нехай сітка* Δ *– рівномірна сітка з кроком* h. *Тоді*

$$\left\| S_1(x) - f(x) \right\| \le \frac{1}{2} h \omega\left(h, f'\right).$$

Наслідок 9.2. *Розглянемо послідовність сіток* Δ_m:

$$a \le x_0^m < x_1^m < \ldots < x_m^m = b$$

і відповідну послідовність лінійних інтерполяційних сплайнів $S_1(x; \Delta_m)$, *що інтерполюють функцію* $f \in C^1[a,b]$ *на* Δ_m. *Якщо*

$\|\Delta_m\| \to 0$ *при* $m \to \infty$ *(або* $h \to 0$ *для рівномірної сітки), то з теореми 9.4 випливає рівномірна збіжність послідовності* $S_1(x; \Delta_m)$ *до* f *при* $m \to \infty$.

9.11. Кубічні інтерполяційні сплайни дефекту 1

Означення 9.4. *Функція* $S_3(x)$ *називається кубічним інтерполяційним сплайном дефекту 1, що інтерполює функцію* f *на сітці* Δ*, якщо виконуються такі умови:*

1^0. $S_3 \in Q_3[x_{i-1}, x_i]$, $i = \overline{1,n}$;

2^0. $S_3 \in C^2[a,b]$;

3^0. $S_3(x_i) = f(x_i)$, $i = \overline{0,n}$.

Отже, кубічний інтерполяційний сплайн $S_3(x)$ – кусково кубічна функція, склеєна у внутрішніх вузлах сітки з гладкістю 2-го порядку, тому дефект $d = 1$. Така конструкція служить математичною моделлю форми гнучкої лінійки, за допомогою якої проводиться плавна лінія через фіксовані точки на площині, а під дією пружної сили мінімізується потенціальна енергія при її деформуванні.

Розглянемо алгоритм побудови кубічного інтерполяційного сплайна дефекту 1. На кожному відрізку $[x_{i-1}, x_i]$ сплайн записується у вигляді

$$\varphi_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \quad i = \overline{1,n}. \quad (9.23)$$

Тому потрібно знайти $4n$ коефіцієнтів a_i, b_i, c_i та d_i. Для цього маємо $3(n-1)$ умов неперервності сплайна

$$\varphi_{i-1}^{(j)}(x_{i-1}) = \varphi_i^{(j)}(x_{i-1}), \; i = \overline{2,n}; \quad j = 0,1,2 \quad (9.24)$$

та $n+1$ умов інтерполювання 3^0, всього є $4n-2$ умов. Ще 2 умови задаються в точках x_0 і x_n, як крайові умови. Наприклад,

$$S''(x_0 + 0) = S''(x_n - 0) = 0, \quad (9.25)$$

або періодичні крайові умови

$$S^{(j)}(x_0 + 0) = S^{(j)}(x_n - 0), \quad j = 0,1,2 \; .$$

Зведемо задачу знаходження коефіцієнтів многочленів φ_i до розв'язування СЛАР з тридіагональною матрицею для невідомих

$c_1,..,c_n$, через які виражатимуться коефіцієнти d_i, b_i. Із умови 3^0 випливає, що $\varphi_1(x_0) = f_0$, $\varphi_i(x_i) = f_i$, $i = \overline{1,n}$. Тому

$$a_i = f_i, \ i = \overline{1,n}. \tag{9.26}$$

Крім того, при $x = x_0$ одержимо

$$a_1 - b_1 h_1 + c_1 h_1^2 - d_i h_1^3 = f_0, \tag{9.27}$$

Оскільки

$$\varphi_i'(x) = b_i + 2c_i(x - x_i) + 3d_i(x - x_i)^2,$$
$$\varphi_i''(x) = 2c_i + 6d_i(x - x_i),$$

то з умов неперервності (9.24) сплайна та першої та другої похідної маємо:

$$a_{i-1} = a_i - b_i h_i + c_i h_i^2 + d_i h_i^3, \tag{9.28}$$
$$b_{i-1} = b_i - 2c_i h_i + 3d_i h_i^2, \tag{9.29}$$
$$2c_{i-1} = 2c_i - 6d_i h_i, \ i = \overline{2,n}. \tag{9.30}$$

Із крайових умов (9.25) випливає, що $\varphi_i''(x_0) = 0$ і $\varphi_i''(x_n) = 0$, тому

$$c_1 - 3d_1 h_1 = 0, \ c_n = 0. \tag{9.31}$$

Уведемо фіктивну невідому $c_0 = 0$. Із (9.31) і з першої рівності (9.30) одержимо

$$d_i = \frac{c_i - c_{i-1}}{3h_i}, \quad i = \overline{1,n}. \tag{9.32}$$

Підставимо $a_i = f_i$ у рівність (9.28)

$$f_{i-1} = f_i - b_i h_i + c_i h_i^2 - d_i h_i^3, \ i = \overline{2,n}.$$

Звідси маємо

$$b_i = f(x_i, x_{i-1}) + c_i h_i - d_i h_i^2, \ i = \overline{2,n}, \tag{9.33}$$

де $f(x_i, x_{i-1}) = f(x_i) - f(x_{i-1}/h_i$. Урахувавши вирази для d_i із (9.32) і (9.33), одержимо, що $b_i = f(x_i, x_{i-1}) + c_i h_i - \frac{c_i - c_{i-1}}{3} h_i$ або

$$b_i = f(x_i, x_{i-1}) + \frac{h_i}{3}(2c_i + c_{i-1}), \quad i = \overline{2,n}. \tag{9.34}$$

Підставимо тепер значення b_i при $k = i - 1$ та i в (9.29). Тоді

$$f(x_{i-1}, x_{i-2}) + \frac{2}{3}c_{i-1}h_{i-1} + \frac{1}{3}c_{i-2}h_{i-1} =$$

$$= f(x_i, x_{i-1}) + \frac{2}{3}c_i h_i + \frac{1}{3}c_{i-1}h_i - 2c_i h_i + (c_i - c_{i-1})h_i.$$

Після перетворень одержимо

$$h_{i-1}c_{i-2} + 2(h_{i-1} + h_i)c_{i-1} + h_i c_i = 3(f(x_i, x_{i-1}) - f(x_{i-1}, x_{i-2}),\ i = \overline{2,n},$$
$$c_0 = c_n = 0. \qquad (9.35)$$

Отже, маємо СЛАР із тридіагональною матрицею (9.35), причому з перевагою головної діагоналі. Тому існує єдиний розв'язок цієї системи, який можна знайти методом прогонки.

Маючи значення $c_1, \ldots, c_{n-1}$, за формулою (9.27) і (9.34) знаходимо значення b_i, $i = \overline{1,n}$. За формулами (9.32) обчислюються d_i, з (9.26) знаходяться a_i, $i = \overline{1,n}$.

Наслідок 9.3. *Нехай* Δ *– рівномірна сітка на* $[a, b]$. *Тоді* $h_i = h$ *і СЛАР (9.35) набуває простішого вигляду*

$$c_{i-2} + 4c_{i-1} + c_i = 6f(x_{i-2}, x_{i-1}, x_i),\ i = \overline{2,n};$$
$$c_0 = c_n = 0.$$

Тут через $f(x_{i-2}, x_{i-1}, x_i)$ *– поділені різниці другого порядку. Коефіцієнтів* a_i, d_i *і* b_i *обчислюються за формулами:*

$$a_i = f(x_i), \quad d_i = \frac{c_i - c_{i-1}}{3h},\ i = \overline{1,n},$$
$$b_i = f(x_{i-1}, x_i) + \frac{h}{3}(2c_i + c_{i-1}),\ i = \overline{1,n}.$$

Нехай $\tilde{C}^2[a,b]$ – простір $(b-a)$-періодичних функцій, які мають неперервну другу похідну $\|f\| = \max\limits_{a \le x \le b} |f(x)|$.

Теорема 9.5 [26, с. 12–13]. *Нехай кубічний сплайн* $S_3(x)$ *інтерполює у вузлах рівномірної сітки* Δ *функцію* $f \in \tilde{C}^2[a,b]$ *і задовольняє періодичні крайові умови*

$$S_3''(x_0) = S''(x_n), \quad S_3''(x_1) = S''(x_{n+1}),\ x_{n+1} = x_n + h,\ y_{n+1} = y_n.$$

Тоді для похибок апроксимації справджуються оцінки

$$\|f^{(i)} - S_3^{(i)}\| \le K_i h^{2-i}\omega(h, f''),\ i = 0,1,2, \quad 8K_0 - K_1 = K_2 = 5. \qquad \blacksquare$$

Зокрема, із теореми 9.6 випливає, що $\|f - S_3\| \le 0.625h^2\omega(h, f'')$.

При наближенні кубічним сплайном $S_3(x)$ функції Рунге при розбитті відрізка на n рівних частини, як показано в праці [10, с. 50], максимальна похибка інтерполювання спадає:

n	$\max\limits_{[-1,1]} \lvert f(x) - S_3(x) \rvert$		
2	0.9246	12	0.02932
4	0.5407	14	0.01661
6	0.2500	16	0.01000
8	0.5562	18	0.006339
10	0.09562	20	0.004195

Кубічні інтерполяційні сплайни володіють важливими екстремальними властивостями. Одна з них пов'язана з тим, що профіль гнучкої лінійки, закріпленої в заданих точках (x_i, y_i), $i = \overline{0,n}$, і задовольняє крайові умови $S''(x_0) = S''(x_n) = 0$, набуває форми, при якій потенціальна енергія лінійки мінімальна. Математично в лінійному випадку це зводиться до нерівності [26]

$$\int_a^b (S_3(x))^2 dx \le \int_a^b (f(x))^2 dx,$$

причому рівність досягається тільки для $f(x) = S_3(x)$.

Нехай $W_2^2[a,b]$ – простір визначених на $[a,b]$ функцій, друга похідна яких інтегрована з квадратом. Розглянемо задачу про мінімізацію функціонала

$$I(f) = \int_a^b (f''(x))^2 dx$$

на множині допустимих функцій $W_2^2[a,b]$.

Теорема 9.6 (екстремальна властивість кубічних сплайнів) [26]. Серед усіх функцій $f \in W_2^2[a,b]$, які інтерполюють значення y_i, $i = \overline{0,n}$, кубічний інтерполяційний сплайн $S_3(x)$ із крайовими умовами (9.25) мінімізує функціонал $I(f)$.

Наслідок 9.4. *У лінійному випадку гнучка стальна лінійка, закріплена в точках* (x_i, y_i), $i = \overline{0,n}$, *набуває форми кубічного інтерполяційного сплайна.* ■

Побудова параболічних інтерполяційних сплайнів має певні особливості. Можна показати, що при деяких умовах, наприклад, при заданні періодичних крайових умов, сплайн не завжди існує. По-друге, для сітки Δ число параметрів дорівнює 3*n,* а кількість

умов (неперервності й інтерполювання) складає $2(n-1)+(n+1)=$ $=3n-1,$ тобто є один вільний параметр. Якщо задавати крайову умову на одному з кінців відрізка, то можна отримати нестійкий обчислювальний процес (швидке зростання похибки і сильну осциляцію сплайна).

У зв'язку з цим вузли параболічного сплайна визначаються сіткою $\bar{\Delta}=\{\bar{x}_i : a=\bar{x}_0<\bar{x}_1<...<\bar{x}_n<b=x_{n+1}\},$ де $x_{i-1}<\bar{x}_i<x_i.$ Зокрема, на рівномірній сітці можна взяти $\bar{x}_i=(x_{i-1}+x_i)/2.$

Функція S_2 називається параболічним сплайном, який інтерполює функцію f на сітці $\bar{\Delta}$, якщо виконуються умови:

1. $S_2 \in P_2, \quad x\in[\bar{x}_{i-1},\bar{x}_i], \quad i=\overline{1,n}.$
2. $S_2 \in C^1[a,b].$
3. $S_2(x_i)=y_i, \quad i=\overline{0,n}.$

При такому виборі сітки існує єдиний інтерполяційний параболічний сплайн [40, 66]. Якщо $f\in C^1[a,b]$ і сплайн періодичний, тобто $S^{(k)}(a+0)=S^{(k)}(b-0), k=0,1,$ то

$$\left|f^{(j)}(x)-S^{(j)}(x)\right|\le K_j h^{1-j}\omega(h,f'),\ j=0,1,\ \forall x\in[a,b],$$

де $2K_0=K_1=5.$

9.12. Середньоквадратичні наближення

9.12.1. Найліпше середньоквадратичне наближення (НСК). Інтерполювання зручно застосовувати при невеликому числі вузлів, значення функції в яких задано точно. Якщо ж ці значення відомі з похибками, наприклад є результатами експерименту чи спостереження, і кількість їх досить велика, то будується НСК.

Нехай функцію $y = f(x)$ задано значеннями $y_0,...,y_n$, якими можуть бути результати обчислення, спостережень або вимірювань у вузлах $x_0,...,x_n$. Припустимо, що $x_0<x_1<...<x_n$. Побудуємо наближення функції у вигляді

$$\varphi(x;a_0,...,a_m)=\sum_{j=0}^{m}a_j\varphi_j(x), \qquad (9.36)$$

де $m<n$, а функції $\varphi_i(x)$ − лінійно незалежні на $[x_0,x_1]$ і досить прості для обчислення функції. Вибір функції φ може ґрунтуватись на аналізі результатів табличного задання функції.

Такою може бути лінійна $\varphi = a_0 + a_1 x$ або квадратична функція $\varphi = a_0 + a_1 x + a_2 x^2$. Коефіцієнти a_i вибираються так, щоб мінімізувати суму квадратів відхилень $y_i - \varphi(x_i; a_0, a_1, ..., a_m)$ між емпіричними і теоретичними значеннями. Тобто потрібно знайти такі значення коефіцієнтів $a_0, a_1, ..., a_m$, щоб

$$\Delta(a_0, a_1, ..., a_m) := \sum_{i=0}^{n} [y_i - \varphi(x_i; a_0, a_1, ..., a_m)]^2 = \min.. \qquad (9.37)$$

Якщо відомі оцінки похибки ε_i значень y_i функції $y = f(x)$, то критерій (9.37) виглядає та:

$$\sum_{i=0}^{n} \rho_i [y_i - \varphi(x_i; a_0, a_1, ..., a_m)]^2 = \min,$$

де вагові коефіцієнти $\rho_i > 0$ і характеризують точність: чим вища точність, тим більші їх значення. Здебільше $\rho_i = \varepsilon_i^{-2}$ [28].

Із необхідної умови екстремуму функції $\Delta(a_0, a_1, ..., a_m)$ одержимо систему рівнянь для знаходження коефіцієнтів $a_0, a_1, ..., a_m$:

$$\frac{\partial \Delta}{\partial a_j} = 0, \quad j = \overline{0, m}. \qquad (9.38)$$

Якщо ввести скалярний добуток

$$(\varphi_i, \varphi_j) = \sum_{k=0}^{n} \varphi_i(x_k) \varphi_j(x_k),$$

то система рівнянь набуває вигляду

$$\sum_{j=0}^{m} (\varphi_i, \varphi_j) a_j = (\varphi_i, f), i = \overline{0, m}. \qquad (9.39)$$

Матриця СЛАР є матрицею Грама, яка симетрична, а також додатно визначена, якщо система функцій $\{\varphi_i\}$ є чебишовською [28]. Нагадаємо, що система функцій $\{\varphi_0(x), ..., \varphi_n(x)\}$ називається системою Чебишева на $[a, b]$, якщо визначник

$$\begin{vmatrix} \varphi_0(x_0) & ... & \varphi_n(x_0) \\ ... & ... & ... \\ \varphi_0(x_n) & ... & \varphi_n(x_n) \end{vmatrix} \neq 0$$

$\forall \{x_0,...,x_n\}, x_i \in [a,b], x_i \neq x_j \quad \forall i \neq j$. Прикладами є алгебраїчна система $\left\{x^k, k=\overline{1,n}\right\}$, система $\{1,\cos x,\sin x,...,\sin nx,\cos nx\}$, многочлени Чебишева $T_k(x)=\cos(k\cdot\arccos x), x\in[-1,1], k=\overline{1,n}.$

Наведений спосіб знаходження апроксимуючої функції називається *методом найменших квадратів*[3].

Для алгебраїчної системи $\left\{x^k, k=\overline{1,n}\right\}$ СЛАР (9.38) набуває вигляду

$$\sum_{j=0}^{m}(x^i,x^j)a_j=(x^i,y), i=\overline{0,m},$$

де $(x^i,x^j)=\sum_{k=0}^{n}x_k^{i+j}, (x^i,y)=\sum_{k=0}^{n}x_k^i y_k.$

Оскільки при великих n система погано обумовлена, то в задачах побудови НСКН $n\approx 1-5.$

9.12.2. Приклади побудови НСКН. Побудуємо лінійне наближення

$$\varphi(x)=a_0+a_1x.$$

У цьому випадку

$$\Delta(a_0,a_1)=\sum_{1=0}^{n}[y_i-(a_0+a_1x_i)]^2$$

і СЛАР (9.39) набуває вигляду

$$\begin{aligned} a_0(n+1)+a_1\sum_{i-0}^{n}x_i&=\sum_{i=0}^{n}y_i,\\ a_0\sum_{i=0}^{n}x_i+a_1\sum_{i-0}^{n}x_i^2&=\sum_{i=0}^{n}x_iy_i. \end{aligned} \qquad (9.40)$$

Для квадратичної функції

$$\varphi(x;a_0,a_1,a_2)=a_0+a_1x+a_2x^2$$

маємо

$$\Delta(a_0,a_1,a_2)=\sum_{i=0}^{n}[y_i-(a_0+a_1x_i+a_2x_i^2)]^2.$$

Коефіцієнти a_0,a_1,a_2 знаходяться із СЛАР

[3] Розробка цього методу пов'язана з іменами К. Ґауса й А. Лежандра.

$$a_o n + a_1 \sum x_i + a_2 \sum x_i^2 = \sum y_i,$$
$$a_o \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 = \sum x_i y_i,$$
$$a_o \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4 = \sum x_i^2 y_i.$$

із симетричною матрицею.

9.12.3. Згладжування даних. Якщо значення y_i відомі зі значними похибками, то можна застосувати процедуру згладжування кількох сусідніх точок за допомогою НСКН із одним або двома коефіцієнтами. Центральній точці приписується значення, одержане в процесі апроксимації. Розглянемо приклад. Нехай значення y_i виміряні на рівномірній сітці з кроком h. За значеннями y_{k-1}, y_k, y_{k+1} побудуємо лінійне наближення $\varphi(x) = a_0 + a_1 x$. Із першого рівняння (9.40) одержимо

$$3a_0 + 3x_k a_1 = y_{k-1} + y_k + y_{k+1},$$

Звідки знаходимо

$$\varphi(x_k) = a_0 + a_1 x_k = (y_{k-1} + y_k + y_{k+1})/3.$$

У радіотехніці такий спосіб згладжування називається *фільтром*, оскільки він послаблює високочастотні коливання, мало впливаючи на високочастотні [28].

Нехай функція $f(x)$ періодична з періодом 2π. такими функціями описуються звукові, світлові та інші періодичні процеси. Якщо період функції $T > 0$, то функція $g(x) = f(Tx/2\pi)$, буде вже 2π-періодичною. Розглянемо питання наближення цієї функції тригонометричним многочленом

$$\varphi(x) = \sum_{k=0}^{M-1} a_k \exp(ikx). \qquad (9.41)$$

Припустимо, що в рівновіддалених вузлах

$$x_j = \frac{2j\pi}{N},\ j = \overline{0, N-1},\ N \geq M+1,$$

задано значення функції y_j функції f. Враховуючи ортогональність системи функцій $\{\exp(ikx)\}$ коефіцієнти a_k знаходяться в явному вигляді [28]

$$a_k = \frac{1}{N} \sum_{l=0}^{N-1} f(x_l) \exp(-ikx_l). \qquad (9.42)$$

Коефіцієнти (9.42) можна одержати також як результат числового інтегрування методом правих прямокутників інтегралів у коефіцієнтах розкладу функції $f(x)$ в ряд Фур'є.

Формулу (9.41) можна застосувати при великих M. Зручно взяти M кратним 6, зокрема для $n=6$ або 12 коефіцієнти (9.42) просто обчислюються.

Зауваження 9.1. *При застосуванні МНК виникає проблема вибору кількості функцій φ_i в (9.36). В окремих випадках такий вибір зрозумілий із аналізу таблиці значень функції або їх візуалізації, наприклад, коли значення y_i прилягають до деякої прямої, то $m=1$, $\varphi_0=1$, $\varphi_1=x$. У загальному випадку для фіксованої системи $\{\varphi_i\}_{i=0}^{m}$, починаємо з деякого $m\geq 0$, знаходимо коефіцієнти a_j й обчислюємо середньоквадратичне відхилення $\delta_m=\sqrt{\Delta/(m+1)}$. Якщо $\delta_m \gg \varepsilon$, то число коефіцієнтів недостатнє і m потрібно збільшити. Якщо $\delta_m \ll \varepsilon$, то старші коефіцієнти недостовірні й m потрібно зменшити. Коли досягається рівність $\delta_m \approx \varepsilon$, то вибір m оптимальний.*

Метод найменших квадратів завдяки широкій сфері застосування посідає виняткове серед методів математичної статистики[4]. Згідно з теоремою Ґауса-Маркова оцінка, одержана методом найменших квадратів, є найліпшою лінійною незміщеною оцінкою.

Приклади розв'язування типових задач

Задача 1. Побудувати інтерполяційний многочлен для наближення функції $y=(1+x)^{-1}$ за значеннями у вузлах $x_n=1+0.2n$, $n=0,1,...,5$.

Розв'язування. Оскільки кількість вузлів дорівнює 6, то інтерполяційний многочлен у формі Ньютона має степінь 5 і набуває вигляду

$$P_5(x)=0.5-0.227(x-1)+0.094(x-1)(x-1.2)-0.036(x-1)(x-1.2)*$$
$$*(x-1.4)+0.013(x-1)(x-1.2)(x-1.4)(x-1.6)-$$
$$+0.004(x-1)(x-1.2)(x-1.4)(x-1.6)(x-1.8).$$

[4] Лоусон Ч., Хенсон Р. Численное решение задач методом наименьших квадратов. – М.: Наука, 1986. – 232 с.

У точці $x=1.5$ маємо $L_5(1.5)=0.400000390234765$. Похибка інтерполювання складає $\left|f(1.5)-L_5(1.5)\right|=3{,}90234765246239\times10^{-7}$.

Задача 2. Проінтерполювати функцію $y=\sin x$ при x=23° за значеннями $\sin 10k, k=\overline{0,5}$.

Розв'язування. Застосуємо схему Ейткена. Результати обчислень наведено в табл. 9.2.

Таблиця 9.2.

k	x_k	l_k	$l_{k,k+1}$	$l_{k,k+1,k+2}$	$l_{k,k+1,k+2,k+3}$
0	10	0.17365			
1	20	0.34202	0.39253		
2	30	0.50000	0.38578	0.39051	
3	40	0.64279	0.37694	0.39019	0.39073
4	50	0.76604	0.36618	0.38990	0.39072
5	60	0.86603	0.35367	0.38962	0.39072

Спостерігається стабілізація значень в останньому стовпці. Згідно з формулою (9.16) за схемою Ейткена $\sin 23^\circ \approx$ $\approx l_{0,1,2}(23^0)=0.39073$. Відхилення від точного значення sin23° не перевищує 10^{-5}.

Задача 3. Побудувати інтерполяційний многочлен Ерміта для таких вхідних даних:

$$\begin{array}{ccc} x_0 & x_1 & x_2 \\ f_0 & f_1 & f_2 \\ & f_1' & \end{array}$$

Розв'язування. Многочлен Ерміта запишемо у вигляді

$$H_3(x)=f_0F_0(x)+f_1F_1(x)+f_2F_2(x)+f_1'F_{11}(x),$$

де $F_i(x)$ і $F_{11}(x)$ − многочлени степеня 3, причому

$$F_i(x_j)=\delta_{ij},\quad F_j'(x_1)=0,\quad F_{11}(x_j)=0,\quad F_{11}'(x_1)=1,\ i,j=0,1,2,$$

δ_{ij} − символ Кронекера. На підставі цих умов функції F_i і F_{11} виберемо у вигляді

$$F_0(x)=C_0(x-x_1)^2(x-x_2),\quad F_3(x)=C_2(x-x_1)^2(x-x_0),$$

$$F_2(x)=(x-x_0)(Ax+B)(x-x_2),\quad F_{11}(x)=C_1(x-x_0)(x-x_1)(x-x_2).$$

Згідно з умовами маємо:

$$C_0=1/((x_0-x_1)^2(x_0-x_2)),\quad C_2=1/((x_2-x_1)^2(x_2-x_0)),$$

$$C_1=1/((x_i-x_0)(x_1-x_2)).$$

Сталі A і B визначаються з системи рівнянь

$$(x_1 - x_0)(Ax_1 - B)(x_1 - x_2) = 1,$$

$$A(x_1 - x_0)(x_1 - x_2) + (Ax_1 + B)(2x_1 - x_0 - x_2) = 0,$$

розв'язком якої є

$$A = \frac{2x_1 - x_0 - x_2}{(x_1 - x_0)^2 (x_1 - x_2)^2}, \quad B = \frac{x_0 x_2 - x_1^2}{(x_1 - x_0)^2 (x_1 - x_2)^2}.$$

Завдання та запитання для самостійної роботи

1. Як можна мінімізувати залишковий член інтерполяційного многочлена Лагранжа?
2. Що зміниться у побудові інтерполяційного полінома за формулою Лагранжа, якщо поміняти систему вузлів, наприклад, додати ще один вузол?
3. Які властивості мають поділені різниці?
4. У яких випадках інтерполяційний многочлен краще використовувати у формі Лагранжа, а коли – у формі Ньютона?
5. Чим відрізняється постановка задачі інтерполювання від постановки задачі середньоквадратичного наближення? В яких випадках використовується кожен із методів?
6. Дати практичну оцінку похибки інтерполювання в ІМЛ та ІМН.
7. Дати геометричну ілюстрацію лінійної та квадратичної інтерполяції.
8. За таблицею значень функції $y = f(x)$

x_i	-1	1	2	3
$f(x_i)$	1	-1	0	-2

   1) побудувати інтерполяційний многочлен Лагранжа і Ньютона;
   2) побудувати лінійну функцію методом найменших квадратів, знайти середню похибку.
   3) побудувати лінійний інтерполяційний сплайн.
9. Довести, що якщо вузли інтерполювання розміщені симетрично відносно середини відрізка і значення в симетричних точках однакові, то інтерполяційний многочлен Лагранжа – парна функція.
10. Для функції $f(x) = x^{3/2}$ зі значеннями $f(0{,}1) = 6.08581$, $f(1{,}2) = 0.76071$, $f(2{,}7) = 0.92540$ і $f(4{,}9) = 0.09220$ побудувати узагальнений інтерполяційний многочлен Лагранжа, якщо $q(x) = 1/x$.
11. Показати, що для поділеної різниці виконується рівність

$$f(x_0,x_1,x_2)=\frac{f(x_0)}{(x_0-x_1)(x_0-x_2)}+\frac{f(x_1)}{(x_1-x_0)(x_1-x_2)}+\frac{f(x_2)}{(x_2-x_0)(x_2-x_1)}.$$

Довести аналогічну властивість для різниці $f(x_0,x_1,\ldots,x_n)$.

12. Порівняти інтерполяційні многочлени Лагранжа і Ньютона за кількістю арифметичних операцій, необхідних для інтерполювання функції в точці x.

13. Оцінити похибку наближення функції e^x многочленом Лагранжа $L_2(x)$, побудованого за вузлами $0,\ 0.1,\ 0.2$ в точках $x=0.05$ *і* $x=0.15$.

14. Обчислити наближене значення $f(x)=e^x$ для $x=1$ і $x=1/3$ за допомогою діагональної апроксимації Паде[5] $f^{[1,1]}(x)=\dfrac{P_p+P_1x}{1+q_1x}$,

якщо $f(-1.2)\approx 0.301,\quad f(0.5)\approx 1.649.$

15. З яким кроком потрібно скласти таблицю $\sin x$ на $[0,\pi/2]$, щоб похибка лінійної інтерполяції не перевищувала $0,5\cdot 10^{-6}$?

16. Побудувати многочлен $P_3(x)=a_0+a_1x+a_2x^2+a_3x^3$, який задовольняє умови: а) $a_2=1,\ P_3(0)=P_3(-1)=P_3(1)=0;$

б) $P_3(-1)=0,\ P_3(1)=1,\ P_3(2)=2,\ P_3'(1)=0.$

17. На проміжку [0;1] побудувати многочлен степеня $n=2$; 3 і 4, які найменше відхиляються від нуля.

18. Побудувати інтерполяційні многочлени у формі Ерміта за такими даними:

а)

x_0	x_1	x_2
$f(x_0)$	$f(x_1)$	$f(x_2)$
$f'(x_0)$	$f'(x_1)$	$f'(x_2)$

б)

x_0	x_1	x_2
$f(x_0)$	$f(x_1)$	$f(x_2)$

19. Для функції $f(x)=\sin x$ визначити крок сітки з рівновіддаленими вузлами для лінійного і квадратичного інтерполювання з точністю 10^{-6}.

20. Оцінити похибку апроксимації функції $f(x)=\sqrt{x}$ у точці $x=111.6$**, а** також на проміжку $[100,144]$ інтерполяційним многочленом $L_2(x)$ з вузлами $\{100,121,144\}$.

21. У табл. 4 задано значення функції $f(x)=2\sin(x\pi/2)+1$. Побудувати середньоквадратичне наближення вигляду $f(x)=\alpha\sin(x\pi/2)+\beta$ та

[5] Cohen H. Numerical Approximation Methods. – New York: Springer, London: Dordrecht Heidelberg, 2011. – 455 p.

обчислити похибку $E = \sum_{k=1}^{N} \varepsilon_k^2$.

Таблиця 9.3

x_i	$f(x_i)$
0.6	-0.895
1.1	0.011
1.4	2.990
1.9	0.691

22. Побудувати систему рівнянь у методі найменших квадратів для функціональної залежності вигляду $\varphi(x; a, b) = ae^{bx}$ і $\varphi(x; a, b, c) = a + b/x + c/x^2$. виконавши заміну так, щоб система нормальних рівнянь була лінійною.

22. Згідно із законом Гука, величина видовження пружини y лінійно залежить від маси $y = gm/k$, де $k-$ характеристика пружини. За результатами вимірювань

m_i	2.1	4.2	9.4	12.3
y_i	1.2	2.3	5.7	8.1

побудувати лінійне середньоквадратичне наближення і лінійний та кубічний інтерполяційні сплайни для величина видовження пружини.

23. Методом найменших квадратів знайти коефіцієнт C у третьому законі Кеплера $T = Cx^{3/2}$, де $x-$відстань планети від Сонця (млн. км.), $T-$ період проходження по орбіті (діб). Відомі дані (x,T) для чотирьох найближчих до Сонця планет Меркурія, Венери, Землі та Марса: (58;88), (108;225), (150;365), (228;687). Обчислити максимальну $\max\left|T_k - \overline{T_k}\right|$ і середню $\left(\sum\left|T_k - \overline{T_k}\right|\right)/N$ та середньоквадратичну $\left(\sum\left(T_k - \overline{T_k}\right)^2 / N\right)^{1/2}$ похибку, де T_k- результат спостережень, $\overline{T_k} = Cx_k^{3/2}, N = 4.$

24. Методом найменших квадратів знайти значення параметрів A і C у третьому законі Кеплера $T = Cx^A$ на підставі значень відстані x і періоду обертання T для дев'яти планет Сонячної системи.

25. За даними Міністерства фінансів України чисельність населення України на 1 січня в 2011–2018 рр. складала відповідно 45778.5, 45633.6, 45533.0, 45426.2, 42928.9, 42760.5, 42584.5 і 42388.4 тис.
Методом найменших квадратів побудувати функцію залежності чисельності населення від часу, підібравши відповідний вигляд цієї функції так, щоб середньоквадратична похибка не перевищувала 0.1.

Розділ 10. Числове диференціювання

Перша і друга різницеві похідні та їх точність. Побудова формул числового диференціювання методом невизначених коефіцієнтів та за допомогою інтерполяційних многочленів. Некоректність операції числового диференціювання. Апроксимація частинних похідних.

Література [13, 22, 28, 45, 59, 73, 78, 100]

10.1. Перша та друга різницеві похідні

Формули числового диференціювання (ФЧД) найчастіше застосовуються для наближеного обчислення значення похідних при табличному заданні функції $u = u(x)$ або при розв'язуванні диференціальних рівнянь. Наприклад, обчислення наближеного розв'язку крайової задачі

$$u'' + p(x)u' + q(x)u = r(x), \quad a < x < b; \tag{10.1}$$
$$u(a) = \mu_0,\ u(b) = \mu_1$$

в точках $x_n = a + nh,\ n = \overline{1, N-1},\ h = (b-a)/N$, зводиться до розв'язування відповідної системи різницевих рівнянь, яка одержується заміною похідних u'_n і u''_n у внутрішніх точках $x_1, ..., x_{N-1}$ різницевими похідними.

Нехай $\Delta_h[a,b] = \{x_n : x_n = x_0 + nh, \quad n = \overline{0, N}; \quad x_0 = a\}$ − рівномірна сітка на $[a,b]$ з кроком h. З означення похідної в точці x_n одержуються такі формули для її наближеного обчислення:

$$u_{x,n} = \frac{u_{n+1} - u_n}{h}, \tag{10.2}$$

$$u_{\bar{x},n} = \frac{u_n - u_{n-1}}{h}, \tag{10.3}$$

де $u_i = u(x_i)$. Формули (10.2) і (10.3) називаються відповідно *правою і лівою різницевою похідною*. Середнє різницевих похідних $u_{x,n}$ і $u_{\bar{x},n}$ визначає *центральну різницеву похідну*

$$u_{\dot{x},n} = \frac{u_{n+1} - u_{n-1}}{2h}. \tag{10.4}$$

Другу різницеву похідну можна одержати аналогічно на підставі перших різницевих похідних, як $u_{\bar{x}x,n} = (u_{x,n} - u_{\bar{x},n})/h.$

У підсумку одержимо формулу

$$u_{\bar{x}x,n} = \frac{u_{n+1} - 2u_n + u_{n-1}}{h^2}, \tag{10.5}$$

яка називається *центральною різницевою похідною*.

На підставі різницевих похідних (10.3) і (10.4) для крайової задачі (10.1) маємо різницеву задачу

$$\frac{1}{h^2}(y_{n-1} - 2y_n + y_{n+1}) + \frac{p_n}{2h}(y_{n+1} - y_{n-1}) + q_n y_n = r_n,$$

$$n = \overline{1, N-1}; \quad y_0 = \mu_0, \quad y_N = \mu_1,$$

тобто СЛАР із тридіагональною матрицею для знаходження наближеного розв'язку.

Знайдемо похибки різницевих похідних (10.2) − (10.5) та їх оцінки. Нехай $u \in C^2[a,b]$, тоді за формулою Тейлора

$$u_{x,n} - u'_n = \frac{1}{h}(u_n + hu'_n + \frac{1}{2}h^2 u''(x_n + \theta_1 h) - u_n) - u'_n = \frac{1}{2}hu''(x_n + \theta_1 h),$$

де $\theta_1 \in (0,1)$. Оскільки функція u'' обмежена на $[a,b]$, то

$$\left|u_{x,n} - u'_n\right| \le M_2 h/2,$$

де $\max\left|u''(x)\right| \le M_2$. Така ж оцінка правильна і для похибки лівої різницевої похідної. Якщо $u \in C^3[a,b]$, то маємо

$$u_{x,n} - u'_n = u''(x_n + \theta_2 h)h^2/6, \quad \theta_2 \in (-1,1) \tag{10.6}$$

і відповідну оцінку

$$\left|u_{\dot{x},n} - u'_n\right| \le M_3 h^2/6.$$

Похибку та її оцінку для другої різницевої похідної $u_{\bar{x}x,n}$ можна одержати, припустивши, що $u \in C^4[a,b]$. Тоді

$$u_{n\pm1} = u_n \pm hu'_n + \frac{h^2}{2}u''_n \pm \frac{h^3}{6}u'''_n + \frac{h^4}{24}u^{(4)}(x_n + \theta_3^{\pm} h),$$

де θ_3^+ і θ_3^- − деякі числа з інтервалу $(0,1)$ і $(-1,0)$ відповідно. Підставивши $u_{n\pm1}$ у вираз для різницевої похідної $u_{\bar{x}x,n}$, одержимо

$$u_{\bar{x}x,n} - u''_n = \frac{h^2}{24}(u^{(4)}(x_n + \theta_3^+ h) + u^{(4)}(x_n + \theta_3^- h)).$$

Оскільки похідна $u^{(4)}(x)$ неперервна, то

$$u_{\bar{x}x,n} - u''_n = \frac{h^2}{12}u^{(4)}(x_n + \theta h), \quad \theta \in (-1,1). \tag{10.7}$$

Звідси випливає, що

$$\left|u_{\bar{x}x,n} - u''_n\right| \le M_4 h^2/12,$$

де сталою M_4 обмежений модуль похідної $u^{(4)}(x)$ на $x \in [a,b]$.

Величина $h^2 u_n^{(4)}/12$ – *головна складова похибки*. Для різницевих похідних (10.1) – (10.3) маємо ГСП $hu_n''/2$, $-hu_n''/2$ і $h^2 u_n''/6$ відповідно.

Підсумкові результати наведені в табл. 10.1.

Таблиця 10.1

Основні перша та друга різницеві похідні

Формула числового диференціювання	Позначення	Порядок	Похибка r_n
$(u_{n+1}-u_n)/h = u_n' + r_n$	$u_{x,n}$	1	$hu''(\xi_n)/2$
$(u_n-u_{n-1})/h = u_n' + r_n$	$u_{\bar{x},n}$	1	$-hu''(\xi_n)/2$
$(u_{n+1}-u_{n-1})/2h = u_n' + r_n$	$u_{\dot{x},n}$	2	$h^2u'''(\xi_n)/6$
$(-u_2+4u_1-3u_0)/2h = u_0' + r_n$	$u_{x,0}$	2	$h^2u'''(\xi_n)/3$
$(3u_N-4u_{N-1}+u_{N-2})/2h = u_N' + r_n$	$u_{\bar{x},N}$	2	$-h^2u'''(\xi_n)/3$
$(u_{n-1}-2u_n+u_{n+1})/h^2 = u_n'' + r_n$	$u_{\bar{x}x,n}$	2	$h^2u^{(4)}(\xi_n)/12$

10.2. Побудова формул числового диференціювання методом невизначених коефіцієнтів

ФЧД можна одержати підбором коефіцієнтів лінійної комбінації значень функції u на сітці Δ_h у відповідній формулі так, щоб похибка апроксимації похідної мала найвищий порядок щодо кроку сітки h. Розглянемо два приклади. Побудуємо різницеву похідну для правої похідної u_0' за вузлами x_0, x_1 і x_2 у вигляді

$$u_0' \approx u_{x,0} = \frac{1}{h}(au_0 + bu_1 + cu_2).$$

Нехай $u \in C^3[a,b]$, тоді похибка ФЧД

$$u_{x,0} - u_0' = \frac{1}{h}\Big[\ au_0 + b\Big(u_0 + hu_0' + \frac{h^2}{2}u_0'' + \frac{h^3}{6}u_0'''\Big) + \ldots\Big) +$$

$$+ c\Big(u_0 + 2hu_0' + \frac{(2h)^2}{2}u_0'' + \frac{(2h)^3}{6}u_0'''\Big)\Big] - u_0' =$$

$$= \frac{u_0}{h}(a+b+c) + u_0'(b+2c-1) + \frac{u_0''h}{2}(b+4c) + \frac{u_0'''h}{2}(b+8c) + \ldots$$

Другий порядок апроксимації досягається, якщо

$$a+b+c=0,\quad b+2c=1,\ b+4c=0.$$

Звідси маємо: $a=-3/2,\quad b=2,\quad c=-1/2$ і різницеву похідну

$$u_0' \approx \bar{u}_{x,0} := \frac{1}{2h}(-3u_0+4u_1-u_2). \tag{10.8}$$

Оскільки $b+8c=-2,$ то найвищий порядок різницевих похідних такого вигляду другий, ГСП дорівнює $-h^2u_0'''/3.$

Аналогічно одержується різницева похідна другого порядку для лівої похідної u_N'

$$u_N \approx \bar{u}_{x,N} := \frac{1}{2h}(3u_N-4u_{N-1}+u_{N-2}). \tag{10.9}$$

Якщо $u\in C^4[a,b]$, то серед формул вигляду

$$\frac{1}{h^2}(au_{n+1}+bu_n+cu_{n-1}),$$

які апроксимують другу похідну, найвищий, другий порядок має різницева похідна (10.5), коли $a=c=1$ і $b=-2$ є розв'язком системи лінійних рівнянь

$$a+b+c=0,\ a-c=0,\ a+c=2.$$

10.3. Застосування інтерполяційних формул

Для функції u на сітці Δ_h побудуємо інтерполяційний многочлен Лагранжа L_n і похідну u_n' апроксимуємо похідною L_n'. Аналогічно, можна скористатись наближеними рівностями похідні $u''(x)\approx L_n''(x),\ u^{(k)}(x)\approx L_n^{(k)}, k\le n.$ Як приклад, розглянемо ФЧД, які одержуються на підставі многочлена Лагранжа $L_{2,n}(x)$, побудованого за вузлами x_{n-1}, x_n, x_{n+1}. Маємо:

$$L_{2,n}(x)=u_{n-1}\frac{(x-x_n)(x-x_{n+1})}{(x_{n-1}-x_n)(x_{n-1}-x_{n+1})}+u_n\frac{(x-x_{n-1})(x-x_{n+1})}{(x_n-x_{n-1})(x_n-x_{n+1})}+$$
$$+\ u_{n+1}\frac{(x-x_{n-1})(x-x_n)}{(x_{n+1}-x_{n-1})(x_{n+1}-x_n)}.$$

Розглянемо випадок рівномірної сітки з кроком h. Оскільки $x_{n+1}-x_n=x_n-x_{n-1}=h,\ \ x_{n+1}-x_{n-1}=2h,$ то

$$L_{2,n}(x)=\frac{1}{2h^2}[u_{n-1}(x-x_n)(x-x_{n+1})-2u_n(x-x_{n-1})(x-x_{n+1})+$$
$$+u_{n+1}(x-x_{n-1})(x-x_n)].$$

Нехай $x_{n\pm 1/2} := x_n \pm h/2$, тоді

$$L'_{2,n}(x) = \frac{1}{2h^2}[u_{n-1}(2x - x_n - x_{n+1}) - 2u_n(2x - x_{n-1} - x_{n+2}) + u_{n+1}(2x - x_{n-1} - x_n)] =$$

$$= \frac{1}{h^2}\left[u_{n-1}\left(x - \frac{x_n - x_{n+1}}{2}\right) - u_n(x - x_n) + u_{n+1}\left(x - \frac{x_{n-1} + x_n}{2}\right) - u_n(x - x_n)\right] =$$

$$= \frac{1}{h}[(x - x_{n-1/2})\frac{u_{n+1} - u_n}{h} + (x_{n+1/2} - x)\frac{u_n - u_{n-1}}{h}]$$

Оскільки $x_n = (x_{n+1} + x_{n-1})/2$, то

$$u'(x) \approx L'_{2,n}(x) = \frac{1}{h}[(x - x_{n-1/2})u_{x,n} + (x_{n+1/2} - x)u_{\bar{x},n}].$$

За одержаною формулою можна обчислити наближене значення похідної $u'(x)$ у довільній точці x. Для $x = x_n$ одержимо центральну різницеву похідну (10.3). Якщо $x = x_{n+1}$ або x_{n+1}, то маємо різницеві похідні вигляду (10.8) і (10.9) відповідно.

Друга похідна $u''(x) \approx L''_{2,n}(x) = (u_{x,n} - u_{\bar{x},n})/h$. Для $x = x_n$ одержимо різницеву похідну $u_{\bar{x}x,n}$, порядок якої другий, що випливає з рівності $u_{\bar{x}x,n} - u''(x) = (x_n - x)u''_n + \frac{1}{12}h^2u_n^{(4)} + h^3...$ У випадку, коли $x \neq x_n$, маємо для $u''(x)$ перший порядок апроксимації.

10.4. Некоректність операції числового диференціювання

Значення функції $u(x)$ у вузлах сітки може обчислюватись з похибкою, зокрема внаслідок заокруглення. У деяких випадках похибка обчислення різницевих похідних може значно перевищувати похибку обчислення значення функції $u(x)$ і зростати при $h \to 0$. У цьому сенсі операція числового диференціювання (на відміну, наприклад, від числового інтегрування) є некоректною. Розглянемо це на прикладі обчислення різницевої похідної $u_{x,n} = (u_{n+1} - u_n)/h$. Нехай замість точних значень u_i функції u маємо наближені значення $\bar{u}_i = u_i + \varepsilon_i$, $|\varepsilon_i| \le \varepsilon$, $i = n, n+1$. Тоді замість $u_{x,n}$ обчислюється значення

$$\bar{u}_{x,n} = u_{x,n} + \frac{\varepsilon_{n+1} - \varepsilon_n}{h}.$$

Похибка обчислення $u_{x,n}$ складає $\varepsilon_{x,n} = \bar{u}_{x,n} - u_{x,n} = (\varepsilon_{n+1} - \varepsilon_n)/h$. Тоді

$$\left|\varepsilon_{x,n}\right| \le 2\varepsilon/h, \qquad (10.10)$$

причому $\varepsilon_{x,n} = 2\varepsilon/h$, коли $\varepsilon_{n+1} = -\varepsilon_n = \varepsilon$.

Якщо ε не залежить від h, то при $h \to 0$ похибка $\varepsilon_{x,n}$ необмежено зростає. При малих h похибка $\varepsilon_{x,n}$ може набагато перевищувати значення різницевої похідної $u_{x,n}$. Тому, щоб не допустити зменшення точності обчислення $u_{x,n}$ похибка заокруглення $\varepsilon_{x,n}$ має не перевищувати похибки апроксимації. Враховуючи оцінку (10.5), одержимо

$$\left|\varepsilon_{x,n}\right| \le 2\varepsilon/h \le M_2 h/2. \qquad (10.11)$$

Звідси маємо

$$\varepsilon \le M_2 h^2/4.$$

Якщо ε фіксоване, то згідно з (10.11), обчислення потрібно проводити з кроком, що задовольняє нерівність

$$h \ge h_0 = 2\sqrt{\varepsilon/M_2}. \qquad (10.12)$$

Отже, сумарна похибка $r(h) \le f(h) = 2\varepsilon/h + M_2 h/2$, а оптимальне значення кроку h_0 досягається у точці мінімуму функції $f(h)$ (рис. 10.1).

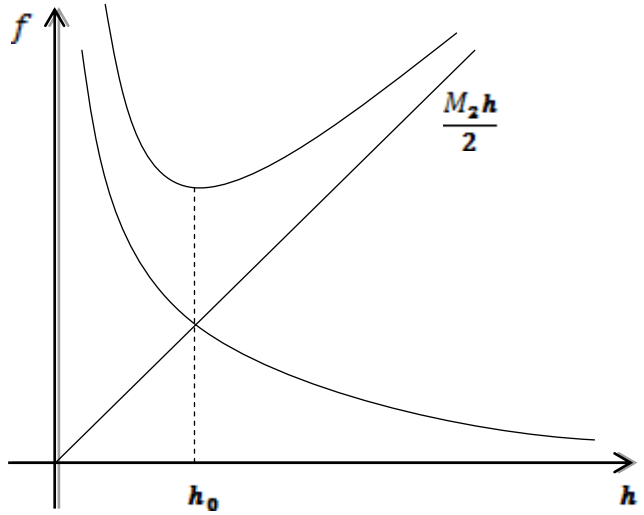


Рис. 10.1. Сумарна похибка ФЧД $u_{x,n}$

Нехай $\varepsilon = 10^{-7}$, $M_2 = 40$. Тоді, згідно з (10.12), оптимальний крок сітки складає $h = h_0 = 2\sqrt{\varepsilon/M_2} = 10^{-4}$. Якщо задати крок

$h=10^{-6}$, то $\varepsilon \le 10h^2 = 10^{-11}$. Така точність досягається при використанні даних типу *double* на мові С.

При обчисленні різницевих похідних вищого порядку вплив похибок заокруглення ще вагоміший, оскільки знаменник різницевих формул складає h^k, $k>1$. Наприклад, для різницевої похідної $u_{\bar{x}x,n}$ похибкою заокруглення є величина $O(\varepsilon / h^2)$, і в цьому випадку потрібно вимагати, щоб $\varepsilon = O(h^4)$.

10.5. Підсумкові зауваження

1. Різницеві похідні другого і вищих порядків. На чотириточковому шаблоні x_0, x_1, x_2, x_3 з кроком h для першої похідної у вузлах сітки маємо ФЧД:

$$u_0' \approx \frac{1}{6h}\left(-11u_0 + 18u_1 - 9u_2 + 2u_3\right), \tag{10.13}$$

$$u_1' \approx \frac{1}{6h}\left(-2u_0 - 3u_1 + 6u_2 - u_3\right), \tag{10.14}$$

$$u_2' \approx \frac{1}{6h}\left(u_0 - 6u_1 + 3u_2 + 2u_3\right), \tag{10.15}$$

$$u_3' \approx \frac{1}{6h}\left(-2u_0 + 9u_1 - 18u_2 + 11u_3\right), \tag{10.16}$$

$$u_2' = \frac{1}{12h}\left(u_0 - 8u_1 + 8u_3 - u_4\right). \tag{10.17}$$

Оцінка похибки різницевих похідних (10.13) і (10.16) складає $M_4 h^3 / 4$, (10.14) і (10.15) – $M_4 h^3 / 12$, різницева похідна (10.17) має четвертий порядок точності.

Різницеві похідні другого порядку на такому ж шаблоні мають вигляд:

$$u_0'' \approx \frac{1}{h^2}\left(2u_0 - 5u_1 + u_2 - u_3\right), \tag{10.18}$$

$$u_1'' \approx \frac{1}{h^2}\left(u_0 - 2u_1 + u_2\right), \tag{10.19}$$

$$u_2'' \approx \frac{1}{h^2}\left(-u_0 + 16u_1 - 30u_2 + 16u_3 - u_4\right). \tag{10.20}$$

Оцінка похибки формул (10.18) дорівнює $11h^2 M_4 / 12$, а (10.19) $h^2 M_4 / 12$. ФЧД (10.20) має третій порядок точності.

Формули наближеного обчислення третьої похідної

$$u_2^{(3)} \approx \left(-u_0 + 3u_1 - 3u_2 + u_3\right) / h^3, \qquad (10.21)$$

$$u_2^{(3)} \approx \left(-u_0 + 2u_1 - 2u_3 + u_4\right) / \left(12h^3\right), \qquad (10.22)$$

$$u_3^{(3)} \approx \left(u_0 - 8u_1 + 13u_2 - 13u_4 + 8u_5 - u_6\right) / \left(8h^3\right) \qquad (10.23)$$

мають відповідно перший, другий і четвертий порядок точності.

Для апроксимації четвертої похідної можна застосувати симетричні ФЧД

$$u_2^{(4)} \approx \left(u_0 - 4u_1 + 6u_2 - 4u_3 + u_4\right) / h^4, \qquad (10.24)$$

$$u_3^{(4)} \approx \left(-u_0 + 12u_1 - 39u_2 + 56u_3 - 39u_4 + 12u_5 - u_6\right) / \left(6h^4\right), \qquad (10.25)$$

які мають відповідно другий і четвертий порядок точності по h.

2. Застосування інтерполяційного многочлена Ньютона зі скінченними різницями. У цьому випадку

$$f(x) \approx N_n(x_0 + th) = y_0 + t\Delta y_0 + \frac{t(t-1)}{2!}\Delta^2 y_0 + \ldots$$

$$+ \frac{t(t-1)\ldots(t-n+1)}{n!}\Delta^n y_0,\ t = (x - x_0)/h$$

Після диференціювання одержимо:

$$f'(x) \approx \frac{1}{h}(\Delta y_0 + \frac{1}{2}(2t-1)\Delta^2 y_0 + \frac{1}{6}(3t^2 - 6t + 2)\Delta^3 y_0 +$$

$$\frac{1}{12}(2t^3 - 9t^2 + 11t - 3)\Delta^4 y_0 + \ldots),$$

$$f''(x) \approx \frac{1}{h^2}(\Delta^2 y_0 + (t-6)\Delta^3 y_0 + \frac{1}{12}(6t^2 - 18t + 11)\Delta^4 y_0 + \ldots),$$

де $\Delta y_0 = y_1 - y_0$, $\Delta^2 y_0 = \Delta y_1 - \Delta y_0 = y_2 - 2y_1 + y_0$ і т.д. Число доданків у цих формулах залежить від кількості вузлів, які використовуються для обчислення похідних.

3. Підвищення точності апроксимації. Точність ФЧД можна підвищити, збільшуючи кількість вузлів. Але це веде до зростання обсягу обчислень. При фіксованій кількості вузлів уточнити результат можна за формулою Рунге, у більш загальному випадку – за формулою Ромберга [6, 13, 28, 43].

Нехай $g(x)$ − похідна функції, яку апроксимуємо різницевою формулою $\varphi(x,h)$, а $h^P\psi(x)$ − ГСП, тобто

$$g(x) = \varphi(x,h) + h^P\psi(x) + o(h^P).$$

Обчислимо різницеву похідну з кроком qh, $1 < q \in N$. Тоді

$$g(x) = g(x, qh) + (qh)^P \psi(x) + \mathrm{O}((qh)^P).$$

Ігноруючи величини, порядок яких вищий h^P, із цих двох формул одержимо вираз для наближеного значення ГСП

$$h^P \psi(x) \approx \frac{\varphi(x,h) - \varphi(x,qh)}{q^p - 1}.$$

Підставивши знайдене значення у вираз для $g(x)$, одержимо формулу Рунге

$$g(x) \approx \varphi(x,h) + \frac{\varphi(x,h) - \varphi(x,qh)}{q^P - 1}. \qquad (10.26)$$

Ця формула (10.26) дозволяє за результатами двох значень $\varphi(x,h)$ і $\varphi(x,qh)$ з порядком p знайти наближене значення похідної з порядком точності $p+1$. Наприклад, для різницевої похідної $u_{\bar{x}x,n}$, визначеної згідно з (10.4), для $k=2$ маємо

$$u_n'' = (4u_{\bar{x}x,n}(h) - u_{\bar{x}x,n}(2h))/3 + o(h^2).$$

4. Апроксимація частинних похідних. Розглянемо функцію двох змінних $u = u(x,y)$. Нехай $u_{ij} = u(x_i, y_j)$, $x_i = x_0 + ih$, $i = \overline{0,N}$; $y_j = y_0 + j\tau$, $j = \overline{0,M}$. Апроксимації частинних похідних з першим порядком набувають вигляду:

$$\left(\frac{\partial u}{\partial x}\right)_{ij} \approx \frac{u_{i+1,j} - u_{i,j}}{h}, \quad \left(\frac{\partial u}{\partial y}\right)_{ij} \approx \frac{u_{i,j+1} - u_{i,j}}{\tau}.$$

За аналогією з центральною різницевою похідною (10.4) можна одержати з похибкою $\mathrm{O}(h^2)$ різницеву похідну

$$\left(\frac{\partial u}{\partial x}\right)_{ij} \approx \frac{u_{i+1,j} - u_{i-1,j}}{2h}.$$

Апроксимація другої частинної похідної з оцінкою похибки $\mathrm{O}(h^2)$:

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{ij} \approx \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2},$$

Застосувавши інші розклади функції $u(x,y)$ в ряд Тейлора, можна вивести формули числового диференціювання з необхідним порядком апроксимації. Запишемо деякі з них:

$$\left(\frac{\partial^2 u}{\partial y^2}\right)_{ij} = \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\tau^2} + \mathrm{O}(\tau^2),$$

$$\left(\frac{\partial^2 u}{\partial x \partial y}\right)_{ij} = \frac{u_{i+1,j+1} - u_{i+1,j-1} - u_{i-1,j+1} + u_{i-1,j-1}}{4h\tau} + \mathrm{O}(h^2 + \tau^2).$$

Одержані формули можна використати для побудови різницевих схем для розв'язування рівнянь із частинними похідними. Розглянемо задачу поширення тепла в обмеженому однорідному стержні [6, 58, 59]

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t), \quad 0 < x < l, \quad t > 0;$$

$$u(0,x) = \varphi(x), \quad 0 \le x \le l;$$

$$u(t,0) = \mu_0(t), \quad u(t,l) = \mu_1(t),$$

де f, φ, μ_0, μ_1 – задані функції. Нехай $u_i^k = u(t_k, x_i) = u(k\tau, ih)$. На підставі апроксимацій похідних одержимо явну різницеву схему

$$\frac{1}{\tau}(y_i^{k+1} - y_i^k) = \frac{1}{h^2}(y_{i+1}^k - 2y_i^k + y_{i-1}^k) + f_i^k,$$

$$k = 0,1,..., \quad i = \overline{1, N-1}, \quad Nh = l;$$

$$y_i^0 = \phi_i, \quad i = \overline{1, N}; \quad y_0^k = \mu_0^k, \quad y_N^k = \mu_1^k, \quad k = 0,1,...$$

Звідси для кожного $k = 0,1,...$ явно знаходиться числовий розв'язок $y_i^{k=1}$, $i = \overline{1, N}$, для кожного $k = 0,1,...$

Розв'язування типових задач

Задача 1. Підвищити порядок ФЧД на підставі правої різницевої похідної з кроком h і $2h$.

Розв'язування. Користуючись виразом для похибки різницевої похідної $u_{x,i}$, для значень кроку сітки h і $2h$ одержимо

$$u_{x,i}^{(1)} = \left(u_{i+1} - u_i\right)/h = u_i' + hu_i''/2 + O\left(h^2\right),$$

$$u_{x,i}^{(2)} = \left(u_{i+2} - u_i\right)/2h = u_i' + hu_i'' + O\left(h^2\right).$$

Тоді з точністю до величин $O\left(h^2\right)$ одержимо

$$u_i' \approx u_{x,i}^{(2)} - 2u_{x,i}^{(1)} = \frac{1}{2h}\left(-3u_i + 4u_{i+1} - u_{i+2}\right).$$

Задача 2. Знайти оптимальний крок сітки для різницевої похідної $u_{\bar{x}x,n}$, яка апроксимує другу похідну u_n'', якщо точність обчислень значень функції $u(x)$ не перевищує ε.

Розв'язування. Нехай $u \in C^4[a,b]$. Згідно з оцінкою (10.7)

$$\left|u_n'' - u_{\bar{x}x,n}\right| \le \frac{M_4 h^2}{6}, \quad M_4 = \max_{[a,b]} \left|u^{(4)}(x)\right|.$$

Якщо замість точних значень u_i використаємо наближені значення $u_i + \varepsilon_i$, $|\varepsilon_i| \le \varepsilon$, то оцінка похибки обчислення $u_{\bar{x}x,n}$ складе

$$\left(|\varepsilon_{n+1}| + 2|\varepsilon_n| + |\varepsilon_{n-1}|\right)/h^2 \le 4\varepsilon/h^2 .$$

Сумарна похибка $\varphi(h) = M_4 h^2/6 + 4\varepsilon/h^2$. Мінімальне значення функції $\varphi(h)$, коли $h > 0$, досягається при $h_0 = 2\sqrt[4]{3\varepsilon}/\sqrt[4]{M_4}$.

Задача 3. Мовою програмування С з типами даних Float і Double обчислити похибки різницевих похідних (10.2)–(10.4) і (10.5) з кроком $h = 10^{-k}$ в точці $x_0 = 1$ для функції $f(x) = (x+1)^{-1}$.

Розв'язування. Результати обчислень наведені в табл. 10.2 і 10.3. Спостерігається спочатку зменшення похибки до певного значення кроку, а відтак досить різке зростання.

Таблиця 10.2

Похибки наближеного обчислення першої та другої похідних функції $f(x) = (x+1)^{-1}$ у точці $x_0 = 1$ з кроком 10^{-k}, тип даних *Float.*

k	Похибка $u_{x,0}$	Похибка $u_{\bar{x},0}$	Похибка $u_{\dot{x},0}$	Похибка $u_{\bar{x}x,0}$
1	0,0119047078	0,0131579607	0,0006266264	0,0006266862
2	0,0012440171	0,0012560405	0,0000060117	0,0000057732
3	0,0001132666	0,0001218406	0,0000042869	0,0148926457
4	0,0000289813	0,0000539896	0,0000414854	0,0000829771
5	0,0003382546	0,0003407614	0,0003395080	0,0006794771
6	0,0115815345	0,0033198686	0,0041308329	14901,153222

Для другої різницевої похідної $u_{\bar{x}x}$ це зростання значно швидше, ніж для перших похідних, для $h = 0.01$ досягається найменше значення похибки 0,00000577, а при $h = 10^{-6}$ її значення досягає 14901,15.

Таблиця 10.3

Похибка наближеного обчислення першої та другої похідних функції $f(x)=(x+1)^{-1}$ у точці $x_0=1$ з кроком 10^{-k}, тип даних *Double*

	$u_{x,0}$	$u_{\bar{x},0}$	$u_{\dot{x},0}$	$u_{\bar{x}\dot{x},0}$
1	0,0119047619	0,0131578947	0,0006265994	0,0006265664
2	0,0012437810	0,0012562814	0,0000061297	0,0000062501
3	0,0001249375	0,0001250625	0,0000015484	0,0000000625
4	0,0000124993	0,0000125006	0,0000207451	0,0000000006
5	0,0000012499	0,0000012500	0,0001697557	0,0000002774
6	0,0000001250	0,0000000125	0,0016598718	0,0000277545
7	0,0000000123	0,0000000123	0,0240116146	0,0027802206
8	0,0000000027	0,0000000025	0,1250000013	0,2774643569
9	0,0000000205	0,0000000069	0,1249999897	27,788735181
10	0,0000000206	0,0000000206	0,1249999896	25000,00000

Задача 4. Задано таблицю значень функції $y=1/x$

Таблиця 10.4

n	0	1	2	3	4
x_n	1.0	1.2	1.4	1.6	1.8
u_n	1.00000	0.83333	0.71429	0.62500	0.55556

1) Обчислити в точці $x=1.4$ значення першої різницевої похідної за формулами (10.2) – (10.4), (10.8), (10.9), (10.14) і (10.17).

2) У цій же точці обчислити другі різницеві похідні (10.18) і (10.20), третю похідну (10.22) та четверту похідну (10.24).

3) Порівняти одержані результати зі значеннями відповідних похідних у точці $x=1.4$.

Розв'язування. Значення похідних $u_2^{(k)}=u^{(k)}$ для $k=\overline{1,4}$ дорівнюють відповідно – 0.51020, 0.728863, –1.56185 і 4.46243.

Таблиця 10.5

Результати наближеного обчислення $u'(1.4)$

ФЧД	(10.2)	(10.3)	(10.4)	(10.8)	(10.9)	(10.15)	(10.21)
Порядок	1	1	2	2	2	3	4
Значення	-0.446	-0.595	-0.52128	-0.49602	-0.47619	-0.51256	-0.50925
Абсолютна похибка	0.064	0.085	0.0108	0.014	0.034	0.0024	0.00095
Відносна похибка (%)	16.66	12.50	2.1	2.7	6.6	0.463	0.186

Результати обчислень перших різницевих похідних згідно з відповідними формулами наведено в табл. 10.4.

Наведені результати ілюструють зростання точності обчислень зі зростанням порядку ФЧД і невеликий тренд значень у межах одного порядку. Зокрема, у ФЧД другого порядку (10.4) і (10.8) розбіжність відносної похибки складає 4.5%.

2) Значення різницевих похідних наближеного обчислення $u^{(k)}$ (1.4), $k = 2,3,4,$ наведено в табл. 10.6.

Таблиця 10.6

Наближені значення похідних u'', u''', $u^{(4)}$

ФЧД	u'' (10.29)		u''' (10.21)	$u^{(4)}$ (10.24)
	(10.19)	(10.20)	(10.22)	(10.24)
Порядок	2	2	2	2
Значення	0.743965	0.72751	-1.73625	4.9875
Абсолютна похибка	0.0151	0.00134	0.174	0.525
Відносна похибка (%)	2.07	0.18	10.04	10.53

Для другої різницевої похідної ріст точності узгоджується з ростом порядку ФЧД. Точність третьої та четвертої різницевих похідних значно нижча порівняно з ФЧД того ж порядку для першої та другої похідних.

Задача 5. На нерівномірній сітці з шаблоном (x_{n-1}, x_n, x_{n+1}) побудувати апроксимації першої і другої похідної.

Розв'язування. Нехай $h_{n+1} = x_{n+1} - x_n$, $h_n = x_n - x_{n-1}$, $h_{n+1} = x_{n+1} - x_n$, $\gamma_{n+1} = h_{n+1}/h_n$ – параметр регуляризації. Якщо $\gamma_{n+1} = 1\ \forall n$, то сітка – рівномірна. Припустимо, що $u \in C^3[a,b]$ і розкладемо за формулою Тейлора функцію $u(x)$ у точці x_{n-1} при $x = x_n$ і x_{n+1}. Одержимо:

$$u_n = u_{n-1} + h_n u'_{n-1} + h_n^2 u''_{n-1}/2 + h_n^3 u'''(\xi_n)/6,$$

$$u_{n+1} = u_{n-1} + h_n u'_{n-1} + h_n^2 u''_{n-1}/2 + h_n^3 u'''(\eta_n)/6,$$

де $\xi_n \in (x_{n-1}, x_n)$, $\eta_n \in (x_n, x_{n+1})$.

