В. О. Любчак
Л. Д. Назаренко

Методи та алгоритми обчислень

Міністерство освіти і науки України
Сумський державний університет

Любчак В.О., Назаренко Л.Д.

МЕТОДИ ТА АЛГОРИТМИ ОБЧИСЛЕНЬ

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

УДК ~~519.6 (075.8)~~
Л93

Рекомендовано Міністерством освіти і науки України (лист № 1.4/18 – Г - 950 від 06.05.2008р.)

Рецензенти:
д-р фіз.-мат. наук, проф. К.Г. Малютін
(Сумський національний аграрний університет);
д-р фіз.-мат. наук, проф. А.М. Чорноус
(Сумський державний університет);
д-р техн. наук, проф. Є.А. Лавров
(Сумський національний аграрний університет).

Любчак В.О. , Назаренко Л.Д.
Л93 Методи та алгоритми обчислень: Навчальний посібник.-Суми: Вид-во СумДУ, 2008.- 313 с.

ISBN 978-966-657-194-9

У посібнику розглянуті найбільш часто використовувані методи та алгоритми чисельної реалізації розв'язування прикладних задач. Наведені приклади та сформульовані завдання для практичних робіт з відповідного курсу.

Для студентів вищих навчальних закладів та широкого кола

УДК 519.6 (075.8)

ISBN 978-966-657-194-9

Зміст

Вступ

Сучасний стан дослідження різноманітних процесів, що взаємодіють між собою, вимагає обґрунтування та побудови складних обчислювальних алгоритмів. З цією метою використовується широкий арсенал чисельних методів. Чисельними методами називають такий розділ математики, у якому предметом вивчення є методи одержання числових, з певним ступенем точності, розв'язків задач, що виникають як у самій математиці, так і в різних її додатках. Неточні (наближені) розв'язки прикладних задач заміняють їх точні розв'язки у практичних або теоретичних застосуваннях.

Створення й удосконалення швидкодіючої обчислювальної техніки дозволило розв'язати багато актуальних і складних прикладних задач, сприяло тому, що чисельні методи перетворилися в життєво необхідну сферу знань. З іншого боку, розвиток обчислювальної техніки став стимулом для критичної переоцінки та вдосконалення існуючих і створення нових чисельних методів. Найважливішим чинником при оцінці ефективності будь-якого чисельного методу в наш час є зручність його реалізації на ЕОМ. Оскільки більшість прикладних задач розв'язується на ЕОМ і надалі кількість таких задач буде збільшуватися, однією з основних проблем є подальше вдосконалення принципів і прийомів користування ЕОМ, полегшення спілкування людини з машиною.

У запропонованому Вашій увазі посібнику викладаються основні групи методів, що найчастіше застосовуються для чисельної реалізації прикладних задач. Розглядаються теоретичні основи методів, оцінки отриманих результатів та комп'ютерна реалізація

чисельних алгоритмів. Для адаптації алгоритмів до різноманітних мов програмування, що використовуються зараз, пропонується використання псевдокоду. Псевдокод являє собою систему позначень і правил, призначену для одноманітного запису алгоритмів. З одного боку, він близький до звичайної природної мови, тому алгоритми можуть на ньому записуватися і читатися як звичайний текст. З іншого, - у псевдокоді використовуються деякі формальні конструкції і математична символіка, що наближає запис алгоритму до загальноприйнятого математичного запису. У псевдокоді не прийняті строгі синтаксичні правила для запису команд, властиві формальним мовам, що полегшує запис алгоритму на стадії його проектування і дає можливість використовувати більш широкий набір команд, розрахований на абстрактного виконавця. Однак використовуються деякі конструкції, властиві формальним мовам, що полегшує перехід від запису на псевдокоді до алгоритму формальною мовою.

Викладення методів та алгоритмів супроводжується прикладами чисельного розв'язання задач із використанням сучасних програмних пакетів, таких як Mathcad, Excel тощо. До кожного розділу посібника наводяться контрольні питання та завдання. Все це може зробити його корисним для використання як з метою ознайомлення з чисельними методами, так і для їх практичного застосування для розв'язання конкретних прикладних задач.

Розділ 1

Основні проблеми чисельного розв'язання задач

При застосуванні чисельних методів розв'язки задач виявляються, як правило, наближеними. Пояснюється це в багатьох випадках тим, що точні методи їх розв'язання дотепер невідомі. Крім того, навіть при застосуванні точного методу задовольняються наближеним розв'язком, зокрема, з таких причин:

— точний розв'язок виявляється трудомістким; тоді як наближений при істотно меншому об'ємі обчислень виявляється цілком прийнятним за своїм характером;

— точність отриманого результату не відіграє істотної ролі, тому що в будь-якому разі заокруглюється до цілого числа (наприклад, при визначенні кількості механізмів, необхідних для виконання даного обсягу робіт).

Наближений розв'язок задачі повинен «не набагато відрізнятися» від точного розв'язку, інакше ним не можна скористатися з конкретною метою. Що означає термін «не набагато відрізняється» або, інакше кажучи, що варто розуміти під неточністю (наближеністю) розв'язку? Кожен чисельний метод дозволяє оцінювати ступінь неточності розв'язку, одержуваного цим методом. У курсі чисельних методів ступінь неточності розв'язку характеризується поняттям похибки розв'язку. Потрібно зазначити, що теорія похибок є одним із основних розділів обчислювальної математики. Очевидно, що відхилення наближеного результату від точного напряму залежить від коректності поставленої задачі та від наявних вхідних даних. Тому актуальним є дослідження збіжності наближеного розв'язку, що пропонує чисельний алгоритм, до точного розв'язку поставленої задачі .

Таким чином, основними проблемами чисельного розв'язання задач можна вважати:
-проблему оцінки похибки наближеного розв'язку;
-проблему коректності та обумовленості поставленої задачі;
-проблему збіжності наближеного методу до точного.

1.1 Класифікація похибок

При розв'язанні прикладних задач дуже важливо мати уявлення про точність отриманих результатів. Похибки, що можуть бути закладені в таких результатах, утворюються з багатьох причин.

Можна визначити чотири основні джерела похибок. результату чисельного методу:

1) вхідні дані;
2) математична модель;
3) наближений метод;
4) округлення при розрахунках.

Проаналізуємо їх.

Похибки вхідних даних

Точні значення багатьох величин практично ніколи не можуть бути введені в процес обчислень, наприклад, ірраціональних величин π, e, $\sqrt{2}$ та ін. У цих випадках неминучі похибки округлення.

При розв'язанні багатьох задач за вхідні беруться значення величин, отриманих з експерименту. З багатьох причин, у тому числі обмеженої точності вимірювальної апаратури і впливу різних випадкових чинників, експериментальні дані завжди мають похибки того або іншого порядку. Так, точність вимірювання температури, відстані, об'єму, ваги залежить від досконалості застосовуваних вимірювальних приладів. Похибки можуть бути у вхідних даних, отриманих теоретично. Природно,

що вони впливають на результати розв'язку задачі, однак жодним чином їх усунути не можна. Тому похибки такого типу часто називають *неусувними*.

Похибки математичної моделі

Необхідно зазначити, що в більшості випадків фахівцю вдається підібрати для розв'язання задачі наближений метод, що дозволяє одержати цілком задовільні за ступенем точності результати. Однак розв'язувана задача є не тим реальним завданням, з яким фахівцю доводиться мати справу, а його спрощеною математичною моделлю. Так, при розрахунку авіаційного двигуна або несучої конструкції промислової споруди неможливо ввести до розгляду їх реальну надзвичайно складну форму, врахувати наявність усіх отворів, деталей сполучення і т.п. При визначенні оптимального складу персоналу універмагу, кас попереднього продажу залізничних квитків доводиться припускати, що покупці приходять через рівні проміжки часу, час обслуговування кожного з них однаковий і таке інше.

Розв'язок реальної задачі не збігається із результатом, отриманим при розгляді її математичної моделі навіть із застосуванням точних методів розв'язку, а похибки, що виникають при цьому, можна назвати *похибками математичного моделювання*.

Похибки наближеного методу

У випадку, коли розв'язати задачу точно неможливо, доводиться застосовувати різні наближені методи. Результати такого підходу завчасно містять похибки, характер яких залежить від використовуваного наближеного методу (*похибки методу*).

При застосуванні наближених методів розв'язання задач, наприклад ітераційних, точні значення шуканих величин можуть бути отримані тільки після виконання нескінченного числа етапів обчислень, що практично

здійснити неможливо. Доводиться задовольнятися певним числом етапів і відповідними наближеними результатами із так званими *залишковими похибками*.

Похибки заокруглень при розрахунках

При реалізації на ЕОМ алгоритмів, що містять велику кількість операцій множення і ділення, типовими є *похибки округлення*. При виконанні операцій множення кількість розрядів може зрости настільки, що всі вони вже не можуть бути розміщені в елементах запам'ятовуючих пристроїв ЕОМ. Частину розрядів праворуч доводиться відкидати, округляти числа. Сам по собі процес округлення числа не обов'язково призводить до внесення в нього якої-небудь істотної похибки. Так, при обчисленні зі звичайною точністю в сучасних ЕОМ можна угримувати, наприклад, дев'ять десяткових розрядів. Природно, що простим відкиданням в ЕОМ десятого і наступних розрядів ми вносимо в число лише дуже незначні зміни. Порівняємо дванадцятирозрядне число 1000000,00297 і округлене дев'ятирозрядне число 1000000,00. Внесена в результаті округлення похибка становить величину 0,00297. Однак у процесі виконання великої кількості арифметичних операцій похибки, послідовно накопичуючись, породжують нові. Таке нагромадження похибок округлення може призвести до дуже істотних помилок в остаточних результатах.

Похибки округлення особливо доводиться враховувати при реалізації нестійких обчислювальних процесів, у яких незначні похибки у вихідних даних або результатах проміжних обчислень можуть призвести до істотних помилок у остаточному результаті.

Приклад. Нехай необхідно обчислити величину c за формулою

$$c = a - b, \qquad (1.1)$$

де $a = 139{,}27$; $b = 138{,}97$. Одержимо $c = 0{,}3$.

Припустимо, що величини a і b обчислені з похибками, що не перевищують 1% їх точних значень, a=140,62, b=37,62. Обчислюючи величину c за формулою (1.1) із наближеними значеннями, одержимо c=140,62-137,62=3,0. Отже, похибки в обчисленні вихідних величин a і b призвели до десятикратного збільшення числа c.

1.2 Абсолютна і відносна похибки

Абсолютна похибка - це модуль різниці між відповідним точним значенням розглянутої величини A і наближеним її значенням a. Вона має вигляд

$$\Delta = |A - a|. \qquad (1.2)$$

Безпосередньо за значенням абсолютної похибки досить важко робити висновок про ступінь розбіжності між точним значенням A величини і його наближеним значенням. Так, похибка 2*м* цілком припустима при визначенні відстані між Києвом і Сумами та абсолютно неприпустима при вимірюванні розмірів кімнати. Тому застосовується ще одна характеристика наближених величин — їх відносна похибка.

Відносною похибкою δ наближеного значення величини, точне значення якої дорівнює A, називається відношення його абсолютної похибки Δ до модуля точного значення, тобто

$$\delta = \frac{\Delta}{|A|}. \qquad (1.3)$$

Наприклад, нехай в результаті вимірювання довжини бігової доріжки отримано значення a=99,1м. Точне значення цієї величини A = 100м. Абсолютна

похибка $\Delta = |100 - 99{,}1| = 0{,}9$. Відносна похибка за формулою (1.3) становить $\delta = \dfrac{0{,}9}{|100|} = 0{,}009$.

Із формул (1.2)–(1.3) бачимо, що абсолютна похибка має розмірність оцінюваних цією похибкою величин, відносна похибка завжди безрозмірна.

Величини Δ і δ можуть бути обчислені точно лише в тих випадках, коли відоме не тільки наближене числове значення розглянутої величини, але і її точне значення. Останнє, однак, можливе далеко не у всіх випадках. Крім того, часто доводиться аналізувати похибки деякої множини наближених величин, наприклад, похибки вимірювання розмірів серії виготовлених деталей, викликані недосконалістю застосовуваних вимірювальних інструментів. Якість серії вимірювань для всіх деталей може оцінюватися найбільшою за модулем величиною абсолютної або відносної похибки їх розмірів. Тому часто вводяться поняття граничних абсолютної та відносної похибок.

За *граничну абсолютну похибку* Δ^* наближеного числа може бути взяте будь-яке число, не менше абсолютної похибки цього числа,

$$\Delta^* \geq \Delta. \qquad (1.4)$$

Аналогічно за *граничну відносну похибку* δ^* наближеного числа може бути взяте будь-яке число, що задовольняє умову

$$\delta^* \geq \delta. \qquad (1.5)$$

При аналізі серії вимірювань за Δ^* і δ^* беремо найбільші з отриманих відповідних значень Δ і δ і тим самим визначаємо межі, всередині яких знаходяться відповідні похибки.

Значущими цифрами числа a називають усі цифри в його записі, починаючи з першої ненульової зліва. Значущу цифру числа a називають *правильною*, якщо абсолютна похибка числа не перевищує одиниці відповідного цій цифрі розряду.

Приклад 1 Для ряду $\sum_{n=0}^{\infty} \frac{72}{n^2+5n+4}$ знайти суму S аналітично. Обчислити значення часткових сум ряду $S_N = \sum_{n=0}^{N} a_n$ і знайти величину похибки при значеннях $N = 10, 10^2, 10^3, 10^4, 10^5$. Побудувати гістограму залежності правильних цифр результату від N.

Знайдемо точну суму цього ряду:

$$S_N = \sum_{n=0}^{N} \frac{72}{n^2+5n+4} = \sum_{n=0}^{N} \frac{72}{(n+1)(n+4)} =$$

$$= 72 \cdot \sum_{n=0}^{N} \frac{1}{3} \cdot \left(\frac{1}{n+1} - \frac{1}{n+4} \right) = = 24 \cdot \left(1 + \frac{1}{2} + \frac{1}{3} - \frac{1}{N+2} - \frac{1}{N+3} - \frac{1}{N+4} \right),$$

Отже, $S = \lim_{N \to \infty} S_N = 44$. Уведемо функцію часткових сум $S(N) = \sum_{n=0}^{N} \frac{72}{n^2+5n+4}$. Тоді абсолютну похибку можна визначити за допомогою функції $d(N) = |S(N) - S|$.

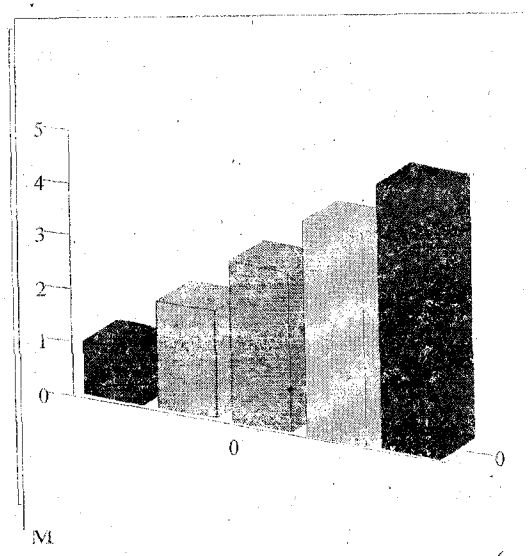
Результати обчислювального експерименту

N	Значення частк. суми ряду $S(N)$	Абсолютна похибка $d(N)$	Кільк. правил. цифр M_i
10	$S(10)$=38.439560439	$d(10)$ =5.56	$M_1 = 1$
10^2	$S(100)$=43.3009269	$d(100)$=0.699	$M_2 = 2$

10^3	$S(1000)=43.9282153$	$d(1000)=0.072$	$M_3=3$
10^4	$S(10000)=43.992802$	$d(10000)=0.0072$	$M_4=4$
10^5	$S(100000)=43.99928021599$	$d(100000)=0.00072$	$M_5=5$

Висновок. Як бачимо з наведеного обчислювального експерименту, збільшення числа членів ряду в 10 разів порівняно з попереднім випадком збільшує число правильних цифр у відповіді на 1.

Гістограма



Приклад 2 Для матриці $A=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$ розв'язати питання про існування оберненої матриці в таких випадках:

1) елементи матриці задані точно;

2) елементи матриці задані наближено з відносною похибкою a) $\delta=\alpha\%$ та b) $\delta=\beta\%$. Знайти відносну похибку результату.

Це питання вирішується шляхом знаходження визначника й порівняння його з нулем. У випадку, коли елементи визначника задані точно, варто обчислити визначник і правильно відповісти на поставлене в задачі питання.

У випадку, коли елементи визначника задані наближено з відносною похибкою δ, питання є складнішим. Нехай елементи матриці позначені через a_{ij}. Тоді кожен елемент матриці a_{ij} тепер уже не дорівнює конкретному значенню, а може набувати будь-якого значення з відрізка $[a_{ij}(1-\delta);\ a_{ij}(1+\delta)]$, якщо $a_{ij}>0$, і з відрізка $[a_{ij}(1+\delta);\ a_{ij}(1-\delta)]$, якщо $a_{ij}<0$. Множина всіх можливих значень елементів матриці являє собою замкнену обмежену множину в 9-вимірному просторі. Сам визначник є неперервною й диференційованою функцією 9 змінних - елементів матриці a_{ij}. За відомою теоремою Вейєрштрасса ця функція досягає на зазначеній множині свого найбільшого та найменшого значень M і m. Якщо відрізок $[m, M]$ не містить точку 0, то це означає, що при будь-яких припустимих значеннях елементів матриці a_{ij} визначник не набуває значення 0. Якщо ж точка 0 належить відрізку $[m, M]$, таке твердження буде неправомірним. Буде мати місце невизначеність.

З'ясувати m і M допомагають наступні міркування. Як функція своїх аргументів (елементів матриці a_{ij}) визначник має таку властивість (принцип максимуму): ця функція досягає свого найбільшого і найменшого значень завжди на границі області. Більше того, можна довести,

що ці значення досягаються в точках, координати яких мають вигляд $a_{ij}(1\pm\delta)$. Таких точок $2^9=512$. У кожній з них варто обчислити визначник, а потім вибрати з отриманих значень найбільше та найменше. Це й будуть числа M і m.

1.3 Середні квадратичні похибки

Нехай передбачається проведення серії вимірів деякої величини X. У кожному з вимірів буде отримане якесь її значення, причому залежно від точності приладу, зокрема, ці значення будуть знаходитися в деякому інтервалі, загальне їх число скінченне. Позначимо ці значення $x_1, x_2, ..., x_n$, їх ймовірності $p_1, p_2, ... p_n$. Оскільки заздалегідь невідомо, яке значення величини X буде отримано в кожному вимірі, ця величина є випадковою.

Математичне очікування X виражається формулою

$$M[X]=\sum_{i=1}^{n} x_i p_i \ . \qquad (1.6)$$

Про якість вимірів, тобто ступінь розкиду помилок виміру, можна роб висновкиити за розмірами дисперсії, або середнього квадратичного відхилення випадкової величини:

$$D[X]=\sigma_x^2=\sum_{i=1}^{n} p_i (x_i - M[X])^2 . \qquad (1.7)$$

Величина σ_x називається в теорії похибок *середньою квадратичною похибкою* вимірювання.

Якщо результати вимірювання є незалежними, тобто результат довільного виміру не залежить від того,

які результати отримані в інших вимірах, для них прийнятні теореми Чебишева і Бернуллі. Зокрема, бувають наступні припущення.

1 Якщо випадкова величина X набуває тільки невід'ємних значень, частина яких менша деякого додатного числа a, то

$$p[(X<a)] \geq 1 - \frac{M[X]}{a} . \qquad (1.8)$$

2 Якщо $a > 0$, то

$$p[|X - M[X]| < a] \geq 1 - \frac{\sigma_x^2}{a} . \qquad (1.9)$$

Відзначимо, що формулою (1.7) користуються для обчислення середніх квадратичних похибок і в детермінованих процесах

$$\sigma_x^2 = \sum_{i=1}^{n} (x_i - A)^2 = \sum_{i=1}^{n} \Delta_i^2 , \qquad (1.10)$$

де A — точне значення числа X, а Δ_i — абсолютні похибки.

1.4 Поширення похибок

Важливим у чисельному аналізі є питання про те, як помилка, що виникла у визначеному місці в ході обчислень, поширюється далі, тобто чи стає її вплив більшим або меншим залежно від того, як виконуються наступні операції. Сформулюємо деякі правила оцінки похибок при виконанні операцій над наближеними числами:

- при додаванні або відніманні чисел їхні абсолютні похибки додаються;
- при множенні або діленні чисел їхні відносні похибки додаються.

Ці правила можна вивести безпосередньо. Нехай є два наближення a_1 і a_2 до чисел x_1 і x_2, а також відповідні абсолютні похибки Δa_1, Δa_2.

Оцінимо, наприклад, похибку суми

$$\Delta(a1+a2)=|(x1+x2)-(a1+a2)|=$$

$$=|(x1-a1)+(x2-a2)|\le|x1-a1|+|x2-a2|\le\Delta a1+\Delta a2.$$

Для визначення оцінок похибки арифметичних дій можна використовувати загальне правило оцінки похибки функції.

Розглянемо функцію $y=f(x)$. Нехай a – наближене значення аргумента x, Δa - його абсолютна похибка. Абсолютну похибку функції можна вважати її приростом, який можна замінити диференціалом $\Delta y \approx dy$. Тоді одержимо $\Delta y = |f'(a)|\Delta a$, $\delta y = |f'(a)\,/\,f(a)|\Delta a$.

Застосуємо загальне правило, наприклад, для оцінки похибки суми $f(x_1,x_2)=x_1+x_2$

$$\Delta(a_1+a_2)=\left|f'_{x_1}\right|*\Delta a_1+\left|f'_{x_2}\right|*\Delta a_2=\Delta a_1+\Delta a_2$$

та добутку $f(x_1,x_2)=x_1x_2$

$$\Delta(a_1a_2)=\left|f'_{x_1}(a_1,a_2)\right|\Delta a_1+\left|f'_{x_2}(a_1,a_2)\right|\Delta a_2=|a_2|\Delta a_1+|a_1|\Delta a_2.$$

Тут через a_1 і a_2 позначені значення величин x_1 і x_2, задані з абсолютними похибками Δa_1 і Δa_2.

Розглянемо віднімання двох майже рівних чисел. Запишемо вираз для відносної похибки різниці у вигляді

$$\delta(a_1-a_2)=\Delta(a_1-a_2)\,/\,|a_1-a_2|=(\Delta a_1+\Delta a_2)\,/\,|a_1-a_2|.$$

При $a_1\approx a_2$ ця похибка може бути як завгодно великою. Нехай a1=2520, a2=2518. Абсолютні похибки вихідних даних Δa1=Δa2=0.5; відносні похибки - δa1$\approx\delta$a2$\approx$0.002 (0.2%). Відносна похибку різниці буде дорівнювати δ(a1-a2)=(0.5+0.5)/2=0.5 (50%). Оскільки в

подальших обчисленнях ця велика відносна похибка буде поширюватися, може виявитися сумнівною точність остаточного результату обчислень.

Наведемо деякі оцінки

Дія (функц.)	Абсолютна похибка	Відносна похибка
$x1+x2$	$\Delta a1+\Delta a2$	$(\lvert a_1\rvert/\lvert a_1+a_2\rvert)\delta a_1+$ $+(\lvert a_2\rvert/\lvert a_1+a_2\rvert)\delta a_2$
$x1-x2$	$\Delta a1+\Delta a2$	$(\lvert a1\rvert/\lvert a1-a2\rvert)*\delta a1+$ $+(\lvert a2\rvert/\lvert a1-a2\rvert)*\delta a2$
$x1*x2$	$\lvert a_2\rvert\Delta a_1+\lvert a_1\rvert\Delta a_2$	$\delta a_1+\delta a_2$
$x1/x2$	$\Delta a_1/\lvert a_2\rvert+\Delta a_2\lvert a_1\rvert/(a_2)^2$	$\delta a_1+\delta a_2$
x^n	$\Delta a*\lvert n*a^{n-1}\rvert$	$\lvert n\rvert\delta a$

У той же час буває і так, що похибки чисел, що беруть участь у тому або іншому обчисленні, взаємно компенсуються. Врахувати це можливо, але досить складно. Для ознайомлення з різними питаннями наближених обчислень можна порекомендувати книги [1],[2].

1.5 Підвищення точності результатів обчислень (рекомендації)

Щоб зменшити можливу похибку результату при розв'язуванні задачі, рекомендується дотримуватися таких правил для практичної організації обчислень.

I Похибка суми кількох чисел при розрахунку на ЕОМ зменшиться, якщо починати додавання з менших за величиною доданків.

Якщо додається досить багато чисел, то їх краще розбити на групи з чисел близьких за величиною, провести додавання в групах за вищезгаданою рекомендацією, після чого отримані суми додати, починаючи з меншої.

Якщо задано n^2 додатних чисел приблизно однакової величини, то загальна помилка округлення зменшиться, якщо числа додати спочатку групами по n-чисел, а потім додати n – часткових сум. При великих n верхня межа округлення при такому способі становить всього $1/n$ від відповідної межі при довільному додаванні чисел одне до одного.

Причина того, що не виконується комутативний закон додавання, полягає в округленні проміжних результатів, коли багатозначні числа не вміщуються в розрядну сітку ЕОМ. Тому і не все одно, в якому порядку необхідно виконувати арифметичні операції, щоб результат був якомога точнішим.

II Варто уникати віднімання двох майже однакових чисел. Обчислюючи різницю двох чисел, доцільно винести за дужки їхній спільний множник. Для прикладу обчислимо величину

$$P = 6.250001*16 - 25.000003*4 = 1*10^{-5}.$$

Винесемо число "4" за дужки, одержимо точний результат: $P = 4(6.250001*4 - 25.000003) = 4*10^{-6}$.

Зменшити похибку різниці дозволяють

перетворення:

$$(a+\varepsilon)^2 - a^2 = \varepsilon(2a+\varepsilon);$$

$$\sqrt{a+\varepsilon} - \sqrt{a} = \varepsilon/(\sqrt{a+\varepsilon} + \sqrt{a});$$

$$a - \sqrt{a^2+\varepsilon} = -\varepsilon/(a+\sqrt{a^2+\varepsilon});$$

$$1 - a/(a+\varepsilon) = \varepsilon/(a+\varepsilon).$$

Тут ε - мале в порівнянні з a число.