Вилучивши з одержаних рівностей u''_{n-1} і розв'язавши відносно u'_{n-1}, одержимо апроксимацію першої похідної у лівій крайній точці

$$\hat{u}_{\bar{x},n-1} = \frac{1}{h_{n+1}}\left[-\left(2+\gamma_{n+1}\right)u_{n-1} + \frac{\left(1+\gamma_{n+1}\right)^2}{\gamma_{n+1}}u_n - \frac{u_{n+1}}{\gamma_{n+1}}\right].$$

Якщо $h = const$, то $\gamma_{n+1} = 1$ і одержимо ФЧД (10.8).

Аналогічно, здійснивши розклад функції $u(x)$ в точках x_{n+1} і x_n, одержимо апроксимацію першої похідної в крайній правій і центральній точках відповідно:

$$\hat{u}_{x,n-1} = \frac{1}{h_{n+1}}\left[\gamma_{n-1}u_{n-1} - \frac{\left(1+\gamma_{n+1}\right)^2}{\gamma_{n+1}}u_n + \frac{\left(2+\gamma_{n+1}\right)}{\gamma_{n+1}}\right],$$

$$\hat{u}_{\overset{\circ}{x},n} = \frac{1}{h_{n+1}}\left[-\gamma_{n+1}u_{n-1} + \frac{\gamma_{n+1}^2 - 1}{\gamma_{n+1}}u_n + \frac{1}{\gamma_{n+1}}u_{n+1}\right].$$

Як показано в [40], похибки одержаних ФЧД мають вигляд:

$$\left|u'_{n-1} - \hat{u}_{\bar{x},n-1}\right| \le \frac{h^2}{6}\left(1+\gamma_{n+1}\right)M_{3,n}, \quad \left|u'_n - \hat{u}_{\overset{\circ}{x},n}\right| \le \frac{h^2}{6}\gamma_{n+1}M_{3,n},$$

$$\left|u'_{n+1} - \hat{u}_{x,n}\right| \le \frac{h^2}{6}\left(1+\gamma_{n+1}\right)\gamma_{n+1}M_{3,n}; \quad M_{3,n} = \max_{[x_{n-1},x_{n+1}]}\left|u'''(x)\right|.$$

Нехай тепер $u \in C^4[a,b]$. Розклавши функцію $u(x)$ для x_{n-1} і x_{n+1} у точці $x = x_n$ до четвертого порядку, додавши одержані вирази і скориставшись ФДЧ для $\hat{u}_{\overset{\circ}{x},n}$, одержимо [40]

$$\hat{u}_{\bar{x}x,n} = \frac{2}{h_n^2}\left[\frac{1}{1+\gamma_{n+1}}u_{n-1} - \frac{1}{\gamma_{n+1}}u_n + \frac{1}{\left(1+\gamma_{n+1}\right)\gamma_n}u_{n+1}\right].$$

Задача 6. Методом невизначених коефіцієнтів побудувати ФДЧ другого порядку (10.21) для наближеного обчислення u'''_2.

Розв'язування. За трьома вузлами x_0, x_1, x_2 на рівномірній сітці з кроком h одержується ФЧД (10.5) другого порядку для наближеного обчислення u''_1. Тому застосовуємо шаблон з 5 вузлів x_i, $i = \overline{0,4}$ і побудуємо ФЧД вигляду

$$u'''_2 \approx \nu_2 = \left(a_0u_0 + a_1u_1 + a_2u_2 + a_3u_3 + a_4u_4\right)/h^3.$$

Нехай $u \in C^5[x_0, x_4]$. Розклавши функцію $u(x+ih)$ в точці $x_2 = x_0 + 2h$ для $i = \pm 1$ і $i = \pm 2$ та прирівнявши коефіцієнти при $k = \overline{-1,3}$, одержимо СЛАР

$$\begin{aligned}
&a_0 + a_1 + a_2 + a_3 + a_4 = 0,\\
&-2a_0 - a_1 + a_3 + 2a_4 = 0,\\
&4a_0 + a_1 + a_3 + 4a_4 = 0,\\
&-8a_0 - a_1 + a_3 + 8a_4 = 1,\\
&16a_0 + a_1 + a_3 + 16a_4 = 0.
\end{aligned}$$

Розв'язок системи рівнянь: $a_4 = -a_0 = 1/12$, $a_3 = -a_1 = -2/12$, $a_2 = 0$, й одержується ФЧД (10.21). При цьому ГСП дорівнює $h^2 u_2^{(5)} / 24$.

Задача 7. За допомогою інтерполяційного многочлена Лагранжа $L_3(x)$ на рівномірній сітці з кроком h одержати ФЧД (10.13) і (10.15) для обчислення u_0' з порядком 3 і u_2' з порядком 4.

Розв'язування. Враховуючи, що $x_i - x_j = (i-j)h$, одержимо

$$L_3(x) = \big[-u_0(x-x_1)(x-x_2)(x-x_3) - 3u_2(x-x_0)(x-x_1)(x-x_3) -$$

$$-3u_2(x-x_0)(x-x_1)(x-x_3) + u_3(x-x_0)(x-x_1)(x-x_2)\big] / (6h^3).$$

Далі маємо

$$L_3'(x) = \left(\sum_{i=0}^{3} a_i u_i \sum_{i \neq j} (x - x_j)\right) \Big/ (6h^3),$$

де $a_0 = -a_3 = -1$, $a_1 = -a_2 = 3$.

Підставимо $x = x_2$. Тоді

$$u_2' \approx L_3(x_2) = \big[-u_0(x_2-x_1)(x_2-x_3) + 3u_1(x_2-x_0)(x_2-x_3) -$$
$$-3u_2\big((x_2-x_1)(x_2-x_3) + (x_2-x_0)(x_2-x_3) + (x_2-x_1)(x_2-x_3)\big) +$$
$$+u_3(x_2-x_0)(x_2-x_1)\big] / (6h^2) = (u_0 - 6u_1 + 3u_2 + 2u_3) / (6h^2).$$

Аналогічно, при $x = x_0$ одержимо ФЧД (10.13).

Завдання та запитання для самостійної роботи

1. Як можна апроксимувати першу і другу похідні на нерівномірній сітці?
2. Методом невизначених коефіцієнтів на рівномірній сітці з кроком h побудувати формули числового диференціювання порядку 2, 3 і 4 для наближеного обчислення похідних u'_n і u''_n.
3. З яким найвищим порядком можна апроксимувати похідну u'_n за допомогою різницевої формули

$$(au_{n-2} + bu_{n-1} + cu_n + du_{n+1} + ed_{n+2}) / h\,?$$

Знайти коефіцієнти цієї формули, якщо функція $u = u(x)$ диференційовна потрібне число разів.

4. За допомогою інтерполяційного многочлена Ньютона на рівномірній сітці з вузлами $x_i = x_0 + ih,\ i = \overline{0,3}$ побудувати:

а) ФЧД (10.13) – (10.17) для першої похідної;

б) ФЧД (10.18) – (10.20) для другої похідної.

5. Побудувати ФЧД (10.20), (10.22) на рівномірному п'ятиточковому шаблоні за допомогою інтерполяційного многочлена Ньютона або методом невизначених коефіцієнтів.
6. Одержати на рівномірній сітці формули четвертого порядку точності для першої і другої похідної вигляду:

$$u'_{5/2} = (-u_3 + 27u_2 - 27u_1 + u_0) / (24h) + O(h^4);$$

$$u''_2 = (-u_4 + 16u_3 - 30u_2 + 16u_1 - u_0) / (12h^2) + O(h^4).$$

6. Вивести формулу $u_{\bar{x}x\bar{x}x,n}$ для наближеного обчислення четвертої похідної $u_n^{(4)}$ на підставі різницевих похідних $u_{\bar{x}x,n}$ і $u_{\bar{x}x,n\pm1}$.
7. За значеннями функції $u = \sin x$ для $x = 0, \pi/6, \pi/3$ і $\pi/2$ знайти наближене значення похідних u' і u'' для $x = \pi/4$.
8. Дослідити зміну похибки апроксимації різницевих похідних $u_{x,n}$, $u_{\bar{x},n}$ і $u_{\bar{x}x,n}$ для функції $u = \cos x$ у точці $x = \pi/2$ зі зменшенням кроку h до мінімально можливого значення для різних типів даних і різних мов програмування.
9. Показати, що на рівномірній сітці формула

$$u'_n \approx \frac{3u_n - 4u_{n-1} + u_{n-2}}{3x_n - 4x_{n-1} + x_{n-2}}$$

має похибку $O(h^2)$ при $h \to 0$.

10. Дослідити похибку різницевих похідних (10.23) і (10.24).

11. Знайти величину оптимального кроку й оцінити точність, яка при цьому досягається, для різних апроксимацій першої і другої похідних.
12. Знайти оптимальний крок для ФЧД

$$u'(x) \approx (u(x-2h) - 8u(x-h) + 8u(x+h) - u(x+2h))/(12h),$$

яка має четвертий порядок, $|u^{(5)}(x)| \le M$ і значення функції обчислюється з точністю ε.

13. Задано таблиці значень функції. За допомогою ФЧД другого порядку обчислити $u'(x)$ і $u''(x)$ в заданих точках.

а)

x	$u(x)$	$u'(x)$	$u''(x)$
0.1	0.1564		
0.2	0.2812		
0.3	0.1920		
0.4	0.1448		

б)

x	$u(x)$	$u'(x)$	$u''(x)$
0.5	2.4834		
0.7	3.5216		
0.9	3.2212		
1.1	3.5682		

14. Показати, що формула

$$u_n^{(4)} \approx u_{\bar{x}x\bar{x}x,n} = \frac{1}{h^2}\left(u_{\bar{x}x,n+1} - 2u_{\bar{x}x,n} + u_{\bar{x}x,n-1}\right)$$

має четвертий порядок точності.
15. Посадка літака на палубу авіаносця характеризується такими даними:

t, сек	0	0.51	1.03	1.74	2.36	3.34	3.82
u, м	154	186	209	250	262	272	274

Тут u – відстань від кінця палуби, t – час посадки на палубу. Знайти швидкість і прискорення у початковий і кінцевий моменти часу, користуючись формулами другого порядку точності.

Рис. 10.2. Авіаносець "Джеральд Форд" (США, 2018 р., (фото з Вікіпедії))

16. Для аналізу точності обчислення різницевих похідних $u_{x,n}$, $u_{\bar{x},n}$, $u_{\bar{x}x,n}$ застосувати правило Рунге із співвідношенням кроків: $h_{n+1} q h_n$, $0 < h < 1$. Знайти наближені значення похибки для q = 1/2 і 1/4.

Розділ 11. Числове інтегрування

Інтерполяційні квадратурні формули (ІКФ), обчислення коефіцієнтів та оцінка похибки. Квадратурні формули (КФ) Ньютона – Котеса. Прості та складені КФ, їх похибки. Практичні способи оцінки похибки складених КФ. Поняття про квадратурні формули найвищого алгебраїчного степеня точності (КФНАСТ). Метод повторного застосування КФ при обчисленні подвійних інтегралів. Кубатурні формули, точні для алгебраїчних многочленів найвищого степеня. Наближене обчислення кратних інтегралів методом Монте-Карло.

Література [5, 13, 43, 44, 46, 59, 64, 73, 82]
Електронні джерела [103, 105–107]

11.1. Квадратурні формули

У прикладних задачах визначені інтеграли найчастіше обчислюються наближено. Це пов'язано з тим, що підінтегральна функція може бути задана таблицею значень або первісна не виражається через елементарні функції, наприклад для інтеграла

$$\Phi(x) = \int_0^x \frac{t^3 dt}{e^t - 1},$$

який у статистичній термодинаміці служить для визначення теплоємності твердого тіла. Ще одним прикладом є функція помилок

$$erf\,(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt.$$

Площа поверхні тіла обертання обчислюється за формулою

$$S = 2\pi \int_a^b f(x)\sqrt{1 + (f'(x))^2}\, dx$$

і, наприклад для функції $f(x) = \sin x^2$, обчислюється наближено.

Первісна для простих підінтегральних функцій може набувати складного аналітичного вигляду, наприклад

$$\int_0^b \frac{dt}{t^4 + 1} = \frac{\sqrt{2}}{4}\left(\frac{1}{2}\ln\left(\frac{b^2 + \sqrt{2}b + 1}{b^2 - \sqrt{2}b + 1}\right) + arctg\left(\frac{\sqrt{2}b}{1 - b^2}\right)\right),$$

тому доцільно застосувати числове інтегрування.

Розглянемо підхід до обчислення визначеного інтеграла

$$I(f) := \int_a^b \rho(x) f(x) dx, \qquad (11.1)$$

що ґрунтується на його заміні сумою

$$I_n(f) := \sum_{k=0}^{n} C_k f(x_k),$$

де x_k – координати різних точок (вузлів) з проміжку $[a,b]$, C_k – коефіцієнти КФ, які не залежать від f. Невід'ємна функція $\rho(x)$ у формулі (11.1) називається ваговою, наближена рівність

$$\int_a^b \rho(x) f(x) dx \approx \sum_{k=0}^{n} C_k f(x_k) \qquad (11.2)$$

– квадратурною формулою, сума в правій частині (11.2) – квадратурна сума, різниця

$$r_n(f) = \mathrm{I}(f) - \mathrm{I}_n(f)$$

називається похибкою КФ.

Простим прикладом КФ, які мають наглядну геометричну ілюстрацію (рис. 11.1), є формула лівих прямокутників

$$\int_a^b f(x) dx \approx f(a)(b-a),$$

правих прямокутників

$$\int_a^b f(x) dx \approx f(b)(b-a)$$

і центральних прямокутників

$$\int_a^b f(x) dx \approx (b-a) f\left(\frac{a+b}{2}\right), \qquad (11.3)$$

а також КФ трапецій

$$\int_a^b f(x) dx \approx \frac{b-a}{2}(f(a) + f(b)). \qquad (11.4)$$

КФ відрізняються за способами вибору вузлів x_k, їх числом і принципами, за якими будуються інтерполяційні многочлени. Наприклад, у складених КФ відрізок $[a,b]$ розбивається на частини й інтеграли обчислюються як сума КФ на кожній з частин.

Простіші КФ використовують рівновіддалені вузли, але ефективнішими є методи, що ґрунтуються на спеціальному їх виборі. Розробка числових методів інтегрування передбачає як побудову КФ для певних класів підінтегральних функцій і

областей інтегрування, так і дослідження їх похибки та стійкості відносно збурень значень $f(x_k)$, вибір вузлів тощо.

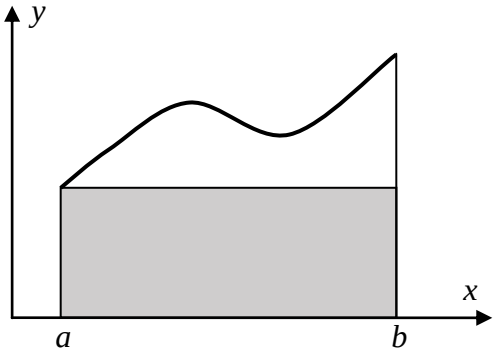


Рис. 11.1а. Метод лівих прямокутників

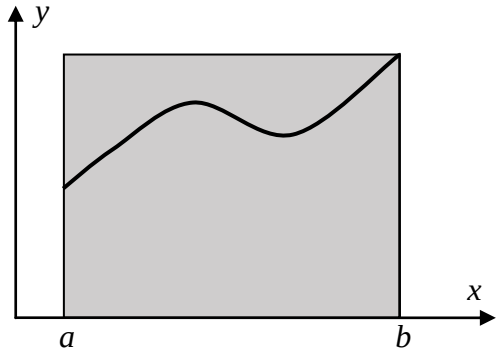


Рис. 11.1б. Метод правих прямокутників

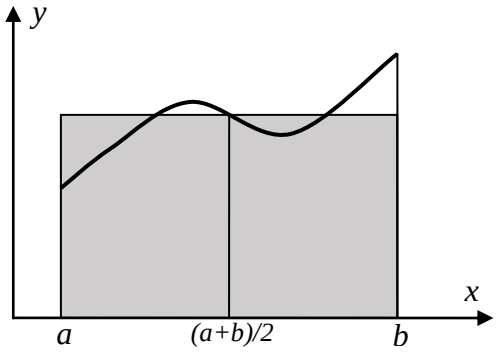


Рис. 11.1в. Метод центральних прямокутників

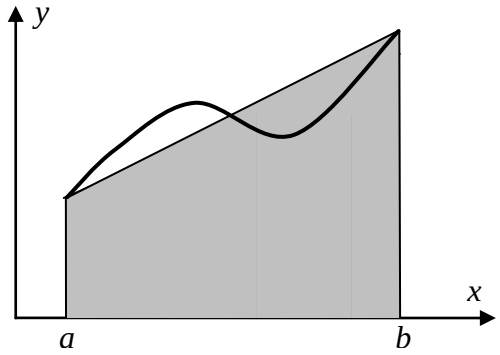


Рис. 11.1г. Метод трапецій

КФ для визначених інтегралів можна використати для обчислення подвійних інтегралів та інтегралів вищої кратності [5, 46, 59]. Але для таких інтегралів ефективніше використовувати методи типу Монте-Карло, які ґрунтуються на випадковому виборі вузлів інтегрування. Швидкість збіжності цих методів невисока, але складність їх значно менша, ніж при повторному застосуванні КФ, і похибка інтегрування тут не залежить від кратності інтеграла [5, 64].

11.2. Інтерполяційні квадратурні формули

За таблицею значень функції f у вузлах сітки $\Delta[a,b] :=$

$= \{x_k : a = x_0 < x_1 < ... < x_n = b\}$ побудуємо інтерполяційний многочлен Лагранжа

$$L_n(x) = \sum_{k=0}^{n} f(x_k)\Phi_k(x), \qquad (11.5)$$

де

$$\Phi_k(x)=\prod_{j\neq k}\frac{x-x_j}{x_k-x_j}=\frac{w_{n+1}(x)}{(x-x_k)w'_{n+1}(x_k)},\ w_{n+1}(x)=(x-x_0)(x-x_1)\cdot\ldots\cdot(x-x_n).$$

Тоді з формул (11.2) і (11.5) одержимо

$$\mathrm{I}_n=\sum_{k=0}^{n}f(x_k)\int_a^b\frac{\omega_{n+1}(x)}{(x-x_k)\omega'_{n+1}(x_k)}dx.$$

Отже, коефіцієнти КФ набувають вигляду

$$C_k=\int_a^b\frac{\omega_{n+1}(x)}{(x-x_k)\omega'_{n+1}(x_k)}dx,\ k=\overline{0,n}. \qquad (11.6)$$

Означення 11.1. *Формула (11.2), в якій коефіцієнти C_k обчислюється згідно з (11.6), називається ІКФ.*

Якщо $f\in C^{n+1}[a,b]$, то для похибки інтерполяційної КФ

$$r_n=\int_a^b(f(x)-L_n(x))dx,$$

враховуючи вигляд (9.7) похибки інтерполяційного многочлена Лагранжа, одержимо

$$r_n(f)=\frac{1}{(n+1)!}\int_a^b f^{(n+1)}(\theta(x))w_{n+1}(x)dx. \qquad (11.7)$$

Якщо $\max\limits_{x\in[a,b]}\left|f^{(n+1)}(x)\right|\le M_{n+1}$, то

$$\left|r_n\right|\le\frac{M_{n+1}}{(n+1)!}\int_a^b\left|\omega_{n+1}(x)\right|dx.$$

Одним із критеріїв точності КФ служить те, для яких класів функцій вони точні, наприклад для алгебраїчних многочленів.

Означення 11.2. КФ *має алгебраїчну степінь точності $m\ge 0$, якщо вона точна для всіх многочленів, степінь яких не перевищує m, і не виконується точно хоча б для одного многочлена, степінь якого більший за m.*

Зауважимо, що перевірити умову алгебраїчного степеня точності КФ рівносильно перевірці її виконання для одночленів $1, x,\ldots,x^m$.

Теорема 11.1. *Якщо $f\in C^{n+1}[a,b]$, то ІКФ (11.2) має алгебраїчний степінь точності n.*

Доведення. Нехай КФ (11.2) − інтерполяційна. З формули (11.7) випливає, що $r_n=0$, якщо $f(x)$ − многочлен степеня $p\le n$,

оскільки у цьому випадку $f^{(n+1)}(x) \equiv 0$. Для $f = x^{n+1}$ похибка $r_n = \int_a^b \omega_{n+1}(x)dx \neq 0$ для довільного вибору вузлів. Отже, $m = n$. ■

Прикладом ІКФ є формула трапецій (11.4), яка одержується при інтерполюванні у вузлах $x_0 = a$ і $x_0 = b$ многочленом

$$L_1(x) = f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a}.$$

КФ (11.4) має алгебраїчну степінь точності *m=1*. Така ж точність і КФ (11.3), а для КФ лівих і правих прямокутників алгебраїчну степінь точності $m = 0$.

11.3. Квадратурні формули Ньютона–Котеса

11.3.1. Побудова КФ. Нехай значення функції $y = f(x)$ задано у рівновіддалених вузлах $x_k = x_0 + kh, k = \overline{0,n};\ x_0 = a, x_n = b,$ $h = (b-a)/n$. ІКФ за такою системою вузлів називаються КФ Ньютона–Котеса. У цьому випадку можна уніфікувати обчислення коефіцієнтів C_k, які визначаються згідно з (11.6), та дослідити їх властивості. Зробимо заміну змінної $x = x_0 + th$. Тоді

$$t_k = (x_k - x_0)/h = k,\ x - x_k = th - kh = (t-k)h.$$

$$w_{n+1}(x) = w_{n+1}(x_0 + th) = th(t-1)h \cdot \ldots \cdot (t-n)h = h^{n+1}\prod_{k=0}^{n}(t-k).$$

Перетворимо вираз для похідної

$$w'_{n+1}(x_k) = \prod_{j\neq k}(x_k - x_j) = \prod_{j\neq k}(k-j)h =$$

$$= h^n k(k-1)\cdot\ldots\cdot 1(-1)(-2)\ldots(-1)^{n-k}(n-k) = (-1)^{n-k}h^n k!(n-k)!$$

КФ Ньютона–Котеса набувають вигляду

$$\int_a^b f(x)dx \approx (b-a)\sum_{k=0}^{n} A_k f(x_k). \qquad (11.8)$$

Множник $b-a$ потрібен для того, щоб коефіцієнти A_k не залежали від меж інтегрування a і b. На підставі (11.6) одержимо

$$C_k = (b-a)A_k = \int_{x_0}^{x_n} \frac{\omega_{n+1}(x)}{(x-x_k)\omega'_{n+1}(x_k)}dx = \int_0^n \frac{h^{n+1}t(t-1)\ldots(t-n)}{h(t-k)h^n k!(n-k)!(-1)^{n-k}}hdt =$$

$$= \frac{h(-1)^{n-k}}{k!(n-k)!}\int_0^n \frac{t(t-1)\ldots(t-n)}{t-k}dt.$$

Оскільки $b-a=nh$, то

$$A_k=\frac{(-1)^{n-k}}{nk!(n-k)!}\int_0^n\prod_{j\neq k}(t-j)dt,\ k=\overline{0,n}. \qquad (11.9)$$

11.3.2. Властивості коефіцієнтів Ньютона–Котеса

1) Сума коефіцієнтів КФ Ньютона–Котеса дорівнює одиниці. Справді, КФ (11.8) точна для многочлена $f(x)=1$. Тому

$$\int_a^b 1dx=(b-a)=(b-a)\sum_{k=0}^n A_k\text{, отже, }\ \sum_{k=0}^n A_k=1.$$

2) $A_{n-k}=A_k,\ k=\overline{0,n}$, що випливає з (11.9) при заміні k на $n-k$.

3) Для $n\geq 8$ існують коефіцієнти $A_k<0$. Зокрема, $A_2=A_6=$ $=-\frac{928}{28350}$, $A_4=-\frac{4540}{28350}$ для $n=8$.

11.3.3. Приклади. Використовуючи властивості 1), 2) і формулу (11.9) можна обчислити коефіцієнти формул Ньютона – Котеса для відповідного n. У таблиці 11.1 для $n=\overline{1,8}$ наведено ці коефіцієнти при наближеному обчислення інтеграла

$$I(h)=\int_{x_0}^{x_0+nh} f(x)dx$$

за допомогою КФ

$$I_n(h)=hq_n\sum_{k=0}^n B_k f(x_k),$$

а також похибку КФ $r_n(h)=I(h)-I_n(h)$.

Для $n=1$ одержимо $A_0=A_1$ і $A_0+A_1=1$, тому $A_0=A_1=1/2$ і маємо формулу трапецій (11.3). Якщо $n=2$, то

$$A_0=A_2=\frac{1}{2\cdot 2!}\int_0^2(t-1)(t-2)dt=\frac{1}{6},\ \ A_1=1-2A_0=\frac{4}{6}$$

і одержується КФ формула Сімпсона

$$\int_a^b f(x)dx\approx\frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(c)\right). \qquad (11.10)$$

Для $n=3$ маємо $A_0=A_3=1/8,\ A_1=A_2=3/8$. Відповідна КФ називається формулою „три восьмих”, або Ньютона

$$\int_a^b(x)dx\approx\frac{b-a}{8}\left(f_0+3f_1+3f_2+f_3\right).$$

Таблиця 11.1

Коефіцієнти і похибка КФ Ньютона-Котеса для $n=\overline{1,8}$

n	q_n	B_0	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	Похибка $r_n(h)$
1	$\frac{1}{2}$	1	1								$-\frac{h^3}{12}f^{(2)}(\xi)$
2	$\frac{1}{3}$	1	4	1							$-\frac{h^5}{90}f^{(4)}(\xi)$
3	$\frac{3}{8}$	1	3	3	8						$-\frac{3h^5}{80}f^{(4)}(\xi)$
4	$\frac{2}{45}$	1	32	7	32	1					$-\frac{8h^7}{945}f^{(6)}(\xi)$
5	$\frac{5}{288}$	19	75	50	50	75	19				$-\frac{275h^7}{12096}f^{(6)}(\xi)$
6	$\frac{1}{140}$	41	216	27	272	27	216	41			$-\frac{9h^9}{1400}f^{(8)}(\xi)$
7	$\frac{7}{17280}$	751	3577	1223	2989	2989	1223	3577	751		$-\frac{8183h^9}{518400}f^{(8)}(\xi)$
8	$\frac{8}{28350}$	989	5888 7	-928	10496	-4540	10496	-928	58887	989	$-\frac{2368h^{11}}{467775}f^{(10)}(\xi)$

Результати обчислення за допомогою деяких КФ центральних прямокутників та КФ Ньютона-Котеса з 2, 3. І 4-ма вузлами інтеграла $erf(1)$ наведені в табл. 11.2.

Таблиця 11.2

Наближене обчислення значення $erf(1)$

Квадратурна формула	Наближене значення	Похибка
Центральних прямокутників	0,8788	–0.0361
Трапецій	0,7717	0.0710
Сімпсона	0,8431	0.0004
„Три восьмих”	0,84289	0.0002

11.4. Стійкість КФ Ньютона–Котеса до похибок в обчисленні значень підінтегральної функції

Припустимо, що значення $f(x_k)$ обчислюється з деякими похибками Δ_i, $k=\overline{0,n}$, $|\Delta_i| \le \Delta$, тобто маємо значення $\overline{f}(x_k) = f(x_k) + \Delta_i$. Тоді замість точного I_n одержимо наближене значення квадратурної суми

$$\overline{I}_n = (b-a)\sum_{k=0}^{n} A_k \overline{f}(x_k). \qquad (11.11)$$

Із формул (11.11) і (11.8) випливає, що

$$|\overline{I}_n - I_n| \le (b-a)\sum_{k=0}^{n}|A_k| \cdot |\overline{f}(x_k) - f(x_k)| \le (b-a)\Delta\sum_{k=0}^{n}|A_k|.$$

Якщо $n<8$, то $A_k > 0$ і на підставі першої властивості маємо

$$|\overline{I}_n - I_n| \le (b-a)\Delta,$$

що характеризує стійкість квадратурних формул (11.8) до похибок в обчисленні f_k. Якщо ж $n \ge 8$, то серед коефіцієнтів A_k є від'ємні. Коефіцієнт $d_n = \sum_{k=0}^{n}|A_k|$ зростає зі збільшенням n, наприклад $d_{10} \approx 3.1$, $d_{20} \approx 500$, що може привести до зростання похибки, якщо збігатимуться знаки коефіцієнтів A_k і похибок обчислення значень f_k.

11.5. Складені КФ Ньютона–Котеса

Для досягнення заданої точності обчислення інтеграла проміжок $[a,b]$ поділимо на N частин довжиною $h=(b-a)/N$, на кожній з яких застосовуємо деяку КФ. Наприклад, якщо просумувати КФ трапецій на відрізках $[x_k, x_{k+1}]$

$$\int_{x_k}^{x_{k+1}} f(x)dx \approx \frac{h}{2}[f_k + f_{k+1}],\ k=\overline{0,N-1},$$

то одержимо складену КФ трапецій

$$\int_a^b f(x)dx \approx \frac{h}{2}[f_0 + 2(f_1 + ... + f_{N-1}) + f_N] \equiv$$
$$= \frac{b-a}{2N}[f_0 + 2(f_1 + ... + f_{N-1}) + f_N]. \qquad (11.12)$$

Найточніший результат для типу даних Float одержується, коли крок сітки дорівнює 2^{-13}, а для типу Double – 2^{-23}.

Таблиця 11.3.

Похибки обчислення значення $erf(1)$ за допомогою складеної КФ трапецій з розбиттям проміжку $[0,1]$ на $N = 2^k$ частин, реалізація мовою С, $erf(1) = 0.84270079295\ldots$

k	Похибка, $I - I_N$		k	Похибка, $I - I_N$	
	Float	Double		Float	Double
2	0.00433303	0.00433302	16	-0.00000022	0.1611 E-10
4	0.00027026	0.00027029	18	-0.00004147	0.1013 E-11
6	0.00001694	0.00001689	20	-0.00003527	0.9026 E-13
8	0.00000121	0.00000106	22	-0.00697325	0.1732 E-13
10	0.00000067	0.00000007	23	0.02172609	**0.7772 E-15**
12	0.00000013	0.4124E-8	24	-0.09673648	0.3708 E-13
13	**0.00000011**	0.1031E-8	26	0.56060599	0.1363 E-12
14	-0.00000017	0.2577 E-9	28	0.77217709	-0.1952E-12

Складена КФ Сімпсона одержується, якщо просумувати формули (11.10) на проміжках довжиною $2h$. У цьому випадку N – парне, Складена КФ Сімпсона набуває вигляду

$$\int_a^b f(x)dx \approx \frac{b-a}{3N}[f_0 + f_N + 4(f_1 + ... + f_{N-1}) + 2(f_2 + ... + f_{N-2})]. \qquad (11.13)$$

Складена КФ "три восьмих" $N = 3l$ має такий вигляд:

$$\int_a^b f(x)dx \approx \frac{3h}{8}\Big[f_0 + f_{3m} + 2\big(f_3 + f_6 + \ldots + f_{3l-3}\big) +$$
$$+3\big(f_1 + f_2 + f_4 + f_5 + \ldots + f_{3l-2} + f_{3l-1}\big)\Big].$$

Зауваження 11.1. *Складена КФ центральних прямокутників набуває вигляду*

$$\int_a^b f(x)dx \approx \frac{b-a}{N}[f(x_0 + \frac{h}{2}) + f(x_1 + \frac{h}{2})f_N + ... + f(x_{N-1} + \frac{h}{2})].$$

11.6. Похибка КФ Ньютона–Котеса

11.6.1. Формула трапецій. У КФ трапецій

$$f(x) = L_1(x) + R_1(x),$$

де $R_1(f) = \frac{1}{2} f''(\theta(x))(x-a)(x-b)$, $\theta(x) \in (a,b)$, тому похибка інтегрування

$$r_1(f) := \int_a^b f(x)dx - \frac{b-a}{2}(f(a)+f(b)) = \frac{1}{2}\int_a^b f''(\theta(x))(x-a)(x-b)dx.$$

Якщо функція φ – неперервна, а ψ – інтегровна на $[a,b]$ і не змінює знак, то згідно з теоремою про середнє

$$\int_a^b \varphi(x)\psi(x)dx = \varphi(\xi)\int_a^b \psi(x)dx.$$

Функція $\psi(x) = (x-a)(x-b) \geq 0$, коли $x \in [a,b]$, тому

$$r_1(f) = \frac{1}{2} f''(\xi)\int_a^b (x-a)(x-b)dx =, \ a < \xi < b.$$

Отже,

$$\frac{1}{2} f''(\xi)\left(\frac{1}{3}x^3 - \frac{1}{2}(a+b)x^2 + abx\right)\Big|_a^b = -\frac{(b-a)^3}{12} f''(\xi).$$

Відповідна оцінка похибки складає $(b-a)^3 M_2/12$, де сталою M_2 обмежений $|f''(x)|$ на $[a,b]$. На проміжку довжиною $b-a=h$, маємо $r_1(f) = -h^3 f''(\xi)/12$.

Для складеної формули трапецій (11.11)

$$\overline{r}_1(f) = -\frac{h^3}{12}[f''(\xi_1) + ... + f''(\xi_N)] =$$

$$-\frac{h^3 N}{12}[f''(\xi_1) + ... + f''(\xi_N)]/N = -\frac{h^3 N}{12} f''(\theta),$$

де $\theta \in (a,b)$. Оскільки $hN = b-a$, то

$$\overline{r}_T(f) = -\frac{h^2(b-a)}{12} f''(\theta).$$

Із цієї рівності випливає оцінка похибки складеної КФ трапецій

$$|\overline{r}_T(f)| \leq \frac{M_2 h^2}{12}. \qquad (11.14)$$

Застосуємо складену КФ трапецій для наближеного обчислення інтеграла в $erf(1)$. Результати наведені в табл. 11.4.

Таблиця 11.4

Кількість відрізків N	Формула трапецій	Похибка $I - I_N$
2	0,855263	0,012502
4	0,838368	0,004333
8	0,841619	0,001082
16	0,842431	0,000260
32	0,8423633	0,000068
64	0,842684	0,000017
128	0,842697	0,000004

11.6.2. Формула Сімпсона. КФ (11.10) точна для довільного многочлена степеня $m \le 2$, як інтерполяційна КФ (теорема 11.1). Але нескладно перевірити, що вона точна і для кубічного многочлена $\overline{P}_3(x) = (x-c)^3, \quad c = (a+b)/2.$ Оскільки кубічний многочлен можна записати у вигляді $P_3(x) = a_0 \overline{P}_3(x) + P_2(x)$, де a_0 – коефіцієнт при x^3, то КФ Сімпсона точна для всіх кубічних многочленів.

Побудуємо для функції $f(x)$ інтерполяційний многочлен Ерміта $H_3(x)$ за значеннями $f(a),\ f(c),\ f(b)$ і $f'(c)$, де $c = (a+b)/2$. Відомо, що похибка інтерполювання такого многочлена

$$R_2(x) = f^{(4)}(\xi(x))(x-a)(x-c)^2(x-b)/4,\ \xi \in (a,\ b).$$

Застосовувавши теорему про середнє і виконавши заміну $x = a + th,$ одержимо похибку КФ Сімпсона

$$r_2 = \frac{1}{24}\int_a^b f^{(4)}(\theta(x))(x-a)(x-c)^2(x-b)dx = -\frac{(b-a)^5}{2^5 \cdot 90} f^{(4)}(\xi)$$

Отже, для КФ Сімпсона похибка

$$r_2(f) = -\frac{(b-a)^5}{2^5 \cdot 90} f^{(4)}(\xi).$$

Для складеної КФ, урахувавши локальні похибки на кожному з проміжків $[x_0, x_0 + 2h]$, $[x_0 + 2h, x_0 + 4h]$,..., $[x_0 + (n-2)h, x_0 + Nh]$, $N = 2l$, одержимо

$$\overline{r}_2(f) = -\frac{h^5}{90}[f^{(4)}(\eta_1) + ... + f^{(4)}(\eta_N)] = -\frac{h^5 l}{90}(f^{(4)}(\eta_1) + ... + f^{(4)}(\eta_N))/l.$$

Оскільки $l = N/2$ і $Nh = b - a$, то

$$\overline{r}_2(f) = -\frac{h^4(b-a)}{180} f^{(4)}(\eta).$$

Відповідна оцінка набуває вигляду

$$|\overline{r}_c(f)| \le \frac{M_4 h^4}{180}. \qquad (11.15)$$

11.6.3. Формула „три восьмих”. Похибка КФ має вигляд [15, 16]

$$r_3(f) = -\frac{(b-a)^5}{6480} f^{(4)}(\xi).$$

Для складеної КФ «три восьмих», коли $N = 3l$ і $b-a = h$, маємо

$$\overline{r}_H(f) = -\frac{(b-a)h^4}{80} f^{(4)}(\eta).$$

Оцінка похибки складеної КФ

$$|\overline{r}_H(f)| \le \frac{M_4 h^4}{80}. \qquad (11.16)$$

Зауважимо, що множник $1/80$ в узагальненій формулі „три восьмих” більший, ніж $1/180$ у відповідній формулі Сімпсона.

11.7. Правило Рунге оцінки похибки складених КФ

Для використання оцінок похибки вигляду (11.15) або (11.16) у методі трапецій і Сімпсона або в інших складених КФ, потрібно знати оцінку модуля відповідної похідної, яка не завжди відома. Тому найчастіше використовуються оцінки, одержані шляхом порівняння наближених значень інтеграла на сітках із різним кроком (методи Рунге, Ромберга та ін.). Нехай

$$I(f) := \int_a^b f(x)dx = I_N(f) + r_N(f),$$

де I_N – деяка складена КФ на сітці з кроком $h = (b-a)/N$. Припустимо, що для похибки виконується умова

$$r_N(f) = Ah^p + o(h^p),$$

де A – деяка стала, що не залежить від h. За тією ж КФ обчислимо наближене значення I_{2N} з кроком $h/2$. Відкинувши величини $o(h^p)$, одержимо

$$I \approx I_N + Ah^p \text{ і } I \approx I_{2N} + 2^{-p} Ah^p.$$

Після вилучення сталої A з цих рівностей маємо

$$I - I_{2N} \approx \frac{I_{2N} - I_N}{2^p - 1}. \qquad (11.17)$$

Для складеної КФ трапецій (11.11), як випливає із (11.15), $p = 2$. Отже,

$$I_N - I_{2N} \approx (I_{2N} - I_N)/3.$$

Для складеної КФ Сімпсона $p = 4$, тому

$$I - I_{2N} \approx (I_{2N} - I_N)/15.$$

Такий спосіб побудови оцінки похибки КФ називають *правилом Рунге*. Уточнене значення інтеграла обчислюється за формулою

$$I \approx I_{2N} + \frac{I_{2N} - I_N}{2^p - 1}. \qquad (11.18)$$

У табл. 11.5 для деяких КФ, наведено значення порядку p, похибки та формули уточнення значення інтеграла.

Отже, обчислення наближеного значення інтеграла з точністю $\varepsilon > 0$ методом подвійного перерахунку полягаєсь у чому:

1) обчислити наближені значення інтегралів I_N та I_{2N} з кроками сітки $h = (b-a)/N$ і $h/2$ відповідно;

Таблиця 11.5

Складена квадратурна формула	p	Наближене значення похибки	Уточнене значення інтеграла
Лівих і правих прямокутників	1	$I_{2N} - I_N$	$(2I_{2N} - I_N)$
Центральних прямокутників і трапецій	2	$(I_{2N} - I_N)/3$	$(4I_{2N} - I_N)/3$
Сімпсона і Ньютона	4	$(I_{2N} - I_N)/15$	$(16I_{2N} - I_N)/15$

2) згідно з формулою (11.17) обчислити наближене значення похибки КФ числового інтегрування;

3) якщо $|I - I_{2N}| \le \varepsilon$, то за формулою (11.18) обчислити уточнене значення інтеграла й процес обчислень зупинити. Інакше, відрізок $[a,b]$ поділити на $4N$ рівних частин, обчислити $I_{4N}(f)$ і перевірити виконання нерівності $|I_{2N} - I_N|/(2^m - 1) \le \varepsilon$. Процес послідовного збільшення удвічі числа вузлів КФ (зменшення удвічі кроку інтегрування) продовжують до тих пір, поки

на певному кроці k не досягається задана точність або коли кількість поділів кроку сітки перевищує деяке задане значення.

Зауваження 11.1. *Якщо асимптотика ГСП невідома, то можна обчислити інтеграл (або іншу величину, залежну від кроку сітки) на сітках з кроками* h, qh *і* q^2h *відповідно. Нехай* $h_i = q^i h$, I_i *– відповідні значення інтеграла,* $i = \overline{0,2}$. *Тоді відносно порядку* p *і сталої* C *маємо систему рівнянь*

$$I = I_1 + Ch^p + o(h^p),$$
$$I = I_2 + Cq^p h^p + o(h^p),$$
$$I = I_3 + Cq^{2p} h^p + o(h^p).$$

Розв'язавши цю систему з точністю $o(h^p)$, *одержимо*

$$I \approx I_1 + \frac{(I_3 - I_1)^2}{2I_2 - I_1 - I_3}, \quad p \approx \frac{1}{\ln p} \ln \frac{I_3 - I_2}{I_2 - I_1}.$$

Наведений алгоритм називається процесом Ейткена.

Зауваження 11.2. *Розглянемо два приклади таких формул. На відрізку* $[a, b]$ *довжиною* $h = b - a$ *побудуємо КФ трапецій* $T(h)$ *і* $T(h/2)$ *з кроками* h *і* $h/2$ *відповідно і підставимо їх у формулу для уточнення інтеграла згідно з правилом Рунге*

$$I \approx T\left(\frac{h}{2}\right) + \frac{1}{3}\left(4T\left(\frac{h}{2}\right) - T(h)\right)$$

У підсумку одержимо $I \approx h\left(f(a) + 4f\left(a + \frac{h}{2}\right) + f(b)\right)$,

тобто КФ Сімпсона з кроком $h/2$. *Застосувавши таку процедуру для* $j = 2,3,...$, *починаючи з кроку* $h = b - a$, *одержимо рекурентну КФ Сімпсона*

$$S_j = \left(4T_j - T_{j-1}\right)/3, \quad j = 2,3,...,$$

де $T_j = T\left((b-a)/2^j\right)$, $S_j = S\left((b-a)/2^j\right)$, *яка має четвертий порядок точності.*

Аналогічно можна побудувати рекурентну формулу

$$B_j = \left(16S_j - S_{j-1}\right)/15, \ j = 2,3,...,$$

яка має шостий порядок точності, якщо $f \in C^6[a, b]$ [13, 46].

11.8. Квадратурні формули найвищого алгебраїчного степеня точності

11.8.1. Постановка задачі і приклади. Одним з критеріїв оцінки КФ є те, наскільки вони точні для певних класів функцій, зокрема для алгебраїчних многочленів. Розглянемо таку задачу: для наближеного обчислення інтеграла

$$\int_a^b \rho(x) f(x) dx,$$

де f – довільна функція з деякого класу, ρ (називатимемо її ваговою) – довільна фіксована, інтегровна і невід'ємна на (a,b) функція, побудувати КФ

$$\int_a^b \rho(x) f(x) dx \approx \frac{b-a}{2} \sum_{k=1}^{n} A_k f(x_k), \qquad (11.19)$$

точну для всіх алгебраїчних многочленів найвищого степеня m. Таку КФ називають *квадратурою Ґауса або КФ* найвищого алгебраїчного степеня точності (КФНАСТ). Якщо (1.19) – ІКФ із заданими вузлами $x_1, x_2, ..., x_n$, то $m = n-1$.

Для побудови КФ (11.19) потрібно вибрати $2n$ параметрів: n вузлів $x_1, x_2, ..., x_n$ і n коефіцієнтів $\mathrm{A}_1, \mathrm{A}_2, ..., \mathrm{A}_n$. Покажемо, що їх можна вказати так, щоб КФ (11.19) мала алгебраїчну точність $m = 2n-1$, тобто була точною для всіх многочленів, степінь яких не перевищує $2n-1$. Зауважимо, що многочлен степеня $2n-1$ визначається $2n$ коефіцієнтами. Точність $2n-1$ досягається, коли КФ (11.19) точна для $1,\ x,\ x^2, ..., x^{2n-1}$. Так одержимо систему $2n$ нелінійних рівнянь із $2n$ невідомими вигляду

$$\sum_{k=1}^{n} \mathrm{A}_k = \frac{2}{b-a} \int_a^b \rho(x) dx,$$

$$\sum_{k=1}^{n} x_k \mathrm{A}_k = \frac{2}{b-a} \int_a^b x \rho(x) dx,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\sum_{k=1}^{n} x_k^{2n-1} \mathrm{A}_k = \frac{2}{b-a} \int_a^b \rho(x) x^{2n-1} dx.$$

Розглянемо приклад, коли $b = -a = 1,\ \rho(x) = 1$. Тоді

$$\int_{-1}^{1} f(x)dx \approx \sum_{k=1}^{n} \mathrm{A}_k f(x_k).$$

Для $n=1$ маємо систему двох рівнянь

$$\mathrm{A}_1 = \int_{-1}^{1} dx = 2 \quad \mathrm{A}_1 x_1 = \int_{-1}^{1} x dx = 0$$

Отже, $A_1 = 2,\ x_1 = 0$ і одержимо КФ центральних прямокутників

$$\int_{-1}^{1} f(x)dx \approx 2f(0),$$

алгебраїчна точність якої $m=1.$

Нехай n = 2. Для визначення параметрів $x_1,\ x_2,\ \mathrm{A}_1,\ \mathrm{A}_2$ маємо систему з чотирьох рівнянь

$$\mathrm{A}_1 + \mathrm{A}_2 = 2, \qquad \mathrm{A}_1 x_1^2 + \mathrm{A}_2 x_2^2 = \frac{2}{3},$$

$$\mathrm{A}_1 x_1 + \mathrm{A}_2 x_2 = 0, \qquad \mathrm{A}_1 x_1^3 + \mathrm{A}_2 x_2^3 = 0.$$

Із симетричності системи випливає, що вона має розв'язок $\mathrm{A}_1 = \mathrm{A}_2 = 1,\ x_1 = -x_2 = \frac{1}{\sqrt{3}}$. Відповідна КФ Ґауса

$$\int_{-1}^{1} f(x)dx \approx f(-\frac{1}{\sqrt{3}}) + f(\frac{1}{\sqrt{3}}). \qquad (11.20)$$

Її степінь точності $m=3$ як і КФ Сімпсона, але з трьома вузлами. Згідно з КФ (11.21), $erf(1) = 0.842442$, що точніше, ніж 0.8431 0.8431 за триточковою КФ Сімпсона.

11.8.2. Існування та єдиність КФНАСТ

Теорема 11.2 [59]. *КФ* (11.19) *точна для довільного алгебраїчного многочлена степеня* m, $m \le 2n-1$, *тоді і тільки тоді, коли виконуються умови:*

1) многочлен $\omega_n(x)$ *ортогональний на* $[a,b]$ *з ваговою функцією* $\rho(x)$ *до будь-якого многочлена* $q(x)$ *степеня* $p \le n-1$, *тобто*

$$\int_{a}^{b} \rho(x) w_n(x) q(x) dx = 0; \qquad (11.21)$$

2) КФ (11.19) – інтерполяційна, тобто її коефіцієнти обчислюються згідно з формулою

$$C_k = A_k(b-a)/2 = \int_{a}^{b} \frac{\omega_n(x)}{(x-x_k)\omega_n'(x_k)} dx,\ k = \overline{0,n}. \qquad (11.22).$$

Доведення. Нехай КФ (11.19) точна для будь-якого многочлена, степінь якого не перевищує $2n-1$. Якщо $q(x)$ – многочлен степеня $p \le n-1$, то $\omega(x)q(x)$ – многочлен степеня $m \le 2n-1$ і

$$\int_a^b \rho(x)\omega(x)q(x)dx = \frac{b-a}{2}\sum_{k=1}^{n} A_k \omega(x_k)q(x_k) = 0.$$

Отже, рівність (11.22) виконується. Те, що (11.19) інтерполяційна КФ, випливає з теореми 11.1, оскільки вона точна для многочленів степеня $n-1$.

2. Нехай тепер виконуються умови 1), 2) і $f(x)$ – многочлен степеня $m \le 2n-1$. Тоді $f(x) = \omega_n(x)q(x) + r(x)$, де $q(x)$ – многочлен степеня не вищий від $n-1$, а $r(x)$ має степінь $p \le 2n-1$. Тоді

$$\int_a^b \rho(x) f(x)dx = \int_a^b \rho(x)\omega_n(x)q(x)dx + \int_a^b \rho(x)r(x)dx = \int_a^b \rho(x)r(x)dx.$$

Оскільки квадратурна формула інтерполяційна, то вона точна для многочлена $r(x)$. Отже,

$$\int_a^b \rho(x)r(x)dx = \frac{b-a}{2}\sum_{k=1}^{n} A_k r(x_k) =$$

$$= \frac{b-a}{2}\sum_{k=1}^{n} A_k (f(x_k) - \omega_n(x_k)q(x_k)) = \frac{b-a}{2}\sum_{k=1}^{n} A_k f(x_k). \quad \blacksquare$$

Так на підставі теореми 11.2 побудова КФНАСТ зводиться до знаходження вузлів $x_1, x_2, ..., x_n$. Коефіцієнти обчислюються згідно з (11.20). Рівняння для $x_1, x_2, ..., x_n$ одержуються з (11.22), якщо взяти замість $q(x)$ функції 1, x, $x^2 ..., x^{n-1}$. Одержимо систему n рівнянь

$$\int_a^b \rho(x)x^k\omega(x)dx = 0, \quad k = \overline{0, n-1}.$$

Якщо $b = -a = 1$, $\rho(x) = 1$, то маємо таку систему рівнянь:

$$\int_{-1}^{1} (x-x_1)...(x-x_n)x^k dx = 0\ , \qquad k = \overline{0, n-1}.$$

Для $n = 2$ маємо

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

Проінтегрувавши, одержимо: $x_1x_2=-1/3,\quad x_1+x_2=0$, звідки випливає, що $x_1=-x_2=1/\sqrt{3}$.

Існування і єдиність такого многочлена, а також те, що всі корені належать відрізку $[a,b]$ дається наступним твердженням.

Теорема 11.3 [59]. *Якщо $\omega_n(x)$ – многочлен степеня n, ортогональний на $[a,b]$ з ваговою функцією $\rho(x)>0$ до довільного многочлена, степінь якого не перевищує $n-1$, то всі його корені різні і належить відрізку $[a,b]$.* ■

11.8.3. Властивості КФ Ґауса

1. Число $m = 2n - 1$ – алгебраїчний степінь точності КФ (11.19). Справді, для многочлена $f(x)=\omega^2(x)$ степеня $2n$ маємо $\int_a^b \rho(x)w^2(x)dx>0$. Оскільки $\omega(x_k)=0$, то

$$\sum_{k=1}^{n}C_k f(x_k)=\sum_{k=1}^{n}C\,\omega^2(x_k)=0.$$

Тобто для інтеграла і відповідної квадратурної суми рівності не досягається.

2. Для довільного $n\ge 1$ коефіцієнти КФ Ґауса додатні. Ця властивість важлива для стійкості обчислень і дозволяє застосовувати такі КФ з довільним числом вузлів.

3. Можна довести, що похибка КФ Ґауса (11.19) має вигляд [31]

$$r_n(f)=\frac{f^{(2n)}(\xi)}{(2n)!}\int_a^b \rho(x)w^2(x)dx,\quad \xi\in(a,b).$$

Зауваження 11.3. *КФ Ґауса належать до формул відкритого типу, тобто коли жоден із вузлів не збігається з кінцем інтегрування. Такі КФ формули особливо корисні тоді, коли підінтегральні функції мають особливості на кінцях відрізка інтегрування, наприклад таких, як*

$$I_1=\int_0^2 \exp\left(-\frac{1}{x^2}\right)dx,\quad I_2=\int_0^1 \frac{\sin x}{\lambda}dx.$$

11.8.4. Частинні випадки

1. КФ Ґауса–Лежандра. Для таких формул $\rho(x)=1$, $-a=b=1$. Ортогональну систему функцій на $[-1,1]$ утворюють многочлени Лежандра (у вигляді формули Родріга)

$$P_n(x)=\frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n. \qquad (11.23)$$

Перші з них: $P_0(x)=1$, $P_1(x)=x$, $P_2(x)=(3x^2-1)/2$, $P_3(x)=(5x^2-3)x/2$. Корені многочленів Лежандра розміщені симетрично на $[-1,1]$. Коефіцієнти КФ Ґауса–Лежандра

$$C_k=\frac{2(1-x_k^2)}{n^2 \mathrm{P}_n(x_k)}, \quad k=\overline{1,n}.$$

Похибка КФ Ґауса–Лежандра [75, 97]

$$r_n(f)=\frac{2^{2n+1}(n!)^4}{\left[(2n)!\right]^3(2n+1)}f^{(2n)}(\xi).$$

Вузли і коефіцієнти КФ Ґауса–Лежандра наведені в табл. 11.6.

Зауваження 11.4. *Для обчислення визначеного інтеграла* $\int_a^b f(x)dx$ *потрібно зробити заміну змінної* $x=\frac{b+a}{2}+\frac{b-a}{2}t$ *. Тоді*

$$\int_a^b f(x)dx=\frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b+a}{2}+\frac{b-a}{2}t\right)dt\approx\frac{b-a}{2}\sum_{i=1}^{n}A_i f(x_i).$$

Таблиця 11.6

Вузли і коефіцієнти формул Ґауса–Лежандра

Вузли	Коефіцієнти
$n=2$	
±0.577350	1.000000
$n=3$	
0.000000 0.555556	0.888889 ±0.774597
$n=4$	
±0.861136 ±0.339981	0.347854 0.652145
$n=8$	
±0.960290 ±0.796666 ±0.525532 ±0.183435	0.101229 0.222381 0.313707 0.362684

Для $n=1$ одержимо КФ центральних прямокутників. Якщо $n=2$, то $A_1=A_2=1,\ x_1=-x_2=-\dfrac{1}{\sqrt{3}}$ і

$$\int_{-1}^{1} f(x)dx \approx f\left(-\frac{1}{\sqrt{3}}\right)+f\left(\frac{1}{\sqrt{3}}\right)$$

Для $n=3$ маємо таку КФ

$$\int_{-1}^{1} f(x)dx \approx \frac{5}{9}\left(f\left(-\sqrt{\frac{3}{5}}\right)+8f(x)+f\left(\sqrt{\frac{3}{5}}\right)\right).$$

2. КФ Ґауса–Чебишева. У цьому випадку вагова функція $\rho(x)=1/\sqrt{1-x^2}$, $-a=b=1$. Ортогональну систему функцій на інтервалі (-1, 1) з такою ваговою функцією утворюють многочлени Чебишева $T_n(x)=2^{1-n}\cos(n\cdot\arccos x)$, коренями яких є $x_k=\cos\dfrac{(2k-1)\pi}{2n},\ k=\overline{1,n}.$ Коефіцієнти КФ Ґауса–Чебишева

$$C_k=\frac{\pi}{n},\ n=\overline{1,n}.$$

КФ набуває вигляду

$$\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}dx \approx \frac{\pi}{n}\sum_{k=1}^{n} f(x_k),$$

а похибка КФ

$$r_n(f)=\frac{\pi}{2^{2n-1}}\frac{f^{(2n)}(\xi)}{(2n)!},\quad \xi\in(-1,\ 1).$$

Для $n=2$ одержимо $\int_{-1}^{1}\dfrac{f(x)}{\sqrt{1-x^2}}dx \approx \dfrac{\pi}{2}\left(f(-\dfrac{1}{\sqrt{2}})+f(\dfrac{1}{\sqrt{2}})\right).$

3. КФ Ґауса–Ерміта. Обчислюється невласний інтеграл

$$I(f)=\int_{-\infty}^{\infty} e^{-x^2} f(x)dx.$$

З ваговою функцією $\rho(x)=\exp(-x^2),\ -a=b=\infty$. Ортогональну систему функцій на $(-\infty,\ \infty)$ із такою ваговою функцією утворюють многочлени Ерміта [68, 100]:

$$H_n(x)=(-1)^n e^{x^2}\frac{d^n}{dx^n}e^{-x^2}.$$

Вузли і коефіцієнти КФ Ґауса–Ерміта наведені в табл. 11.7.

Таблиця 11.7

Вузли і коефіцієнти формул Ґауса–Ерміта

n	x_k	A_k
2	$\pm 0{,}707107$	0,886227
3	0,0000000 $\pm 2{,}224745$	1,181636 0,295409
4	$\pm 0{,}524648$ $\pm 1{,}650680$	0,804914 0,081313
5	0.000000 ± 0.958572 ± 2.020183	0.945309 0.393619 0.019953
6	± 0.436074 ± 1.335849 ± 2.350605	0.724630 0.157067 0.004530

Оцінка похибки КФ Ґауса–Ерміта

$$|r_n| \le \frac{n!\sqrt{\pi}}{2^n (2n)!} \sup_{x\in(-\infty,\infty)} \left| f^{(2n)}(\xi) \right|, \ \xi \in (-\infty, \ \infty)$$

11.9. Наближене обчислення кратних інтегралів

Розглянемо спочатку методи наближеного обчислення подвійних інтегралів вигляду

$$I(f) = \iint_D f(x,y)dxdy, \qquad (11.23)$$

де D − замкнена область в R^2, функція f визначена в D і володіє достатнім запасом гладкості. Існують різні підходи до побудови формул вигляду

$$I(f) \approx \sum_i \sum_j C_{ij} f(x_i, y_j), \qquad (11.24)$$

які називається *кубатурними*. Основні з цих способів такі: 1) повторне застосування КФ; 2) побудова КФНАСТ; 3) імовірносні методи типу Монте–Карло.

11.9.1. Повторне застосування КФ. Нехай D − прямокутник $\{(x,y): a \le x \le b, \ c \le y \le d\}$. Побудуємо в D прямокутну сітку з кроками $h_1 = (b-a)/N_1, \quad h_2 = (d-c)/N_2$. Записавши інтеграл (11.23) у вигляді повторного інтеграла

$$\iint_D f(x,y)dxdy = \int_a^b \left(\int_c^d f(x,y)dy \right) dx$$

і застосувавши для наближеного обчислення інтегралів

$$F(x)=\int_c^d f(x,y)dy,\quad I(f)\approx\int_a^b F(x)dx \tag{11.25}$$

складені КФ, одержимо кубатурну формулу (11.24). Якщо кроки $h_1=h_2$, то доцільно використати КФ одного порядку.

Розглянемо приклади. Застосуємо для наближеного обчислення інтегралів (11.25) складену КФ центральних прямокутників. Тоді одержимо

$$I(f)\approx h_1h_2\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} f(\overline{x}_i,\overline{y}_j),$$

де $(\overline{x}_i,\overline{y}_j)=(x_{i-1}+h_1/2, y_{j-1}+h_2/2)$ – координати центрів прямокутників, на які розбитий прямокутник D. При застосуванні складеної КФ трапецій одержимо таку кубатурну формулу

$$I(f)\approx\frac{h_1h_2}{4}\sum_{i=0}^{N_1}\sum_{j=0}^{N_2}C_{ij}f(x_i,y_j), \tag{11.26}$$

де $C_{00}=C_{0N_2}=C_{N_10}=C_{N_1N_2}=1$, в інших вузлах на сторонах прямокутника $C_{ij}=2$, а у внутрішніх вузлах $C_{ij}=4$. Повторне застосування складеної КФ Сімпсона з кроком сітки $h_1=(b-a)/4$ і $h_2=(d-c)/4$ приводить до кубатурної формули

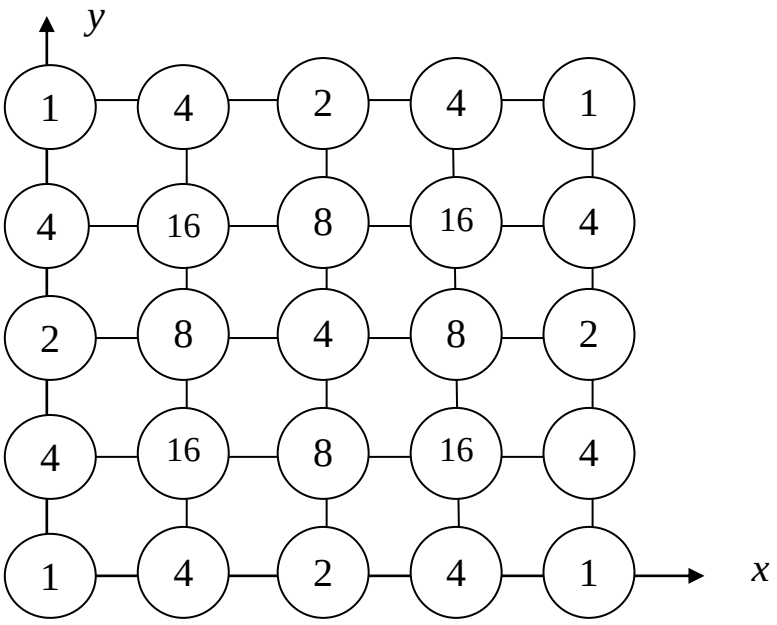


Рис. 11.1. Коефіцієнти кубатурної формули (11.27)

$$I(f) \approx \frac{h_1 h_2}{9} \sum_{i=0}^{4} \sum_{j=0}^{4} D_{ij} f(x_i, y_j), \qquad (11.27)$$

коефіцієнти якої D_{ij} показані на рис. 11.1.

Недоліком повторного застосуванням КФ є те, що зі зростанням кратності інтеграла швидко зростає кількість вузлів, отже, і кількість обчислень значень підінтегральної функції.

11.9.2. Оцінка похибки кубатурних формул. Нехай у напрямі кожної змінної інтегрування застосовується складена КФ трапецій. Наведемо оцінку похибки кубатурної формули (11.26), коли $D = [0,1] \times [0,1]$ і $h_i = 1/N_i$. Розглянемо інтеграл (11.23), де похідні $f_{x_1 x_1}$ і $f_{x_2 x_2}$ підінтегральної функції – неперервні в квадраті D, $M_k = \max_D \left| f_{x_k x_k}(x_1, x_2) \right|$, $k = 1,2$. Для наближеного обчислення інтеграла застосуємо кубатурну формулу (11.26). Для оцінки похибки r маємо [1]:

$$|r| \le \frac{1}{12} \left(\frac{M_1}{N_1^2} + \frac{M_2}{N_2^2} \right).$$

Якщо $N_1 = N_2 = N$, то $|r| \le \frac{M_1 + M_2}{12N^2}$. Нехай $M_1 = M_2 = 60$. Для досягнення точності $\varepsilon = 0.00001$ число N знаходиться з нерівності $10N^{-2} \le \varepsilon$, звідки $N \ge 1000$. Отже, кількість вузлів сітки на D складе $(N+1)^2 \approx 10^6$. Для шестикратного інтеграла, коли $M_1 = M_2 = 100$ і $\varepsilon = 0.001$, число вузлів уже досягає 10^{60} і практично реалізувати обчислення за такою кубатурою формулою не вдасться.

11.9.3. Кубатурні формули найвищого алгебраїчного степеня точності. Як і у випадку визначеного інтеграла для наближеного обчислення подвійного інтеграла (11.23), можна побудувати кубатурну формулу вигляду

$$\iint_D f(x,y)dxdy \approx \sum_{i=1}^{n} C_i f(x_i, y_i), \qquad (11.28)$$

де коефіцієнти C_i і точки $M_i(x_i, y_i)$ вибираються так, щоб кубатурна формула (11.28) мала алгебраїчну точність m. Такі формули розраховані на підінтегральні функції високої гладкості і вимагають значно менше обчислень, ніж при повторному застосуванні КФ. Побудова кубатурних формул такого типу

залежить від області інтегрування. Якщо D - одиничний круг із центром у початку координат, то можна застосувати кубатурну формулу Люстерника–Діткіна

$$\iint_D f(x,y)dxdy \approx \frac{\pi}{8}(2f(0)+\sum_{k=0}^{5} f(M_k)),$$

де M_i – точки з полярними координатами $\rho_i = \sqrt{\frac{2}{3}}, \quad \varphi_k = \frac{\pi}{3}k,$ $k = \overline{0,5},$ тобто точки M_k є вершинами правильного шестикутника, вписаного в коло радіуса $\sqrt{2/3}$.

Коефіцієнти та вузли кубатурних формул для квадрата з вершинами в точках $(1,1),(-1,1),(-1,-1),(1,-1)$ й одиничного круга з центром у початку координат, показані на рис. 11.2-11.5. Тут символами □, Δ і О позначено вузли з коефіцієнтами, які вказані стрілками, m – алгебраїчний степінь точності.

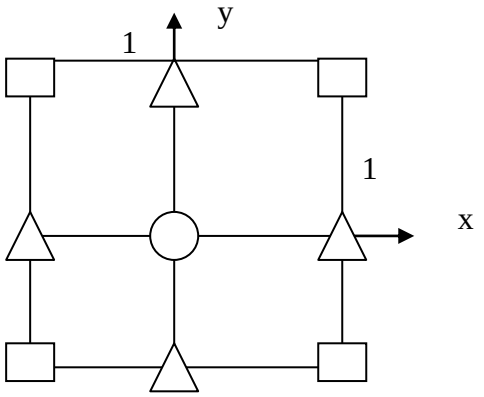


Рис.11.2. $m=3, \square \to \frac{4}{9}, \Delta \to \frac{1}{9}, \text{O} \to \frac{16}{9}$

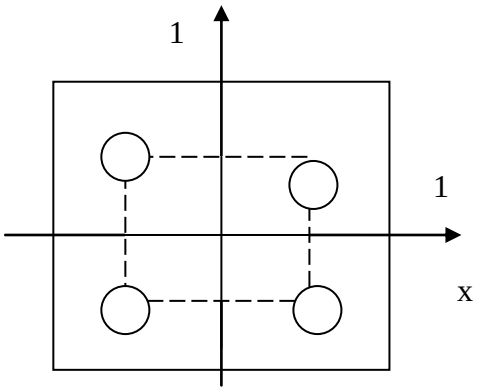


Рис.11.3. $\text{O} \to 1, \ \xi = 1/\sqrt{3}$

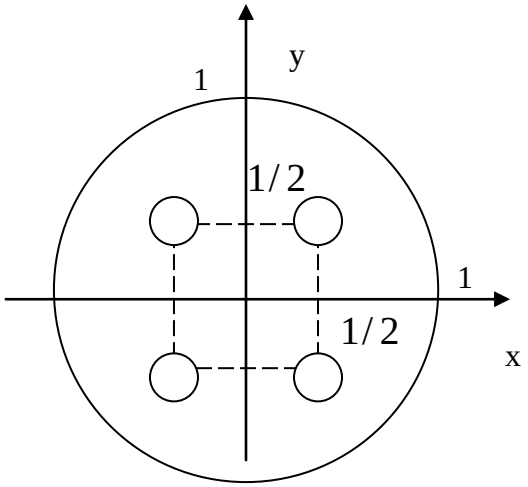


Рис.11.4. $m=3, \text{O} \to \pi/4$

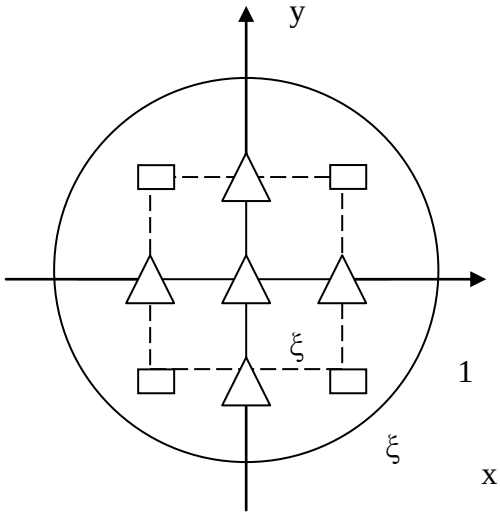


Рис. 11.5. $m=5, \ \xi = 0{,}707107,$ $\square \to 0{,}1309, \Delta \to 0{,}523599$

Зауважимо, що розглянуті кубатурні формули можна застосувати для довільного круга або еліпса, квадрата або прямокутника після відповідного перетворення системи координат або розбиття області D на області такого вигляду [35, 44, 63].

11.10. Метод Монте–Карло

Повторне застосування КФ для обчислення кратних інтегралів веде до різкого зростання вузлів сітки. Кубатурні формули найвищого алгебраїчного степеня точності орієнтовані на спеціальні області інтегрування і для функцій високої гладкості. Ефективним методом наближеного обчислення кратних інтегралів є метод Монте–Карло, який має ймовірнісну природу. У цьому методі кількість вузлів не пов'язане з кратністю інтеграла, але оцінка похибки одержується тільки з деякою ймовірністю. Застосуванням такого методу можна наближено обчислити потенціал одного тіла $V_1 \subset R^3$ на інше $V_1 \subset R^3$, який визначається шестикратним інтегралом

$$W = \iint\limits_{V_1 \times V_2} \iint\limits \iint \frac{\rho_1 \rho_2}{r_{1,2}} dx_1 dy_1 dz_1 dx_2 dy_2 dz_2 ,$$

де $r_{1,2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$, функції $\rho_i (x_i, y_i, z_i)$, $i = 1, 2$, задають густину розподілу мас в кожному із тіл.

Розглянемо кратний інтеграл вигляду

$$I(f) = \int_0^1 ... \int_0^1 f(x_1, ..., x_n) dx_1 ... dx_n .$$

Нехай $x = (x_1, ..., x_n)$, $G = [0, 1]^n$ – n-вимірний куб. Тоді інтеграл матиме вигляді

$$I(f) = \int\limits_G f(x) dx.$$

Задамо N випадкових попарно незалежних точок $\bar{x}_1, ..., \bar{x}_N$, які рівномірно розподілені в G. Нехай M і D – символи математичного сподівання і дисперсії. На підставі рівномірного розподілу точок $\bar{x}_i$ і того, що міра G дорівнює 1, маємо

$$M(y_i) = \int\limits_G f(x) dx = I(f), \quad D(y_i) = M(y_i^2) - M^2(y_i) = D(f).$$

Обчислимо наближене значення інтеграла $I(f)$ згідно з формулою

$$S_N(f)=\frac{1}{N}\sum_{i=i}^{N} f(\overline{x}_i). \qquad (11.29)$$

Тоді $M(S_N)=\frac{1}{N}\sum_{i=i}^{N}M(y_i)=I(f), \quad D(S_N)=\frac{1}{N^2}\sum_{i=i}^{N}D(y_i)=\frac{1}{N}D(f).$

Згідно з нерівністю Чебишева з імовірністю $1-q$, де $q\in(0,1)$, маємо

$$|S_N(f)-I(f)|\le\sqrt{\frac{D(f)}{qN}}.$$

Наприклад, для $q=0.01$ з імовірністю 0.99 правильна оцінка

$$|S_N(f)-I(f)|\le 10\sqrt{D(f)}/\sqrt{N}.$$

Дисперсію $D(f)$ можна наближено обчислити за формулою

$$D(f)\approx\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i-S_N(f)\right)^2.$$

Тобто порядок оцінки похибки наближеного обчислення інтеграла має порядок $1/\sqrt{N}$, на відміну від детермінованих кубатурних формул, у яких порядок оцінок швидкості збіжності погіршується зі зростанням кратності інтеграла. Важливо, що оцінка не залежить від кратності інтеграла. З іншого боку, ця оцінка ймовірнісна. Крім того, випадкові величини $\overline{x}_1,...,\overline{x}_N$ одержуються за допомогою датчика випадкових, або, як їх називають, псевдовипадкових чисел, які генерують послідовність випадкових чисел, рівномірно розподілених на проміжку $[0,1]$. Але не завжди вдається досягнути бажаних статистичних властивостей таких чисел. Наприклад, деякі датчики псевдовипадкових чисел генерують послідовність чисел, яку можна розглядати тільки як попарно незалежні, а не як незалежні за сукупністю.

Приклади розв'язування типових задач

Задача 1. Знайти оцінку похибки простої та складеної КФ центральних прямокутників.

Розв'зування. Нехай $f \in C^2[a,b]$, $c=(a+b)/2$ і $\max\limits_{x\in[a,b]}\left|f''(x)\right| \le M_2$. Для простої КФ центральних прямокутників із розкладу функції f у точці $x=c$ та теореми про середнє випливає

$$\int_a^b f(x)dx-(b-a)f(c)=\int_a^b\left[f(c)+f'(c)(x-c)+\frac{1}{2}f''\left(\bar{\xi}(x)\right)(x-c)^2\right]dx-$$

$$-(b-a)f(c)=\frac{1}{2}f''(\xi)\int_a^b(x-c)^2dx=\frac{(b-a)^3}{24}f''(\xi),\quad \xi\in(a,b)$$

.

Для складеної КФ (рис. 11.6)

$$\int_a^b f(x)dx \approx \frac{b-a}{N}[f(x_0+\frac{h}{2})+f(x_1+\frac{h}{2})f_N+...+f(x_{N-1}+\frac{h}{2})]$$

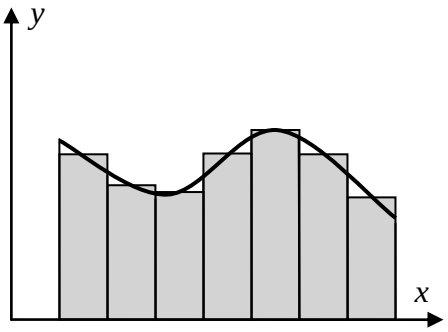


Рис. 11.6. Метод центральних прямокутників

Маємо

$$r_N := \int_a^b f(x)dx-h\sum_{k=0}^{N-1}f\left(x_k+\frac{h}{2}\right)=\frac{h^3}{24}\left[f''(\eta_0)+...+f''(\eta_{N-1})\right]=$$

$$=\frac{(b-a)h^2}{24}f''(\eta),\quad \eta\in(a,b).$$

Оцінка похибки складеної КФ центральних прямокутників

$$\left|r_N\right| \le (b-a)h^2M_2/24.$$

Задача 2. Оцінити похибку КФ трапецій, Сімпсона і «три восьмих» при обчисленні інтеграла

$$erf\,(1)=\frac{2}{\sqrt{\pi}}\int_0^1 e^{-x^2}dx,\ \ \pi\approx 3.1415927.$$

Розв'язування. На проміжку $[0,1]$ знайдемо оцінки другої і четвертої похідної функції $f(x)=e^{-x^2}$. Маємо: $f''(x)=2\left(2x^2-1\right)e^{-x^2}$, $f'''(x)=4x\left(3-2x^2\right)e^{-x^2}$, $f^{(4)}(x)=4\left(3-12x^2+4x^4\right)e^{-x^2}$. Оскільки похідна $f'''(x)\neq 0$, при $x\in[0,1]$, то найбільше значення досягається на кінцях відрізка, при цьому $\max\limits_{x\in[0,1]}\left|f''(x)\right|=$ $\max\left(\left|f''(0)\right|,\left|f''(1)\right|\right)=2=:M_2$. Далі,

$$\left|f^{(4)}(x)\right|\le 4\max_{x\in[0,1]}\left|3-12x^2+4x^4\right|=$$

$$4\max\left(\left|f^{(4)}(0)\right|,\left|f^{(4)}(1)\right|\right)=20=:M_4.$$

На підставі оцінки (11.14) для КФ трапецій

$$\left|r_T\right|\le\frac{2}{\sqrt{\pi}}\frac{(1-0)^3\cdot 2}{12}=\frac{1}{3\sqrt{\pi}}<0.18806.$$

Із формули для похибки для КФ Сімпсона випливає оцінка $\left|r_C\right|\le\frac{2}{2^5\cdot 90\sqrt{\pi}}20<0.007846$. Для КФ «три восьмих» маємо $\left|r_H\right|\le\frac{2}{\sqrt{\pi}}\cdot\frac{27}{720}\cdot\left(\frac{1}{3}\right)^5\cdot 20<0.00348$. Одержані оцінки, як видно з табл. 11.2, перевищують точні оцінки.

Задача 3. Для наближеного обчислення інтеграла $erf\,(1)$ записати СКФ трапецій, центральних прямокутників і Сімпсона з кроком сітки h. Для кожної із СКФ вибрати крок сітки так, щоб похибка не перевищувала 10^{-6}.

Розв'язування. Розіб'ємо проміжок $[0,1]$ на $N=2l$ відрізків довжиною $h=1/N$. Вузли одержаної сітки $x_k=kh,\ k=\overline{0,N}$. Складена КФ трапецій, центральних прямокутників і Сімпсона набувають відповідно вигляду:

$$I_T:=\frac{h}{\sqrt{\pi}}\left(1+e^{-1}+2\sum_{k=1}^{N-1}e^{-(kh)^2}\right);\ I_{ЦП}:=\frac{2h}{\sqrt{\pi}}\sum_{k=0}^{N-1}e^{-\left(k+\frac{1}{2}\right)^2h^2};$$

$$I_C := \frac{2h}{3\sqrt{\pi}}\left[1+e^{-1}+4\sum_{k=1}^{N/2}e^{-h^2(2k-1)^2}+\sum_{k=2}^{N/2-1}e^{-4h^2k^2}\right],\ N=2l.$$

Для СКФ трапецій з похибки (11.14) випливає, що $\frac{2}{\sqrt{\pi}}\cdot\frac{h^2}{12}\cdot 2\le 10^{-6}$. Тоді $h^2\le 3\cdot 10^{-6}\sqrt{\pi}$ і $h\le\sqrt{3}\cdot 10^{-3}\sqrt[4]{\pi}=$ $=0,0023059$. Тому похибки СКФ трапецій не перевищує 10^{-6}, якщо взяти $h_0=0,002$.

Для СКФ центральних прямокутників $h\le\sqrt{6}\cdot 10^{-3}\sqrt[4]{\pi}=$ $=0,00326$ і точність забезпечується, якщо $h_0=0,0025$.

На підставі оцінки (11.16) для похибки СКФ Сімпсона маємо $\frac{2}{\sqrt{\pi}}\cdot\frac{h^4}{180}\cdot 20\le 10^{-6}$, тому $h\le\sqrt[4]{\frac{9}{2}}\cdot\sqrt[8]{\pi}\cdot\sqrt[4]{10^{-6}}=0,05314...$ Отже, для кроку сітки $h_0=0,05$ досягається задана точність і при цьому $N=h_0^{-1}=20.$

Задача 4. Обчислити наближене значення інтеграла

$$I=\int_{-2}^{2}\frac{\sin x+x^2}{\sqrt{4-x^2}}dx,$$

застосувавши КФ Гауса–Чебишева із чотирма вузлами й оцінку похибки.

Розв'язування. Заміною $x=2t$ перетворимо інтеграл до вигляду

$$I=\int_{-1}^{1}\frac{\sin 2t+4t^2}{\sqrt{1-t^2}}dt.$$

Тут $f(t)=\sin 2t+4t^2$, $f^{(8)}(t)=-2^8\sin 2t$, $\max\limits_{[-1,1]}\left|f^{(8)}(t)\right|<256.$

Коефіцієнти КФ $A_k=\pi/4,$ вузли $t_1=-t_4=\cos\frac{\pi}{8}\approx 0.9239,$

$t_2=-t_3=\cos\frac{3\pi}{8}\approx 0.3827.$ Отже, $I_4=\frac{\pi}{4}\cdot 8\left[\cos^2\frac{\pi}{8}+\cos^2\frac{3\pi}{8}\right]=2\pi.$

Оцінка похибки

$$\left|I_4-I\right|\le\frac{\pi}{2^7}\frac{\max\left|f^{(8)}(x)\right|}{8!}<\frac{\pi}{2^7}\cdot\frac{2^8}{8!}=1.56\cdot 10^{-4}.$$

Задача 5. За допомогою методу Монте–Карло з кількістю вузлів $N=2^k$, де $k=3,6,9$ і 12, обчислити інтеграл

$$J := \int_0^1\int_0^1\int_0^1\int_0^1 e^{xy}\cos\left(\frac{\pi}{2}uv\right)dxdydudv,$$

у значенні якого 1.150073 всі цифри правильні.

Розв'язування. Координати вузлів вибираються як псевдо-випадкові числа [82]. Для вказаних N отримуються результати: $J_8 = 1.027190$, $J_{64} = 1.149216$, $J_{512} = 1.120108$, $J_{4096} = 1.149970$. Похибка наближеного значення J_{4096} складає ≈ 0.0001.

Завдання та запитання для самостійної роботи

1. Що таке КФ і з яких міркувань вибираються її коефіцієнти та вузли?
2. Що таке алгебраїчна степінь точності КФ? Навести приклади.
3. Яка КФ називається інтерполяційною? Навести приклади. Чому дорівнює її алгебраїчний степінь точності?
4. Який вигляд має КФ Ньютона–Котеса? Обчислення та властивості коефіцієнтів.
5. Дати геометричну ілюстрацію КФ прямокутників, трапецій і Сімпсона та порівняти їх за алгебраїчним степенем точності.
6. Як вибираються вузли та коефіцієнти у КФНАСТ?
7. Одержати вирази для похибки простої і СКФ Сімпсона.
8. Записати складену КФ центральних прямокутників, трапецій і Сімпсона для обчислення з точністю $\varepsilon = 0.0001$ таких інтегралів:

$$1)\ \int_0^1 \sin^2 x dx;\ 2)\ \int_0^2 x^2 e^{-x} dx;\ 3)\ \int_0^1 x\cos x dx.$$

9. Побудувати прості і складені КФ Ньютона-Котеса, кількість вузлів у яких 5 і 6.
10. Побудувати складені КФ лівих і правих прямокутників Сімпсона і Ньютона ("три восьмих ").
11. Побудувати СКФ, наблизивши функцію f інтерполяційним кубічним сплайном $S_3(x)$. Відомо, що $\pi = \int_0^1 \frac{4}{1+x^2}dx$. Знайти наближене значення числа π, застосувавши: а) складену КФ Сімпсона з $n = 2^k$ відрізками, $k = \overline{1,7}$; б) КФ Ґауса з 2 і 4 вузлами.
12. Показати, що між складеними КФ центральних прямокутників $I_П(h)$, трапецій $I_Т(h)$ і Сімпсона $I_С(h)$ виконуються співвідношення:

1) $I_C(h) = \frac{2}{3} I_П(h) + \frac{1}{3} I_T(h),\ h = (b-a)/N; \qquad I_C(h) = I_T(h) + R_T(h),$

де $R_T(h) = (I_T(h) - I_T(2h))/3$ – поправка Річардсона.

13. Нехай Т – трикутник на площині, де A, B, C – середини його сторін. Показати, що кубатурна формула

$$\iint_T f(x)dx \approx \tfrac{1}{3} S_T (f(A) + f(B) + f(C))$$

точна для всіх поліномів змінних x_1 і x_2 степеня 2. Тут S_T – площа трикутника.

14. Нехай П – прямокутник на площині, A,B,C і D – середини його сторін, E – центр. Показати, що кубатурна формула

$$\iint_\Delta f(x)dx \approx \tfrac{1}{6} S_П (f(A) + f(B) + f(C) + f(D) + 2f(E))$$

точна для всіх алгебраїчних многочленів двох змінних третього степеня.

15. Знайти оцінку похибки обчислення інтеграла $\int_0^1 f(x)dx, f(x) = (1+x^2)^{-1}$

за складеною КФ

$S(f) = [f(0) + 2f(0{,}1) + 2f(0{,}2) + ... + 2f(0{,}9) + f(1{,}0)]/20.$

16. Оцінити мінімальне число вузлів складеної КФ Сімпсона для обчислення інтеграла $\int_0^2 f(x)dx$, що забезпечує точність $\varepsilon \le 10^{-4}$ на класі функцій, які задовольняють умову: $\max_{x \in [0,2]} |f^{(4)}(x)| \le 1.$

17. Знайти коефіцієнти КФ

$$\int_0^2 f(x)dx \approx C_1 f(0) + C_2 f(1) + C_3 f(2),$$

точної для алгебраїчних многочленів найвищого степеня.

18. Для обчислення інтеграла $\int_0^2 f(x)dx$ застосувати складені КФ трапецій і центральних прямокутників. Оцінити мінімальне число N розбиття відрізка інтегрування, яке забезпечує точність 10^{-3} на двох класах функцій: $\|f''(x)\| \le 1$ і $\int_0^2 |f''(x)|dx \le 1.$

19. За допомогою складених КФ центральних прямокутників, трапецій і Сімпсона обчислити з точністю 10^{-8} інтеграл

$$\int_0^{0.5} \frac{dx}{\sqrt{1-x^2}}.$$

точне значення якого $\pi/6$, $\pi = 3.141592654\ldots$

20. Довести, що в трикутнику з вершинами $(0,0)$, $(0,1)$ і $(1,0)$ кубатурні формули (рис.).

$$J_3 = \frac{1}{6}\left[f\left(\frac{1}{2},0\right) + f\left(0,\frac{1}{2}\right) + f\left(\frac{1}{2},\frac{1}{2}\right)\right]; \quad J_3 = \frac{1}{6}\left[f\left(\frac{1}{6},\frac{1}{6}\right) + f\left(\frac{2}{3},\frac{1}{6}\right) + f\left(\frac{1}{6},\frac{2}{3}\right)\right]$$

мають алгебраїчну степінь точності $m = 2$.

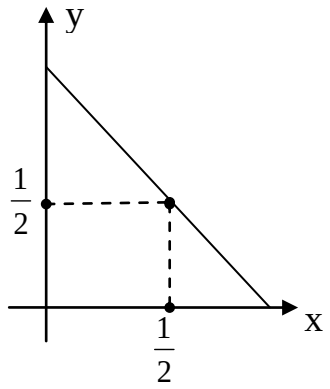


Рис. 11.6

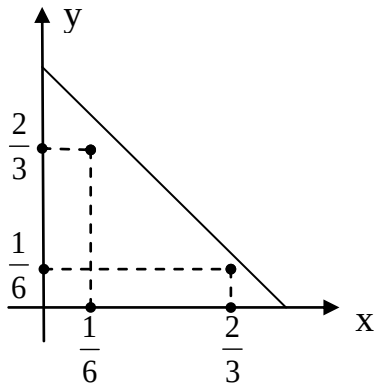


Рис. 11.7

21. Побудувати КФ Ґауса з двома і трьома вузлами для обчислення інтегралів вигляду:

$$1)\ \int_{-1}^{5} \sin x^2 dx; \quad 2)\ \int_{-\infty}^{\infty} e^{-x^2} f(x)dx; \quad 3)\ \int_{-1}^{1} \frac{\sin 2x}{\sqrt{1-x^2}} dx.$$

22. Застосувати КФ Ґауса з одним, двома і чотирма вузлами для обчислення інтегралів I_1 і I_2 із зауваження 11.1.

23. Із точністю 10^{-6} обчислити за допомогою складеної квадратурної формули трапецій площу фігури, обмеженої сигмоїдою[1] $\sigma(x) = (1+e^{-x})^{-1}$ і віссю Q_x на відрізку [-1,1].

24. Обчислити за допомогою будь- якої складеної КФ інтеграл [35, с. 190]

$$I = \frac{1}{\sqrt{2\pi}} \int_{-200\,000}^{200\,000} t^2 \exp\left(-t^2/2\right) dt \approx 1.0.$$

Якщо це не вдасться зробити, то можна змінити підінтегральну функцію, оскільки вона мало відрізняється від нуля на відрізку , що складає 0,01% від заданого.

[1] Простий вираз для похідної $\sigma'(x) = \sigma(x)(1-\sigma(x))$ через саму функцію спричинило її використання в нейронних мережах через зменшення обчислювальної складності методу зворотного поширення помилки.

25. Вивести формули для наближеного значення інтеграла та порядку ГСП, за його значеннями на сітках з кроками h, qh та q^2h, $q-$ ціле додатне число.
26. Довести, що всі коефіцієнти КФ Гауса з ваговою функцією $\rho(x) > 0$ додатні.
27. На відрізку $[x_{i-1}, x_i]$ підінтегральна функція апроксимується кубічним сплайном

$$S_{3,i}(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3.$$

Побудувати формулу сплайн-квадратури та дослідити її точність.
28. Методом Монте–Карло обчислити інтеграл

$$\iint_D \exp\left(x^2 y^2\right) dxdy,$$

де D – одиничний круг з центром у точці $(0,0)$, кількість вузлів $N = 10^2, 10^3$ і 10^4. Порівняти одержані значення з результатами при застосуванні кубатурних формул найвищого алгебраїчного степеня точності з чотирма вузлами.
28. Знайти параметри B_1, B_2 і $x_2 \in (a,b]$ так, щоб КФ

$$\int_a^b f(x)dx \approx B_1 f(a) + B_2 f(x_2)$$

була точною для для алгебраїчних многочленів найвищого степеня.

Розділ 12. Однокрокові числові методи розв'язування задачі Коші для ЗДР

Числовий розв'язок задачі Коші. Явний та неявний методи Ейлера. Похибка апроксимації і збіжність явного методу Ейлера. Загальна схема явних методів Рунге–Кутти. Явні методи Рунге–Кутти другого, третього і четвертого порядку точності. Огляд явних методів Рунге–Кутти вищих порядків. Апостеріорна оцінка похибки числового розв'язку. Стійкість явних методів Рунге – Кутти. Неявні методи Рунге–Кутти.

Література [4, 5, 13, 16, 21, 48, 59, 65, 73, 75, 77, 84, 98]
Електронні джерела [103−108]

12.1. Числовий розв'язок диференціальної задачі

Математичні моделі фізичних, біологічних, екологічних та інших процесів формулюються у вигляді диференціальних рівнянь. Дослідження цих моделей вимагає знаходження розв'язків, виразити які через відомі функції можливо лише у випадках для спрощених (модельних) задач. Тому актуальна задача побудови числових розв'язків диференціальних задач, визначених на дискретній множині точок – сітці.

Розглянемо диференціальне рівняння першого порядку

$$\dot{u} = f(t,u), \tag{12.1}$$

де $\dot{u} := \frac{du}{dt}$, $t \in \left[t_0, t_f \right]$, $u = u(t)$ – шукана функція незалежної змінної t, функція f визначена в області $G = [t_0, t_f] \times D$, $D \subseteq R$. Якщо (12,1) – система d звичайних диференціальних рівнянь (ЗДР), то $u = col(u_1,...,u_d)$, $f = col(f_1,..., f_d)$, $D \subseteq R^d$.

Рівняння (12.1) має сім'ю розв'язків, що залежить від довільної сталої. Тому для знаходження частинних розв'язків потрібно задавати додаткові умови, наприклад початкову умову

$$u(t_0) = u_0. \tag{12.2}$$

Розв'язком задачі Коші або початкової задачі (12.1), (12.2) на відрізку $[t_0, t_f]$ називається неперервно диференційовна функція $u = u(t)$, така, що $\left(t, u\left(t\right)\right) \in G$ і $\dot{u}(t) = f\left(t, u\left(t\right)\right)$ для

$t \in [t_0, t_f]$, і $u(t_0) = u_0$. У точках t_0 і t_f під похідною будемо розуміти праву й ліву похідні.

Як відомо [61], в разі неперервності функції f в області G, через кожну точку $(t_0, u_0) \in G$ проходить хоча б одна інтегральна лінія рівняння (12.1). Якщо, крім того, функція $f(t,u)$ в області G задовольняє по другому аргументу умову Ліпшиця або похідна $f_u(t,u)$ неперервна в G, то існує єдиний розв'язок задачі Коші (12.1), (12.2) в деякому околі довільної точки $\tau \in [t_0, t_f]$.

У більшості математичних моделей, які описуються диференціальними рівняннями, не вдається знайти точний розв'язок в аналітичному вигляді. Наприклад модель взаємодії хижака і жертви, яку запропонували А. Лотка (1925 р.) і В. Вольтерра (1926 р.) формулюється за допомогою системи двох нелінійних диференціальних рівнянь

$$\dot{u} = (a - bv)u,$$
$$\dot{v} = (-cu + dv)v,$$

де $u(t)$ і $v(t)$ – величини популяцій (чисельність або біомаса), $t \in [0, T]$, $a, b, c,$ і d – додатні сталі.

Із другого закону Ньютона одержується рівняння коливання математичного маятника. Нехай на невагомій нитці довжиною l підвішений вантаж масою m. За відсутності опору руху маятника диференціальне рівняння другого порядку, яке визначає його рух, записується у вигляді [61]

$$\ddot{u}(t) + \omega^2 \sin u(t) = 0,$$

де $u(t)$ – кут між ниткою і вертикаллю в момент часу t, $\omega^2 = g/l$, $g \approx 9.8 м/с^2$ – прискорення земного тяжіння.

У деяких випадках, наприклад для рівняння $\dot{u} = \dfrac{u - t}{u + t}$, можна знайти інтеграл $\ln(t^2 + u^2) + 2arctg\dfrac{u}{t} = C$, але для обчислення значень розв'язку $u = u(t)$ у точці t потрібно наближено розв'язати трансцендентне рівняння.

Для розв'язування диференціальних задач застосовуються числові методи, за допомогою яких знаходяться наближені значення розв'язку у вузлах t_n сітки $a = t_0 < t_1 < ... < t_N = b$, яку позна-

чимо через $\Delta_h\left[t_0, t_f\right]$. Для рівномірної сітки $h=(t_f-t_0)/N$, тоді $t_n=t_0+nh$, $\overline{n=0,N}$. Для нерівномірної сітки $h=\left(h_1,...,h_N\right)$ – вектор кроків, $h_n=t_n-t_{n-1}$, $n=\overline{1,N}$. Отже, числовий розв'язок диференціальної задачі – це сіткова функція $y_n=y\left(t_n\right)$, визначена на сітці $\Delta_h\left[t_0,t_f\right]$. Якщо задано умову (12.2), то $y_0=u_0$.

В однокрокових числових методах наближення y_{n+1} для точного значення $u_{n+1}=u\left(t_{n+1}\right)$ обчислюється на підставі відомого наближення y_n у вузлі t_n. Явні однокрокові методи можна записати у вигляді

$$y_{n+1}=F\left(y_n,h\right),$$

а неявні так

$$y_{n+1}=G\left(y_n,y_{n+1},h\right),$$

де функції F і G визначаються відповідним числовим методом.

У багатокрокових методах стартовими є m, $m>1$, значень $y_{n-m+1},...,y_n$ для обчислення наближеного розв'язку y_{n+1}.

12.2. Числові методи розв'язування задачі Коші, які ґрунтуються на формулі Тейлора

Такі методи будуються шляхом розкладу розв'язку $u=u\left(t+h\right)$ рівняння (12.1), (12.2) за степенями h в точці $t=t_n$ згідно з формулою Тейлора. Припустимо, що функція $f\left(t,u\right)$ має в області G неперервні частинні похідні до порядку r. Тоді розв'язок $u\left(t\right)$ рівняння (12.1) диференційовний до порядку $r+1$. Застосувавши формулу Тейлора, одержимо

$$u_{n+1}=u_n+h\dot{u}_n+\frac{h^2}{2}\ddot{u}_n+...+\frac{h^r}{r!}u_n^{(s)}+\frac{h^{r+1}}{(r+1)!}u^{(r+1)}\left(t_n+\theta\cdot h\right), \quad (12.3)$$

де $u_n^{(i)}=u^{(i)}\left(t_n\right), i=\overline{0,\,n+1}$, $\theta\in(0,1)$.

За наближене значення розв'язку в точці t_{n+1} можна взяти $m+1$, $m\le r$, перших доданків у правій частині рівності (12.3)

$$u_n+h\dot{u}_n+...+\frac{h^m}{m!}u_n^{(m)}.$$

Значення похідних обчислюються згідно з рівнянням (12.1):

$$\dot{u}_n = f\left(t_n, u_n\right), \quad \ddot{u}_n = \left(f_t + f_u f\right)_n, \tag{12.4}$$
$$\dddot{u}_n = \left(f_{tt} + 2ff_{tu} + f^2 f_{uu} + \left(f_t + f_u f\right) f_u\right)_n, \ldots$$

Якщо значення u_n відомо, точно чи наближено, то одержимо формулу для обчислення наближеного розв'язку

$$y_{n+1} = y_n + h\dot{y}_n + \ldots + \frac{h^m}{m!} y_n^{(m)}, \quad n = \overline{0, N-1}, \quad y_0 = u_0,$$

де похідні $y_n^{(i)}$ обчислюються згідно з формулами (12.4).

Недоліками цього методу є потреба обчислення функції та її похідних (12.4). Це вимагає написання блоків обчислення похідних, що суперечить тенденції спрощення стосунків між користувачем і комп'ютером. Наприклад, для рівняння $\dot{u} = t^2 + u^2$ із досить простою правою частиною маємо

$$\ddot{u} = 2u^3 + 2ut^2 + 2t,$$
$$\dddot{u} = 6u^4 + 8u^2t^2 + 4ut + 2t^4 + 2,$$
$$u^{(4)} = 4u(6u^2 + 4t^2 + t)(t^2 + u^2) + 4(4u^2t + u + 2t^3).$$

Функцію $y_h = y_h\left(t\right)$, визначену на сітці $\Delta_h\left[t_0, t_f\right]$, будемо називати сітковою. Для таких функцій природно визначається операція множення на число й арифметичні операції

$$\left(y_h \circ z_h\right)\left(t\right) = \{y_h\left(t\right) \circ z_h\left(t\right), \ t \in \Delta_h\},$$

де символом $\circ$ позначено операції + або $-$. Простір сіткових функцій можна нормувати. Надалі використовуватимемо такі норми для сіткових функцій:

$$\left\|y_h\right\|_1 = \max_{t \in \bar{\Delta}_h} \left|y_n\left(t\right)\right|, \quad \left\|y_h\right\|_2 = \left(\sum_{t \in \bar{\Delta}_h} h y_n^2\left(t\right)\right)^{1/2}.$$

12.3. Методи Ейлера

12.3.1. Побудова різницевих схем. У вузлі t_n, $n = \overline{0, N-1}$, сітки $\Delta_h\left[t_0, t_f\right]$ апроксимуємо похідну $\dot{u}_n$ правою різницевою похідною (10.2). Тоді у цьому вузлі рівняння (12.1) можна замінити дискретним аналогом

$$\frac{y_{n+1} - y_n}{h} = f\left(t_n, y_n\right), \ n = \overline{0, N-1}; \ \ y_0 = u_0. \tag{12.5}$$

Наближений розв'язок рівняння (12.5) послідовно обчислюється

згідно з рекурентною формулою, якою визначається явний метод Ейлера

$$y_{n+1} = y_n + hf\left(t_n, y_n\right),\ n = \overline{0, N-1},\ y_0 = u_0. \qquad (12.6)$$

Враховуючи, що $f\left(t_n, y_n\right)$ – кутовий коефіцієнт дотичної до інтегральної кривої у точці $\left(t_n, y_n\right)$, явний метод Ейлера має просту геометричну ілюстрацію (рис. 12.1).

Якщо ж апроксимувати похідну у вузлі t_{n+1}, $n = \overline{0, N-1}$, лівою різницевою похідною (10.3), то наближене значення розв'язку y_{n+1} визначається із рівняння

$$\frac{y_{n+1} - y_n}{h} = f\left(t_{n+1}, y_{n+1}\right),\ n = \overline{0, N-1};\ \ y_0 = u_0$$

і маємо неявний метод Ейлера

$$y_{n+1} = y_n + hf\left(t_{n+1}, y_{n+1}\right),\ n = \overline{0, N-1},\ y_0 = u_0. \qquad (12.7)$$

Ілюстрації методів наведені на рис. 12.2.

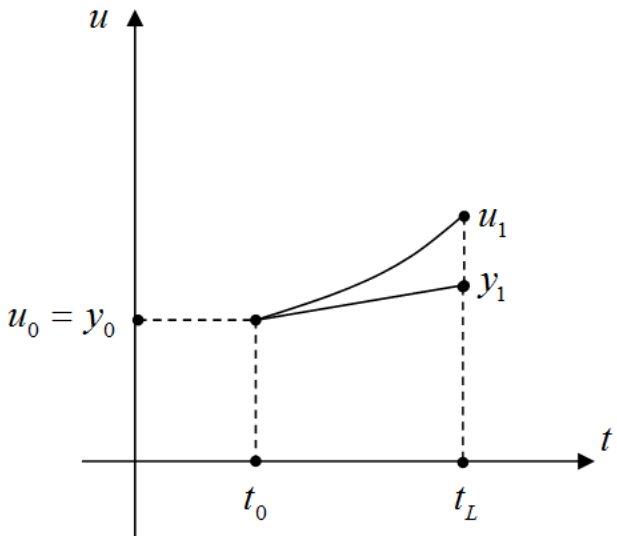


Рис. 12.1. Явний метод Ейлера

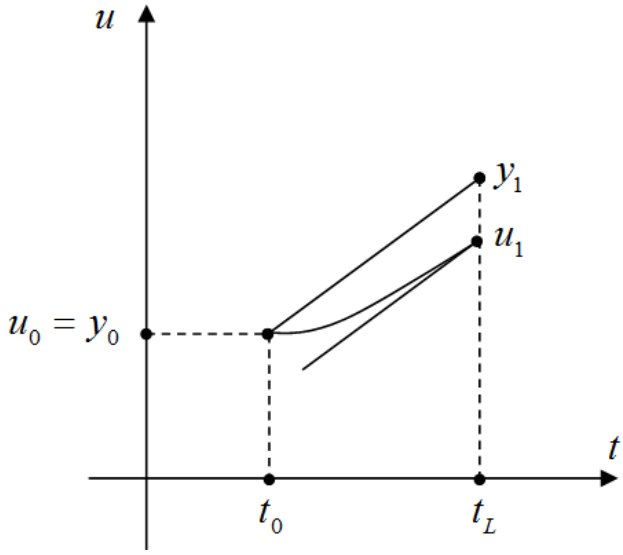


Рис. 12.2. Неявний метод Ейлера

Із лінійної комбінації явного і неявного методів Ейлера одержимо симетричну РС (метод трапецій)

$$\frac{y_{n+1} - y_n}{h} = \frac{1}{2}[\, f(t_n, y_n) + f(t_{n+1}, y_{n+1})],\ n = \overline{0, N-1}, \qquad (12.8)$$

яка є неявною і має на порядок вищу точність, ніж методи Ейлера.

Для кожного n рівняння (12.7) можна розв'язати, наприклад, методом простої ітерації (5.13)

$$y_{n+1}^{(k+1)} = y_n + h \cdot f\left(t_{n+1}, y_{n+1}^{(k)}\right) \equiv g\left(y_{n+1}^{(k)}\right),\ k = 0,1,... \qquad (12.9)$$

Початкове наближення $y_{n+1}^{(0)}$ обчислюється за формулою явного методу Ейлера (12.6)

$$y_{n+1}^{(0)} = y_n + h \cdot f\left(t_n, y_n\right).$$

Якщо похідна $f_u(t,u)$ обмежена в області G сталою $M_1 > 0$, то досить малого h

$$\left|g'(u)\right| = h \cdot \left|f_u(t,u)\right| \le hM_1 < 1.$$

Тому ітераційний метод (12.9) збіжний при $k \to \infty$.

Для початкового наближення $y_{n+1}^{(0)}$, близького до точного розв'язку, із порядком 2 збігається метод Ньютона:

$$y_{n+1}^{(k+1)} = y_n^{(k)} + [1 - h\frac{\partial f}{\partial y}(t_{n+1}, y_{n+1}^{(k)})]^{-1}(hf(t_{n+1}, y_{n+1}^{(k)}) + y_n - y_{n+1}^{(k)}), k = 0,1,... \quad (12.10)$$

Формула (12.10) одержується при застосуванні методу Ньютона до рівняння $F(y) \equiv y - y_i - hf(t_{n+1}, y) = 0$.

Зауваження 12.1. *Формули (12.6) і (12.7) можна одержати й іншими способами. Наприклад. проінтегрувавши рівняння (12.1) у межах від t_n до t_{n+1} і застосувавши для обчислення інтеграла*

$$\int_{t_n}^{t_{n+1}} f(t, u(t))dt$$

формули лівих або правих прямокутників, дістанемо відповідно явний і неявний методи Ейлера. ■

У числовому методі можуть використовуватись і значення розв'язку не у вузлах сітки Δ_h. Наприклад, обчислюючи інтеграл методом центральних прямокутників (11.3) одержимо формулу:

$$y_{n+1} = y_n + hf(t_n + h/2, y(t_n + h/2)),$$

де $y(t_n + h/2) = y_n + hf(t_n, y_n)/2$, якою визначається метод Рунге–Кутти другого порядку.

12.3.2. Похибка апроксимація РС Ейлера. Позначимо через $z_h(t) = y_h(t) - u_h(t),\ t \in \Delta_h\left[t_0, t_f\right]$, сіткову функцію, яка характеризує похибку методу у вузлах сітки. Нехай $\|\cdot\|$ − деяка сіткова норма. Величина $z_n = y_n - u_n$ − похибка методу у вузлі t_n, а $\|z_h\| = \|y_h - u_h\|$ − глобальна похибка на сітці. Виведемо рівняння

для похибки явного методу Ейлера у вузлі t_n. Підставивши $y_n = u_n + z_n$ у формулу (12.6), одержимо

$$\frac{z_{n+1} - z_n}{h} = \psi_n^{(1)} + \psi_n^{(2)}, \qquad (12.11)$$

де $\psi_n^{(1)} = -\frac{u_{n+1} - u_n}{h} + f\left(t_n, u_n\right)$, $\psi_n^{(2)} = f\left(t_n, u_n + z_n\right) - f\left(t_n, u_n\right)$.

Функція $\psi_n^{(1)}$, $n = \overline{0, N-1}$, називається *нев'язкою.* або *похибкою апроксимації різницевого методу на розв'язку диференціальної задачі.* Значення $\psi_n^{(1)}$ одержується як результат підстановки значення точного розв'язку $u(t)$ задачі (12.1), (12.2) у РС (12.6), записану у вигляді

$$-\frac{y_{n+1} - y_n}{h} + f\left(t_n, y_n\right) = 0.$$

Означення 12.1. *Різницевий метод або РС апроксимує диференціальну задачу, якщо у кожному вузлі* $\psi_n^{(1)} \to 0$ *при* $h \to 0$ *і має* p*–й порядок апроксимації,* $p > 0$*, якщо* $\psi_n^{(1)} = O\left(h^p\right)$ *при* $h \to 0$ *або те ж саме, що для досить малих* h *виконується нерівність*

$$\left|\psi_n^{(1)}\right| \le Mh^p, \; M = const > 0. \qquad (12.12)$$

Нехай $u \in C^2[a,b]$, тоді за формулою Тейлора,

$$u_{n+1} = u_n + h \cdot \dot{u}_n + \frac{h^2}{2}\ddot{u}\left(t_n + \theta \cdot h\right), \; \theta \in (0,1)\cdot$$

де $M_2 = \max\limits_{t_0 \le t \le t_f} \left|\ddot{u}(t)\right|$, $M = M_2 / 2$. Враховуючи, що $f\left(t_n, u_n\right) = \dot{u}_n$, одержи-мо $\psi_n^{(1)} = -h\ddot{u}(t_n + \theta \cdot h)/2$ і оцінка похибки апроксимації

$$\left\|\psi_h^{(1)}\right\| = \max_{0 \le n \le N}\left|\psi_h^{(1)}\right| \le M_2 h / 2 = Mh.$$

Отже, метод має перший порядок апроксимації. Такий же порядок апроксимації має і неявний метод Ейлера (12.8).

Для $\psi_n^{(2)}$ одержимо

$$\psi_n^{(2)} = f_u(t_n, u_n + \theta z_n) z_n \;, \; \theta \in (0,1).$$

Тобто доданок $\psi_n^{(2)}$ пропорційний похибці z_n і має порядок, не менший, ніж z_n. Зокрема, $\psi_n^{(2)} \equiv 0$, якщо права частина диференціального рівняння не залежить від u.

12.3.3. Збіжність методу Ейлера. Доведемо збіжність явного методу Ейлера (12.7), тобто покажемо, що в кожній точці t_n сітки

$$z_n := y_n - u_n \to 0 \text{ при } h \to 0.$$

Означення 12.2. *Метод збігається зі швидкістю* $\mathrm{O}(h^p)$ *при* $h \to 0$, $p > 0$, *до розв'язку диференціальної задачі (має порядок точності* p*), якщо виконується рівність* $|y_n - u_n| = \mathrm{O}(h^m)$ *при* $h \to 0$, *або, що те ж саме, для досить малих* h *справджується нерівність* $|y_n - u_n| \le Ch^p$, *де стала* $C > 0$ *і не залежить від* h.

Теорема 12.1 [59]. *Якщо частинні похідні* $f_t(t,u)$, $f_u(t,u)$ *неперервні в області* G *і обмежені разом із функцією* $f(t,u)$ *сталою* $A > 0$, *то метод Ейлера має перший порядок точності.*

Доведення. Згідно з формулою (12.12), для похибки $z_n = y_n - u_n$ маємо $|\psi_n^{(1)}| \le Mh$, де $M = \max\limits_{t\in[a,b]} |\ddot{u}(t)| / 2 =$

$= \max\limits_{G} |f_t(t,u) + f_u(t,u) f(t,u)| / 2 \le A(1+A)/2$. Для $\psi_n^{(2)}$ одержимо

$$|\psi_n^{(2)}| = |f(t_n, u_n + z_n) - f(t_n, u_n)| \le A|z_n|.$$

Звідси маємо

$$|z_{n+1}| \le |z_n| + h|\psi_n^{(1)}| + hA|z_n| = (1+hA)|z_n| + h^2 M.$$

Аналогічну оцінку можна записати і для $z_n, z_{n-1}, \ldots, z_1$.

Оскільки

$z_0 = y_0 - u_0 = 0$ і $n+1 \le N$, то

$$|z_{n+1}| \le (1+hA)^2 |z_{n-1}| + (1+hA)h^2 M + h^2 M \le (1+hA)^{n+1} |z_0| +$$

$$+h^2 M \sum_{\nu=0}^{n} (1+hA)^{\nu} = h^2 M \frac{(1+hA)^{n+1} - 1}{Ah} \le \frac{Mh(1+hA)^N}{A}.$$

На підставі нерівності $1 + t \le e^t$, коли $t \ge 0$, і враховуючи, що $hN = t_f - t_0$, одержимо для довільного n, $0 \le n \le N-1$,

$$|z_{n+1}| \le A^{-1} Mhe^{hNA} = A^{-1} Me^{(t_f - t_0)A} h = Ch, \quad C = A^{-1} Me^{(t_f - t_0)A}.$$

Отже, метод Ейлера має перший порядок точності. ■

Зауваження 12.2. *Оскільки* $z_0 = 0$, *то*

$$z_1 = y_1 - u_1 = u_0 + hf_0 - u_1 =$$

$$= h\left[-\frac{u_1 - u_0}{h} + f(t_0, u_0) \right] = h\psi_0^{(1)}.$$

Тому, $|z_1| \le Mh^2$. *Нехай* $u \in C^3\left[t_0, t_f\right]$. *Тоді можна записати*

$$z_1 = -\ddot{u}(t_0)h^2 / 2 + o(h^2). \qquad \blacksquare$$

Вираз $e_1(h) := -\ddot{u}(t_0)h^2 / 2$ називається *головною складовою похибки* (ГСП) явного методу Ейлера на кроці й використовується для аналізу швидкості збіжності різницевого методу та порівняння його з іншими методами. Тобто, метод Ейлера збігається з першим порядком точності, а похибка цього методу на кроці має другий порядок щодо кроку сітки Δ_h.

12.4. Явні методи Рунге–Кутти

12.4.1. Загальна схема явних методів Рунге-Кутти. Оскільки для глобальної похибки явного методу Ейлера виконується оцінка $\|y_h - u_h\| \le Ch$, то, щоб одержати у наближеному розв'язку шість правильних десяткових цифр, потрібно орієнтуватися на крок сітки $h_1 = 10^{-6}$. На сітці $\Delta_h[0,1]$ для цього потрібно обчислити 10^6 значень функції $f(t,u)$. У 1895 р. К. Рунге запропонував для обчислення y_{n+1} застосувати формулу

$$y_{n+1} = y_n + hf\left(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n) \right), \quad n = 0,1,\ldots \quad y_0 = u_0,$$

яку зручно записати у вигляді

$$y_{n+1} = y_n + hk_2(h), \quad n = 0,1,\ldots, \tag{12.13}$$

де $k_2(h) = f\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1 \right)$, $k_1 = f(t_n, y_n)$.

Формулу (12.13) в методі Рунге можна одержати, якщо проінтегрувати рівняння (12.1) у межах від t_n до t_{n+1} і застосувати для обчислення інтеграла формулу центральних прямокутників

$$u_{n+1} \approx u_n + f\left(t_n + \frac{h}{2}, u\left(t_n + \frac{h}{2} \right) \right).$$

Відтак значення розв'язку $u\left(t_n + \frac{h}{2} \right)$ наближено обчислити явним методом Ейлера

$$u\left(t_n+\frac{h}{2}\right)\approx u_n+\frac{h}{2}\,f\left(t_n,u_n\right).$$

Точність методу (12.13) $p=2$, і щоб одержати шість правильних цифр потрібно орієнтуватись уже на крок $h_2\approx 10^{-3}=1000h_1$. Це ефективніше, порівняно з методом Ейлера, хоч і вимагає у кожному вузлі обчислювати два значення функції f.

У 1901 р. К. Кутта сформулював загальну схему явних однокрокових методів, відомі як методи Рунге–Кутти, і побудував методи порядку 3, 4 і 5.

Зафіксуємо ціле $s\geq 1$, яке назвемо числом стадій, а відповідний метод s-стадійним. Введемо позначення:

$$\begin{aligned}
&k_1(h)=f\left(t_n,y_n\right),\\
&k_2\left(h\right)=f\left(t_n+c_2h,y_n+a_{21}hk_1\right),\\
&k_3\left(h\right)=f\left(t_n+c_3h,y_n+a_{31}hk_1+a_{32}hk_2\right),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&k_s\left(h\right)=f\left(t_n+c_sh,y_n+a_{s1}hk_1+\ldots+a_{s,s-1}hk_{s-1}\right),
\end{aligned}\qquad (12.14)$$

де коефіцієнти c_i,a_{ij} не залежать від кроку h і правої частини рівняння (12.1).

Згідно з методом Рунге–Кутти, наближене значення розв'язку y_{n+1} за відомим значенням y_n обчислюється згідно з формулою

$$y_{n+1}=y_n+h\cdot\left(b_1k_1(h)+\ldots+b_sk_s(h)\right),\qquad (12.15)$$

де коефіцієнти $b_1,\ldots,b_s$ також не залежать від h і функції f.

Для заданого s підлягає визначенню $\left(s^2+3s-2\right)/2$ коефіцієнтів a_{ij},b_i та c_j, які зручно записати у таблиці Бутчера [76, 83].

Таблиця 12.1
Коефіцієнти явних методів Рунге–Кутти

c_2	a_{21}				
c_3	a_{31}	a_{32}			
...	...	...			
c_s	a_{s1}	a_{s2}	..	$a_{s,s-1}$	
	b_1	b_2	..	b_{s-1}	b_s

Коефіцієнти a_i, b_{ij}, c_j визначаються так, щоб на точному розв'язку задачі (12.1), (12.2) розклад похибки апроксимації методу

$$\psi_n^{(1)} = -\frac{u_{n+1}-u_n}{h} + b_1 k_1(h) + \ldots + b_s k_s(h), \qquad (12.16)$$

за степенями h починався із максимального високого степеня h для довільного кроку сітки і довільної правої частини рівняння (12.1) із деякого класу функцій. Найчастіше такий клас функцій визначається їх гладкістю за змінними t, u. У формулі (12.16) вирази для $k_\nu(h)$ залежать від u_n.

Зауваження 12.1. *Коефіцієнти явних методів Рунге–Кутти здебільше задовольняють умови*
або $c_2 = a_{21},\ c_3 = c_{31} + c_{32},\ \ldots,\ c_s = a_{s1} + \cdots + a_{s,s-1}$

$$c_i = \sum_{j=1}^{i-1} a_{ij}, \quad i = \overline{2,s}.$$

Як зазначено в монографії [75]*, ці умови були прийняті Куттою без пояснень. Зміст їх полягає в тому, що всі точки, в яких обчислюється функція f, є наближеннями першого порядку до розв'язку. У* [75, розділ II.2] *наведені загальні умови порядку для s-стадійних методів Рунге–Кутти порядку p.*

Доведено [2, 75, 84], що $m = s$ для $s = \overline{1,4}$. Позначимо такі методи через РКm. Методи із $s \geq 5$ стадіями мають порядок апроксимації $m < s$.

Нехай $s = 1$ і $u \in C^2[a,b]$. Тоді

$$y_{n+1} = y_n + b_1 h k_1(h),\ \ k_1(h) = f(t_n, y_n),$$

$$\psi_n^{(1)} = -\frac{u_{n+1}-u_n}{h} + b_1 f(t_n, u_n) = (1-b_1)\dot{u}_n - 0.5h\ddot{u}(t_n + \theta h),\ \ \theta \in (0,1).$$

Тільки при $b_1 = 1$ метод має перший порядок апроксимації, отже, метод РК1 – явний метод Ейлера (12.6).

12.4.2. Метод Рунге–Кутти другого порядку. У цьому випадку

$$y_{n+1} = y_n + h \cdot \left(b_1 k_1(h) + b_2 k_2(h)\right), \qquad (12.17)$$

де $k_1(h) = f(t_n, y_n),\ k_2(h) = f(t_n + c_2 h, y_n + a_{21} h k_1)$. Задача полягає у визначенні коефіцієнтів b_1, b_2, a_{21}, c_2 так, щоб розклад за степенями h похибки апроксимації

$$\psi_n^{(1)} = -\frac{u_{n+1}-u_n}{h} + b_1 f_n(t_n,u_n) + b_2 f\left(t_n + c_2 h, u_n + a_{21} h f_n\right),$$

починався з максимально високого степеня h.

Припустимо, що $u \in C^3\left[t_0, t_f\right]$. Для цього досить, щоб існували другі неперервні частинні похідні функції $f(t,u)$ в області G. Нехай у цій області $\max\left(\left|f_{uu}\right|,\left|f_{tu}\right|,\left|f_{tt}\right|\right) \le M_2$. Згідно з формулою Тейлора

$$u_{n+1} = u_n + h\dot{u}_n + \frac{1}{2}h^2\ddot{u}_n + \frac{1}{6}h^3\dddot{u}\left(t_n + \theta_1 \cdot h\right),\ \ \theta_1 \in (0,1);$$

$$k_2(h) = f_n + \left(f_t\right)_n c_2 h + \left(f_u\right)_n f_n a_{21} h + h^2 r_2(h),$$

де $f_n = f(t_n,u_n)$, $(f_t)_n = f_t(t_n,u_n)$ $(f_u)_n = f_u(t_n,u_n)$,

$$r_2(h) = \left[c_2^2 f_{tt}(A_n) + 2c_2 a_{21} f_n(A_n) f_{tu}(A_n) + a_{21}^2 f_n^2 f_{uu}(A_n)\right]/2,$$

$A_n := \left(t_n + \theta_2 h, u_n + \theta_2 h a_{21} f_n\right)$ – точка в області G, $\theta_2 \in (0,1)$.

Урахувавши ці розклади і замінивши похідні $\dot{u}_n, \ddot{u}_n$ значеннями f_n і $\left(f_t\right)_n + \left(f_u\right)_n f_n$ $(f_t + f_u f)_n$ відповідно, одержимо

$$\psi_n^{(1)}(h) = \left(b_1 + b_2 - 1\right) f_n + \left(b_2 c_2 - \frac{1}{2}\right) h\left(f_t\right)_n + \left(a_{21} b_2 - \frac{1}{2}\right) h\left(f_{tu}\right)_n +$$

$$+\left(b_2 r_2(h) - \frac{1}{6}\dddot{u}\left(t_n + \theta_1 \cdot h\right)\right) h^2.$$

Другий порядок апроксимації досягається, якщо коефіцієнти задовольняють систему рівнянь

$$\begin{aligned} b_1 + b_2 &= 1, \\ b_2 c_2 &= 1/2, \\ a_{21} b_2 &= 1/2. \end{aligned} \qquad (12.18)$$

Справді, для кожного такого набору коефіцієнтів b_1, b_2, c_2, a_{21} маємо

$$\left\|\psi_n^{(1)}\right\|_1 \le \left|b_2\right|\left(M_3 + 3M_2\left(c_2^2 + 2\left|c_2 a_{21}\right| M_2 + a_{21}^2 M_1^2\right)\right) h^2/6 =: Mh^2.$$

Нехай $b = b_2$. Із другого рівняння випливає, що $a_{21} = a_2$. Одержимо однопараметричну сім'ю розв'язків:

$$b_1 = 1 - b,\ \ c_2 = a_{21} = 1/(2b),\ \ b \ne 0.$$

Якщо $u \in C^4\left[t_0, t_f\right]$, то можна виділити ГСП апроксимації, яка набуває вигляду $(bcr_2(0) - \ddot{u}_n / 6)h^2 =$

$= [(3c_2 - 2) f_{tt} + (3a_{21} - 2)(2 f_{tu} - ff_{uu}) f - 2 f_u (f_t - ff_u)]_n h^2 / 12.$

Якщо функція у правій частині рівняння (5.1) не залежить від u, тобто $f = f(t)$, то $f_u \equiv 0$ і для $c_2 = a_{21} = 2/3$ тоді ГСП$= 0$ і метод має порядок вищий, ніж 2.

Найчастіше використовуються методи порядку 2, одержані Рунге[1], коли $b = 1$, $b = 1/2$ і $b = 3/4$ (табл. 12.2-12.4).

Таблиця 12.2.

$\frac{1}{2}$	$\frac{1}{2}$	
	0	1

Таблиця 12.3.

1	1	
	$\frac{1}{2}$	$\frac{1}{2}$

Таблиця 12.4.

$\frac{2}{3}$	$\frac{2}{3}$	
	$\frac{1}{4}$	$\frac{3}{4}$

У загальному випадку для двостадійного методу $(s = 2)$ порядок не може бути більшим, ніж $m = 2$. Справді, для задачі

$\dot{u} = u, \ \ t \in [0,1]; \ \ u(0) = 1$

похибка апроксимації

$$\psi_n^{(1)} = (1 - b_1 - b_2) \cdot u_n + \left(b_2 a_{21} - 1/2\right) \cdot h u_n - \frac{1}{6} h^2 \ddot{u}(t_n + \theta \cdot h).$$

Згідно з (12.18) і враховуючи, що $\ddot{u}(t) = u(t) = e^t \geq 1$, одержимо

$$\left\| \psi_h^{(1)} \right\| = \frac{h^2}{6} \max_{0 \leq t \leq 1} |u(t)| \geq \frac{h^2}{6}.$$

12.4.3. Методи Рунге–Кутти третього порядку. Розглянемо тристадійний явний метод Рунге−Кутти (s = 3):

$y_{n+1} = y_n + h(b_1 k_1(h) + b_2 k_2(h) + b_3 k_3(h)), \ m = 0,1,...; \ y_0 = u_0;$

$k_1(h) = f(t_n, y_n), \ \ k_2(h) = f(t_n + c_2 h, y_n + a_{21} h k_1),$

$k_3(h) = f(t_n + c_3 h, y_n + a_{31} h k_1 + a_{32} h k_2).$

Нехай $f \in C^4(G)$. Розкладемо за формулою Тейлора $u(t_{n+1}) = u(t_n + h)$ в точці t_n до величин порядку $O(h^4)$ і використаємо в одержаному розкладі вирази (12.5) для похідних $\dot{u}_n$,

[1] В літературі метод з коефіцієнтами в табл. 12.2 називають методом Рунге або Ейлера-Коші, в табл. 12.2 – методом Гюна [45].

$\ddot{u}_n, \ddot{u}_n$. Якщо також розкласти $k_2(h)$ і $k_3(h)$ у точці (t_n, u_n) до величин порядку $O(h^3)$ і прирівняти до нуля коефіцієнти при h^ν, $\nu = 0,1,2$ у похибці апроксимації і $f_n, (f_u)_n, (f_t)_n, \ldots$, то в підсумку одержимо систему рівнянь [4, 6, 59, 75]:

$$b_1 + b_2 + b_3 = 1, \tag{12.20}$$

$$b_2 c_2 + b_3 c_3 = b_2 a_{21} + b_3 (a_{31} + a_{32}) = 1/2, \tag{12.21}$$

$$c_2^2 b_2 + c_3^2 b_3 = c_2 b_2 a_{21} + c_3 b_3 (a_{31} + a_{32}) = b_2 a_{21}^2 + b_3 (a_{31} + a_{32})^2 = 1/3, \tag{12.22}$$

$$a_2 c_3 b_{32} = c_3 b_{21} b_{32} = 1/6. \tag{12.23}$$

Із системи рівнянь (12.23) випливає, що

$$c_2 = a_{21}$$

і матимемо одне рівняння

$$c_2 b_3 a_{32} = 1/6.$$

Із (12.22) одержується $a_{31} + a_{32} = c_3$.

Отже, для знаходження восьми параметрів маємо систему із шести рівнянь. Визначимо через a_2, a_3 всі інші невідомі. Із системи рівнянь (12.21) і (12.22) знаходимо значення

$$b_3 = \frac{2 - 3c_2}{6c_3(c_3 - c_2)}, \; b_2 = \frac{3c_3 - 2}{6c_2(c_3 - c_2)}.$$

Тоді з (12.20) маємо $b_1 = 1 - b_2 - b_3$. Далі знаходимо $a_{21} = c_2$, $a_{32} = (6c_2 b_3)^{-1}$, $a_{31} = c_3 - a_{32}$, $c_3 \neq 0$, $c_2 \neq c_3$, $c_2 \neq 2/3$.

Найчастіше використовуються методи, коли $c_2 = 1/2$, $c_3 = 1$ (табл. 12.5) і $c_2 = 1/3$, $c_3 = 2/3$ (табл. 12.6). У першому випадку

$$y_{n+1} = y_n + \frac{h}{6}(k_1 + 4k_2 + k_3),$$

у другому випадку

$$y_{n+1} = y_n + \frac{h}{4}(k_1 + 3k_3).$$

Існує ще дві однопараметричні сім'ї розв'язків. Якщо припустити, що $c_3 = 0$, то $a_{31} + a_{32} = 0$,

$$c_2^2 b_2 = \frac{1}{3}, \; c_2 b_2 = \frac{1}{2}, \; c_2 b_3 a_{32} = \frac{1}{6}, \; c_2 = a_{21}, \; b_1 = 1 - b_1 - b_2.$$

Таблиця 12.5.
Метод Рунге

$\frac{1}{2}$	$\frac{1}{2}$		
1	–1	2	
	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$

Таблиця 12.6
Метод Хойна

$\frac{1}{3}$	$\frac{1}{3}$		
$\frac{2}{3}$	0	$\frac{2}{3}$	
	$\frac{1}{4}$	0	$\frac{3}{4}$

Звідси одержимо таку сім'ю розв'язків:

$$c_2 = a_{21} = 2/3, \quad b_2 = 3/4, \quad a_{32} = -a_{21}, \quad b_1 = 1/4 - b_3, \quad b_3 \in R/\{0\}.$$

Якщо $c_2 = c_3$, то $c_2 = a_3 = a_{21} = 2/3$,

$$a_{32} = 1/4b_3,\ a_{31} = \frac{8b_3 - 3}{4b_3},\ b_2 = \frac{3}{4} - b_3,\ b_1 = 1/4,\ b_3 \in R\backslash\{0\}.$$

Отже, для $s = 3$ порядок апроксимації $p = 3$ і маємо одну двопараметричну і дві однопараметричні сім'ї розв'язків. Існують і чотиристадійні методи третього порядку, запропоновані Рунге (12.7), Хойном (12.8) та ін.

12.4.4. Метод Рунге–Кутти четвертого порядку. Розглянемо чотиристадійний метод

$$y_{n+1} = y_n + h \cdot (b_1k_1 + b_2k_2 + b_3k_3 + b_4k_4),$$

де $k_1 = f(t_n, y_n)$, $k_2 = f(t_n + c_2h, y_n + a_{21}hk_1)$,

$k_3 = f(t_n + c_3h, y_n + a_{31}hk_1 + a_{32}hk_2)$,

$k_4 = f(t_n + c_4h, y_n + a_{41}hk_1 + a_{42}hk_2 + a_{43}hk_3)$.

У випадку $s = 4$ вдається знайти коефіцієнти методу так, щоб порядок апроксимації дорівнював числу стадій, тобто $p = s = 4$.

Таблиця 12.7
Явний чотиристадійний метод Рунге-Кутти

c_2	a_{21}			
c_3	a_{31}	a_{32}		
c_4	a_{41}	a_{42}	a_{43}	
	b_1	b_2	b_3	b_4

Похибка апроксимації методу набуває вигляду

$$\psi_n^{(1)} = -\frac{u_{n+1}-u_n}{h} + b_1k_1 + b_2k_2 + b_3k_3 + b_4k_4 .$$

Припустимо, що функція $f(t,u)$ за змінними t,u має неперервні похідні до четвертого порядку включно. Розкладемо u_{n+1} в точці t_n до величини порядку $O(h^5)$, функції $k_i(h)$, $i=2,3,4$ до величин $O(h^4)$ і врахуємо вирази для $u_n^{(i)}$, $i=\overline{1,4}$. Прирівнявши вирази біля $h^i, i=\overline{0,3}$ до нуля, одержимо для знаходження 13-ти коефіцієнтів систему з одинадцяти рівнянь [2, 4, 59, 75]:

$$\begin{aligned}
&b_1+b_2+b_3+b_4=1\ ,\ a_{21}=c_2,\\
&a_{31}+a_{32}=c_3\ ,\ a_{41}+a_{42}+a_{43}=c_4,\\
&b_2c_2+b_3c_3+b_4c_4=\frac{1}{2},\ b_2c_2^2+b_3c_3^2+b_4c_4^2=\frac{1}{3},\\
&b_2c_2^3+b_3c_3^3+b_4c_4^3=\frac{1}{4},\ (b_3a_{32}+b_4a_{42})c_2+c_4a_{43}c_3=\frac{1}{6},\\
&(b_3a_{32}+b_4a_{42})c_2^2+b_4a_{43}c_3^2=\frac{1}{12},\\
&b_3a_{32}c_2c_3+b_4a_{42}c_2c_4+b_4a_{43}c_3c_4=\frac{1}{8},\\
&c_2c_4a_{43}a_{32}=\frac{1}{24}.
\end{aligned} \qquad (12.24)$$

Система (12.24) має одну двопараметричну і три однопараметричних сім'ї розв'язків.