1.6 Машинна арифметика

В ЕОМ для кодування дійсних чисел використовується двійкова система зчислення й прийнята форма подання чисел із плаваючою точкою $x = \mu \cdot 2^p$, $\mu = \pm(\gamma_1 \cdot 2^{-1} + \gamma_2 \cdot 2^{-2} + ... + \gamma_t \cdot 2^{-t})$. Тут μ - мантиса ; $\gamma_1, \gamma_2, .. \gamma_t$ - двійкові цифри, причому завжди γ_1=1, p-ціле число, що називається двійковим порядком. Кількість t цифр, що приділяється для запису мантиси, називається розрядністю мантиси. Діапазон подання чисел в ЕОМ обмежений кінцевою розрядністю мантиси й значенням числа p. Усі числа, що подані в ЕОМ, задовольняють нерівності: $0 < X_0 \le |x| < X_\infty$, де $X_0 = 2^{-(p_{max}+1)}$, $X_\infty = 2^{p_{max}}$. Усі числа, за модулем більші X_∞, не подані в ЕОМ і розглядаються як машинна нескінченність. Усі числа, за модулем менші X_0, для ЕОМ не відрізняються від нуля й розглядаються як машинний нуль. Машинним іпсилоном ε_M називається відносна точність ЕОМ, тобто границя відносної похибки подання чисел в ЕОМ. Покажемо, що $\varepsilon_M \approx 2^{-t}$. Нехай $x^* = \mu \cdot 2^p$, тоді границя абсолютної похибки подання цього числа дорівнює

$\overline{\Delta}(x^*) \approx 2^{-t-1} \cdot 2^p$. Оскільки $\frac{1}{2} \le \mu < 1$, то величина відносної похибки подання оцінюється так:

$$\overline{\delta}(x^*) \approx \frac{\overline{\Delta}(x^*)}{|x^*|} \approx \frac{2^{-t-1} \cdot 2^p}{\mu \cdot 2^p} = \frac{2^{-t-1}}{\mu} \le \frac{2^{-t-1}}{2^{-1}} = 2^{-t}.$$

Машинний іпсилон визначається розрядністю мантиси та способом округлення чисел, реалізованим на ЕОМ.

Візьмемо такі способи визначення наближених значень параметрів, необхідних у задачі:

1 Покладемо $X_\infty = 2^n$, де n - перше натуральне число, при якому відбувається переповнення.

2 Покладемо $X_0 = 2^{-m}$, де m – перше натуральне число , при якому 2^{-m} збігається з нулем.

3 Покладемо $\varepsilon_M = 2^{-k}$, де k – найбільше натуральне число, при якому сума обчисленого значення $1+ 2^{-k}$ ще більша за 1. Фактично ε_M є границя відносної похибки подання числа $x^* \approx 1$.

Приклад. Для пакета MATHCAD знайти значення машинного нуля, машинної нескінченності, машинного іпсилон.

Машинна нескінченність $\inf(n) = 2^n$.

Машинний нуль $zero(m) = 2^{-m}$.

Машинний іпсилон $eps(k) = 2^{-k}$.

$\inf(1019) = 5.618 \times 10^{306}$; $zero(1019) = 1.78 \times 10^{-307}$;

$zero(3020) = 0$; $eps(50) = 8.8817841970018510^{-16}$.

Результати обчислювального експерименту:

Машинна нескінченність $X_\infty \approx 10^{307}$ Машинний нуль $X_0 \approx 10^{-306}$ Машинний іпсилон $\varepsilon_M \approx 10^{-15}$.

1.7 Обумовленість обчислювальної задачі

Під *обумовленістю* обчислювальної задачі розуміють чутливість її розв'язку до малих похибок вхідних даних. Нехай установлена нерівність $\Delta(y^*) \le \nu_\Delta \Delta(x^*)$, де $\Delta(x^*)$ - абсолютна похибка вхідних даних, а $\Delta(y^*)$ - абсолютна похибка розв'язку. Тоді ν_Δ називається абсолютним числом обумовленості задачі. Якщо ж установлена нерівність $\delta(y^*) \le \nu_\delta \delta(x^*)$ між відносними похибками даних і розв'язку, то ν_δ називають відносним числом обумовленості задачі.

Як правило, під *числом обумовленості* ν розуміють відносне число обумовленості. Якщо $\nu >> 1$, то задачу називають погано обумовленою.

Як уже зазначалося основними джерелами похибок в обчислювальних задачах є помилки у вихідних даних, помилки округлення під час машинних обчислень і обмеження точності використовуваної обчислювальної схеми (відмінність *обчислювальної схеми* від *чисельного методу* цілком символічна – у випадку класичних завдань чисельного аналізу ми говоримо про чисельний метод, *схему ж чисельного розв'язку конкретної прикладної задачі називаємо обчислювальною схемою*). Похибки, що виникають через помилки у вхідних даних, і похибки, пов'язані з округленнями, є неусувними з погляду обчислювача. У той же час при конструюванні обчислювальної схеми ми маємо деяку свободу у виборі використовуваних чисельних методів, які відрізняються один від одного *точністю* та *стійкістю*.

Точність чисельного методу, як правило, визначається порядком використовуваного в ньому наближення. Наприклад, при інтегруванні замість методу прямокутників можна використовувати метод парабол , що має більш високий порядок і забезпечує кращу точність. Часто також у нашому розпорядженні є параметр, що дозволяє керувати точністю одержуваного розв'язку (у випадку чисельного інтегрування це крок інтегрування). Точність обчислювальної схеми, як правило, обирають так, щоб відповідна похибка була принаймні вдвічі меншою за неусувну похибку у вхідних даних.

Важливо також розуміти значення стійкості використовуваної обчислювальної схеми. Похибки у вхідних даних і похибки округлення можуть по-різному впливати на точність результату залежно від використовуваної схеми обчислень. Під *стійкістю* обчислювальної схеми розуміють її стійкість стосовно похибок у вхідних даних і похибок округлення – для стійкої схеми малі похибки в даних і типові похибки округлення не призводять в остаточному підсумку до більших похибок. У той же час використання нестійкої обчислювальної схеми може призвести до значного зростання похибки результату.

Нестійкість також може мати різну природу. Іноді нестійкість обумовлена сутністю розв'язуваної задачі. У цьому випадку говорять, що задача *погано обумовлена* або *некоректна*. Більшість так званих *обернених задач* є прикладом поганої обумовленості. У той же час навіть у випадку коректної (добре обумовленої) задачі невдалий вибір обчислювальної схеми може призвести до нестійкості. Отже,

- нестійкість може бути властива завданню (слабка обумовленість);
- навіть добре обумовлену задачу можна зіпсувати

невдало підібраним чисельним методом.

1.8 Приклад втрати точності

Іноді втрата точності під час розв'язання задачі виявляється неминучою. Класичний приклад – пошук екстремуму функції.

Розглянемо задачу мінімізації функції $f(t)$ і припустимо, що мінімум досягається в точці t_0. В околі такої точки збільшення df функції $f(t)$ при збільшенні аргумента t на мале dt виражається так:
$df = f'(t_0)dt + (1/2)f''(t_0)dt^2 + o(dt^2) = (1/2)f''(t_0)dt^2 + o(dt^2)$, оскільки $f'(t_0) = 0$. Таким чином, в околі екстремуму збільшення функції пропорційно квадрату збільшення аргумента. З цього випливає принципова неможливість локалізувати екстремум точніше, ніж із половиною доступних значущих цифр – менші зміни аргумента не будуть приводити до помітної зміни значення функції.

Часто при розв'язанні рівняння $f(t) = 0$ пропонується перейти до еквівалентної задачі пошуку мінімуму функції $s(t) = [f(t)]^2$. Якщо мінімум досягається в точці t_0 і дорівнює нулю, то в цій точці рівняння має корінь. Легко зрозуміти, чому без достатніх підстав не варто застосовувати цей метод. Для лівої частини рівняння в околі простого кореня ми маємо представлення $df = f'(t_0)dt + 0(dt)$ і можемо локалізувати корінь із точністю, що збігається з точністю використовуваної машинної арифметики.

1.9 Погана обумовленість задачі

Погано обумовлені задачі є з погляду обчислювача великою проблемою. У таких задачах неминучі помилки

округлення й похибки у вхідних даних катастрофічно наростають, роблячи одержувану відповідь неприйнятною. **Погано обумовлені задачі становлять небезпеку при отриманні числового розв'язку. Потрібно вміти виявляти погану обумовленість і боротися з нею.**

Відмінна риса погано обумовлених задач така: *мале відхилення у вхідних даних призводить до суттєвих змін відповіді.* Якщо розглядати результат обчислень як функцію $y(x)$ від вхідних даних, то мале відхилення dx аргумента приведе до зміни обчисленого значення на dy, причому $|dy|$ можна оцінити як добуток модуля похідної та абсолютної величини відхилення: $|dy| = |y'(x)| \cdot |dx|$. У випадку поганої обумовленості малі зміни аргументу (вхідних даних) приводять до суттєвих змін у відповіді, тобто похідна $|y'(x)|$ набуває більших за модулем значень. Як правило, працюють із відносними приростами

$$Dx = dx/x,\ Dy = dy/y.$$

У цьому випадку $|Dy| = |f'(x)| \cdot (|x|/|y|) \cdot |Dx|$. Число $CD = |y'(x)| \cdot (|x|/|y|)$ називається *числом обумовленості* задачі. Отже,

$$|Dy| = CD \cdot |Dx| \text{ або } |dy| = |y| \cdot CD \cdot |Dx|.$$

Якщо помилки у вихідних даних обумовлені винятково використанням машинної арифметики, то $Dx = \varepsilon_{min}$, де ε_{min} – точність машинної арифметики. Машинним іпсилоном ε_{min} називається мінімальне додатне число, що при додаванні до числа 1.0 дає результат, більший ніж 1.0. Величина машинного іпсилона визначає відносну похибку введених чисел і, таким чином, є найважливішою характеристикою машинної арифметики. У цьому випадку $|Dy| = CD\varepsilon_{min}$ і $|dy| = |y|CD\varepsilon_{min}$. Таким чином, чим більше число обумовленості, тим сильніший

вплив похибки вхідних даних і похибки, внесеної округленнями.

Іноді дуже прості задачі демонструють погану обумовленість.

Приклад. Розглянемо задачу знаходження кореня многочлена

$$f(t)=(t-2)^9-t^9-18t^8+144t^7-672t^6+2016t^5-4032t^4+5376t^3-4608t^2+2304t-512.$$

Якщо многочлен заданий за допомогою першої з наведених формул, то задача знаходження кореня не становить проблеми. Нехай многочлен заданий за допомогою іншої формули, тобто задані його коефіцієнти й відомо, що один із коренів локалізований на відрізку [1,3]. Тоді значення многочлена можна обчислювати за схемою Горнера

$$f(t) = (((((((t - 18)t + 144)t - 672)t + 2016)t - 4032)t + 5376)t - 4608)t + 2304)t - 512,$$

а корінь можна знайти за методом дихотомії. Розглянемо поведінку функції на відрізку [1.97,2.03]. Обчислення виконувалися з подвійною точністю (double). Значення функції обчислювалися із кроком 0.0001. На першому малюнку показані результати при обчисленні за першою з наведених формул, а на другому – при обчисленні за схемою Горнера.

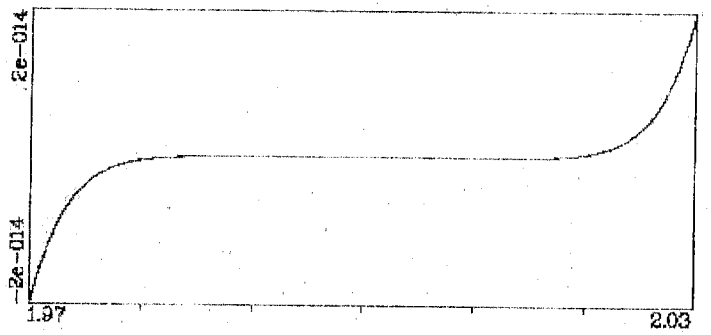


Рис. 1.1

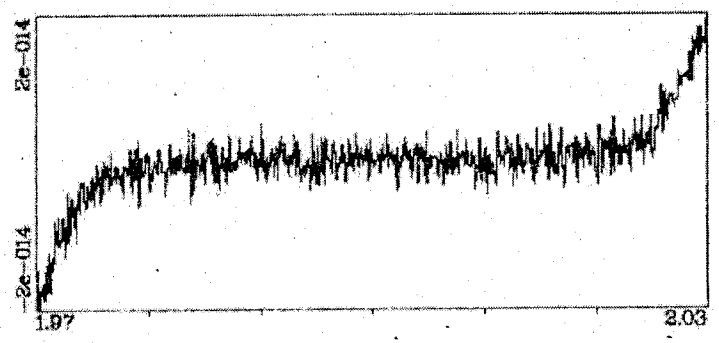


Рис. -1.2

Бачимо, що в околі точки t=2 функція, обчислена за схемою Горнера, поводиться непередбачувано. Якщо ми спробуємо знайти чисельно корінь рівняння $f(t) = 0$ для функції $f(t)$, що обчислюється за схемою Горнера, то одержимо випадкову точку з відрізка [1.98, 2.02]. На цьому відрізку зміна функції становить менш ніж 4e-14, але відрізок невизначеності для кореня виходить великий – довжиною 0.04.

Обмежимося простим і нестрогим зауваженням, що з'ясовує природу поганої обумовленості цієї задачі. Якщо точка t пробігає діапазон $[2-e,2+e]$ при деякому досить малому значенні e (наприклад, при e<0.01), то значення функції $f(t)$ змінюється дуже мало. Відповідно, мала зміна функції $f(t)$ (або коефіцієнтів многочлена) приведе до великої зміни значення кореня.

1.10 Прості тести на обумовленість

У погано обумовленій задачі відповідь дуже залежить від будь-яких похибок як у вхідних даних, так і внесених у процесі обчислень. Це дозволяє використовувати наступні прості тести на погану обумовленість.

•Можна спробувати змінити вхідні дані, вносячи в них малі випадкові помилки. У випадку поганої обумовленості відповідь повинна при цьому суттєво змінюватися.

•Можна спробувати повторити розрахунок з використанням машинної арифметики різної точності. Саме із цього погляду корисно визначати свій дійсний тип так, щоб зміна точності арифметики зводилася просто до перекомпіляції. Сильна залежність відповіді від точності використовуваної арифметики свідчить про слабку обумовленість.

•Можна внести штучну похибку у деякі із проміжних результатів. Для цього можна просто округляти обчислені значення до заданої кількості значущих цифр.

1.11 Загальна схема розв'язання задач чисельного аналізу. Апроксимація, стійкість, збіжність

Більшість задач чисельного аналізу в загальному вигляді можна записати у вигляді рівняння

$$y=F(x), \qquad (1.11)$$

де $F:X\to Y$ — деякий оператор, що задає відображення метричного простору X у метричний простір Y(дивись Додаток 1.1).

Загальний підхід, що реалізується в наближених методах розв'язання таких задач, полягає в заміні рівняння (1.11) близьким йому, простішим (як правило, скінченновимірним) рівнянням

$$y_n=F_n(x_n)\ . \qquad (1.12)$$

Тут $F_n:X_n\to Y_n$ – оператор, що відповідає вихідному оператору F. При цьому елементи $x_n \in X_n$ та $y_n \in Y_n$ розглядаються як образи елементів $x \in X$ та $y \in Y$. Цей роз'язок можна визначити через відповідні оператори

$$x_n=\varphi(x),\ y_n=\psi(y)\ . \qquad (1.13)$$

Як відомо, заміна одних математичних об'єктів іншими, чимось близькими до них, називається *апроксимацією*.

Визначення 1 Рівняння

$$F_n(x_n) = \psi_n(y) \qquad (1.14)$$

апроксимує рівняння 1.11 (оператор F_n апроксимує F), якщо для будь-яких елементів x з $D(F) \subseteq X$ міра апроксимації

$$\rho_{Y_n}(F_n(\varphi_n(x)), \psi_n(F(x))) \to 0 \text{, коли } n \to \infty . \qquad (1.15)$$

(Тут $\rho_{Y_n}(\alpha, \beta)$ визначає метрику, тобто відстань між елементами $\alpha, \beta \in Y_n$).

Щоб можна було порівнювати якість різних моделей вигляду (1.14) для задачі (1.11), користуються поняттям порядку апроксимації. Ця характеристика пов'язує прямування до 0 міри апроксимації (1.15) з порядком зменшення деякої залежної від n малої величини, наприклад, кроку апроксимації.

Будемо вважати, що розв'язки $x^* \in X$, та $x_n^* \in X_n$ рівнянь відповідно (1.11) та (1.14) існують і єдині. Наближеним розв'язком задачі (1.11) вважається елемент $x^{(n)} = \varphi_n^{-1}(x_n^*)$, де φ_n^{-1} обернене до φ відображення $x = \varphi_n^{-1}(x_n)$.

Головним питанням будь-якої теорії наближених методів розв'язання задач вигляду (1.11) є питання про те, чи можна наближеним розв'язком $x^{(n)}$ як завгодно добре відобразити поведінку точного розв'язку x^*. Це питання про збіжність $x^{(n)}$ до x^*.

Визначення 2 Має місце збіжність наближеного розв'язку $x^{(n)}$ до точного розв'язку x^* рівняння (1.11), якщо

$$\rho_X(x^*, x^{(n)}) \underset{n \to \infty}{\to} 0 .$$

Наявність фактичних оцінок величин

$\rho_{X_n}(x_n^*, \varphi_n(x^*))$ дозволяє не тільки робити висновки про збіжність наближених розв'язків, але і визначати похибки отриманих наближень до розв'язку.

Питання про збіжність розв'язків $y^{(n)}$ до y^* тісно пов'язане з тим, чи можна надійно розв'язати спрощену задачу (1.14). Адже спрощена задача також розв'язується наближено. Покращання якості апроксимації шляхом зменшення її міри (1.15) спричинює збільшення розмірності n для задачі (1.14), а отже, збільшення об'єму обчислень, що може призвести до збільшення обчислювальних похибок.

Визначення 3 Обчислювальний процес називається стійким, якщо малі похибки вхідних даних викликають малі похибки результатів.

Питання і завдання до розділу 1

1 Джерела й класифікація похибок. Наближені числа. Абсолютна й відносна похибки. Правильні й значущі цифри. Способи округлення.

2 Подання чисел в ЕОМ. Машинний нуль, машинна нескінченність, машинний епсилон. Алгоритми обчислення.

3 Похибки арифметичних операцій над наближеними числами.

4 Похибки обчислення функцій однієї та декількох змінних.

5 Похибки обчислення неявно заданої функції.

6 Числа a, b, c задані наближено: $a = 255.651, b = 0.9386, c = -5.1486$. Відомо, що $\Delta a = 2.1$, $\delta b = 0.02, \Delta c = 0.06$. Записати ці числа з усіма правильними знаками.

7 Наближене число a містить 5 правильних цифр. Що можна сказати про відносну похибку числа a?

8 З якою відносною похибкою потрібно знайти наближене значення числа a, щоб правильними виявилися 5 значущих цифр?

9 Для наближених чисел a та b $(a>b>0)$ відомо, що $\delta(a)=\delta(b)=\delta$. Оцінити похибки: a) $\delta(a+b)$, b) $\delta(a-b)$, c) $\delta(a*b)$, d) $\delta(a/b)$.

10 Числа a та b задані наближено: $a=1.137$, $b=1.073$, $\Delta a=\Delta b=0.011$. Оцінити похибки: а)різниці $c=a-b$, b) добутку $d=ab$. Записати відповідь з урахуванням правильних цифр.

11 Визначити правила оцінки абсолютних і відносних похибок функцій

a) $y=x^a$; b) $y=a^x$, $a>0$; c) $y=e^x$.

12 Функція $y=\dfrac{x_1-2x_2}{x_3}$ обчислюється при значеннях $x_1=2{,}5\pm0{,}1$, $x_2=2{,}0\pm0{,}2$, $x_3=1{,}7\pm0{,}3$. Знайти значення $y,\Delta y,\delta y$. Записати результат з усіма правильними цифрами.

13 Коефіцієнти a,b,c обчислюються з відносною похибкою $\delta(a)=\delta(b)=\delta(c)=\delta$. Знайти максимальну похибку, з якою можуть обчислюватися корені рівнянь: a) $ax^2+c=0$; b) $ax^2+bx=0$.

14 Функція $y=\dfrac{2x_1+x_2}{x_3}$ обчислюється при значеннях $x_1\approx2.7$, $x_2\approx-3.1$, $x_3\approx1.8$. Визначити при яких значеннях $\delta x_1,\delta x_2,\delta x_3$ відповідь буде містити 3 правильні цифри.

Розділ 2
Числове розв'язання нелінійних рівнянь

У традиційних курсах математики найчастіше ставиться задача пошуку точного розв'язку таких рівнянь. Однак у прикладних задачах виникають такі рівняння, коли, як правило, неможливо його знайти.

Якщо обчислювальна техніка дозволяє впоратися з великою кількістю складних рівнянь, то абстрактні методи дозволяють для різноманітних класів рівнянь (таких, як алгебраїчні, диференціальні, інтегродиференціальні і т.д.) знайти загальні підходи до їх розв'язку.

Спробуємо дати уявлення про деякі результати, пов'язані з питаннями існування і побудови розв'язків нелінійних рівнянь. Визначимося з математичним апаратом, необхідним для обґрунтування чисельних методів розв'язання рівнянь.

2.1 Відображення множин

Бажання охопити якнайбільше рівнянь призводить до необхідності розглядати відображення (функції), за допомогою яких записуються рівняння досить загальної природи, і застосовувати теоретико-множинний підхід. Для цього залучимо поняття відображення однієї множини в іншу, причому із самого початку підкреслимо, що це поняття буде вважатися первинним, що не підлягає формально строгому визначенню (так само, як і поняття числа, точки, множини і т.д.). Інтуїтивний зміст слова "відображення" відбиває наявність відповідності між елементами двох множин (при бажанні можна обмежитися тільки розглядом числових множин).

Нехай: 1) задана (непуста) множина X; 2) задана непуста множина Y; 3) для кожного елемента множини X

зазначений один цілком визначений елемент множини Y. У такому випадку будемо говорити про відображення множини X у множину Y.

Інакше кажучи, поняття "відображення" містить у собі нероздільний опис двох множин X (область визначення відображення), Y (область значення) і опис правила (способу, закону), за яким для кожного елемента x з множини X задається певний елемент y з множини Y, у який елемент x відображається.

Закон (правило) відповідності, за яким для кожного елемента області визначення множини відображення X задається рівно один елемент області значень Y, позначимо буквою f. Тоді відображення множини X у множину Y можна записати так (це позначення прийняте в сучасній математиці):

$$f : X \to Y \text{ чи } X \xrightarrow{f} Y . \qquad (2.1)$$

Елемент $y \in Y$, у який при даному відображенні (2.1) переходить елемент $x \in X$, називається образом елемента x (чи значенням відображення на елементі x) і позначається символом $f(x)$. Часто, коли ясний вибір множин X і Y, для розглянутої функції використовується позначення f.

З'ясувавши зміст поняття "відображення", можна дати визначення функції: ***функцією називається однозначне відображення однієї множини чисел в іншу множину чисел.*** Таким чином, функція — той окремий випадок відображення, коли область визначення й область значень є числовими множинами.

2.2 Розв'язок рівнянь і нерухомі точки відображень

Одна з простих теорем існування розв'язку формулюється так:

Теорема 1 *Нехай f: $[a, b] \to R$ — неперервна на відрізку $[a, b]$ функція і $f(a)f(b) < 0$ (тобто на кінцях відрізків вона набуває значення різних знаків), тоді рівняння*

$$f(x)=0 \qquad (2.2)$$

має на відрізку $[a, b]$ хоча б один розв'язок.

Теореми існування, як правило, формулюють як теореми існування нерухомих точок відображень.

Нехай f: $X \to X$ — деяке відображення множини X в себе. ***Елемент x_0 з X називається нерухомою точкою відображення f, якщо має місце рівність $f(x_0) = x_0$.***

Якщо f: $[a, b] \to R$ — функція, що задовольняє умови теореми 1, то розглянемо на відрізку $[a, b]$ функцію $F(x) = \lambda f(x)+x$, де параметр λ виберемо так, щоб усі значення функції F належали відрізку $[a, b]$. Наприклад, можна покласти $\lambda=M/m$,

де $M=\max\limits_{x\in[a,b]} f(x)=f(x_2)$ і $m=\min\limits_{x\in[a,b]} f(x)=f(x_1)$,

x_1 і x_2 — точки з відрізка $[a, b]$, у яких досягаються мінімум і максимум функції f відповідно.

Безпосередньо з вигляду функції F випливає, що x_0 — розв'язок рівняння (2.2) тоді і тільки тоді, коли x_0 — нерухома точка для функції f. Таким чином, теорема 1 еквівалентна теоремі, що формулюється у вигляді принципу нерухомої точки.

Теорема 2 *Неперервне відображення F відрізка в себе має нерухому точку.*

Аналогічно задачу про можливість розв'язання систем рівнянь з кількома невідомими можна звести до питання існування нерухомої точки відображення багатомірного куба (чи кулі) у себе.

Крім проблеми існування нерухомої точки, виникає природне запитання про способи її наближеного

обчислення. У наступній теоремі одночасно вирішуються обидві проблеми.

Теорема 3 *Нехай φ: [a, b] → [a, b] — функція, що задовольняє умову: існує число q з інтервалу (0,1) таке, що*

$$|\varphi(x_1) - \varphi(x_2)| \le q|x_1 - x_2| \qquad (2.3)$$

для всіх чисел x_1, x_2 з [a, b]. Тоді функція φ має єдину нерухому точку $x^* \in [a, b]$, що є границею послідовності $x_0, x_1, x_2, \ldots$, де x_0 — довільна точка з [a,b], а інші члени послідовності визначаються за правилом:

$$x_n = \varphi(x_{n-1}),\ n > 1.$$

Викладений у теоремі ***метод побудови нерухомої точки називається методом простих ітерацій або методом послідовних наближень, а послідовність $\{x_n\}$ — ітераційною послідовністю***.

2.3 Теореми про стискаючі відображення

Розглянемо повний метричний простір X з відстанню ρ. ***Відображення $f: X \to X$ називається стискаючим, якщо існує таке число $q \in (0, 1)$, що***

$$\rho(f(x_1), f(x_2)) \le q\rho(x_1, x_2) \qquad (2.4)$$

для всіх елементів $x_1, x_2 \in X$.

Має місце наступне узагальнення теореми 3.

Теорема 4 *(про стискаючі відображення). Стискаюче відображення $f: X \to X$ має єдину нерухому точку $x^* \in X$, яку можна знайти як границю послідовності*

$$x_{n+1} = f(x_n), n = 0,1,2,\ldots, \qquad (2.5)$$

де x_0 — довільний елемент із X. Крім того виконується

$$\rho(x_m, x_n) \le \frac{q^m}{1-q}\rho(x_1, x_0), m = 1,2,\ldots \qquad (2.6)$$

Доведення. Покажемо, що послідовність $\{x_n\}$ фундаментальна. З (2.4) і (2.5) одержуємо оцінки

$$\rho(x_{k+1}, x_k) \le \rho(f(x_k), f(x_{k-1})) \le q\rho(x_k, x_{k-1}) \le q^k \rho(x_1, x_0).$$

Вважаючи для визначеності, що $n > m$, з нерівності трикутника одержуємо оцінки

$$\rho(x_n, x_m) \le \rho(x_m, x_{n-1}) + \rho(x_{n-1}, x_n) \le$$
$$\le \rho(x_n, x_{n-1}) + \rho(x_{n-1}, x_{n-2}) + \ldots + \rho(x_{m+1}, x_m) \le$$
$$\le q^n \rho(x_1, x_0) + q^{n-1}\rho(x_1, x_0) + \ldots + q^{m-1}\rho(x_1, x_0) \le$$
$$\le \frac{q^m - q^n}{1-q}\rho(x_1, x_0) \le \frac{q^m}{1-q}\rho(x_m, x_n) \to 0$$

при $m \to \infty$.