Двопараметрична сім'я розв'язків, залежна від параметрів c_2 і c_3, $c_2\neq 1/2$, $c_2\neq c_3$ і $c_3\neq 0$, набуває вигляду

$$\begin{aligned}
&c_4=1,\ b_2=\frac{2c_3-1}{12c_2(c_3-c_2)(1-c_2)},\quad b_3=\frac{1-2c_2}{12c_3(c_3-c_2)(1-c_3)},\\
&b_4=\frac{6c_2c_3-4c_2-4c_3+3}{12(1-c_2)(1-c_3)},\ a_{42}=-\frac{4c_3^2-c_2-5c_3+2}{24b_4c_2(c_3-c_2)(1-c_3)},\\
&a_{43}=\frac{1-2c_2}{12b_4c_3(c_3-c_2)},\ a_{41}=1-a_{42}-a_{43},\ a_{31}=c_2-a_{32}\ ,\ a_{21}=c_2,\\
&b_1=1-b_2-b_3-b_4.
\end{aligned}$$

Якщо $2c_2 = c_3 = 2/3$, то одержимо метод, який має назву „три восьмих” (табл. 12.9).

Якщо $a_3 = 0$, то одержується однопараметрична сім’я розв’язків, залежна від параметра $c_3 \neq 0$:

$$c_2 = \frac{1}{2} = a_{21},\ c_3 = 0,\ \ c_4 = 1,\ \ b_1 = \frac{1}{6} - b_3,\ \ b_2 = \frac{2}{3},\ \ b_4 = \frac{1}{6},$$

$$a_{31} = -a_{32} = \frac{1}{12b_3},\ \ a_{41} = -\frac{1}{2} - 6b_3,\ \ a_{42} = \frac{3}{2},\ \ a_{43} = 6b_3.$$

Нехай $c_2 = c_3$. Тоді є сім’я розв’язків, також залежна від $b_3 \neq 0$:

$$c_2 = c_3 = a_{21} = \frac{1}{2},\ c_4 = 1,\ b_2 = \frac{2}{3} - b_3,\ \ b_1 = b_4 = \frac{1}{6};$$

$$a_{31} = \frac{1}{2} - \frac{1}{6b_3},\ a_{32} = \frac{1}{6b_3},$$

$$a_{41} = 0,\ \ a_{42} = 1 - 3b_3,\ \ a_{43} = 3b_3,\ \ a_{42} = 1 - a_{43}.$$

Для $b_3 = 1/3$ одержується метод четвертого порядку (табл. 12.8), побудований Куттою в 1901 р., і який найчастіше використовується в обчислювальній практиці: $a_{31} = a_{41} = a_{42} = 0$,

$$c_2 = c_3 = a_{21} = \frac{1}{2},\ \ c_4 = a_{43} = 1,\ \ b_1 = b_2 = \frac{1}{6},\ \ b_2 = b_3 = \frac{4}{6}.$$

Таблиця 12.8

Класичний метод Рунге–Кутти

	$\frac{1}{2}$			
$\frac{1}{2}$				
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

Таблиця 12.9

Схема “три восьмих”

$\frac{1}{3}$	$\frac{1}{3}$			
$\frac{2}{3}$	$-\frac{1}{3}$	1		
1	1	–1	1	
	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Ще одна сім’я розв’язків для $c_2 = c_4$, залежна від параметра $b_4 \neq 0$, набуває вигляду:

$$c_2 = c_4 = a_{21} = 1,\ \ c_3 = \frac{1}{2},\ b_1 = \frac{1}{6},\ b_2 = \frac{1}{6} - b_4,\ b_3 = \frac{2}{3},$$

$$a_{31} = 3a_{32} = \frac{3}{8},\ a_{42} = -\frac{1}{6b_4},\ a_{41} = 1 + a_{42},\ a_{43} = -2a_{42}.$$

На прикладі задачі Коші $\dot{u} = (1-u)u,\ u(0) = 2,$ якою описується динаміка ізольованої популяції за умови конкуренції, порівняємо точність обчислення числового розв'язку методами Рунге–Кутти різного порядку. Точний розв'язок задачі $u(t) = 2e^t / (1 + 2(e^t - 1)).$ Числові розв'язки, обчислені методами РК2-4, коефіцієнти яких наведені в табл. 12.2 і 12.3, 12.5 і 12.6 , 12.8 і 12.9 подані у табл. 12.10.

Шість правильних цифр у наближеному розв'язку для $t = 5$ одержано методами РК3. Методом РК4-2 значення $u(5)$ знайдено з сімома правильними цифрами після крапки.

Таблиця 12.10

t[i]	Точний розв'язок	Метод Ейлера	Рунге-Кутти 2-1	Рунге-Кутти 2-2	Рунге-Кутти 3-1	Рунге-Кутти 3-2	Рунге-Кутти 4-1	Рунге-Кутти 4-2
0.0	2.00000	2.00000	2.00000	2.00000	2.00000	2.00000	2.00000	2.0000000
0.5	1.43527	1.39500	1.43915	1.43786	1.43513	1.43499	1.41490	1.4352685
1.0	1.22540	1.19663	1.22789	1.22713	1.22532	1.22524	1.21062	1.2254009
1.5	1.12557	1.10642	1.12710	1.12667	1.12553	1.12549	1.11565	1.1255756
2.0	1.07258	1.05991	1.07353	1.07327	1.07255	1.07253	1.06598	1.0725794
2.5	1.04280	1.03443	1.04339	1.04325	1.04278	1.04277	1.03842	1.0427994
3.0	1.02553	1.02002	1.02591	1.02582	1.02552	1.02551	1.02263	1.0255292
3.5	1.01533	1.01171	1.01557	1.01552	1.01532	1.01532	1.01342	1.0153303
4.0	1.00924	1.00688	1.00939	1.00936	1.00924	1.00924	1.00798	1.0092425
4.5	1.00559	1.00405	1.00568	1.00566	1.00558	1.00558	1.00476	1.0055856
5.0	1.00338	1.00239	1.00344	1.00343	1.00338	1.00338	1.00285	1.0033804

12.5. Методи Рунге–Кутти для систем диференціальних рівнянь

Методи Рунге-Кутти переносяться на системи диференціальних рівнянь. Розглянемо систему рівнянь порядку d у нормальній формі

$$\dot{u}_\nu = f_\nu(t, u_1, \ldots, u_d),\ t \in [t_0, t_f],$$

із початковими умовами $u_\nu(t_0) = u_{\nu,0},\ \nu = \overline{1,d}.$

У загальній схемі явного методу Рунге-Кутти позначимо функції $k_1(h),\ldots,k_s(h)$, які відповідають компоненті розв'язку u_ν, через $k_1^{(\nu)}(h),\ldots,k_s^{(\nu)}(h)$. Тоді схема методу набуде вигляду

$$y_{n+1}^{(\nu)} = y_n^{(\nu)} + h(b_1 k_1^{(\nu)} + \ldots + b_s k_s^{(\nu)}),\ \ \nu = \overline{1,d},$$

де

$$k_1^{(\nu)} = f_\nu(t_n, y_n^{(1)},\ldots, y_n^{(d)}),$$

$$k_2^{(\nu)} = f_\nu(t_n + c_2 h, y_n^{(1)} + a_{21} h k_1^{(1)},\ldots, y_n^{(d)} + a_{21} h k_1^{(d)}),$$

. . . . . .

$$k_s^{(\nu)} = f_\nu(t_n + b_s h, y_n^{(1)} + a_{s,1} h k_1^{(1)} + \ldots + a_{s,s-1} h k_{s-1}^{()},\ \ldots,$$

$$y_n^{(d)} + a_{s,1} h k_1^{(d)} + \ldots + a_{s,s-1} h k_{s-1}^{(d)}).$$

Запишемо формули явного методу Ейлера (РК1) і методу Рунге–Кутти порядку 4 для системи двох диференціальних рівнянь

$$\dot{u}_1 = f_1(t, u_1, u_2),$$

$$\dot{u}_2 = f_2(t, u_1, u_2).$$

Явний метод Ейлера набуває вигляду:

$$y_{n+1}^{(1)} = y_n^{(1)} + h f_1(t_n, y_n^{(1)}, y_n^{(2)}),$$

$$y_{n+1}^{(2)} = y_n^{(2)} + h f_2(t_n, y_n^{(1)}, y_n^{(2)}).$$

Формули класичного методу Рунге–Кутти порядку 4:

$$y_{n+1}^{(1)} = y_n^{(1)} + \frac{h}{6}(k_1^{(1)} + 2k_2^{(1)} + 2k_3^{(1)} + k_4^{(1)}),$$

$$y_{n+1}^{(2)} = y_n^{(2)} + \frac{h}{6}(k_1^{(2)} + 2k_2^{(2)} + 2k_3^{(2)} + k_4^{(2)}),$$

де

$$k_1^{(\nu)} = f_\nu(t_n, y_n^{(1)}, y_n^{(2)}),$$

$$k_2^{(\nu)} = f_\nu(t_n + \frac{h}{2}, y_n^{(1)} + \frac{h}{2}k_1^{(1)}, y_n^{(2)} + \frac{h}{2}k_1^{(2)}),$$

$$k_3^{(\nu)} = f_\nu(t_n + \frac{h}{2}, y_n^{(1)} + \frac{h}{2}k_2^{(1)}, y_n^{(2)} + \frac{h}{2}k_2^{(2)}),$$

$$k_4^{(\nu)} = f_\nu(t_n + h, y_n^{(1)} + h k_3^{(1)}, y_n^{(2)} + h k_3^{(2)}),\ \nu = 1, 2.$$

Математичними моделями сучасних прикладних задач служать системи диференціальних рівнянь високого порядку. Прискорення обчислювального процесу при розв'язуванні таких систем здійснюється через розпаралелювання обчислення правих

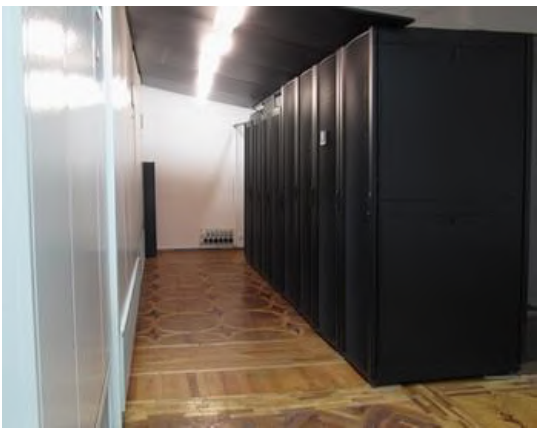


Рис. 12.3. Суперкомп'ютерний комплекс
СКІТ Інституту кібернетики НАН України

частин і матриці Якобі. Нехай кількість рівнянь n, а q – кількість процесорів, $l = \left[\frac{n}{q}\right]$ – ціла частина числа, $s = p(l+1) - n$. Тоді на $p - s$ реалізуються блоки по $l+1$ рівнянь, а на s процесорах по l рівнянь. Пакети програм розв'язування систем диференціальних рівнянь реалізовані на суперкомп'ютерному комплексі СКІТ, опис їх наведено в [51].

12.6. Явні методи Рунге–Кутти вищих порядків

Кількість стадій і порядок апроксимації явних методів Рунге – Кутти збігаються тільки для порядку 1–4. Ще в 1901 р. В. Кутта намагався побудувати п'ятистадійний метод п'ятого порядку, для якого потрібно визначити 15 параметрів. Негативну відповідь на це питання одержано тільки в 60-х роках [85]. Шестистадійний метод порядку $p = 5$, позначимо його через РК6(5), побудований у 1925 Х. Нюстремом (табл. А13). Такий же метод запропонований Р. Інглендом (табл. 12.11). Добре зарекомендував себе семистадійний метод п'ятого порядку Дормана–Прінса (табл. 12.13).

Теорема 12.2 [75, с. 198]. *Якщо* $p \geq 5$, *то не існує явних методів Рунг –Кутти із числом стадій* $s = p$, *тобто* $s \geq p+1$. ■

Методи шостого порядку з числом стадій $s = 7$ побудовані в 1964 р. Й. Бутчером [85], коефіцієнти в табл. А14 і А15.

Виявляється, що восьмистадійних методів сьомого порядку також не існує.

Теорема 12.3. [86]. *Для* $p \geq 8$ *не існує явних методів Рунге – Кутти порядку* p *таких, що* $s = p+1$. ■

Дев'ятистадійний метод сьомого порядку Бутчера з 55 коефіцієнтами наведено в табл. А16.

Методи восьмого порядку з 11 стадіями побудували А. Куртіс [88], Г. Купер й І. Вернер [87]. Тільки в 1985 р. Й. Бутчер довів, що не існує десятистадійних методів 8-го порядку. Таблиця коефіцієнтів методу 7-го й 8-го порядків Нормана–Прінса з 11 стадіями наведена в [75, с. 208].

Теорема 12.4 [89]. *Для* $p \geq 8$ *не існує явних методів Рунге-Кутти порядку m, які мають* $s = p+2$ *стадії.* ■

Таблиця 12.11

Метод Інгленда ($p = 5$)

$\frac{1}{2}$	$\frac{1}{2}$					
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$				
1	0	–1	2			
$\frac{2}{3}$	$\frac{7}{27}$	$\frac{10}{27}$	0	$\frac{1}{27}$		
$\frac{1}{5}$	$\frac{28}{625}$	$-\frac{125}{625}$	$\frac{54}{625}$	$\frac{546}{625}$	$-\frac{378}{625}$	
	$\frac{14}{336}$	0	0	$\frac{35}{336}$	$\frac{163}{336}$	$\frac{125}{336}$

У книгу рекордів Гіннеса занесений результат, одержаний у 1975 р. А. Куртісом [89]. Він побудував явний метод порядку 10 з 18 стадіями. Е. Хайрер у 1978 р. побудував метод 10-го порядку із 17 стадіями (табл. А19), але вже з ірраціональними коефіцієнтами [75, с. 203–204; 95].

Метод порядку 12 наведено в [99]. Нові явні методи Рунге–Кутта високого порядку побудував Т. Фейгін, застосувавши методику Е. Хайрера побудови явного методу Рунге–Кутти десятого порядку, і які відомі як *m*-симетричні методи [101]. Поняття *m*-симетрії значно спрощує генерування методів високого порядку з розумною кількістю стадій. Для таких методів 12-го порядку потрібно 25 стадій, а для методу 14-го порядку – 35. Посібник_20181123 10, 12 і 14 наведені у [101]. Функціонує сайт товариства методів Рунге–Кутти [102].

Коефіцієньт явних методів явних методів Рунге–Кутти, порядок яких 1–8 і 10 наведені в додатку А.

12.7. Збіжність явних методів Рунге–Кутти

Нехай наближене значення розв'язку задачі Коші (12.1), (12.2) обчислюється s-стадійним методом Рунге–Кутти (12.15), коефіці-єнти якого визначені в табл. 12.1, а функції k_ν згідно з (12.14). Як і раніше, позначимо через $z_n = y_n - u_n$ похибку методу у вузлі t_n, а через e_n – похибку методу на кроці, тобто похибку після одного кроку методу, за умови, що $y_n = u_n$. Зрозуміло, що $z_0 = e_0 = 0$, $z_1 = e_1$. Ці похибки проілюстровано на рис. 12.3, де u_N^ν – значення у вузлі t_N точного розв'язку із початковою умовою $u(t_\nu) = y_\nu$, $\nu = \overline{0, N-1}$.

Теорема 12.5 [59]**.** *Якщо порядок апроксимації методу Рунге–Кутти дорівнює m, то порядок похибки на кроці складає* $m+1$. *Тобто* $e_n(h) = O(h^{m+1})$ *або те ж саме, що для досить малих* h *виконується нерівність* $|e_n(h)| \le Ch^{m+1}$, $C = const > 0$. ■

У підрозділі 12.3.3 доведено збіжність явного методу Ейлера, тобто явного методу Рунге–Кутти першого порядку, і показано, що порядок точності $p = 1$. Сформулюємо теорему про збіжність й оцінку для глобальної похибки $\|z_n\| = \max\limits_{0 \le n \le N} |z_n|$ явних методів Рунге–Кутти. На рис. 12.4 глобальна похибка методу дорівнює $y_N - u_N^0$.

Введемо позначення: $A = \max\limits_{\substack{2 \le i \le s \\ 1 \le j \le s-1}} |a_{ij}|$, $B_0 = \max\limits_{1 \le i \le s} |b_i|$.

Теорема 12.6 [59]. *Нехай функція* f *задовольняє по другому аргументу умову Ліпшіця*

$$|f(t, u_2) - f(t, u_1)| \le L|u_2 - u_1|.$$

Тоді для похибки z_n *методу Рунге – Кутти на сітці* Δ_h *правильна оцінка*

$$|y_n - u_n| \le T \exp(\alpha T) \cdot \max_{0 \le i \le n-1} |\psi_i^{(1)}|, \ \alpha = sBL(1 + ALBh)^{s-1}, \ n = \overline{1, N}.$$

Якщо метод Рунге–Кутти апроксимує задачу (12.1), (12.2), то він збіжний при $h \to 0$, *причому порядок точності дорівнює порядку апроксимації.* ■

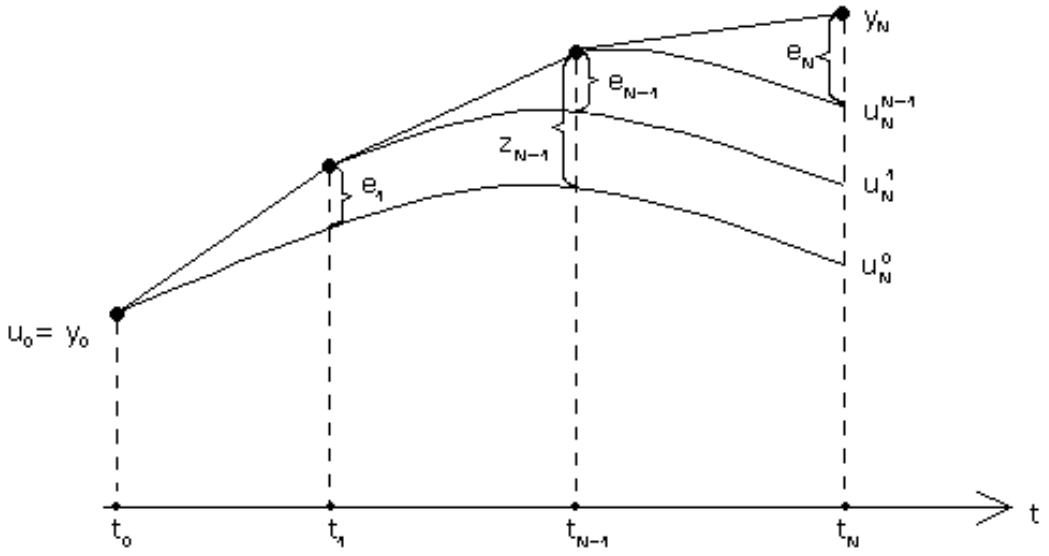


Рис. 12.4. Похибки методу Рунге–Кутти

12.8. Практичні способи оцінки похибки числового розв'язку

У теоремі 12.6 дається оцінка глобальної похибки числового розв'язку без знання значення y_n (апріорна оцінка), наприклад, оцінка (12.6) в теоремі 2.1 для явного методу Ейлера. Такого типу оцінки одержуються на підставі оцінок відповідних похідних розв'язку, які складно одержати і можуть бути завищеними. Тому в обчислювальній практиці найчастіше користуються наближеними оцінками, побудованими на підставі порівняння числових розв'язків, одержаних в одному й тому ж вузлі сітки з різними кроками або різними методами, тобто за результатами проведених обчислень (апостеріорні оцінки).

12.8.1. Оцінка похибки розв'язку за правилом Рунге[2]. Припустимо, що похибка однокрокового методу на кроці $z(h) = y(t_n + h) - u(t_n + h)$, де $y_n = u_n$, має вигляд

$$z(h) = Ah^{m+1} + o(h^{m+1}),\ A = A(t_n, u_n), \qquad (12.25)$$

[2] К. Рунге використовував з 1895 р. обчислення зі зменшенням удвічі кроку сітки, Л. Річардсон розвинув ідею оцінки похибки з різними кроками в працях 1910 і 1927 рр. [76, с. 175].

Доданок $A = A(t_n, u_n)h^{m+1}$ – ГСП. Формулу (12.25) можна одержати, припустивши, що функція $f(t,u)$ та всі її похідні до порядку m+1 неперервні в області G. Тоді $u \in C^{m+2}[a,b]$ і за формулою Тейлора

$$z(h) = \frac{z^{(m+1)}(0)}{(m+1)!}h^{m+1} + \frac{z^{(m+2)}(\theta h)}{(m+2)!}h^{m+2},\ \theta \in (0,1).$$

Похідна $z^{(m+1)}(0) = y^{(m+1)}(t) - u^{(m+1)}(t)$ явно виражається через значення в точці (t,u) функції f та її похідних до порядку, який не перевищує $m+1$. Розглянемо метод Ейлера (12.7), який має перший порядок точності. Якщо $u \in C^3$ і $y(t) = u(t)$, то похибка методу на кроці

$$z(h) = y(t+h) - u(t+h) = -\frac{1}{2}\ddot{u}(t)h^2 - \frac{1}{6}\dddot{u}(t+h\theta)h^3, \theta \in (0,1).$$

Отже, $m = 1$, $A = -\frac{1}{2}\ddot{u}(t)$, $\ddot{u}(t) = f_t(t,u) + f_u(t,u)f(t,u)$.

Для методу РК2, згідно з (12.19), головна складова похибки дорівнює $(C_2 r_2(0) - \dddot{u}(t)/6)h^3$.

Правило Рунге полягає в тому, що числовий розв'язок у вузлі t_{n+1} обчислюється одним і тим же методом із різними кроками й одержані значення розв'язку використовуються для оцінки похибки. Нехай у вузлі $t_{n+1} = t_n + h$ однокроковим методом із порядком точності p обчислено значення $y = y(t+h)$ наближеного розв'язку. Ігноруючи в (12.25) величину $O(h^{m+2})$, одержимо з точністю до величин $O(h^{m+1})$

$$y(t+h) - u(t+h) \approx Ah^{m+1}. \qquad (12.26)$$

Нехай $u = u(t+h)$, значення $y_1 = y(t+qh_1)$ обчислене з кроком $h_1 = h/q$, де q – ціле число, $q \geq 2$. Оскільки на кожному кроці довжиною h_1 похибка складає $O(h_1^{m+1})$, то за q кроків

$$y_1 - u \approx qA(h/q)^{m+1}. \qquad (12.27)$$

Помноживши (12.27) на q^m і віднявши від одержаного результату рівність (12.26), одержимо

$$q^m(y_1 - u) - (y - u) \approx 0, \text{ або } (q^m - 1)(y_1 - u) + y_1 - y_h \approx 0.$$

Звідси маємо апостеріорну похибку числового розв'язку y_1:

$$y_1 - u \approx \frac{1}{q^m - 1}(y - y_1). \qquad (12.28)$$

Оскільки значення y_h і y_{h_1} відомі, тому формула (12.28) використовується як для наближеного обчислення похибки $y_h - u$ в точці $t + h$, так і для уточнення y_{h_1} за формулою

$$u(t+h) \approx y_1 + \frac{1}{q^m - 1}(y_1 - y). \qquad (12.29)$$

У формулах (12.26) і (12.27) допущена похибка, порядок якої $O(h^{m+2})$, тому, згідно з (12.29), наближений розв'язок знаходиться з похибкою $O(h^{m+1})$, що на порядок точніш, ніж y_1.

В обчислювальній практиці часто $q = 2$, тобто крок ділиться пополам. Тоді для методу Ейлера ($m = 1$) маємо

$$y_1 - u \approx y - y_1.$$

У методі РК2 знаменник у формулі (12.29) дорівнює 3, а для РК3 ділити потрібно на 7. Для методу РК4

$$y_1 - u \approx \frac{1}{15}(y - y_1).$$

Якщо задана точність обчислення числового розв'язку $\varepsilon > 0$, то при виконанні нерівності

$$(2^m - 1)^{-1} \left| y - y_1 \right| < \varepsilon$$

вважається, що точність досягнена. Інакше, крок знову ділиться і в точці $t + h$ порівнюється з ε значення $\left| y_{h/2} - y_{h/4} \right| / (2^m - 1)$. Процес поділу кроку продовжується доти, доки не буде досягнута задана точність числового розв'язку. Однак варто обмежувати число послідовних поділів кроку, що здійснюється в одній точці. Наприклад, обмежуючи кількість поділів числом 10, ми допускаємо максимальне зменшення кроку в $2^{10} \approx 10^6$ разів. У більшості випадків цього досить, якщо тільки крок h не стає настільки малим, що не викликає зміни значення аргументу x. Останнє означає, що $x \oplus h = x$ для машинної арифметики додавання чисел із плаваючою крапкою.

Простіший, але менш строгий спосіб оцінки малості кроку h при обчисленні y_1 за методом РК4, полягає в обчисленні величини [73]

$$\theta = \left| \frac{k_2 - k_3}{k_1 - k_2} \right|.$$

Якщо величина не перевищує кількох сотих, то можна продовжувати обчислення з даним кроком або збільшувати його, інакше крок варто зменшити.

12.8.2. Застосування методів різного порядку точності. Апостеріорну оцінку похибки можна одержати, застосувавши методи різного порядку точності, наприклад *m* і *m*+1. Нехай $y^{(1)}$ знайдено в точці $t_{n+1} = t_n + h$ із порядком точності на кроці $m + 1$, а $y^{(2)}$ із порядком $m + 2$, тобто

$$y^{(1)} - u = O(h^{m+1}), \quad y^{(2)} - u = O(h^{m+2}).$$

Тоді $y^{(1)} - y^{(2)} = O(h^{m+1})$ і похибка наближеного розв'язку $y^{(1)}$ на кроці при малих *h* може бути досить точно наближена величиною

$$y^{(1)} - u \approx y^{(1)} - y^{(2)}.$$

Таблиця 12.12

Вкладені методи Фельдберга 4(5)

$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{2197}$	$\frac{7296}{2197}$				
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4105}$		
$\frac{1}{2}$	$\frac{16}{135}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	
c_ν	$\frac{25}{210}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0
$\bar{c}_\nu$	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$

Для методів Рунге-Кутти основні обчислення пов'язані із знаходженням величин $k_\nu(h)$. Тому для мінімізації обсягу обчислень доцільно використовувати *вкладені методи Рунге–Кутти* [76, 84], які для підвищення порядку точності на одиницю вимагають додаткового обчислення мінімальної кількості величин $k_\nu(h)$. У таблиці 12.12 подані коефіцієнти методів

Рунге–Кутти четвертого і п’ятого порядків точності, які відповідно мають вигляд

$$y_{n+1}^{(4)} = y_n + h\sum_{\nu=1}^{5} c_\nu k_\nu(h), \quad y_{n+1}^{(5)} = y_n + h\sum_{\nu=1}^{6} \bar{c}_\nu k_\nu(h).$$

Ще одним прикладом є вкладені формули четвертого і п’ятого порядку Дормана – Прінса (табл. 12.13):

$$y_{n+1}^{(4)} = y_n + h\sum_{\nu=1}^{7} c_\nu k_\nu(h) \qquad \text{(четвертий порядок)},$$

$$y_{n+1}^{(5)} = y_n + h\sum_{\nu=1}^{7} \bar{c}_\nu k_\nu(h) \qquad \text{(п’ятий порядок)}.$$

Цей метод має ту перевагу, що сталі множники в похибці формули п’ятого порядку мінімальні (серед аналогічних методів). Відзначимо також, що значення k_7 для $t = t_n$ дорівнює k_1 для наступного вузла t_{n+1}, оскільки коефіцієнти $b_{7\nu}$ і c_ν, $\nu = \overline{1,6}$ збігаються. Тому на кожному кроці потрібно виконати лише шість обчислень значень функції f. Інші методи наведені у додатку В.

Таблиця 12.13

Вкладені методи Дормана–Прінса 4(5)

$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$-\frac{2187}{6784}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
c_ν	$\frac{35}{384}$	0	$\frac{500}{1113}$	$-\frac{2187}{6784}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
$\bar{c}_\nu$	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$

12.9. Стійкість методів Рунге–Кутти

Показниками якості заміни диференціальної задачі різницевою служить порядок апроксимації на точному розв'язку або похибки методу на кроці, збіжність різницевого методу та порядок точності. Важливо також, щоб дискретна задача зберігала властивість стійкості розв'язку диференціальної задачі. Розглянемо це на прикладі задачі Коші

$$\dot{u} = \lambda u,\ t > 0,\ u(0) = u_0, \tag{12.30}$$

яку називатимемо модельною. Розв'язок задачі $u(t) = u_0 e^{\lambda t} \to 0$ при $t \to \infty$, якщо $\lambda < 0$ або $\mathrm{Re}\lambda < 0$ для $\lambda \in C$. Тобто розв'язок $u = 0$ диференціального рівняння $\dot{u} = \lambda u$ асимптотично стійкий і для довільних $\tau > 0$ і $u_0 \in R$ виконується нерівність

$$\left|u(t+\tau)\right| \le \left|u(t)\right|. \tag{12.31}$$

Якщо $u_0 \neq 0$ і $\tau > 0$, то нерівність строга. Природно вимагати такої ж поведінки розв'язку і для різницевої задачі.

Означення 12.3. *Розв'язок різницевої задачі називається стійким (метод стійкий або РС стійка), якщо виконується нерівність*

$$\left|y_{n+1}\right| \le \left|y_n\right|, \quad n = 0,1... \tag{12.32}$$

Розглянемо, як реалізується властивість стійкості розв'язку на прикладі явного методу Ейлера (РК1) для задачі (12.30). На підставі (12.7) і (12.30) маємо

$$y_{n+1} = (1+\lambda h) y_n.$$

Розв'язок РС стійкий, якщо $\left|1+\lambda h\right| \le 1$. Якщо λ набуває дійсних значень, то $-1 \le 1+\lambda h \le 1$ або $-2 \le \lambda h \le 0$. Отже, для $\lambda < 0$ метод стійкий, якщо крок сітки задовольняє умову

$$0 < h \le -2/\lambda. \tag{12.33}$$

Нехай $\lambda \in C$, $\mu := \lambda h = s + i\sigma$, де i – уявна одиниця. Тоді умова $\left|1+\mu\right| = \sqrt{(1+s)^2 + \sigma^2} \le 1$ виконується для всіх μ з одиничного круга комплексної площини із центром у точці $(-1,0)$ (рис. 12.5).

У випадку диференціального рівняння

$$\dot{u} = f(t,u)$$

вибір кроку для забезпечення стійкості визначається умовою [48]

$$h_{\min} < \frac{c}{\max|\lambda|},$$

де стала $c < 3$, $\max|\lambda| = \max\left|\frac{\partial f}{\partial u}\right|$ для скалярного рівняння і $\max|\lambda|$ – максимальне по модулю власних значення матриці Якобі для системи диференціальних рівнянь.

Для неявного методу Ейлера для модельної задачі

$$\frac{y_{n+1} - y_n}{h} = \lambda y_{n+1}, \quad y_0 = u_0, \tag{12.34}$$

маємо $y_{n+1} = (1-\mu)^{-1} y_n$. Умова стійкості набуває вигляду

$$|1-\mu| \geq 1 \text{ або } (1-s)^2 + s^2 \geq 1,$$

тобто стійкість досягається для всіх точок комплексної площини за винятком одиничного круга із центром у т. (1, 0) (рис. 12.6).

Означення 12.4. *Множина точок μ, для яких метод стійкий, називається областю стійкості різницевого методу.*

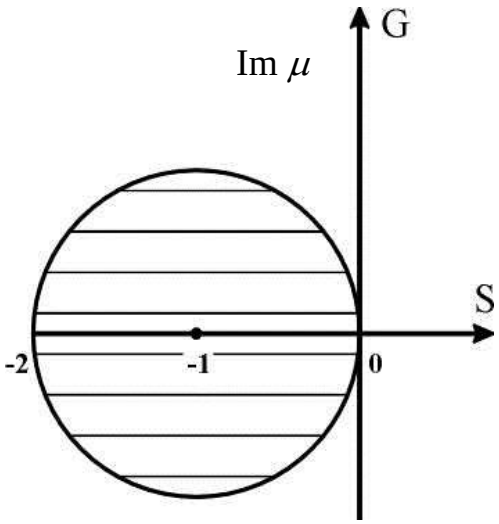


Рис. 12.5. Область стійкості явного методу Ейлера

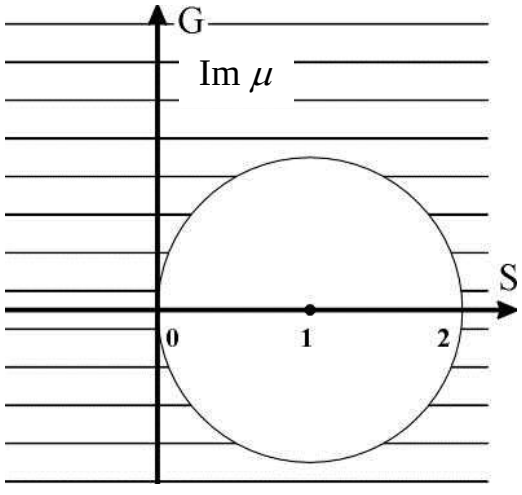


Рис. 12.6. Область стійкості неявного методу Ейлера

О**значення 12.5.** *Різницевий метод абсолютно стійкий, якщо він стійкий для довільного $h > 0$, і умовно стійкий в разі обмеження на крок сітки h.*

ААбсолютно стійким є неявний метод Ейлера, явний метод Ейлера – умовно стійкий. Проілюструємо стійкість методів Ейлера на прикладі числових розв'язків для початкової задачі [9]

$$\dot{u} = -10^4 u + (10^4 - 1)e^{-t}, t \in [0,1], u(0) = 2,$$

точний розв'язок якої $u(t) = e^{-10^4 t} + e^{-t}$. Наближений розв'язок обчи-слюється явним і неявним методами Ейлера відповідно:

$$\bar{y}_{n+1} = (1 - 10^4 h) y_n + 9999 h e^{-nh},$$
$$\tilde{y}_{n+1} = (1 + 10^4 h)^{-1} (y_n + 9999 h e^{-(n+)h}).$$

Крок $h = 10^{-3}$ не задовольняє умову стійкості, для $h = 10^{-4}$ маємо $-\lambda h = 1$, а для $h = 10^{-5}$ виконується нерівність $-\lambda h < 1$. Результати обчислень і точний розв'язок із шістьма правильними цифрами наведені в таблиці 12.14.

Таблиця 12.14

Стійкість методів Ейлера в залежності від кроку сітки

t_n	h	Точний розв'я-зок	Явний метод Ейлера		Неявний метод Ейлера	
			Числовий розв'язок	Відносна похибка, %	Числовий розв'язок	Відносна по-хибка, %
0.001	10^{-3}	0.999046	-8.001000	901.864	-8.001000	901.864
	10^{-4}		-0.998901	0.015	-0.998901	0.015
	10^{-5}		0.999617	0.003	0.999017	0.003
0.1	10^{-3}	0.904837	$2.65 \cdot 10^{95}$	$2.93 \cdot 10^{97}$	0.903933	0.168
	10^{-4}		0.904747	0.010	0.904747	0.068
	10^{-5}		0.904828	0.001	0.904828	0.001
0.5	10^{-3}	0.606531	$-1.66 \cdot 10^{206}$	$4.15 \cdot 10^{207}$	0.604131	0.101
	10^{-4}		0.606470	0.010	0.606570	0.011
	10^{-5}		0.606525	0.001	0.606521	0.001
1	10^{-3}	0.367879	$1.17 \cdot 10^{308}$	$3.48 \cdot 10^{309}$	0.367860	0.012
	10^{-4}		0.367843	0.099	0.367809	0.001
	10^{-5}		0.367876	0.001	0.367875	0.000

Розглянемо метод Рунге–Кутти другого порядку (12.17).

$$y_{n+1} = y_n + h(k_1 + k_2)/2.$$

Для рівняння (12.30) він набуває вигляду

$$y_{n+1} = y_n + 0.5h(\lambda y_n + \lambda(y_n + \lambda h y_n)) = (1 + \mu + 0.5\mu^2) y_n.$$

Умова стійкості $\left|1 + \mu + 0.5\mu^2\right| \le 1$ рівносильна системі нерів-ностей

$$\begin{cases} \mu + 0.5\mu^2 \le 0, \\ 0.5\mu^2 + \mu + 2 \ge 0, \end{cases}$$

розв'язком якої для $\lambda \in R$ є відрізок $-2 \le \mu \le 0$, як і для явного методу Ейлера.

Метод РК3 для рівняння (12.30) набуває вигляду

$$y_{n+1} = (1 + \mu + \frac{1}{2}\mu^2 + \frac{1}{6}\mu^3) y_n .$$

Умова стійкості методу задається системою нерівностей

$$\begin{cases} \mu(\mu^2 + 3\mu + 6) \le 0, \\ \mu^3 + 3\mu^2 + 6\mu + 7 \ge 0. \end{cases}$$

Розв'язком системи є відрізок $-2.51 \le \mu \le 0$, тобто область стійкості S_4 дещо ширша, ніж область S_2 для методу РК2 (рис 12.6).

Для класичного методу РК4

$$y_{n+1} = y_n + h(k_1 + 2k_2 + 2k_3 + k_4)/6$$

область стійкості визначається нерівністю

$$\left| 1 + \mu + \frac{1}{2}\mu^2 + \frac{1}{6}\mu^3 + \frac{1}{24}\mu^4 \right| \le 1.$$

На дійсній осі розв'язком є проміжок $[\mu_0, 0]$, де $\mu_0 \approx -2.78$.

Отже, явні методи Рунге–Кутти умовно стійкі. З ростом порядку апроксимації методу область стійкості розширюється, що проілюстровано на рис. 12.7.

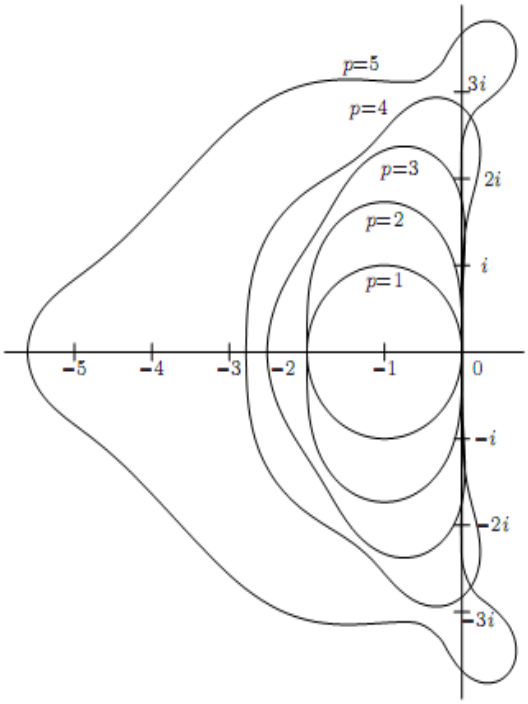


Рис. 12.7. Області стійкості методів Рунге–Кутти порядку 1-5

При розв'язуванні дискретної задачі на комп'ютері до похибки методу додається похибка заокруглення. Розглянемо вплив обчислювальної похибки при реалізації явної схеми Ейлера (12.7), яка в цьому випадку набуде вигляду:

$$y_{n+1} = y_n + hf(t_n, y_n) + \delta_n,\ n = \overline{0, N-1}, \qquad (12.35)$$

$$y_0 = u_0 + \delta_0;\ \ |\delta_n| \le \delta,\ n = \overline{0, N}.$$

Для похибки у вузлі t_{n+1} маємо співвідношення

$$z_{n+1} = z_n + h\psi_n^{(1)} + h\psi_n^{(2)} + \delta_n,\ n = \overline{0, N-1}.$$

Враховуючи, що $|\psi_n^{(1)}| \le Mh^2$, $\max\limits_{t\in[ab]}|\psi_n^{(2)}| \le M_1|z_n|$, де $M_1 = \max\limits_{t\in[a,b]}|\dot{u}(t)|$, $M = 0.5\max\limits_{t\in[a,b]}|\ddot{u}(t)|$, аналогічно як при доведенні теореми 12.1, одержимо $|z_{n+1}| \le (1+M_1h)|z_n| + Mh^2 + |\delta_n| \le$

$$\le (1+M_1h)^{n+1}|z_0| + \sum_{\nu=0}^{n}(Mh^2 + |\delta_\nu|)(1+M_1h)^\nu \le$$

$$\le (1+M_1h)^{n+1}|\delta_0| + (Mh^2 + |\delta|)(M_1h)^{-1}((1+M_1h)^N - 1) \le$$

$$\le e^{(b-a)M_1}|\delta_0| + e^{(b-a)M_1}\left(\frac{M}{M_1}h + \frac{\delta}{M_1h}\right) = e^{(b-a)M_1}\left(\frac{M}{M_1}h + |\delta_0| + \frac{\delta}{M_1h}\right).$$

Якщо крок сітки $h \to 0$, то похибка методу Ейлера (12.35) прямує до нуля (теорема 12.1), а похибка реалізації на комп'ютері зростає. Тому, числовий розв'язок завжди буде відрізнятися від математичного розв'язку. Похибка машинної реалізації дається взнаки для кроків сітки, порівняних із точністю обчислень δ і великому числі кроків $N \approx \delta^{-1}$. Тому при виборі кроку сітки потрібно враховувати як точність числового розв'язку, так і вплив збурень внаслідок обчислень на комп'ютері.

У табл. 12.15 наведено результати обчислення явним методом Ейлера в точці $t = 2$ значення числового розв'язку y_h задачі Коші

$$\dot{u} = 2t^2 + 2u,\ u(0) = 1,$$

та похибки z_h на сітці із кроком $h = 10^{-k}$, $k = \overline{3,8}$.

Таблиця 12.15

h	0.001	0.0001	0.00001	0.000001	0.0000001	0.00000001
$y_h(2)$	75.048088	75.378071	75.395309	75.396881	75.396179	75.386163
$z_h(2)$	0.349137	0.019154	0.001916	0.000344	0.001046	0.011062

Спостерігається зменшення похибки $z_h(2) = y_h(2) - u_h(2)$ при зменшенні кроку h від 10^{-3} до 10^{-6}. Подальше зменшення кроку веде до зростання похибки.

Приклади розв'язання типових задач

Задача 1. Дослідити стійкість явного методу РК4

Розв'язування. Для рівняння $\dot{u} = \lambda u$ метод РК4 набуває вигляду

$$y_{n+1} = q(\mu)\, y_n,$$

де $\mu = \lambda h$, $q(\mu) = 1 + \mu + \frac{1}{2}\mu^2 + \frac{1}{6}\mu^3 + \frac{1}{24}\mu^4$. Умова стійкості випливає з нерівності $|q(\mu)| \le 1$ (рис. 12.) На дійсній осі наближене значення лівої межі інтервалу стійкості $[\mu_0, 0]$ знаходимо, застосувавши метод половинного поділу до рівняння $q(\mu) = 1$ або $((\mu + 4)\mu + 12)\mu + 24 = 0$ на проміжку $[-3, -2]$. На десятому поділі кроку маємо $\mu_0 \approx -2.7832$.

Задача 2. Одним із методів Рунге–Кутти другого порядку записати формулу для обчислення наближеного розв'язку моделі коливання плоского маятника

$$\ddot{u} + \omega^2 \sin u = 0,\ \omega^2 = g/l;\ u(0) = u_0,\ \dot{u}(0) = \dot{u}_0.$$

Розв'язування. Запишемо рівняння у вигляді системи рівнянь

$$\dot{u} = v,\ \dot{v} = -\omega^2 \sin u$$

з початковими умовами $u(0) = u_0$, $v(0) = \dot{u}_0$.

У системі рівнянь (12.18) покладемо $b = 1/4$. Тоді $b_1 = 1 - b = 3/4$, $b_2 = a_{21} = 1/(2c) = 2$ і формула (12.17) набуває вигляду

$$y_{n+1} = y_{n-1} + \frac{h}{4}(3k_1 + k_2).$$

Для системи рівнянь маємо:

$$k_1^{(1)} = v_n,\ k_1^2 = -\omega^2 \sin u_n,$$

$$k_2^1 = v_n + 2hk_1^2,\ k_2^2 = -\omega^2 \sin(u_n + 2hk_1^1).$$

Тоді наближений розв'язок обчислюється за формулами

$$\bar{u}_{n+1} = \bar{u}_n + \frac{h}{4}\left(3k_1^{(1)} + k_2^{(1)}\right),\ \ \bar{u}_0 = u_0;$$

$$\bar{v}_{n+1} = \bar{v}_n + \frac{h}{4}\left(3k_1^{(2)} + k_2^{(2)}\right),\ \ \bar{v}_0 = \dot{u}_0,\ n = 0,1,\ldots.$$

Задача 3. Записати формули методу Рунге–Кутти четвертого порядку для обчислення числового розв'язку моделі хижака та жертви

$$\dot{u} = (a - bv)u,\ \ u(0) = u_0;$$
$$\dot{v} = (-c + du)v,\ \ v(0) = v_0,$$

де $u(t)$ і $v(t)$ – величини популяцій «жертви» і «хижака» відповідно в момент часу $t \in [0,T]$, a,b,c і d – додатні сталі.

Розв'язування. Позначимо наближений розв'язок на сітці з кроком $h = T/N$ через $\bar{u}_n$ і $\bar{v}_n$, $n = \overline{0,n}$. Застосуємо метод Рунге–Кутти порядку 4 з коефіцієнтами з табл. 12.8. Тоді

$$\bar{u}_{n+1} = \bar{u}_n + h\left(k_1^{(1)} + 2k_2^{(1)} + 2k_3^{(1)} + k_4^{(1)}\right)/6,\ \ \bar{u}_0 = u_0;$$

$$\bar{v}_{n+1} = \bar{v}_n + h\left(k_1^{(2)} + 2k_2^{(2)} + 2k_3^{(2)} + k_4^{(2)}\right)/6,\ \ \bar{v}_0 = v_0;$$

$$k_1^{(1)} = (\beta_1 - \gamma_1\bar{v}_n)\bar{u}_n,\ \ k_1^{(2)} = (-\beta_2 + \gamma_2\bar{u}_n)\bar{v}_n;$$

$$k_2^{(1)} = \left(\beta_1 - \gamma_1\left(\bar{v}_n + \tau k_1^{(2)}\right)\right)\left(\bar{u}_n + \tau k_1^{(1)}\right),\ \ \tau = 0.5h;$$

$$k_2^{(2)} k_{2,2} = \left(-\beta_2 + \gamma_2\left(\bar{u}_n + \tau k_1^{(1)}\right)\right)\left(\bar{v}_n + \tau k_1^{(2)}\right);$$

$$k_3^{(1)} = \left(\beta_1 - \gamma_1\left(\bar{v}_n + \tau k_2^{(2)}\right)\right)\left(\bar{u}_n + \tau k_2^{(1)}\right);$$

$$k_3^{(2)} = \left(-\beta_2 + \gamma_2\left(\bar{u}_n + \tau k_2^{(1)}\right)\right)\left(\bar{v}_n + \tau k_2^{(2)}\right);$$

$$k_4^{(1)} = \left(\beta_1 - \gamma_1\left(\bar{v}_n + hk_3^{(2)}\right)\right)\left(\bar{u}_n + hk_3^{(1)}\right),$$

$$k_4^{(2)} = \left(-\beta_2 + \gamma_2\left(\bar{u}_n + hk_3^{(1)}\right)\right)\left(\bar{v}_n + hk_3^{(2)}\right).$$

Зауважимо, що алгоритм можна записати компактніше, ввівши функції $f(h,z) = \beta_1 + \gamma_1 hz$ і $g(h,z) = -\beta_2 + \gamma_2 hz$.

Завдання та запитання для самостійної роботи

1. Наскільки відрізняються порядки локальної та глобальної похибок неявного методу Ейлера? Проілюструвати цей метод.

2. Використавши формули числового інтегрування лівих і правих прямокутників та трапецій, одержати відповідно явний і неявний методи Ейлера та симетричну РС.
3. Який порядок апроксимації симетричної РС? Проілюструвати цю неявну РС.
4. Застосувати ітераційні методи простої ітерації та Ньютона для знаходження розв'язку неявної РС Ейлера та симетричної РС.
5. Записати формули методів Ейлера та РК2 для обчислення наближеного розв'язку таких математичних моделей [11, 42]:

   а) взаємодії популяцій "жертви" і "хижака"

$$\begin{cases} \dot{N}_1 = (a - bN_2)N_1, \\ \dot{N}_2 = (-c + dN_1)N_2; \end{cases}$$

   б) хімічної кінетики Ледивера і Ніколіса

$$\dot{u} = a + u^2 v - (b+1)u, \dot{v} = bu - u^2 v;$$

6. Застосувати метод Рунге–Кутти четвертого порядку для обчислення розв'язку моделі Лоренца[3]

$$\dot{x} = \sigma(y - x),$$
$$\dot{y} = rx - y - xz,$$
$$\dot{z} = bz + xy$$

   на відрізку $[0, 50]$ з початковими умовами $x(0) = 3.05$, $y(0) = 1.58$, $z(0) = 15.62$, кроком сітки $h = 0.01$ та значеннями коефіцієнтів $\sigma = 10$, $r = 28$, $b = 8/3$. Побудувати у фазовому просторі траєкторію динамічної системи. При вказаних параметрах має місце ефект, який називається «дивним атрактором».
7. Методом Рунге–Кутти другого порядку на відрізку $[0, 50]$ з кроком сітки $h = 0.01$ побудувати числовий розв'язок моделі Чуа[4]

$$\dot{x} = \alpha\left(y + d_1 x - d_3 x^3\right),$$
$$\dot{y} = x - y + z, \quad \dot{z} = -\beta y,$$

в якій виникають режими хаотичних коливань, $d_1 = 1/6$, $d_3 = 1/16$, $\beta = 14$. Проаналізувати динаміку системи при значеннях параметра $\alpha = 6.578$, 8.198 і 10.769, які є точками біфуркації.

---

[3] Lorenz Edward N. Deterministic Nonperiodic Flow // *J. Atmos. Sci.*, 1963, **20**. – P. 130–141.

[4] Matsumoto, T. A Chaotic Attractor from Chua's Circuit // IEEE Transactions on Circuits & Systems. – 1984/ – vol.CAS-31, no.12. – P. 1055-1058

8. Методи Рунге–Кутти порядку 1 і 4 побудувати на проміжку $[0,10]$ числовий розв'язок моделі Лотки–Вольтерри з кроком $h=0.01$

$$\dot{u}=u-uv-u/10,\ u(0)=2;$$

$$\dot{v}=-v+uv-v^2/20,\ v(0)=1.$$

9. Записати задачу Коші , розв'язком якої є щільність нормального розподілу

$$u(t)=\frac{1}{2}+\frac{1}{\sqrt{\pi}}\int_0^t e^{-\tau^2/2}d\tau,\quad 0\le t\le 3.$$

Застосувати метод РК4 з кроком $h=0.1$ для знаходження $u(1)$ та $u(3)$. В значеннях $u(1)=0.8413448$ і $u(3)=0.9986501$ всі цифри правильні.

10. Коливання плоского математичного маятника без урахування опору середовища описується рівнянням

$$\frac{d^2\theta}{d^2t}+\frac{g}{l}\sin\theta=0,$$

де l – довжина невагомого стержня, $g\approx 9.8$ м/сек^2. Застосувати методи Рунге–Кутти порядку 2 і 4 для обчислення розв'язку з кроком $h=0.1$ на відрізку часу $[0,3]$, якщо $l=1$, $\theta(0)=0$ і $\dot{\theta}(0)=1$.

11. Дослідити стійкість неявного методу Ракитського

$$y_{n+1}=y_n+hf\,(t_n+\frac{h}{2},\frac{y_n+y_{n+1}}{2}).$$

12. Побудувати всі методи Рунге–Кутти другого порядку з трьома стадіями, коефіцієнти яких задані в таблиці, та дослідити, чи можна побудувати метод Рунге–Кутти третього порядку за таблицею коефіцієнтів такого вигляду

c_2	c_2		
c_3	0	c_3	
	0	0	1

13. Дослідити стійкість неявного методу Ейлера, методу трапецій і методу середньої точки на прикладі задачі Коші [4, с. 143]

$$\frac{du}{dt}=\lambda(t)u,\ u(0)=u_0.$$

14. Серед методів РК2 (табл. 12.2) вибрати метод із найменшими значеннями коефіцієнта ГСП для таких задач:

а) $\dot{u}=u\lambda,\ u(0)=u_0>0$, коли $\lambda>1$ і $\lambda<1$;

б) $\dot{u} = g(u)$, $u(0) = u_0$, де $g(u)$ – задана гладка функція.

15. Звести рівняння $u''' + a(x,u)u'' + b(x,u)u' = c(x,u)$ до системи трьох рівнянь першого порядку і записати явний метод Ейлера і Рунге–Кутти 2-го порядку для побудови числового розв'язку.
16. Розв'язавши систему рівнянь (12.20)–(12.23) знайти коефіцієнти двопараметричної і двох однопараметричних сімей методів Рунге–Кутти третього порядку. Виділити з кожної з сімей по одному набору числових коефіцієнтів.
17. Розв'язавши систему рівнянь (12.24) знайти коефіцієнти двопараметричної і трьох однопараметричних сімей методів Рунге–Кутти четвертого порядку. Виділити з кожної з сімей по одному набору числових коефіцієнтів.
18. Знайти порядок апроксимації методу Гюна

$$y_{n+1} = y_n + \frac{h}{2}\left[f\left(t_n, y_n\right) + f\left(t_{n+1}, y_n + hf\left(t_n, y_n\right)\right)\right]$$

і дослідити його на стійкість.

19. Отримати формулу для s-стадійного явного методу Рунге–Кутти для рівняння $\dot{u} = \lambda u$.
20. Задача Дж. Бутчера: показати, що для довільного методу Рунге–Кутти порядку 5

$$\sum_i b_i \left(\sum_j a_{ij} c_j - c_i^2 / 2\right)^2 = 0.$$

Звідси випливає, що не існує явних методів РК5, в яких усі $b_i > 0$.

21. Для скалярного рівняння

$$\dot{u} = u^\alpha, \quad \alpha > 0$$

записати двостадійний метод Рунге–Кутти і знайти його похибку на кроці.

22. Політ снаряда масою m, зосередженою в одній точці, та з врахування опору повітря, пропорціонального квадрату швидкості снаряда ($\left|\vec{F}\right| = cv^2$), описується системою диференціальних рівнянь

$$m\ddot{x} = -c\sqrt{\dot{x}^2 + \dot{y}^2}\,\dot{x},$$

$$m\ddot{y} = -c\sqrt{\dot{x}^2 + \dot{y}^2}\,\dot{y} - mg,$$

де $x(t)$ і $y(t)$ – проекції на осі Ox і Oy, $0 \le t \le T$. Для самохідної артилерійської установки 2C1 «Гвоздика» маса снаряда $m = \ = 21.7$ кг, $-3^0 \le \alpha \le 70^0$, $v_0 = 690$ м/сек. Скласти програму розрахунку висоти і дальності польоту снаряда, якщо $c = 0.01$, $T = 300$ сек, при $t = 0$

початкові умови: $x(0)=y(0)=0,\quad \dot{x}(0)=\nu_0\cos\alpha,\quad \dot{y}(0)=\nu_0\sin\alpha,$ $\alpha=60^0$.

23. У розкладі похибки $y_n-u_n=c_1h_1+c_2h^2+\cdots$ для модельної задачі $\dot{u}=\lambda u$ з початковою умовою $u(0)=1$ знайти сталу c_1 для $t_n=1$ при апроксимації її різницевою схемою

$$\frac{y_{n-1}-y_n}{h}=\frac{y_{n+1}+y_n}{2},\quad n\geq 0.$$

24. Математична модель Кермана–Маккендрика поширення епідемії має вигляд

$$\dot{S}=-\beta SI,$$
$$\dot{I}=\beta SI-\gamma I,$$

де $S(t)$ – кількість особин, схильних до інфікування, а $I(t)$ – кількість інфікованих. На часовому відрізку $[0,20]$ методом Рунге–Кутти знайти наближений розв'язок із початковими умовами $S(0)=9,\ I(0)=1,$ якщо $\beta=-0.1,\quad \gamma=0.08.$

25. Дослідити стійкість явного методу Ейлера на прикладі розв'язку задачі Коші $0.01\dot{u}+u=1, u(0)=0.$

26. Знайти явним і неявним методами Ейлера значення розв'язку в точці $t_f=1$ початкової задачі

$$\dot{u}_1+u_2=t^2, u_1(0)=1;$$
$$\dot{u}_2-u_2=1, u_2(0)=0.$$

27. Протестувати методи Рунге–Кутти порядку 1–4 для системи диференціальних рівнянь, запропонований у [75],

$$\dot{u}_1=2tu_1u_4, \dot{u}_2=10tu_1^5u_4,$$
$$\dot{u}_3=2tu_4, \dot{u}_4=-2t(u_3-1),$$

розв'язок якої задовольняє початкові умови $u_i(0)=1, i=\overline{1,4}.$ Точний розв'язок задачі:

$u_1(t)=\exp(\sin t^2), u_2(t)=\exp(5\sin t^2),\ \ u_3(t)=\sin t^2+1,\ \ u_4(t)=\cos t^2.$

28. Розглянути застосування явного і неявного методів Ейлера для числового розв'язування задачі Коші для диференціального рівняння $F(t,u,\dot{u})=0$, яке не розв'язане відносно похідної.

Розділ 13. Багатокрокові різницеві методи розв'язування задачі Коші

Багатокрокові різницеві схеми (РС). Побудова та приклади явних і неявних РС Адамса. РС Штермера. Стійкість багатокрокових РС. Умова коренів. Область стійкості. Стійкість РС Адамса.

Література [5, 13, 28, 43, 58, 59, 65, 73, 75, 83, 84]

Електронні джерела [103, 105–108]

13.1. Постановка задачі

В однокрокових методах для знаходження наближеного значення розв'язку задачі Коші

$$\dot{u} = f(t,u),\ t_0 \le t \le t_f, \qquad u(t_0) = u_0, \tag{13.1}$$

у вузлі $t_n = t_{n-1} + h$ використовувалось тільки значення y_{n-1} у вузлі t_{n-1}. Нехай в m $(m \ge 1)$ вузлах $t_{n-m},...,t_{n-1}$ сітки задано значення розв'язку $y_{n-m},...,y_{n-1}$. Задача полягає у знаходженні за цими значеннями наближеного розв'язку у вузлі t_n. Такі методи називаються m-кроковими. Виділимо із цього класу лінійні m-крокові, при $m \ge 2$ – багатокрокові, методи вигляду

$$\frac{1}{h}(a_0 y_n + a_1 y_{n-1} + ... + a_m y_{n-m}) = \sum_{\nu=0}^{m} b_\nu f_{n-\nu},\ a_0 \ne 0, \tag{13.2}$$

$$n = \overline{m,N},\ \ N = (t_f - t_0)/h,$$

де $f_{n-\nu} = f(t_{n-\nu}, y_{n-\nu})$. Коефіцієнти a_ν і b_ν, $\nu = \overline{0,m}$, визначаються таким чином, щоб похибка апроксимації РС (13.2)

$$\psi_n^{(1)} = -\frac{1}{h}\sum_{\nu=0}^{m} a_\nu u_{n-\nu} + \sum_{\nu=0}^{m} b_\nu f(t_{n-\nu}, u_{n-\nu}) = -\frac{1}{h}\sum_{\nu=0}^{m} a_\nu u_{n-\nu} + \sum_{\nu=0}^{m} b_\nu \dot{u}_{n-\nu} \tag{13.3}$$

на точному розв'язку задачі Коші (13.1) мала найвищий порядок. Тобто $\psi_n^{(1)} = O(h^p)$, де $0 < p$ – найбільше число для деякого класу функцій $f(t,u)$.

Якщо $b_0 = 0$, то РС явна, оскільки з (13.2) знаходимо

$$y_n = (-a_1 y_{n-1} - ... - a_m y_{n-m} + \sum_{\nu=1}^{m} b_\nu f_{n-\nu})/a_0.$$

Якщо ж $b_0 \ne 0$, то РС − неявна, тому y_n визначається з рівняння

$$y_n = \frac{hb_0}{a_0} f(t_n, y_n) + F(y_{n-1},...,y_{n-m}),$$

де $F = (-a_1 y_{n-1} - ... - a_m y_{n-m} + h\sum_{\nu=1}^{m} b_\nu f_{n-\nu}) / a_0$.

Історично багатокрокові методи розв'язування задачі Коші передували методам Рунге–Кутти. У 1855 р. Джон Адамс запропонував цей метод для числового інтегрування задачі про форму поверхні краплі рідини [75, с. 323], яка зводилась до системи диференціальних рівнянь вигляду

$$\frac{dr}{ds} = \cos\varphi,\ \frac{dz}{ds} = -\sin\varphi,\ \frac{d\varphi}{ds} = -\frac{\sin\varphi}{r} - 2z,$$

де z – висота краплі, φ – кут між нормаллю до поверхні краплі і віссю z, r – радіус кривизни, s – довжина дуги (рис. 13.1).

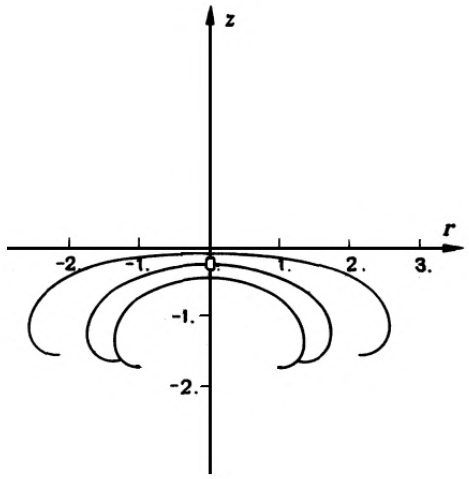


Рис. 13.1. Профілі краплі рідини

Перевагою багатокрокових методів над однокроковими, наприклад методами Рунге–Кутти, є те, що використання відомих значень y_ν в m вузлах підвищує порядок апроксимації. Крім того, при розв'язуванні скалярного рівняння для визначення наступного значення y_n потрібно обчислити тільки одне значення функції f, тоді як в s-стадійному методі Рунге–Кутти на кожному кроці обчислюється $s \geq m$ значень функції, причому у вузлах не більше двох. При знаходженні значення y_n наближеного розв'язку системи d диференціальних рівнянь обчислюється sd значень функцій з правої частини системи рівнянь (13.1).

Недоліком багатокрокових методів є потреба попередньо знаходити "стартові" значення $y_1,...,y_{m-1}$. Це можна зробити, наприклад, методом Рунге–Кутти того ж порядку. Крім того, тут

складніше контролювати точність. Наприклад, зміна кроку при застосуванні правила Рунге потребує перерахунку попередніх значень розв'язку.