Таким чином, побудована послідовність $\{x_n\}$ фундаментальна. Оскільки метричний простір X повний, то вона має границю $x^* \in X$. Оскільки

$$\rho(x^*, x_m) \le \rho(x^*, x_n) + \rho(x_m, x_n),$$

то при $n \to \infty$ з наведених оцінок одержуємо доведену оцінку (2.6). Оскільки при $m \to \infty$

$$\rho(x^*, f(x^*)) \le \rho(x^*, x_{m+1}) + \rho(x_{m+1}, f(x^*)) =$$
$$= \rho(x^*, x_{m+1}) + \rho(f(x^*), f(x_{m+1})) \le$$
$$\le \rho(x^*, x_{m+1}) + \rho(x_m, x^*) \to 0,$$

то $f(x) = x^*$, тобто x^* — нерухома точка відображення f.

Доведемо, що вона — єдина. Нехай x^{**} — інша нерухома точка для f. Тоді з оцінки

$$\rho(x^*, x^{**}) = \rho(f(x^*), f(x^{**})) \le q\rho(x^*, x^{**})$$

випливає, що $\rho(x^*, x^{**}) = 0$ і тому $x^* = x^{**}$.

Теорема доведена.

Нерівність (2.6) дозволяє визначити, скільки потрібно знайти послідовних наближень, щоб знайти нерухому точку x^* відображення f із заданою точністю $\varepsilon > 0$. Наприклад, нерівність $\rho(x^*, x_m) < \varepsilon$ буде виконана, якщо

$$m > m(\varepsilon) = \frac{1}{\ln q}\ln\frac{\varepsilon(1-q)}{\rho(x_1, x_0)}.$$

Теорема 4 широко застосовується для пошуку розв'язків систем алгебраїчних, диференціальних та інтегральних рівнянь, а також покладена в основу різних методів, які використовуються в обчислювальній математиці.

2.4 Критерій існування нерухомих точок

Розглянемо узагальнення теорем 1 і 2, що пов'язані з питаннями існування нерухомих точок відображень. Формулювання відповідних результатів використовує наступні важливі поняття.

Множина K з метричного простору X, на якому введена відстань ρ, називається компактною, якщо з будь-якої послідовності $\{x_n\}$ елементів цієї множини можна виділити підпослідовність, що збігається, границя якої також належить K.

Множина B з лінійного простору X (тобто множина, на якій визначені операції додавання і множення на числа з R) називається опуклою, якщо разом з будь-якими двома точками $x_1, x_2 \in B$ множині B належать усі точки $ax_1+(1-a)x_2$, $a \in [0,1]$ (множина таких точок називається відрізком, що з'єднує точки $x_1, x_2 \in B$).

Теорема 5. *Будь-яке неперервне відображення $f: K \to K$ опуклої компактної множини K з лінійного простору R^n має нерухому точку.*

Ця теорема доведена голландським математиком Л.Э.Я. Брауером. Оскільки будь-який відрізок $[a, b]$ з R є опуклою компактною множиною, то теорема 2 випливає з теореми Брауера. Розглянемо деякі наслідки теореми 5.

Теорема Борсука-Улама для кола. *Нехай $f: C \to R$ — функція на колі C. Тоді існує пара точок-антиподів x, x^* така, що $f(x) = f(x^*)$.*

Наслідком цієї теореми є той факт, що на будь-якому колі Земної кулі (наприклад, на екваторі) знайдеться пара

антиподів, у яких температура повітря однакова.

Теорема *(про млинці). Якщо A і B — обмежені фігури на площині, то існує пряма, що поділяє кожну з цих фігур на дві рівновеликі за площею частини.*

Тут передбачається, що кожна фігура має площу (що може бути тоді, коли фігура має складну форму). Образно кажучи, теорема стверджує, що два млинці, що лежать на тарілці, можна розрізати точно навпіл одним змахом ножа.

Теорема Брауера має різноманітні узагальнення, що знаходять застосування в питаннях існування розв'язків диференціальних та інтегральних рівнянь, у математичній економіці (теорія економічної рівноваги) і теорії ігор.

2.5 Розв'язання нелінійних рівнянь у комплексній площині

Нехай задана функція $f(x)$ дійсної змінної. Потрібно знайти нулі функції $f(x)$ або, що те ж саме, корені рівняння

$$f(x) = 0 \qquad (2.7)$$

Уже на прикладі алгебраїчного многочлена відомо, що нулі $f(x)$ можуть бути як дійсними, так і комплексними. Тому більш точною є постановка задачі у визначенні коренів рівняння (2.7), розміщених у заданій області комплексної площини. Можна розглядати задачу визначення дійсних коренів, розміщених на заданому відрізку. Іноді, нехтуючи точністю формулювань, будемо говорити, що потрібно розв'язати рівняння (2.7).

У загальному випадку задача визначення коренів рівняння (2.7), як правило, розв'язується в два етапи. На першому етапі вивчається розміщення корені і проводиться їх відділення, тобто виділяються області в комплексній площині, що містять тільки один корінь. Крім того, вивчається питання про кратність коренів. Знаходяться деякі початкові наближення для коренів

рівняння (2.7). На другому етапі, використовуючи задане початкове наближення, будується ітераційний процес, що дозволяє уточнити значення невідомого кореня.

Не існує якихось загальних регулярних прийомів розв'язання задачі про розміщення коренів довільної функції $f(x)$. Найбільш повно вивчене питання про розміщення коренів алгебраїчних многочленів

$$f(x)=a_0+a_1x+a_2x^2+\ldots+a_mx^m. \qquad (2.8)$$

Наприклад відомо, що якщо для многочлена (2.8) з дійсними коефіцієнтами виконані нерівності $f(c)>0, f'(c)>0, \ldots, f^{(m)}(c)>0$, то додатні корені $f(x)$ не перевершують числа c. Дійсно, з формули Тейлора

$$f(x)=f(c)+(x-c)f'(x)+\frac{(x-c)^2}{2!}f''(c)+\ldots+\frac{(x-c)^m}{m!}f^{(m)}(c)+\ldots$$

одержуємо, що $f(x)>0$ при $x>c$.

Чисельні методи розв'язання нелінійних рівнянь є, як правило, ітераційними методами, що припускають завдання досить близьких до шуканого розв'язку початкових даних. Перш ніж переходити до викладу конкретних ітераційних методів, відзначимо два прості прийоми відділення дійсних коренів рівняння (2.7). Припустимо, що $f(x)$ визначена і неперервна на $[a,b]$.

Перший прийом полягає в тому, що обчислюється таблиця значень функції $f(x)$ у заданих точках $x_k\in[a, b]$, $k=0, 1,\ldots, n$. Якщо виявиться, що при якомусь k числа $f(x_k)$, $f(x_{k+1})$ мають різні знаки, то це буде означати, що на інтервалі (x_k,x_{k+1}) рівняння (2.7) має принаймні один дійсний корінь (точніше, має непарне число коренів на (x_k, x_{k+1})). Потім можна розбити інтервал (x_k,x_{k+1}) на більш дрібні інтервали і за допомогою аналогічної процедури уточнити розміщення кореня.

Більш регулярним способом відділення дійсних коренів є *метод бісекції (розподілу навпіл)*. Припустимо,

що на (a, b) розміщений лише один корінь x рівняння (2.7). Тоді $f(a)$ і $f(b)$ мають різні знаки. Нехай для визначеності $f(a)>0$, $f(b)<0$. Покладемо $x_0=0,5\cdot(a+b)$ і обчислимо $f(x_0)$. Якщо $f(x_0)<0$, то шуканий корінь знаходиться на інтервалі (a,x_0), якщо ж $f(x_0)>0$, то на (x_0, b). Далі, із двох інтервалів (a,x_0) і (x_0,b) вибираємо той, на кінцях якого функція $f(x)$ має різні знаки, знаходимо точку x_1-*середину* обраного інтервалу, обчислюємо $f(x_1)$ і повторюємо зазначений процес. У результаті одержуємо послідовність інтервалів, що містять шуканий корінь x_1, причому довжина кожного наступного інтервалу вдвічі менша за попередній. Процес закінчується, коли довжина знову отриманого інтервалу стане менше заданого числа $\varepsilon>0$, і за корінь x приблизно береться середина цього інтервалу.

Помітимо, що якщо на (a, b) міститься кілька коренів, то зазначений процес зійдеться до одного з коренів, але заздалегідь невідомо, до якого саме. Можна використовувати прийом виділення коренів: якщо корінь $x=x^*$ кратності m знайдений, то розглядається функція $g(x)=f(x)/(x-x^*)^m$ і для неї повторюється процес визначення кореня.

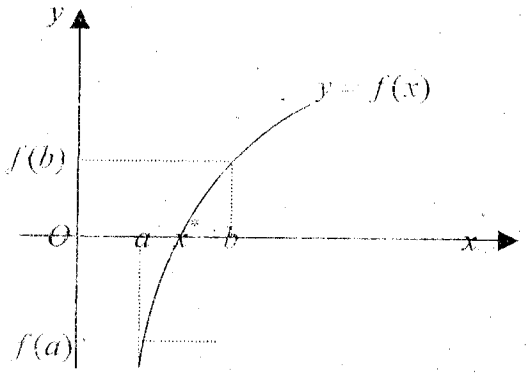


Рис. - 2.1

Обмеження кореня функції, якщо вона визначення на необмеженому інтервалі здійснюється так.

1 Для початкового наближення x_0, знайти $f_0=f(x_0)$, задати інтервал пошуку D і його інкремент $d>1$.

2 Обчислити $a=x_0-D$, $b=x_0+D$; $f_a=f(a)$, $f_b=f(b)$.

3 Збільшити інтервал пошуку: $D=D\times d$.

4a Якщо знаки f_a і f_0 різні, то вважати корінь обмеженим на $[a,x_0]$ -> вихід.

4b Якщо знаки f_b і f_0 відрізняються, то вважати корінь обмеженим на $[x_0,b]$ -> вихід.

5 Перевіряється, яке з f_a і f_b найменше. Якщо вони однакові, то переходимо до 6a (двосторонній пошук), якщо f_b – робимо пошук вправо 6b, інакше - пошук уліво 6c.

6a Знаходимо $a=a-D$, $b=b+D$, $f_a=f(a)$, $f_b=f(b)$, йдемо до пункту 3.

6b Присвоюємо послідовно $a=x_0$, $x_0=b$, $f_a=f_0$, $f_0=f_b$; знаходимо $b=b+D$, $f_b=f(b)$, йдемо до пункту 3.

6c. Аналогічно 6b, тільки напрямок пошуку - вліво.

2.6 Метод простих ітерацій

Замінимо рівняння $f(x)=0$ еквівалентним йому рівнянням $x=\varphi(x)$.

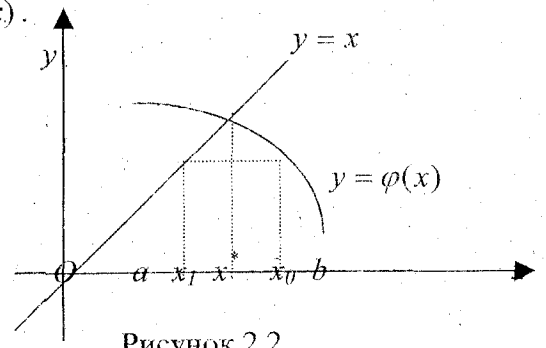


Рисунок 2.2

Виберемо деяке наближення $x_0 \in [a,b]$ кореня і підставимо його в праву частину. Одержимо $x_1=\varphi(x_0)$. Далі обчислюємо за формулами: $x_n=\varphi(x_{n-1})$, $n=2,3,\ldots$. Отримуємо послідовність наближень $\{x_n\}$ до кореня, що у випадку її збіжності до кореня x^* може дати наближене його значення із заданою точністю ε.

Теорема 6 *Нехай функція $\varphi(x)$ визначена і диференційована на відрізку $[a,b]$, причому всі значення $\varphi(x)\in[a,b]$. Тоді якщо існує правильний дріб q, такий, що*

$$|\varphi'(x)|\le q<1 \qquad (2.9)$$

при $a<x<b$, то: процес ітерації

$$x_n=\varphi(x_{n-1}) \quad (n=1,2,\ldots) \qquad (2.10)$$

1) збігається незалежно від початкового значення $x_0\in[a,b]$;

2) граничне значення $\xi=\lim\limits_{n\to\infty}x_n$ є єдиним коренем рівняння

$$x=\varphi(x) \qquad (2.11)$$

на відрізку $[a,b]$.

Доведення. Розглянемо два послідовних наближення $x_n=\varphi(x_{n-1})$ і $x_{n+1}=\varphi(x_n)$ (які внаслідок умов теореми існують). Звідси $x_{n+1}-x_n=\varphi(x_n)-\varphi(x_{n-1})$.

Застосовуючи теорему Лагранжа, будемо мати:

$$x_{n+1}-x_n=(x_n-x_{n-1})\,\varphi'(\bar{x}_n),\ \text{де}\ \bar{x}_n\in[x_{n-1},x_n].$$

Отже, на підставі умови (2.9) одержимо

$$|x_{n+1}-x_n|\le q|x_n-x_{n-1}|. \qquad (2.12)$$

Звідси, надаючи значення $n=1,2,3,\ldots$, отримаємо:

$$|x_2-x_1|\le q|x_1-x_0|;$$

$$|x_3-x_2|\le q|x_2-x_1|\le q^2|x_1-x_0|;$$

.....

$$|x_{n+1} - x_n| \le q^n |x_1 - x_0|. \qquad (2.13)$$

Розглянемо ряд:

$$x_0 + (x_1 - x_0) + (x_2 - x_1) + \ldots + (x_n - x_{n-1}) + \ldots, \qquad (2.14)$$

для якого наші послідовні наближення x_n є $(n+1)$-ми частковими сумами, тобто $x_n = S_{n+1}$. За нерівністю (2.13) члени ряду (2.14) за абсолютною величиною менші відповідних членів геометричної прогресії зі знаменником $q<1$, тому ряд (2.14) збігається, до того ж абсолютно. Отже, існує $\lim\limits_{n\to\infty} S_{n+1} = \lim\limits_{n\to\infty} x_n = \xi$, причому, вочевидь, $\xi \in [a,b]$. Переходячи до границі в рівності (2.10), зважаючи на неперервність функції $\varphi(x)$ одержуємо

$$\xi = \varphi(\xi). \qquad (2.15)$$

У такий спосіб ξ є корінь рівняння (2.11). Іншого кореня на відрізку $[a,b]$ рівняння (2.11) не має. Дійсно, якщо

$$\bar{\xi} = \varphi(\bar{\xi}), \qquad (2.16)$$

то з рівностей (2.15) і (2.16) одержимо

$$\bar{\xi} - \xi = \varphi(\bar{\xi}) - \varphi(\xi)$$

і отже,

$$(\bar{\xi} - \xi)[1 - \varphi'(c)] = 0, \qquad (2.17)$$

де $c \in [\xi, \bar{\xi}]$. Оскільки вираз у квадратній дужці в рівності (2.17) не дорівнює нулю, то $\bar{\xi} = \xi$, тобто корінь ξ - єдиний.

Зауваження 1 *Теорема залишається правильною, якщо функція $\varphi(x)$ визначена і диференційована на інтервалі $-\infty < x < \infty$, причому при $x \in (-\infty;+\infty)$ виконана нерівність (2.9).*

Зауваження 2 *В умовах теореми метод ітерації збігається при будь-якому виборі початкового значення $x_0 \in [a,b]$. Завдяки цьому він є самовиправним, тобто окрема помилка в обчисленнях, що не виводить за межі відрізка*

$[a,b,]$ не вплине на кінцевий результат, тому що помилкове значення можна розглядати як нове початкове значення x_0. Можливо, зросте лише обсяг роботи . Властивість самовиправлення робить метод ітерації одним із найнадійніших методів обчислень.

Оцінка наближення. З формули (2.13) маємо:

$$|x_{n+p} - x_n| \le |x_{n+p} - x_{n+p-1}| + |x_{n+p-1} - x_{n+p-2}| + \ldots + |x_{n+1} - x_n| \le$$

$$\le q^{n+p-1}|x_1 - x_0| + q^{n+p-2}|x_1 - x_0| + \ldots + |x_{n+1} - x_n| + + q^n|x_1 - x_0| =$$

$$= q^n|x_1 - x_0|(1 + q + q^2 + \ldots + q^{p-1}).$$

Застосувавши формулу суми геометричної прогресії, одержимо:

$$|x_{n+p} - x_n| \le q^n|x_1 - x_0|\frac{1-q^p}{1-q} < \frac{q^n}{1-q}|x_1 - x_0|.$$

Спрямовуючи число p до нескінченності і з огляду на те, що $\lim\limits_{p\to\infty} x_{n+p} = \xi$, знаходимо остаточно:

$$|\xi - x_n| \le \frac{q^n}{1-q}|x_1 - x_0|. \qquad (2.18)$$

Звідси ясно, що збіжність процесу ітерації буде тим швидшою, чим менше число q.

Для оцінки наближення можна дати іншу формулу, корисну в деяких випадках. Представимо $f(x) = x - \varphi(x)$.

Очевидно, що $f'(x) = 1 - \varphi'(x) \ge 1 - q$. Звідси, з огляду на те, що $f(\xi) = 0$, одержимо:

$$|x_n - \varphi(x_n)| = |f(x_n) - f(\xi)| = |x_n - \xi|f'(\overline{x_n}) \ge (1-q)|x_n - \xi|,$$

де $\overline{x_n} \in (x_n, \xi)$, і, отже,

$$|x_n - \xi| \le \frac{|x_n - \varphi(x_n)|}{1-q}, \qquad (2.19)$$

тобто

$$|\xi - x_n| \le \frac{|x_{n+1} - x_n|}{1-q}. \qquad (2.20)$$

Використовуючи формулу (2.12), маємо також

$$|\xi - x_n| \le \frac{q}{1-q}|x_n - x_{n-1}| . \qquad (2.21)$$

Звідси, зокрема, випливає, що якщо $q \le \frac{1}{2}$, то $|\xi - x_n| \le |x_n - x_{n-1}|$. В цьому випадку з нерівності $|x_n - x_{n-1}| < \varepsilon$ випливає нерівність $|\xi - x_n| < \varepsilon$.

Зауваження. *Існує поширена думка, що якщо при застосуванні методу ітерації два послідовні наближення x_{n-1} і x_n збігаються між собою із заданою точністю ε (наприклад, для цих наближень установилися т перших десяткових знаків), то з тією самою точністю справедлива рівність $\xi \approx x_n$ (тобто, зокрема, у наведеному прикладі т знаків наближеного числа x_n є правильними!). У загальному випадку це твердження помилкове. Більше того, легко показати, що якщо $\varphi'(x)$ близька до 1, то величина $|\xi - x_n|$ може бути великою, хоча $|x_n - x_{n-1}|$ дуже мала.*

Формула (2.20) дає можливість оцінити похибку наближеного значення x_n за різницею двох послідовних наближень x_{n-1} і x_n.

Процес ітерації варто продовжувати доти, поки для двох послідовних наближень x_{n-1} і x_n не буде забезпечене виконання нерівності

$$|x_n - x_{n-1}| \le \frac{1-q}{q}\varepsilon ,$$

де ε - задана гранична абсолютна похибка кореня ξ і $|\varphi'(x)| \le q$. Тоді за формулою (2.21) будемо мати нерівність $|\xi - x_n| \le \varepsilon$, тобто $\xi = x_n \pm \varepsilon$

Зауважимо, що якщо $x_n = \varphi(x_{n-1})$ і $\xi = \varphi(\xi)$, то $|\xi - x_n| = |\varphi(\xi) - \varphi(x_{n-1})| = |\xi - x_{n-1}|\,|\varphi'(\bar{x}_{n-1})| \le q|\xi - x_{n-1}|$, $(\bar{x}_{n-1} \in (x_{n-1}, \xi))$, тобто $|\xi - x_n| \le |\xi - x_{n-1}|$.

Таким чином, при ітераційному процесі, що збігається, похибка $|\xi - x_n|$ прямує до нуля монотонно, тобто кожне наступне значення x_n є більш точним, ніж попереднє значення x_{n-1}. Як правило, при всіх цих висновках ігноруються похибки округлень, тобто передбачається, що послідовні наближення знаходяться точно.

На практиці здебільшого буває так, що грубим прийомом встановлюється існування кореня рівняння (2.7) і методом ітерації потрібно одержати досить точне наближене значення кореня, причому нерівність (2.9) виконується лише в деякому околі *(a, b)* цього кореня. Тут при невдалому виборі початкового значення x_0 послідовні наближення $x_n = \varphi(x_{n-1})$ $(n = 1,2,\ldots)$ можуть залишити інтервал *(a, b)* чи навіть втратити сенс.

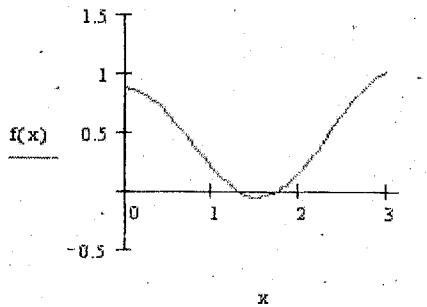
Приклад. Розв'язати рівняння $f(x) = 0$ на заданому відрізку [a,b]=[0, π], де $f(x) = (\cos x)^2 - \frac{1}{12}\cos x - \frac{1}{24} = 0$,

Аналітичне розв'язання задачі. Розкладемо функцію $f(x) = \left(\cos x - \frac{1}{4}\right)\cdot\left(\cos x + \frac{1}{6}\right)$. Точні значення коренів $x_1 = \arccos\left(\frac{1}{4}\right) = 1.31811607652818$,

$x_2 = \pi - \arccos\left(\frac{1}{6}\right) = 1.738244406014586$.

Чисельне розв'язання задачі. Локалізація кореня для чисельного розв'язання задачі

$$f(x) := (\cos(x))^2 - \frac{\cos(x)}{12} - \frac{1}{24}$$



[1,1.5], [1.5,2]

Метод бісекції, зреалізований у пакеті Mathcad, дає

Перший корінь

$$\text{bisec}\left(f,1,1.5,10^{-10}\right) = \begin{bmatrix} 1.318116071692202 \\ 32 \end{bmatrix}.$$

Обравши $x0 := 1$ - задання початкового наближення, користуємось убудованою функцією пакета MATHCAD

$root(f(x0),x0) = 1.317959944516193$.

Значення кореня відрізняється від знайденого за допомогою функції bisec , тому що за замовчуванням величина похибки при роботі вбудованих функцій дорівнює 0.001.

Перевизначимо параметр для задання похибки

$TOL := 10^{-10}$

$root(f(x0),x0) = 1.318116071652817$

Значення кореня із заданою точністю 1.3181160717.

Другий корінь

$$\text{bisec}\left(f,1.5,2,10^{-10}\right) = \begin{bmatrix} 1.738244406005833 \\ 32 \end{bmatrix}.$$

Значення кореня із заданою точністю 1.7382444060, число ітерацій 32; $x0 := 1.8$ - задання початкового наближення; $root(f(x0),x0) = 1.738244406014586$.

Значення кореня у межах заданої точності збігаються.

```
bisec(f, a, b, ε) :=
    an ← a
    bn ← b
    k ← 0
    while (bn − an) > 2·ε
        xn ← (an + bn)/2
        fa ← f(an)
        fb ← f(bn)
        fxn ← f(xn)
        bn ← xn if fa·fxn < 0
        an ← xn otherwise
        k ← k + 1
    xn ← (an + bn)/2
    res ← [xn; k]
    res
```

2.7 Метод Ньютона

Метод Ньютона (метод дотичних) для наближеного розв'язку рівняння $f(x) = 0$ полягає в побудові ітераційної послідовності

$$x_{n+1} = x_n - f(x_n)/f'(x_n), \qquad (2.22)$$

що збігається до кореня рівняння, на відрізку $[a,b]$ локалізації кореня.

Теорема 7 *Якщо $f(a)\,f(b)<0$, причому $f'(x)$ і $f''(x)$ не дорівнюють нулю і зберігають певні знаки при $a \le x \le b$, то, виходячи з початкового наближення $x_0 \in [a,b]$, що задовольняє нерівність*

$$f(x_0)\,f''(x_0) > 0, \qquad (2.23)$$

можна обчислити методом Ньютона єдиний корінь ξ рівняння $f(x) = 0$ з будь-яким ступенем точності.

Доведення. Нехай, наприклад, $f(a)<0$, $f(b)>0$, $f'(x)>0$, $f''(x)>0$ при $a \le x \le b$ (інші випадки розглядаються аналогічно). Відповідно до нерівності (2.23) маємо $f(x_0)>0$. (наприклад, можна взяти $x_0=b$). Методом математичної індукції доведемо, що всі наближення $x_n>\xi$ $(n=0,1,2\ldots)$ і, отже, $f(x_n)>0$. Справді, насамперед, $x_0>\xi$. Нехай тепер $x_n>\xi$. Покладемо $\xi = x_n + (\xi - x_n)$.

Застосовуючи формулу Тейлора, одержимо

$$0 = f(\xi) = f(x_n) + f'(x_n)(\xi - x_n) + \frac{1}{2} f''(c_n)(\xi - x_n)^2, \qquad (2.24)$$

де $\xi<c_n<x_n$. Оскільки $f''(x)>0$, маємо

$$f(x_n) + f'(x_n)(\xi - x_n) < 0 \quad \text{і отже,} \quad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} > \xi,$$

що і потрібно було довести.

З огляду на знаки $f(x_n)$ та $f'(x_n)$, маємо $x_{n+1}<x_n$ $(n=0,1,\ldots)$, тобто послідовні наближення $x_0, x_1, \ldots, x_n, \ldots$ утворять обмежену монотонно спадну послідовність. Отже, існує $\bar{\xi} = \lim\limits_{n\to\infty} x_n$.

Переходячи до границі в рівності (2.22), будемо мати

$$\bar{\xi} = \bar{\xi} - \frac{f(\bar{\xi})}{f'(\bar{\xi})},$$

тобто $f(\bar{\xi})= 0$. Звідси $\bar{\xi} = \xi$, що і потрібно було довести.

Тому, застосовуючи метод Ньютона, варто керуватися таким правилом: *за вихідну точку x_0*

вибирається той кінець інтервалу (a,b), якому відповідає ордината того самого знака, що і знак $f''(x)$.

Зауваження. *З формули (2.22) бачимо, що чим більше числове значення $f'(x)$ в околі кореня, тим меншою є поправка, яку треба додати до попереднього наближення, щоб отримати наступне. З цієї причини метод Ньютона особливо зручний тоді, коли в околі кореня графік функції має велику крутизну. Якщо ж $f'(x)$ біля кореня – мала, то застосовувати даний метод не рекомендується.*

Для оцінки похибки n-го наближення x_n можна скористатися формулою

$$|\xi - x_n| \le \frac{|f(x_n)|}{m_1}, \qquad (2.25)$$

де m_1 найменше значення $|f'(x)|$ на відрізку $[a,b]$.

Виведемо ще одну формулу для оцінки точності наближення x_n.