13.2. Різницеві схеми Адамса

13.2.1. Побудова РС схем методом невизначених коефіцієнтів. РС Адамса одержуються із лінійної РС (13.2), коли $a_0 = -a_1 = 1$, $a_2 = ... = a_m = 0$. Тоді РС набуває вигляду

$$\frac{y_n - y_{n-1}}{h} = \sum_{\nu=0}^{m} b_\nu f_{n-\nu}\ ,\ n = \overline{m, N}\,. \qquad (13.4)$$

Припустимо, що $u \in C^{p+1}[a,b]$, де $p \geq m$ для явної і $p \geq m+1$ для неявної РС. Для цього достатньо, щоб в області G існували неперервні похідні функції $f(t,u)$ за змінними t і u до порядку p. Покажемо, що вибором коефіцієнтів $b_0, b_1, ..., b_m$, незалежних від правої частини рівняння (13.1) і кроку сітки h, можна досягнути порядку апроксимації, що дорівнює p, тобто

$$\psi_n^{(1)} = -\frac{u_n - u_{n-1}}{h} + \sum_{\nu=0}^{m} b_\nu \dot{u}_{n-\nu} = O(h^p).$$

Згідно з формулою Тейлора маємо

$$u(t_n - h) = u_n - h\dot{u}_n + \frac{h^2 \ddot{u}_n}{2!} - ... + \frac{(-h)^{p+1}}{(p+1)!} u_n^{(p+1)} + o(h^{p+1}),$$

$$\dot{u}_{n-\nu} = \sum_{l=0}^{p} \frac{\nu^l \cdot (-h)^l}{l!} \cdot u_n^{(l+1)} + o(h^{p+1})\,,\ \nu = \overline{1, m}.$$

Для виразу $\sum_{\nu=0}^{m} b_\nu \cdot \dot{u}_{n-\nu}$ одержимо

$$b_0 \cdot \dot{u}_n + \sum_{\nu=1}^{m} b_\nu \cdot \left[\sum_{l=0}^{p} \frac{\nu^l \cdot (-h)^l}{l!} \cdot u_n^{(l+1)} + o(h^p) \right] =$$

$$= b_0 \cdot \dot{u}_n + \sum_{l=0}^{p} \frac{(-h)^p}{l!} \cdot u_n^{(l+1)} \sum_{\nu=1}^{m} b_\nu \cdot \nu^p + o(h^p) =$$

$$\dot{u}_n \cdot \sum_{\nu=0}^{m} b_\nu + \sum_{l=1}^{p-1} \frac{(-h)^l}{l!} \cdot u_n^{(l+1)} \sum_{\nu=1}^{m} b_\nu \cdot \nu^l + o(h^p).$$

Перетворимо вираз для різницевої апроксимації похідної

$$-\frac{1}{h}(u_n-u_{n-1})=-\frac{1}{h}\left[u_n-(u_n-h\cdot\dot u_n+\frac{h^2\cdot\ddot u_n}{2!}-...+\frac{(-h)^{p+1}}{(p+1)!}\cdot u_n^{(p)}+o(h^{p+1}))\right]=$$

$$=-\sum_{l=0}^{p}\frac{(-h)^l}{(l+1)!}\cdot u_n^{(l+1)}+o(h^p)=-\dot u_n-\sum_{l=1}^{p-1}\frac{(-h)^l}{(l+1)!}\cdot u_n^{(l+1)}+o(h^p).$$

Підставивши одержані вирази у співвідношення для похибки апроксимації, матимемо: $\psi_n^{(1)}=(-1+\sum_{\nu=0}^{p}b_\nu)\dot u_n+$

$$+\sum_{l=1}^{p}\frac{(-h)^l}{l!}\cdot u_n^{(l+1)}(-\frac{1}{l+1}+\sum_{\nu=1}^{m}b_\nu\cdot\nu^l)+o(h^p).$$

Прирівнявши коефіцієнти біля $\dot u_n,\ddot u_n,...,u_n^{(p)}$ до нуля, одержимо систему p лінійних рівнянь для знаходження $b_0,b_1,...,b_m$. Для неявної РС $(b_0=0)$ $p=m+1$ і система рівнянь набуває вигляду

$$\begin{aligned}
b_0 \;+\; b_1+\; b_2+...+\;\; b_m&=1,\\
b_1+\;\; 2b_2+...+\;\; mb_m&=\frac{1}{2},\\
&.....\\
b_1+2^m b_2+...+m^m b_m&=\frac{1}{m+1}.
\end{aligned}\qquad(13.5)$$

Оскільки визначник системи дорівнює $\prod_{0\le j<i\le m}(i-j)\neq 0$, то розв'язок СЛАР існує і єдиний. Отже, для фіксованого m неявна РС Адамса, коефіцієнти якої є розв'язком системи (13.5), має порядок $p=m+1$.

ГСП має вигляд

$$\frac{(-h)^{m+1}}{(m+1)!}\cdot u_n^{(m+2)}\left(\sum_{\nu=1}^{m}b_\nu\cdot\nu^{m+1}-\frac{1}{m+2}\right).$$

Для явної РС $(b_0=0)$ відповідна система лінійних рівнянь для визначення коефіцієнтів має порядок $p=m$ і набуває вигляду:

$$\begin{aligned}
b_1+\;\; b_2+...+\;\; b_m&=1,\\
b_1+\;\; 2b_2+...+\;\; mb_m&=\frac{1}{2},\\
&.....\\
b_1+2^{m-1}b_2+...+m^{m-1}b_m&=\frac{1}{m}.
\end{aligned}\qquad(13.6)$$

Для кожного $m \geq 1$ існує єдиний розв'язок системи (13.6), а порядок явної РС Адамса дорівнює m. ГСП похибки апроксимації набуває вигляду

$$\frac{(-h)^m}{m!} \cdot u_n^{(m+1)} \left(\sum_{\nu=1}^{m} b_\nu \cdot \nu^m - \frac{1}{m+1} \right).$$

13.2.2. Приклади різницевих схем Адамса. Якщо $m = 1$, то для неявної РС Адамса порядок апроксимації $p = 2$. Коефіцієнти b_0 і b_1 визначаються із системи рівнянь

$$b_0 + b_1 = 1, \quad b_1 = 1/2 .$$

Отже, $b_0 = b_1 = 1/2$. Відповідна РС називається симетричною, або РС трапецій і набуває вигляду

$$\frac{y_n - y_{n-1}}{h} = \frac{1}{2}(f_n + f_{n-1}), \quad n = \overline{1, N}; \quad y_0 = u_0 . \qquad (13.7)$$

ГСП одержаної РС складає $h^2 u_n''' / 12$.

Табл. 13.1
Неявні РС Адамса

Неявні РС Адамса					
p	b_0	b_1	b_2	b_3	b_4
1	1				
2	$\frac{1}{2}$	$\frac{1}{2}$			
3	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$		
4	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$	
5	$\frac{251}{720}$	$\frac{646}{720}$	$\frac{-264}{720}$	$\frac{-106}{720}$	$\frac{-19}{720}$

Табл. 13.2
Явні РС Адамса

Явні РС Адамса					
p	b_1	b_2	b_3	b_4	b_5
1	1				
2	$\frac{3}{2}$	$-\frac{1}{2}$			
3	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$		
4	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$	
5	$\frac{1901}{720}$	$\frac{-2774}{720}$	$\frac{2616}{720}$	$\frac{-1274}{720}$	$\frac{251}{720}$

Коефіцієнти явної РС другого порядку визначаються із системи рівнянь

$$b_1 + b_2 = 1,$$
$$b_1 + 2b_2 = 1/2.$$

Звідси $b_1 = 3/2$, $b_2 = -1/2$. Одержимо явну РС другого порядку

$$y_n = y_{n-1} + \frac{h}{2}(3f_{n-1} - f_{n-2}), n = \overline{2, N}. \qquad (13.9)$$

ГСП для цієї РС набуває вигляду $-5h^2u_n'''/12$.

Коефіцієнти явних і неявних РС Адамса порядку 1–5 наведені в табл. 13.1 і 13.2.

Порівняння числових розв'язків задачі Коші,

$$\dot{u} = 2u + t^2,\ t \in [0,2];\ u(0) = 1,$$

обчислених за методом Рунге–Кутти порядку 4 (табл. 12.8) та явним методом Адамса того ж порядку зі стартовими значеннями, знайденими МРК4, наведені в табл. 13.3, де $\varepsilon = \max|y_n - u_n|,\ n = \overline{1,20}$.

Табл. 13.3

Числові розв'язки задачі методами Рунге–Кутти та Адамса порядку 4

t_n	$h = 0.1$		$h = 0.01$	
	Метод Рунге-Кутти	Метод Адамса	Метод Рунге-Кутти	Метод Адамса
0.0	1.00000	1.00000	1.00000	1.00000
0.2	1.49773	1.49773	1.49774	1.49774
0,4	2.27829	2.27805	2.27831	2.27831
0.6	3.5203	3.51893	3.52018	3.52018
0.8	5.48944	5.48649	5.48955	5.48955
1.0	8.58338	8.57720	8.58358	8.58358
1.2	13.39439	13.38253	13.39476	13.39736
1.4	20.80631	20.78467	20.80697	20.80697
1.6	32.13764	32.09951	32.13886	32.13839
1.8	49.35539	49.28975	49.35735	49.35734
2.0	75.39393	75.28296	75.39722	75.39721
Похибка ε	0,00331	0.11426	0.000004	0.0000175

Як видно з табл. 13.3 точність обчислення методом Рунге–Кутти на два порядки вища, ніж у методі Адамса.

13.2.3. Збіжність РС Адамса. Рівняння для похибки числового розв'язку $z_n = y_n - u_n$ набуває вигляду

$$\frac{1}{h}(z_n - z_{n-1}) = \psi_n^{(1)} + \psi_n^{(2)},\ n = m, m+1, \ldots,$$

$$\psi_n^{(1)} = -\frac{1}{h}(u_n - u_{n-1}) + \sum_{\nu=0}^{m} b_\nu f(t_{n-\nu}, u_{n-\nu}),$$

$$\psi_n^{(2)} = \sum_{\nu=0}^{m} b_\nu \left(f(t_{n-\nu}, y_{n-\nu}) - f(t_{n-\nu}, u_{n-\nu})\right).$$

Теорема 13.1 [59]. *Нехай* $u \in C^p$, $p \ge m = 1$, $\max\limits_{(t,u)\in G} |f_u(t,u)| \le A$ *і виконується нерівність* $2Ah|b_0| \le 1$. *Тоді для похибки РС Адамса (13.5) справджується оцінка*

$$|z_n| \le c_1 \max_{0\le\nu\le m-1} |z_\nu| + hc_1 \sum_{k=1}^{n-m} |\psi_{n-k}^{(1)}|,$$

де $c_1 = \exp[A(b-a)c_2]$, $c_2 = 2\sum\limits_{\nu=1}^{m} |b_\nu|$. ■

Тож, при $h \to 0$ і $z_\upsilon \to 0$, $\upsilon = \overline{0, m-1}$, похибка $z_n \to 0$ $\forall n = \overline{m, N}$.

13.2.4. Реалізація неявних РС. Знаходження розв'язку різницевого рівняння (13.5) ($b_0 \ne 0$) зводиться до розв'язування нелінійного рівняння. Для цього можна застосувати метод простої ітерації:

$$y_n^{(k+1)} = \frac{hb_0}{a_0} f(t_n, y_n^{(k)}) + F(y_{n-1},..., y_{n-m}),\ i = 0,1,...$$

Якщо в області G похідна $|f_u(t,u)| \le M_1$, то достатня умова збіжності

$$\left|\frac{\partial}{\partial u}\left(\frac{hb_0}{a_0} f(t_n,u) + F(y_{n-1},..., y_{n-m})\right)\right| = h\left|\frac{b_0}{a_0} f_u(t,u)\right| \le hM_1\left|\frac{b_0}{a_0}\right| < 1$$

виконується для досить малого кроку сітки $h \le |a_0|/(M_1|b_0|)$.

За початкове значення $y_n^{(0)}$ можна взяти y_{n-1} або обчислити його за явною багатокроковою схемою. Розглянемо це на прикладі РС Адамса третього порядку. Явна РС називається в цьому випадку *прогнозом*, або *предиктором*. Нехай

$$y_n^{(прог)} = y_{n-1} + \frac{h}{12}(23 f_{n-1} - 16 f_{n-2} + 5 f_{n-3}). \qquad (13.9)$$

Уточнюється y_n за формулою (13.9) неявного методу Адамса третього порядку вигляду

$$y_n^{кор} = y_{n-1} + \frac{h}{12}(5 f(t_n, y_n^{прог}) + 8 f_{n-1} - f_{n-2}). \qquad (13.10)$$

Формула (13.10) називається *коректором*, а схема (13.9), (13.10) − *предиктор-коректором*. Подальше уточнення розв'язку здійснюється ітераційним методом:

$$y_n^{(i+1)} = y_{n-1} + \frac{h}{12}(5 f(t_n, y_n^{(i)}) + 8 f_{n-1} - f_{n-2}),\ i = 0,1,...,\ y_n^{(0)} = y_n^{кор}.$$

На підставі явної та неявної РС одного порядку можна підви-щити порядок РС. Розглянемо це на прикладі РС Адамса

другого порядку. Нехай u_n – точне, а v_n і w_n – наближені значення розв'язку, знайдені за явною і неявною РС Адамса відповідно. Враховуючи вигляд ГСП цих РС, одержимо

$$u_n = v_n - 5h^2 u_n'''/12 + o(h^2),$$
$$u_n = w_n + h^2 u_n'''/12 + o(h^2).$$

Помноживши другу рівність на 5 і додавши до першої, матимемо $u_n = (v_n + 5w_n)/6 + o(h^2)$. Отже, з точністю до величин порядку h^3 на кроці

$$u_n \approx (5w_n + v_n)/6.$$

З явної та неявної РС Адамса третього порядку з точністю до величин четвертого порядку одержимо

$$u_n \approx (251w_n + 19v_n)/270.$$

13.3. Різницеві схеми Штермера

Диференціальне рівняння *n*-го порядку

$$u^{(n)} = f(t, u, \dot{u}, ..., u^{(n-1)})$$

замінами можна звести до системи n рівнянь першого порядку. Нехай $u = v_1$. Тоді маємо систему рівнянь

$$\dot{v}_1 = v_2,\ \dot{v}_2 = v_3, , \ldots,\ \dot{v}_{n-1} = v_n,\ \dot{v}_n = f(t, v_1, v_2, ..., v_n).$$

Але часто зручніше застосувати різницевий метод безпосередньо до рівняння другого або вищого порядку. Розглянемо задачу Коші для диференціального рівняння другого порядку

$$\ddot{u} = f(t,u),\ t_0 \le t \le t_f; \qquad (13.11)$$
$$u(t_0) = u_0,\ \dot{u}(t_0) = \dot{u}_0.$$

Найчастіше для розв'язування цієї задачі використовуються РС Штермера [13, 28, 76]. Цей метод розроблений у 1907 р. для опису руху заряджених частинок у магнітному полі Землі при дослідженні полярного сяйва. Наведемо дві із цих схем відповідно другого і третього порядку:

$$y_n - 2y_{n-1} + y_{n+1} = h^2 f_{n-1},\ n = 2,3,...$$
$$y_n - 2y_{n-1} + y_{n+1} = \frac{h^2}{12}(13 f_{n-1} - f_{n-2} + f_{n-3}),\ n = 3,4,... \qquad (13.12)$$

Для РС Штермера другого порядку потрібно знайти значення y_1 із другим порядком. Для цього введемо фіктивне значення $u_{-1} = u(-h)$. Тоді

$$\dot{u}_0 = \left(u_1 - u_{-1}\right)/\left(2h\right) + O\left(h^2\right),$$
$$u_1 - u_0 + u_{-1} = h^2 f\left(t_0, u_0\right) + O\left(h^3\right).$$

З цих рівностей маємо

$$u_1 \approx u_0 + h\dot{u}_0 + h^2 f\left(t_0, u_0\right)/2. \qquad (13.13)$$

Разом (13.12) і (13.13) утворюють РС Штермера другого порядку.

Проілюструємо методи Штермера й Адамса 2-го порядку на прикладі розв'язування задачі Коші

$$\ddot{u} = u,\ t \in [0;1];\quad u(0) = 0, \dot{u}(0) = 1.$$

У табл. 13.3 показані наближені значення розв'язку в точці t = 1, знайдені з кроками $h = 1/N$. За стартові значення взято значення точного розв'язку $u(t) = \sin t$.

Табл. 13.3

Порівняння методів Штермера та Адамса

Кількість відрізків розбиття	Похибка методу Штермера	Похибка методу Адамса 2-го порядку
5	0.0001385033129889	0.0012902443618832
100	0.0000000340824339	0.000000307094352614001310000
200	0.000000004321346169788853300	0.000000038914655386967922000
500	0.000000000278762346539451760	0.000000002510851859760521200
1000	0.000000000034709346508066119	0.000000000314705705939388740
5000	0.000000000004581002244208320	0.000000000002522759778855743
20000	0.000000000021241564063245733	0.000000000000040190073491430
50000	0.000000000101187946910386020	0.000000000000001221245327087
100000	0.000000000048163029120473766	0.000000000000013544720900426
200000	0.000000000537218936003114320	0.000000000000001554312234475
500000	0.000000000225347074334081300	0.000000000000028199664825478
1000000	0.000000000158495105928579960	0.000000000000018984813721090
2000000	0.000000003994793718042899400	0.000000000000058064664187895
5000000	0.000000184825435622926190000	0.000000000000071831429693247
10000000	0.000003000704598021819400000	0.000000000000030198066269804
50000000	0.001489944381329655300000000	0.000000000000238475905689483

Результати обчислень свідчать, що метод Адамса для даного прикладу стійкіший щодо нагромадження обчислювальної похибки порівняно з методом Штермера.

Крім методів Адамса та Штермера, відомі багатокрокові методи Мілна, Нюстрема, Хемінга [75, 78] та ін. Для розв'язування жорстких систем ЗДР широко застосовуються чисто неявні різницеві схеми [57, 75. 84]. Коефіцієнти деяких із багатокрокових РС наведено в додатку Г.

13.4. Стійкість багатокрокових РС

Дослідження стійкості багатокрокових, як і однокрокових різницевих методів розв'язування задачі Коші, здійснюється на модельному рівнянні

$$\dot{u} = \lambda u,\ \lambda \in C. \tag{13.14}$$

Якщо записати РС (13.2) для рівняння (13.14), тобто коли $f = \lambda u$, то одержимо однорідне різницеве рівняння порядку m

$$\sum_{\nu=0}^{m} (a_\nu - \mu b_\nu) y_{n-\nu} = 0,\ \mu = \lambda h. \tag{13.15}$$

Тому дослідження стійкості РС (13.2) рівносильно дослідженню стійкості різницевого рівняння (13.15), що добре вивчено [58, 59].

Розглянемо лінійне різницеве рівняння порядку m

$$c_0 y_n + c_1 y_{n-1} + \ldots + c_{m-1} y_{n-m+1} + c_m y_{n-m} = 0,\ c_0 \neq 0. \tag{13.16}$$

Розв'язок цього рівняння будується у вигляді $y_n = q^n$, де $q \neq 0$ і підлягає визначенню. Підставивши $y_i = q^i$ в (13.16) і скоротивши на q^{n-m} одержимо для q алгебраїчне рівняння порядку m

$$g(q) := c_0 q^m + c_1 q^{m-1} + \ldots + c_{m-1} q + c_m = 0, \tag{13.17}$$

яке називається характеристичним.

Якщо число q є коренем рівняння (13.17) кратності $r \geq 1$, то різницеве рівняння має частинні розв'язки $y_n = n^j q^n,\ j = \overline{0,1}$.

Означення 13.1. *Різницеве рівняння (13.16) (відповідний різницевий метод) називається стійким за початковими умовами* $y_0, y_1, \ldots, y_{m-1}$, *якщо виконується оцінка*

$$|y_n| \leq B_1 \max_{0 \leq \nu \leq m-1} |y_\nu|,\ n = m, m+1, \ldots, \tag{13.18}$$

де B_1 *– незалежна від* n *стала.*

Отже, стійкість означає рівномірну по n обмеженість розв'язку задачі Коші для різницевого рівняння (13.16).

Умову (13.18) можна інтерпретувати ще так. Внесемо в y_ν похибки ε_ν, $\nu = \overline{0, m-1}$, наслідком яких є похибка ε_n у визначенні y_n при $n \geq m$. У термінах похибок умова (13.18) набуває вигляду:

$$|\varepsilon_n| \leq B_1 \max_{0 \leq \nu \leq m-1} |\varepsilon_\nu|.$$

Виявляється, що стійкість чи нестійкість рівняння (13.16) за початковими даними на модельному рівнянні цілком визначається розміщенням коренів характеристичного рівняння (13.17).

Означення 13.2. *Рівняння (13.16) задовольняє умову коренів, якщо корені характеристичного рівняння (*13.17*) лежать в одиничному крузі* $|q| \le 1$ *комплексної площини, причому на межі круга немає кратних коренів.*

Теорема 13.2 [59]**.** *Умова коренів необхідна і достатня для стійкості різницевого рівняння* (13.16) *за початковими даними.* ■

Пояснити умову коренів можна таким чином. Похибка РС виражається через розв'язки q^ν, і якщо корінь характеристичного рівняння (13.17) $|q| > 1$, то похибка зростатиме зі зростанням n. Якщо ж $|q| = 1$ і корінь має кратність $r > 1$, то похибка також зростатиме із-за наявності розв'язків $nq^n, ..., n^{r-1}q^n$.

13.5.2. Неоднорідне рівняння. Розглянемо задачу Коші (13.1) для неоднорідного рівняння

$$y_n + a_1 y_{n-1} + ... + a_m y_{n-m} = hg_{n-m}, \qquad n = m, m+1, ..., \tag{13.19}$$

де величини $y_0, y_1, ..., y_{m-1}$ і сіткова функція g_{n-m} – задані величини.

Означення 13.3. *Рівняння (13.19) називається стійким, якщо для будь-яких початкових даних* $y_0, y_1, ..., y_{m-1}$ *і для будь-якої правої частини* g_{n-m} *справджується оцінка*

$$|y| \le B_1 \|y^0\|_2 + B_2 \|g\|_3, \tag{13.20}$$

де невід'ємні сталі B_1 *і* B_2 *не залежать від значень* $y^0 := (y_0, ..., y_{m-1})$ *і* $g := (g_0, ..., g_{m-1})$.

Другий доданок в нерівності (13.20) характеризує стійкість за правою частиною. Зазначимо, що нерівність (13.20) задає стійкість як щодо початкових умов. так правої частини.

Теорема 13.3 [59]**.** *Якщо однорідне рівняння (*13.16*) – стійке за початковими умовами (виконується умова коренів), то неоднорідне рівняння (*13.19*) стійке і виконується оцінка*

$$|y_n| \le B_1 \max_{0 \le \nu \le m-1} |y_\nu| + B_2 \sum_{k=0}^{n-m} h|g_k|. \qquad ■$$

Означення 13.4. *Областю стійкості різницевого методу (РС) розв'язування задачі Коші називається множина всіх точок*

μ, $\mu = \lambda h$, комплексної площини, для яких цей метод при застосуванні до модельного рівняння (13.4) стійкий.

Приклад 13.1. Знайдемо область стійкості явної РС Адамса другого порядку $y_n = y_{n-1} + h(3f_{n-1} - f_{n-2})/2$.

Записавши РС для модельного рівняння (13.14). одержимо різницеве рівняння вигляду

$$y_n - \left(1 + \frac{3}{2}\mu\right) y_{n-1} + \frac{1}{2}\mu y_{n-2} = 0.$$

Відповідне характеристичне рівняння

$$g(q) := q^2 - \left(1 + \frac{3}{2}\mu\right) q + \frac{1}{2}\mu = 0$$

має два дійсні корені, оскільки дискримінант

$$D = \left(1 + \frac{3}{2}\mu\right)^2 - 2\mu = \left(\frac{3}{2}\mu + \frac{1}{3}\right)^2 + \frac{8}{9} > 0 \quad \forall \mu \in R.$$

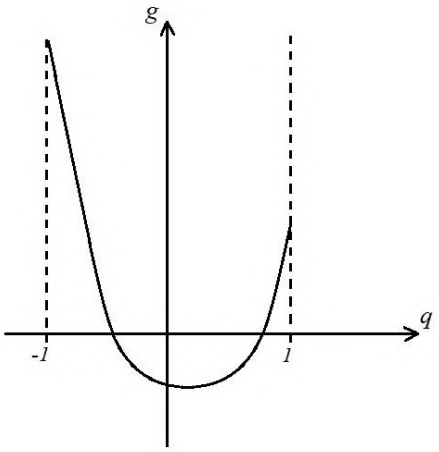


Рис. 13.2

Якщо $\mu > 0$, то більший з коренів характеристичного рівняння $q = \frac{1}{2}\left[1 + \frac{3}{2}\mu + \sqrt{\frac{9}{4}\mu^2 + \mu + 1}\right] > 1$ й умова коренів не виконується.

Для $\mu = 0$ корені 1 і 0. Якщо ж $\mu < 0$, то корені мають різні знаки, і $|q| \le 1$ тоді і тільки тоді, коли $g(-1) \ge 0$ і $g(1) \ge 0$ (рис. 13.2). Звідси маємо систему нерівностей $\mu + 1 \ge 0$, $\mu \le 0$. Тобто на дійсній прямій R стійкість досягається на проміжку [–1,0]. Нагадаєм, що для явних методів Ейлера і Рунге–Кутти другого порядку ширша область стійкості на R і складає [–2,0].

Можна також показати, що всі РС Адамса (явні й неявні) порядку $p \geq 3$ умовно стійкі [75, 84]. Області стійкості явних і неявних РС Адамса порядку 2-6 показані на рис. 13.2 і 13.3 відповідно [75].

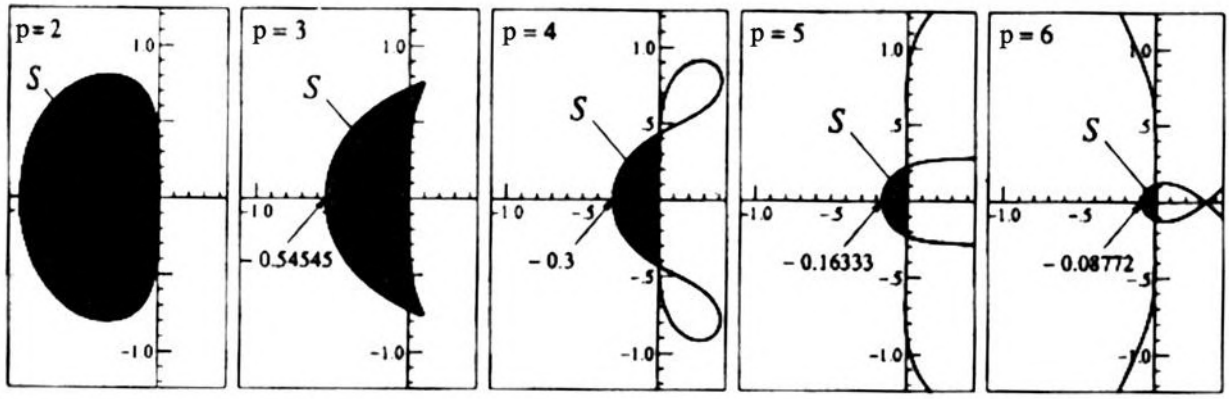


Рис. 13.3. Області стійкості явних різницевих схем Адамса

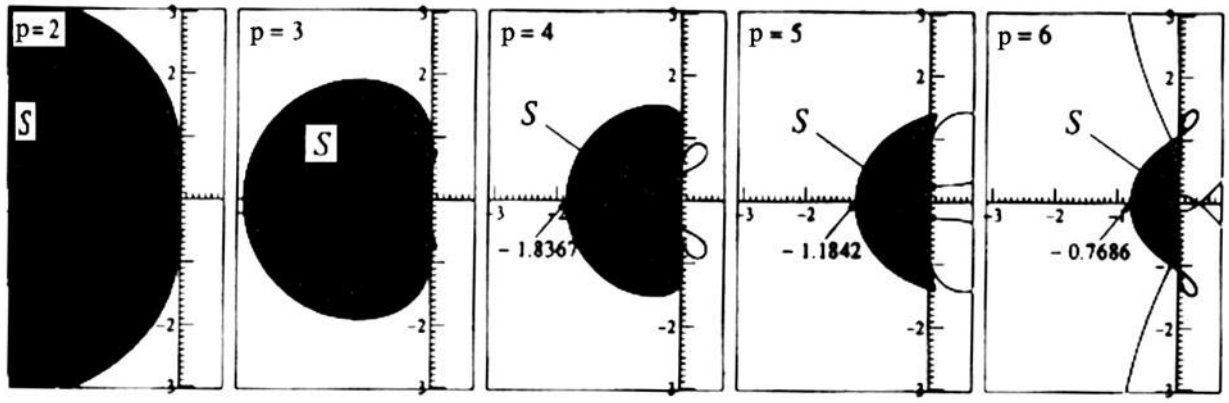


Рис. 13.4. Області стійкості неявних різницевих схем Адамса

Порівняння областей стійкості явних і неявних методів Адамса та явних методів Рунге–Кутти на дійсній осі наведено в табл. 13.4.

Таблиця 13.4

Області стійкості на R багатокрокових методів

Порядок РС	**1**	**2**	**3**	**4**
Явні методи Адамса	$[-2,0]$	$[-1,0]$	$[-6/11,0]$	$[-3/10,0]$
Неявні методи Адамса	$(-\infty,0]$	$(-\infty,0]$	$[-6,0]$	$[-49/11,0]$
Методи Рунге-Кутти	$[-2,0]$	$[-2,0]$	$[-2.51,0]$	$[-2.78,0]$

Нагадаємо, що для явного і неявного методів Ейлера областями стійкості на R є інтервали $[-2, 0]$ і $(-\infty, 0]$ відповідно (рис. 13.3). Як випливає з табл. 13.5, зі зростанням порядку область стійкості методів Адамса звужується, а методів Рунге–

Кутти розширюється. Якщо $p=4$, то область стійкості ширша в методі РК4, ніж в неявному методі Адамса того ж порядку і його довжина в 9 разів перевищує відповідне значення для явного методу Адамса.

Приклади розв'язування типових задач

Задача 1. В явному методі Штермера другого порядку одержати формулу для обчислення y_1 зі збереженням другого порядку.

Розв'язування. Виберемо значення γ так, щоб похибка апроксимації похідної $(y_1-y_0)/h=f(t_0,y_0)+\gamma$ була $O(h^2)$ при $h\to 0$. Нехай $u\in C^3[t_0,t_f]$, розклавши $u_1=u_1(t_0+h)$ за формулою Тейлора в точці t_0 одержимо $\gamma=hf(t_0,u_0)/2$. Звідки маємо $y_1=u_0+ +h\dot{u}_0(1+h)/2$.

Задача 2. Побудувати триточковий метод розв'язування Коші із використанням КФ Сімпсона

Розв'язування. Інтегруючи рівняння (13.1) у межах від t_{n-1} до t_{n+1} одержимо

$$\frac{1}{2h}(u_{n+1}-u_{n-1})=\frac{1}{2h}\int_{t_{n-1}}^{t_{n+1}} f(t,u(t))dt.$$

Застосуємо у правій частині формули Сімпсона

$$\frac{1}{2h}\int_{t_{n-1}}^{t_{n+1}} f(t,u(t))dt=\frac{1}{6}(f_{n+1}+4f_n+f_{n-1})+O(h^2).$$

У підсумку одержимо РС четвертого порядку апроксимації

$$\frac{1}{2h}(y_{n+1}-y_{n-1})=\frac{1}{6}(f_{n+1}+4f_n+f_{n-1})+O(h^2),\ n=1,2,...$$

Задача 3. Побудувати явну і неявну РС Адамса третього порядку та знайти їх ГСП.

Розв'язування. Коефіцієнти неявної РС знаходяться зі СЛАР

$$b_0+b_1+b_2=1,$$
$$b_1+2b_2=1/2,$$
$$b_1+4b_2=1/3,$$

і набувають вигляду $b_0 = 5/12,\ b_1 = 8/12,\quad b_2 = -1/12$ і для знаходження y_n маємо рівняння

$$y_n = y_{n-1} + \frac{h}{12}(5f_n + 8f_{n-1} - f_{n-2}),\quad n = \overline{2, N}.$$

Для коефіцієнтів явного методу Адамса третього порядку $(p = m = 3)$ маємо таку СЛАР

$$b_1 + b_2 + b_3 = 1,$$
$$b_1 + 2b_2 + 3b_3 = 1/2,$$
$$b_1 + 4b_2 + 9b_3 = 1/3,$$

звідки знаходимо $b_1 = 23/12, b_2 = -16/12, b_3 = 5/12$. Наближене значення розв'язку початкової задачі обчислюється за з явною формулою

$$y_n = y_{n-1} + \frac{h}{12}(23f_{n-1} - 16f_{n-2} + 5f_{n-3}),\quad n = \overline{2, N}.$$

Задача 4. Знайти область стійкості симетричної РС.

Розв'язування. Симетрична РС (неявна РС Адамса порядку 2) для рівняння (22) набуває вигляду

$$y_n = y_{n-1} + \mu(y_n + y_{n-1})/2,\ n = 1, 2, \ldots$$

Звідси одержимо

$$y_{n+1} = \left[\left(1 + \frac{\mu}{2}\right)/\left(1 - \frac{\mu}{2}\right)\right] y_n.$$

Відповідне різницеве рівняння: $q = \left(1 + \frac{\mu}{2}\right)/\left(1 - \frac{\mu}{2}\right)$. РС стійка, якщо $|q| \le 1$. У цьому випадку $|y_n| \le |y_{n-1}|$, $n = 1, 2, \ldots$ Отже. для μ маємо нерівність $|1 + 0.5\mu| \le |1 - 0.5\mu|$. Розв'язком цієї нерівності у комплексній області C служить ліва півплощина $s = \mathrm{Re}\,\mu = h\,\mathrm{Re}\,\lambda \le 0$, як і для неявного методу Ейлера.

Задача 5. Дослідити стійкість явної РС Адамса порядку 3.

Розв'язування. Підставивши $f_\nu = \alpha y_\nu$ у РС Адамса, одержимо різницеве рівняння

$$12y_n - (12 + 23\mu)y_{n-1} - 16\mu y_{n-2} + 5\mu y_{n-3} = 0.$$

Відповідне характеристичне рівняння має вигляд

$$g(q) := 12q^3 - (12 + 23\mu)q^2 - 16\mu q + 5\mu = 0.$$

Нехай q_1, q_2 і q_3 – корені характеристичного рівняння. Оскільки $q_1 q_2 q_3 = 5\mu < 0$, то існує хоча б один від'ємний корінь. Якщо $12 + 23\mu > 0$, тобто $\mu > -12/23$, то число знакозмін в системі коефіцієнтів дорівнює 2. На підставі теореми Декарта існує 2 додатні коренів або немає жодного. Оскільки $g(0) = -q_1 q_2 q_3 = -5\mu > 0$ є одна знакозміна в системі коефіцієнтів при заміні q на $-q$, тобто є один від'ємний корінь і два додатних. Це означає, що виконуються умови:

$$q(-1) \le 0 \text{ і } q(1) \ge 0$$

або $\mu \le 0$ і $\mu \ge -6/11 > -12/23$. Таким чином відрізок $[-6/11, 0]$ є областю стійкості РС.

Задача 6. Дослідити стійкість неявної двокрокової ($m = 2$) РС Адамса третього порядку.

Розв'язування. На модельному рівнянні маємо

$$y_n = y_{n-1} + \frac{\mu}{12}(5y_n + 8y_{n-1} - y_{n-2}),\ n = 2,3,...$$

Відповідне характеристичне рівняння

$$g(q) \equiv \left(1 - \frac{5}{12}\mu\right)q^2 - \left(1 + \frac{8}{12}\mu\right)q + \frac{1}{12}\mu = 0,\ \mu \in R.$$

Умова коренів виконується, якщо

$$\begin{cases} \mu \le 0, \\ g(-1) = 2 + \dfrac{1}{3}\mu \ge 0, \\ g(1) = -\mu \ge 0. \end{cases}$$

Звідки одержимо, що $\mu \in [-6, 0]$.

Якщо $\mu > 0$, то для квадратного рівняння $g(q) = 0$ дискримінант $D = 7\mu^2/12 + \mu + 1 > 0$, тому корені рівняння дійсні. Більший корінь $q = (1 + 2\mu/3 + \sqrt{D}) > 1$, тому умова коренів не виконується. Отже, неявна РС Адамса третього порядку вже є умовно стійкою.

Завдання та запитання для самостійної роботи

1. Побудувати явні і неявні РС Адамса третього і четвертого порядку та знайти ГСП на кроці.
2. Побудувати РС Ністрема другого і третього порядку, знайти ГСП і дослідити на стійкість.
3. Чи є серед явних багатокрокових РС вище другого порядку A-стійкі РС? Обґрунтувати відповідь.
4. Дослідити порядок апроксимації та стійкість РС Штермера

$$y_{n+1}^2 - 2y_n + y_{n-1} = \frac{h^2}{12}\left(f\left(t_{n+1}, y_{n+1}\right) + 10 f\left(t_n, y_n\right) + f\left(t_{n-1}, y_{n-1}\right)\right).$$

5. Чи є A-стійкими явна і неявна РС Адамса третього порядку? Якщо ні, то знайти область стійкості на дійсній осі.
6. З'ясувати, чи апроксимують диференціальне рівняння (13.1) такі РС:

   1) $y_n - y_{n-3} = 3hf_{n-1}$;

   2) $y_n - 3y_{n-2} + 2y_{n-1} = 3h\left(f_{n-1} + f_{n-2}\right)$;

   3) $3y_n - 4y_{n-1} + y_{n-2} = 3hf_n$.

7. Для диференціального рівняння $\dot{u} = f\left(t, u\right)$ побудувати РС

$$\left(y_n - y_{n-2}\right) / 2h = a_0 f_n + a_1 f_{n-1} + a_2 f_{n-2}$$

   з найвищим порядком апроксимації.

8. Для задачі $\dot{u} + 4u = \sin t, u(0) = -1/17$ побудувати триточкову РС другого порядку.
9. За допомогою явної РС Адамса другого і третього порядку знайти числовий розв'язок моделі епідемій Кермака–Макендрика

$$\dot{S} = -SI\,,\ \dot{I} = SI - I\,,\ \dot{R} = I$$

10. На відрізку $\left[0, 20\right]$ з кроком $h = 0.01$ і початковими умовами $S\left(0\right) = 4.9$, $I\left(0\right) = 0.1$, $S\left(0\right) = 0$. Дати інтерпретацію результатів.
11. Дослідити на стійкість РС Штермера другого порядку.
12. На дійсній прямій знайти область стійкості РС:

   1) $5y_n + 3y_{n-1} - 3y_{n-2} = 2h(f_n + f_{n-1})$;

   2) $6y_n + 4y_{n-1} - 4y_{n-2} = h(f_n + 5f_{n-1})$;

   3) $3y_n + 5y_{n-1} - 5y_{n-2} = h(3f_n + f_{n-1})$;

   4) $2y_n + 6y_{n-1} - 6y_{n-2} = 2h(2f_n + f_{n-1})$;

   5) $y_n + 6y_{n-1} - 6y_{n-2} = 2h(f_n + 3f_{n-1})$.

13. Довести стійкість РС, які використовуються для розв'язування задачі Коші $\dot{u} + au = \varphi(t), t \geq 0, u(0) = u_0,$ та знайти сталу C, яка входить у визначення стійкості $\|y_h\| \leq C\|f_h\|$, де $\|y_h\| = \max_n |y_n|$, $\|f_h\| = (|y_0|, \max_n |\varphi_n|)$, $y_0 = u_0$, для таких РС:

1) $(y_{n+1} - y_n)/h + ay_n = \varphi_n, n = 0,1,...;$
2) $(y_{n+1} - y_n)/h + ay_{n+1} = \varphi_n, n = 0,1,...;$
3) $(y_{n+1} - y_n)/h + a(y_{n+1} + y_n)/2 = \varphi(t_n + h/2), n = 0,1,...$

14. Дослідити порядок апроксимації та стійкості РС, які апроксимують рівняння (13.1):

1) $\frac{y_n - y_{n-2}}{2h} = f_{n-1}$ (РС Мілна) ;
2) $\frac{3y_n - 4y_{n-1} + y_{n-2}}{2h} = f_n$;
3) $\frac{-3y_{n-2} + 4y_{n-1} - y_n}{2h} = f_{n-2}$.

15. Дослідити РС на стійкість за Дальквістом:

1) $y_{n+1} = 5y_{n-1} - 4y_n + h(4f_n + 2f_{n-1});$
2) (метод Коуелла четвертого порядку)

$$y_{n+1} = y_n + \frac{h}{24}(-f_{n+2} + 13f_{n+1} + 13f_n - f_{n-1}).$$

16. Знайти порядок апроксимації та дослідити точність і стійкість методу Гюна

$$y_{n+1} = y_n + \frac{h}{2}\left[f(t_n, y_n) + f(t_{n+1}, y_n + hf(t_n, y_n)) \right].$$

17. Для наведених різницевих апроксимацій диференціального рівняння $\dot{u} = f(t,u)$ знайти похибку апроксимації та її порядок, дослідити стійкість РС:

1) $(8u_{n+1} - 9u_n + u_{n-2})/6h = (f_{n+1} + 2f_n - f_{n-1})/2;$
2) $(u_{n+1} - u_{n-2})/3h = (f_n + f_{n-2})/2;$
3) $(u_{n+1} + 4u_n - 5u_{n-1})/6h = (2f_n + f_{n-1})/3.$

18. Математична модель росту пухлини має вигляд

$$\frac{dR}{dt} = -\frac{SR}{3} + \frac{2\lambda\nu}{\mu R + \sqrt{(\mu R)^2 + 4\nu}}, R(0) = a,$$

де $R(t)$ – радіус пухлини, яка вважається сферичною, S – смертність клітин в ядрі пухлини, S – рівень поживних речовин, а μ і λ – скалярні параметри. Припускаючи, що $S = 0.90,$ $\nu = 0.05,$ $a = 0.25,$ а

$\mu = \lambda = 1$, знайти радіус пухлина, коли $t = 2$. Застосувати явну РС Адамса порядку 2.

19. Записати неявну РС Адамса порядку 2 для моделі Лоренца (див. задачу 6 розділу 12), скласти програму її реалізації на відрізку $[0, 20]$ з кроком сітки $h = 0.01$ та значеннями коефіцієнтів $\sigma = 10$, $r = 28$, $b = 8/3$ і початковими значеннями $x(0) = 3.05$, $y(0) = 1.58$, $z(0) - 15.32$. Проаналізувати фазовий портрет динамічної системи (рис. 13.5).

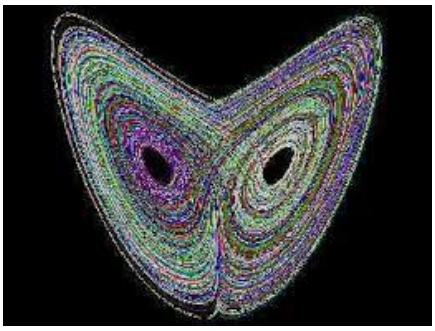


Рис. 13.5

20. Побудувати двокроковий метод вигляду
$$y_n + a_1 y_{n-1} + a_2 y_{n-2} = h(b_0 f_{n-1} + b_1 f_{n-2}),$$
який має порядок 2. Дослідити його стійкість.
21. Знайти головну складову похибки апроксимації та дослідити на стійкість різницевої схеми
$$\frac{y_n - y_{n-2}}{2h} = \frac{f_n + 4 f_{n-1} + f_{n-2}}{6}.$$
22. Дослідити на стійкість РС
$$\theta \frac{y_{n+1} - y_n}{h} + (1-\theta)\frac{y_n - y_{n-1}}{h} = f_n, \quad \theta \in (0,1).$$
23. Показати, що необхідною і достатньою умовою апроксимації диференціального рівняння (13.1) різницевою схемою (13.2) є виконання рівностей
$$\sum_{\nu=0}^{m} a_\nu = 0, \quad \sum_{\nu=0}^{m} \nu a_\nu = -1, \quad \sum_{\nu=0}^{m} b_\nu = 0.$$
24. Математичною моделлю радіоактивного розпаду є рівняння
$$\dot{u} = -ku,$$
де стала k залежить від періоду напіврозпаду Δt і дорівнює $(\ln 2)/\Delta t$. Для елемента радій Δt =1600 років. Узявши за одиницю часу 100 років, обчислити наближений розв'язок на проміжку $[0,100]$ явним і неявним методами Адамса другого порядку з різними кроками та порівняти одержані результати з точним розв'язком $u(t) = \exp(-kt)$, для якого $u(0) = 1$.
25. Проілюструвати метод предиктор-коректор (13.9), (13.10) з кроками $h = 0.1$ і $h = 0.01$ для задачі $\dot{u} = 2t(1+u^2), t \in [0,1]; u(0) = 0$. Точний розв'язок задачі $u(t) = tg(t^2)$.

Розділ 14. Числові методи розв'язування жорстких систем диференціальних рівнянь

Приклади та поняття жорстких систем диференціальних рівнянь. A та $A(\alpha)$-стійкість РС. Чисто неявні різницеві схеми. Неявні методи Рунге–Кутти, їх реалізація та стійкість. Огляд інших методів числового розв'язування жорстких систем.

Література [57, 59, 65, 73, 75, 76, 94]
Електронні джерела [102, 103, 105, 107]

14.1. Приклади жорстких систем

Термін «жорсткі рівняння» (stiff equations) введений у 1952 р. Куртісом і Гіршфельдером[1]. Такі задачі виникають у різноманітних прикладних галузях, зокрема при моделюванні кінетики хімічних реакцій, розрахунку електронних схем й електричних мереж, у біології та економіці, моделюванні імунних процесів та ін. У теорії нелінійних коливань [62] жорсткими є системи ЗДР вигляду

$$\frac{da}{d\tau} = X(\tau, a, \varphi), \ \frac{d\varphi}{d\tau} = \frac{\omega(\tau)}{\varepsilon} + Y(\tau, a, \varphi),$$

де a – вектор амплітудних, а φ – фазових змінних, $\omega(\tau)$ – вектор частот, ε – малий додатний параметр.

Приклад 14.1. Електричне коло на рис. 14.1 описується системою лінійних диференціальних рівнянь

$$\dot{u}_1 = -\frac{R_1 + R_2}{R_1 R_2 C_1} + \frac{u_2}{R_2 C_2} + \frac{e(t)}{R_1 C_1}, \ \dot{u}_2 = \frac{u_1}{R_1 C_1} - \frac{u_2}{R_2 C_2},$$

де $u_1(t), u_2(t)$ – значення напруги. Нехай $R_1 = 1$ кОм, $R_2 = 3$ мОм, $C_1 = C_2 = 1$ мкФ, $e(t) = 1$ В. При таких значеннях параметрів одержується система лінійних диференціальних рівнянь

$$\dot{u}_1 = -1001u_1 + u_2 + 1000,$$
$$\dot{u}_2 = u_1 - u_2.$$

[1] Curtiss C.F., Hirschfelder J.O. Integration of stiff equations // Proc. of the National Academy of Sciences of U.S. – 1952, vol. 38. – P. 235-243.

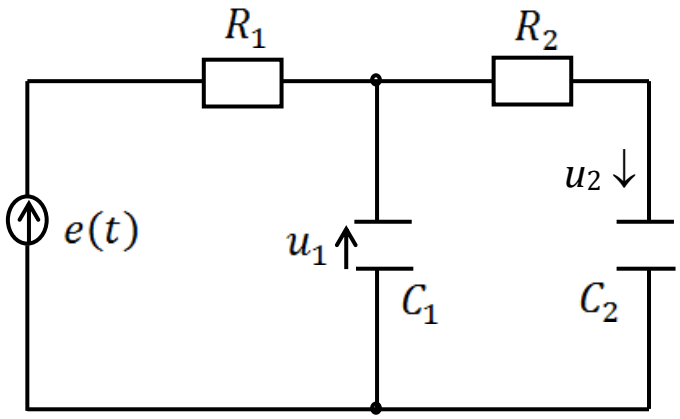


Рис. 14.1. Схема ланцюга

Власні значення $\lambda_1 = -501 + \sqrt{250001} \approx -0.9890$ і $\lambda_2 = -501 - \sqrt{250001} \approx -1001.0010$ матриці однорідної системи є розв'язками алгебраїчного рівняння

$$\begin{vmatrix} -1000-\lambda & 1 \\ 1 & -1-\lambda \end{vmatrix} = 0.$$

Розв'язок системи рівнянь із початковими умовами $u_1(0) = u_2(0) = 0$ має вигляд

$$u_1(t) = \frac{\lambda_2(1-\lambda_1)}{\lambda_1-\lambda_2}e^{\lambda_1 t} + \frac{\lambda_1(1-\lambda_2)}{\lambda_2-\lambda_1}e^{\lambda_2 t} + 1,\ u_2(t) = \frac{\lambda_2}{\lambda_1-\lambda_2}e^{\lambda_1 t} + \frac{\lambda_1}{\lambda_2-\lambda_1}e^{\lambda_2 t} + 1.$$

Отже, власні значення матриці від'ємні, їх відношення $\lambda_2/\lambda_1 \approx 1002.003$ досить велике число, а швидкості спадання компонент розв'язку значно відрізняються.

Приклад 14.2. Відомим прикладом жорсткої системи є нелінійна модель хімічної кінетики [65]

$$\dot{u}_1 = 10^4 u_2 u_3 - 0.04u_1,$$
$$\dot{u}_2 = -0.04u_1 - 10^4 u_2 u_3 - 3\cdot 10^7 u_2^2,$$
$$\dot{u}_3 = 3\cdot 10^7 u_2^2,$$
$$u_1(0) = 1, \quad u_2(0) = u_3(0) = 0.$$

Матриця Якобі системи рівнянь має одне нульове власне значення (оскільки $u_1(t) + u_2(t) + u_3(t) = 1, t \geq 0$) і два дійсних від'ємних власних значення, відношення яких змінюються вздовж інтегральної кривої від $O(10^4)$ до $O(10^5)$.

У літературі є різні означення жорсткості [57, 65, 73, 76]. Це явище полягає в тому, що розв'язок системи змінюється повільно, але існують швидко згасаючі компоненти. Наявність таких

компонент приводить до того, що одержати числовий розв'язок стандартними явними методами числового інтегрування ЗДР не вдається через суттєве зменшення кроку сітки. Розглянемо ще такий модельний приклад

$$\dot{u}_1 = -u_1 + \varepsilon, \quad \dot{u}_2 = -u_2/\varepsilon,$$

де ε – малий параметр, $0 < \varepsilon << 1$, початкові умови $u_1(0) = u_2(0) = 1$. Розв'язком задачі є функції

$$u_1(t) = e^{-t} + \left(e^{-t} - e^{-\frac{t}{\varepsilon}} \right) / (1-\varepsilon), u_2(t) = e^{-\frac{t}{\varepsilon}}.$$

Компонента $u_2(t)$ при малому $\varepsilon > 0$ швидко спадає зі зростанням t, і поведінка розв'язку системи визначається компонентою $u_1(\varepsilon)$. При застосуванні явного методу Ейлера стійкість для u_1 досягається для кроку сітки $h \le 2\varepsilon$. Для першої компененти $h < 2$, що практично не обмежує крок. Зауважимо, що для системи рівнянь $\dot{u} = Au$ з матрицею

$$A = \begin{bmatrix} -1 & \varepsilon^{-1} \\ 0 & -\varepsilon^{-1} \end{bmatrix},$$

власні значення $\lambda_1 = -1, \lambda_2 = -\varepsilon^{-1}$, причому $\lambda_2/\lambda_1 \ge 1/\varepsilon \gg 1$.

14.2. Поняття жорсткої системи

Розглянемо спочатку систему d лінійних рівнянь зі сталими коефіцієнтами

$$\dot{u} = Au, \tag{14.1}$$

де A – матриця порядку d, $u = [u_1, ..., u_d]^T$. Позначимо власні значення матриці A через $\lambda_1, ..., \lambda_d$, вони є коренями алгебраїчного рівняння

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

Означення 14.1. *Система диференціальних рівнянь (14.1) називається жорсткою, якщо:*

1) $\mathrm{Re}(\lambda_\nu) < 0, \nu = \overline{1,d}$;

2) *число жорсткості* $\chi = \dfrac{\max\limits_{\nu}(-\mathrm{Re}\,\lambda_\nu)}{\min\limits_{\nu}(-\mathrm{Re}\,\lambda_\nu)}$ *досить велике.*

Оскільки тут фігурує поняття "досить велике число", то й поняття жорсткості чітко не визначено. Ч. Гір відзначав: "Хоча

загальноприйнято говорити про “жорсткість диференціальних рівнянь”, але рівняння само по собі не є жорстким. Жорсткою може бути конкретна задача Коші для цього рівняння, причому в певних областях, розміри яких залежать від початкових даних і допустимих похибок”.

Якщо система (14.1) неавтономна, тобто $A = A(t)$, то і власні значення $\lambda_\nu = \lambda_\nu(t)$, тому вводиться поняття жорсткості на часовому відрізку $[0,T]$. Нехай

$$\chi(t) = \frac{\max\limits_{\nu} |\operatorname{Re} \lambda_\nu(t)|}{\min\limits_{\nu} |\operatorname{Re} \lambda_\nu(t)|}. \tag{14.2}$$

Означення 14.2. *Система лінійних рівнянь* (14.2) *називається жорсткою на* $[0,T]$, *якщо при всіх* $t \in [0,T]$

$$\operatorname{Re} \lambda_\nu(t) < 0,\ \nu = \overline{1,d},$$

і число $\max\limits_{t\in[0,T]} \chi(t)$ *досить велике.*

Уведемо поняття жорсткості на випадок нелінійної системи

$$\dot{u} = F(t,u),\ t \ge 0, \tag{14.3}$$

де вектор-функція $F(t,u) := (F_1(t,u),...,F_d(t,u))^T$ визначена на множині $[0,T]\times D$, $D \subset R^d$ і двічі неперервно диференційовна за змінними $u_1,...,u_d$. Зафіксуємо деякий розв’язок $u = u_0(t)$ системи (14.3), а різницю $z(t) = u(t) - u_0(t)$ розглядатимемо як збурення розв’язку $u_0(t)$. Тоді $\dot{z}(t) = \dot{u}(t) - \dot{u}_0(t) = F(t,u_0 + z) - F(t,u_0)$. За формулою Тейлора

$$F_\nu(t,u_0+z) - F(t,u_0) = \sum_{k=1}^{d} \frac{\partial F_\nu}{\partial u_k}(t,u_0(t)) z_k(t) + o(\|z\|),\ \nu = \overline{1,d}.$$

Припустимо, що норма збурення $z(t)$ досить мала на $[0,T]$. Відкинувши величину $o(\|z\|)$, для $z(t)$ одержимо лінійну систему рівнянь вигляду

$$\dot{z} = A(t)z,\ A(t) = \frac{\partial F}{\partial u}(t,u_0(t)), \tag{14.4}$$

для опису жорсткості якої можна застосувати означення 14.2.

Означення 14.3. *Система (14.3) називається жорсткою на розв’язку* $u_0(t)$ *при* $t \in [0,T]$, *якщо:*

1) для власних значень матриці A(t)

$$\operatorname{Re} \lambda_\nu(t) < 0,\ \nu = \overline{1,d};\ t \in [0,T];$$

2) число $\max\limits_{t\in[0,T]}\chi(t)$ *досить велике, де* $\chi(t)$ *визначене згідно з формулою* (14.2).

Приклад 14.3. Розглянемо систему рівнянь

$$\dot{u}_1 = u_1u_2 - e^t u_1,$$
$$\dot{u}_2 = u_2^2 - e^t u_2,\ t\in[0,T],$$

яка має розв'язок $u_1 = u_2 = 0$. Матриця Якобі

$$\frac{\partial F}{\partial u} = \begin{bmatrix} u_2 - e^t & u_1 \\ -1 & 2u_2 - e^{-t} \end{bmatrix}.$$

Для розв'язку $u_1 = u_2 = 0$ маємо:

$$A(t) := \frac{\partial F(t,u)}{\partial u}\Big|_{u=0} = \begin{bmatrix} -e^{+t} & 0 \\ -1 & -e^{-t} \end{bmatrix}.$$

Власні значення матриці $A(t)$ такі: $\lambda_1(t) = -e^{-t} < 0$, $\lambda_2(t) = = -e^t < 0$. Число жорсткості $\chi(t) = \lambda_2(t)/\lambda_1(t) = e^{2t}$ і на відрізку $[0, T]$ одержимо $\max\limits_{t\in[0,T]}\chi(t) = e^{2T}$. Отже, на $[0, 1]$ система не жорстка, оскільки $\chi(1) \approx 6.3$. Якщо $T = 2$, то $\chi(2) \approx 38.7$ і число жорсткості зростає. На відрізку $[0, 1]$ система жорстка, оскільки число жорсткості $\chi(2) \approx 4.85 * 10^8 \gg 1.$

14.3. Спеціальні означення стійкості

Усі явні РС умовно стійкі [65]. Неявна РС Ейлера (12.7) – приклад абсолютно стійкої РС з областю стійкості на рис. 12.5, яка містить ліву півплощину $C^- = \{\lambda : \mathrm{Re}\,\lambda < 0\}$.

Розв'язок $u = 0$ диференціального рівняння

$$\dot{u} = \lambda u,\ \lambda \in C, \tag{14.5}$$

асимптотично стійкий, якщо $\lambda \in C^-$ (рис. 14.2), тобто будь-який інший розв'язок $u(t) = u_0 e^{\lambda t} \to 0$ при $t \to \infty$. Доцільно зберегти таку властивість і для РС, які апроксимують рівняння (14.5).

Розглянемо багатокрокову РС (13.2), яка для рівняння (14.5) набуває вигляду

$$\sum_{v=0}^{m}(a_v - \mu b_v)y_{n-v} = 0,\ \mu = \lambda h. \tag{14.6}$$

Якщо класифікувати стійкість за виглядом її області, то по аналогією з областю стійкості рівняння (14.5) приходимо до такого означення.

Означення 14.4. (Далквіст, [90]). *РС, область стійкості якої містить півплощину* $\lambda h \in C^-$, *називається А-стійкою*.

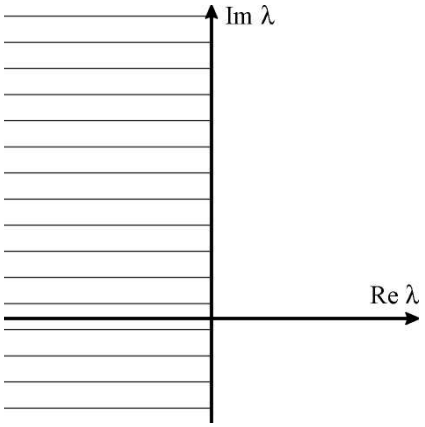


Рис. 14.2. Область стійкості рівняння (14.5)

Прикладами A-стійких РС є неявна схема Ейлера (12.7) і симетрична РС (12.8). Доведено [65, 76], що серед явних ($b_0 = 0$) багатокрокових РС вигляду (14.5) немає A-стійких. Більше того, найвищий порядок неявної A-стійкої різницевої схеми (14.5) дорівнює 2.

Теорема 14.1 (Далквіст, [90, 91]). *Будь-який А-стійкий багатокроковий метод має порядок* $p \le 2$. *Якщо порядок дорівнює 2, то стала похибки задовольняє нерівність* $C \le -1/12$. *Неявний метод Адамса порядку 2 (правило трапецій) – єдиний А-стійкий метод другого порядку із такою сталою в похибці.* ■

Оскільки клас A-стійких РС досить вузький, то введено поняття стійкості з необмеженою, але вужчою за C^- (рис. 14.3), областю стійкості.

Означення 14.6. (О.Б. Відлунд[2]). *РС називається* $A(\alpha)$*-стійкою, якщо сектор* $S_\alpha = \{\mu : |\arg(-\mu)| < \alpha,\ 0 < \alpha \le \frac{\pi}{2}\}$ *міститься в її області стійкості.*

[2] Widlund O.B. A note on unconditionally stable linear multistep methods // BIT. – 1967, vol. 7. – P. 65-70.

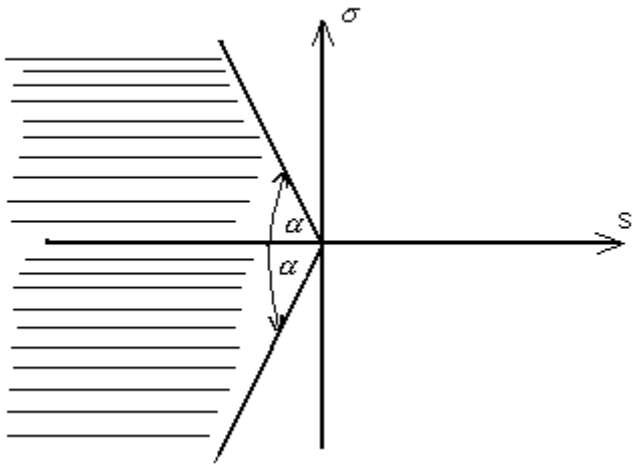


Рис. 14.3. Область $A(\alpha)$-стійкості

Якщо $\alpha = \frac{\pi}{2}$, то РС A-стійка. До класу $A(\alpha)$-стійких належать чисто неявні різницеві схеми (ЧНРС), уведені Кюртісом і Хіршфельдером у 1952 р.). Серед лінійних явних ($b_0 = 0$) багатокрокових методів вигляду (14.5) не існує жодного $A(\alpha)$-стійкого методу. Справді, із відповідного (14.6) характеристичного рівняння одержимо

$$\mu = \frac{a_0 q^m + a_1 q^{m-1} + \ldots + a_m}{b_1 q^{m-1} + b_2 q^{m-2} + \ldots + b_m}.$$

Права частина рівності зростає лінійно зі зростанням $|q|$, якщо $b_1 \neq 0$ або степінь $|q|$ вищий, якщо $b_1 = 0$. Тому для досить великого $|\mu|$ знайдеться таке q, що $\operatorname{Re} q < 0$ і $|q| > 1$. Отже, умова коренів не виконується і РС A-стійкою не буде. Подібно доводиться, що не існує явних $A(\alpha)$-стійких РС.

Крім A і $A(\alpha)$, для неявних РС введено ще поняття L, G і S-стійкості [65] та ін. Зокрема, область стійкості *жорстко стійкої* неявної РС показана на рис. 14.4, де області S_1 і S_2 визначаються умовами [94]:

$$S_1 = \{\mu : -a \le \operatorname{Re}\mu \le b, |\operatorname{Im}\mu| \le d\},\ S_2 = \{\mu : \operatorname{Re}\mu < -a\}.$$

Якщо метод жорстко стійкий, то він $A(\alpha)$-стійкий, але не навпаки.

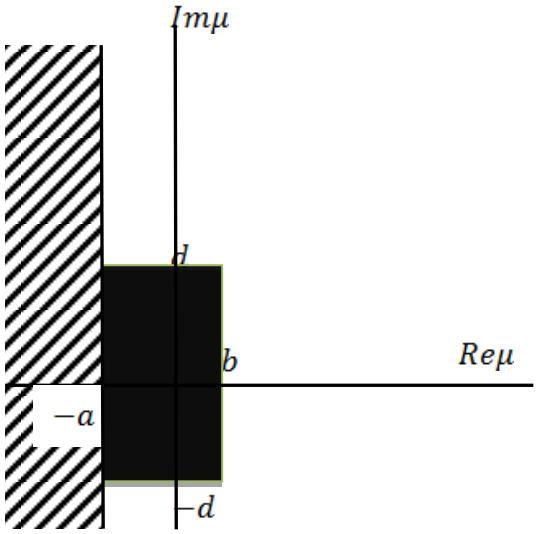


Рис. 14.4. Область жорстко стійкої неявної РС

14.4. Чисто неявні різницеві схеми

Серед багатокрокових методів, які мають ширшу область стійкості, ніж методи Адамса, для числового розв'язування жорстких систем застосовуються ЧНРС[3] [59]. Для диференціальних рівнянь

$$\dot{u} = f(t, u) \tag{14.7}$$

ЧНРС набувають вигляду

$$\frac{1}{h}(a_0 y_n + a_1 y_{n-1} + \ldots + a_m y_{n-m}) = f(t_n, y_n), \quad a_0 \neq 0,\ n = m, m+1, \ldots \tag{14.8}$$

На відміну від РС (13.2), назва "ЧНРС" пояснюється тим, що в правій частині (14.8) міститься тільки значення $f(t_n, y_n)$ з невідомим значенням y_n. РС (14.8) зручно записати ще в такому вигляді

$$y_n = \sum_{\nu=1}^{m} \alpha_\nu y_{n-\nu} + h\beta f_n, \tag{14.9}$$

де $\alpha_\nu = -a_\nu / a_0, \quad \beta = 1/a_0.$

[3] Такі схеми називають ще формулами диференціювання назад (ФДН- або BDF-методами), наприклад у [75] . У даній роботу використовуємо назву ЧНРС [59].

Сім'я $A(\alpha)-$стійких ЧНРС (14.8) називається *РС Гіра* [57, 656]. Такі РС мають порядок 1–6, коефіцієнти наведені в додатку Г, табл. Г4. ЧНРС для $m \geq 7$, уже не будуть навіть $A(0)$-стійкими.

Теорема 14.2 ([75], теорема 3.4). *Чисто неявна m-крокова РС (14.8) стійка при $m \leq 6$ і нестійка при $m \geq 7$.* ■

Коефіцієнти a_ν в (14.8) знаходяться з умови досягнення найвищого порядку апроксимації РС. Нехай $u \in C^{m+1}[a,b]$. Підставивши $y_\nu = z_\nu + u_\nu$, де $z_\nu -$ локальна похибка РС, одержимо

$$\frac{1}{h}\sum_{\nu=0}^{m} a_\nu z_{n-\nu} = \psi_n^{(1)} + \psi_n^{(2)},$$

де $\psi_n^{(1)} = -\frac{1}{h}\sum_{\nu=0}^{m} a_\nu u_{n-\nu} + f(t_n, y_n)$, $\psi_n^{(2)} = f(t_n, u_n + z_n) - f(t_n, u_n)$.