Застосовуючи формулу Тейлора, маємо:

$$f(x_n) = f[x_{n-1} + (x_n - x_{n-1})] =$$

$$= f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) + \frac{1}{2} f''(\xi_{n-1})(x_n - x_{n-1})^2, \qquad (2.26)$$

де $\xi_{n-1} \in (x_{n-1}, x_n)$. Оскільки з визначення наближення x_n маємо

$f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1})^2 = 0$, то з (2.26) знаходимо: $|f(x_n)| \le \frac{1}{2} M_2 (x_n - x_{n-1})^2$, де M_2 – найбільше значення $|f''(x)|$ на відрізку $[a,b]$. Отже, на підставі формули (26) остаточно одержуємо

$$|\xi - x_n| \le \frac{M_2}{2m_1}(x_n - x_{n-1})^2. \qquad (2.27)$$

Якщо процес збігається, то $x_n - x_{n-1} \to 0$ при $n\to\infty$. Тому при $n \ge N$ маємо $|\xi - x_n| \le |x_n - x_{n-1}|$, тобто «усталені»

початкові десяткові знаки наближень x_{n-1} і x_n, починаючи з деякого наближення, є правильними.

Зауважимо, що в загальному випадку збіг з точністю до ε двох послідовних наближень x_{n-1} і x_n зовсім не гарантує, що з тією самою точністю збігаються значення x_n і точний корінь ξ.

Проаналізуємо абсолютні похибки двох послідовних наближень x_n і x_{n+1}. З формули (2.24) одержуємо

$$\xi = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{1}{2}\frac{f''(c_n)}{f(x_n)}(\xi - x_n)^2,$$

де $c_n \in (x_n, \xi)$. Звідси, з огляду на формулу (2.22), будемо мати

$$\xi - x_{n+1} = -\frac{1}{2}\cdot\frac{f''(c_n)}{f'(x_n)}(\xi - x_n)^2$$

і, отже,

$$|\xi - x_{n-1}| \le \frac{M_2}{2m_1}(\xi - x_n)^2. \qquad (2.28)$$

Формула (2.28) забезпечує швидку збіжність процесу Ньютона, якщо початкове наближення x_0 таке, що $\frac{M_2}{2m_1}|\xi - x_0| \le q < 1$. Зокрема, якщо $\mu = \frac{M_2}{2m_1} \le 1$ і $|\xi - x_n| < 10^{-2m}$, тобто наближення x_n мало m правильних десяткових знаків після коми, то наступне наближення x_{n+1} буде мати не менше $2m$ правильних знаків; іншими словами, якщо $\mu \le 1$, то за допомогою методу Ньютона число правильних знаків після коми шуканого кореня ξ подвоюється на кожному кроці.

Приклад. Знайти корінь рівняння $e^x - 10x = 0$ з точністю $\varepsilon = 10^{-3}$.

1 Це рівняння має один корінь на $[0,1]$. (f(0)f(1))<0)

Знайдемо похідні

$$f(x) = e^x - 10x;\ f'(x) = e^x - 10;\ f''(x) = e^x.$$

2 Вибираємо початкове наближення кореня $x_0 \in [0,1]$ так, щоб $f(x_0) * f''(x_0) \rangle 0$. Обираємо $x_0 = 0$, тому що $f(0)f''(0) > 0$.

3 Будуємо ітераційну послідовність

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 0 - \frac{e^0}{e^0 - 10} = \frac{-1}{-9} = 0.1111,$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 0.1111 - \frac{e^{0.1111} - 1.111}{e^{0.1111} - 10} = 0.1111 + \frac{0.00640}{8.8824} =$$

$$= 0.1111 + 0.00072 = 0.11183.$$

4 Обчислення припиняємо, тому що $|x_2 - x_1| < \varepsilon$, і за наближене значення кореня з точністю 10^{-3} беремо $\xi = x_2 = 0.11183$.

Приклад реалізації чисельного алгоритму розв'язування нелінійних рівнянь на псевдокоді

```
//Метод Ньютона. Вважаємо, що умова збіжності
методу перевірена
f(x):
  //повертає значення функції для даного x
end
f1(x):
//повертає значення першої похідної функції для данного x
end
f2(x):
//повертає значення другої похідної функції для даного x
end
//a,b – границі відрізка, eps – точність розв'язку
```

```
Solve_Nonlinear(a,b,eps):
1   if f1(a)<f1(b) then
2     min:=abs(f1(a))
3   else
4     min:=abs(f1(b))
5   fi
6   if f2(a)>f2(b) then
7     max:=abs(f2(a))
8   else
9     max:=abs(f2(b))
10  fi
11  fault:=sqrt(2*min*eps)
12  if f(b)*f2(b)>0 then
13    x:=b
14  else
15    x:=a
16  fi
17  repeat
18    n++
19    if n>1 then
20      x:=xn
21    fi
22    xn:=x-f(x)/f1(x)
23  until abs(xn-x)<=fault
```



```
24  return x
end
```

2.8 Обумовленість задачі визначення кореня

Нехай $\bar{x}$ – корінь, що підлягає визначенню. Будемо вважати, що вхідними даними для задачі обчислення кореня є значення функції $f(x)$. Оскільки $f(x)$ обчислюється наближено, то позначимо функцію, отриману в дійсності через $f^*(x)$. Припустимо, що в малому околі кореня виконується нерівність: $|f(x) - f^*(x)| < \Delta(f^*)$. Для близьких до $\bar{x}$ значень x справедлива рівність $f(x) \approx f(\bar{x}) + f'(\bar{x})(x - \bar{x})$, отже, $\Delta(x^*) \approx \frac{1}{|f'(\bar{x})|}\Delta(f^*)$. Це означає, що число обумовленості задачі знаходження кореня дорівнює $\nu_\Delta = \frac{1}{|f'(\bar{x})|}$. З останньої формули можна зробити висновок, що чим менше значення похідної функції в точці кореня, тим задача гірше обумовлена. Зокрема, задача знаходження кратного кореня має число обумовленості – нескінченність.

Інтервал невизначеності кореня. Якщо функція $f(x)$ неперервна, то знайдеться такий малий окіл $(\bar{x} - \varepsilon, \bar{x} + \varepsilon)$ кореня $\bar{x}$, що має радіус ε, у якому виконується нерівність $|f(x)| < \bar{\Delta}$. Це означає, що для $x \in (\bar{x} - \varepsilon, \bar{x} + \varepsilon)$ знак обчисленого значення $f^*(x)$ взагалі не зобов'язаний збігатися зі знаком $f(x)$, і, отже, стає неможливим визначити, яке саме значення x з інтервалу $(\bar{x} - \varepsilon, \bar{x} + \varepsilon)$ обертає функцію f в нуль. Цей інтервал

називається *інтервалом невизначеності* кореня. Очевидно, що радіус інтервалу невизначеності для простого кореня дорівнює $\varepsilon \approx \Delta(x^*) \approx \frac{1}{f'(\bar{x})}\Delta(f^*)$. Аналогічно можна показати, що для кратного кореня $\varepsilon \approx \left|\frac{m!}{f^{(m)}(\bar{x})}\right|^{\frac{1}{m}} \Delta^{\frac{1}{m}}$. Це означає, що для простого кореня радіус інтервалу невизначеності прямо пропорційний похибці обчислення функції $\Delta(f^*)$, а для кратного кореня $\left(\Delta(f^*)\right)^{\frac{1}{m}}$.

Приклад. Теоретична оцінка радіуса інтервалу невизначеності кореня.

Нехай $f(x) = x^3 - 3x + 1$. Корінь рівняння простий і дорівнює $\bar{x} = -0.34729635\,533861$. Тоді $f'(x) = 3x^2 - 3$ і $f'(\bar{x}) = -2.64$. Якщо $\Delta(f^*) \cong 10^{-6}$, то $\varepsilon \approx 4 \cdot 10^{-7}$. Це означає , що знайти корінь із точністю меншою, ніж радіус інтервалу невизначеності, не вдасться.

2.9 Метод Ньютона для знаходження кратного кореня

Метод Ньютона на випадок кратного кореня має лише лінійну швидкість збіжності. Щоб зберегти квадратичну збіжність, його модифікують у такий спосіб:

$x^{(n+1)} = x^{(n)} - m\frac{f\left(x^{(n)}\right)}{f'\left(x^{(n)}\right)}$, де m – кратність кореня. Як правило, значення m невідоме. Використовуючи метод Ньютона, можна знайти кратність кореня. Для цього будемо задавати значення $m = 1,2,3$ і обчислювати

значення кореня із заданою точністю, одночасно підраховуючи кількість ітерацій для кожного значення m. При деякому значенні m число ітерацій буде мінімальним. Це значення m і є кратністю кореня.

Питання та завдання до розділу 2

1 Постановка задачі розв'язання нелінійних рівнянь. Основні етапи розв'язання задачі.

2 Ітераційне уточнення кореня: порядок збіжності методу, апріорні й апостеріорні оцінки похибки.

3 Метод бісекції: опис методу, швидкість збіжності, критерій закінчення.

4 Метод простої ітерації розв'язку нелінійного рівняння: опис методу, умова й швидкість збіжності, критерій закінчення, геометрична ілюстрація, приведення до вигляду, зручного для ітерацій.

5 Метод Ньютона розв'язку нелінійного рівняння: опис методу, теорема про збіжність, критерій закінчення, геометрична ілюстрація.

6 Недоліки методу Ньютона. Модифікації методу Ньютона. Модифікація методу Ньютона для пошуку кратних коренів.

7 Інтервал невизначеності кореня.

8 Визначити кількість коренів рівняння й для кожного кореня знайти відрізки локалізації:

а) $\sin x - (x-\pi)^3 = 0$, b) $\sin x - x^3 = 0$.

9 Знайти дійсний корінь рівняння $x^3 + 3x - 5 = 0$ методом бісекції з точністю $\varepsilon = 10^{-4}$.

10 Визначити порядок p і знаменник q швидкості збіжності методу бісекції.

11 Виписати ітераційну формулу і вказати початкове наближення для розв'язку рівняння

$$\sqrt{x+1}-\frac{1}{x-1}=0$$

12 Рівняння $\cos x - x^2 + 2x - 1 = 0$ має 2 корені: $x_1 = 0$, $x_2 \approx 1.5$. Для уточнення кореня застосовується метод простої ітерації $x_{n+1} = \frac{1}{2}(1 + x_n^2 - \cos x_n)$. До якого кореня зійдеться процес? Запропонувати ітераційний процес для уточнення другого кореня.

13 Розв'язується рівняння $x^3 + x - 1000 = 0$. Визначити, який з ітераційних процесів збігається до кореня $x \approx 10$:

$$x_{n+1} = \sqrt[3]{1000 - x_n},$$

$$x_{n+1} = 1000 - x_n^3,$$

$$x_{n+1} = x_n - 0.0002(x_n^3 + x_n - 10^3).$$

14 Нехай рівняння $f(x)=0$ має на відрізку $[a,b]$ єдиний корінь x і для його обчислення використається метод простої ітерації $x_{n+1} = \varphi(x_n)$. Показати, що якщо $\varphi(x)$ - неперервна функція на $[a,b]$ і $|\varphi'(x)| < 1$ на ньому, то для будь-якого початкового наближення з відрізка локалізації ітераційна послідовність збігається до кореня.

Розділ 3
Розв'язування систем лінійних алгебраїчних рівнянь (СЛАР)

Розглянемо систему вигляду

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ a_{m1}x_1 + a_{n2}x_2 + \ldots + a_{mn}x_n = b_n. \end{cases} \quad (3.1)$$

Її матричний вигляд

$$AX=B. \quad (3.2)$$

Тут $A - \{[a_{ij}],(i,j=\overline{1,n})\}$-матриця системи,

$$B=\begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_n \end{bmatrix}, \qquad X=\begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{bmatrix}$$ - вектори-стовпці.

Відомо, що система (3.1) має єдиний розв'язок, якщо її матриця невироджена (тобто визначник матриці A відмінний від нуля). У випадку виродженості матриці система може мати безліч розв'язків (якщо ранг матриці A і ранг розширеної матриці, отриманої додаванням до A стовпця вільних членів, однакові) або ж не мати розв'язків узагалі (якщо ранги матриці A і розширеної матриці не збігаються).

Методи чисельного розв'язання СЛАР поділяються на точні і наближені. Метод вважають точним, якщо, нехтуючи похибками округлення, він дає точний результат після виконання певної кількості обчислювальних операцій. Математичні пакети прикладних програм для ПЕОМ містять стандартні процедури розв'язання СЛАР

такими поширеними точними методами, як метод Гауса, метод Жордана-Гауса, квадратного кореня та інші.

До наближених методів розв'язання СЛАР належать метод простої ітерації, метод Зейделя, метод релаксації та інші. Вони дозволяють отримати послідовність $\{X^k\}$ наближень до розв'язку X^* таку, що $\lim_{k\to\infty} X^k = X^*$.

Ітераційні методи прості, легко програмуються і мають малу ***похибку округлення***, ***яка не накопичується,*** але вони дають збіжну послідовність наближень тільки за виконання певної умови, що гарантує виконання принципу стискаючих відображень (дивись пункти 2.2 -2.5).

Розглянемо більш детально ці дві різні групи підходів до розв'язання СЛАР.

3.1 Метод Гауса

Цей метод базується на приведенні шляхом еквівалентних перетворень вихідної системи (3.1) до вигляду з верхньою трикутною матрицею.

$$\left.\begin{array}{l} c_{11}x_1 + c_{12}x_2 + \ldots + c_{1n}x_n = d_1 \\ 0 + c_{22}x_2 + \ldots + c_{2n}x_n = d_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 0 + \ldots 0 + c_{n-1,n-1}x_{n-1} + c_{n-1,n}x_n = d_{n-1} \\ 0 + 0 + \ldots + 0 + c_{nn}x_n = d_n \end{array}\right\} \quad (3.3)$$

Тоді з останнього рівняння відразу визначаємо $x_n = \frac{d_n}{c_{nn}}$. Підставляючи його в попереднє рівняння, знаходимо x_{n-1} і т.д. Загальні формули для отримання розв'язку мають вигляд

$$x_k = \frac{1}{c_{kk}}(d_k - \sum_{j=k+1}^{n} c_{kj}x_j), k = n, n-1, \ldots, 1. \quad (3.4)$$

При обчисленнях за формулами (3.4) треба буде виконати приблизно $1/2n^2$ арифметичних дій. Зведення системи (3.1) до вигляду (3.3) можна виконати, послідовно заміняючи рядки матриці системи їх лінійними комбінаціями. Перше рівняння не змінюється. Віднімемо з другого рівняння системи (3.1) перше, помножене на таке число, щоб звернувся в нуль коефіцієнт при x_1. Потім у такий самий спосіб віднімемо перше рівняння з третього, четвертого і т.д. Таким чином обнуляються всі коефіцієнти першого стовпця, що лежать нижче головної діагоналі. Потім за допомогою другого рівняння виключимо з третього, четвертого і т.д. рівнянь коефіцієнти другого стовпця. Послідовно продовжуючи цей процес, виключимо з матриці всі коефіцієнти, що лежать нижче головної діагоналі.

Запишемо загальні формули процесу. Нехай проведене виключення коефіцієнтів з *k-1* стовпця. Тоді залишилися такі рівняння з ненульовими елементами нижче головної діагоналі:

$$\sum_{j=k}^{n} a_{ij}^{(k)} x_j = b_i^{(k)}, k \le i \le n. \quad (3.5)$$

Помножимо *k*-й рядок на число

$$c_{mk} = \frac{a_{mk}^{(k)}}{a_{kk}^{(k)}}, m > k \quad (3.6)$$

і віднімемо від *m*-го рядка. Перший ненульовий елемент цього рядка звернеться в нуль, а інші зміняться за формулами

$$a_{ml}^{(k+1)} = a_{ml}^{(k)} - c_{mk}a_{kl}^{(k)}, \quad (3.7)$$

$$b_m^{(k+1)} = b_m^{(k)} - c_{mk} b_k^{(k)}, k < m, l \le n.$$

Виконуючи обчислення при всіх зазначених індексах, виключимо елементи k-го стовпця. Будемо називати таке виключення циклом процесу. Виконання всіх циклів називається прямим ходом виключення.

Після виконання всіх циклів утвориться система, матриця якої має трикутний вигляд . Її легко розв'язати зворотним ходом за формулами (3.4).

Виключення за формулами (3.7) не можна проводити, якщо в ході розрахунків на головній діагоналі виявиться нульовий елемент $a_{kk}^{(k)} = 0$ Але в першому стовпці проміжної системи (3.5) всі елементи не можуть бути нулями: це означало б, що $detA=0$. Перестановкою рядків можна перемістити ненульовий елемент на головну діагональ і продовжити розрахунки.

Для зменшення обчислювальної похибки можна кожне повторення зовнішнього циклу починати з вибору максимального за модулем елемента в k-му стовпці (головного елемента) і перестановки рівняння з головним елементом так, щоб він виявився на головній діагоналі. Цей варіант називається методом Гауса з вибором головного елемента.

Однією з характеристик ефективності того чи іншого алгоритму вважають обчислювальні витрати, що визначаються кількістю елементарних операцій, які необхідно виконати для одержання розв'язку. Для прямого ходу методу Гауса число арифметичних операцій, відповідно до (3.6), (3.7), становить

$$Q_1(n) = \sum_{k=1}^{n-1} \sum_{m=k+1}^{n} \left[\textit{ділення} + \sum_{p=k}^{n+1} (\textit{множення} + \textit{віднімання}) \right] =$$

$$= \frac{1}{3} n(n-1)(2n + \frac{13}{2}) = \frac{2}{3} n^3 + \frac{3}{2} n^2 - \frac{13}{6} n.$$

Для зворотного ходу за формулами (3.4) число арифметичних операцій дорівнює

$$Q_2(n) = \sum_{k=1}^{n} (\textit{ділення} + \textit{віднімання} + \sum_{j=k+1}^{n} \textit{множення}) =$$

$$= \frac{1}{2} n(n+3) = \frac{1}{2} n^2 + \frac{3}{2} n.$$

Загальні обчислювальні витрати методу Гауса становлять

$$Q(n) = \frac{2}{3} n^3 + 2n^2 - \frac{2}{3} n, \text{ тобто } Q(n) \approx \frac{2}{3} n^3 = O(n^3).$$

Зауваження 1. *Якщо елементи будь-якого рядка матриці системи в результаті перетворень стали дорівнювати нулю, то СЛАР несумісна, оскільки не виконуються умови теореми Кронекера-Капеллі.*

Зауваження 2. *Якщо елементи будь-якого рядка матриці системи і права частина в результаті перетворень стали дорівнювати нулю, то СЛАР сумісна, але має безліч розв'язків, які можна отримати за допомогою методу Гауса для СЛАР порядку r, де r - ранг матриці заданої СЛАР.*

3.2 Додаткові застосування методу Гауса

Виконання прямого ходу методу Гауса дозволяє також **обчислити значення визначника матриці системи.**

При заміні рядків матриці їхніми лінійними комбінаціями значення визначника не змінюється. Знак змінюється при кожній перестановці рядків. Для трикутної матриці величина визначника дорівнює добутку елементів, що стоять на головній діагоналі. Тому визначник обчислюється за формулою

$$\det A = \pm \prod_{k=1}^{n} a_{kk}^{(k)}.$$

Метод Гауса може бути використаний для **отримання оберненої матриці.** Позначимо її елементи через a_{jm}. Тоді співвідношення $AA^{-1}=E$ можна записати так:

$$\sum_{k=1}^{n} a_{ik}a_{kj} = \delta_{ij}, \qquad 1 \le i, j \le n .$$

Якщо розглядати j-й стовпець оберненої матриці як вектор, то він є розв'язком лінійної системи вигляду (3.1) з матрицею A і спеціальною правою частиною (у якій на j-му місці стоїть одиниця, а на інших нулі). Таким чином, для обертання матриці треба розв'язати n систем лінійних рівнянь з однаковою матрицею A і різними правими частинами. Зведення матриці A до трикутної виконується при цьому тільки один раз, а праві частини перетворюються за формулами (3.6)-(3.7).

Перетворення матриці вимагає порядку $\frac{2}{3}n^3$ операцій. Дії з перетворення правих частин систем і зворотний хід методу Гауса повторюються $7n$ разів, а однократне перетворення правих частин і зворотний хід вимагають порядку $\frac{3}{2}n^2$ операцій. Отже, сумарні обчислювальні витрати на пошук оберненої матриці становлять: $\frac{2}{3}n^3+\frac{3}{2}n^3=\frac{13}{6}n^3$.

Обертання матриці зводиться до розв'язання n систем лінійних рівнянь, але вимагає лише приблизно втроє більше дій, ніж розв'язання однієї системи рівнянь. Це обумовлюється тим, що при розв'язку лінійної системи велика частина обчислень пов'язана з приведенням матриці до трикутного вигляду, а це при обертанні матриці робиться тільки один раз. Зворотний хід і перетворення правих частин виконуються набагато швидше.

3.3 Метод Краута

Суть методу Краута, або LU-розкладання, полягає в тому, що це своєрідний перезапис методу Гауса. Він дозволяє зробити зручною комп'ютерну реалізацію методу Гауса. Можна явно виділити два етапи, у яких один робить перетворення з матрицею A системи, інший – з вектором правих частин b. Отже, нехай дана СЛАР $Ax=b$, наприклад, система розміром 4×4. Запишемо розширену матрицю системи

$$[Ab]=\left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ a_{22} & a_{23} & a_{24} & a_{25} & b_2 \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \\ a_{41} & a_{42} & a_{43} & a_{44} & b_4 \end{array}\right]$$

Тоді, за Гаусом можна явно виділити два етапи (тобто два кроки) – прямий хід (ПХ) і зворотний (ЗХ):

1) ПХ: $a_{kl}^{(i)} = a_{kl}^{(i-1)} - \frac{a_{il}^{(i-1)}}{a_{il}^{(i-1)}} a_{kl}^{(i-1)}$

2) ЗХ: $x_i = b_i^{(i)} - \sum_{j}^{i-1} a_{i,j}x_j$

На прямому ході ми робимо так звані "виключення", тобто приводимо матрицю до трикутного вигляду. Тепер легко знайти x_4, а потім і x_3 і т.д. Це був зворотний хід методу Гауса. Всі ці перетворення виконувалися не із самою матрицею, а з розширеною матрицею.

Головна ідея і потреба методу LU – декомпозиції полягає в тому, щоб розділити окремо етап перетворення коефіцієнтів матриці і окремо етап перетворення вектора правих частин.

Розглянемо k -ий крок методу Гауса, на якому здійснюється занулення піддіагональних елементів k-го

стовпчика матриці $A^{(k-1)}$. Як було зазначено раніше, з цією метою використовується операція

$$a_{ml}^{(k)} = a_{ml}^{(k-1)} - c_{mk} a_{kl}^{(k-1)},$$

$$c_{mk} = \frac{a_{mk}^{(k-1)}}{a_{kk}^{(k-1)}}, m > k \quad m = \overline{k+1, n}, \; l = \overline{k, n}.$$

У термінах матричних операцій така операція еквівалентна множенню $A^{(k)} = M_k A^{(k-1)}$, де елементи матриці M_k визначаються таким чином:

$$m_{ij}^k = \begin{cases} 1, \; i = j; \\ 0, \; i \neq j, \; j \neq k; \\ -c_{k+1,k}, \; i \neq j, \; j = k, \end{cases}$$ тобто матриця M_k має

вигляд $$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & ...-c_{k+1,k} & 1 & 0 & 0 \\ .. & .. & .. & .. & .. & .. \\ 0 & 0 & ...-c_{n,k} & 0 & 0 & 1 \end{pmatrix}.$$

При цьому вираз для зворотної операції запишеться у вигляді $A^{(k-1)} = M_k^{-1} A^{(k)}$, де

$$M_k^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & ...c_{k+1,k} & 1 & 0 & 0 \\ .. & .. & .. & .. & .. & .. \\ 0 & 0 & ...c_{n,k} & 0 & 0 & 1 \end{pmatrix}.$$

У результаті прямого ходу методу Гауса отримаємо

$$A^{(n-1)} = U,$$

$$A = A^{(0)} = M_1^{-1} A^{(1)} = M_1^{-1} M_2^{-1} A^{(2)} = M_1^{-1} M_2^{-1} ... M_{n-1}^{-1} A^{(n-1)},$$

де $A^{(n-1)} = U$ - верхня трикутна матриця, а $L = M_1^{-1} M_2^{-1} ... M_{n-1}^{-1}$ - нижня трикутна матриця, що має вигляд

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ c_{21} & 1 & 0 & 0 & 0 & 0 \\ c_{31} & c_{32} & 1 & 0 & 0 & 0 \\ .. & .. & ...c_{k+1,k} & 1 & 0 & 0 \\ .. & .. & .. & .. & .. & .. \\ c_{n1} & c_{n2} & ...c_{nk} & c_{n,k+1} & ...c_{n,n-1} & 1 \end{pmatrix}.$$

У подальшому LU - розкладання може бути ефективно використано для розв'язання систем лінійних алгебраїчних рівнянь. Це дозволяє один раз перетворити матрицю системи, а потім неодноразово розв'язувати декілька систем з різними правими частинами. Обчислювальні витрати при цьому будуть зводитися тільки до зворотного ходу.

Запишемо $A \cdot x = b$, як

$$L \cdot U \cdot x = b.$$

Позначимо

$$U \cdot x = y.$$

І, отже, $$L \cdot y = b.$$

Таким чином, прямий хід методу LU-декомпозиції складається з розкладу матриці A на нижню L та верхню U трикутні матриці – це прямий хід.

Потім визначається вектор y на основі співвідношень: $y_1 = \frac{b_1}{l_{11}}$, $y_i = \frac{1}{l_{ii}} \left(b_i - \sum_{j=1}^{i-1} l_{ij} y_j \right)$.

На зворотному ході методу *LU*-декомпозиції розв'язується рівняння $U \cdot x = y$. З урахуванням того, що U – трикутна матриця,

$$x_N = \frac{y_N}{u_{NN}} \quad , x_i = \frac{1}{u_{ii}}\left(y_i - \sum_{j=i+1}^{N} u_{ij} x_j\right).$$

Отже, *LU*-розкладання є просто свого роду іншою формою запису еквівалентних перетворень матриці за методом Гауса, але проведених з урахуванням умови $A = L \cdot U$.

Теорема 1. Для існування *LU*-розкладання матриці A необхідно й достатньо, щоб у матриці A всі головні мінори були відмінні від нуля.

У довільної невиродженої матриці А головні мінори, тобто $|a_{11}|$, $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$, …, $\begin{vmatrix} a_{11} & \dots & a_{1n} \\ a_{n1} & \dots & a_{nn} \end{vmatrix}$, можуть дорівнювати нулю. Тоді необхідно переставити рядки так, щоб головні мінори стали відмінними від нуля.

Звичайно перестановка рядків не проводиться окремо від процедури вилучення, ці два процеси поєднуються в один. Якщо $a_{11}=0$, переставимо рядки матриці A так, щоб у лівому верхньому куті виявився ненульовий елемент. У першому стовпці такий елемент завжди знайдеться, інакше $detA = 0$. Якщо після першого кроку дістанемо $a_{22}^{(1)} = 0$, то виконаємо, як і вище, переставлення: у другому стовпці завжди знайдеться ненульовий елемент, інакше два перші стовпці були б лінійно залежні і $det\ A = 0$. Помістимо рядок з ненульовим елементом у другому стовпці на місце другого рядка, тоді $a_{22}^{(1)} \neq 0$. Продовжуючи цей процес вилучення й перестановки рядків (якщо елемент $a_{kk}^{(k-1)}=0$) до $k=n$, дістанемо *LU*-розкладання матриці A з додатковою матрицею P перестановок рядків

$$PA = LU.$$

Матриця P одержується з одиничної матриці E перестановкою тих самих рядків. Наприклад, перестановці другого та четвертого рядків матриці відповідає

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

Таким чином, ми отримали LUP-розкладання.