Підставивши розклад $u_{n-\nu}$ у точці t_n за формулою Тейлора у вираз $\psi_n^{(1)}$ для похибки апроксимації, одержимо $\psi_n^{(1)} =$

$$-\frac{1}{h}\left(a_0 u_n + \sum_{\nu=1}^{m} a_\nu \sum_{l=0}^{p} \frac{u_n^{(l)}}{l!}\nu^l (-h)^l + \sum_{\nu=1}^{m} a_\nu \frac{\nu^{p+1}(-h)^p}{(p+1)!} u^{(p+1)}(t_n - \theta_\nu \nu h) - \right) =$$

$$-f(t_n, u_n) = -\frac{u_n}{h}\sum_{\nu=0}^{m} a_\nu + \sum_{l=1}^{p} \frac{u_n^{(l)}}{l!}(-h)^{l-1} \sum_{\nu=1}^{m} a_\nu \nu^l + \dot{u}_n + O(h^p).$$

РС (14.7) містить $m+1$ коефіцієнтів $a_0, a_1, ..., a_m$. Для їх визначення складемо систему з $m+1$ рівнянь. Тоді $p = m$ і $\psi_n^{(1)} = O(h^m)$, тобто різницева схема (14.8) матиме порядок m. Одержимо систему $m+1$ рівнянь

$$\sum_{\nu=0}^{m} a_\nu = 0,\ \sum_{\nu=1}^{m} a_\nu \nu + 1 = 0,\ \sum_{\nu=1}^{m} a_\nu \nu^l = 0,\quad l = \overline{2,m},$$

яку запишемо у вигляді

$$\begin{aligned} &a_0 + a_1 + a_2 + ... + a_m = 0 \\ &a_1 + 2a_2 + ... + m a_m = -1 \\ &a_1 + 4a_2 + ... + m^2 a_m = 0 \\ &..... \\ &a_1 + 2^m a_2 + ... + m^m a_m = 0. \end{aligned} \qquad (14.9)$$

Визначником системи є визначник типу Вандермонда. Отже, розв'язок системи (14.9) існує і єдиний. Тобто доведена наступна теорема.

Теорема 14.3. *Нехай розв'язок диференціального рівняння* (14.7) $u \in C^{m+1}[a,b]$. *Тоді чисто неявна РС* (14.8)*, коефіцієнти якої визначаються із системи* (14.10), *має порядок* $p = m$.

Якщо $p \geq m+1$, *то з (31) і виразу для* $\psi_n^{(1)}$ *одержимо значення головної складової похибки апроксимації*

$$\frac{(-h)^m}{(m+1)!}\left(\sum_{\nu=1}^{m} a_\nu \nu^{m+1}\right)\dot{u}_n^{(m+1)}. \qquad \blacksquare$$

Розглянемо приклади ЧНРС. Для $m=1$ маємо $a_0 = -a_1 = -1$, тобто неявну РС Ейлера. Якщо $m=2$, то з (14.9) одержимо систему лінійних рівнянь

$$a_1 + 2a_2 = -1,$$
$$a_1 + 4a_2 = 0,$$
$$a_0 = -(a_1 + a_2).$$

Звідси знаходимо $a_2 = \frac{1}{2}$, $a_1 = -2$, $a_0 = \frac{3}{2}$ і маємо РС другого порядку

$$\frac{3}{2}y_n - 2y_{n-1} + \frac{1}{2}y_{n-2} = hf(t_n, y_n),\ n = 2,3,... \qquad (14.10)$$

Різницеві схеми 3-го і 4-го порядків набувають відповідно вигляду

$$\frac{11}{6}y_n - 3y_{n-1} + \frac{3}{2}y_{n-2} - \frac{1}{3}y_{n-3} = hf(t_n, y_n),\ n = 3,4,...$$

$$\frac{25}{12}y_n - 4y_{n-1} + 3y_{n-2} - \frac{4}{3}y_{n-3} + \frac{1}{4}y_{n-4} = hf(t_n, y_n),\ n = 4,5,...,$$

або у вигляді (14.9)

$$y_n = \frac{4}{3}y_{n-1} - \frac{1}{3}y_{n-2} + \frac{2}{3}hf(t_n, y_n),\ n = 2,3,...;$$

$$y_n = \frac{18}{11}y_{n-3} - \frac{9}{11}y_{n-2} + \frac{2}{11}y_{n-1} + \frac{6}{11}hf(t_n, y_n),\ n = 3,4,...$$

У табл. 14.1 наведено абсолютні похибки обчислення розв'язку задачі Коші

$$\dot{u}_1 = -10u_1,\ u_1(0) = 1,$$
$$\dot{u}_2 = -0.2u_2,\ u_2(0) = 1,$$

методами другого порядку: Адамса, явним Рунге–Кутти і чисто неявним методом. Систему можна вважати жорсткою, оскільки

$(-\lambda_1)(-\lambda_2) = 50$. Ліпші результати спостерігаються для неявного методу Адамса і ЧНРС порядку 2.

Таблиця 14.1.

t_n	Явний метод Адамса		Неявний метод Адамса		Явний метод Рунге–Кутти		Чисто неявний метод	
	1	2	1	2	1	2	1	2
1	0,088926	0,000025	0,000444	0,000005	0,000932	0,000011	0,000017	0,000019
2	0,088081	0,000043	0,000000	0,000008	0,000001	0,000018	0,000000	0,000033

14.5. Неявні методи Рунге–Кутти

14.5.1. НМРК нижчих порядків. Методи порядку 1 і 2 просто одержується при застосуванні КФ. Із КФ правих прямокутників випливає неявний метод Ейлера (12.7). На підставі КФ центральних прямокутників одержується неявний метод порядку 2, який називається *правилом середньої точки*

$$k_1 = f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1), \qquad y_{n+1} = y_n + hk_1. \tag{14.11}$$

Урахувавши, що $k_1 = (y_{n+1} - y_n)/h$, одержимо

$$y_{n+1} = y_n + hf\left(t_n + \frac{h}{2}, \frac{1}{2}(y_n + y_{n+1})\right).$$

Із КФ трапецій маємо симетричну РС, тобто неявну РС Адамса порядку 2

$$y_{n+1} = y_n + \frac{h}{2}\left[f(t_n, y_n) + f(t_{n+1}, y_{n+1})\right]. \tag{14.12}$$

Згідно з КФ Радо,

$$u_{n+1} - u_n = \int_{t_n}^{t_n+h} f(t, u(t_1))dt \approx$$

$$\approx \frac{h}{4}\left(f(t_n, y_n) + 3f(t_n + \frac{2}{3}h, u(t_n + \frac{2}{3}h)\right).$$

Наблизивши $y(t_n + \frac{2}{3}h)$ з допомогою інтерполяційного многочлена Ерміта, побудованого за значеннями u_n, $\dot{u}_n$ і u_{n+1},

$$u(t_n + \frac{2}{3}h) \approx (5u_n + 4u_{n+1} + 2hf(t_n, u_n)/9,$$

одержимо *неявний метод Хаммера і Холінгсуорта*

$$k_1 = f(t_n, y_n), \quad k_2 = f\left(t_n + \frac{2}{3}h, y_n + \frac{h}{3}(k_1 + k_2)\right),$$

$$y_{n+1} = y_n + \frac{h}{4}(k_1 + 3k_2).$$

Як видно з одержаних формул, для знаходження розв'язку потрібно розв'язати нелінійне рівняння.

14.5.2. Загальна схема НМРК. Метод Рунге–Кутти з s стадіями задається формулами вигляду

$$y_{n+1} = y_n + h\sum_{i=1}^{s} b_i k_i(h),$$

$$k_i = f\left(t_n + c_i h, y_n + h\sum_{j=1}^{s} a_{ij} k_i(h)\right), \ i = \overline{1,s}. \qquad (14.13)$$

Якщо $a_{ij} = 0$ при $i \le j$, то маємо явні методи. Якщо ж $a_{ij} = 0$ при $i > j$, і хоча б одне $a_{ij} \neq 0$, то напівнеявні НМРК. Коефіцієнти Бутчера НМРК наведена в табл.14.2.

Таблиця 14.2

Коефіцієнти неявних методів Рунге–Кутти

c_1	a_{11}	...	a_{1s}
c_2	a_{21}	...	a_{2s}
…	…	…	…
c_s	a_{s1}	...	a_{ss}
	b_1	…	b_s

Коефіцієнти у формулі (14.13), як правило, задовольняють умови [75] $c_i = \sum_{j=1}^{s} a_{ij}, \quad i = \overline{1,s}.$

Теорема 14.4 (Бутчер, [75]). *Нехай f – неперервна функція і задовольняє умову Ліпшіця за змінною u зі сталою L у деякому околі початкових умов. Якщо виконується нерівність*

$$h < 1/L \max_i \sum_j |a_{ij}|, \qquad (14.14)$$

то існує єдиний розв'язок системи рівнянь (14.13), який можна знайти ітераційним методом. ■

Зауважимо, що для жорстких систем стала $L \gg 1$, тому умова (14.14) накладає сильне обмеження на крок сітки. Деталі реалізації НМРК наведено в [76, розділ IV.8].

Скориставшись формулою Тейлора для функції $f(t-\tau, u_{n+1})$ у точці (t_{n+1}, u_{n+1}), можна одержати формули НМРК, аналогічні явним методам [59]. Метод першого порядку (неявний метод Ейлера) і другого порядку

$$y_{n+1} = y_n + h(k_1 + 3k_2),$$
$$k_1 = f(t_{n+1,} y_{n+1}), k_2 = f(t_{n,} y_{n+1} - hk_1),$$

які є A-стійкими, методи порядку 3 і 4 – $A(\alpha)$-стійкі.

Отже, у НМРК знаходження y_{n+1} зводиться до розв'язання s нелінійних рівнянь відносно $k_1,...,k_s$. Для системи d рівнянь одержимо систему sd у загальному випадку нелінійних рівнянь.

14.5.3. Методи Кунцмана–Бутчера. Неявні методи найчастіше будуються з використанням КФ високого порядку. Методи Кунцмана–Бутчера ґрунтуються на КФ Ґауса. Значеннння коефіцієнтів c_i є нулями полінома Ґауса–Лежандра (10.23), «зсунутого» на відрізок [0,1].

Теорема 14.5 (Бутчер, [76]). *S-стадійний метод Кунцмана–Бутчера має порядок* $2s$ *і* A*-стійкий.* ■

Таблиця 14.3

НМРК порядку 4

$\frac{1}{2} - \frac{\sqrt{3}}{6}$	$\frac{1}{4}$	$\frac{1}{4} - \frac{\sqrt{3}}{6}$
$\frac{1}{2} + \frac{\sqrt{3}}{6}$	$\frac{1}{4} - \frac{\sqrt{3}}{6}$	$\frac{1}{4}$
	$\frac{1}{2}$	$\frac{1}{2}$

Розглянемо приклади. Нехай $s=1$, тоді $c_1 = 1/2 = a_{11}$, $b_1 = 1$. Одержимо неявний метод *середньої точки* (14.11).

Для $s=2$ НМРК порядку 4 задається таблицею Бутчера [76]. Розглянемо приклади. Нехай $s=1$, тоді $c_1 = 1/2 = a_{11}$, $b_1 = 1$. Одержимо неявний метод *середньої точки* (14.11).

Для $s=2$ НМРК порядку 4 задається таблицею Бутчера [76].

14.5.4. Методи Радо. Ще один клас неявних методів одержується з КФ Радо (R. Radau), які характеризуються тим, що коефіцієнт $c_1 = 0$ або $c_s = 1$ (один із вузлів попадає на кінець інтервалу інтегрування). Порядок таких методів дорівнює $2s-1$. Приклади A-стійких методів Радо порядку 3
наведено в таблиці 14.4.

Таблиці 14.4
Методи Радо порядку 3

0	$\frac{1}{4}$	$-\frac{1}{4}$
$\frac{2}{3}$	$\frac{1}{4}$	$\frac{5}{12}$
	$\frac{1}{4}$	$\frac{3}{4}$

$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{3}{4}$	$\frac{1}{4}$
	$\frac{1}{4}$	$\frac{3}{4}$

14.5.4. Методи Лобатто. Якщо вузлами є кінці відрізка, то одержимо неявні методи Лобатто порядку $2s-2$, які є A-стійкими. Для першої з таблиць $y_{n+1} = y_n + (k_1 + 3k_2)/4,$ де

$$k_1 = f(t_n, y_n),\ k_2 = f(t_n + h, y_n + h(k_1 + k_2)/2).$$

Таблиці 14.5
Методи Лобатто порядку 2

0	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{4}$	$\frac{3}{4}$

0	$\frac{1}{2}$	0
1	$\frac{1}{2}$	0
	$\frac{1}{2}$	$\frac{1}{2}$

14.5.3. Діагонально неявні методи Рунге–Кутти. У загальному випадку в таблиці коефіцієнтів НМРК досить мало нулів, тому на кожному кроці потрібно розв'язувати систему sd нелінійних рівнянь, що є технічно складною задачею. Побудова таких методів, за висловлюванням Р. Александера[4], "…заводить у нелінійні

[4] Alexander R. Diagonally implicit Runge-Kutta methods for stiff ODE. // SIAM J. Numer. Nnal. – 1977, v.14. – P. 1006-1021.

алгебраїчні джунглі, цивілізацію і порядок в які були внесені в піонерській праці Дж. Бутчера". У діагонально неявних методах Рунге–Кутти значення $k_1, k_2, ..., k_s$ знаходяться із ситеми рівнянь меншого порядку, ніж у загальному випадку. Такими є *діагонально неявні методи Рунге–Кутти* [21, 65, 76]. S-стадійна схема діагонального НМРК показана в табл. 14.6. Коефіцієнти такого A-стійкого методу третього порядку, побудованого Ньорсетом (Nørsett S.P.), наведені в табл. 14.7. При інших значеннях параметра γ метод має другий порядок. Тристадійний A-стійкий метод Нюрсета порядку 4 наведено в табл. Б9 додатку Б. Щоб обчислити

$$y_{n+1} = y_n + (k_1 + k_2)/2,$$

потрібно розв'язати рівняння

$$k_1 = f(t_n + \gamma h, y_n + \gamma h k_1)$$

і знайти k_1, і ще одне рівняння

$$k_2 = f(t_n + \gamma h, y_n + (1-2\gamma)hk_1 + \gamma h k_2).$$

Таблиця 14.6

Коефіцієнти діагонально неявних методів Рунге–Кутти

c_1	γ		
c_2	a_{21}	γ	
…	…	…	
c_s	a_{s1}	...	γ
	b_1	b_2	b_s

Таблиця 14.7

Метод Ньорсета третього порядку $\gamma = \frac{1}{2} \pm \frac{\sqrt{3}}{6}$

γ	γ	0
γ	$1-2\gamma$	γ
	1/2	1/2

Приклад L-стійкого діагонально неявного методу порядку 4 наведено в додатку 6. Оцінка похибки

$$y_1 - \hat{y}_1 = \sum_{\nu=1}^{5} (b_\nu - \hat{b}_\nu) k_\nu = (-6k_1 - 27k_2 + 25k_3 + 8k_5)/32.$$

14.6. Огляд інших методів розв'язування жорстких систем

14.6.1. Однокрокові ітераційні методи Розенброка. Перевагою таких неявних методів є те, що на кожному кроці замість нелінійної системи розв'язується система лінійних рівнянь. Розглянемо автономну систему ЗДР

$$\dot{u} = f(u), \quad u := (u_1, \ldots, u_d)^T.$$

Якщо система неавтономна, тобто набуває вигляду $\dot{u} = f(t,u)$, то введемо змінну $u_{d+1} = t$ і долучимо ще одне рівняння $\dot{u}_{n+1} = 1$.

Метод Розенбрака, порядок точності якого 3, має вигляд [34]

$$\begin{aligned} y_{n+1} &= y_n + b_1 k_1 + b_2 k, \\ (I - d_1 h J(y_n))k_1 &= hf(y_n), \\ (I - hd_2 J(y_n + c_1 k_1))k_2 &= hf(y_n + c_1 k_1), \end{aligned} \tag{14.15}$$

де $J(y_n) := \dfrac{\partial f}{\partial u}(y_n)$ – матриця Якобі. Параметри $b_1, b_2, c_1, c_2, d_1, d_2$ вибираються так, щоб одержати максимально високий порядок точності і набувають значень:

$$a_1 = 1 + \frac{\sqrt{6}}{6} \approx 1.408248, \quad a_2 = 1 - \frac{\sqrt{6}}{6} \approx 5.917517,$$

$$b_1 = -0.413154, \quad b_2 = 1.413154,$$

$$c_1 = d_1 = \frac{-6 - \sqrt{6} + \sqrt{58 + 20\sqrt{6}}}{6 + 2\sqrt{6}} \approx 0.173787.$$

У цьому методі матриця Якобі обчислюється тільки один раз. Очевидна подібність формул (14.15) з явним методом Рунге – Кутти, але з тією різницею, що k_1 і k_2 є розв'язками СЛАР.

14.6.2. Явні РС Федули [73]. Крім лінійних багатокрокових методів розв'язування жорстких задач, останнім часом розроблені явні нелінійні методи, область стійкості, близьку до неявних методів. Одним із підходів до побудови таких методів є апроксимація розв'язку на проміжку $[t_n, t_{n+1}]$ вигляду

$$u_k(t) = A / \left(1 + \sum_{i=1}^{k} a_i t^i\right),$$

де A і a_i – деякі коефіцієнти. Для $k = 1$ маємо гіперболічну апроксимацію. Відповідний явний метод першого порядку запропонований Федулою

$$y_{n+1} = \frac{y_n^2}{y_n - hf_n},\ n = 1,...$$

Область стійкості цього методу така ж, як і неявного методу Ейлера (рис. 12.).

14.6.3. **Неявні методи Ракитського** [73]. Такі методи не вимагають багаторазового обчислення функції f. Метод другого порядку збігається з неявним методом Адамса такого ж порядку, а метод четвертого порядку має вигляд

$$y_{n+1} = y_n + \frac{h}{6}\left[f_{n+1} + f_n + f\left(t_n + \frac{h}{2}, \frac{u_{n+1} + u_n}{2} \right) - \frac{h}{8}(f_{n+1} - f_n) \right].$$

14.6.4. Явний метод Глинського. Це приклад явного нелінійного методу другого порядку. Якщо $y_{n-1} \neq 0$, то

$$y_n = y_n / \left[1 - \frac{h}{2y_{n-1}}\left(3f_{n-1} - f_{n-2}\right) + \left(\frac{h}{2y_{n-1}}\left(3f_{n-1} - f_{n-2}\right) \right)^2 \right], \quad (14.16)$$

інакше $y_n = h\left(3f_{n-1} - f_{n-2}\right)/2$.

У пакеті Mathematica [106] для розв’язування жорстких задач можна використати стандартний оператор NDSolve. Налаштування на використання неявних методів Брайтона з автоматичною зміною порядку методу, від першого до п’ятого, і кроком сітки здійснюється за допомогою опції Method $\rightarrow$ BDF. У разі використання стандартного оператора NDSolve передбачено можливість автоматичного переходу від явних методів Адамса, порядок яких від першого до дванадцятого, до неявних методів Брайтона до п’ятого порядку і навпаки. Для цього відповідна програма аналізує жорсткість задачі, а потім на основі оцінки жорсткості використовується явний чи неявний метод. Значення кроку і порядок методу змінюється, виходячи з умови забезпечення потрібної точності. Результати обчислень зі змінними кроком і порядком використовуються для побудови інтерполяційного полінома Лагранджа або Ерміта (оператор InterpolatingFuntion) для побудови відповідних таблиць значень розв’язку та їх графіків.

Коефіцієнти неявних методів Рунге–Кутти наведені в додатку Б, ЧНРС – у додатку Г.

Приклади розв'язування типових задач

Задача 1. На модельному рівняння (14.5) дослідити стійкість РС Гіра

$$3y_n - 4y_{n-1} + y_{n-2} = 2hf_n \qquad (14.15)$$

та побудувати область стійкості.

Розв'язування. Для рівняння (14.5) РС набуває вигляду

$$\left(1-\frac{2}{3}\mu\right)y_n - \frac{4}{3}y_{n-1} + \frac{1}{3}y_{n-2} = 0, \mu = \lambda h.$$

З характеристичного рівняння $\left(1-\frac{2}{3}\mu\right)y^2 - \frac{4}{3}y - \frac{1}{3} = 0$ маємо

$$\mu = \frac{3}{2} - \frac{2}{q} + \frac{1}{2q^2}.$$

Покажемо, що границя області стійкості – замкнена лінія, а область стійкості – зовнішність цієї лінії. На границі області $|q| = 1$, тому $q = e^{-it}, t \in [0, 2\pi)$, тому

$$\mu(t) = \frac{3}{2} - 2e^{it} + \frac{1}{2}e^{2it}, \quad t \in [0, 2\pi). \qquad (14.16)$$

Оскільки $\mu(0) = \mu(2\pi) = 0,$ то границя області Γ – замкнена лінія. Згідно з формулою Ейлера, $e^{it} = \cos t + i\sin t$

$$\mu(t) = \frac{3}{2} - 2\cos t + \frac{1}{2}\cos 2t + \left(\frac{1}{2}\sin 2t - 2\sin t\right)i.$$

Якщо замінити t на $2\pi - t$, то $\mu(2\pi - t) - \overline{\mu}(t)$, де $\overline{\mu} = s - i\sigma$. Отже, лінія Γ симетрична відносно дійсної осі, $\mu(\pi) = 4$.

Найбільшого значення уявна частина $\sigma(t) = \sin 2t - 2\sin t$ досягає, при $t = t_1$, коли $\cos t_1 = 1 - \sqrt{3} \approx -0.7,$ а $\sin t_1 = \sqrt{2\sqrt{3} - 1} \approx$ ≈ 1.2. Тоді $s(t_1) = 4(\sqrt{3} - 1) \approx 2.9$. Візьмемо $\mu = 1$. Тоді корінь характеристичного рівняння $q = 2 + \sqrt{5} > 1$. Отже, областю стійкості є $C \setminus G,$ G – область у правій напівплощині, обмежена лінією (14.16), тому РС (14.15) є A-стійкою.

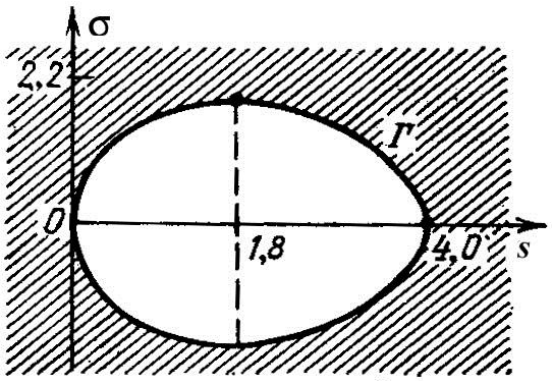


Рис. 14.6. Область стійкості ЧНРС порядку 2

Задача 2. Для ЧНРС третього порядку побудувати область жорсткої стійкості.

Розв'язування. Границя області стійкості досягається при $|q|=1$, коли $q=e^{-it}$, $t\in[0,2\pi)$, і набуває вигляду

$$\mu(t)=\frac{11}{6}-3e^{it}+\frac{3}{2}e^{2it}-\frac{1}{3}e^{3it}=S(t)+i\sigma(t),$$

$$S(t)=\frac{11}{6}-3\cos t+\frac{3}{2}\cos 2t-\frac{1}{3}sos3t,$$

$$\sigma(t)=-3\sin t+\frac{3}{2}\sin 2t-\frac{1}{3}\sin 3t.$$

Найменше значення функції $\sigma(t)$ досягається при $\cos t=1/12$ і дорівнює $-1/12$. Відповідне значення уявної частини $\sigma(t)$ при цьому $d=3\sqrt{3}/4.$ Отже, $a=1/12,\quad b=0,\quad d=3\sqrt{3}/4$ (рис. 14.7).

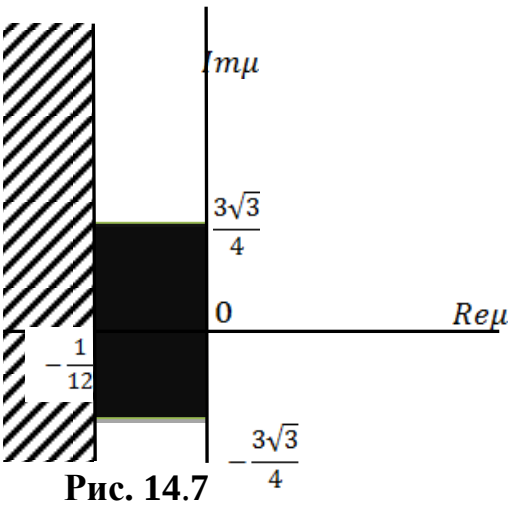


Рис. 14.7

Завдання та запитання для самостійної роботи

1. Довести, що серед явних лінійних багатокрокових РС немає $A(\alpha)$-стійких.
2. Показати, що ЧНРС порядку 3 є $A(\alpha)$-стійкою та знайти область стійкості, побудувавши границю цієї області.

$$\dot{u}_1 = 2000u_1 + 1000u_2 + 1,$$
$$\dot{u}_2 = u_1 + u_2.$$

3. Дослідити стійкість неявних методів Рунге-Кутти порядку 1 – 4. Показати, що методи порядку 1 – 2 є A-стійкими, а 3 і 4 – $A(\alpha)$-стійкими. Побудувати області стійкості.
4. Одержати умови $A(\alpha)$ - стійкості ЧНРС четвертого порядку точності.
5. Для яких *a, b* і *c* різницева схема

$$\frac{1}{h}(y_n + ay_{n-1} - ay_{n-3} - y_{n-4}) = bf_{n-1} + cf_{n-2} + bf_{n-3}$$

має найвищий порядок апроксимації? Чи виконана умова α-стійкості?

6. Знайти значення коефіцієнтів у різницевих схемах Гіра третього, четвертого і п’ятого порядку для сталого кроку сітки.
7. Записати формули неявних методів Ейлера та НРК2 для обчислення наближеного розв’язку математичної моделі Зімана роботи серця

$$\varepsilon\dot{x} = x - x^3 - y,\ \dot{y} = x - x_0.\ \sqrt{3}x_0 > 3$$

8. Довести, що явний нелінійний метод Федули A – стійкий.
9. Довести стійкість явного методу Глинського (14.16) на модельному рівнянні, перевіривши, що для двох сусідніх ітерацій виконується умова $K_n(h\lambda)K_{n-1}(h\lambda) < 1.$
10. Застосувати метод симетричну РС для побудови з кроком $h = 0.1$ числового розв’язку рівняння Лейна-Емдена

$$\ddot{u} + \frac{2}{t}\dot{u} + u^n = 0, \quad t \in (0,8],$$

з початковими умовами $u(0) = 1, \dot{u}(0) = 0.$ Це рівняння виникає в астрофізиці при вивченні розподілу температури в межах однієї зірки. Зокрема, задача має точний розв’язок $u(t) = (\sin t)/t$ для $n = 1$ і $u(t) = (1 + t^2/3)^{-1/2}$ для $n = 5.$

11. Для автономного рівняння $\dot{u} = f(u)$ показати, що явний метод Рунге-Кутти з таблицею коефіцієнтів 12.8 має четвертий порядок.

12. Для крайової задачі

$$\varepsilon u'' - u' = -1, \quad 0 < x < 1, \; u(0) = 1, \quad u(1) = 3$$

знайти числовий розв'язок, застосувавши неявні РС Ейлера та Адамса порядку 2 та чисто неявну РС другого порядку. Для $N =$ 10 і 20 порівняти одержані числові розв'язки з різними кроками, а також з точним розв'язком $u^*(x) = 1 + x + (e^{1/\varepsilon} - 1)^{-1}(e^{x/\varepsilon} - 1)$, коли $\varepsilon = 0.1$ і 0.01.

13. Довести, що метод першого порядку Федули A-стійкий.

14. Розглянути застосування явного і неявного методу Ейлера, неявних методів Адамса та Ґіра другого порядку при числовому інтегруванні задачі Коші

$$\dot{u}_1 = -au_2, \quad u_1(0) = 1,$$
$$\dot{u}_2 = au_1 - u_2, \quad u_2(0) = 1$$

на відрізку [0,1]. Точний розв'язок

$$u_1(t) = e^{-\frac{t}{2}}\left[b^{-1}(1-2a)\sin\frac{bt}{2} + \cos\frac{bt}{2} \right],$$

$$u_2(t) = e^{-\frac{t}{2}}\left[b^{-1}(2a-1)\sin\frac{bt}{2} + \cos\frac{bt}{2} \right],$$

де $b = \sqrt{4a^2 - 1}$. Проаналізувати випадки $a = 1$; 10; 100 і 1000 для кроків $h = 10^{-5}, 10^{-4}$ і 10^{-3}. Для $a = 1000$ задача жорсткоосцилююча.

15. Проаналізувати похибки обчислення розв'язку задачі Коші

$$\dot{u} = 998u + 1998v, \quad u(0) = 1;$$
$$\dot{v} = -999u - 1999v, \quad v(0) = 1$$

на відрізку [0,1] явним і неявним методами Ейлера з кроками 0.1, 0.001 і 0.001. Порівняти числові розв'язки зі значеннями розв'язку задачі

$$u(t) = 4e^{-t} - 3e^{-1000t}, \; v(t) = 2e^{-t} + 3e^{-1000t}.$$

16. Проінтегрувати рівняння[5]

$$\dot{u} = -50(4 - \cos t)$$

з початковою умовою $u(0) = 0$ явним і неявним методом Ейлера на відрізку [0,1] із різними кроками. Знайти числовий розв'язок «назад» неявним методом Ейлера, починаючи зі значення $u(1.5) = 0$ з кроками 0.5 і 0.25, та зробити висновки щодо похибки між числовим і точним розв'язками.

[5] На цьому прикладі Кюртіс і Хіршфельдер (1952 р.) пояснювали властивість жорсткості.

Розділ 15. Числові методи розв'язування двоточкових крайових задач для ЗДР

Зведення крайової задачі до задачі Коші методом уточнення початкових умов. Елементи теорії лінійних різницевих схем (РС)**:** похибка апроксимації, стійкість, збіжність, теорема Лакса. Різницева апроксимація лінійної двоточкової крайової задачі: похибка апроксимації, існування і єдиність розв'язку РС, стійкість і збіжність. РС для нелінійної двоточкової крайової задачі: похибка апроксимації, збіжність і побудова розв'язку. Інтегро-інтерполяційний метод побудови РС.

Література [3, 13, 15, 28, 59, 65, 73, 75, 83] Електронні джерела [105–108]

15.1. Постановка крайових задач

При розв'язуванні задачі Коші умови задаються в одній точці. Для диференціального рівняння другого порядку

$$u'' = f(x,u,u'),\ a < x < b, \tag{15.1}$$

початкові умови набувають вигляду: $u(x_0) = u_0,\ u'(x_0) = u'_0$, де $a \le x_0 \le b$. У двоточкових крайових задачах значення розв'язку або його похідних, або функцій від них задаються у двох точках, що обмежують інтервал знаходження розв'язку. Наприклад, умови

$$u(a) = \mu_0,\ u(b) = \mu_1 \tag{15.2}$$

є найпростішими крайовими умовами (1-го роду). Умови вигляду

$$u'(a) = \nu_0,\ u'(b) = \nu_1. \tag{15.3}$$

називаються крайовими умовами 2-го роду. До 3-го роду належать крайові умови

$$\alpha_0 u(a) + \beta_0 u'(a) = \gamma_0,\ \alpha_1 u(b) + \beta_1 u'(b) = \gamma_1, \tag{15.4}$$
$$|\alpha_i| + |\beta_i| \ne 0,\ i = 0,1.$$

Прикладом нелінійних крайових умов для рівняння (15.1) є умови вигляду

$$f(u(a),u'(a)) = 0,\ \ g(u(b),u'(b)) = 0, \tag{15.5}$$

де f і g – задані функції.

У математичній моделі коливань плоского маятника

$$\frac{d^2u}{dt^2} + \omega^2 \sin u = 0,\ 0 < t < T,$$

умовами $u(0)=u_0, u(T)=u_1$ задається початкове і кінцеве його відхилення, а умовами

$$\dot{u}(0)=\nu_0, \quad \dot{u}(T)=\nu_1$$

– швидкості в ці моменти часу. Такого вигляду умови появляються при визначенні положення матеріальної точки, рух якої згідно з другим законом Ньютона $F=ma$, описується диференціальним рівнянням вигляду $\ddot{u}=f(t,u,\dot{u})$.

Статичний прогин навантаженого пружного бруска описується диференціальним рівнянням четвертого порядку

$$u^{(4)}=f(x), a<x<b.$$

Якщо брусок у точках x_i лежить як на опорах, то умови можуть задаватися більше, ніж у двох точках, наприклад,

$$u(x_i)=0, i=\overline{1,4}, \ a<x_1<x_2<x_3<x_4<b.$$

В моделі Лотки–Вольтерри

$$\dot{u}=(a-bv)u, \dot{v}=(-cu+dv)v,$$

крайовими умовами можуть задаватися величини жертви і хижака, наприклад, $u(0)=\mu_0, \ v(T)=\mu_1$ або деяка їх залежність:

$$u(0)+v(0)=\gamma_0, u(T)-v(T)=\gamma_1.$$

Умови можуть задаватись на всьому відрізку $[a,b]$, наприклад у формі нормування у статистичній фізиці

$$\int_a^b u^2(x)dx=1.$$

Умови існування та єдиності розв'язку крайових задач значно складніші, порівняно із початковими задачами. Наприклад, задача

$$u''=1, \ 0\le x\le\pi; \ u(0)-u(\pi)=1, \ u'(0)-u'(\pi)=0,$$

не має розв'язку. Для рівняння

$$u''+u=0$$

з такими ж лінійними крайовими умовами існує єдиний розв'язок $u(x)=(\cos x)/2$. Для періодичної двоточкової крайової задачі

$$u''+u=0, \ u(0)=u(\pi)=0,$$

існує безліч розв'язків $u(x)=C\sin x, \ C\in R$.

15.2. Розв'язування крайової задачі методом уточнення початкових умов[1]

15.2.1. Нелінійна крайова задача. Розглянемо спочатку нелінійну крайову задачу (15.1), (15.5). Задамо деяке початкове значення $u(a)=\eta$. Тоді з першої з умов (15.5), як рівняння відносно u', знайдемо $u'(a,\eta)=\xi(\eta)$. Розв'яжемо початкову задачу

$$u''=f(x,u,u'),\ \ a\le x\le b;\ u(a)=\eta,\ \ u'(a)=\xi(\eta), \tag{15.6}$$

наприклад, методом Рунге-Кутти. Нев'язка

$$\delta(\eta)=g\big(y(b,\eta),y'(b,\eta)\big) \tag{15.7}$$

характеризує, наскільки точно задовольняється друга крайова умова (15.5). Якщо $|\delta(\mu)|\le\varepsilon$, де ε характеризує точність нев'язки, тоді задача (15.1), (15.5) наближено розв'язана, інакше потрібно уточнити значення η. Зауважимо, що нам відомо тільки наближене значення розв'язку $y(b,\eta)$ та його похідної $y'(b,\eta)$, а не функціональна залежність від η. Тому рівність (15.7) – це не рівняння, з якого можна знайти η.

Значення η можна уточнити за формулою січних

$$\eta_{i+1}=\eta_i-\frac{(\eta_i-\eta_{i-1})\delta(\eta_i)}{\delta(\eta_i)-\delta(\eta_{i-1})},\ \ i=1,2,..., \tag{15.8}$$

якщо розв'язано задачу Коші (15.6) для двох значень η_1 і η_2, $\eta_1\ne\eta_2$. Якщо $\delta(\eta_i)\delta(\eta_{i-1})<0$, то (15.8) – формула методу лінійної інтерполяції. У цьому випадку можна застосувати також метод половинного поділу відрізка.

У разі досягнення потрібної точності нев'язки (15.7) наближений розв'язок крайової задачі одержується як розв'язок задачі Коші з уточненими початковими умовами.

15.2.2. Лінійна крайова задача. Розглянемо лінійне диференціальне рівняння

$$u''+p(x)u'+q(x)u=r(x),\ a<x<b \tag{15.9}$$

із лінійними крайовими умовами (15.4). Лінійна структура задачі дозволяє звести її до двох задач Коші.

Справді, якщо знайти розв'язок $u_0(x)$ однорідного рівняння $(r=0)$ з початковими умовами $u_0(0)=\xi,\ \ u'(0)=(\gamma_0-\alpha_0\xi)/\beta_0$,

[1] Цей метод ще має назву "метод стрільби" [18]

(якщо $\beta_0 \neq 0$, то задамо $u'(a) = \eta_1$ а $u(a) = \gamma_0 / \alpha_0$) і розв'язок $u_1(x)$ неоднорідного рівняння з нульовими початковими умовами $u_1(0) = u_1'(0) = 0,$ то функція $u(x) = Cu_0(x) + u_1(x), C \in R$ задовольняє рівняння (15.9) і першу з умов (15.4). Із другої з умов (15.4) знаходимо значення

$$C = \frac{\gamma_1 - \alpha_1 u(b) - \beta_1 u'(b)}{\alpha_1 u_0(b) + \beta_1 u_0'(b)},$$

якщо знаменник відмітний від нуля. Якщо знаменник у виразі для C дорівнює нулю, то потрібно задати інше значення ξ.

Метод уточнення початкових даних успішно застосовується, коли розв'язок задачі Коші не дуже сильно залежить від збурень початкових умов. Розглянемо це на прикладі крайової задачі

$$u'' - a^2 u = 0,\ 0 < x < 1,\ a > 0;\ u(0) = \mu_0, u(1) = \mu_1. \qquad (15.10)$$

Точний розв'язок задачі (15.10)

$$u(x; \mu_0, \mu_1) = \frac{e^{-ax} - e^{ax-2a}}{1 - e^{-2a}} \mu_0 + \frac{e^{-a+ax} - e^{-a-ax}}{1 - e^{-2a}} \mu_1.$$

Зауважимо, що коефіцієнти при μ_0, μ_1 невід'ємні і не перевищують 1 при $x \in [0,1]$. Оскільки $1 - x \geq 0$, то для першого із них

$$0 \leq \frac{e^{-ax}(1 - e^{2ax-2a})}{1 - e^{-2a}} \leq \frac{1 - e^{-2a(1-x)}}{1 - e^{-2a}} \leq 1.$$

Для збурених крайових умов $u(0) = \mu_0 + \delta_0$, $u(1) = \mu_1 + \delta_1$, $|\delta_i| \leq \delta, i = 0,1$ Тоді $|u(x, \mu_0 + \delta_0, \mu_1 + \delta_1) - u(x, \mu_0, \mu_1)| \leq$

$$|u(x, \mu_0 + \delta_0, \mu_1 + \delta_1) - u(x, \mu_0, \mu_1)| \leq \frac{e^{-ax} - e^{ax-2a}}{1 - e^{-2a}} |\delta_0| +$$

$$+ \frac{e^{-a+a} - e^{-a-ax}}{1 - e^{-2a}} |\delta_1| \leq |\delta_0| + |\delta_1| \leq 2\delta.$$

Отже, малим збуренням крайових умов відповідає мале збурення розв'язку крайової задачі.

У методі уточнення початкових даних припускаємо $u(0) = \mu_0$, $u'(0) = \nu.$ Тоді

$$u(x, \nu) = \frac{a\mu_0 + \nu}{2a} e^{ax} + \frac{a\mu_0 - \nu}{2a} e^{-ax}.$$

Якщо в початкову умову внести похибку ε, то одержимо розв'язок вигляду

$$u(1,\nu+\varepsilon)-u(1,\nu)=\frac{\varepsilon}{2a}e^{a}-\frac{\varepsilon}{2a}e^{-a}=\frac{\varepsilon}{2a}(e^{a}-e^{-a}).$$

Навіть для невеликих ε похибка $\left|u(1,\nu+\varepsilon)-u(1,\nu)\right|$ може набувати досить великих значень. Наприклад, для $\varepsilon=10^{-6}$ і $a=15$ похибка розв'язку в точці $x=1$ складає 0.1, що в 100000 разів перевищує похибку в початковій умові. Отже, крайова задача може бути стійкою, задача Коші – ні. У цьому випадку розв'язок матиме невисоку точність. Також стійкість залежить від вибору початкової точки a або b і методу розв'язування задачі Коші.

15.3. Елементи теорії лінійних РС

Розглянемо диференціальну задачу, яка містить диференціальне рівняння і додаткові умови

$$(Lu)(x)=r(x),\ \ x\in D \tag{15.11}$$

$$(lu)(x)=\mu(x),\ \ x\in\gamma, \tag{15.12}$$

де L – лінійний диференціальний оператор, лінійним оператором l задаються додаткові умови, u,r,μ – задані функції, визначені у відповідних областях.

Прикладом є задача Коші

$$(lu)(x):=u''+\rho(x)u'+q(x)u=r(x),\ \ x\in(a,b], \tag{15.13}$$

$$(lu)(a):=\begin{cases}u(a),\\ u'(a),\end{cases}\qquad \mu(a):=\begin{bmatrix}u_0,\\ u_0'.\end{bmatrix}$$

Тут $\gamma=\{a\}$, u_0 і u_0' – задані числа.

Якщо умова (15.12) набуває вигляду

$$(lu)(x):=\begin{cases}\alpha_0u(a)+\beta_0u'(a),\\ \alpha_1u(b)+\beta_1u'(b),\end{cases}\qquad \mu(x):=\begin{bmatrix}\gamma_0,\\ \gamma_1,\end{bmatrix}$$

то маємо лінійну крайову задачу (15.9), (15.4).

Нагадаємо, що L – лінійний оператор над полем дійсних (комплексних) чисел, якщо для довільних u і v з деякого лінійного простору і $\alpha,\beta\in R$ виконується рівність $L(\alpha u+\beta v)=\alpha L(u)+\beta L(v)$.

Задамо на множині D сітку $\overline{\Delta}_h$. Наприклад, рівномірну сітку з кроком h на відрізку $[a, b]$. Апроксимувавши на рівномірній сітці $\overline{\Delta}_h$ задачу (15.11), (15.12), одержимо різницеву задачу

$$\left(L_h y_h\right)(x) = \varphi_h(x), \quad x \in \Delta_h, \tag{15.14}$$

$$\left(l_h y_h\right)(x) = \xi_h(x), \quad x \in \gamma_h, \tag{15.15}$$

де $y_h = y_h(x)$ – сіткова функція, яка апроксимує $u(x)$ на сітці $\overline{\Delta}_h$, φ_h і ξ_h – апроксимації r і μ на Δ_h і γ_h відповідно.

Апроксимувати диференціальне рівняння (15.13) у вузлах сітки $x_n = x_0 + nh,\ n = \overline{1, N-1},\ x_0 = a$ можна, використавши центральну різницеву похідну (10.4) і симетричну різницеву похідну (10.5) для апроксимації u' і u'' відповідно. У підсумку одержимо різницеву апроксимацію рівняння (15.13)

$$\left(L_h y_h\right)(x_n) := \frac{1}{h^2}\left(y_{n-1} - 2y_n + y_{n+1}\right) + \frac{p_n}{2h}\left(y_{n+1} - y_{n-1}\right) + q_n y_n = h^2 r_n. \tag{5.16}$$

Увівши позначення

$$\Im := \begin{bmatrix} L \\ l \end{bmatrix}, \quad f := \begin{bmatrix} r \\ \mu \end{bmatrix}, \quad \Im_h := \begin{bmatrix} L_h \\ l_h \end{bmatrix}, \quad g_h := \begin{bmatrix} \varphi_h \\ \xi_h \end{bmatrix},$$

задачі (15.11), (15.12) і (14.14), (15.15) запишемо у вигляді

$$\left(\Im u\right)(x) = f(x), \ x \in \overline{D} = D \cup \gamma, \tag{15.17}$$

$$\left(\Im_h y_h\right)(x) = g_h(x), \ x \in \overline{\Delta_h} = \Delta_h \cup \gamma_h. \tag{15.18}$$

Означення 15.1. *Система рівнянь (15.18), яка залежить від кроку сітки h як параметра, називається різницевою схемою.*

Означення 15.2. *Сіткова функція $z_h(x) = y_h(x) - u_h(x)$, $x \in \overline{\Delta_h}$, називається похибкою розв'язку РС у точці x.*

Підставивши $y_h = z_h + u_h$ в рівняння (15.18), одержимо РС

$$\left(\Im_h z_h\right)(x) = g_h(x) - \left(\Im_h u_h\right)(x). \tag{15.19}$$

Означення 15.3. *Сіткова функція*

$$\psi_h^{(1)}(x) = g_h(x) - \left(\Im_h u\right)(x), \quad x \in \overline{\Delta}_h$$

називається похибкою апроксимації РС.

Означення 15.4. *РС (15.18) апроксимує диференціальну задачу (14.17) на сітці $\overline{\Delta}_h$, якщо $\left\|\psi_h^{(1)}\right\| \to 0$ при $h \to 0$. РС має*

порядок $p>0$, *якщо* $\left\|\psi_h^{(1)}\right\| \le Ch^p$, *де стала* $C>0$ *і не залежить від* h.

Тут $\|\cdot\|$ – деяка сіткова норма, наприклад, $\|u_h\| = \max\limits_{x\in\bar{\Delta}_h} |u_h(x)|$.

Означення 15.5. *РС (15.18) називається стійкою, якщо для досить малого кроку сітки* h *і довільної сіткової функції* g_h

$$\|y_h\| \le B\|g_h\|, \tag{15.20}$$

де стала B *не залежить від* h.

Означення 15.6. *РС збігається до розв'язку диференціальної задачі на сітці* $\bar{G}_h$ *(РС збіжна), якщо* $\|z_h\| \to 0$ *при* $h \to 0$. *РС має порядок* $p>0$, *якщо* $\|z_h\| \le Mh^p$, *де* M *– стала, не залежна від* h.

Між властивостями апроксимації, стійкості і збіжності РС є зв'язок, який можна записати у вигляді символічної формули: "*апроксимація+стійкість=збіжність*".

Теорема 15.1 (Лакса). *Нехай диференціальна задача (15.14) коректна, а РС (15.15) апроксимує диференціальну задачу і стійка. Тоді РС збіжна. Якщо порядок апроксимації РС дорівнює* p, *то й порядок її точності також дорівнює* p.

Доведення. Для похибки z_h РС маємо рівняння (15.19). Оскільки РС (15.18) апроксимує диференціальну задачу (15.17), то $\left\|\psi_h^{(1)}\right\| \to 0$ при $h \to 0$. На підставі стійкості РС маємо

$$\|z_h\| \le B\left\|\psi_h^{(1)}\right\|.$$

Звідси й випливає, що $\|z_h\| \to 0$ при $h \to 0$. Якщо похибка апроксимації має порядок $p>0$, то на підставі нерівності (15.20) маємо

$$\|z_h\| \le B\left\|\psi_h^{(1)}\right\| \le BMh^p.$$

Отже, РС має порядок точності p. ■

15.4. Різницева схема для лінійної крайової задачі

15.4.1. Різницева схема. Розглянемо лінійну крайову задачу

$$u'' + p(x)u' + q(x)u = r(x), \quad x \in (a,b), \tag{15.21}$$

$$u(a) = \mu_0, \; u(b) = \mu_1, \tag{15.22}$$

де p, q і r − визначені на проміжку $[a, b]$ функції, μ_0 і μ_1 – задані числа. Урахувавши (15.16) відповідна крайовій задачі (15.21), (15.22) РС набуде вигляду

$$\frac{1}{h^2}\left(y_{i+1}-2y_i+y_{i-1}\right)+\frac{p_i}{2h}\left(y_{i+1}-y_{i-1}\right)+q_i y_i=r_i,\ i=\overline{1,N-1}; \qquad (15.23)$$

$$y_0=\mu_0,\ y_N=\mu_1,$$

де p_i, q_i, r_i – значення функцій у вузлі. $x=x_i\ x=x_i$

Систему рівнянь (15.23) для знаходження розв'язку зручніше записати у вигляді

$$\left(1-\frac{1}{2}hp_i\right)y_{i-1}-\left(2-h^2q_i\right)y_i+\left(1+\frac{1}{2}hp_i\right)y_{i+1}=h^2r_i,\ i=\overline{1,\ N-1}; \qquad (15.24)$$

$$y_0=\mu_0,\ y_N=\mu_1.$$

Матриця СЛАР (15.24) тридіагональна, тому її можна розв'язувати методом прогонки, формули якого наведені у розділі 2.

Розглянемо випадок крайових умов третього роду (15.4). Похідні в точках $x_0=a,\ x_N=b$ можна апроксимувати з першим порядком за допомогою формул

$$u_0'=\frac{u_1-u_0}{h}+O(h),\ u_N'=\frac{u_N-u_{N-1}}{h}+O(h).$$

У цьому випадку знаходження значень розв'язку $y_0,...,y_N$ система рівнянь (14.24) доповнюється ще двома рівняннями

$$\begin{aligned}&\left(\alpha_0h-\beta_0\right)y_0+\beta_0y_1=h\gamma_0,\\&-\beta_1y_{N-1}+\left(\alpha_1h+\beta_1\right)y_N=h\gamma_1.\end{aligned} \qquad (15.25)$$

Матриця системи (15.24), (15.25) має порядок $N+1$ і залишається при цьому тридіагональною.

Похідні u_0' і u_N' із другим порядком апроксимуються за допомогою формул числового диференціювання

$$u_{x,0}\approx\frac{-3u_0+4u_1-u_2}{2h},\ u_{x,N}\approx\frac{3u_N-4u_{N-1}+u_{N-2}}{2h}.$$

Тоді крайовим умовам відповідають різницеві рівняння

$$\begin{aligned}&\left(2\alpha_0h-3\beta_0\right)y_0+4\beta_0y_1-\beta_0y_2=2h\gamma_0,\\&\beta_1(y_{N-2}-4y_{N-1})+\left(\alpha_1h+\beta_1\right)y_N=2h\gamma_1.\end{aligned} \qquad (15.26)$$

Матриця системи рівнянь (15.24), (15.26) уже не буде тридіагональною, якщо β_0 і β_1 не дорівнюють нулю. Звести систему рівнянь до тридіагонального вигляду можна, вилучивши y_2 із першого з рівнянь (15.26) та y_{N-2} із другого рівняння за допомогою першого і останнього рівнянь системи (15.24) відповідно. Одержимо систему рівнянь

$$\bar{\alpha}_0 y_0 + \bar{\beta}_0 y_1 = \bar{\gamma}_0,\ \bar{\beta}_1 y_{N-1} + \bar{\alpha}_1 y_N = \bar{\gamma}_1,$$

де $\bar{\alpha}_0 = \alpha_0 h\left(2 + hp_1\right) - 2\beta_0(1 + hp_1),\ \bar{\beta}_0 = \beta_0(2 + 2hp_1 + h^2 q_1),$

$\bar{\gamma}_0 = \gamma_0 h\left(2 + hp_1\right) + \beta_0 r_1 h^2,\ \bar{\alpha}_1 = \alpha_1 h(2 - hp_{N-1}) + 2\beta_0(1 - hp_{N-1}),$

$$\bar{\beta}_1 = -\beta_1\left(2 - 2hp_{N-1} - h^2 q_{N-1}\right),\ \bar{\gamma}_1 = \gamma_1 h(2 - hp_{N-1}) - \beta_1 r_{N-1} h^2.$$

15.4.2. Похибка апроксимації. Позначимо через L_h – лінійний оператор

$$(L_h y_h)_n := y_{\bar{x},n} + p_n y_{\dot{x},n} + q_n y_n,\ \ n = \overline{1, N-1}.$$

Тоді РС схема (15.24) набуде вигляду

$$(L_h y_h)_n = r_n\ ,\ n = \overline{1, N-1};\ y_0 = \mu_0,\ y_N = \mu_1. \qquad (15.27)$$

Запишемо рівняння для похибки z_h

$$(L_h z_h)_n = (L_h y_h)_n - (L_h u_h)_n = r_n - (L_h u_h)_n = \psi_n^{(1)},$$

де $\psi_n^{(1)}$ – похибка апроксимації РС (15.27). Зрозуміло, що $z_h(a) = z_h(b) = 0.$ Отже, порядок апроксимації РС визначається порядком апроксимації похідних u_n' та u_n''.

Теорема 15.2. *Нехай* $p, q, r \in C^2[a,b]$. *Тоді РС (15.24) має другий порядок апроксимації.*

Доведення. Із умови гладкості функцій p, q, r на підставі рівняння (15.21) випливає, що $u \in C^4[a,b]$. Тоді

$$u_{\dot{x},i} - u_i' = \frac{1}{6}h^2 u'''\left(\xi_i\right),\ u_{\bar{x}\dot{x},i} - u_i'' = \frac{h^2}{12}u^{(4)}\left(\eta_i\right),\ \xi_i, \eta_i \in (x_{i-1}, x_{i+1}).$$

Тобто, $\psi_i^{(1)} = r_i - u_{\bar{x}\dot{x},i} - p_i u_{\dot{x},i} - q_i u_i =$

$$= r_i - \left(u_i'' + \frac{h^2}{12}u^{(4)}\left(\eta_i\right)\right) - p_i\left(u_i' + \frac{1}{6}h^2 u'''\left(\xi_i\right)\right) - q_i u_i =$$

$$= \left(r - u'' - pu' - qu\right)_i - \frac{1}{12}h^2 u^{(4)}\left(\eta_i\right) - \frac{1}{6}p_i h^2 u'''\left(\xi_i\right) =$$

$$= -\frac{1}{12}h^2\left(u^{(4)}(\eta_i)+2u'''(\xi_i)\right).$$

Для сіткової функції $\psi_h^{(1)}$ маємо

$$\left\|\psi_h^{(1)}\right\| = \max_{x\in\omega_h}\left|\psi_h(x)\right| \le \left(M_4 + 2M_3 \max_{a\le x\le b}\left|p(x)\right|\right)h^2/12 \equiv Ch^2,$$

$M_i = \max\limits_{a\le x\le b}\left|u^{(i)}(x)\right|$, тобто РС має другий порядок апроксимації. ■

15.4.3. Існування та єдиність розв'язку РС. Розглянемо першу крайову задачу (15.21), (15.22) і відповідну РС (15.24). Визначимо для сіткової функції $y_h = y_h(x), x\in\bar{\Delta}_h$ лінійний оператор Λ_h

$$(\Lambda_h y_h)_n := a_n y_{n-1} - c_n y_n + b_n y_{n+1},\ n=\overline{1,N}. \qquad (15.28)$$

Тут a_n, b_n, c_n – деякі задані числа, $y_0=\mu_0,\ y_N=\mu_1$.

Зауважимо, що оператор Λ_h задається тридіагональною матрицею порядку $N-1$. Вивчимо питання існування та єдиності розв'язку $y_1,...,y_{N-1}$ системи (15.24) на підставі наступної теореми.

Теорема 15.3 (принцип максимуму, [58]). *Нехай виконуються умови:*

1) $a_n>0, b_n>0, c_n \ge a_n+b_n,\ n=\overline{1,N-1};$ (15.29)

2) $(\Lambda_h y_h)_n \ge 0,\ n=\overline{1,N-1};$ (15.30)

$((\Lambda_h y_h)_n \le 0,\ n=\overline{1,N-1}).$

Тоді сіткова функція набуває найбільшого (найменшого) значення в точках a *або* b. ■

Наслідок 15.1. *Нехай* $(\Lambda_h y_h)_n \ge 0,\ n=\overline{1,N-1};\ \ \mu_0\ge 0,\ \mu_1\ge 0.$ *Тоді* $y_n\ge 0$ *для* $n=\overline{1,N-1}.$

Справді, якби для деякого $1<n_0<N-1$ *було б* $y_{n_0}<0$, *що суперечить висновку теореми 15.3, бо* $\min\limits_{0<n<N} y_i \ge \min(\mu_0,\mu_1)\ge 0.$

Теорема 15.4. *Нехай виконуються такі умови:*

1) $p,q\in C[a,b];$

2) $h\max\limits_{x\in[a,b]}\left|p(x)\right|<2;$

3) $q(x)\le 0,\ x\in[a,b].$

Тоді існує єдиний розв'язок СЛАР (15.24).

Доведення. Розглянемо відповідну (*15.24*) однорідну задачу

$$\left(1-\frac{1}{2}p_n h\right)y_{n-1}-\left(2-h^2 q_n\right)y_n+\left(1+\frac{1}{2}p_n h\right)y_{n+1}=0,\ \mu_0=\mu_1=0. \quad (15.31)$$

Для системи (15.31) сумісно виконуються обидві умови (15.30). Коефіцієнти

$$a_n=1-\frac{1}{2}hp_n>0,\ \ b_n=1+\frac{1}{2}hp_n>0,\ \ c_n=2-h^2q_n\geq a_n+b_n=2$$

задовольняють нерівності (15.29). Отже, на підставі теореми 15.3 $\max\limits_{0\leq n\leq N} y_i=\min\limits_{0\leq n\leq N} y_i=0$. Звідси випливає, що $y_h(x)\equiv 0$. Якщо розв'язок однорідної СЛАР тільки нульовий, то розв'язок відповідної неоднорідної системи існує, й до того ж тільки один. ■

Наближений розв'язок крайової задачі (15.21), (15.22) знаходиться як розв'язок СЛАР (15.24). Вплив обчислювальної похибки на розв'язок цієї системи характеризується числом обумовленості $\kappa(A)=\|A\|\cdot\left\|A^{-1}\right\|$, де A – матриця системи (15.24). Згідно з нерівністю (3.4),

$$\frac{\|\delta x\|}{\|x\|}\leq\kappa(A)\frac{\|\delta b\|}{\|b\|},$$

де δb і δx – збурення правої частини і розв'язку СЛАР.

Розглянемо число обумовленості матриці на прикладі автономної задачі

$$u''=-f(x),\ x\in(0,1);\ u(0)=u(1)=0,$$

для якої різницева схема набуває вигляду

$$y_{\bar{x}x,n}=-f\left(x_n\right),\ n=\overline{1,N-1};\ \mu_0=\mu_1=0. \quad (15.32)$$

Матриця системи (15.32) тоді така:

$$A=\frac{1}{h^2}\begin{bmatrix} -2 & 1 & 0 & 0 & \dots & 0 & 0 & 0\\ 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0\\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1\\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -2 \end{bmatrix}$$

Для симетричної додатно визначеної матриці $\|A\|_2=\max\limits_{1\leq n\leq N-1}|\lambda_n|=\lambda_{N-1}$. Матриця A^{-1} також симетрична і додатно визначена, тому $\left\|A^{-1}\right\|_2=\max\limits_{1\geq n\leq N-1}|\mu_n|=\mu_{N-1}=\lambda_1^{-1}$, де μ_i – власні числа матриці A^{-1}. Отже, $\|A\|_2=\lambda_{N-1}/\lambda_1$.

Відомо, що $\lambda_n = \frac{4}{h^2}\sin^2\frac{\pi n h}{2}$, $i = \overline{1, N-1}$. Тому при малих h $\sin\frac{\pi h}{2} \approx \frac{\pi h}{2}$, отже $\lambda_1 \approx \frac{4}{h^2}\frac{\pi^2 h^2}{4} = \pi^2$, $\lambda_{N-1} = \frac{4}{h^2}\sin^2\frac{\pi(N-1)h}{2} = \frac{4}{h^2}\cos^2\frac{\pi h}{2} \approx \frac{4}{h^2}$. Число обумовленості $\kappa(A) = \frac{\lambda_{N-1}}{\lambda_1} \approx \frac{4}{\pi^2 h^2} \to \infty$, коли $h \to 0$.

Отже, зі зменшенням кроку h обумовленість системи зростає, оскільки $\kappa(A) = O(h^{-2})$. Тому практичне розв'язування крайової задачі потребує вибору сітки з таким кроком, щоб збільшити стійкість розв'язку до похибок округлень (крок h малий) і зменшити число обумовленості (крок h не дуже малий). Відзначимо, що для більшості практичних задач вибір величини кроку h не дається взнаки.

15.5. Стійкість різницевої схеми

15.5.1. Теорема порівняння. Стійкість і збіжність різницевої схеми (15.24) будемо досліджувати в просторі сіткових функцій із рівномірною нормою $\|y_h\| = \max\limits_{0\le n\le N}|y_n|$. Розглянемо СЛАР

$$(\Lambda_h y_h)_n = 0,\ n = \overline{1, N-1};\ y_0 = \mu_0,\ y_N = \mu_1,$$

де оператор Λ_h визначений згідно з (15.28), і відповідну систему порівняння

$$(\Lambda_h \bar{y}_h)_n = -\bar{\varphi}_n,\ n = \overline{1, N-1};\ \bar{y}_0 = \bar{\mu}_0,\ \bar{y}_N = \bar{\mu}_1,$$

сіткові функції φ_h і $\bar{\varphi}_h$ задані на сітці $\bar{\Delta}_h$.

Теорема 15.5 (порівняння, [59]). *Нехай виконуються такі умови:*

1) $a_n > 0, b_n > 0, c_n \ge a_n + b_n,\ n = \overline{1, N-1}$;

2) $\bar{\varphi}_n \ge 0,\ n = \overline{1, N-1},\ \bar{\mu}_0 \ge 0,\ \bar{\mu}_1 \ge 0$;

$|\varphi_n| \le \bar{\varphi}_n,\ n = \overline{1, N-1};\ |\mu_0| \le \bar{\mu}_0,\ |\mu_1| \le \bar{\mu}_1$;

Тоді справджуються оцінки

$|y_n| \le \bar{y}_n,\ n = \overline{1, N-1}$. ■

15.5.2. Стійкість за крайовими умовами

Означення 15.7. *РС* (15.24) *стійка за крайовими умовами* $\mu := (\mu_0, \mu_1)$*), якщо*

$$\|y_h\|_1 \le B_2 \|\mu\|_3, \qquad 15.33)$$

де $\|\cdot\|_1$, $\|\cdot\|_2$ *– деякі сіткові норми, стала* $B_2 > 0$ *не залежить від* h.

Означенню 15.7 можна дати таку інтерпретацію. Внесемо в граничні умови збурення δ_1 і δ_2: $\bar{\mu}_0 = \mu_0 + \delta_0$, $\bar{\mu}_1 = \mu_1 + \delta_1$. Одержимо відповідний збурений розв'язок

$$\bar{y}_n = y + \varepsilon_n, \ n = \overline{1, N-1}.$$

Збурення розв'язку ε_i при фіксованих φ_i задовольняють систему рівнянь

$$(\Lambda_h \varepsilon_h)_n = 0, \ n = \overline{1, N-1}; \ \varepsilon_0 = \delta_0, \varepsilon_N = \delta_1. \qquad (15.34)$$

Для (15.34) означення 15.7 формулюється так.

Означення 15.8. *РС (15.24) стійка за крайовими умовами, якщо виконується нерівність*

$$\|\varepsilon_h\|_1 \le B_1 \|\delta_h\|_2. \qquad (15.35)$$

Теорема 15.6. *Нехай виконуються умови 1) – 3) теореми 15.4. Тоді РС (15.24) для крайової задачі (15.21),(15.22) стійка за крайовими умовами.*

Доведення. Побудуємо відповідну (*15.34*) систему порівняння вигляду

$$(\Lambda_h \bar{\varepsilon}_h)_n = 0, \bar{\varepsilon}_0 = \bar{\varepsilon}_N = \|\delta\|,$$

$$a_i = 1 - \frac{1}{2} p_n h > 0, b_i = 1 + \frac{1}{2} p_n h > 0, c_n = 2 + h^2 q_n > a_n + b_n.$$

На підставі наслідку з теореми 15.4 маємо $\bar{\varepsilon}_n \ge 0, n = \overline{1, N}$. Згідно з теоремою 15.5 $\max|\bar{\varepsilon}_i| \le \max(\bar{\varepsilon}_0, \bar{\varepsilon}_N) = \|\delta\|$.

Тепер застосуємо теорему порівняння. Маємо, $\varphi_n = \bar{\varphi}_n, \tilde{\varphi}_n = 0, \ n = \overline{1, N}$; $|\varepsilon_\nu| \le \bar{\varepsilon}_\nu = \|\delta\|, \nu = 0,1$. Отже, $\max\limits_{0 \le n \le N} |\varepsilon_n| \le \|\delta\|$ і нерівність (15.35) виконується зі сталою $B_1 = 1$.

15.7.3. Стійкість за правою частиною. Зафіксуємо значення μ_0 і μ_1 у крайових умовах (15.22).

Означення 15.9. *РС (15.24) стійка за правою частиною, якщо*

$$\|y_h\|_1 \le B_2 \|\varphi_h\|_3, \qquad (15.36)$$

де $B_2 > 0$ *і не залежить від* h.

Якщо розглянути відповідну (15.24) збурену РС

$$(\Lambda_h \tilde{y}_h)_n = -\tilde{\varphi}_n, \tilde{\varphi}_n = \bar{\varphi}_n + \eta_n, \ n = \overline{1, N-1};$$

$\tilde{y}_0 = \mu_0;\ \tilde{y}_N = \mu_1,\ n = \overline{1, N-1},$

то для збурення розв'язку $\xi_n = \tilde{y}_n - y_n$ одержимо співвідношення

$$(\Lambda_h \xi_h)_n = -\eta_n,\ \ n = \overline{1, N-1};\ \ \xi_0 = \xi_N = 0.$$

Стійкість за правою частиною в цьому випадку означає, що малим за нормою $\|\cdot\|_3$ збуренням правої частини (15.24) відповідає мале, за нормою $\|\cdot\|_1$, збурення розв'язку. Вважатимемо, що стала B_2 не дуже велика, інакше стійкість умовна.