Приклад. Розв'яжемо СЛАР за схемою *LU*-розкладання:

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

Виконаємо дії за алгоритмом і отримаємо матриці L та U у вигляді:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -3/7 & 1 \end{bmatrix}, \; U = \begin{bmatrix} 1 & -2 & 3 \\ 0 & 7 & -7 \\ 0 & 0 & 1 \end{bmatrix}.$$

Спочатку знаходимо розв'язок системи $Lg = b$

$$\begin{cases} g_1 = 1 \\ 2g_1 + g_2 = 2 \\ -g_1 - 3/7 g_2 + g_3 = 3 \end{cases}$$

Отримаємо: $g = \{1, 0, 4\}$.

Тепер реалізуємо зворотний хід методу Гауса, розв'язуючи систему $Ux = g$:

$$\begin{cases} x_1 - 2x_2 + 3x_3 = 1 \\ 7x_2 - 7x_3 = 0 \\ x_3 = 4 \end{cases}$$

Отже, остаточна відповідь: $x_1 = 4$, $x_2 = 4$, $x_3 = -3$.

3.4 Метод прогонки

Це - ще одна модифікація методу Гауса для систем лінійних алгебраїчних рівнянь спеціального вигляду. Нехай потрібно знайти розв'язок системи так званих триточкових рівнянь:

$$\begin{aligned} &c_0y_0-b_0y_1=f_0, && i=0;\\ &-a_iy_{i-1}+c_iy_i-b_iy_{i+1}=f_i, && 1\le i\le N-1;\\ &-a_Ny_{N-1}+c_Ny_N=f_N, && i=N. \end{aligned} \qquad (3.8)$$

Такі системи виникають як результат апроксимації крайових задач для звичайних диференціальних рівнянь другого порядку, а також при реалізації різницевих схем для рівнянь у часткових похідних.

Переходимо до побудови алгоритму розв'язування системи (3.8). На першому кроці з усіх рівнянь системи (3.8) для $i=1,2,...,N$ виключається за допомогою першого рівняння (3.8) невідоме y_0, потім з перетворених рівнянь для $i=2,3, ...,N$ за допомогою рівняння, що відповідає $i=1$, виключається невідоме y_1 і т.д. У результаті одержимо одне рівняння відносно y_N. На цьому прямий хід методу закінчується. На зворотному ході для $i=N-1,N-2, ...,0$ знаходиться y_i через уже знайдені y_{i+1}, y_{i+2}, ..., y_N і перетворені праві частини.

Використовуючи підходи методу Гауса, проведемо виключення невідомих з (3.8). Позначивши $\alpha_1=b_0/c_0$, $\beta_1=f_0/c_0$, запишемо (3.8) :

$$\begin{aligned} &y_0-\alpha_1y_1=\beta_1, && i=0;\\ &-a_iy_{i-1}+c_iy_i-b_iy_{i+1}=f_i, && 1\le i\le N-1;\\ &-a_Ny_{N-1}+c_Ny_N=f_N, && i=N. \end{aligned} \qquad (3.9)$$

Візьмемо перші два рівняння системи (3.9)

$$y_0-\alpha_1y_1=\beta_1, \qquad -a_iy_{i-1}+c_iy_i-b_iy_{i+1}=f_i.$$

Помножимо перше рівняння на a_1 і додамо до другого рівнянням. Будемо мати $(c_1-a_1\alpha_1)y_1-b_1y_2=f_1+a_1\beta_1$ або після ділення на $c_1-a_1\alpha_1$

$$y_1-\alpha_2y_2=\beta_2, \quad \alpha_2=\frac{b_1}{c_1-\alpha_1a_1}, \quad \beta_2=\frac{f_1+a_1\beta_1}{c_1-\alpha_1a_1}.$$

Усі інші рівняння системи (3.9) y_0 не містять, тому на цьому перший крок процесу виключення закінчується. В результаті одержимо нову "скорочену" систему

$$\begin{aligned} &y_1-\alpha_2y_2=\beta_2, && i=1;\\ &-a_iy_{i-1}+c_iy_i-b_iy_{i+1}=f_i, && 2\le i\le N-1;\\ &-a_Ny_{N-1}+c_Ny_N=f_N, && i=N, \end{aligned} \qquad (3.10)$$

яка не містить невідоме y_0 і має аналогічну (3.9) структуру. Якщо ця система буде розв'язана, то невідоме y_0 знайдеться із рівняння $y_0-\alpha_1y_1=\beta_1$. До системи (3.10) можна знову застосувати описаний спосіб виключення невідомих. На другому кроці буде виключене невідоме y_1, на третьому – y_2 і т.д. У результаті l-го кроку одержимо систему для невідомих y_l, y_{l+1}, ..., y_N

$$\begin{aligned} &y_l-\alpha_{l+1}y_{l+1}=\beta_{l+1}, && i=l;\\ &-a_iy_{i-1}+c_iy_i-b_iy_{i+1}=f_i, && l+1\le i\le N-1;\\ &-a_Ny_{N-1}+c_Ny_N=f_N, && i=N. \end{aligned} \qquad (3.11)$$

і формули для знаходження y_i з номерами $i\le l-1$

$$y_i=\alpha_{i+1}y_{i+1}+\beta_{i+1}, \quad i=l-1, l-2, ..., 0. \qquad (3.12)$$

Коефіцієнти α_i і β_i знаходяться за формулами

$$\alpha_{i+1}=\frac{b_i}{c_i-a_i\alpha_i}, \quad \beta_{i+1}=\frac{f_i-a_i\beta_i}{c_i-a_i\alpha_i}, \ i=1,2,...,$$

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

Вважаючи в (3.12) $l=N-1$, одержимо систему рівнянь для y_N і y_{N-1}

$$\begin{aligned} &y_{N-1}-\alpha_Ny_N=\beta_N,\\ &-a_Ny_{N-1}+c_Ny_N=f_N, \end{aligned}$$

з якої знайдемо $y_N=\beta_{N+1}$, $y_{N-1}=\alpha_Ny+\beta_N$.

Таким чином, одержимо остаточні формули для знаходження невідомих

$$y_i=\alpha_{i+1}y_{i+1}+\beta_{i+1}, \quad i=N-1, N-2, \ldots, 0,$$
$$y_N=\beta_{N+1}, \qquad (3.13)$$

де коефіцієнти знаходяться за рекурентними співвідношеннями:

$$\alpha_{i+1}=\frac{b_i}{c_i-a_i\alpha_i}, \quad i=1, 2, \ldots, N-1, \quad \alpha_1=\frac{b_0}{c_0} \qquad (3.14)$$

$$\beta_{i+1}=\frac{f_i-a_i\beta_i}{c_i-a_i\alpha_i}, \quad i=1, 2, \ldots, N-1; \quad \beta_1=\frac{f_0}{c_0} \qquad (3.15)$$

Отже, формули (3.13)-(3.15) описують метод Гауса, що у застосуванні до системи (3.8) одержав спеціальну назву - метод прогонки. Коефіцієнти α_i і β_i називаються прогонковими коефіцієнтами, формули (3.14), (3.15) описують прямий хід прогонки, а (3.13) - зворотний хід. Оскільки значення y_i знаходяться тут послідовно при переході від $i+1$ до i, то формули (3.13)-(3.15) називають іноді правою прогонкою.

Метод зустрічних прогонок. Аналогічно до правої виводяться формули лівої прогонки:

$$\xi_i=\frac{a_i}{c_i-b_i\xi_{i+1}}, \quad i=N-1, N-2, \ldots, 1; \quad \xi_N=\frac{a_N}{c_N} \qquad (3.16)$$

$$\eta_i=\frac{f_i+b_i\eta_{+1}}{c_i-b_i\xi_{i+1}}, \quad i=N-1, N-2, \ldots, 0; \quad \eta_N=\frac{f_N}{c_N}, \qquad (3.17)$$

$$y_{i+1}=\xi_{i+1}y_i+\eta_{i+1}, \quad i=0, 1, \ldots, N-1; \quad y_0=\eta_0. \qquad (3.18)$$

Тут значення y_i знаходяться послідовно при зростанні індексу i (зліва направо).

Іноді виявляється зручним комбінувати праву і ліву прогонки, одержуючи так званий *метод зустрічних прогонок*. Цей метод доцільно застосовувати, якщо треба знайти тільки одне невідоме, наприклад y_m ($0\le m\le N$), або групу невідомих. Одержимо формули методу зустрічних прогонок. Нехай $1\le m\le N$ і за формулами (3.14)-(3.17) знайдені $\alpha_1, \alpha_2, \ldots, \alpha_m, \beta_1, \beta_2, \ldots, \beta_m$ і $\xi_N, \xi_{N-1}, \ldots, \xi_m, \eta_N, \eta_{N-1}, \ldots, \eta_m$. Випишемо формули (3.13), (3.18) для зворотного ходу правої і лівої прогонок для $i=m-1$. Будемо мати систему, з якої знайдемо y_m

$$y_m=\frac{\eta_m+\xi_m\beta_m}{1-\xi_m\alpha_m}.$$

Використовуючи знайдене y_m, за формулами (3.13) для $i=m-1, m-2, \ldots, 0$ знайдемо послідовно $y_{m-1}, y_{m-1}, \ldots, y_0$, а за формулами (3.18) для $i=m, m+1, \ldots, N$ обчислимо інші $y_{m+1}, y_{m+2}, \ldots, y_N$.

Отже, формули методу зустрічних прогонок мають вигляд:

$$\alpha_{i+1}=\frac{b_i}{c_i-a_i\alpha_i}, \quad i=1, 2, \ldots, m-1, \quad \alpha_1=\frac{b_0}{c_0},$$

$$\beta_{i+1}=\frac{f_i-a_i\beta_i}{c_i-a_i\alpha_i}, \quad i=1, 2, \ldots, m-1, \quad \beta_1=\frac{f_0}{c_0},$$

$$\xi_i=\frac{a_i}{c_i-b_i\xi_{i+1}}, \quad i=N-1, N-2, \ldots, m, \quad \xi_N=\frac{a_N}{c_N}, \qquad (3.19)$$

$$\eta_i=\frac{f_i+b_i\eta_{i+1}}{c_i-b_i\xi_{i+1}}, \quad i=N-1, N-2, \ldots, m, \quad \eta_N=\frac{f_N}{c_N},$$

для обчислення прогонкових коефіцієнтів і

$$y_i=\alpha_{i+1}y_{i+1}+\beta_{i+1}, \quad i=m-1, m-2, \ldots, 0;$$
$$y_{i+1}=\xi_{i+1}y_i+\eta_{i+1}, \quad i=m, m+1, \ldots, N-1; \qquad (3.20)$$
$$y_m=\frac{\eta_m+\xi_m\beta_m}{1-\xi_m\alpha_m}$$

для визначення розв'язків.

3.5 Ітераційні методи розв'язання СЛАР. Метод простих ітерацій

При великій кількості рівнянь прямі методи розв'язання СЛАР (за винятком методу прогонки) стають важко реалізованими на ЕОМ насамперед через складність

зберігання й обробки матриць великої розмірності. У той же час характерною рисою багатьох СЛАР, що виникають у прикладних задачах є розрідженість матриць. Число ненульових елементів таких матриць є малим у порівнянні з їхньою розмірністю. Для розв'язання СЛАР з розрідженими матрицями краще використати ітераційні методи.

Методи послідовних наближень, у яких при обчисленні наступного наближення розв'язку використовуються попередні, уже відомі наближення розв'язку, називаються ітераційними (дивись 2.4).

Розглянемо СЛАР (3.1) з невиродженою матрицею *(det A ≠ 0)*. Розв'яжемо систему (3.1) щодо невідомих при ненульових діагональних елементах $a_{ii} \neq 0,\ i = 1...n$ (якщо який-небудь коефіцієнт на головній діагоналі дорівнює нулю, досить відповідне рівняння поміняти місцями з будь-яким іншим рівнянням). Одержимо систему у вигляді

$$\begin{cases} x_1 = \beta_1 + \alpha_{11}x_1 + \ldots + \alpha_{1n}x_n \\ x_2 = \beta_2 + \alpha_{21}x_1 + \ldots + \alpha_{2n}x_n \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ x_n = \beta_n + \alpha_{n1}x_1 + \ldots + \alpha_{nn}x_n \end{cases} \qquad (3.26)$$

або у векторно-матричній формі $X = \beta + \alpha X$.

Тут $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$, $\alpha = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{n1} \\ \vdots & \cdots & \vdots \\ \alpha_{n1} & \cdots & \alpha_{nn} \end{pmatrix}$.

Вирази для компонентів вектора β та матриці α еквівалентної системи:

$$\beta_i = \frac{b_i}{a_{ii}}; \alpha_{ij} = -\frac{a_{ij}}{a_{ii}}, i, j = 1..n, i \neq j; \alpha_{ij} = 0, i = j, i = 1..n. \qquad (3.27)$$

При такому способі приведення вихідної СЛАР до еквівалентного вигляду метод простих ітерацій ще називають методом Якобі.

За нульове наближення $X^{(0)}$ вектора невідомих візьмемо вектор правих частин $X^{(0)} = \beta$ або $(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})^* = (\beta_1, \beta_2, \ldots, \beta_n)^*$. Тоді метод простих ітерацій набере вигляду

$$\begin{cases} X^{(0)} = \beta \\ X^{(1)} = \beta + \alpha X^{(0)} \\ X^{(2)} = \beta + \alpha X^{(1)} \\ \ldots\ldots\ldots\ldots \\ X^{(k)} = \beta + \alpha X^{(k-1)} \end{cases} \qquad (3.28)$$

З (3.28) бачимо перевагу ітераційних методів у порівнянні, наприклад, з розглянутим вище методом Гауса. В обчислювальному процесі беруть участь тільки добутки матриці на вектор, що дозволяє працювати тільки з ненульовими елементами матриці, значно спрощуючи процес зберігання й обробки матриць. При цьому не відбувається накопичення похибки заокруглення.

Визначення збіжності ітераційного процесу можна знайти в 2.3-2.5. З огляду на сформульовані там теореми, має місце достатня умова збіжності методу простих ітерацій для СЛАР.

Теорема. *Метод простих ітерацій (3.28) збігається до єдиного розв'язку СЛАР(3.26)(а отже, й до розв'язку вихідної СЛАР (3.1)) при будь-якому початковому наближенні $X^{(0)}$, якщо яка-небудь норма матриці α еквівалентної системи менше одиниці* $\|\alpha\| < 1$.

Якщо ж розглядати СЛАР у вигляді (3.1), то достатньою умовою збіжності є діагональна перевага матриці A, тобто $|a_{ii}| > \sum_{j=1, i \neq j}^{n} |a_{ij}| \ \forall i$ (для кожного рядка матриці A модулі елементів, що розміщені на головній діагоналі, більше суми модулів недіагональних елементів).

Очевидно, що в цьому випадку $\|\alpha\|_c$ менше одиниці й, отже, ітераційний процес (3.28) збігається.

При виконанні достатньої умови збіжності оцінка похибки розв'язку на k – й ітерації дається виразом

$$\|X^{(k)} - X^*\| \le \varepsilon^{(k)} = \frac{\|\alpha\|}{1-\|\alpha\|}\|X^{(k)} - X^{(k-1)}\|, \qquad (3.29)$$

де X^* - точний розв'язок СЛАР.

Процес ітерацій зупиняється при виконанні умови $\varepsilon^{(k)} \le \varepsilon$, де ε - задана точність розв'язку.

Беручи до уваги, що з (3.29) випливає нерівність

$$\|X^{(k)} - X^*\| \le \varepsilon^{(k)} = \frac{\|\alpha\|^k}{1-\|\alpha\|}\|X^{(1)} - X^{(0)}\|,$$

можна одержати апріорну оцінку необхідного для досягнення заданої точності числа ітерацій. При використанні за початкове наближення вектора β будемо мати нерівність

$$\frac{\|\alpha\|^{k+1}}{1-\|\alpha\|}\|\beta\| \le \varepsilon,$$

звідки отримаємо апріорну оцінку числа ітерацій k при $\|\alpha\| < 1$

$$k+1 \ge \frac{\lg\varepsilon - \lg\|\beta\| + \lg(1-\|\alpha\|)}{\lg\|\alpha\|}.$$

Варто підкреслити, що ця нерівність дає завищене число ітерацій k, тому рідко використовується на практиці.

Зауваження. Оскільки $\|\alpha\| < 1$ є тільки достатньою (не необхідною) умовою збіжності методу простих ітерацій, то ітераційний процес може збігатися іноді, незважаючи на невиконання умови.

Приклад. Методом простих ітерацій розв'язати СЛАР з точністю $\varepsilon = 0,01$.

$$\begin{cases} 10x_1 + x_2 + x_3 = 12 \\ 2x_1 + 10x_2 + x_3 = 13 \\ 2x_1 + 2x_2 + 10x_3 = 14 \end{cases}$$

Розв'язання. Приведемо СЛАР до еквівалентного вигляду

$$\begin{cases} x_1 = 1,2 - 0,1x_2 - 0,1x_3 \\ x_2 = 1,3 - 0,2x_1 - 0,1x_3 \\ x_3 = 1,4 - 0,2x_1 - 0,2x_2 \end{cases}$$ або $X = \beta + \alpha X$, де

$$\alpha = \begin{pmatrix} 0 & -0,1 & -0,1 \\ -0,2 & 0 & -0,1 \\ -0,2 & -0,2 & 0 \end{pmatrix}; \quad \beta = (1,2 \;\; 1,3 \;\; 1,4)^T.$$

Визначимо $\|\alpha\|_c = 0,4 < 1$. Отже, достатня умова збіжності методу простої ітерації виконана.

Ітераційний процес має в такий вигляд:

$x^{(0)} = \beta$; $x^{(1)} = \beta + \alpha\beta = (0,93 \;\; 0,92 \;\; 0,9)^T$; $\varepsilon^{(1)} = 0,333 > \varepsilon$

$x^{(2)} = \beta + \alpha x^{(1)} = (1,018 \;\; 1,024 \;\; 1,03)^T$; $\varepsilon^{(2)} = 0,0867 > \varepsilon$

$x^{(3)} = \beta + \alpha x^{(2)} = (0,9946 \;\; 0,9934 \;\; 0,9916)^T$; $\varepsilon^{(3)} = 0,0256 > \varepsilon$

$x^{(4)} = \beta + \alpha x^{(3)} = (1,0015 \;\; 1,00192 \;\; 1,0024)^T$; $\varepsilon^{(4)} = 0,0072 < \varepsilon$

Таким чином, обчислювальний процес завершений за 4 ітерації. Відзначимо, що точний розв'язок СЛАР в цьому випадку відомий $X^* = (1,1,1)^*$. Звідси випливає, що задану точність $\varepsilon = 0,01$ задовольняло наближення, отримане вже на третій ітерації.

Відзначимо, що апріорна оцінка необхідної кількості ітерацій дає $k+1 \ge (-2+\lg 0,6-\lg 1,4)/\lg 0,4 = 5,95$, тобто для досягнення точності $\varepsilon = 0,01$, відповідно до апріорної оцінки, необхідно зробити не менше п'яти ітерацій, що ілюструє характерну для апріорної оцінки тенденцію до завищення числа ітерацій.

3.6 Метод Зейделя розв'язання СЛАР

Метод простої ітерації досить повільно збігається. Для його прискорення існує метод Зейделя. Суть його в тому, що при обчисленні компонентів $x_i^{(k+1)}$ вектора невідомих на *(k+1)*-ій ітерації використовуються $x_1^{(k+1)}$, $x_2^{(k+1)}$,...,$x_{i-1}^{(k+1)}$, уже обчислені на *(k+1)*-ій ітерації. Значення інших компонентів беруться з попередньої ітерації. Так само, як і у методі простих ітерацій, будується еквівалентна СЛАР (3.26) і за початкове наближення береться вектор правих частин $X^0=(\beta_1,\beta_2,\ldots,\beta_n)^*$.

Тоді метод Зейделя для пошуку наближення $X^{(k+1)}$ має вигляд

$$\begin{cases} x_1^{k+1} = \beta_1 + \alpha_{11}x_1^k + \alpha_{12}x_2^k + \ldots + \alpha_{1n}x_n^k \\ x_2^{k+1} = \beta_2 + \alpha_{21}x_1^{k+1} + \alpha_{22}x_2^k + \ldots + \alpha_{2n}x_n^k \\ x_3^{k+1} = \beta_3 + \alpha_{31}x_1^{k+1} + \alpha_{32}x_2^{k+1} + \alpha_{33}x_3^k + \ldots + \alpha_{3n}x_n^k \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n^{k+1} = \beta_n + \alpha_{n1}x_1^{k+1} + \alpha_{n2}x_2^{k+1} + \ldots + \alpha_{nm-1}x_{n-1}^{k+1} + \alpha_{nn}x_n^k \end{cases}$$

Із цієї системи бачимо, що $X^{k+1}=\beta+BX^{k+1}+CX^k$, де B - нижня трикутна матриця з діагональними елементами, що дорівнюють нулю, а C - верхня трикутна матриця з діагональними елементами, відмінними від нуля, $\alpha=B+C$.

Отже, $$(E-B)X^{k+1} = CX^k + \beta,$$

звідки $$X^{k+1} = (E-B)^{-1}CX^k + (E-B)^{-1}\beta.$$

Таким чином, метод Зейделя є методом простих ітерацій з матрицею правих частин $\alpha=(E-B)^{-1}C$ і вектором правих частин $(E-B)^{-1}\beta$, й, отже, збіжність і похибку методу Зейделя можна досліджувати за допомогою формул, виведених для методу простих ітерацій, у яких замість матриці α підставлена матриця $(E-B)^{-1}C$, а замість вектора правих частин - вектор $(E-B)^{-1}\beta$. Для практичних обчислень важливо, що як достатні умови збіжності методу Зейделя можуть бути використані умови, наведені вище для методу

простих ітерацій ($\|\alpha\| < 1$ або, якщо використовується еквівалентна СЛАР у формі (3.1), -діагональна перевага матриці A). У випадку виконання цих умов для оцінки похибки на k -ій ітерації можна використати вираз

$$\varepsilon^{(k)} = \frac{\|C\|}{1-\|\alpha\|}\left\|X^{(k)} - X^{(k-1)}\right\|.$$

Відзначимо, що, як і метод простих ітерацій, метод Зейделя може збігатися й при порушенні умови $\|\alpha\| < 1$.

Приклад. Методом Зейделя розв'язати СЛАР із попереднього прикладу.

Розв'язання. Діагональна перевага елементів вихідної матриці СЛАР гарантує збіжність методу Зейделя.

Ітераційний процес будуємо в такий спосіб:

$$x^{(0)} = (1{,}2 \quad 1{,}3 \quad 1{,}4)^T$$

$$\begin{cases} x_1^{(1)} = 1{,}2 - 0{,}1*1{,}3 - 0{,}1*1{,}4 = 0{,}93 \\ x_2^{(1)} = 1{,}3 - 0{,}2*0{,}93 - 0{,}1*1{,}4 = 0{,}974 \\ x_3^{(1)} = 1{,}4 - 0{,}2*0{,}93 - 0{,}2*0{,}974 = 1{,}0192 \end{cases}$$

$$\begin{cases} x_1^{(2)} = 1{,}2 - 0{,}1*0{,}974 - 0{,}1*1{,}0192 = 1{,}0007 \\ x_2^{(2)} = 1{,}3 - 0{,}2*1{,}0007 - 0{,}1*1{,}0192 = 0{,}998 \\ x_3^{(2)} = 1{,}4 - 0{,}2*1{,}0007 - 0{,}2*0{,}998 = 1{,}0003. \end{cases}$$

Таким чином, уже на другій ітерації похибка $\left\|x^{(2)} - x^{(*)}\right\| < 10^{-2} = \varepsilon$, тобто метод Зейделя в цьому випадку збігається швидше ніж метод простих ітерацій.

Приклад . Розв'язання СЛАР **Ax=b**, отримане за допомогою вбудованої функції ***lsolve*** (пакет Mathcad).

Перевірка достатньої умови збіжності методу Зейделя

$$normB(B,n) := \left| \begin{array}{l} \text{for } i \in 1..n \\ \quad s_i \leftarrow \sum_{j=1}^{n} |B_{i,j}| \\ \max(s) \end{array} \right.$$

$normB(B,4) = 0.8$ *Достатня умова виконана.*

$$A := \begin{pmatrix} 15 & 3 & 4 & 5 \\ 2 & 16 & 4 & 5 \\ 2 & 3 & 17 & 5 \\ 2 & 3 & 4 & 18 \end{pmatrix} \qquad b := \begin{pmatrix} 13 \\ -1 \\ 17 \\ -50 \end{pmatrix}$$

Перетворення системи $Ax=b$ до вигляду $x=Bx+c$, зручного для ітерацій.

$x := lsolve(A,b)$

$$PB(A,n) := \left| \begin{array}{l} \text{for } i \in 1..n \\ \quad \text{for } j \in 1..n \\ \quad\quad B_{i,j} \leftarrow 0 \text{ if } i = j \\ \quad\quad B_{i,j} \leftarrow \frac{-A_{i,j}}{A_{i,i}} \text{ if } i \neq j \\ B \end{array} \right.$$

$$Pc(A,b,n) := \left| \begin{array}{l} \text{for } i \in 1..n \\ \quad c_i \leftarrow \frac{b_i}{A_{i,i}} \\ c \end{array} \right. \qquad c := Pc(A,b,4)$$

ORIGIN:= 1 - нумерація масивів починається з одиниці.

$B := PB(A,4)$

$$B = \begin{pmatrix} 0 & -0.2 & -0.2666666667 & -0.3333333333 \\ -0.125 & 0 & -0.25 & -0.3125 \\ -0.1176470588 & -0.1764705882 & 0 & -0.2941176471 \\ -0.1111111111 & -0.1666666667 & -0.2222222222 & 0 \end{pmatrix}$$

$x =$

Алгоритм методу Зейделя

Вхідні параметри: **B** та **c** - матриця **B** та вектор правої частини **c** системи **x=Bx+c**; **n** - порядок матриці **B**; **k**- число ітерацій; **x0** - вектор початкового наближення.

Функція zeid повертає двовимірний масив розмірності kxn; i-й рядок якого – це i-е наближення.

$$zeid(B,c,n,k,x0) := \left| \begin{array}{l} y \leftarrow x0 \\ \text{for } m \in 1..k \\ \quad x \leftarrow y \\ \quad \text{for } i \in 1..n \\ \quad\quad u \leftarrow 0 \\ \quad\quad j \leftarrow 1 \\ \quad\quad \text{while } 1 \le j < i \\ \quad\quad\quad u \leftarrow u + B_{i,j} \cdot y_j \\ \quad\quad\quad j \leftarrow j + 1 \\ \quad\quad j \leftarrow i + 1 \\ \quad\quad \text{while } i < j \le n \\ \quad\quad\quad u \leftarrow u + B_{i,j} \cdot x_j \\ \quad\quad\quad j \leftarrow j + 1 \\ \quad\quad y_i \leftarrow u + c_i \\ \quad \text{for } i \in 1..n \\ \quad\quad rez_{m,i} \leftarrow y_i \\ rez \end{array} \right.$$

Початкове наближення: $x0 := \begin{pmatrix} 0 \\ -1 \\ 1.2 \\ 2 \end{pmatrix}$

Результат роботи функції zeid - 10 перших наближень $y := zeid(B, c, 4, 10, x0)$

$y =$

	1	2	3	4
1	0.08	-0.9975	0.5783823529	-2.748946078
2	1.828246732	0.4234192198	1.5187046629	-3.388976098
3	1.5066536121	0.4285471636	1.7438783399	-3.404136781
4	1.4506352706	0.3839937506	1.7627901245	-3.394689571
5	1.4513537406	0.3762237425	1.7612981853	-3.393142858
6	1.4527900215	0.3759338442	1.7607254539	-3.393126855
7	1.4529953951	0.3760463543	1.7606767307	-3.393157598
8	1.4529961338	0.3760680502	1.7606818574	-3.39316243
9	1.45299204	0.3760687919	1.7606836308	-3.393162498
10	1.4529914397	0.3760684432	1.7606837815	-3.393162407

Абсолютна похибка:

$i := 1..4 \quad d_i := |y_{10,i} - x_i| \qquad \max(d) = 6.714953038 \times 10^{-8}$

3.7 Оцінка похибки і міра обумовленості

Припустимо, що матриця системи лінійних рівнянь і вектор правих частин задані неточно і замість пред'явленої до розв'язку системи

$$AX=b \tag{3.30}$$

у дійсності розв'язується деяка інша система

$$A_1X=b_1, \tag{3.31}$$

де $A_1=A+\Delta$,$b_1=b+\eta$. Позначимо розв'язки (3.30) і (3.31) через X і X_1 відповідно. Оцінимо похибку розв'язку $z = X_1 - X$. Підставимо вирази для A_1, b_1 і X_1 у (3.31)

$$(A+\Delta)(X+z) = b + \eta.$$

Віднімаючи (3.30), одержимо

$$A z + \Delta x + \Delta z = \eta,$$

$$z = A^{-1}(\eta - \Delta x - \Delta z),$$

$$||z|| <= ||A^{-1}|| (||\eta|| + ||\Delta|| \cdot ||x|| + ||\Delta|| \, ||z||). \tag{3.32}$$

Якщо малі $\|\Delta\|$ і $\|\eta\|$, то варто очікувати і малості $||z||$. Тоді доданок Δz має більш високий порядок малості. Звідси випливає оцінка похибки

$$\|z\| \le \frac{\|A^{-1}\|(\|\eta\| + \|\Delta\|\|x\|)}{1 - \|A^{-1}\|\|\Delta\|}. \tag{3.33}$$

Досить поширений випадок, коли похибка матриці системи істотно менша похибки правої частини; як модель цієї ситуації будемо розглядати випадок точного задання матриці системи. Тоді, вважаючи $\Delta = 0$ у (3.33), маємо

$$||z|| <= ||A^{-1}|| \, ||\eta||. \tag{3.34}$$

Для якісної характеристики зв'язку між похибками правої частини і розв'язку вводиться поняття *обумовленості* матриці системи. Абсолютні похибки правої частини і розв'язку системи залежать від масштабів, якими вимірюються ці величини і матриця системи. Тому правильніше характеризувати властивості системи через зв'язок між відносними похибками правої частини і розв'язку. Для відносної похибки розв'язку з (3.34) маємо

$$\frac{\|z\|}{\|x\|} \le \|A^{-1}\| \frac{\|\eta\|}{\|x\|}. \qquad (3.35)$$

Підставляючи оцінку для $||x||$ у (3.35), маємо

$$\frac{\|z\|}{\|x\|} \le \|A^{-1}\| \|A\| \frac{\|\eta\|}{\|b\|}. \qquad (3.36)$$

Величину $||A^{-1}|| ||A|| = condA$ *називають мірою обумовленості матриці.* Величина відносної похибки розв'язку при фіксованій величині відносної похибки правої частини може стати як завгодно великою при досить великій мірі обумовленості матриці. Число обумовленості залежить від вибору норми матриці. Будь-яка норма матриці не менша від її найбільшого за модулем власного значення, тобто $||A|| >= max|\lambda_A|$. Власні значення матриці A і A^{-1} взаємно обернені; тому $\|A^{-1}\| \ge \frac{1}{|\lambda_A|} = \frac{1}{\min|\lambda_A|}$.

Таким чином, $condA \ge \frac{\max|\lambda_A|}{\min|\lambda_A|} \ge 1$.

Системи рівнянь і матриці з великими значеннями мір обумовленості прийнято називати *погано обумовленими*, а з малими - *добре обумовленими*. Отже, при розв'язуванні СЛАР на ЕОМ обов'язково виникають похибки заокруглення. Тому фактично маємо розв'язок $\tilde{X}$ деякої іншої системи $\tilde{A}\tilde{X} = \tilde{b}$. На практиці важливо знати відносну похибку $\delta X = \frac{\|X - \tilde{X}\|}{\|X\|}$. Якщо замість $\tilde{A}\tilde{X} = \tilde{b}$ брати модель $A\tilde{X} = \tilde{b}$, тобто A в ЕОМ задається точно, то з попередніх співвідношень випливає

$\delta X \le condA\,\delta b$ ($condA$ - міра невизначеності розв'язку системи при неточних вхідних даних).

Якщо брати систему $\tilde{A}\tilde{X} = b$, в якій збурені лише елементи A, а b – точне, то, використовуючи співвідношення $C^{-1} - B^{-1} = B^{-1}(B - C)C^{-1}$, дістаємо $\tilde{X} - X = [\tilde{A}^{-1} - A^{-1}]b = -A^{-1}(\tilde{A} - A)\tilde{A}^{-1}b = -A^{-1}(\tilde{A} - A)\tilde{X}$;

$$\|\tilde{X} - X\| \le \|A^{-1}\| \|\tilde{A} - A\| \|\tilde{X}\| \Rightarrow$$

$$\|\tilde{X} - X\| \le cond\,A \frac{\|\tilde{A} - A\|}{\|A\|}.$$

Лема. Якщо C-квадратна матриця така, що $\|C\| < 1$, то існує $(I+C)^{-1}$, причому $\|(I+C)^{-1}\| \le \frac{1}{1 - \|C\|}$.

Доведення.

$$\|(I + C) \times X\| = \|X + C \times X\| \ge \|X\| - \|C \times X\| \ge (1 - \|C\|) \times \|X\|$$

Оскільки $(I - \|C\|) > 0 \Rightarrow \|(I + C)X\| > 0 \quad (X \ne 0) \Rightarrow$, СЛАР $(I + C) \times X = 0$ має лише тривіальний розв'язок, що й означає невиродженість матриці $I+C$.

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

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

$$= \|(I + C)^{-1}\| - \|C\| \times \|(I + C)^{-1}\| = \|(I + C)^{-1}\| \times (1 - \|C\|) > 0 \Rightarrow \Delta$$

Теорема. Нехай A – невироджена квадратна матриця. Тоді , якщо X та $\tilde{X} = X + \Delta X$ є відповідно розв'язками систем $AX=b$ та $\tilde{A}\tilde{X} = \tilde{b}$, де $\tilde{A}=A+\Delta A$

$\widetilde{b} = b + \Delta b$, причому $\|\Delta A\| < \frac{1}{\|A^{-1}\|}$, то можлива оцінка

$$\frac{\|\Delta X\|}{\|X\|} \le \frac{condA}{1 - condA\frac{\|\Delta A\|}{\|A\|}}\left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|}\right).$$

Доведення. Оскільки $\|A^{-1} \times \Delta A\| \le \|A^{-1}\| \times \|\Delta A\| < 1$, то внаслідок леми існує $(I + A^{-1}\Delta A)$ причому

$$\left\|(I + A^{-1}\Delta A)^{-1}\right\| \le \frac{1}{1 - \|A^{-1}\Delta A\|} \le \frac{1}{1 - \|A^{-1}\| \times \|\Delta A\|}$$

$$A^{-1}\widetilde{A}\widetilde{X} = A^{-1}\widetilde{b}; \quad X = A^{-1}b.$$

Знайдемо

$$\Delta X : (I + A^{-1}\Delta A)X + (I + A^{-1}\Delta A)\Delta X = A^{-1}b + A^{-1}\Delta b \Rightarrow$$

$$\Rightarrow \Delta X = (I + A^{-1}\Delta A)^{-1} A^{-1}(\Delta b - \Delta A X) \Rightarrow$$

$$\frac{\|\Delta X\|}{\|X\|} \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\Delta A\|}\left(\frac{\|\Delta b\|}{\|X\|} + \|\Delta A\|\right); \text{ оскільки } \|\Delta X\| \ge \frac{\|b\|}{\|A\|},$$

дістаємо шукане.

Приклад реалізації чисельного алгоритму розв'язання СЛАР на псевдокоді.

//Метод Зейделя. Вважаємо, що умова збіжності методу перевірена

//повертає норму матриці. A – матриця

```
NormaMatrix(A):
1     temp:=0
2     for i:=1 to A.lengthi do
3       sum:=0
4       for j:=1 to A.lengthj do
5         sum+=abs(A[i,j])
6       done
7       if sum>temp then temp:=sum
8       fi
9     done
10    return temp
end
```



```
//повертає норму вектора. B – вектор
NormaVector(B):
1   sum:=0
2   for i:=1 to B.length do
3     sum+=sqr(abs(B[i]))
4   done
5   return sqrt(sum)
end
```

//повертає обернену матрицю до даної A0 з точністю eps

```
gaussinv(A0,eps):
//доступна в модулі naz.pas
end
```

```
//розв'язує систему методом Зейделя
//A – матриця коефіцієнтів
//B – стовпець вільних членів
//X – вектор відповідей
```

```
//eps – точність обчислень
// змінні, що обчислюються, але не повертаються функцією як
// результат можуть бути використані в конкретних реалізаціях // алгоритму
Solve_Zeidel(A,B,X,eps):
1   n:=A.lengthI;
2   for i:=1 to n do  //AX=B приводимо до вигляду X=CX+D
3     for j:=1 to n do
4       if i<>j then
5     C[i,j]:=(-1)*A[i,j]/A[i,i]
6       else C[i,j]:=0
7       fi
8     done
9   done
10  delta:=(1-NormaVector(C))*eps/NormaVector(C)
11  for i:=1 to n do //обчислюємо A^-1
12    for j=1 to n do
13      A0[i,j]:=A[i,j]
14    done
15  done
16  gaussinv(A0,eps)
17  condA:=norma(A)*norma(A0)
```



```
18  fault:=(condA*((0.01/norma(B))+(0.01/norma(A))))/(1-(condA*0.04)/norma(A)); //обчислюємо похибку
19  for i:=1 to n do
20    X[i]:=B[i]
21  done
22  k:=0
23  repeat  //ітераційний процес
24  k++
25  maxr:=0
26  r:=0
27  for i:=1 to n do
28    xk:=X[i]
29    s:=0
30    for j:=1 to n do
31      s+=C[i,j]*X[j]
32      X[i]:=s+D[i]
33    done
34    r:=abs(xk-X[i])
35    if maxr<r then
36      maxr:=r
37    fi
38  done
39  until maxr<=delta;
end
```

Питання і завдання до теми

"Розв'язання систем лінійних алгебраїчних рівнянь точними методами"

1 Норми векторів і матриць. Абсолютна й відносна похибки вектора.

2 Обумовленість задачі розв'язання системи лінійних алгебраїчних рівнянь. Оцінка похибки розв'язку за похибками вхідних даних: $\delta(x^*) \le cond(A)(\delta(b^*)+\delta(A^*))$.

3 Метод Гауса (схема єдиного ділення):опис методу, трудомісткість методу.

4 Метод Гауса з вибором головного елемента за стовпцем (схема часткового вибору): опис методу, його обчислювальна стійкість.

5 Застосування методу Гауса для розв'язання інших задач обчислювальної алгебри.

6 Метод прогонки з тридіагональною матрицею: опис методу, умови його застосування і переваги.

7 Трудомісткість методу прогонки.

8 Матрична форма запису методу Гауса.

9 LU-розкладання матриці. Теорема про можливості застосування LU- розкладання (без доведення).

10 Застосування методу LU- розкладання для розв'язку задач обчислювальної алгебри.

11 Стратегії вибору провідного елемента в методі Гауса.

12 Метод Гаусса із частковим вибором у матричній формі.

13 Обчислити норми векторів:a) $\|a\|_m$, a=(-3, 0, 4, -5); b) $\|a\|_l$, a=(2, 6, 0); c) $\|a\|_k$, a=(-13 , 7, -4,

14 Обчислити норми матриць

a) $\|A\|_m$, де $A=\begin{pmatrix} 3 & -1 & 1 \\ 9 & 4 & -1 \\ 1 & -2 & 4 \end{pmatrix}$,

b) $\|A\|_l$, де $A=\begin{pmatrix} 2 & -1 & 1 \\ 0 & 2 & -1 \\ 1 & -2 & 0 \end{pmatrix}$,

c) $\|A\|_k$, де $A=\begin{pmatrix} 7 & -7 & 6 \\ 10 & 11 & -4 \\ 3 & -20 & 40 \end{pmatrix}$.

15 Чи є вираз $\min\{|x_1|+2|x_2|,2|x_1|+|x_2|\}$ нормою вектора $x \in R^2$?

16 Довести властивість норм матриць A й B: $\|AB\| \le \|A\| \cdot \|B\|$.

17 Нехай $A = A^T$. Довести, що $A > 0$, тоді й тільки тоді, коли $\lambda_i(A) > 0 \ \forall i$, де $\lambda_i(A)$ - власні значення матриці A.

18 Перевірити справедливість властивостей числа обумовленості:

a) $cond(E) = 1$, b) $cond(A) \ge 1$,

c) $cond(\alpha A) = cond(A) \ \forall \alpha \ne 0$.

19 Оцінити кількість правильних значущих цифр розв'язку системи лінійних алгебраїчних рівнянь, якщо матриця системи A задана точно, елементи вектора правих частин задані із трьома правильними значущими цифрами, а $cond(A) = 10^3$.

Питання і завдання до теми "Розв'язання систем лінійних алгебраїчних рівнянь ітераційними методами"

1 Розв'язати систему $\begin{cases} 10x_1 + x_2 - x_3 = 11 \\ x_1 + 10x_2 - x_3 = 10 \\ -x_1 + x_2 + 10x_3 = 10 \end{cases}$ методом простої ітерації (методом Якобі) з точністю 0.01.

2 Зробити 3 ітерації за методом Зейделя, попередньо перетворивши системи до вигляду, зручного для ітерації. За початкове наближення взяти нульовий вектор. Зобразити графічно поведінку ітераційного процесу. Проаналізувати отримані результати з погляду збіжності (розбіжності) методу.

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

$$\begin{cases} 2x_1 - 0.5x_2 = 3 \\ 2x_1 + 0.5x_2 = 1 \end{cases}$$

3 Перетворити систему до вигляду, зручного для ітерації:

$$\begin{cases} 2x_1 - 1.8x_2 + 0.4x_3 = 1 \\ 3x_1 + 2x_2 - 1.1x_3 = 0 \\ x_1 - x_2 + 7.3x_3 = 0 \end{cases}$$

Перевірити виконання достатньої умови збіжності.

4 Переконатися в тім, що якщо А - нижня трикутна матриця, з ненульовими діагональними елементами, то метод Зейделя збігається за одну ітерацію.

5 Переконатися в тім, що якщо А - діагональна матриця з ненульовими діагональними елементами, то метод Зейделя збігається за одну ітерацію.

6 Переконатися в тім, що якщо А - верхня трикутна матриця, з ненульовими діагональними елементами, то метод Зейделя збігається за скінченне число ітерацій. Знайти цю кількість ітерацій.

7 При яких значеннях a і b метод простої ітерації, застосований для розв'язання системи $x = Bx + c$ з $B = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$ і деяким вектором c, збігається?

8 Нехай система $Ax = b$ розв'язується методом Якобі $x^{(n+1)} = Bx^{(n)} + c$, n=0,1,..... Показати, що достатня умова збіжності методу $\|B\| < 1$ (при $\|B\| = \|B\|_1$ й $\|B\| = \|B\|_\infty$) еквівалентна умові діагональної переваги матриці A.

У задачах 9-13 передбачається, що ітераційні методи розв'язання системи $Ax = b$ записані в канонічній формі $B\dfrac{x^{(n+1)} - x^{(n)}}{\tau} + Ax^{(n)} = b$, n=0,1,..., *де* B *й* τ - *ітераційні параметри.*

9 Нехай всі власні значення матриці А дійсні й додатні. Довести збіжність методу $\dfrac{x^{(n+1)} - x^{(n)}}{\tau} + Ax^{(n)} = b$ при $\tau = \|A\|^{-1}$ з будь-якою матричною нормою.

10 Нехай А - матриця простої структури й всі власні числа $\lambda(A) \in [m, M]$, m>0. Довести, що ітераційний метод із задачі 9 збігається при $0 < \tau < 2/M$.

11 Довести, що для систем $Ax = b$ 2-го порядку метод простої ітерації (метод Якобі)

$$Dx^{(n+1)} + (A_1 + A_2)x^{(n)} = b$$

і метод Зейделя:

$$(D + A_1)x^{(n+1)} + A_2 x^{(n)} = b$$

збігаються і розбігаються одночасно. Тут $A = D + A_1 + A_2$, D-діагональна матриця, A_1- нижня трикутна матриця, A_2 - верхня трикутна матриця.

12 Довести, що для методу Зейделя необхідною й достатньою умовою збіжності є така умова: всі корені λ рівняння

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \dots & a_{1m} \\ a_{21}\lambda & a_{22}\lambda & \dots & a_{2m} \\ & \dots & & \\ a_{m1}\lambda & a_{m2}\lambda & \dots & a_{mm}\lambda \end{vmatrix} = 0$$

за модулем повинні бути менше 1. Тут a_{ij}, $i, j = 1,\dots,m$ – елементи матриці A вихідної системи $Ax = b$.

13 Довести, що якщо $A = A^T > 0$, то справедлива оцінка

$\lambda_{\min}\|x\|_2^2 \le (Ax, x) \le \lambda_{\max}\|x\|_2^2$, $\forall x \in R^m$, де $\lambda_{\min}$ й $\lambda_{\max}$ - мінімальне й максимальне власні значення матриці A.

Розділ 4
Чисельне розв'язування систем нелінійних рівнянь

Розглянемо систему нелінійних рівнянь

$$\begin{cases} f_1(x_1, x_2, \dots x_n) = 0 \\ f_2(x_1, x_2, \dots x_n) = 0 \\ \dots\dots\dots \\ f_n(x_1, x_2, \dots x_n) = 0. \end{cases} \quad (4.1)$$

Представимо цю систему в матричному вигляді

$$\vec{f}(\vec{x}) = 0\ , \quad (4.2)$$

де $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$, $\vec{f} = \begin{pmatrix} f_1 \\ f_2 \\ \dots \\ f_n \end{pmatrix}$.

Очевидно, для нелінійного рівняння (4.2) можна застосувати підходи, викладені в розділі 2 нашої книги, а саме там йшлося про ітераційні методи отримання наближень до кореня для нелінійних рівнянь, визначених на множинах довільної природи. Тут же йдеться про розв'язок на множині елементів з R^n. Розглянемо особливості застосування ітераційних методів для розв'язання систем нелінійних алгебраїчних рівнянь (СНАР).

4.1 Метод простих ітерацій

Нехай система нелінійних рівнянь (4.2) приведена до спеціального вигляду $\vec{x} = \vec{\varphi}(\vec{x})$, де функції φ_i дійсні,

визначені і неперервні в деякій області ізольованого розв'язку $\vec{x}^*$ цієї системи.

Як відомо з розділу 2, для визначення вектора-кореня $\vec{x}^*$ цієї системи зручно користуватися *методом простої ітерації*, де ітераційний процес організується за формулою $\vec{x}^{(p+1)} = \vec{\varphi}\left(\vec{x}^{(p)}\right)$, обравши якесь початкове наближення $x^0 \approx \vec{x}^*$.

Уведемо $\rho(x', x'') = \max_i |x_i' - x_i''|$. Нехай кожне рівняння системи має вигляд $x_i = \varphi_i(x_1, x_2, ..., x_n)$, $i = 1, ..., n$, причому задовольняє умову Ліпшиця $|\varphi_i(x') - \varphi_i(x'')| \le K\rho(x', x'')$, тоді при $K < 1$ цей ітераційний процес збігається. Цей факт випливає з принципу стискаючих відображень, причому $\rho(x^k, X^*) = \rho(\varphi(x^{k-1}), \varphi(X^*)) \le K\rho(x^{k-1}, X^*) \le K^k \rho(x^0, X^*)$.

Якщо R – сукупність векторів x, для яких $\rho(x, y_0) \le r$, то в R є єдиний розв'язок. Розглянемо матриці

$$A_k = \left\{ \frac{\partial \varphi^{(x^k)}}{\partial x_j} \right\}, \text{ такі що}$$

$$x^{k+1} - X^* = A_k(x^k - X^*) = A_k A_{k-1} ... A_0 (x^0 - X^*).$$

Далі – як в ітераційних процесах. Для того щоб ітераційний процес збігався, необхідно й достатньо виконання умови $\|A_k A_{k-1} ... A_0\| \to 0$, при $k \to \infty$. Цю умову важко перевірити, тому використовується достатня умова $\|A_k\| < 1$ при будь-якому k.

При виконанні умов збіжності ітераційного процесу для розв'язання системи нелінійних рівнянь можна застосовувати аналог методу Зейделя:

$$x_1^{k+1} = \varphi_1(x_1^k, x_2^k, ..., x_n^k)$$
$$x_2^{k+1} = \varphi_2(x_1^{k+1}, x_2^k, ..., x_n^k)$$
$$.....$$
$$x_n^{n+1} = \varphi_n(x_1^{k+1}, ..., x_{n-1}^{k+1}, x_n^k)$$

Теорема 1 *Нехай область G замкнена і відображення $\vec{y} = \vec{\varphi}(\vec{x})$ є стискаючим у G, тобто виконана умова $\|\vec{y}_1 - \vec{y}_2\| \le q\|\vec{x}_1 - \vec{x}_2\|$. Тоді, якщо для ітераційного процесу $\vec{x}^{(p)} = \vec{\varphi}\left(\vec{x}^{(p-1)}\right)$ всі послідовні наближення $x^{(p)} \in G$, то:*

1) *незалежно від вибору початкового наближення ітераційний процес збігається, тобто існує*

$$\vec{x}^* = \lim \vec{x}^{(p)} \text{ при } p \to \infty;$$

2) *граничний вектор x^* є єдиним розв'язком рівняння $\vec{x} = \vec{\varphi}(\vec{x})$ в G;*

3) *справедлива оцінка* $\| \vec{x}^* - \vec{x}^{(p)} \| \le \frac{q^p}{1-q} \| \vec{x}^{(1)} - \vec{x}^0 \|$.

Теорема 2 *Нехай $\vec{\varphi}(\vec{x})$ і $\vec{\varphi}'(\vec{x})$ неперервні в області G, причому в G виконується нерівність*

$$\| \vec{\varphi}'(\vec{x}) \| \le q < 1.$$

Якщо послідовні наближення $\vec{x}^{(p+1)} = \vec{\varphi}\left(\vec{x}^{(p)}\right)$ не виходять з G, то процес ітерації збігається і граничний вектор $\vec{x}^* = \lim \vec{x}^{(p)}$ при $p \to \infty$ є в G єдиним розв'язком.

4.2 Ітераційний метод Ньютона для СНАР

Нехай, керуючись підходами, викладеними в розділі 2, знайдено p-е наближення

$$\vec{x}^{(p)} = \left(x_1^{(p)}, \ x_2^{(p)}, \ \ldots \ x_n^{(p)}\right)$$

одного з ізольованих коренів $\vec{x} = (x_1, x_2, \ldots x_n)$ векторного рівняння (4.2). Тоді точний корінь можна подати у вигляді

$$\vec{x} = \vec{x}^{(p)} + \vec{\varepsilon}^{(p)}, \qquad (4.3)$$

де $\vec{\varepsilon}^{(p)} = (\varepsilon_1^{(p)}, \varepsilon_2^{(p)}, \ldots \varepsilon_n^{(p)})$ - похибка кореня.

Підставимо (4.3) у (4.2):

$$\vec{f}(\vec{x}^{(p)} + \vec{\varepsilon}^{(p)}) = 0. \qquad (4.4)$$

Нехай $\vec{f}(\vec{x})$ – неперервна диференційована функція в деякій опуклій області, що містить $\vec{x}$ і $\vec{x}^{(p)}$, розкладемо ліву частину (4.4) в ряд за степеням малого вектора $\vec{\varepsilon}^{(p)}$, обмежившись лінійними членами:

$$\vec{f}(\vec{x}^{(p)} + \vec{\varepsilon}^{(p)}) = \vec{f}(\vec{x}^{(p)}) + \vec{f}'(\vec{x}^{(p)})\vec{\varepsilon}^{(p)} = 0 \qquad (4.5)$$

З (4.5) випливає, що під $\vec{f}'(\vec{x}^{(p)})$ треба розуміти матрицю Якобі системи функцій $f_1, f_2, \ldots, f_n$ щодо $x_1, x_2, \ldots, x_n$

$$W(\vec{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \ldots & \frac{\partial f_1}{\partial x_n} \\ & \ldots\ldots & \\ \frac{\partial f_n}{\partial x_1} & \ldots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

Система (4.5) являє собою систему лінійних рівнянь відносно похибок $\varepsilon_i^{(p)}$ $(i=1,2,\ldots,n)$ з матрицею $W(x)$, тому формула (4.5) набере вигляду

$$\vec{f}(\vec{x}^{(p)}) = W(\vec{x}^{(p)})\vec{\varepsilon}^{(p)} = 0.$$

Допускаючи, що $W(x)$– невироджена, знаходимо

$$\vec{\varepsilon}^{(p)} = -W^{-1}(\vec{x}^{(p)})\vec{f}(\vec{x}^{(p)}),$$

значить,

$$\vec{x}^{(p+1)} = \vec{x}^{(p)} - W^{-1}(\vec{x}^{(p)})\vec{f}(\vec{x}^{(p)}). \qquad (4.6)$$

Отримали інтерполяційну формулу Ньютона для СНАР. Очевидно, формула (4.6) дозволить побудувати

збіжну до кореня ітераційну послідовність за умови, що відображення $\Phi(\vec{x}) = \vec{x} - W^{-1}(\vec{x})\vec{f}(\vec{x})$ буде стискаючим. Для цього треба правильно обрати нульове наближення $\vec{x}^0$.

Теорема 3. *Маємо нелінійну систему рівнянь з дійсними коефіцієнтами (4.2), де вектор-функція визначена і неперервна разом зі своїми частковими похідними 1-го і 2-го порядків в області ω. Вважатимемо, що $\vec{x}^0$ є точка, яка лежить у ω разом зі своїм замкненим* H *-околом. Причому виконані умови*:

1) Матриця Якобі при $\vec{x} = \vec{x}^0$ має обернену Γ_0, де $||\Gamma_0|| <= A_0$, (в змісті m-норми)

2) $||\Gamma_0 f(x^0)|| <= B_0 <= H/2$

3) $$\sum_{k=1}^{n} \left| \frac{\partial^2 f_i(x)}{\partial x_j \partial x_k} \right| <= C \text{ при } i,j=1,2,\ldots,n$$

4) постійні A_0, B_0 і C задовольняють нерівність $\mu_0 = 2n_0 B_0 C <= 1$

Тоді процес Ньютона (4.6) при початковому наближенні $\vec{x}^0$ збігається і граничний вектор $\vec{x}^* = \lim \vec{x}^{(p)}$ є розв'язком системи таким, що $||x^* - x^0|| <= 2B_0 <= H$.