Теорема 15.7 [58]. *Нехай виконуються умови 1), 2) теореми 15.4 і при* $x \in [a, b]$ *справджується нерівність* $q(x) < 0.$

*Тоді РС (*15.24*) стійка за правою частиною зі сталою, де* $B_2 = \min\limits_{x \in [a,b]} |q(x)|.$ ■

Зауваження 15.1. *Із виконання умов теореми 15.7 випливає і стійкість за граничними даними. Тому, можна стверджувати про стійкість різницевої схеми за крайовими умовами і правою частиною, що визначається нерівністю*

$$\|y_h\|_1 \le B_1 \|\mu_h\|_2 + B_2 \|\varphi_h\|_3.$$

15.6. Збіжність різницевої схеми

Теорема 15.8. *Нехай виконуються умови:*

1) $p, q \in C^2[a,b];$

2) $h \max\limits_{x \in [a,b]} |p(x)| < 2;$

3) $q(x) < 0,\ x \in [a,b].$

Тоді РС (15.24) збіжна в рівномірній нормі зі швидкістю $O(h^2)$, *тобто має другий порядок точності.*

Доведення. Застосуємо для доведення теорему Лакса. Умова 1) забезпечує другий порядок апроксимації (теорема 15.2) зі сталою $C = \left(M_4 + 2M_3 \max\limits_{a \le x \le b} |p(x)|\right)/12$, де сталими M_3 і M_4 обмежені на відрізку $[a, b]$ похідні $|u^{(3)}(x)|$ і $|u^{(4)}(x)|$ відповідно.

Із умов 1) – 3) на підставі теорем 15.6 і 15.7 випливає стійкість РС (15.24) за крайовими умовами і правою частиною зі сталими $M_1 = 1$ і $M_2 = \left(\min\limits_{a \le x \le b} |p(x)|\right)^{-1}$. Застосування теореми Лакса дає змогу записати нерівність

$$\|z_h\| \le \|y_h - u_h\| \le (1 + M_2)h^2,$$

що й завершує доведення теореми. ■

Зауваження 15.2. *Для практичної оцінки похибки числового розв'язку можна скористатися правилом Рунге, тобто обчислити розв'язок на сітках з кроком* h *і* $h/2$ *та порівняти одержані розв'язки у спільних вузлах.*

15.7. Інтегро-інтерполяційний метод побудови РС

Для деяких крайових задач, наприклад тих, що виводяться з інтегральних законів збереження, похідна розв'язку може існувати не в усіх точках. Наприклад, для крайової задачі

$$\begin{aligned} &-(k(x)u')' + q(x)u = f(x),\ 0 < x < l; \\ &u(0) = \mu_0,\ u(l) = \mu_1, \end{aligned} \tag{15.37}$$

якою описується стаціонарний розподіл температури у стержні, коефіцієнти можуть мати розриви першого роду. Тут $u(x)$ – температура в точці з координатою x, $k(x)$ – коефіцієнт теплопровідності, функція $q(x) \ge 0$ задає теплообмін із зовнішнім середовищем, а $f(x)$ характеризує густину теплових джерел. Якщо стержень скомпонований з різних матеріалів, тому $k(x)$ – кусково неперервна функція і розв'язок задовольняє рівняння (15.37) в деякому узагальненому сенсі.

Інтегро-інтерполяційний метод побудови РС ґрунтується на тих же фізичних законах збереження, на яких одержане диференціальне рівняння (15.37). РС такого типу називаються *консервативними*.

Якщо функція $k(x)$ має розриви в деяких точках, то РС, побудовані шляхом апроксимації похідних, може не збігатись. В [58, с. 147] наведено приклад такої РС для крайової задачі

$$-\left(k(x)u'\right)' = 0,\ 0 < x < 1;\ u(0) = 1,\ u(1) = 0,$$

де $k(x) = k_1$ при $0 \le v \le 1/\sqrt{2}$ і $k(x) = k_2 \ne k_1$ при $1/\sqrt{2} < x \le 1$.

Побудуємо на $[0, l]$ рівномірну сітку з кроком $h = 1/N$ і введемо такі позначення: $x_{n\pm1/2} = x \pm h/2$, $u_{n\pm1/2} = u(x \pm h/2)$, $W(x) = -k(x)u'$ – тепловий потік, $n = \overline{1, N-1}$. Проінтегруємо рівняння (15.37) на проміжку $[x - h/2, x + h/2]$ (рис. 15.1).

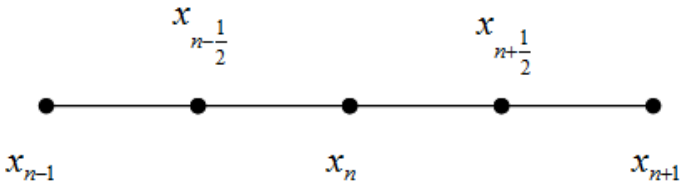


Рис.15.1

Одержимо таку рівність

$$W_{n-1/2} - W_{n+1/2} + \int_{x_{n-1/2}}^{x_{n+1/2}} f(x)dx = \int_{x_{n-1/2}}^{x_{n+1/2}} q(x)u(x)dx. \qquad (15.38)$$

Рівністю (15.38) описується баланс тепла на проміжку $[x-h/2, x+h/2]$. Перші два доданки в лівій частині (15.38) – це кількість тепла, яка надходить і виходить з елемента стержня, третій доданок описує кількість тепла, що виділяється джерелами. Права частина в (15.38) визначає теплообмін між елементом стержня і зовнішнім середовищем.

Інтеграл у правій частині рівності (15.38) обчислимо за формулою центральних прямокутників, для чого введемо ще такі позначення:

$$\varphi_n = \frac{1}{h}\int_{x_{n-1/2}}^{x_{n+1/2}} f(x)dx,\ d_n = \frac{1}{h}\int_{x_{n-1/2}}^{x_{n+1/2}} q(x)dx,\ a_n = \left(\frac{1}{h}\int_{x_{n-1}}^{x_n}\frac{dx}{k(x)}\right)^{-1}. \qquad (15.39)$$

Тоді з (15.38) одержимо

$$W_{n-1/2} - W_{n+1/2} + h\varphi_n - hd_n u_n \approx 0. \qquad (15.40)$$

Оскільки $u'(x) = -W(x)/k(x)$, то знову застосуємо формулу центральних прямокутників і матимемо

$$u_n - u_{n-1} = -\int_{x_{n-1}}^{x_n}\frac{W(x)}{k(x)}dx \approx -W_{n-1/2}\int_{x_{n-1}}^{x_n}\frac{dx}{k(x)} = -W_{n-1/2}ha_n^{-1}.$$

У такий спосіб одержимо, що, $u_{n+1} - u_n \approx W_{n+1/2}ha_{n+1}^{-1}$. Звідси випливає, що

$$W_{n-1/2} \approx -a_n\frac{u_n - u_{n-1}}{h},\ W_{n+1/2} \approx a_{n+1}\frac{u_{n+1} - u_n}{h}.$$

Підставивши $W_{n\pm1/2}$ в (15.39), одержимо апроксимацію рівняння в точці x_n

$$a_{n+1}\frac{y_{n+1} - y_n}{h} - a_n\frac{y_n - y_{n-1}}{h} - hd_n y_n = -h\varphi_n \qquad (15.41)$$

або

$$\frac{1}{h}\left(a_{n+1}y_{x,n}-a_n y_{\bar{x},n}\right)=-d_n y_n=-\varphi_n,\ n=\overline{1,N-1},\ y_0=\mu_0,\ y_N=\mu_1.$$

Система (15.41) – СЛАР з тридіагональною матрицею.

Нехай на лівому кінці стержня задана крайова умова третього роду

$$-k(0)u'(0)+\beta u(0)=\gamma. \tag{15.42}$$

Після інтегрування на $\left[0,x_{1/2}\right]$ одержимо рівняння

$$a_1 y_{\bar{x},1}+\left(\beta+0.5hd_0\right)y_0=\gamma-0.5h\varphi_0, \tag{15.43}$$

де $\varphi_0=\frac{2}{h}\int\limits_0^{x_{1/2}} f(x)dx,\ d_0=\frac{2}{h}\int\limits_0^{x_{1/2}} q(x)dx.$

Тепер для задачі (15.37), (15.42) маємо РС (15.43), (15.41), яка є СЛАР порядку N.

15.8. Нелінійна крайова задача

15.8.1. Різницева схема. Розглянемо диференціальне рівняння другого порядку

$$u''=f(x,u,u'),\ a<x<b, \tag{15.44}$$

із крайовими умовами

$$u(a)=\mu_0,\ u(b)=\mu_1. \tag{15.45}$$

Припустимо, що функція $f(x,u,v)$ визначена в деякій області G, похідна $f_v\in C^1(G)$ й існує єдиний розв'язок задачі (15.44), (15.45), такий, що $u\in C^4[a,b]$.

У вузлах сітки $x_n\in\Delta_h$ апроксимуємо похідні u'_n та u''_n різницевими похідними

$$u'_n=\left(u_{n+1}-u_{n-1}\right)/(2h)+O(h^2),\ u''_n=\left(u_{n+1}-2u_n+u_{n-1}\right)/h^2+O(h^2).$$

Одержимо нелінійну РС:

$$\frac{1}{h^2}\left(y_{n-1}-2y_n+y_{n+1}\right)=f\left(x_n,y_n,\frac{y_{n+1}-y_{n-1}}{2h}\right),\ n=\overline{1,N-1}, \tag{15.46}$$

$$y_0=\mu_0,\ y_N=\mu_1.$$

Для знаходження наближеного розв'язку $y_1,\ldots,y_{N-1}$ на сітці Δ_h потрібно розв'язати систему $N-1$ нелінійних рівнянь (15.46).

Підставимо $y_n=u_n+z_n$, де z_n – похибка різницевої схеми, у систему (15.46). Одержимо рівняння для похибки

$$\left(z_{n-1}-2z_n+z_{n+1}\right)/h^2=\psi_n^{(1)}+\psi_n^{(2)},\ n=\overline{1,N-1};\ z_0=z_N=0,$$

$$\psi_n^{(1)}=-\frac{1}{h^2}\left(u_{n-1}-2u_n+u_{n+1}\right)+f\left(x_n,u_n,\frac{u_{n+1}-u_{n-1}}{2h}\right),$$

$$\psi_n^{(2)}=-f\left(x_n,u_n+z_i,\frac{u_{n+1}-u_{n-1}}{2h}+\frac{z_{n+1}-z_{n-1}}{2h}\right)-f\left(x_n,u_n,\frac{u_{n+1}-u_{n-1}}{2h}\right).$$

На підставі зроблених припущень щодо гладкості розв'язку крайової задачі похибку апроксимації $\psi_n^{(1)}$ зведемо до вигляду

$$\psi_n^{(1)}=-u_n''-\frac{h^2}{12}u^{(4)}(\eta_n)+f\left(x_n,u_n,u_n'-\frac{h^2}{6}u'''(\xi_n)\right)=$$

$$=\left(-u_n''+f\left(x_n,u_n,u_n'\right)\right)-\frac{h^2}{12}u^{(4)}(\eta_n)-f_{u'}\left(x_n,u_n-\theta\frac{h^2}{6}u'''(\xi_n)\right)\times$$

$$\frac{h^2}{6}u'''(\xi_n)=-\frac{h^2}{12}\left(u^{(4)}(\eta_n)+2f_{u'}\left(x_n,u_n-\theta\frac{h^2}{6}u'''(\xi_n)\right)u'''(\xi_n)\right),$$

де $\eta_n,\ \xi_n\in(x_{n-1},x_n),\ \theta\in(0,1)$. Звідси одержимо оцінку

$$\left|\psi_n^{(1)}(h)\right|\le Ch^2,$$

де $C=\left(\max\limits_{a\le x\le b}\left|u^{(4)}(x)\right|+2\max\limits_{G}\left|f_v(x,u,v)\right|\right)/12$. Отже, РС (15.46) має другий порядок апроксимації.

15.8.2. Збіжність різницевої схеми. Розглянемо дещо простішу крайову задачу

$$u''=f(x,u),\ a<x<b;\ u(a)=\mu_0,\ u(b)=\mu_1. \qquad (15.47)$$

Дослідимо збіжність відповідної РС

$$\frac{1}{h^2}\left(y_{n-1}-2y_n+y_{n+1}\right)=f(x_n,y_n),\ u(a)=\mu_0,\ u(b)=\mu_1. \quad (15.48)$$

Теорема 15.9. *Нехай виконуються такі умови:*

1) *Існує єдиний розв'язок крайової задачі* (15.47) *і* $u\in C^4[a,b]$;
2) $f_u\in C^1(G)$ *і* $f_u\ge m>0$ *в* G;

*Тоді РС (*15.48*) збіжна і має другий порядок точності.*

Доведення. Покажемо, що для похибки РС справджується оцінка

$$\left\|z_h\right\|=\max_{x\in\Delta_h}\left|z_h(x)\right|=\left\|y_h-u_h\right\|\le C_1h^2,\ C_1=const>0.$$

Оскільки

$$\varphi_n^{(2)} = f(x_n, u_{n+1} + z_{n+1}) - f(x_n, u_n) = f_u(u_n + \theta_n z_n) z_n \equiv d_n z_n,$$

то система (15.48) набуде вигляду

$$z_{n-1} - (2 + d_n h^2) z_n + z_{n+1} =$$

$$= \frac{h^2}{12} u^{(4)}(\eta_n), \text{ де } d_n = f_u(x_n, u_n + \theta_n z_n)\,), 0 < \theta_n < 1.$$

Із умови 2) теореми випливає, що $d_n > 0$ і матриця СЛАР є тридіагональною з перевагою головної діагоналі.

Для деякого $n_0, 1 \le n_0 \le N-1,$ маємо $\|z_h\| = |z_{n_0}| > 0.$ Інакше, $y_h(x) = u_h(x)$ коли $x \in \overline{\Delta_h}.$ Запишемо рівняння для похибки для $n = n_0$ у такому вигляді

$$\left(2 + d_{n_0} h^2\right) z_{n_0} = z_{n_0 - 1} + z_{n_0 + 1} + \frac{h^4}{12} u^{(4)}\left(\eta_{n_0}\right).$$

Звідси випливає, що

$$\left(2 + d_{n_0} h^2\right)\|z_h\| \le |z_{n_0 - 1}| + |z_{n_0 + 1}| + \frac{h^4}{12} M_4 \le 2\|z_h\| + \frac{h^4}{12} M_4,$$

де $\max\limits_{a \le x \le b} |u^{(4)}(x)| = M_4.$ Отже,

$$\|z_h\| \le \frac{M_4}{12 d_{i_0}} h^2 \le \frac{M_4}{12m} h^2 = C_1 h^2,\ C_1 = M_4/(12m). \qquad \blacksquare$$

Наслідок 15.2. *Розглянемо лінійне диференціальне рівняння*

$$u'' + q(x)u = r(x),\ a < x < b$$

*з крайовими умовами (14.46). Це рівняння набуває вигляду (*15.48*), якщо функція* $f(x,u) = -q(x)u + r(x).$ *Умова 2 теореми 15.9 виконується, якщо* $q(x) < 0,\ x \in [a,b],$ *оскільки* $f_u(x,u) = -q(x).$

15.8.3. Обчислення розв'язку системи (15.48). Нехай, крім похідної $f_u(x,u),$ існує обмежена в області G друга похідна $f_{uu}(x,u).$ Припустимо, що відоме k-те наближення розв'язку $y_n^{(k)}$, $n = \overline{1, N-1};\ k \ge 0,$ системи (15.47). Вважатимемо

$$y_n = y_n^{(k)} + \Delta_n,\ i = \overline{1, N-1};\ \Delta_0 = \Delta_N = 0.$$

Застосувавши метод Ньютона, побудуємо систему лінійних рівнянь для наближеного обчислення $\Delta_N.$ За формулою Тейлора

$$f(x_n, y_n) = f(x_n, y_n^{(k)} + \Delta_n) = f(x_n, y_n^{(k)}) + f_u(x_n, y_n^{(k)})\Delta_n + \alpha(\Delta_n),$$

де $\alpha(\Delta_n) = f_{uu}(x_n, y_n^{(k)} + \theta_n\Delta_n)\Delta_n^2$. Відкинувши величину $\alpha(\Delta_i) = O(\Delta_i^2)$, одержимо систему рівнянь для наближень $\Delta_i^{(k)}$:

$$\Delta_{i-1}^{(k)} - \left(2 + h^2 d_i^{(k)}\right)\Delta_i^{(k)} + \Delta_{i+1}^{(k)} = g_i^{(k)},\ i = \overline{1, N-1}$$
$$\Delta_0^{(k)} = \Delta_N^{(k)} = 0,$$

де $d_i^{(k)} = f_u\left(x_i, y_i^{(k)}\right),\ g_i^{(k)} = h^2 f\left(x_i, y_i^{(k)}\right) - \left(y_{i-1}^{(k)} - 2y_i^{(k)} + y_{i+1}^{(k)}\right)$.

За формулою Тейлора

$$f(x_i, y_i) = f(x_i, y_i^{(k)} + \Delta_i) = f(x_i, y_i^{(k)}) + f_u(x_i, y_i^{(k)}) + \alpha(\Delta_i),$$

де $\alpha(\Delta_n) = f_{uu}(x_n, y_n^{(k)} + \theta_n\Delta_n)\Delta_n^2 / 2$. Відкинувши величину $\alpha(\Delta_n) = O(\Delta_n^2)$, одержимо систему рівнянь для наближень $\Delta_n^{(k)}$:

$$\Delta_{n-1}^{(k)} - \left(2 + h^2 d_n^{(k)}\right)\Delta_n^{(k)} + \Delta_{n+1}^{(k)} = g_n^{(k)},\ i = \overline{1, N-1},\ \Delta_0^{(k)} = \Delta_N^{(k)} = 0,$$

де $d_n^{(k)} = f_u\left(x_n, y_n^{(k)}\right),\ g_n^{(k)} = h^2 f\left(x_n, y_n^{(k)}\right) - \left(y_{n-1}^{(k)} - 2y_n^{(k)} + y_{n+1}^{(k)}\right)$.

Оскільки $f_u(x, u) > 0$, то одержана система має перевагу головної діагоналі, що служить достатньою умовою існування та єдиності розв'язку системи лінійних рівнянь і стійкості методу прогонки. Після обчислення $\Delta_n^{(k)}$ маємо (k + 1)-е наближення

$$y_n^{(k+1)} = y_n^{(k)} + \Delta_n^{(k)},\ n = \overline{1, N-1},$$

яке можна уточнити у такий спосіб. Процес уточнення можна зупинити, наприклад, коли

$$\max\left|y_i^{(k+1)} - y_i^{(k)}\right| = \max\left|\Delta_i^{(k)}\right| < \varepsilon,$$

де ε − задана точність. Якщо початкові наближення $y_n^{(0)}$ вибрані близькими до y_n, то метод володіє квадратичною збіжністю.

15.9. Огляд аналітичних методів розв'язування крайових задач

Варіаційними методами (Рітца, найменших квадратів) та проекційних (Гальоркіна, колокації, скінченних елементів) наближений розв'язок крайової задачі

$$Au = f \qquad (15.49)$$

будується в аналітичному вигляді. Ідея варіаційних методів полягає у зведенні крайової задачі (15.49) до варіаційної задачі, тобто мінімізації деякого функціоналу

$$F(u) \to \min, \quad u \in H. \qquad (15.50)$$

Нехай H – гільбертовий простір [34], тобто нескінченновимір

ний евклідів простір із скалярним добутком $(\cdot,\cdot)$, повний відносно метрики $\rho(u,v) = \sqrt{(u-v,u-v)}$. Тоді задача (15.49) рівносильна варіаційній задачі з функціоналом вигляду [16]

$$Lu := -(k(x)u')' + q(x)u = f(x), \quad 0 < x < l; \qquad (15.51)$$

$$u(0) = u(l) = 0. \qquad (15.52)$$

Для простоти викладу припустимо, що k – кусково диференційовнаа, а q і f – кусково неперервні функції на $[0,l]$ зі скінченною кількістю точок розриву першого роду, і також

$$k(x) \ge k_0 > 0; \quad q(x) > 0, \quad x \in [0,l].$$

Уведемо позначення:

$$(\varphi,\psi) := \int_0^l \varphi(x)\psi(x)dx, \qquad (15.53)$$

$$a(\varphi,\psi) := \int_0^l [k(x)\varphi'(x)\psi'(x) + q(x)\varphi(x)\psi(x)]dx. \qquad (15.54)$$

Метод Рітца. Нехай $\{\varphi_i\}_{i=1}^n$ – ортогональна система лінійно незалежних диференційовних функцій на $[0,l]$, наприклад алгебраїчних або тригонометричних. Припустимо, що система $\{\varphi_i\}$ повна в $C^2[a,b]$ і

$$\varphi_i(0) = \varphi_i(l) = 0, \quad i = \overline{1,n}. \qquad (15.55)$$

У методі Рітца наближений розв'язок будується у вигляді лінійної комбінації

$$u_n = c_1\varphi_1 + \ldots + c_n\varphi_n, \qquad (15.56)$$

з невідомими коефіцієнти c_i, які знаходимо при розв'язуванні варіаційної задачі. Для крайової задачі (15.51), (15.52) фунціонал F набуває вигляду

$$F(u) = (Lu, u) - 2(r, u) = a(u, u) - 2(f, u).$$

У підсумку для визначення коефіцієнтів одержимо СЛАР [16]

$$\sum_{j=1}^{n} c_j (L\varphi_j, \varphi_i) = (f, \varphi_i), \quad i = \overline{1, n}. \tag{15.57}$$

Оскільки матрицею цієї системи є матриця Грама і система функцій $\{\varphi_i\}$ – лінійно незалежна, то існує єдиний розв'язок, отже, єдиний наближений розв'язок варіаційної задачі. При досить деяких умовах на функції k, q і f доведено [13, 16, 78] збіжність методу Рітца до розв'язку задачі (15.51), (15.52).

Метод Гальоркіна. Цей метод належить до класу проекційних методів. Його можна застосувати до ширшого класу задач, ніж метод Рітца, зокрема, для нелінійних крайових задач. Ідея методу ґрунтується на ортогональності нев'язки $f - Au_n$ до функцій $\varphi_1, ..., \varphi_n$. Справді, якщо система $\{\varphi_i\}$ повна в H, то в просторі H немає відмінного від нуля елемента, ортогонального до всіх φ_i при $i \to \infty$. Тому можна розраховувати на те, що наближений розв'язок u_n, який будується у вигляді (15.56), прямує до точного розв'язку u, оскільки з $(Au - r, \varphi_i) = 0, i = 1, 2, ...$

Отже, для фіксованого n і лінійно незалежної системи функцій $\{\varphi_i\}_{i=1}^{n}$ на $[0, l]$, коефіцієнти c_i знаходяться із системи рівнянь

$$(Au_n, \varphi_i) = (f, \varphi_i), \quad i = \overline{1, n}, \tag{15.58}$$

де матриця системи рівнянь може не бути симетричною, як у методі Рітца.

Метод скінченних елементів. Побудова наближеного розв'язку методом Рітца або Гальоркіна для задачі (15.51), (15.51) зводиться до розв'язування СЛАР порядку n із заповненою матрицею, де $n-$ кількість координатних функцій. Застосування ме-

тоду Ґауса для такої СЛАР вимагає $O(n^3)$ арифметичних операцій. Крім того, при великих n система рівнянь може бути погано обумовленою. Ці методи також накладають обмеження на вибір системи координатних функцій, які залежать від крайових умов.

Різницевий та інтегро-інтерполяційний методи побудови РС породжують СЛАР із тридіагональною матрицею, розв'язування яких методом прогонки зводиться до виконання $O(n)$ операцій. Але у цьому разі наближений розв'язок одержується на сітці Δ_h, а не в аналітичному вигляді. Розв'язок задачі в аналітичному вигляді зі СЛАР із тридіагональною матрицею одержується, якщо спеціальним чином вибрати координатні функції. Нехай φ_i – кусково лінійні функції вигляду:

$$\varphi_i(x) = \begin{cases} 0, \; x \in [0, x_{i-1}) \bigcup (x_{i+1}, l]; \\ 1 + (x - x_i), \; x \in [x_{i-1} x_i]; \\ 1 - (x - x_i), \; x \in [x_i, x_{i+1}], \end{cases}$$

де $i = \overline{1, N-1}, \quad Nh = l.$ Зауважимо, що $\varphi_i(x_i) = 1$ (рис. 15.2).

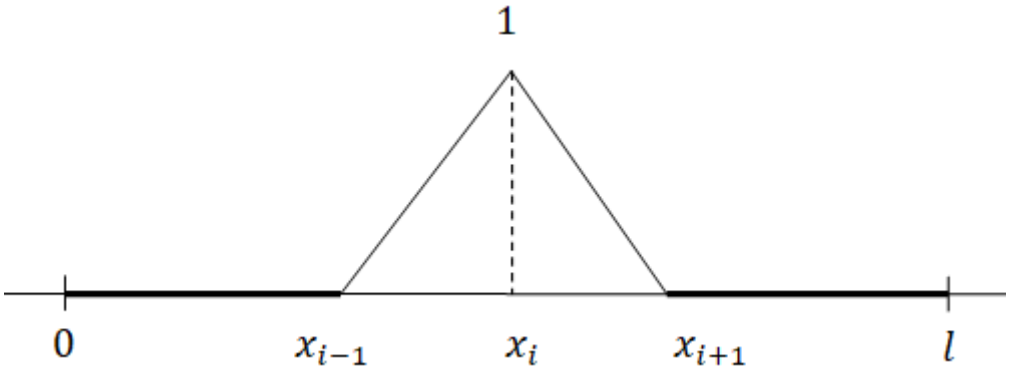


Рис. 15.2. Координатні функції φ_i

Наближений розв'язок $u_{N-1}(x)$ побудуємо у вигляді

$$u_{N-1}(x) = \sum_{i-1}^{N-1} c_i \varphi_i(x).$$

Оскільки $u_{N-1}(x_i) = c_i$ і функція φ_i ортогональна до всіх функцій φ_i, для яких $|i - j| > 1$, тому СЛАР (15.57) для знаходження коефіцієнтів є тридіагональною. Визначником СЛАР служить

визначник Грама, побудований за системою лінійно незалежних функцій φ_i, $i=\overline{1,n}$, тому він не дорівнює нулю і СЛАР має єдиний розв'язок. Коефіцієнти матриці в системі (15.57) обчислюються за формулами

$$(L\varphi_j,\varphi_i)=\int_0^l (L\varphi_j)\varphi_i(x)dx.$$

Доведено, збіжність методу скінченних елементів та показано, що $\|u-v\|=\sqrt{(u-v,u-v)}\le Ch,\quad C=const>0$ [17,67].

Приклади розв'язування типових задач

Задача 1. Побудувати РС і знайти наближене значення розв'язку з кроками 0.1 та 0.01 лінійної крайової задачі

$$u''+(x+1)u'-2u=2,\ 0<x<1;$$
$$u(0)-u'(0)=1,\ u(1)=4.$$

Розв'язування. На підставі (15.24), (15.27) РС для крайової задачі набуває вигляду

$$(3+2h)y_0-4y_1+y_2=2h,$$
$$\left(1-\frac{1}{2}h(1+nh)\right)y_{n-1}-2(1+h^2)y_n\left(1+\frac{1}{2}h(1+nh)\right)y_{n+1}=2h^2,$$
$$n=\overline{1,N-1};\ y_N=4.$$

Таблиця 15.2

Числові розв'язки із кроками $h_1=0.1$ і $h_2=0.01$

x_n	$u(x_n)$	$y_{h_1}(x_n)$	$y_{h_2}(x_n)$
0.0	1.0000	1.0322	1.0330
0.1	1.2100	1.2354	1.2124
0.2	1.4400	1.4598	1.4419
0.3	1.6900	1.7052	1.6914
0.4	1.9600	1.9714	1.9611
0.5	2.2500	2.2584	2.2508
0.6	2.5600	2.5659	2.5606
0.7	2.8900	2.8939	2.8904
0.8	3.2400	3.2423	3.2402
0.9	3.6100	3.6111	3.6101
1.0	4.0000	4.0000	4.0000
	Норма похибки на сітці	0.0322	0.0030

Після вилучення y_2 із першого рівняння одержимо

$$(3+2h-c(1-h(1+h)))y_0-(4+2c(1+h^2))y_1=-2h(1+ch),$$

де $c=1/(1+h(1+h)/2).$

Значення розв'язку одержаної СЛАР з тридіагональною матрицею порядку 10 і 100 в точках $0.1n$ із кроками наведені в таблиці 15.2. Максимальні значення похибки числового розв'язку з кроками 0.1 і 0.01 складає відповідно 0.0322 і 0.0030.

Задача 2. Побудувати РС для нелінійної крайової задачі

$$u''=3u+10u^2+x^2,\ 0<x<1;\quad u(0)=u(1)=0$$

та знайти її розв'язок на сітці з кроками $h=0.01$ і $h=0.001$.

Розв'язування. Запишемо відповідну РС у вигляді

$$y_{n-1}-2y_n+y_{n+1}=h^2(3y_n+10y_n^2+nh),\quad n=\overline{1,N-1};\ y_0=y_N=0.$$

Уведемо позначення:

$$A=\begin{bmatrix}-2&1&0&\cdots&0&0&0\\1&-2&1&\cdots&0&0&0\\\cdots&\cdots&\cdots&\cdots&\cdots&\cdots&\cdots\\0&0&0&\cdots&1&-2&1\\0&0&0&\cdots&0&1&-2\end{bmatrix},\ D(Y)=\begin{bmatrix}y_1&&0\\&\ddots&\\0&&y_{N-1}\end{bmatrix},$$

$$Y=\begin{bmatrix}y_1\\\dots\\y_{N-1}\end{bmatrix},\ F(Y)=\begin{bmatrix}3y_1+10y_1^3+h^2\\3y_2+10y_2^3+4h^2\\\dots\dots\dots\dots\\3y_{N-1}+10y_{N-1}^3+(N-1)^2h^2\end{bmatrix}.$$

Тоді СЛАР для обчислення вектора $\Delta^{(k)}=\left[\Delta_1^{(k)},\dots,\Delta_{N-1}^{(k)}\right]^T$ для уточнення наближення $Y^{(k)}$ набуде вигляду

$$\left[A-(3I+20D(Y^{(k)}))h^2\right]\Delta^{(k)}=h^2F(Y^{(k)})-AY^{(k)}.$$

Матриця системи тридіагональна, оскільки недіагональні елементи дорівнюють 1, а діагональні із протилежним знаком дорівнюють $2+(3+20y_i)h^2$. Наступне наближення $Y^{(k+1)}=Y^{(k)}+\Delta^{(k)}$.

Результати обчислень у вузлах $x_i=0.1i$ із початковим вектором $Y^{(0)}$ наведені в табл 15.6. Обчислення проводилися на сітках із кроком $h=10^{-k}$, $k=1,2,3$. Ітерації припинялись, якщо всі координати вектора $\Delta^{(k)}$ ставали за модулем менше 10^{-4}.

Таблиця 15.6

x_i	h = 0.1	h = 0.01	h = 0.001
0.1	-0.0058	-0.0058	-0.003821
0.2	-0.0116	-0.0118	-0.011807
0.3	-0.0174	-0.0176	-0.017625
0.4	-0.0223	-0.0230	-0.023061
0.5	-0.0274	-0.0276	-0.027630
0.6	-0.0302	-0.0304	-0.030403
0.7	-0.0303	-0.0305	-0.030514
0.8	-0.0265	-0.0276	-0.026629
0.9	-0.0170	-0.0170	-0.017106

Задача 3. Побудувати РС розподілу температури в тонкому однорідному стержні довжиною 1, який складається з двох частин із різними коефіцієнтами теплопровідності k_1 і k_2, сталим коефіцієнтом теплообміну із зовнішнім середовищем q і сталою щільністю джерел f. На кінцях стержня підтримується стала температура μ_0 і μ_1.

Розв'язування. Розподіл температури у стержні є розв'язком крайової задачі (15.37) із розривною функцією $k(x)$, тому застосуємо інтегро-інтерполяційний метод. Розглянемо два випадки. У першому (рис. 15.3) крок сітки $h = 1/N$ такий, що $c = mh, m \in N$.

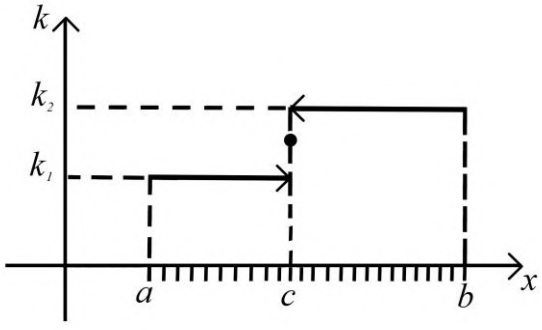


Рис. 15.3

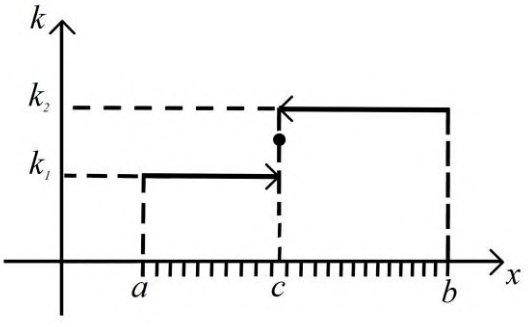


Рис. 15.4

Згідно з формулами (15.39) маємо:

$$a_n = \left(\frac{1}{h} \int\limits_{x_{n-1}}^{x_n} \frac{dx}{k_1} \right)^{-1} = \frac{k_1 h}{x_n - x_{n-1}} = k_1,\ n = \overline{1,m}; \qquad (15.49)$$

$$a_n = \left(\frac{1}{h} \int\limits_{x_{n-1}}^{x_n} \frac{dx}{k_2} \right)^{-1} = k_2,\ n = \overline{m+1, N}; \qquad (15.50)$$

$$d_n = q, \quad \varphi_n = f, \quad n = \overline{1, N}.$$

РС у цьому випадку набуває вигляду

$$-\frac{k_1}{h^2}(y_{n-1} - 2y_n + y_{n+1}) + qy_n = \varphi, \quad n = \overline{1, m-1};$$

$$-\frac{k_2}{h^2}(y_{m+1} - y_m) + \frac{k_1}{h^2}(y_m - y_{m-1}) + qy_m = \varphi, \quad n = m;$$

$$-\frac{k_2}{h^2}(y_{n-1} - 2y_n + y_{n+1}) + qy_n = \varphi, \quad n = \overline{m+1, N-1}.$$

Урахувавши крайові умови одержимо, СЛАР із тридіагональною матрицею

$$\begin{aligned}
&(2k_1 + h^2 q)y_1 - k_1 y_2 = h^2\varphi + k_1\mu_0, \quad n = 1;\\
&-k_1 y_{n-1} + (2k_1 + h^2 q)y_n - k_1 y_{n+1} = h^2\varphi, \quad n = \overline{2, m-1};\\
&-k_1 y_{m-1} + (k_1 + k_2 + h^2 q)y_m - k_2 y_{m+1} = h^2\varphi, \quad n = m; \qquad (15.51)\\
&-k_2 y_{n-1} + (2k_2 + h^2 q)y_n - k_2 y_{n+1} = h^2\varphi, \quad n = \overline{m+1, N-2};\\
&(2k_2 + h^2 q)y_{N-2} - k_2 y_{N-1} = h^2\varphi + k_2\mu_1, \quad n = N-1.
\end{aligned}$$

Нехай тепер $mh < c < (m+1)h$ (рис. 15.4). Тоді $c = x_m + \xi h$, де $0 < \xi < 1$. Коефіцієнти a_n для $n = \overline{1, m}$ і $n = \overline{m+2, N}$ обчислюються за формулами (15.49) і (15.50) відповідно. Коефіцієнт

$$a_{m+1} = \left(\frac{1}{h} \int\limits_{x_m}^{c} \frac{dx}{k_1} + \frac{1}{h} \int\limits_{c}^{x_{m+1}} \frac{dx}{k_2} \right)^{-1} = \frac{hk_1}{c - x_m} + \frac{hk_2}{x_{m+1} - c} = \frac{(1-\xi)k_1 + \xi k_2}{\xi(1-\xi)} =: \bar{k}.$$

У цьому випадку одержується СЛАР, яка відрізняється від СЛАР (15.50) рівняннями з номерами m і $m+1$, які набувають вигляду

$$\begin{aligned}
&-k_1 y_{m-1} + (k_1 + \bar{k} + h^2 q)y_m - \bar{k} y_{m+1} = h^2\varphi, \quad n = m+1;\\
&-\bar{k} y_m + (k_1 + \bar{k} + h^2 q)y_{m+1} - k_2 y_{m+2} = h^2\varphi, \quad n = m+2.
\end{aligned}$$

Задача 4. Методом Рітца побудувати наближений розв'язок крайової задачі

$$-((t^2+1))' + \frac{2}{t^2} z = \frac{2}{t}, \quad 1 < t < 2;$$

$$z(1) = 2, \quad z(2) = 2.5.$$

Точний розв'язок задачі $u(t) = t + 1/t.$

Розв'язування. Виконавши заміну

$$t = x + 1, \; z = 2.5x - 2(x-1) + 4 = 2 + 0.5x + 4,$$

одержимо крайову задачу

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

$$u(0) = u(1) = 0,$$

де $v(x) = x^2 + 2x + 2$, $q(x) = 2(x+1)^{-2}$, $q(x) = -x - 1 + (x+1)^{-1} -$ $-3(x+1)^{-2}$, $0 \le x \le 1$. За формулами (15.53), (15.54) маємо:

$$F(u) = (Lu, u) - 2(f, u),$$

$$(Lu, u) = \int_0^1 \left[(x^2+2x+2)(u')^2 + 2(x+1)^{-2} u^2 \right] dx,$$

$$(f, u) = \int_0^1 \left(-x - 1 + \frac{u}{x+1} - \frac{3}{(x+1)^2} \right) u dx.$$

Координати функції $\varphi_1(x) = x(1-1)$ і $\varphi_2(x) = x^2(1-x)$ задовольняють нульові крайові умови і лінійно незалежні. Коефіцієнти c_1 і c_2 наближеного розв'язку

$$u_2(x) = c_1 x(1-x) + c_2 x^2(1-x)$$

знаходяться із СЛАР

$$(L\varphi_1, \varphi_2)c_1 + (L\varphi_1, \varphi_2)c_2 = (r, \varphi_1),$$

$$(L\varphi_1, \varphi_2)c_1 + (L\varphi_1, \varphi_2)c_2 = (r, \varphi_2).$$

Обчисливши відповідні інтеграли, одержимо СЛАР

$$1.1645c_1 + 0.6807c_2 = -0.7952,$$

$$0.6807c_1 + 0.5597c_2 = -0.1332,$$

звідки знаходимо $c_1 = -1.8799, c_2 = 2.0479$ і наближений розв'язок $u_2(x) = -1.8799x(1-x) + 2.0479x^2(1-x)$ або $u_2(x) = (2.0479x - 1.8799)(1-x)x.$ Похибка розв'язку $u_2(x)$ у

точках $0.25k, k=\overline{1,3}$ дорівнює 0.0389б 0.1730 і 0.0109 відповідно.

Задача 5. Задача 5. Методом Гальоркіна знайти наближений розв'язок $u_2(x)$, коли $n=2,$ крайової задачі

$Lu := u'' - xu' + 2u = 2, \quad 0 < x < 1;$

$u(0) + u'(0) = 1, \quad u'(1) = 2,$

точний розв'язок якої $u(x) = x^2 + 1.$

Розв'язування. У цьому випадку оператор L не симетричний, тобто $(Lu, v) \neq (u, Lv),$ тому метод Рітца застосувати не можна. Оскільки крайові умови неоднорідні, то наближений розв'язок побудуємо у вигляді

$$u_2(x) = \varphi_0(x) + c_1\varphi_1(x) + c_2\varphi_2(x),$$

де $\varphi_0(x) = ax + b$ і задовольнятиме крайові умови, а $\varphi_1(x) = 1 + a_1x + b_2x^2,\ \varphi_2(x) = a_2x^2 + b_2x^3.$

Значення $a = 2$ і $b = -1$ знаходимо із системи рівнянь $a + b = 1, \quad a = 2,$ тобто $\varphi_0(x) = 2x - 1.$ Із умов (15.55) маємо:

$$\varphi_1(x) = 1 - x + x^2/2, \quad \varphi_2(x) = -3x^2/2 + x^3.$$

Система (15.58) набуває вигляду

$$\frac{41}{24}c_1 - \frac{1}{8}c_2 = \frac{25}{12},$$

$$\frac{9}{20}c_1 - \frac{27}{140}c_2 = \frac{1}{10},$$

звідки $c_1 = 108/109, \quad c_2 = -1022/327.$ Отже, $u_2(x) = \varphi_0(x) +$

$$+\frac{108}{109}\varphi_1(x) - \frac{1022}{327}\varphi_2(x) = (-3 + 330x + 3128x^2 - 1022x^3)/327.$$

Похибка наближеного розв'язку в точці 0.5 дорівнює

$$u_2(0.5) - u(0.5) = 1.4006 - 1.2500 = 0.1506.$$

Завдання та запитання для самостійної роботи

1. Проілюструвати метод уточнення початкових даних на прикладі крайової задачі

$$u'' = f(x, u, u'),\ 0 < x < 1;\ u(0) = \mu_0,\ u(1) = \mu_1.$$

2. Побудувати різницеву схему другого порядку для таких крайових задач:

1) $u'' - u' = f(x),\ 0 < x < 1;\ \ u(0) = u(1),\ u'(0) = u'(1);$

2) $u'' + p(x)u' + q(x)u = r(x),\ 0 < x < 1;$

$$u(0) = a,\ \int_0^1 u(x)dx = b;$$

3) $u^{(4)} = f(x),\ 0 < x < 1;$

$u(0) = \mu_0;\ -u'(0) = \nu_0,\ u(1) = \mu_1\, u'(1) = \nu_1.$

3. Побудувати РС і знайти її розв'язок із різними кроками для одновимірної задачі конвенкції-дифузії

$$-au'' + bu' = 0,\ 0 < x < 1;\ u(0) = 0,\ u(1) = 1.$$

Порівняти одержаний розв'язок з точним розв'язком

$$u(x) = \frac{1 - e^{-Rz}}{1 - e^{-R}},\ R = \frac{b}{a}.$$

4. Побудувати РС, яка апроксимує лінійну крайову задачу (15.21), (15.22) з четвертим порядком.

5. Різницевим методом знайти наближений розв'язок нелінійної крайової задачі

$$u'' = -u + 2(u')^2 / u,\ -1 < x < 1,$$

$$u(-1) = u(1) = \left(e + e^{-1}\right)^{-1} = 0.324027137.$$

Порівняти одержані результати з точним розв'язком $u(x) = \left(e^x + e^{-x}\right)^{-1}$ крайової задачі. За початкове наближення в методі Ньютона взяти $y_n^{(0)} = \left(e + e^{-1}\right)^{-1}$, ітерації припинити, коли $\max_n \left| y_n^{(k+1)} - y_n^{(k)} \right| \le 10^{-10}$.

6. Побудувати РС другого порядку для таких крайових задач:

1) $u'' - u' = f(x), 0 \le x \le 1;$

$u(0) = u(1);\quad u'(0) = u'(1);$

2) $u'' + p(x)u' + q(x)u = r(x),\quad 0 \le x \le 1;$

$u(o) = a;\quad \int_0^1 u(x)dx u'(0) = b;$

3) $u^{IV} = f(x),\quad 0 \le x \le 1;$

$u(0) = \mu_0;\quad -u'(0) = \nu_0;\quad u(1) = \mu_1;\quad u'(1) = \nu_1.$

7. Для яких значень α, β і γ різницева схема

$$-\frac{1}{h^2}(y_{n+1} - 2y_n + y_{n-1}) + (2y_{n+1} - \beta y_n + \gamma y_{n-1}) = f_n + \frac{h^2}{12} f_n'',\ y_0 = y_N = 0,$$

апроксимує крайову задачу $-u''+u=f(x);\quad u(0)=u(1)=0$ із четвертим порядком?

8. Побудувати РС для задачі деформації пружної струни під дією поперечного навантаження, яка описується диференціальним рівнянням

$$-u''=a^2(u')^2+1,\quad 0<x<1,$$

з крайовими умовами $u(0)=u(1)=0..$

9. Дослідити на стійкість РС:

а) $-\frac{1}{h^2}(y_{n+1}-2y_n+y_{n-1})=f_n,\ y_0=y_1,\ y_N=0,\ Nh=1;$

б) $-\frac{1}{h^2}(y_{n+1}-2y_n+y_{n-1})=f_n,\ y_0=y_N=0.$

10. Побудувати РС для крайової задачі

$u^{(4)}+xu^{(3)}-2u''+(x+1)u'-x^2u=(3-x^2)\sin x+\cos x,$

$u(0)=0,\ u'(0)=1, u(\frac{\pi}{2})=1, u'(\frac{\pi}{2})=0,$ точний розв'язок якої $u(x)=\sin x.$

11. Побудувати РС другого порядку для рівняння

$$-u''+p(x)u=f(x),\quad 0<x<1,$$

з крайовими умовами:

а) $u'(0)=u(1),\quad u'(1)=u(0);$

б) $u'(0)=u'(1),\quad u(0)=u(1);$

в) $2u'(0)=u(0)+u(1),\quad u'(1)=\mu.$

12. Апроксимувати за двома точками з другим порядком крайову умову $u'-4u=2,$ задану при $x=2,$ для рівняння $u''=\sin\frac{\pi x}{4}+2.$

13. Дослідити стійкість РС

$$h^{-2}(y_{n-1}-2y_n+y_{n+1})=-f(x_n),\quad n=\overline{1,N-1};$$
$$y_0=y_N=0,$$

і показати, що при $h\to 0$ число обумовленості алгебраїчної системи для знаходження y_n має порядок $O(h^{-2}).$

14. При яких a,b,c РС

$$-h^{-2}(y_{n-1}-2y_n+y_{n+1})+(ay_{n-1}+by_n+cy_{n+1})=f_n+\frac{h^2}{12}f''(x_n),$$
$$y_0=y_N=0,$$

апроксимує з четвертим порядком крайову задачу

$-u'' + u = f(x), \quad u(0) = u(1) = 0$?

15. Різницевим методом і методом уточнення початкових даних розв'язати з кроком $h = 0.01$ крайову задачу

$$u'' + u + x = 0, \quad x \in (0,1);$$
$$u(0) = 0, \quad u(1) = 0.5.$$

Порівняти одержаний числовий розв'язок з точним розв'язком $u(x) = (\sin x)/\sin 1 - 0.5x$ крайової задачі.

16. Математичною моделлю малих коливань плоского маятника під дією періодичної сили служить диференціальне рівняння $u'' + \omega^3 u = \sin \nu t,$ де $u(t)$ - кутове відхилення, $\omega^2 = g/l, \quad \nu$ – частота зовнішньої сили. Побудувати РС другого порядку, якщо для розв'язку задано крайові умови:

а) $u(0) = -1, \quad u(1) = 1;$

б) $u'(0) = 0, \quad u(1) = 1;$

в) $u(0) = u(1), \quad u'(0) = u'(1).$

17. Електростатичний потенціал між двома сферами радіусів $r = 1$ і $r = 4$ визначається як розв'язок крайової задачі

$$\frac{d^2u}{dr^2} + \frac{2}{r}\frac{du}{dr} = 0,\ u(1) = 50,\ u(4) = 100.$$

Знайти числовий розв'язок, розбивши відрізок $[1,4]$ на 30 частин.

18. Побудувати РС з кроком h для рівняння Дюфінга

$$u'' + u = \varepsilon u^3,\ \varepsilon = const > 0,$$

з крайовими умовами $u(0) = u(1) = \mu.$ Записати СЛАР для уточнення розв'язку РС методом Ньютона.

19. Побудувати РС порядку 2 та відповідну СЛАР для наближеного обчислення числового розв'язку для моделі поширення епідемії в задачі 24 розділу 12.

20. Дослідити стійкість РС (15.24) для КЗ

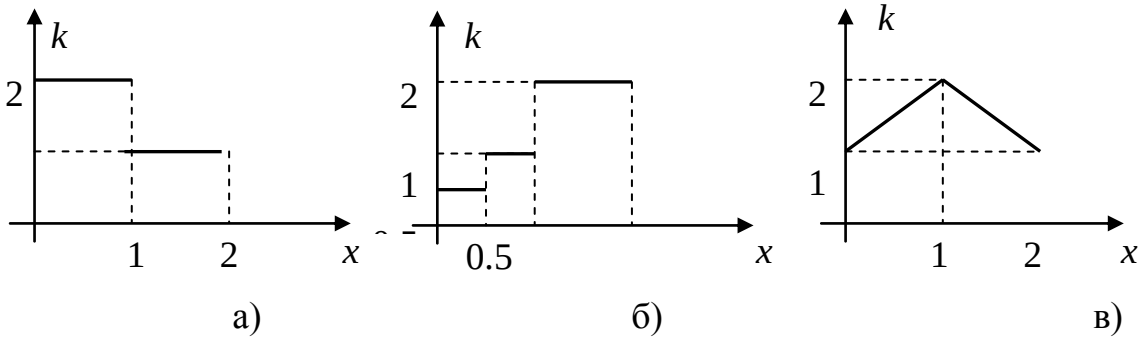
$$u'' - \omega^2 u = f(x), 0 < x < l, \omega > 0;$$
$$u(0) = \alpha, \quad u(l) = \beta.$$

21. Побудувати РС і знайти числовий розв'язок з кроком 0.1 для КЗ

$$u'' = -\gamma e^u, 0 < x < 4, u(0) = u(4) = 0,$$

де параметр $\gamma \le \gamma_c \approx 3.5.$ Для $\gamma < \gamma_c \approx 3.5$ задача має два розв'язки, для $\gamma = \gamma_c$ розв'язок єдиний, коли $\gamma > \gamma_c,$ то розв'язку не існує [11].

22. Інтегро-інтерполяційним методом побудувати СЛАР для знаходження числового розв'язку крайової задачі $(k(x)u')' = 0, u(0) = \mu_0,\ u(2) = \mu_1$, де функція $k(x)$ задана на рис. 15.5.



а) б) в)

Рис. 15.5

23. Побудувати РС і знайти числові розв'язки для $N = 10$, 25 і 50, порівняти одержані результати між собою, а також з точним розв'язком $u^*(x)$ для крайових задач:

1) $u'' - u = \sin(2\pi x), 0 < x < 1, u(0) = u(1) = 0;$

$u^*(x) = \sin(2\pi x)/(1 + 4\pi^2);$

2) $u'' - (2x - 1)u' - 2u = 0, 0 < x < 1, u(0) = u(1) = 1;$

$u^*(x) = \exp(x(x - 1)).$

24. Вертикальне відхилення закріпленого кабелю є розв'язком крайової задачі

$$u'' = \mu\sqrt{1 + (u')^2},\ 0 < x < l;\quad u(0) = \mu_0,\quad u(l) = \mu_1,$$

де μ, α і β - додатні сталі.

1) Знайти розв'язок задачі вигляду $u(x) = A + \mu^{-1}ch(\mu x + B)$, де A і B визначаються крайовими умовами.

2) Побудувати РС (15.24), знайти числовий розв'язок для $N = 10, 20, 40$.

3) Порівняти одержані числові розв'язки між собою і з точним розв'язком.

25. Побудувати РС другого порядку для ЗДР $u'' - u = -100x, 0 < x < 1,$

з інтегральними умовами

$$\int_0^1 u(x)dx = 0,\ \int_0^1 xu(x)dx = 1.$$

Знайти точний розв'язок $u^*(x)$ задачі та похибку $\max\limits_{i=0,N} |y(x_i) - u^*(x_i)|$ для $N = 4, 8$ і 16.

Короткі відомості про вчених, які згадуються у посібнику

Абель Нільс Генрік (норв. *Niels Henrik Abel*, 1802–1829 рр.) – норвезький математик. Довів у 1824 р. теорему про те, що для рівнянь п'ятого і вищих степенів не існує загальної формули для вираження коренів через коефіцієнти в радикалах.

Адамс Джон Кауч (англ. *John Couch Adams*, 1819–1892 рр.) – англійський астроном, математик і механік. У 1845 р. передбачив планету Нептун на основі дослідження збурень руху планети Уран. У 1882 р. використав метод Ньютона для розв'язування проблеми Кеплера. Розробив багатокроковий метод числового інтегрування диференціальних рівнянь для дослідження капілярних процесів. Розрахував сталу Ейлера $\gamma = 0.577215665...$ з точністю до 272 десяткових знаків.

Бернштейн Сергій Натанович (1880–1968 рр.) розв'язав дев'ятнадцяту і двадцяту проблеми Гільберта, розвинув ідеї П.Л. Чебишева про наближення функцій многочленами, йому належить термін «конструктивна теорія функцій». Автор базисних функцій, названих його іменем.

Буль Джордж (англ. *George Boole*, 1815–1864 рр.) – англійський математик і філософ. Один із засновників математичної логіки, застосував символьний метод у логіці. До обчислювальної математики належить праця «Трактат про скінченні різниці» (1860 р.). Його іменем названа квадратурна формула із рівновіддаленими вузлами шостого порядку точності.

Бутчер Джон (англ. *John Charles Butcher*, нар. 1933 р.) – новозеландський математик. Фахівець з розробки методів Рунге–Кутти та багатокрокових методів для задачі Коші. Запропонував таблицю для запису коефіцієнтів цих методів. Координатор клубу з обміну інформацією про методи Рунге–Кутти http://www.jcbutcher.com/rkclub/

Валліс Джон (англ. *John Wallis*, 1616–1703 рр.) – англійський математик, один з попередників математичного аналізу. Ввів терміни: мантиса, неперервний дріб, символ ∞, а в 1656 р. – термін "інтерполювання". Вивів рекурентні співвідношення для відповідних дробів неперервного дробу, а також формулу для обчислення числа $\pi/2$.

Вронський Юзеф (польськ. *Józef Hoene-Wroński*, 1776–1853 рр.) – польський математик. У 1812 р. уперше ввів поняття функціонального визначника (вронськіана).

Галуа Еварист (фр. *Évariste Galois*, 1811–1832 рр.) – французький математик. Запровадив термін “група”, використав властивості груп для розв’язування проблеми розв’язності алгебраїчних рівнянь (1830 р.), розвинувши дослідження Нільса Абеля.

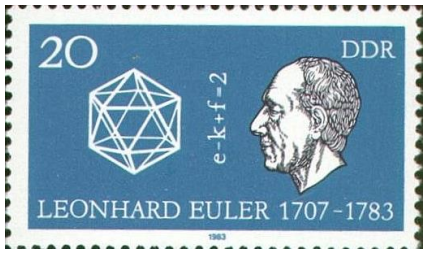
Ґаус Іоган Карл Фрідріх (нім. *Johann Carl Friedrich Gaus*, 1777–1855 рр.) – німецький математик, астроном і фізик. Побудував інтерполяційну формулу, яка носить його ім’я. Метод Ґауса – класичний метод розв’язування СЛАР, уперше опублікований у 1801 р. У 1814 р. Ґаус опублікував квадратурну формулу найвищого алгебраїчного степеня точності, що було одним з найбільших досягнень обчислювальної математики у XIX столітті. У 1795 р., незалежно від Лежандра, розробив метод найменших квадратів для розв’язування задачі аналізу топографічних даних в астрономічних обчисленнях.

Ґір Чарль Вільям (англ. *Charles William Gear*, нар. у 1935 р.) – британсько-американський математик. Фахівець з обчислювальної математики, комп’ютерної графіки та комп’ютерних наук. Відомий розробкою методу числового розв’язування жорстких систем диференціальних рівнянь (1966 р.).

Горнер Вільям Джордж (англ. *William George Horner*, 1786–1837 рр.) − англійський математик. Основні праці з алгебри. У 1819 р. опублікував метод наближеного обчислення дійсних коренів многочлена, названий методом Руффіні–Горнера. Його ім’ям названа схема виділення лінійного множника алгебраїчного многочлена.

Далквіст Гермунд (англ. *Germund Dahlquist*, 1925–2005 рр.) – шведський математик, зробив значний внесок в аналіз стійкості багатокрокових різницевих методів для звичайних диференціальних рівнянь (бар’єр Далквіста щодо порядку стійкого лінійного k-крокового різницевого методу).

Декарт Рене (фр. *René Descartes*, 1596–1650 рр.) – французький філософ, фізик, математик, засновник аналітичної геометрії. Запровадив систему координат, яка носить його ім’я, ввів поняття змінної величини і функції, сформулював «правило знаків» для визначення числа додатних коренів алгебраїчного рівняння й основну теорему алгебри.



Ейлер Леона́рд (нім. *Leonhard Euler*, 1707–1783 рр.) – швейцарський, російський та німецький математик і фізик. Автор 866 наукових публікацій, зокрема в галузях математичного аналізу, диференціальної геометрії, теорії чисел, теорії графів, наближених обчислень, небесної механіки, математичної фізики, оптики, балістики, теорії музики. У його працях набули поширення символи $f(x), e, \pi, i, \Sigma$. В обчислювальній математиці відомий метод Ейлера числового інтегрування початкової задачі для звичайних диференціальних рівнянь (1768 р.), наближення функцій ланцюговими дробами.

Ейткен Олександр (англ. *Alexander Craig Aitken*, 1895–1967 рр.) – новозеландський математик. Запропонував оцінку похибки при невідомому порядку точності, схему побудови інтерполяційного многочлена (схема Ейткена), метод побудови ітерацій вищого порядку (Δ^2 – процес), узагальнив метод найменших квадратів.

Ерміт Шарль (фр. *Charles Hermite*, 1822–1901 рр.) – французький математик. Праці з теорії чисел, алгебри та теорії еліптичних функцій. Звів загальне алгебраїчне рівняння п'ятого степеня до вигляду, що розв'язується в еліптичних модулярних функціях, увів новий клас ортогональних многочленів (многочлени Ерміта), інтерполяційних многочленів із кратними вузлами, довів трансцендентність числа e (1873 р.).

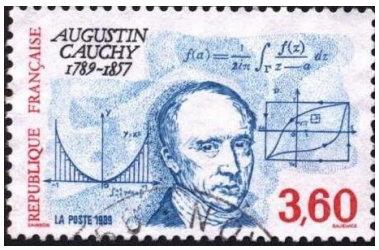
Зейдель Філіп Людвіг (нім. *Philipp Ludwig von Seidel*, 1821–1896 рр.) – німецький математик і астроном. Його іменем названо ітераційні методи розв'язування систем лінійних і нелінійних рівнянь, у яких використовуються вже знайдені наближення (1874 р.). Метод Ґауса – Зейделя можна розглядати як модифікацію методу Якобі.

Йордан Вільгельм (нім. *Wilhelm Jordan*, 1842–1899 рр.) – німецький геодезист. Автор методу Ґауса–Йордана (називають ще методом Ґауса–Жордана) розв'язування СЛАР із квадратною матрицею, знаходження оберненої матриці, координат вектора у заданому базисі та рангу матриці.

Канторович Леонід Віталійович (1912–1986 рр.) – радянський математик і економіст. Уперше застосував функціональний аналіз в обчислювальній математиці. Розвинув загальну теорію наближених методів, побудував ефективні методи розв'язування операторних рівнянь, зокрема метод найшвидшого спуску і метод Ньютона для таких рівнянь (1948 р.).

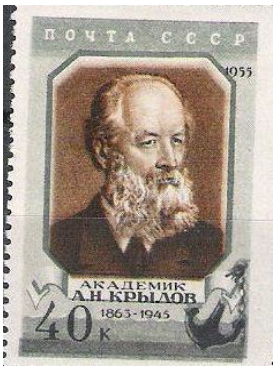
Котес Роджер (англ. *Roger Cotes*, 1682–1716 рр.) – англійський математик і філософ. Розробив метод, який був розвинениий як метод найменших квад-ратів. У праці "Гармонія мір" (1722 р.) наведено формули розкладу

$x^n \pm a^n$ та коренів рівняння $x^n - 1 = 0$, а також формулу для наближеного обчислення визначеного інтеграла на рівномірній сітці.



Коші Огюсте́н Луї (фр. *Augustin Louis Cauchy*, 1789–1857 рр.) – французький математик, член Паризької академії наук. Опублікував понад 800 праць з арифметики і теорії чисел, алгебри, математичного аналізу, диференціальних рівнянь, теоретичної і небесної механіки, математичної фізики тощо. Неявний метод розв'язування початкової задачі наведено в його праці 1824 р. при застосуванні теореми про середнє для оцінки похибок. Тоді ж ним застосовано метод послідовних наближень ("…on la résoudra facilement par des approximatons successives[1]…"). У 1829 р. запропонував багатовимірне узагальнення методу Ньютона.

Крамер Габріель (нім. *Gabriel Cramer*, 1704–1752 рр.) – швейцарський мате-матик, один із творців лінійної алгебри. Метод Крамера розв'язування СЛАР з квадратною матрицею опубліковано у праці "Introduction à l'analyse des lignes courbes algébraique" в 1750 р.



Крилов Олексій Миколайович (1863–1945 рр.) – російський математик. Запропонував метод розв'язування вікового рівняння. Одержав низку результатів щодо раціональної організації числових методів. Автор першої у світовій практиці книги з числових методів «Лекции о приближенных вычислениях» (1911 р.).

Кутта Мартін Вільге́льм (нім. *Martin Wilhelm Kutta*, 1867–1944 рр.) – німецький математик, співавтор відомої сім'ї методів наближеного інтегрування звичайних диференціальних рівнянь (методи Рунге–Кутти). Також відомий завдяки аеродинамічній поверхні Жуковського–Кутти й аеродинамічній умові Кутти.

Лаґранж Жозеф Луї (фр. *Joseph Louis Lagrange*, 1736–1813 рр.) – французський математик і механік. Автор інтерполяційної формули для наближення функції многочленом. У трактаті "Про розв'язування числових рівнянь" (1767 р.) зробив припущення, що не всі алгебраїчні рівняння вище четвертого степеня розв'язуються в радикалах. Знайшов спосіб наближеного обчислення коренів алгебраїчного рівняння за допомогою неперервних дробів і розробив метод відокремлення коренів таких рівнянь.

[1] …легко розв'язати за допомогою послідовних наближень…

Лалан Леон Луї Кретє'н (фр. *Léon-Louis Chrétien-Lalanne*, 1811–1892 рр.) − французький інженер. Винахідник обчислювальних пристроїв. У 1840 р. запропонував скорочену систему числення за основою три.

Лежандр Адрієн-Марі (фр. *Adrien-Marie Legendre*, 1752–1833 рр.) – французький математик, з 1783 р. член Французької академії наук. Уперше дав повний виклад теорії чисел, у теорії геодезичних вимірювань першим відкрив і застосував в обчисленнях метод найменших квадратів. Його іменем названі квадратурних формул Ґауса–Лежандра.

Лейбніц Ґотфрід Вільгельм (нім. *Gottfried Wilhelm Leibniz*, 1646–1716 рр.) – провідний німецький філософ, логік, математик, фізик. Заклав основи двійкової системи числення. Створив першу механічну лічильну машину, здатну виконувати додавання, віднімання, множення й ділення чисел, а також добування коренів і піднесення до степеня. Незалежно від Ньютона створив диференціальне й інтегральне числення. У 1684 р. опублікував першу у світі велику працю з диференціального числення: «Новий метод максимумів і мінімумів». У 1675 р. увів символ для запису інтеграла.

Ліпшіц Рудольф Отто Сігізмунд (нім. *Rudolf Otto Sigismund Lipschitz*, 1832–1903 рр.) – німецький математик. Учень Діріхле. Основні праці в галузі математичного аналізу, теорії диференціальних рівнянь, алгебри. Його іменем названа умова обмеження на поведінку приросту функції.

Ньютон Ісаа́к (англ. *Sir Isaac Newton*, 1643–1727 рр.) − видатний англійський учений, який заклав основи математичного аналізу і класичної фізики. Для розв'язування нелінійних рівнянь 1669 р. він розробив метод обчислення коренів кубічного рівняння, в основі якого знаходиться ітераційний процес лінеаризації. Його іменем названо інтерполяційні многочлени зі скінченними і поділеними різницями. Квадратурні формули Ньютона–Котеса застосовуються для наближеного обчислення визначених інтегралів на рівномірній сітці.

Остроградський Михайло Васильович (1801–1861 рр.) – український математик, механік і фізик. У курсі лекцій з алгебраїчного і трансцендентного аналізу (1836 і 1837 рр.) навів нові ідеї та методи в теорії алгебраїчних рівнянь, одержаних Лагранжем, Коші, Штурмом, Ґаусом, Абелем.

Рафсон Джон (англ. *Joseph Raphson*, 1648–1715 рр.) − англійський учений, який у 1690 р. сформулював ідею Ньютона ітераційного розв'язування нелінійних рівнянь для випадку многочлена довільного степеня у формі, близькій до сучасного вигляду. Тому цей метод ще називають методом Ньютона–Рафсона.

Річардсон Льюїс Фрай (англ. *Lewis Fry Richardson*, 1881–1953 рр.) – англійський геофізик. Запропонував поправку для уточнення значення інтеграла, обчисленого на сітці з різними кроками, метод розв'язування СЛАР – модифіковані ітерації Річардсона (1910 р.). Застосував метод скінченних різниць наближеного інтегрування диференціальних рівнянь для прогнозу погоди.

Рітц Вальтер (нім. *Walter Ritz*, 1878–1909рр.) – швейцарський математик і фізик-теоретик. Запропонував у 1909 р. метод наближеного розв'язування варіаційних задач (метод Рітца).