4.3 Модифікований метод Ньютона

При побудові процесу Ньютона (4.6) істотною незручністю є необхідність для кожного кроку заново обчислювати обернену матрицю Якобі. Якщо ця матриця неперервна в околі шуканого розв'язку x^0, досить близького до x^*, то приблизно можна покласти

$$W^{-1}(\vec{x}^{(p)}) \approx W^{-1}(\vec{x}^0)$$

і ми приходимо до *модифікованого методу Ньютона*

$$\vec{\xi}^{(p+1)} = \vec{\xi}^{(p)} - W^{-1}(\vec{x}^0)\vec{f}(\vec{\xi}^{(p)}).$$

4.4 Метод градієнтного спуску

Припустимо, що в системі нелінійних алгебраїчних рівнянь (4.2) функції f_i дійсні і неперервно диференційовані в їхній загальній області визначення. Розглянемо функцію

$$U(\vec{x}) = \sum_{i=1}^{n} [f_i(\vec{x})]^2 = \left(\vec{f}(\vec{x}),\ \vec{f}(\vec{x})\right). \qquad (4.7)$$

Очевидно, що кожен розв'язок системи (4.2) перетворює в нуль функцію $U(x)$; і навпаки, числа $x_1, x_2, \ldots, x_n$, для яких функція $U(x)$ дорівнює нулю, є коренем системи (4.2).

Припустимо, що система має лише ізольований розв'язок, що являє собою точку строгого мінімуму функції $U(x)$ у n-вимірному просторі $R^n = \{x_1, x_2, \ldots, x_n\}$.

Нехай $\vec{x}^*$ - корінь системи (4.2) і $\vec{x}^0$ - його нульове наближення. Через точку $\vec{x}^0$ проведемо поверхню рівня функції $U(\vec{x})$. Якщо точка $\vec{x}^0$ досить близька до кореня $\vec{x}^*$; то при наших припущеннях поверхня рівня $U(\vec{x}) = U(\vec{x}^0)$ буде схожа на еліпсоїд.

З точки $\vec{x}^0$ рухаємося по нормалі до поверхні $U(\vec{x}) = U(\vec{x}^0)$ доти, поки ця нормаль не доторкнеться в деякій точці $\vec{x}^1$ іншої поверхні рівня $U(\vec{x}) = U(\vec{x}^1)$.

Потім, відправляючись від точки $\vec{x}^1$, знову рухаємося по нормалі до поверхні рівня $U(\vec{x}) = U(\vec{x}^1)$ доти, поки ця нормаль не доторкнеться в деякій точці $\vec{x}^2$ нової поверхні рівня $U(\vec{x}) = U(\vec{x}^2)$, і т.д.

Оскільки $U(\vec{x}^0) > U(\vec{x}^1) > U(\vec{x}^2) > \ldots$, то, рухаючись таким чином, ми швидко наближаємося до точки з найменшим значенням U ("дно ями"), що відповідає кореневі системи (4.2). Позначимо через

$$\nabla U(\vec{x}) = \begin{bmatrix} \dfrac{\partial U}{\partial x_1} \\ \ldots \\ \dfrac{\partial U}{\partial x_n} \end{bmatrix}$$

градієнт функції $U(\vec{x})$. Визначимо описаний алгоритм пошуку точок-наближень за формулою

$$\vec{x}^{(p+1)} = \vec{x}^{(p)} - \lambda_p \nabla U\left(\vec{x}^{(p)}\right). \qquad (4.8)$$

Залишається визначити множники λ_p. Для цього розглянемо скалярну функцію $\Phi(\lambda) = U\left[\vec{x}^{(p)} - \lambda \nabla U\left(\vec{x}^{(p)}\right)\right]$.

Функція $\Phi(\lambda)$ дає зміну рівня функції U уздовж відповідної нормалі до поверхні рівня в точці $\vec{x}^{(p)}$. Множник $\lambda = \lambda_p$ потрібно вибрати таким, щоб $\Phi(\lambda)$ мала мінімум. Керуючись необхідною умовою екстремуму функції, одержуємо рівняння

$$\Phi'(\lambda) = \frac{d}{d\lambda} U\left[\vec{x}^{(p)} - \lambda \nabla U\left(\vec{x}^{(p)}\right)\right] = 0. \qquad (4.9)$$

Найменший додатний корінь цього рівняння і дасть нам значення λ_p. Будемо вважати, що λ - мала величина, квадратом і вищими ступенями якої можна зневажити. Маємо $\Phi(\lambda) = \sum_{i=1}^{n} \left\{ f_i\left[\vec{x}^{(p)} - \lambda \nabla U\left(\vec{x}^{(p)}\right)\right]\right\}^2$. Розкладаючи функції f_i за степенями λ з точністю до лінійних членів, одержимо

$$\Phi(\lambda) = \sum_{i=1}^{n} \left[f_i\left(\vec{x}^{(p)}\right) - \lambda \frac{\partial f_i\left(\vec{x}^{(p)}\right)}{\partial \vec{x}} \nabla U\left(\vec{x}^{(p)}\right)\right]^2,$$

де $\dfrac{\partial f_i}{\partial \vec{x}} = \left[\dfrac{\partial f_i}{\partial x_1}, \ldots, \dfrac{\partial f_i}{\partial x_n}\right]$. Звідси

$$\Phi'(\lambda)=-2\sum_{i=1}^{n}\left[f_i\left(\vec{x}^{(p)}\right)-\lambda\frac{\partial f_i\left(\vec{x}^{(p)}\right)}{\partial\vec{x}}\nabla U\left(\vec{x}^{(p)}\right)\right]\frac{\partial f_i\left(\vec{x}^{(p)}\right)}{\partial\vec{x}}\nabla U\left(\vec{x}^{(p)}\right)=0$$

Отже,

$$\lambda_p=\frac{\sum_{i=1}^{n}f_i\left(\vec{x}^{(p)}\right)\frac{\partial f_i\left(\vec{x}^{(p)}\right)}{\partial\vec{x}}\nabla U\left(\vec{x}^{(p)}\right)}{\sum_{i=1}^{n}\left[\frac{\partial f_i\left(\vec{x}^{(p)}\right)}{\partial\vec{x}}\nabla U\left(\vec{x}^{(p)}\right)\right]^2}=\frac{\left(\vec{f}\left(\vec{x}^{(p)}\right),W\left(\vec{x}^{(p)}\right)\nabla U\left(\vec{x}^{(p)}\right)\right)}{\left(W\left(\vec{x}^{(p)}\right)\nabla U\left(\vec{x}^{(p)}\right),W\left(\vec{x}^{(p)}\right)\nabla U\left(\vec{x}^{(p)}\right)\right)},$$

де $W(\vec{x})=\begin{pmatrix}\frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n}\\ \dots\dots & & \\ \frac{\partial f_n}{\partial x_1} & \dots & \frac{\partial f_n}{\partial x_n}\end{pmatrix}$ - матриця Якобі. Далі маємо

$$\frac{\partial U}{\partial x_j}=\frac{\partial}{\partial x_j}\left\{\sum_{i=1}^{n}\left[f_i(\vec{x})\right]^2\right\}=2\sum_{i=1}^{n}f_i(\vec{x})\frac{\partial f_i(\vec{x})}{\partial x_j}.$$

Звідси $\nabla U(\vec{x})=2\begin{bmatrix}\sum_{i=1}^{n}\frac{\partial f_i(\vec{x})}{\partial x_1}f_i(\vec{x})\\ \dots\dots\\ \sum_{i=1}^{n}\frac{\partial f_i(\vec{x})}{\partial x_n}f_i(\vec{x})\end{bmatrix}=2W'(\vec{x})\vec{f}(\vec{x}),$

де W`(x) - транспонована матриця Якобі. Тому остаточно

$\mu_p=2\lambda_p=\frac{\left(\vec{f}^{(p)},W_pW_p'\vec{f}^{(p)}\right)}{\left(W_pW_p'\vec{f}^{(p)},W_pW_p'\vec{f}^{(p)}\right)}$, а $\vec{x}^{(p+1)}=\vec{x}^{(p)}-\mu_pW_p'\vec{f}^{(p)}$.

Отримали розрахункову формулу методу градієнтного спуску з визначенням кроку.

Сучасна комп'ютерна техніка дозволяє суттєво спростити цей метод розв'язання нелінійних систем. Множник λ_p у формулі (4.8) обирають як достатньо малий постійний крок у напрямку антиградієнта. Наприклад, $\lambda_p=0.00001$.

Приклад. Розв'язати систему нелінійних рівнянь з точністю ε=0,0001

$$\begin{cases}2x-y^2+z=1\\ 3x^2-2y+3z=1{,}75\\ x^2+y+z^2=2{,}25.\end{cases}$$

Скористаємося методом градієнтного спуску. Для цього побудуємо функцію $U(x)=\sum_{i=1}^{n}(f_i(\bar{x}))^2=\left(\bar{f}(\bar{x}),\bar{f}(\bar{x})\right)$,

де $\bar{f}(\bar{x})=\begin{pmatrix}2x-y^2+z-1\\ 3x^2-2y+3z-1{,}75\\ x^2+y+z^2-2{,}25\end{pmatrix}$, а $\bar{x}=\begin{pmatrix}x\\ y\\ z\end{pmatrix}$.

Кожен розв'язок системи - це нуль функції $U(\bar{x})$ і навпаки. Виберемо початкове наближення $\bar{x}^0=(0{,}5;\ 0{,}5;\ 0{,}5)$. Позначимо $\bar{\nabla}U=\left(\frac{\partial U}{\partial x};\frac{\partial U}{\partial y};\frac{\partial U}{\partial z}\right)$. Пошук розв'язку проводимо за формулою

$$\bar{x}^{(p+1)}=\bar{x}^{(p)}-\lambda^{(p)}\bar{\nabla}U\left(\bar{x}^{(p)}\right).$$

Множник λ(p) визначається так:

$\lambda^{(p)}=\frac{\left(\bar{f}\left(\bar{x}^{(p)}\right),W\left(\bar{x}^{(p)}\right)\times\bar{\nabla}U\left(\bar{x}^{(p)}\right)\right)}{\left(W\left(\bar{x}^{(p)}\right)\times\bar{\nabla}U\left(\bar{x}^{(p)}\right),W\left(\bar{x}^{(p)}\right)\times\bar{\nabla}U\left(\bar{x}^{(p)}\right)\right)}$, де:

$$\bar{\nabla}U=2\begin{pmatrix}\sum_{i=1}^{n}f_i(\bar{x})\frac{\partial f_i(\bar{x})}{\partial x_1}\\ \cdots\cdots\\ \sum_{i=1}^{n}f_i(\bar{x})\frac{\partial f_i(\bar{x})}{\partial x_n}\end{pmatrix}=2\cdot W(\bar{x})\bar{f}(\bar{x}).$$

Тоді процес здійснюється за формулою

$\bar{x}^{(p+1)}=\bar{x}^{(p)}-2\lambda^{(p)}W\left(\bar{x}^{(p)}\right)\bar{f}\left(\bar{x}^{(p)}\right)$;

W – матриця Якобі: $W = \begin{pmatrix} 2 & -2y & 1 \\ 6x & -2 & 3 \\ 2x & 1 & 2z \end{pmatrix}$.

Пошук наближень до розв'язку припиняється за умови $\left\| x^{(p+1)} - x^{(p)} \right\| < \varepsilon$.

Реалізація алгоритму на псевдокоді:

```
VectorF(F,x):
//розраховуємо коефіцієнти вектора F при x
end
VectorW(W,x):
//розраховуємо коефіцієнти матриці W при x
end
//W – матриця Якобі
//F – матриця системи
//Z – результат множення
multWF(W,F,Z): // множення W на F
        1   for i:=1 to W.lengthI do
        2     for j:=1 to W.lengthJ do
        3       Z[i]+=W[i,j]*F[j]
        4     done
        5   done
end
//A – вихідна матриця
//X – транспонована матриця
Transpon(A,X):
```



```
        1   for i:=1 to A.lengthI do
        2     for j:=1 to A.lengthJ do
        3       if i=j then
        4         X[i,j]:=A[i,j]
        5       else
        6         X[i,j]:=A[j,i]
        7       fi
        8     done
        9   done
end
Scalar(A,B): //скалярний добуток векторів a та b
        1   s:=0
        2   for i:=1 to A.length do
        3     s+=A[i]*B[i]
        4   done
        5   return s
end
// розв'язання нелінійної системи методом градієнтного спуску
Solve_NonLinear_System(F,W,X):
        1   n:=A.lengthI
        2   for i:=1 to n do
        3     X[i]:=-1;
        4   done
```

```
5     k:=0
6     repeat
7       k++
8       for i:=1 to n do
9         XK[i]:=X[i]
10      done
11    Vectorf(F,xk)
12    MatrixW(W,xk)
13    MultWF(wt,f,u)
14    MultWF(w,u,z)
15    mu:=(scalar(f,z))/(scalar(z,z))
16    maxr:=0
17    for i:=1 to n do
18      X[i]:=XK[i]-mu*U[i]
19      if abs(X[i]-XK[i])>maxr then
20        maxr:=abs(X[i]-XK[i])
21      fi
22    until maxr<eps
23    done
end.
```

Відповідь: X=0.50, Y=1.00, Z=1.00.

4.5 Метод релаксацій

Перепишемо систему (4.1) у вигляді

$$\vec{x} = \vec{x} + \tau\ \vec{f}(\vec{x}),$$

де τ - деяка константа, і побудуємо ітераційний процес за схемою

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} + \tau\ \vec{f}(\vec{x}^{(k)}).$$

Параметр τ повинен бути таким, щоб в околі розв'язку виконувалася достатня умова збіжності

$$\|\mathbf{E} + \tau\ \mathbf{W}\| < 1,$$

де **E**- одинична матриця, а **W** –матриця Якобі. На практиці виконання цієї умови досить складно перевірити, тому значення параметра τ вибирають пробним шляхом, перевіряючи виконання необхідної умови збіжності після здійснення кожної ітерації

$$\|\vec{x}^{(k)} - \vec{x}^{(k-1)}\| < \|\vec{x}^{(k-1)} - x^{(k-2)}\|.$$

Якщо виявиться, що на якій-небудь ітерації ця умова не виконується, то необхідно змінити значення параметра τ.

Приклад. Знайти з точністю $\varepsilon = 10^{-6}$ всі корені системи нелінійних рівнянь $\begin{cases} f_1(x_1, x_2) = 0, \\ f_2(x_1, x_2) = 0, \end{cases}$ використовуючи метод Ньютона для системи нелінійних рівнянь. Знайти корінь за допомогою убудованого блоку розв'язку рівнянь **Given Find** пакета MATHCAD. Рівняння системи:

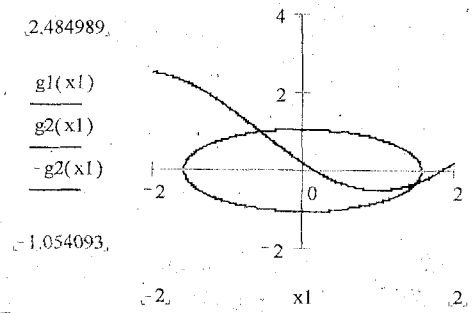
$$f1(x1, x2) := x2 + 1.5 \cdot \cos(x1 - 1) - 1$$

$$f2(x1, x2) := 0.9 \cdot x2^2 + 0.4 \cdot x1^2 - 1$$

Локалізація кореня

Перше рівняння, визначене відносно *x2:* $g1(x1) := 1 - 1.5 \cdot \cos(x1 - 1)$. Друге рівняння, визначене відносно *x2:* $g2(x1) := \sqrt{\frac{1 - 0.4 \cdot x1^2}{0.9}}$.

Маємо $x1 := -2, -2 + 0.01 \ldots 2$.



Перший корінь

Початкове наближення: $x1 := 1.7 \quad x2 := -0.5$.

Точність для блоку **Given Find**: $TOL := 10^{-6}$.

Розв'язання системи f(x1,x2)=0 за допомогою убудованого блоку MATHCAD:

Given

$$f1(x1, x2) = 0$$

$$f2(x1, x2) = 0$$

$$xr1 := \text{Find}((x1, x2))$$

Отриманий наближений розв'язок $xr1 = \begin{bmatrix} 1.5124471 \\ -0.3073209 \end{bmatrix}$.

Питання і завдання до розділу 4

1 Постановка задачі розв'язання системи нелінійних рівнянь. Основні етапи розв'язування задачі.

2 Метод простої ітерації: опис методу, умова й швидкість збіжності, критерій закінчення, приведення до вигляду, зручного для ітерацій.

3 Метод Ньютона: опис методу, теорема про збіжність, критерій закінчення.

4 Недоліки методу Ньютона. Модифікації методу Ньютона.

5 Застосування методів розв'язання систем нелінійних рівнянь для задачі мінімізації функцій.

6 Розв'язати методом Ньютона з точністю $\varepsilon = 0.05$ системи рівнянь:

a) $\begin{cases} x^3 - y^2 - 1 = 0 \\ xy^3 - y - 4 = 0 \end{cases}, \quad \begin{pmatrix} x^{(0)} \\ y^{(0)} \end{pmatrix} = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix};$

b) $\begin{cases} x^3 - y^3 + 0.1 = 0 \\ xy - 0.95 = 0 \end{cases}, \quad \begin{pmatrix} x^{(0)} \\ y^{(0)} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$

7 Чи можна стверджувати, що система має, й до того ж єдиний, розв'язок?

$$\begin{cases} x = 0.1\sin(x) + 0.3\cos(y) - 0.4 \\ y = 0.2\cos(x) - 0.1\sin(y) - 0.3 \end{cases}$$

8 Для системи рівнянь виписати розрахункові формули методу релаксацій:

$$\begin{cases} x - 6y + 4z + 3xy - 8 = 0 \\ 2x + 3y - z + 5yz - 24 = 0 \\ -x + 4y + z + 8xz - 21 = 0 \end{cases}$$

9 Розв'язати методом простої ітерації такі системи:

a) $\begin{cases} x = \lg\dfrac{y}{z} + 1 \\ y = 0.4 + z^2 - 2x^2 \\ z = 2 + \dfrac{xy}{20} \end{cases}, \quad \begin{pmatrix} x^{(0)} \\ y^{(0)} \\ z^{(0)} \end{pmatrix} = \begin{pmatrix} 1 \\ 2.2 \\ 2 \end{pmatrix};$

b) $\begin{cases} x + x^2 - 2yz = 0.1 \\ y - y^2 + 3xz = -0.2 \\ z + z^2 + 2xy = 0.3 \end{cases}, \quad \begin{pmatrix} x^{(0)} \\ y^{(0)} \\ z^{(0)} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$

10 Для функції $f(x_1, x_2) = x_1^4 + x_2^4 - x_1^2 - x_2^2$ знайти точки мінімуму, звівши задачу до розв'язання системи рівнянь.

Розділ 5
Апроксимація функцій

Апроксимація (від лат. approximo - наближаюся) - заміна одних математичних об'єктів іншими, якомось чином близькими до вихідних. Апроксимація дозволяє досліджувати числові характеристики і якісні властивості об'єкта, зводячи задачу до вивчення більш простих або більш зручних об'єктів (наприклад таких, характеристики яких легко обчислюються або властивості яких уже відомі). У теорії чисел вивчаються діофантові наближення, зокрема наближення ірраціональних чисел раціональними. У геометрії і топології розглядаються апроксимації кривих, поверхонь, просторів і відображень. Деякі розділи математики цілком присвячені апроксимації, наприклад *наближення функцій.*

5.1 Поняття про наближення функцій

Нехай величина y є функцією аргумента x. Це означає, що будь-якому значенню x з області визначення поставлено у відповідність значення y. Разом з тим на практиці часто невідомий явний зв'язок між y та x, тобто неможливо записати цей зв'язок у вигляді деякої залежності $y=f(x)$. У деяких випадках навіть при відомій залежності $y=f(x)$ вона настільки громіздка (наприклад, містить вирази, що важко обчислюються, складні інтеграли і т.п.), що її використовувати в практичних розрахунках важко.

Найбільш поширеним і практично важливим випадком, коли вигляд зв'язку між параметрами x та y невідомий, є його завдання у вигляді деякої таблиці $\{x_i, y_i\}$. Це означає, що дискретній множині значень аргумента $\{x_i\}$ поставлена у відповідність множина значень функції $\{y_i\}$

$(i=0,1,\ldots,n)$. Ці значення - або результати розрахунків, або експериментальні дані. На практиці нам можуть знадобитися значення величини y і в інших точках поза вузлами x_i. Однак одержати ці значення можна лише шляхом дуже складних розрахунків або проведенням дорогих експериментів.

Таким чином, з огляду економії часу і засобів ми приходимо до необхідності використання наявних табличних даних для наближеного обчислення невідомого параметра y при будь-якому значенні (з деякої області) визначального параметра x, оскільки точний зв'язок $y = f(x)$ - невідомий.

Цій меті підпорядкована задача про наближення (апроксимацію) функцій: задану функцію $f(x)$ потрібно приблизно замінити (апроксимувати) деякою функцією $F(x)$ так, щоб відхилення (у деякому змісті) $F(x)$ від $f(x)$ у заданій області було найменшим. Функція $F(x)$ при цьому називається апроксимуючою.

Апроксимуючими функціями можуть бути поліноміальні, тригонометричні, експонентні та ін.

Якщо наближення будується на заданій дискретній множині точок $\{x_i\}$, то апроксимація називається ***точковою.*** До неї належать інтерполяція, середньоквадратичне наближення та ін. При побудові наближення на неперервній множині точок (наприклад, на відрізку $[a,b]$) апроксимація називається ***неперервною*** (або інтегральною).

Одним із основних типів точкової апроксимації є інтерполяція. У цьому випадку апроксимуюча функція проходить через задані вузлові точки. Іноді наближення табличних даних методом інтерполяції проводити незручно. Так, наприклад, якщо дані в таблиці неточні, то збіг значень інтерполяційної функції у вузлах з табличними даними означає, що вона точно повторює

помилки таблиці. У таких випадках використовують інші види апроксимації, наприклад, метод найменших квадратів. Цим методом апроксимуюча функція будується так, щоб сума квадратів відстаней від ординат точок до лінії графіка апроксимуючої функції для однакових абсцис була найменшою.

5.2 Інтерполювання функції

Загальна постановка задачі інтерполювання така. Задані значення $y_1, y_2, ..., y_n$ функції аргумента x при відповідних його значеннях $x_1, x_2, ..., x_n$ $(x_i < x_{i+1})$. Побудувати неперервну функцію $F(x)$, що належить до заданого класу функцій, таку, що вона збігається з y_i при значеннях аргумента x_i $(i = 1,2,...n)$. Така функція називається ***інтерполюючою***. Точки x_i, $i=1,...,n$ називаються вузлами інтерполяції і вони утворюють сітку розбиття ω_n, а y_i - вузловими значеннями.

У такому формулюванні розв'язок задачі є невизначеним, бо крізь задані точки можна провести безліч кривих. Тому загальну постановку дещо звужують, задаючи не тільки клас інтерполюючої функції, але й додаткову умову мінімальної її складності.

Наприклад, для найбільш поширеного поліноміального інтерполювання (при якому інтерполююча функція обирається серед поліномів аргумента x), додатковою умовою є мінімальний порядок інтерполюючого полінома. З цього випливає, що якщо первісну функцію задано лише двома точками, її треба інтерполювати поліномом першого порядку (через дві задані точки проходить єдина пряма), якщо трьома - параболою другого порядку і так далі. Взагалі функція,

задана своїми n значеннями (у n точках), інтерполюється однозначно поліномом $(n-1)$-го порядку, тобто таким

$$F(x) = a_0 + a_1 \cdot x + a_2 \cdot x^2 + ... + a_{n-1} \cdot x^{n-1}. \quad (5.1)$$

Тепер задача інтерполювання звелася до пошуку значень n невідомих коефіцієнтів полінома (5.1) з умови набуття ним значень y_i при значеннях аргумента x_i $(i = 1,2,...n)$.

Існують кілька способів визначення цих коефіцієнтів. Вони відрізняються методикою обчислень, зручною в одних і незручною в інших випадках. Але при ідеальних обчисленнях вони, природно, призводять до тих самих результатів, тобто до того самого полінома.

5.2.1 Інтерполювання за Лагранжем

За цією методикою попередньо визначають допоміжні поліноми $(n-1)$-го порядку $L_{n-1}(x, x_i)$ такі, що

$$L_{n-1}(x_k, x_i) = \begin{cases} 0, & k \neq i \\ 1, & k = i \end{cases}. \quad (5.2)$$

Тобто кожен із них набуває значення 1 тільки при $x = x_i$, а для решти заданих значень аргумента він дорівнює нулю. Такі поліноми одержали назву лагранжевих коефіцієнтів, або множників впливу відповідних вузлів інтерполювання.

Щоб виконувалася перша умова (5.2), поліном $L_{n-1}(x, x_i)$ повинен мати такий вигляд

$$L_{n-1}(x, x_i) = A(x - x_1)(x - x_2)...(x - x_{i-1})(x - x_{i+1})...(x - x_n), \quad (5.3)$$

тобто добутку $(n-1)$ різниць між поточним значенням аргумента й одним із заданих, окрім i-го, з деяким коефіцієнтом A. Друга умова (5.2) дозволяє визначити цей

коефіцієнт A. Для цього в (5.3) слід покласти $x = x_i$ і прирівняти результат до одиниці. З цього матимемо

$$A = \frac{1}{(x_i - x_1)(x_i - x_2)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)}.$$

Враховуючи це, одержимо остаточний вигляд допоміжного полінома

$$L_{n-1}(x, x_i) = \frac{(x - x_1)(x - x_2)...(x - x_{i-1})(x - x_{i+1})...(x - x_n)}{(x_i - x_1)(x_i - x_2)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)}. \quad (5.4)$$

Тепер шуканий інтерполюючий поліном можна подати у вигляді

$$F(x) = \sum_{i=1}^{n} L_{n-1}(x, x_i) \cdot y_i = L_{n-1}(x, x_1) \cdot y_1 + L_{n-1}(x, x_2) \cdot y_2 + ... + L_{n-1}(x, x_n) \cdot y_n =$$

$$= \frac{(x - x_2)(x - x_3)...(x - x_n)}{(x_1 - x_2)(x_1 - x_3)...(x_1 - x_n)} y_1 + \frac{(x - x_1)(x - x_3)...(x - x_n)}{(x_2 - x_1)(x_2 - x_3)...(x_2 - x_n)} y_2 +$$

$$+ ... + \frac{(x - x_1)(x - x_2)...(x - x_{n-1})}{(x_n - x_1)(x_n - x_2)...(x_n - x_{n-1})} y_n. \quad (5.5)$$

Це і є інтерполяційний поліном Лагранжа степеня $n-1$. Кількість арифметичних операцій для його обчислення дорівнює n^2. Інтерполювання за Лагранжем зручно використовувати тоді, коли ведеться багаторазове інтерполювання різних функцій за однакових значень масиву аргументів. Тоді можна заздалегідь одноразово обчислити коефіцієнти Лагранжа, оскільки вони не залежать від функції, що інтерполюється.