Рунге Карл Давид Тольме (нім. *Carl David Tolmé Runge*, 1856–1927 рр.) − німецький математик, фізик. Співавтор відомої сім'ї методів наближеного інтегрування ЗДР (методи Рунге–Кутти). В обчислювальній математиці використовується правило Рунге оцінки похибок. Феномен Рунге – проблема, що виникає в обчислювальній математиці при використанні поліноміальної інтерполяції за допомогою поліномів високих порядків (функції Рунге).

Самарський Олександр Андрійович (1919 – 2008) – російський математик, спеціаліст у галузі обчислювальної математики, математичної фізики, теорії математичного моделювання. Створив теорію операторно-різницевих схем і загальної теорії стійкості таких схем. Співавтор класичних підручників із числових методів.

Сімпсон Томас (англ. *Thomas Simpson*, 1710–1761 рр.) – англійський математик. У 1743 р. вивів квадратурну формулу наближеного обчислення визначеного інтеграла, названу його іменем. У 1740 р. у сучасному вигляді сформулював метод Ньютона як ітераційний метод розв'язування нелінійних рівнянь і системи двох нелінійних рівнянь.

Стевін Сімон (нід. *Simon Stevin*, 1548–1620 рр.) – Нідерландський математик та інженер. У трактаті «Десята» (*De Thiende*) у 1585 р. узагальнив поняття цілого й дробового десяткового числа, по суті, винайшовши дійсні числа, використав від'ємні числа, ввів дробові показники степеня і многочлен однієї змінної.

Стефенсен Йохан Фредерік (англ. *Johan Frederik Steffensen*, 1873–1961 рр.) – датський математик, статистик і актуарій. Провів дослідження із застосування скінченних різниць та інтерполяції. Розробив ітераційний метод другого прядку точності розв'язування нелінійних рівнянь (1933 р.), який носить його ім'я.

Тихонов Андрій Миколайович (1906−1993 рр.) – російський математик і геофізик. Автор праць з теорії диференціальних рівнянь, обчислювальної математики, математичної фізики. Заклав основи теорії та методів розв'язування некоректних задач (1943 р.).

Уілкінсон Джеймс Гарді (англ. *James Hardy Wilkinson*, 1919−1986 рр.). Учений у галузі обчислювальної математики, прикладної математики та інформатики. Премія Тюрінга (1970 р.) за дослідження алгоритмів обчислювальної математики для високошвидкісних комп'ютерів і лінійної алгебри. Автор методики оберненого аналізу похибок при розв'язуванні рівнянь і систем.

Фібоначчі – псевдонім італійського математика Леонардо Пізанського (італ. *Leonardo Pisano*, близько 1170 р. – близько 1250 р.). Його іменем названі числа, які мають застосування у багатьох прикладних задачах, зокрема при оцінці швидкості збіжності методу січних.

Фур'є Жан Батист Жозеф (фр. *Jean Baptiste Joseph Fourier*, 1768–1830 рр.) – французький математик і фізик, започаткував використання рядів Фур'є для розв'язування задач математичної фізики. У 1818 р. довів квадратичну збіжність методу Ньютона. Його іменем названа умова вибору початкового наближення в цьому методі.

Хайрер Ернст (англ. *Ernst Hairer*, нар. 1949 р.) – професор Женевського університету, автор явного методу Рунге–Кутти десятого порядку із 17 стадіями та інших методів із цієї сім'ї розв'язування ЗДР. Співавтор двотомника [75, 76], який на даний час є найповнішим викладом числового розв'язування ЗДР.

Холецький Андре–Луї (фр. *André–Louis Cholesky*, 1875–1918 рр.) – французький військовий геодезист. Його іменем названі схема розкладу матриці СЛАР на добуток двох трикутних матриць і метод квадратного кореня, який ґрунтується на цій ідеї.

Чебишев Пафнутій Львович (1821–1894 рр.) – російський математик і механік, автор праць з теорії наближення функцій, теорії чисел і теорії ймовірностей. У 1873 р. побудував КФ з однаковими коефіцієнтами. У 1859 р. показав, що похибка інтерполювання для функції, заданої на $[-1,1]$, мінімальна, якщо вузлами інтерполювання є корені введеного ним многочлена. Побудував многочлени, які найменше відхиляються від нуля.

Шонберг Якоб Ісаак (англ. *Isaac Jacob Schoenberg*, 1903–1990 рр.) – румунський і американський математик, увів термін «spline», автор першої публікації з теорії сплайнів (1946 р.).

Штермер Фредерік Карл Мюлерц (норв. *Frederik Carl Mülertz Størmer*, 1874–1957 рр.) – норвезький математик і геофізик. Запропонував метод розрахунку траєкторій космічних променів, який увійшов у сучасну математику як метод числового інтегрування ЗДР другого порядку (метод Штермера).

Штурм Шарль Франсуа (фр. *Sturm, Charles-François*, 1803–1855 рр.) – французький математик, член Паризької академії наук. Його праці присвячені теорії коливання струни (спільно з Ж. Ліувіллем розв'язав проблему знаходження власних значень і власних функцій для ЗДР) та числовим методам (Mémoire sur la résolution des équations, 1829 р.). У 1835 р. розв'язав задачу про існування та кількість дійсних коренів на інтервалі (a,b) для алгебраїчних рівнянь із дійсними коефіцієнтами.

Якобі Карл Густав Якоб (нім. *Carl Gustav Jacob Jacobi*, 1804–1851 рр.) – німецький математик і механік. Зробив внесок у комплексний аналіз, лінійну алгебру, механіку. Йому належать історично перший (1846 р.) метод знаходження власних значень матриці без зведення її до спеціального вигляду, а також ітераційний метод розв'язування СЛАР (1845 р.).

Список літератури та електронних джерел

1. Алберт Д. Теория сплайнов и ее приложения / Д. Алберт, Э. Нильсон, Д. Уолш. – М.: Мир, 1972. – 318 с.
2. Арушанян О.Б. Численное решение обыкновенных дифференциальных уравнений / О.Б. Арушанян, С.Ф. Залеткин. – М. МГУ, 1990. – 336 с.
3. Бабенко К.И. Основы численного анализа / К.И. Бабенко. – М.; Ижевск: НИЦ «Регулярная и хаотическая динамика», 2002. – 848 с.
4. Бабушка И. Численные процессы решения дифференциальных уравнений / И. Бабушка, Э. Витасек, М. Прагер. – М.: Мир, 1969. – 327 с.
5. Бахвалов Н.С. Численные методы / Н.С. Бахвалов, Н.П. Жидков, Г.М. Кобельков. – М.:, СПб.: Физматлит, 2003. – 632 с.
6. Бахвалов Н.С. Численные методы в задачах и упражнениях / Н.С. Бахвалов, Е.В. Лапин, Е.В. Чижонков. – М.: Высшая школа, 2000. – 190 с.
7. Бігун Я.Й. Числові методи розв'язування нелінійних рівнянь і систем: навч. посібник / Я.Й. Бігун, І.В. Березовська. – Чернівці: Чернівецький націо-нальний ун-т, 2011. – 103 с.
8. Бігун Я.Й. Числові методи. Системи лінійних алгебраїчних рівнянь: навч. посібник / Я.Й. Бігун, Л.М. Сергєєва. – Чернівці: Рута, 2008. – 152 с.
9. Бойко Л.Т. Основи чисельних методів / Л.Т. Бойко. – Дніпропетровськ: ДНУ, 2009. – 244 с.
10. Бор К. Практическое руководство по сплайнам / К. де Бор. – М.: Радио и связь, 1985. – 304 с.
11. Братусь А.С., Динамические системы и модели / А.С. Братусь, А.С. Новожи-лов, А.П. Платонов. – М.: Физматлит, 2011. – 401 с.
12. Васильев Ф.П. Численные методы решения экстремальных задач / Ф.П. Васильев. – М.: Наука, 1988. – 552 с.
13. Вержбицкий В.М. Основы численных методов / В.М. Вержбицкий. – М.: Высшая школа, 2005. – 840 с.
14. Воеводин В.В. Вычислительные основы линейной алгебры / В.В. Воеводин. – М.: Наука, 1977. – 303 с.
15. Гаврилюк І.С. Методи обчислень / І.С. Гаврилюк, В.Л. Макаров. – К.: Вища школа, 1995. – Ч.1. – 456 с.
16. Гаврилюк І.С. Методи наближених обчислень / І.С. Гаврилюк, В.Л. Мака-ров. – К.: Вища школа, 1995. – Ч.2. – 488 с.
17. Галагер Р. Метод конечных элементов /Р. Галагер. – М.: Мир, 1983. – 428 с.
18. Галанин М.П. Методы численного анализа математических моделей / М.П. Галанин, Е.Б. Савенков. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2010. – 591 с.

19. Гантмахер Ф.Р. Теория матриц / Ф.Р. Гантмахер. – М.: Физматлит, 2003. – 564 с.
20. Голуб Дж. Матричные вычисления / Дж. Голуб, Ч. Ван Лоун. – М.: 1999. – 548 с.
21. Деккер К. Устойчивость методов Рунге-Кутты для жестких нелинейных дифференциальных уравнений / К. Деккер, Я. Вервер. – М.: Мир, 1988. – 334 с.
22. Демидович Б.П. Основы вычислительной математики / Б.П. Демидович, И.А. Марон. – М.: Наука, 1966. – 664 с.
23. Деммель Дж. Вычислительная линейная алгебра. Теория и приложения / Дж. Деммель. – М.: Мир, 2001. – 429 с.
24. Джордж А., Лю Дж. Численное решение больших разреженных систем уравнений / А. Джордж, Дж. Лю. – М.: Мир, 1984. – 333 с.
25. Дэннис Дж. Численные методы безусловной оптимизации и решение нелинейных уравнений / Дж. Дэннис, Р. Шнабель. – М.: Мир, 1988. – 367 с.
26. Завьялов Ю.С. Методы сплайн-функций / Ю.С. Завьялов, Б.И. Квасов, Б.Л. Мирошниченко. – М.: Наука, 1980. – 350 с.
27. Калайда О.Ф. Чисельні методи (основи обчислювальної математики) / О. Ф. Калайда. – К. : ВПЦ "Київський університет", 2000. – 250 с.
28. Калиткин Н.Н. Численные методы / Н.Н. Калиткин. – Спб: БХВ-Петербург, 2011. – 592 с.
29. Канторович Л.В. Функциональный анализ / Л.В. Канторович, Г.П. Акилов. – М.: Наука, 1977. – 742 с.
30. Каханер Д. Численные методы и программное обеспечение / Д. Каханер, К. Моулер, С. Нэш. – М.: Мир, 1998. – 575 с.
31. Киреев В.И. Численные методы в примерах и задачах / В.И. Киреев, А.В. Пантелеев. – М.: Высшая школа, 2006. – 480 с.
32. Кнут Д. Искусство программирования / Д. Кнут. – М.: С-Пб.: Вильямс. – Т.2, 2002. – 820 с.
33. Коллатц Л. Функциональный анализ и вычислительная математика / Л. Коллатц. – М.: Мир, 1969. – 447 с.
34. Колмогоров А.Н. Элементы теории функций и функционального анализа. / А.Н. Колмогоров, С.В. Фомин. – М.: Наука, 1976. – 543 с.
35. Крылов В.И. Начала теории вычислительных методов. Интерполирование и интегрирование / В.И. Крылов, В.В Бобков, П.И. Монастырный. – Минск: Наука и техника, 1983. – 287 с.
36. Крылов В.И. Начала теории вычислительных методов. Линейная алгебра и нелинейные вычисления / В.И. Крылов, В.В Бобков, П.И. Монастырный. – Минск: Наука и техника, 1982. – 286 с.

37. Кубрак, А.І. Числові методи. Алгоритми і програми / А. І. Кубрак А.І. Жученко, О.В. Ситніков. – К. : Політехніка, 2013. – 389 с.
38. Курош А.Г. Алгебраические уравнения произвольных степеней / А.Г. Курош. – М.: Наука, 1983. – 332 с.
39. Лященко М.Я. Чисельні методи / М.Я. Лященко, М.С. Голованъ. – К. : Либідь, 1996. – 287 с.
40. Макаров В.Л. Сплайн-аппроксимация функций / В.Л. Макаров, В.В. Хлобыстов. – М.: Высшая школа, 1983. – 173 с.
41. Маценко В.Г. Комп'ютерна графіка: навч. посібник / В.Г. Маценко. – Чернівці: Чернівецький національний університет, 2009. – 343 с.
42. Маценко В.Г. Математичне моделювання: навч. посібник / В.Г. Маценко. – Чернівці: Чернівецький національний університет, 2013. – 519 с.
43. Молчанов И.Н. Машинные методы решения прикладной математики. Алгебра, приближение функций, обыкновенные дифференциальные уравнения / И.Н. Молчанов. – К.: Наук. думка, 2007. – 552 с.
44. Мысовских И.П. Интерполяционные кубатурные формулы / И.П. Мысовских. – М.: Наука, 1981. – 287 с.
45. Мэтьюз Дж.Г. Численные методы. Использование MathLab / Дж.Г. Мэтьюз, К.Д. Финк. – М.: Вильямс, 2001. – 720 с.
46. Никольский С.М. Квадратурные формулы / С.М. Никольский. – М.: Наука, 1987. – 363 с.
47. Норри Д. Введение в метод конечных элементов / Д. Норри, Ж. де Фриз. – М.: Мир, 1981. – 304 с.
48. Ортега Дж. Введение в численные методы решения дифференциальных уравнений / Дж. Ортега, У. Пул. – М.: Наука, 1986. – 288 с.
49. Ортега Дж. Итерационные методы решения нелинейных систем уравнений со многими неизвестными / Дж. Ортега, В. Рейнболдт. – М.: Мир, 1975. – 560 с.
50. Островский А.М. Решение уравнений и систем уравнений / А.М. Островский. – М.: Изд-во иностранной литературы, 1963. – 219 с.
51. Параллельные алгоритмы решения задач вычислительной математики / [Химич А.Н., Молчанов И.Н., Попов А.В., Чистякова Т.В., Яковлев М.Ф.]. – К.: Наук. думка, 2008. – 247 с.
52. Парлетт Б. Симметричная проблема собственных значений / Б. Парлетт. – М.: Мир, 1983. – 382 с.
53. Писанецки С. Технология разреженных матриц / С. Писанецки. – М.: Мир, 1988. – 411 с.
54. Попов В.В. Методи обчислень / В.В. Попов. – К.: ВПЦ «Київський університет», 2012. – 303 с.
55. Програмування числових методів мовою Python: [підруч.] / А.В. Анісімов,

А.Ю. Дорошенко, С.Д. Погорілий, Я.Ю. Дорогий. – К.: ВПЦ «Київський університет», 2014. – 560 с.
56. Райс Дж. Матричные вычисления и математическое обеспечение / Дж. Райс. – М.: Мир, 1984. – 264 с.
57. Ракитский Ю.В. Численные методы решения жестких систем / Ю.В. Ракитский, С.М. Устинов, И.Г. Черноруцкий. – М.: Наука, 1974. – 808 с.
58. Самарский А.А. Теория разностных схем / А.А. Самарский. – М.: Наука, 1989. – 616 с.
59. Самарский А.А. Численные методы / А.А. Самарский, А.В. Гулин. – М.: Наука, 1989. – 430 с.
60. Самарский А.А. Методы решения сеточных уравнений / А.А. Самарский, Е.С. Николаев. – М.: Наука, 1978. – 591 с.
61. Самойленко А.М. Диференціальні рівняння: підручник / А.М. Самойленко, М.О. Перестюк, І.О. Парасюк. – К.: Видавничо-поліграфічний центр «Київський університет», 2010. – 527 с.
62. Самойленко А.М. Математичні аспекти теорії нелінійних коливань / А.М. Самойленко, Р.І. Петришин. − Київ: Наукова думка, 2004. – 475 с.
63. Соболев С.Л. Введение в теорию кубатурных формул / С.Л. Соболев. – М.: Наука, 1974. – 808 с.
64. Соболь И.М. Численные методы Монте-Карло И.М. / И.М. Соболь. – М.: Наука, 1973. – 311 с.
65. Современные численные методы решения дифференциальных уравнений / Под. Ред. Дж. Холла и Дж. Уатта. – М.: Мир, 1979. – 312 с.
66. Стечкин С.Б. Сплайны в вычислительной математике / С.Б. Стечкин, Ю.Н. Субботин. – М.: Наука, 1976. – 248 с.
67. Стренг Г. Теория метода конечных элементов / Г. Стренг, Дж. Фикс. – М.: Мир, 1977. – 349 с.
68. Суэтин П.К. Классические ортогональные многочлены / П.К. Суэтин. – М.: Наука, 1979. – 415 с.
69. Тихонов А.Н. Методы решения некорректных задач / А.Н. Тихонов, В.Я. Арсенин. – М.: Наука, 1979. – 288 с.
70. Трауб Дж. Итерационные методы решения уравнений / Дж. Трауб. – М.: Мир, 1985. – 464 с.
71. Уилкинсон Дж. Алгебраическая проблема собственных значений / Дж. Уилкинсон. – М.: Наука, 1970. – 430 с.
72. Фаддеев Д.К. Вычислительные методы линейной алгебры / Д.К. Фаддеев, В.Н. Фаддеева. – М.: Физматгиз, 1963. – 736 с.
73. Фельдман Л.П. Чисельні методи в інформатиці / Л.П. Фельдман, А.І. Петренко, О.А. Дмитрієва – К.: Видавнича група BHV, 2006. – 480 с.

74. Форсайт Дж. Машинные методы математических вычислений / Дж. Форсайт, М. Малькольм, К. Моулер. – М.: Мир, 1980. – 279 с.
75. Хайрер Э. Решение обыкновенных дифференциальных уравнений. Нежёсткие задачи/Э. Хайрер, С. Нёрсетт, Г. Ваннер. – М.: Мир, 1990. – 512 с.
76. Хайрер Э. Решение обыкновенных дифференциальных уравнений. Жёсткие задачи / Э. Хайрер, Г. Ваннер. – М.: Мир, 1999. – 685 с.
77. Хейгеман Л., Янг Д. Прикладные итерационные методы / Л. Хейгеман, Д. Янг. – М.: Мир, 1986. – 448 с.
78. Цегелик Г.Г. Чисельні методи / Г.Г. Цегелик. – Львів: Видавничий центр Львівського національного університету, 2004. – 408 с.
79. Шахно С.М. Практикум з чисельних методів: навч. посібник / С.М. Шахно, А.Т. Дудекевич, С.М. Левицька. – Львів: ЛНУ імені Івана Франка, 2013. – 432 с.
80. Шахно С.М. Чисельні методи лінійної алгебри: навч. посібник / С.М. Шахно. – Львів: Видавничий центр ЛНУ імені Івана Франка, 2007. – 245 с.
81. Эстербю О. Прямые методы для разреженных матриц / О. Эстербю, З. Златев З. – М.: Мир, 1987. – 118 с.
82. Ясинська Л.І. Імітаційне моделювання на ЕОМ / Л.І. Ясинська, В.К. Ясинський, І.В. Юрченко. – Чернівці: Зелена Буковина, 1999. – 346 с.
83. Buchanan Ja.I. Numerical Methods and Analysis / Ja.I. Buchanan. – New York: McGrw-Hill, Ins.1992. – 751 p.
84. Butcher J.C. Numerical methods for ordinary differential equations / J.C. Butcher. − John Wiley & Sons Ltd, 2008. − 463 p.
85. Butcher J.C. On Runge−Kutta processes of high order / J.C. Butcher // J. Austral. Math. Soc. − 1964. − 4, №3. – P. 179–194.
86. Butcher J.C. The non-existence of ten stage eighth order explicit Runge–Kutta methods / J.C. Butcher // BIT, 1985, N 25. – P. 521–540.
87. Cooper G.J. Some explicit Runge–Kutta methods of high order / G.J. Cooper, J.H. Verner // SIAM J. Numer. Anal. – 1972, N 9. – P. 389–405.
88. Curtis A.R. An eighth order Runge–Kutta processes with eleven function evaluations per step / A.R. Curtis // Numer. Math. – 1970. – 16. – P. 268–277.
89. Curtis A.R. High-order explicit Runge–Kutta formula, their uses and limitations / A.R. Curtis // J. Inst. Maths. Applics, 1975, N 16. – P. 35–55.
90. Dahlquist G.A. special stability problem for linear multistep methods methods / G.A. Dahlquist // BIT. – 1963, V. 3. – P. 27–43.
91. Dahlquist G.A. Numerical Methods / G.A. Dahlquist, Ă. Bjőrck. – New Jersey: Prentice–Hall. Inc, 1974. – 574 p.
92. Fillipson P.E, Modeling by nonlinear differential equations / P.E. Fillipson, P. Schuster. – Singapore: World Scientific Publishing Co.R.Ltd, 2009 – 225 p.
93. Gautschi W. Numerical analysis / W. Gautschi. – New York, Berlin, London:

Springer Dordrecht Heidelberg, 2012. – 588 p.
94. Gear C.W. Numerical initial value problems in ordinary differential equations / C.W. Gear. − New Jersey: Printice, 1971. – 253 p.
95. Haurer E. A Runge–Kutta method of order 10 / E. Haurer // J. Inst. Maths. Applics, 1978, N21. – P. 47–59.
96. Handbook of Floating-Point Arithmetic / Muller J.-M., Brisebarre N., Dinechin F. and other. – Boston, Basel, Berlin, Birkhauser Boston, 2010. − 572 p.
97. Higham J.N. Accuracy and stability of numerical algorithms / J. Nicholas. Higham. – Philadephia: SIAM, 2002. – 680 p.
98. Lambert J.D. Numerical for ordinary differential equations / J.D. Lambert. – NewYork: John Wiley & Sons Ltd, 2000. − 293 p.
99. Truong Nguyen-Ba Nine-stage multi-derivative Runge–Kutta method of order 12 / Truong Nguyen-Ba, Vl. Božić, Em. Kengne and R. Vaillancourt // Publications de l'Inst. Math. Nouvelle série. – 2009, V. 86(100). – P. 75–96.
100. Quarteroni A. Numerical Mathematics / A. Quarteroni, R. Sacco, F. Saleri. – New York, Berlin, London: Springer Dordrecht Heidelberg, 2012. – 588 p.

Електронні джерела

101. Методи Рунге–Кутти порядку 12–14 [Електронний ресурс]. – Режим доступу: URL **:** http://sce.uhcl.edu/rungekutta/
102. The Runge–Kutta Club [Electronic Resource]**. –** Mode of access : URL **:** http://www.math.auckland.ac.nz/~butcher/RKclub/
103. Numerical Methods with Applications [Electronic Resource]**. –** Mode of access: URL **:** http://autarkaw.com/books/numericalmethods/index.html/
104. WEB-сторінку кластерних систем Інституту кібернетики імені В.М. Глушкова НАН України [Електронний ресурс]. – Режим доступу: http://icybcluster.org.ua/
105. Сторінка MATLAB на сайті The MathWorks [Електронний ресурс]. – Режим доступу : URL **:** http://www.mathworks.com/
106. Комп'ютерна система Mathematica14 [Електронний ресурс]. – Режим доступу: URL **:** www.wolfram.com/
107. Комп'ютерна система MathCad 14 [Електронний ресурс]. – Режим доступу: URL **:** http://www.ptc.com/product/mathcad/
108. Програми на Фортрані і Matlab-коди розв'язування ЗДР [Електронний ресурс]. – Режим доступу :
URL : http://www.unige.ch/~hairer/software.html/

Додаток А. Коефіцієнти явних методів Рунге–Кутти

Таблиця А1. Явний метод Ейлера ($p = s = 1$)

0	
	1

Методи Рунге 2-го порядку ($p = s = 2$)

Таблиця А2.

0		
$\frac{1}{2}$	$\frac{1}{2}$	
	0	1

Таблиця А3.

0		
1	1	
	$\frac{1}{2}$	$\frac{1}{2}$

Таблиця А4.

0		
$\frac{2}{3}$	$\frac{2}{3}$	
	$\frac{1}{4}$	$\frac{3}{4}$

Методи Рунге–Кутти третього порядку

Таблиця А5. Метод Хойна($p = s = 3$)

0			
$\frac{2}{3}$	$\frac{2}{3}$		
$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	
	$\frac{1}{4}$	0	$\frac{3}{4}$

Таблиця А6. Метод Рунге ($p = s = 3$)

0			
$\frac{1}{2}$	$\frac{1}{2}$		
1	-1	2	
	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$

Таблиця А7. Метод Рунге ($p = 3, s = 4$)

0				
$\frac{1}{2}$	$\frac{1}{2}$			
1	0	1		
1	0	0	1	
	$\frac{1}{6}$	$\frac{4}{6}$	0	$\frac{1}{6}$

Методи Рунге–Кутти четвертого порядку ($p = s = 4$)

Таблиця А8. “Класичний” метод Кутти

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

Таблиця А9. Метод „3/8”

0				
$\frac{1}{3}$	$\frac{1}{3}$			
$\frac{2}{3}$	$-\frac{1}{3}$	1		
1	1	-1	1	
	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Таблиця А10. Метод Гілла

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	$\frac{\sqrt{2}-1}{2}$	$\frac{2-\sqrt{2}}{2}$		
1	0	$-\frac{\sqrt{2}}{2}$	$\frac{2+\sqrt{2}}{2}$	
	$\frac{1}{6}$	$\frac{2-\sqrt{2}}{6}$	$\frac{2+\sqrt{2}}{6}$	$\frac{1}{6}$

Таблиця А11.

0				
1	1			
$\frac{1}{2}$	$\frac{3}{8}$	$\frac{1}{8}$		
1	$-\frac{1}{2}$	$-\frac{1}{2}$	2	
	$\frac{1}{6}$	0	$\frac{2}{3}$	$\frac{1}{6}$

Методи 5-го порядку ($s \geq p+1$)

Таблиця А12. Метод Кутти ($p = 5, s = 6$)

0						
$\frac{1}{5}$	$\frac{1}{5}$					
$\frac{2}{5}$	0	$\frac{2}{5}$				
1	$\frac{9}{4}$	-5	$\frac{15}{4}$			
$\frac{3}{5}$	$-\frac{63}{100}$	$\frac{9}{5}$	$-\frac{13}{20}$	$\frac{2}{25}$		
$\frac{4}{5}$	$-\frac{18}{75}$	$\frac{4}{5}$	$\frac{2}{15}$	$\frac{8}{75}$	0	
	$\frac{17}{144}$	0	$\frac{100}{144}$	$\frac{2}{144}$	$-\frac{50}{144}$	$\frac{75}{144}$

Таблиця A13. Метод Нюстрема $(p=5, s=6)$

0						
$\frac{1}{3}$	$\frac{1}{3}$					
$\frac{2}{5}$	$\frac{4}{25}$	$\frac{6}{25}$				
1	$\frac{1}{4}$	-3	$\frac{15}{4}$			
$\frac{2}{3}$	$\frac{2}{27}$	$\frac{10}{9}$	$-\frac{50}{81}$	$\frac{8}{81}$		
$\frac{4}{5}$	$\frac{2}{25}$	$\frac{12}{25}$	$\frac{2}{15}$	$\frac{8}{75}$	0	
	$\frac{23}{192}$	0	$\frac{125}{192}$	0	$-\frac{27}{64}$	$\frac{125}{192}$

Методи Бутчера 6-го порядку $(s \geq p+1)$

Таблиця A14 $(p=6, s=7)$

0							
$\frac{1}{2}$	$\frac{1}{2}$						
$\frac{2}{3}$	$\frac{2}{9}$	$\frac{4}{9}$					
$\frac{1}{3}$	$\frac{7}{36}$	$\frac{2}{9}$	$-\frac{1}{12}$				
$\frac{5}{6}$	$-\frac{35}{144}$	$-\frac{55}{36}$	$\frac{35}{48}$	$\frac{15}{8}$			
$\frac{1}{6}$	$-\frac{1}{360}$	$-\frac{11}{36}$	$-\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{10}$		
1	$-\frac{41}{260}$	$\frac{22}{13}$	$\frac{43}{156}$	$-\frac{118}{39}$	$\frac{32}{195}$	$\frac{80}{39}$	
	$\frac{13}{200}$	0	$\frac{11}{40}$	$\frac{11}{40}$	$\frac{4}{25}$	$\frac{4}{25}$	$\frac{13}{200}$

Таблиця A15 ($p=6, s=7$)

0							
$\frac{1}{3}$	$\frac{1}{3}$						
$\frac{2}{3}$	0	$\frac{2}{3}$					
$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{3}$	$-\frac{1}{12}$				
$\frac{5}{6}$	$\frac{25}{48}$	$-\frac{55}{24}$	$\frac{35}{48}$	$\frac{15}{8}$			
$\frac{1}{6}$	$\frac{3}{20}$	$-\frac{11}{24}$	$-\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{10}$		
1	$-\frac{261}{260}$	$\frac{33}{13}$	$\frac{43}{156}$	$-\frac{118}{39}$	$\frac{32}{195}$	$\frac{80}{39}$	
	$\frac{13}{200}$	0	$\frac{11}{40}$	$\frac{11}{40}$	$\frac{4}{25}$	$\frac{4}{25}$	$\frac{13}{200}$

Метод 7-го порядку ($s \geq p+2$)

Таблиця A16. Метод Бутчера ($p=7, s=9$)

0								
$\frac{1}{6}$	$\frac{1}{6}$							
$\frac{1}{3}$	0	$\frac{1}{3}$						
$\frac{1}{2}$	$\frac{1}{8}$	0	$\frac{3}{8}$					
$\frac{2}{11}$	$\frac{148}{1331}$	0	$\frac{150}{1331}$	$-\frac{56}{1331}$				
$\frac{2}{3}$	$-\frac{404}{243}$	0	$-\frac{170}{27}$	$\frac{4024}{1701}$	$\frac{10648}{1701}$			
$\frac{6}{7}$	$\frac{2466}{2401}$	0	$\frac{1242}{343}$	$-\frac{19176}{16807}$	$-\frac{51909}{16807}$	$\frac{1053}{2401}$		
0	$\frac{5}{154}$	0	0	$\frac{96}{539}$	$-\frac{1815}{20384}$	$-\frac{405}{2464}$	$\frac{49}{1144}$	
1	$-\frac{113}{32}$	0	$-\frac{195}{22}$	$\frac{32}{7}$	$\frac{29403}{3584}$	$-\frac{729}{512}$	$\frac{1029}{1408}$	$\frac{21}{16}$
	0	0	0	$\frac{32}{105}$	$\frac{1771561}{6289920}$	$\frac{243}{2560}$	$\frac{16807}{74880}$	$\frac{77}{1440}$	$\frac{11}{270}$

Методи 8-го порядку ($s \geq p+3$)

Таблиця А17. Метод Шенкса ($p=8,\ s=12$)

0												
$\frac{1}{9}$	$\frac{1}{9}$											
$\frac{1}{6}$	$\frac{1}{24}$	$\frac{1}{8}$										
$\frac{1}{4}$	$\frac{1}{16}$	0	$\frac{3}{16}$									
$\frac{1}{10}$	$\frac{29}{500}$	0	$\frac{33}{500}$	$-\frac{3}{125}$								
$\frac{1}{6}$	$\frac{11}{324}$	0	0	$\frac{1}{243}$	$\frac{125}{972}$							
$\frac{1}{2}$	$-\frac{7}{12}$	0	0	$\frac{19}{9}$	$\frac{125}{36}$	$-\frac{9}{2}$						
$\frac{2}{3}$	$-\frac{10}{81}$	0	0	$-\frac{32}{243}$	$\frac{125}{243}$	0	$\frac{11}{27}$					
$\frac{1}{3}$	$\frac{1175}{324}$	0	0	$-\frac{32}{3}$	$-\frac{3125}{162}$	26	$\frac{121}{162}$	$-\frac{1}{12}$				
$\frac{5}{6}$	$\frac{293}{324}$	0	0	$-\frac{71}{27}$	$-\frac{1375}{324}$	$\frac{51}{9}$	$-\frac{59}{162}$	$\frac{1}{2}$	1			
$\frac{5}{6}$	$\frac{1303}{1620}$	0	0	$-\frac{71}{27}$	$-\frac{1375}{324}$	$\frac{37}{6}$	$\frac{103}{162}$	0	0	$\frac{1}{10}$		
1	$-\frac{955}{492}$	0	0	$\frac{2560}{369}$	$\frac{8125}{738}$	$-\frac{612}{41}$	$\frac{7}{82}$	$-\frac{27}{164}$	$-\frac{18}{41}$	$-\frac{12}{41}$	$\frac{30}{41}$	
	$\frac{41}{840}$	0	0	0	0	$\frac{216}{840}$	$\frac{272}{840}$	$\frac{27}{840}$	$\frac{27}{840}$	$\frac{36}{840}$	$\frac{180}{840}$	$\frac{41}{840}$

Таблиця A18. Метод Купера–Вернера $(p=8, s=11)$

0											
$\frac{1}{2}$	$\frac{1}{2}$										
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$									
$\frac{7-\sqrt{21}}{14}$	$\frac{1}{7}$	$\frac{-7+3\sqrt{21}}{98}$	$\frac{21-5\sqrt{21}}{49}$								
$\frac{7-\sqrt{21}}{14}$	$\frac{11-\sqrt{21}}{84}$	0	$\frac{18-4\sqrt{21}}{63}$	$\frac{21+\sqrt{21}}{252}$							
$\frac{1}{2}$	$\frac{5-\sqrt{21}}{48}$	0	$\frac{9-\sqrt{21}}{36}$	$\frac{-231-14\sqrt{21}}{360}$	$\frac{63+7\sqrt{21}}{80}$						
$\frac{7+\sqrt{21}}{14}$	$\frac{10+\sqrt{21}}{42}$	0	$\frac{-432-92\sqrt{21}}{315}$	$\frac{633+145\sqrt{21}}{90}$	$\frac{-504-115\sqrt{21}}{70}$	$\frac{63+13\sqrt{21}}{35}$					
$\frac{7+\sqrt{21}}{14}$	$\frac{1}{14}$	0	0	0	$\frac{14+3\sqrt{21}}{126}$	$\frac{13+3\sqrt{21}}{63}$	$\frac{1}{9}$				
$\frac{1}{2}$	$\frac{1}{32}$	0	0	0	$\frac{91+21\sqrt{21}}{576}$	$\frac{11}{72}$	$\frac{-385+75\sqrt{21}}{1152}$	$\frac{63-13\sqrt{21}}{128}$			
$\frac{7-\sqrt{21}}{14}$	$\frac{1}{14}$	0	0	0	$\frac{1}{9}$	$\frac{-733+147\sqrt{21}}{2205}$	$\frac{515-111\sqrt{21}}{504}$	$\frac{-51+11\sqrt{21}}{56}$	$\frac{132-28\sqrt{21}}{245}$		
1	0	0	0	0	$\frac{-42-7\sqrt{21}}{18}$	$\frac{-18-28\sqrt{21}}{45}$	$\frac{-273+53\sqrt{21}}{72}$	$\frac{301-53\sqrt{21}}{72}$	$\frac{28+28\sqrt{21}}{45}$	$\frac{49+7\sqrt{21}}{18}$	
	$\frac{1}{20}$	0	0	0	0	0	0	$\frac{49}{180}$	$\frac{16}{45}$	$\frac{49}{180}$	$\frac{1}{20}$

A19. Метод Хайрера[1] 10-го порядку ($p=10, s=17$)

Коефіцієнти подані з 60 правильними цифрами

c[1] = 0
c[2] = 0.500000000000000000000000000000000000000000000000000000000000
c[3] = 0.526509100141612162751191034851811859742209850194804733855006
c[4] = 0.789763650212418244126786552277717789613314775292207100782509
c[5] = 0.393923570125671611282770592155157086836785959703592730218258
c[6] = 0.7666539862535488
c[7] = 0.289763650212418244126786552277717789613314775292207100782507
c[8] = 0.108477689219566212940847926866966428088773728748584448916927
c[9] = 0.357384241759677451842924502979560464040498263636787304090125
c[10] = 0.882527661964732346425501486979669075182867844268052119663791
c[11] = 0.642615758240322548157075497020439535959501736363212695909875
c[12] = 0.117472338035267653574498513020330924817132155731947880336209
c[13] = 0.7666539862535488
c[14] = 0.289763650212418244126786552277717789613314775292207100782507
c[15] = 0.526509100141612162751191034851811859742209850194804733855006
c[16] = 0.500000000000000000000000000000000000000000000000000000000000
c[17] = 1

b[1] = 0.033333333333333333333333333333333333333333333333333333333333
b[2] = -0.033333333333333333333333333333333333333333333333333333333333
b[3] = -0.120000000000000000000000000000000000000000000000000000000000
b[4] = b[5] = 0
b[6] = -0.130000000000000000000000000000000000000000000000000000000000
b[7] = -0.180000000000000000000000000000000000000000000000000000000000
b[8] = 0
b[9] = 0.27742918851774317650836026256065434042850431971840836339461
b[10] = 0.18923747814892349015830640410601232623816234694862583032718 6
b[11] = 0.27742918851774317650836026256065434042850431971840836339461
b[12] = 0.18923747814892349015830640410601232623816234694862583032718 6
b[13] = 0.13, b[14] = 0.18, b[15] = 0.12
b[16] = 0.033333333333333333333333333333333333333333333333333333333333
b[17] = 0.033333333333333333333333333333333333333333333333333333333333

a[2][1] = 0.500000000000000000000000000000000000000000000000000000000000
a[3][1] = 0.249297267609681978012770798836949850852762744727995584459954
a[3][2] = 0.277211832531930184738420236014862008889447105466809149395052
a[4][1] = 0.197440912553104561031696638069429447403328693823051775195627
a[4][2] = 0
a[4][3] = 0.592322737659313683095089914208288342209986081469155325586884
a[5][1] = 0.197320548628702140899690086861353747358770450201528525984281
a[5][2] = 0
a[5][3] = 0.295083334092671853710828447833096210342908115497316514674583

[1] 1978 р., 169 коефіцієнтів

a[5][4] = -0.0984803125957023833277479425392928708648926059952523104406 06
a[6][1] = 0.131313417344461520076336778998284793307104801194610995172297
a[6][2] = a[6][3] = 0
a[6][4] = 0.110154439538638507039579681263290188617727424904457619127907
a[6][5] = 0.525186129370448772884083539738425018075167773900931385699766
a[7][1] = 0.134200341846322406193523571556538810809682280358769153934056
a[7][2] = a[7][3] = 0
a[7][4] = 0.696088703288076908068781383509747545922284685096513437617197
a[7][5] = 0.250497721570339375352478192091560594379279460030814125387197
a[7][6] = -0.791023116492320445487996594880129161497931650193889616155943
a[8][1] = 0.072218274189662145445099670444212796765974993995168482122971
a[8][2] = a[8][3] = a[8][4] = 0
a[8][5] = -0.058336322936455036913377508391576010317463044557635281319970
a[8][6] = 0.003047557668574494379291469893838362961393915887951023386923
a[8][7] = 0.091548180297784610029834294920491278678867863423100224727003
a[9][1] = 0.031255008135165617062363364144360909510377723542879033066377
a[9][2] = a[9][3] = a[9][4] = a[9][5] = 0
a[9][6] = 0.000109123821542419946869294595176439827041028849283869573166
a[9][7] = 0.156725758630995015163428123239419606745477617728875204132576
a[9][8] = 0.169294351171974399670263721000603507957601893515749197317902
a[10][1] = 0.011906604414675032143567980566018907965795279427286722551640
a[10][2] = a[10][3] = a[10][4] = a[10][5] = 0
a[10][6] = 0.283437082024606548111158799175805892098647137522471712426518
a[10][7] = -0.41631216757056131507274318900799605125082734477463205055540 6
a[10][8] = 0.264646333949743004842122357733181166192796563855712621386704
a[10][9] = 0.738849809146269076401395538512659160176456208237213113854335
a[11][1] = 0.023406573691335449391473625259504364790836965582732726515266
a[11][2] =a[11][3] = a[11][4] = a[11][5] = 0
a[11][6] = 0.094493130189496180237115961328822134600313674649913746742381
a[11][7] = -0.272872055901956418864682670202868234952633322515255977036325
a[11][8] = 0.224022046115592207351971806319306580738071745039347414295235
a[11][9] = 0.604381441075135095305897333729761437318011101339504960867870
a[11][10] = -0.030815376929279965264700559414086746535098427733030175474552
a[12][1] = 0.045443775310176369942806803893104543458338111866468981747004
a[12][2] = a[12][3] = a[12][4] = a[12][5] = 0
a[12][6] = -0.001187996671864415676041636994855456361339596915346007818933
a[12][7] = 0.012035654990928113493648304752436307386546037675792982032503
a[12][8] = 0.075126902987647924054107277470638404347912348750999242251640
a[12][9] = -0.018220924098884569051697315534231472982896496949076759684261
a[12][10] = -0.000257152854084065043410107101784645506674343404229260965934
a[12][11] = 0.004532078371348295855085186535023244475246094707338702774190
a[13][1] = 0.178401086400436429271810252236105036974681175327876024199032
a[13][2] = a[13][3] = 0
a[13][4] = 0.110154439538638507039579681263290188617727424904457619127907
a[13][5] = 0.525186129370448772884083539738425018075167773900931385699766
a[13][6] = -0.489148591820436212817412947822614869893806907334840001353455
a[13][7] = 0.932443612635135732902219456897349102947542447872601542391818

a[13][8] = -0.774475053439839525351939491600259737086503063606661168648411
a[13][9] = -1.0549021781393582425937994254969188331798776260496196838361 33
a[13][10] = 0.131046712034157154515966324465601614648136847647750441792119
a[13][11] = 0.587049777599487392267978695578709837405311995961302793122857
a[13][12] = 0.620898052074878791881513914740312641491619931376201047504500
a[14][1] = 0.130220806600497793489850897668368548052241092768846863090306
a[14][2] = a[14][3] = 0
a[14][4] = 0.696088703288076908068781383509747545922284685096513437617197
a[14][5] = 0.250497721570339375352478192091560594379279460030814125387197
a[14][6] = -0.758948987129607342656068221840345352199042576536499357879795
a[14][7] = -0.171517208463488383623083873537419777707371265796495694789097
a[14][8] = -0.370217673678906704670291354494340072195473492233820139184160
a[14][9] = 0.124981008574747347838896675044479814593797663434805537555884
a[14][10] = 0.003353109248372670741534775312156870143841271203588569182358
a[14][11] = -0.006632546136761535819147201694291880906694816676094469246359
a[14][12] = 0.429116573121617904714636512492799112778719296257597242114120
a[14][13] = -0.037177856782469789310801232274997613248266542257049013065144
a[15][1] = 0.249297267609681978012770798836949850852762744727995584459954
a[15][2] = 0.277211832531930184738420236014862008889447105466809149395052
a[15][3] = a[15][4] = a[15][5] = 0
a[15][6] = -0.145940595936085218185544908546664251487835151696104976496855
a[15][7] = -0.799015893511029475357365128687567486685020877796267011619833
a[15][8] = a[15][9] = a[15][10] = a[15][11] =a[15][12] = 0
a[15][13] = 0.145940595936085218185544908546664251487835151696104976496855
a[15][14] = 0.799015893511029475357365128687567486685020877796267011619833
a[16][1] = 0.500000000000000000000000000000000000000000000000000000000000
a[16][2] = 0
a[16][3] = -0.807097076095341093249737398720457841858098658277918854066158
a[16][4] = a[16][5] = a[16][6] = a[16][7] = a[16][8] = a[16][9] = a[16][10] = 0
a[16][11] = a[16][12] = a[16][13] = a[16][14] = 0
a[16][15] = 0.807097076095341093249737398720457841858098658277918854066158
a[17][1] = 0.057320795432057541096538894884632859560006508826368053069170
a[17][2] = -0.499999999999999999999999999999999999999999999999999999999984
a[17][3] = -0.897470163394855120845974875813019463069945881020784104055832
a[17][4] = a[17][5] = 0
a[17][6] = -1.039910049226953433635817598029320374005958706733760229012810
a[17][7] = -0.407357014288385810046479045129055783867364682542885522639949
a[17][8] = -0.182830236640741849250386135478138753893364006257772618117464
a[17][9] = -0.333659270649225020341781909749343709488434667585788318361911
a[17][10] = 0.395648542376057924045745488431211325217210441572302981789688
a[17][11] = 0.695057049459982281781252754578863682006506102115089723858592
a[17][12] = 0.271487376457383239111724412549073964264941175168838760307907
a[17][13] = 0.585423734866589756810364983516995806184432782082086755163155
a[17][14] = 0.958819072213235370428838154425080984022025053355520413943622
a[17][15] = 0.897470163394855120845974875813019463069945881020784104055832
a[17][16] = 0.499999999999999999999999999999999999999999999999999999999984

Додаток Б. Коефіцієнти неявних методів Рунге-Кутти

Таблиця Б1. Правило середньої точки ($p = 2$)

$\frac{1}{2}$	$\frac{1}{2}$
	1

Таблиця Б2. Метод Хамера – Холінгсуорта ($p = 4$)

$\frac{1}{2}-\frac{\sqrt{3}}{6}$	$\frac{1}{4}$	$\frac{1}{4}-\frac{\sqrt{3}}{6}$
$\frac{1}{2}+\frac{\sqrt{3}}{6}$	$\frac{1}{4}+\frac{\sqrt{3}}{6}$	$\frac{1}{4}$
	$\frac{1}{2}$	$\frac{1}{2}$

Таблиця Б3. Метод Кунцмана–Бутчера ($p = 6$)

$\frac{1}{2}-\frac{\sqrt{15}}{10}$	$\frac{5}{36}$	$\frac{2}{9}-\frac{\sqrt{15}}{15}$	$\frac{5}{36}-\frac{\sqrt{15}}{30}$
$\frac{1}{2}$	$\frac{5}{36}-\frac{\sqrt{15}}{24}$	$\frac{2}{9}$	$\frac{5}{36}-\frac{\sqrt{15}}{24}$
$\frac{1}{2}+\frac{\sqrt{15}}{10}$	$\frac{5}{36}+\frac{\sqrt{15}}{30}$	$\frac{2}{9}+\frac{\sqrt{15}}{15}$	$\frac{5}{36}$
	$\frac{5}{18}$	$\frac{4}{9}$	$\frac{5}{18}$

Методи Радо 3-го порядку

Таблиця Б4а

0	$\frac{1}{4}$	$-\frac{1}{4}$
$\frac{2}{3}$	$\frac{1}{4}$	$\frac{5}{12}$
	$\frac{1}{4}$	$\frac{3}{4}$

Таблиця Б4б

$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{3}{4}$	$\frac{1}{4}$
	$\frac{3}{4}$	$\frac{1}{4}$

Методи Радо 5-го порядку

Таблиця Б5а

0	$\frac{1}{9}$	$\frac{-1-\sqrt{6}}{18}$	$\frac{-1+\sqrt{6}}{18}$
$\frac{6-\sqrt{6}}{10}$	$\frac{1}{9}$	$\frac{88+7\sqrt{6}}{360}$	$\frac{88-43\sqrt{6}}{360}$
$\frac{6+\sqrt{6}}{10}$	$\frac{1}{9}$	$\frac{88+43\sqrt{6}}{360}$	$\frac{88-7\sqrt{6}}{360}$
	$\frac{4}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{16-\sqrt{6}}{36}$

Таблиця Б5б

$\frac{4-\sqrt{6}}{10}$	$\frac{88-7\sqrt{6}}{360}$	$\frac{4-\sqrt{6}}{10}$	$\frac{88-7\sqrt{6}}{360}$
$\frac{4+\sqrt{6}}{10}$	$\frac{269+169\sqrt{6}}{1800}$	$\frac{4+\sqrt{6}}{10}$	$\frac{269+169\sqrt{6}}{1800}$
1	$\frac{16-\sqrt{6}}{36}$	1	$\frac{16-\sqrt{6}}{36}$
	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{4}{36}$

Методи Лобатто 2-го порядку

Таблиця Б6а

0	0	0
1	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$

Таблиця Б6б

0	$\frac{1}{2}$	0
1	$\frac{1}{2}$	0
	$\frac{1}{2}$	$\frac{1}{2}$

Таблиця Б6в

0	$\frac{1}{2}$	$-\frac{1}{2}$
1	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$

Методи Лобатто 4-го порядку

Таблиця Б7а

0	0	0	0
$\frac{1}{2}$	$\frac{5}{24}$	$\frac{1}{3}$	$-\frac{1}{24}$
1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$

Таблиця Б7б

0	$\frac{1}{6}$	$-\frac{1}{6}$	0
$\frac{1}{2}$	$\frac{5}{24}$	$\frac{1}{3}$	0
1	$\frac{1}{6}$	$\frac{5}{6}$	0
	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$

Таблиця Б7в

0	$\frac{1}{6}$	$-\frac{1}{3}$	$\frac{1}{6}$
$\frac{1}{2}$	$\frac{1}{6}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$

Діагонально неявні методи Ньорсета

Таблиця Б8 (p = 3)

γ	γ	0
$\frac{1}{6\gamma}$	$-\frac{\sqrt{3}}{3}$	γ
	$\frac{1}{2}$	$\frac{1}{2}$

$$\gamma = \frac{3+\sqrt{3}}{6}$$

Таблиця Б9 (p = 4)

γ	γ	0	0
$\frac{1}{2}$	$\frac{1}{2}-\gamma$	γ	0
γ	2γ	$1-4\gamma$	γ
	β	$1-2\beta$	β

$$\beta = \frac{1}{12(2-4\gamma)^2}, \quad \gamma = \frac{1}{2} + \frac{1}{\sqrt{3}}\cos\frac{\pi}{18}.$$

Вкладений діагонально-неявний метод 4-го порядку

Таблиця Б10

$\frac{1}{4}$	$\frac{1}{4}$				
$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$			
$\frac{11}{20}$	$\frac{17}{50}$	$-\frac{1}{25}$	$\frac{1}{4}$		
$\frac{1}{2}$	$\frac{371}{1360}$	$-\frac{137}{2720}$	$\frac{15}{544}$	$\frac{1}{4}$	
1	$\frac{25}{24}$	$-\frac{49}{48}$	$\frac{125}{16}$	$-\frac{85}{12}$	$\frac{1}{4}$
y_{n+1}	$\frac{25}{24}$	$-\frac{49}{48}$	$\frac{125}{16}$	$-\frac{85}{12}$	$\frac{1}{4}$
$\hat{y}_{n+1}$	$\frac{59}{48}$	$-\frac{17}{96}$	$\frac{225}{32}$	$-\frac{85}{12}$	0

Додаток В. Коефіцієнти вкладених методів Рунге–Кутти

Таблиця В1 РК2(3)

$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{3}{4}$	$\frac{3}{16}$	$\frac{9}{16}$		
1	$\frac{5}{18}$	$\frac{1}{6}$	$\frac{5}{9}$	
$p=3$	$\frac{5}{18}$	$\frac{1}{6}$	$\frac{5}{9}$	
$p=4$	$\frac{5}{18}$	0	$\frac{8}{9}$	$-\frac{1}{6}$

Таблиця В2. РК3(4)

$\frac{2}{7}$	$\frac{2}{7}$				
$\frac{7}{15}$	$\frac{77}{900}$	$\frac{343}{900}$			
$\frac{35}{38}$	$\frac{805}{1444}$	$-\frac{77175}{54872}$	$\frac{97125}{54872}$		
1	$\frac{79}{490}$	0	$\frac{2175}{3626}$	$\frac{2166}{9065}$	
$p=3$	$\frac{79}{490}$	0	$\frac{2775}{3626}$	$\frac{2166}{9065}$	
$p=4$	$\frac{229}{1470}$	0	$\frac{1125}{1813}$	$\frac{13718}{81585}$	$\frac{1}{18}$

Таблиця В3. РК4(5)

$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$			
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$		
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	
$p=4$	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	
$p=5$	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$

Таблиця В4. РК5(6)

$\frac{1}{6}$	$\frac{1}{6}$						
$\frac{4}{15}$	$\frac{4}{75}$	$\frac{16}{75}$					
$\frac{2}{3}$	$\frac{5}{6}$	$-\frac{8}{3}$	$\frac{5}{2}$				
$\frac{4}{5}$	$-\frac{8}{5}$	$\frac{144}{25}$	-4	$\frac{16}{25}$			
1	$\frac{361}{320}$	$-\frac{18}{5}$	$\frac{407}{128}$	$-\frac{11}{80}$	$\frac{55}{128}$		
0	$-\frac{11}{640}$	0	$\frac{11}{256}$	$-\frac{11}{160}$	$\frac{11}{256}$	0	
1	$\frac{93}{640}$	$-\frac{18}{5}$	$\frac{803}{256}$	$-\frac{11}{160}$	$\frac{99}{256}$	0	1
$p=5$	$\frac{31}{384}$	0	$\frac{1125}{2816}$	$\frac{9}{32}$	$\frac{125}{768}$	$\frac{5}{66}$	
$p=6$	$\frac{7}{1408}$	0	$\frac{1125}{2816}$	$\frac{9}{32}$	$\frac{125}{768}$	0	$\frac{5}{66}$	$\frac{5}{66}$

Таблица B5. РК6(7)

$\frac{2}{33}$	$\frac{2}{33}$									
$\frac{4}{33}$	0	$\frac{4}{33}$								
$\frac{2}{11}$	$\frac{1}{22}$	0	$\frac{3}{22}$							
$\frac{1}{2}$	$\frac{43}{64}$	0	$-\frac{165}{64}$	$\frac{77}{32}$						
$\frac{2}{3}$	$-\frac{2383}{486}$	0	$\frac{1067}{54}$	$-\frac{26312}{1701}$	$\frac{2176}{1701}$					
$\frac{6}{7}$	$\frac{10077}{4802}$	0	$-\frac{5643}{686}$	$\frac{116259}{16807}$	$-\frac{6240}{16807}$	$\frac{1053}{2401}$				
1	$-\frac{733}{176}$	0	$\frac{141}{8}$	$-\frac{335763}{23296}$	$\frac{216}{77}$	$-\frac{4617}{2816}$	$\frac{7203}{9152}$			
0	$\frac{15}{352}$	0	0	$-\frac{5445}{46592}$	$\frac{18}{77}$	$-\frac{1215}{5632}$	$\frac{1029}{18304}$	0		
1	$-\frac{1833}{352}$	0	$\frac{141}{8}$	$-\frac{51237}{3584}$	$\frac{18}{7}$	$-\frac{729}{512}$	$\frac{1029}{1408}$	0	1	
$p = 6$	$\frac{77}{1440}$	0	0	$\frac{1771561}{6289920}$	$\frac{32}{105}$	$\frac{243}{2560}$	$\frac{16807}{74880}$	$\frac{11}{270}$		
$p = 7$	$\frac{11}{864}$	0	0	$\frac{1771561}{6289920}$	$\frac{32}{105}$	$\frac{243}{2560}$	$\frac{16807}{74880}$	0	$\frac{11}{270}$	$\frac{11}{270}$

Таблиця В6. РК7(8)

$\frac{2}{27}$	$\frac{2}{27}$												
$\frac{1}{9}$	$\frac{1}{36}$	$\frac{1}{12}$											
$\frac{1}{6}$	$\frac{1}{24}$	0	$\frac{1}{8}$										
$\frac{5}{12}$	$\frac{5}{12}$	0	$-\frac{25}{16}$	$\frac{25}{16}$									
$\frac{1}{2}$	$\frac{1}{20}$	0	0	$\frac{1}{4}$	$\frac{1}{5}$								
$\frac{5}{6}$	$-\frac{25}{108}$	0	0	$\frac{125}{108}$	$-\frac{65}{27}$	$\frac{125}{54}$							
$\frac{1}{6}$	$\frac{31}{300}$	0	0	0	$\frac{61}{225}$	$-\frac{2}{9}$	$\frac{13}{900}$						
$\frac{2}{3}$	2	0	0	$-\frac{53}{6}$	$\frac{704}{45}$	$-\frac{107}{9}$	$\frac{67}{90}$	3					
$\frac{1}{3}$	$-\frac{91}{108}$	0	0	$\frac{23}{108}$	$-\frac{976}{135}$	$\frac{311}{54}$	$-\frac{19}{60}$	$\frac{17}{6}$	$-\frac{1}{12}$				
1	$\frac{2383}{4100}$	0	0	$-\frac{341}{164}$	$\frac{4496}{1025}$	$-\frac{301}{82}$	$\frac{2133}{4100}$	$\frac{45}{82}$	$\frac{45}{164}$	$\frac{18}{41}$			
0	$\frac{3}{205}$	0	0	0	0	$-\frac{6}{41}$	$-\frac{3}{205}$	$-\frac{3}{41}$	$\frac{3}{41}$	$\frac{6}{41}$	0		
1	$-\frac{1777}{4100}$	0	0	$-\frac{341}{164}$	$\frac{4496}{1025}$	$-\frac{289}{82}$	$\frac{2193}{4100}$	$\frac{51}{82}$	$\frac{33}{164}$	$\frac{12}{41}$	0	1	
$p = 7$	$\frac{41}{840}$	0	0	0	0	$\frac{34}{105}$	$\frac{9}{35}$	$\frac{9}{35}$	$\frac{9}{280}$	$\frac{9}{280}$	$\frac{41}{840}$		
$p = 8$	0	0	0	0	0	$\frac{34}{105}$	$\frac{9}{35}$	$\frac{9}{35}$	$\frac{9}{280}$	$\frac{9}{280}$	0	$\frac{41}{840}$	$\frac{41}{840}$

Додаток Г. Коефіцієнти багатокрокових різницевих схем

Таблиця Г1

Явні РС Адамса (Адамса–Башфорта) порядку $p = m$,

$$y_n = y_{n-1} + \sum_{\nu=1}^{m} b_\nu f(t_{n-\nu}, y_{n-\nu}),\ C \text{ – стала у ГСП апроксимації РС } Ch^m u^{(m+1)}$$

m	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	C
1	1								$-\frac{1}{2}$
2	$\frac{3}{2}$	$-\frac{1}{2}$							$\frac{5}{12}$
3	$\frac{23}{12}$	$-\frac{4}{3}$	$\frac{5}{12}$						$-\frac{3}{8}$
4	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{3}{8}$					$\frac{251}{720}$
5	$\frac{1901}{730}$	$-\frac{1387}{360}$	$\frac{109}{30}$	$-\frac{637}{360}$	$\frac{251}{720}$				$-\frac{95}{288}$
6	$\frac{4277}{1440}$	$-\frac{2641}{480}$	$\frac{4991}{720}$	$-\frac{3619}{720}$	$\frac{959}{480}$	$-\frac{95}{288}$			$\frac{19087}{60480}$
7	$\frac{198721}{60480}$	$-\frac{18637}{2520}$	$\frac{135183}{20160}$	$-\frac{10754}{945}$	$\frac{135713}{20160}$	$-\frac{5603}{2520}$	$\frac{19087}{60480}$		$-\frac{5257}{17280}$
8	$\frac{16083}{4480}$	$-\frac{1152169}{120960}$	$\frac{242653}{13440}$	$-\frac{296053}{13440}$	$\frac{2102243}{120960}$	$-\frac{115747}{13440}$	$\frac{32863}{13440}$	$-\frac{5257}{17280}$	$\frac{1070017}{3628800}$

Таблиця Г2

Неявні РС Адамса (Адамса–Мултона) порядку $p = m+1$

$$y_n = y_{n-1} + \sum_{\nu=0}^{m} b_\nu f(t_{n-\nu}, y_{n-\nu}).$$

m	b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7	C
0	1								$\frac{1}{2}$
1	$\frac{1}{2}$	$\frac{1}{2}$							$-\frac{1}{12}$
2	$\frac{5}{12}$	$\frac{2}{3}$	$-\frac{1}{12}$						$\frac{1}{24}$
3	$\frac{3}{8}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$					$-\frac{19}{720}$
4	$\frac{251}{720}$	$\frac{323}{360}$	$-\frac{11}{30}$	$\frac{53}{360}$	$-\frac{19}{720}$				$\frac{3}{160}$
5	$\frac{95}{288}$	$\frac{1427}{1440}$	$-\frac{133}{240}$	$\frac{241}{720}$	$-\frac{173}{1440}$	$\frac{3}{160}$			$-\frac{863}{60480}$
6	$\frac{19087}{60480}$	$\frac{2713}{2520}$	$-\frac{15487}{20160}$	$\frac{586}{945}$	$\frac{6737}{20160}$	$\frac{263}{2520}$	$-\frac{863}{60480}$		$\frac{275}{24192}$
7	$\frac{5257}{17280}$	$\frac{139849}{120960}$	$-\frac{4511}{4480}$	$\frac{123133}{120960}$	$-\frac{88547}{120960}$	$\frac{1537}{4480}$	$-\frac{11351}{120960}$	$\frac{275}{24192}$	$-\frac{33953}{3628800}$

Таблиця ГЗ

Багатокрокові РС Нюстрема, Мілна, Хемінга

№	Формули	Порядок	Головна складова похибки на кроці y_n-u_n
Явні різницеві схеми Нюстрема			
5.	$y_n = y_{n-2} + 2hf_{n-1}$	2	$-h^3 u_n^{(3)} / 3$
6.	$y_n = y_{n-2} + h(7f_{n-1} - 2f_{n-2} + f_{n-3})/3$	3	$-h^4 u_n^{(4)} / 3$
7.	$y_n = y_{n-2} + h(8f_{n-1} - 5f_{n-2} + 4f_{n-3} - f_{n-4})/3$	4	$-29h^5 u_n^{(5)} / 90$
Явні різницеві схеми Мілна			
8.	$y_n = y_{n-4} + 4h(2f_{n-1} - f_{n-2} + 2f_{n-3})/3$	4	$-14h^5 u_n^{(5)} / 45$
9.	$y_n = y_{n-3} + 3h(7f_{n-1} - 3f_{n-2} + 5f_{n-3} - f_{n-4})/8$	4	$-27h^5 u_n^{(5)} / 80$
Явні різницеві схеми Хемінга			
10.	$y_n = (y_{n-1} + y_{n-2})/2 + h(119f_{n-1} - 99f_{n-2} + 69f_{n-3} - 17f_{n-4})/48$	4	$-161h^5 u_n^{(5)} / 480$
11.	$y_n = (2y_{n-2} + 4y_{n-3})/3 + h(191f_{n-1} - 107f_{n-2} + 109f_{n-3} - 25f_{n-4})/72$	4	$-707h^5 u_n^{(5)} / 2160$
Неявні різницеві схеми Мілна			
12.	$y_n = y_{n-2} + h(f_n + 4f_{n-1} + f_{n-2})/3$	4	$h^5 u_n^{(5)} / 90$
13.	$y_n = y_{n-3} + 3h(f_n + 3f_{n-1} + 3f_{n-2} + f_{n-3})/8$	4	$3h^5 u_n^{(5)} / 80$
Неявні різницеві схеми Хемінга			
18.	$y_n = (y_{n-1} + y_{n-2})/2 + h(17f_n + 51f_{n-1} + 3f_{n-2} + f_{n-3})/48$	4	$3h^5 u_n^{(5)} / 160$
19.	$y_n = (2y_{n-2} + y_{n-3})/3 + h(25f_n + 91f_{n-1} + 43f_{n-2} + 9f_{n-3})/72$	4	$43h^5 u_n^{(5)} / 2160$

Таблиця Г4

Коефіцієнти $A(\alpha)$ – стійких чисто неявних різницевих схем порядку $p = m$

$$y_n = \sum_{\nu=1}^{m} \alpha_\nu y_{n-\nu} + h\beta f_n,$$

$C = \beta / (p+1)$ – стала в головній складовій похибки на кроці $Ch^{m+1}u^{(m+1)}$

m	α	β	α_1	α_2	α_3	α_4	α_5	α_6	C
1	90°	1	1						$\frac{1}{2}$
2	90°	$\frac{2}{3}$	$\frac{4}{3}$	$-\frac{1}{3}$					$\frac{2}{9}$
3	$86^\circ 03'$	$\frac{6}{11}$	$\frac{18}{11}$	$-\frac{9}{11}$	$\frac{2}{11}$				$\frac{3}{22}$
4	$73^\circ 35'$	$\frac{12}{25}$	$\frac{48}{25}$	$-\frac{36}{25}$	$\frac{16}{25}$	$\frac{3}{25}$			$\frac{12}{125}$
5	$51^\circ 84'$	$\frac{60}{137}$	$\frac{300}{137}$	$-\frac{300}{137}$	$\frac{200}{137}$	$-\frac{75}{137}$	$\frac{12}{137}$		$\frac{10}{137}$
6	$17^\circ 84'$	$\frac{60}{147}$	$\frac{360}{147}$	$-\frac{450}{147}$	$\frac{400}{147}$	$-\frac{225}{147}$	$\frac{72}{147}$	$-\frac{10}{147}$	$\frac{20}{343}$

Навчальне видання

Ярослав Йосипович Бігун

Числові методи

Навчальний посібник

Літературний редактор **Лупул О.В.**

Комп’ютерна верстка **Романенко Н.В.**

Дизайн обкладинки **Книгніцької Т.В.**

Підписано до друку . .2018. Формат 60х84/16.
Папір офсетний. Друк різографічний. Умов.-друк. арк. .
Обл.-вид. арк. . Тираж 100. Зам. Н- .

Видавництво та друкарня Чернівецького національного університету.
58012, Чернівці, вул. Коцюбинського, 2.
Свідоцтво суб’єкта видавничої справи ДК № 891 від 08.04.2002.