Розглянемо деякі часткові випадки.

1 Лінійна інтерполяція

У цьому разі маємо $n = 2$ два вузли інтерполяції. Інтерполяційний поліном Лагранжа має вигляд:

$$F(x) = L_1(x, x_1) \cdot y_1 + L_1(x, x_2) \cdot y_2 = \frac{x - x_2}{x_1 - x_2} y_1 + \frac{x - x_1}{x_2 - x_1} y_2 = \quad (5.6)$$

$$= \frac{1}{x_2 - x_1}[(y_1 x_2 - y_2 x_1) + (y_1 - y_2)x].$$

2 Квадратична інтерполяція

У цьому випадку є три вузли інтерполяції ($n = 3$). Інтерполяційний поліном Лагранжа набирає вигляду:

$$F(x) = L_2(x, x_1) \cdot y_1 + L_2(x, x_2) \cdot x_2 + L_2(t, t_3) \cdot x_3 = \quad (5.7)$$

$$= \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} y_1 + \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} y_2 + \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)} y_3$$

5.2.2. Інтерполювання за Ньютоном

Недоліком інтерполювання за Лагранжем є те, що якщо для поліпшення наближення додати ще один вузол інтерполювання, доведеться всі обчислення проводити заново.

На практиці часто трапляються випадки, коли вузли інтерполяції стають відомими не одразу, а поступово, один за одним, наприклад, у процесі вимірювання. Тоді зручно побудувати процес інтерполювання у такий спосіб, щоб поява даних про новий вузол інтерполювання, призводила б до необхідності мінімального перерахунку попередніх обчислень. Саме таку властивість має інтерполювання за Ньютоном.

Нехай вузли інтерполяції рівновіддалені один від одного за аргументом, тобто виконується умова

$$x_i - x_{i-1} = h; \qquad (i = 2,3,...n). \quad (5.8)$$

Різниці

$$\Delta y_i = y_{i+1} - y_i \quad (5.9)$$

називають ***скінченними різницями першого порядку***.

Різниці сусідніх скінченних різниць першого порядку

$$\Delta^2 y_i = \Delta y_{i+1} - \Delta y_i = y_{i+2} - 2y_{i+1} + y_i \quad (5.10)$$

називають ***скінченними різницями другого порядку***.

Аналогічно

$$\Delta^k y_i = \Delta^{k-1} y_{i+1} - \Delta^{k-1} y_i \qquad (5.11)$$

є скінченними різницями k-го порядку. Вони визначаються за формулою $\Delta^k y_0 = \sum_{i=0}^{k} (-1)^i c_k^i y_{k-i}$, де c_k^i - біноміальні коефіцієнти.

Розглянемо поліном

$$F(x) = q_0 + q_1(x - x_1) + q_2(x - x_1)(x - x_2) + \\ + \ldots + q_{n-1}(x - x_1)(x - x_2)\ldots(x - x_{n-1}) \qquad (5.12)$$

Визначимо його коефіцієнти. Коефіцієнт q_0 визначимо з умови проходження полінома через першу точку (x_1, y_1)

$$q_0 = F(x_1) = y_1. \qquad (5.13)$$

З умови проходження полінома через точку (x_2, y_2) одержимо значення q_1

$$q_0 + q_1(x_2 - x_1) = F(x_2) = y_2; \Rightarrow q_1 = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y_1}{h}. \qquad (5.14)$$

Аналогічно визначається решта коефіцієнтів

$$q_i = \frac{\Delta^i y_1}{i! h^i}. \qquad (5.15)$$

Підставляючи отримані вирази у (5.12), одержуємо

$$F(x) = y_1 + \sum_{k=1}^{n-1} \frac{\Delta^k y_1}{k! h^k} (x - x_1)\ldots(x - x_k). \qquad (5.16)$$

Це є ***перша інтерполяційна формула Ньютона*** (формула інтерполювання вперед).

Як бачимо, особливостями інтерполювання за Ньютоном є:

- при появі нового вузла додається лише новий член, решта не перераховується;
- коефіцієнти швидко зменшуються зі зростанням k, бо у знаменнику міститься факторіал від k.

Іноді використовується ***формула для інтерполювання назад***

$$F(x) = y_n + \sum_{k=1}^{n-1} \frac{\Delta^k y_{n-k}}{k! h^k} (x - x_n)\ldots(x - x_{n-k}).$$

Візьмемо деяку функцію $f(x) \in R$ і систему вузлів інтерполяції $x_0, x_1, x_2, \ldots, x_n$, $x_i \neq x_j$, $i, j = \overline{1,n}$ при $i \neq j$. Вузли інтерполяції не є рівновіддаленими. Для цієї функції і вузлів утворимо відношення

$$\left.\begin{aligned} &\frac{f(x_1) - f(x_0)}{x_1 - x_0} = f(x_0, x_1), \\ &\frac{f(x_2) - f(x_1)}{x_2 - x_1} = f(x_1, x_2), \ldots, \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} = f(x_{n-1}, x_n) \end{aligned}\right\}.$$

Вони називаються ***розділеними різницями першого порядку***. Одержавши їх, ми можемо утворити нові відношення

$$\left.\begin{aligned} &\frac{f(x_1, x_2) - f(x_0, x_1)}{x_2 - x_0} = f(x_0, x_1, x_2) \\ &\frac{f(x_2, x_3) - f(x_1, x_2)}{x_3 - x_1} = f(x_1, x_2, x_3), \ldots, \\ &\frac{f(x_{n-1}, x_n) - f(x_{n-2}, x_{n-1})}{x_n - x_{n-2}} = f(x_{n-2}, x_{n-1}, x_n). \end{aligned}\right\}$$

Вони називаються ***розділеними різницями другого порядку***. Взагалі, якщо ми уже визначили розділені різниці k-го порядку $f(x_i, x_{i+1}, \ldots, x_{i+k})$, то ***розділені різниці (k+1)-го порядку*** знаходяться за допомогою формули

$$\frac{f(x_i, x_{i+1}, \ldots, x_{i+k}) - f(x_{i-1}, x_i, \ldots, x_{i+k-1})}{x_{i+k} - x_{i-1}} = f(x_{i-1}, x_i, \ldots, x_{i+k}).$$

Іноді замість $f(x_i, x_{i+1}, \ldots, x_{i+k})$ для позначення розділених різниць використовують позначення $[x_i, x_{i+1}, \ldots, x_{i+k}]$.

Домовимося розміщувати таблиці розділених різниць у такий спосіб:

$f(x_0)$			
	$f(x_0;x_1)$		
$f(x_1)$		$f(x_0;x_1;x_2)$	
	$f(x_1;x_2)$		$f(x_0;x_1;x_2;x_3)$
$f(x_2)$		$f(x_1;x_2;x_3)$	
	$f(x_2;x_3)$		$f(x_1;x_2,x_3;x_4)$
$f(x_3)$		$f(x_2;x_3;x_4)$	
	$f(x_3;x_4)$		
$f(x_4)$			

При $x_i = x_0 + ih \quad (i = \overline{0,n})$ скінченні і розділені різниці пов'язані співвідношенням у вигляді

$$f(x_0, x_1, ..., x_n) = \frac{\Delta^n y_0}{n! h^n}.$$

Розділені різниці порядку n від многочлена n-го степеня постійні, а різниці більш високого порядку дорівнюють нулю. Останнім зауваженням можна скористатися для виявлення помилок у таблицях многочленів чи функцій, близьких до них.

За допомогою розділених різниць можна побудувати інтерполяційний многочлен Ньютона

$$P_n(x) = f(x_0) + (x - x_0) \cdot f(x_0, x_1) + (x - x_0)(x - x_1) \cdot f(x_0, x_1, x_2) + ...$$
$$+ (x - x_0)(x - x_1) \ ... \ (x - x_{n-1}) \cdot f(x_0, x_1, \ ... \ , x_n).$$

Варто зазначити, що при збільшенні кількості вузлів процес обчислення скінченних та поділених різниць стає все більш обчислювально нестійким - похибка визначення скінченних різниць великого порядку різко зростає зі збільшенням порядку скінченної різниці. Тому метод

Ньютона може бути застосований лише для невеликої кількості вузлів.

5.2.3 Інтерполювання за Ермітом

У більш загальному випадку потрібно, щоб у вузлах інтерполяції збігалися не лише значення інтерполюючої функції і функції, яку необхідно інтерполювати, але й значення їхніх похідних до деякого порядку. У цьому випадку застосовують інтерполювання за Ермітом.

Інтерполяційним поліномом Ерміта s-го порядку називають поліном $H_s(x)$ аргумента x, який визначається з умов

$$H_s(x_1) = y_1; \frac{dH_s}{dx}(x_1) = y_1'; \frac{d^{(\alpha_1-1)}H_s}{dx^{(\alpha_1-1)}}(x_1) = y_1^{(\alpha_1-1)};$$
$$H_s(x_2) = y_2; \frac{dH_s}{dx}(x_2) = y_2'; . \frac{d^{(\alpha_2-1)}H_s}{dx^{(\alpha_2-1)}}(x_2) = y_2^{(\alpha_2-1)};$$
$$. \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \qquad (5.17)$$
$$H_s(x_k) = y_k; \frac{dH_s}{dx}(x_k) = y_k'; . \frac{d^{(\alpha_k-1)}H_s}{dx^{(\alpha_k-1)}}(x_k) = y_k^{(\alpha_k-1)};$$
$$. \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ .$$
$$H_s(x_n) = y_n; \frac{dH_s}{dx}(x_n) = y_n'; \frac{d^{(\alpha_n-1)}H_s}{dx^{(\alpha_n-1)}}(x_n) = y_n^{(\alpha_n-1)}.$$

Тут, як і раніше, n - кількість вузлів інтерполяції.

Якщо у вузлі $x_k (k = 1,...,n)$ поліном і функція, яка інтерполюється, збігаються до похідної порядку $\alpha_k - 1$, то число α_k називається кратністю вузла x_k. При цьому $\alpha_1 + \alpha_2 + ... + \alpha_k + ... + \alpha_n = s + 1$.

Інтерполяційний поліном Ньютона (5.12) узагальнюється на випадок кратних вузлів таким чином:

$$F(x) = q_0 + q_1(x-x_1) + q_2(x-x_1)^2 + \dots + q_{\alpha_1}(x-x_1)^{\alpha_1} +$$
$$+ q_{\alpha_1+1}(x-x_1)^{\alpha_1}(x-x_2) + \dots + q_{\alpha_1+\alpha_2}(x-x_1)^{\alpha_1}(x-x_2)^{\alpha_2} + \quad (5.18)$$
$$+ q_{\alpha_1+\alpha_2+1}(x-x_1)^{\alpha_1}(x-x_2)^{\alpha_2}(x-x_3) + \dots +$$
$$\dots + q_{\alpha_1+\alpha_2+\alpha_3}(x-x_1)^{\alpha_1}(x-x_2)^{\alpha_2}(x-x_3)^{\alpha_3} +$$
$$+ \dots + q_s(x-x_1)^{\alpha_1}\dots(x-x_n)^{\alpha_n-1}.$$

Інтерполювання за Ермітом зводиться до визначення $s+1$ коефіцієнтів q_0, q_1, ..., q_s з умов (5.17).

5.2.4 Похибка інтерполяції та способи її зменшення

Зафіксуємо точку x та визначимо похибку інтерполяції $r_n(x)=f(x)-P_n(x)$. Нехай $f(x) \in C^{n+1}[a;b]$, $x_i \in [a,b]$ та введемо функцію $g(s)=f(s)-P_n(s)-kw(s)$, де $w(s)=(s-x_0)(s-x_1)\dots(s-x_n)$. При $s=x_i$, $i=0,1,\dots,n$, $w(x_i)=0$. Тому $g(x_i)=0$, бо $f(x_i)=P_n(x_i)$. Виберемо деяку точку $x \neq x_i$, $i = 0,1,\dots,n$ та виберемо коефіцієнт k так, щоб $g(x)=0$. Тоді $f(x)-P_n(x)-kw(x)=0$; $k=(f(x)-P_n(x))/w(x)$.

Враховуючи, що $w^{(n+1)}(s) = (n+1)!$ та те, що $g(s)$ має $n+2$ нулі на $[a;b]$, то $g'(s)$ має $n+1$ нуль, $g''(s)$ має n нулів, …, $g^{(n+1)}(s)$ має принаймні один нуль. Нехай це буде при $s=\xi$. Тоді

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - 0 - \frac{f(x)-P_n(x)}{\omega(x)}(n+1)!.$$

Звідси отримуємо оцінку для похибки інтерполювання

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi)\omega(x)}{(n+1)!}.$$

Тоді оцінка для абсолютної похибки поліноміальної інтерполяційної формули має вигляд

$$|f(x) - P_n(x)| \le \frac{\sup_{[a,b]}|f^{(n+1)}(x)|}{(n+1)!}|\omega(x)|. \quad (5.19)$$

Як бачимо з (5.19), похибка заміни функції $y=f(x)$ інтерполяційним многочленом залежить від вибору вузлів інтерполяції $x_0, x_1, x_2, \dots, x_n$. Перш ніж перейти до питання про раціональний вибір вузлів інтерполяції, розглянемо деякі властивості одного з найважливіших і добре вивчених зараз класів спеціальних функцій – многочленів Чебишева першого роду, що часто використовуються для наближення функцій. Многочлен Чебишева $n-$ го степеня визначається за формулою

$$T_n(x) = \frac{2^n n!}{(2n)!}\sqrt{x^2-1}\frac{d^n}{dx^n}((x^2-1)^{n-\frac{1}{2}}). \quad (5.20)$$

Для визначення многочленів Чебишева часто користуються тригонометричною формою запису

$$T_n(x) = \cos(n \arccos x), |x| \le 1, \quad (5.21)$$

що приводить до таких самих виразів для $T_n(x)$, як і в (5.20).

Із тотожності $\cos(n+1)\theta = 2\cos\theta\cos n\theta - \cos(n-1)\theta$ при $\theta = \arccos x$ маємо рекурентну формулу

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

Многочлен $T_n(x)$ має n коренів, які можна отримати, розв'язавши рівняння $\cos(n\arccos x) = 0$, або $n\arccos x = \frac{\pi}{2}(2m+1)$;

$$x = \cos\frac{(2m+1)\pi}{2n}, m = 0,1,\dots,n-1. \quad (5.22)$$

Як бачимо з (5.22), всі n коренів, що відповідають значенням $m = 0,1,\dots,n-1$, знаходяться на відрізку [-1,1], причому ці точки не рівновіддалені, а згущуються ближче

до кінця даного відрізка. З формули (5.21) також очевидно, що на відрізку [-1,1]

$$\max|T_n(x)| = 1. \qquad (5.23)$$

Доведено, що серед усіх можливих n значень x на відрізку $[-1,1]$ корені $x_0^{(T)}, x_1^{(T)}, ..., x_{n-1}^{(T)}$ многочлена $T_n(x)$ мають ту чудову властивість, що для них величина

$$\omega_n(x) = (x-x_0)(x-x_1)...(x-x_{n-1}) = \frac{1}{2^{n-1}} T_n(x) \qquad (5.24)$$

має найменше за модулем максимальне значення.

Беручи до уваги (5.24), запишемо

$$\max|\omega_n(x)| = \frac{1}{2^{n-1}}. \qquad (5.25)$$

Виходячи з властивостей коренів многочленів Чебишева першого роду і визначення інтерполяційного многочлена $P_n(x)$ n-го степеня на відрізку $[-1,1]$, можна стверджувати, що якщо за n вузлів інтерполювання взяти корені многочлена $T_n(x)$, то максимальне значення похибки на цьому відрізку буде найменшим для всіх можливих варіантів вибору n вузлів інтерполювання. Інтерполяційний многочлен, наділений такою властивістю, називається многочленом найкращого наближення. Оцінка (5.19) при цьому набирає вигляду

$$|f(x) - P_n(x)| \le \frac{M_{n+1}}{2^n (n+1)!}, \text{ де } M_{n+1} = \sup_{[a,b]} |f^{(n+1)}(x)|.$$

Якщо інтерполювання проводиться на довільному відрізку $[a,b]$, то заміною змінної

$$x = \frac{1}{2}((b-a)z + (b+a)), z = \frac{1}{b-a}(2x-b-a)$$

цей відрізок можна звести до відрізка $[-1,1]$. При цьому корені многочлена $T_n(x)$ будуть знаходитися в точках

$$x_m = \frac{1}{2}(b-a)\cos\frac{2m+1}{2n}\pi + (b+a).$$

Оцінка похибки має вигляд

$$|f(x) - L_n(x)| \le \frac{M_{n+1}}{2^{2n+1}(n+1)!}(b-a)^{n+1}.$$

5.2.5 Збіжність процесу інтерполяції

Розглянемо послідовність сіток

$$w_n: a=x_0<x_1<...<x_{n-1}<x_n=b.$$

Кажуть, що інтерполяційний поліном $L_n(x)$ рівномірно збігається до заданої функції $f(x)$, якщо при $max\ (x - x_{n-1}) \to 0$ $\max_{x\in[a,b]} |L_n(x) - f(x)| \to 0$. Справедливі такі теореми.

Теорема Фабера. Для будь-якої послідовності сіток w_n знайдеться $f(x) \in C^1[a;b]$ така, що збіжність відсутня.

Теорема Марцинкевича. Для будь-якої функції $f(x) \in C^1[a;b]$ знайдеться послідовність сіток $\omega_n : L_n \to f$.

Приклад. Використовуючи інтерполяційний поліном Ньютона, визначити *f(0.14)*, де *y=f(x)* задана таблично.

x	0	0.1	0.2	0.3	0.4	0.5
y	0	0.1002	0.2013	0.8045	0.4108	0.5211

Розв'язання. Складаємо таблицю скінченних різниць, користуючись пакетом Excel:

	A	B	C	D	E	F	G
1	x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$
2	0	0	=B3-B2	=C3-C2	=D3-D2	=E3-E2	=F3-F2
3	0,1	0,1002	=B4-B3	=C4-C3	=D4-D3	=E4-E3	
4	0,2	0,2013	=B5-B4	=C5-C4	=D5-D4		
5	0,3	0,3045	=B6-B5	=C6-C5			
6	0,4	0,4108	=B7-B6				
7	0,5	0,5211					

У результаті отримаємо таке:

	A	B	C	D	E	F	G
1	x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$
2	0	0	0,1002	0,0009	0,0012	-0,0002	0,0001
3	0,1	0,1002	0,1011	0,0021	0,0010	-0,0001	
4	0,2	0,2013	0,1032	0,0031	0,0009		
5	0,3	0,3045	0,1063	0,0040			
6	0,4	0,4108	0,1103				
7	0,5	0,5211					

Для розрахунку *f(0.14)* скористаємося інтерполяційним поліномом Ньютона, покладаючи, що $x_0=0.1$ та h=0.1; тоді $q=(x-x_0)/h$=(0,14-0,1)/0,1=0,4. Звідси за формулою визначаємо:

$f(0.14)\approx$0,1002+0,1011*0,4+0,0021*0,4*(-0,6)/2+
+0,1010*0,4*(-0,6)*(-1,6)/6 ≈ 0,1405, або у MS Excel у даному випадку, формула матиме такий вигляд:
=B3+Q*C3+Q*(Q-1)*D3/ФАКТР(2)+Q*(Q-1)*(Q-2)*
*E3/ФАКТР(3), де Q - адреса комірки із розрахованим значенням q.

При цьому похибка наближення дорівнює $R4=|f(0.14)-P_3(0.14)|<0.0001*0.4*0.6*1.6*2.6/4!=$ 4.16*10-6, або у MS Excel =ABS(F3*Q*(Q-1)*(Q-2)*(Q-3))/ФАКТР(4).

5.2.6 Інтерполяція за допомогою сплайнів

Підвищення точності інтерполювання вимагає збільшення вузлів інтерполяції. Це призведе до зростання степеня інтерполяційних многочленів. Але в умовах відсутності додаткової інформації про задану таблично функцію останні дають досить значну похибку. На практиці рідко проводять інтерполяцію поліномами степенів вище третього, тому що, по-перше, вони дають значні похибки й, по-друге, при нескінченному збільшенні порядку n інтерполяційного полінома $P_n(x)$ послідовність P_n не є збіжною (відповідно до теореми Фабера). Цей факт уперше виявив Рунге в 1901 р. В цьому випадку більш ефективним є використання сплайнів, що на проміжку між вузлами інтерполювання є поліномами невисокого степеня. На всьому проміжку інтерполяції $[a,b]$ сплайн - це функція, що склеєна з різних частин поліномів. Отже, розглянемо на відрізку $[a,b]$ систему вузлів $a=x_0<x_1<x_2<\ldots<x_n=b$. *Сплайном $S_m(x)$ називається функція, що визначена на $[a,b]$, має на ньому неперервні похідні $m-1$ порядку і на кожному частковому відрізку $[x_i,x_{i+1}]$ збігається з деяким многочленом степеня не вище m. При цьому хоча б на одному з відрізків степінь многочлена дорівнює m. Якщо $S_m(x_i)=f(x_i)$, маємо інтерполюючий сплайн.* Визначити сплайн можна також так. ***Поліноміальним сплайном порядку m та дефекту k*** називається функція $S_{m,k}(x)$ на сітці $a=x_0<x_1<\ldots<x_{n-1}<x_n=b$ така, що:

1) кожному проміжку $x\in[x_i,x_{i+1}]; S_{m,k}(x)\in P_m$;

2) число k називається дефектом сплайна, якщо $S_{m,k} \in C^{m-k}[a,b]$, $0<k<m$;

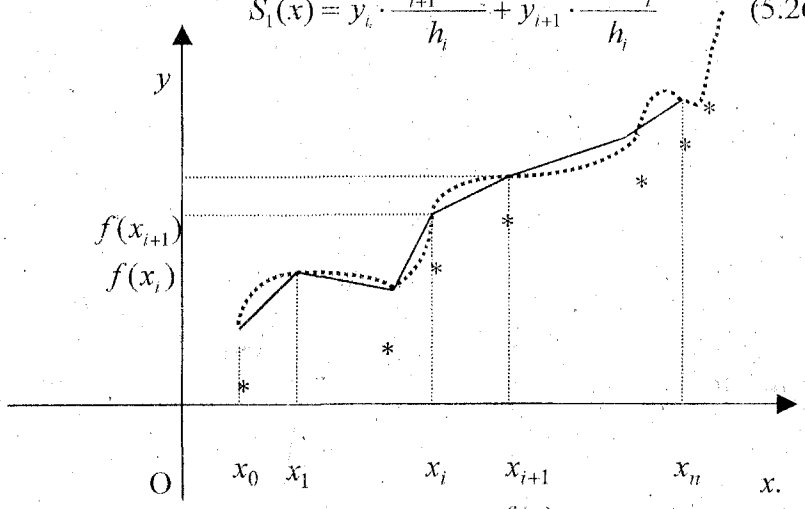
3) розглянемо сплайн дефекту 1. $S_{m,1} = S_m$. Інтерполяційним сплайном називається $S_m(x)$, якщо $S_m(x_i)=y_i$, $i = 0, 1, \ldots, n$.

Лінійний інтерполяційний сплайн

Нехай ω - розбиття відрізка $x_0, x_1, \ldots, x_n$: $\omega : a < x_0 < x_1 < \ldots < x_n = b$, $y_i = f(x_i)$ - задані значення.

Сплайном першого степеня називається :неперервна на відрізку $[a,b]$, лінійна на кожному частковому проміжку функція $f(x)$. Його позначення $S_1(x)$. Нехай $x \in [x_i, x_{i+1}]$, $h_i = x_{i+1} - x_i$. Вираз для сплайна $S_1(x)$ на цьому проміжку

$$S_1(x) = y_i \cdot \frac{x_{i+1} - x}{h_i} + y_{i+1} \cdot \frac{x - x_i}{h_i} \qquad (5.26)$$



$f(x)$ - ·····

Рис. – 5.1 $S_1(x)$ - ——

Залишковий член такої інтерполяції $R_1(x) = S_1(x) - f(x)$.

Оцінка залишкового члена залежить від диференціальних властивостей функції $f(x)$.

Нехай $y = f(x) \in C[a,b]$. Позначимо

$\omega_i(x) = \max_{x', x'' \in [x_i, x_{i+1}]} |y(x'') - y(x')|$ - коливання функції $f(x)$ на проміжку $[x_i, x_{i+1}]$. Тоді використовується лема.

Лема (варіант теореми про середнє).

Нехай $y = f(x) \in C[a,b]$. Якщо величини α, β однакового знака, то існує точка $\xi \in [a,b]$ така, що $\alpha y(a) + \beta y(b) = (\alpha + \beta) y(\xi)$.

За допомогою цієї леми доводиться теорема про оцінку залишкового члена лінійного інтерполяційного сплайна.

Теорема. Якщо $y = f(x) \in C[a,b]$, то $\|R_1\| \le \omega(f)$.

Дійсно,

$R_1(x) - S_1(x) - f(x) = (1-t) \cdot y_i + t \cdot y_{i+1} - f(x)$, де $t = \frac{x - x_i}{h_i}$.

З наведеної вище леми маємо $R_1(x) = f(\xi) - f(x)$, де $\xi \in [x_i, x_{i+1}]$. Отже,

$$|R_1| \le \omega_i(f) - \omega(f).$$

З поліпшенням гладкості функції $f(x)$ оцінка похибки її інтерполяції лінійними сплайнами також поліпшується. А саме, якщо $y \in C^1[a,b]$, то $\|R_1\| \le \frac{\bar{h}}{2} \cdot \|f'(x)\|$, де $\bar{h} = \max_i h_i$.

Для $f(x) \in C^2[a,b]$ можна одержати оцінку $\|R_1\| \le \frac{\bar{h}^2}{8} \cdot \|f''(x)\|$.

Подальше збільшення гладкості функції $f(x)$ не дає підвищення порядку апроксимації. Відбувається насичення алгоритму.

Збіжність

Нехай на $[a,b]$ задана послідовність сіток $\Delta_k : a = x_{k,0} < x_{k,1} < \ldots < x_{k,n} = b$, $k=1,2,3\ldots$, які задовольняють умову $\bar{h}_k \to 0$ при $k \to \infty$. Для Δ_k будується інтерполяційний сплайн $S_{1,\Delta_k}(x)$. Інтерполяційний процес вважається збіжним, якщо $\left\| S_{1,\Delta_k}(x) - f(x) \right\| \to 0$ при $k \to \infty$ для будь-якої функції $f(x)$ з деякого класу. Звідси випливає можливість інтерполяції з наперед заданою точністю

$$\forall \varepsilon \; \exists \Delta_k : \left\| S_{1,\Delta_k} - y \right\| \le \varepsilon .$$

Перевага лінійних сплайнів у порівнянні з інтерполяційними многочленами полягає в тому, що з оцінки похибки випливає збіжність.

Нехай $f(x) \in C[a,b]$. За доведеною теоремою $\left\| R_1 \right\| \le \omega(f)$. За визначенням $\omega(f) \to 0$ при $h \to 0$, тому процес інтерполяції лінійними сплайнами збігається на множині неперервних функцій по довільній послідовності сіток Δ_k.

Якщо $f(x) \in C^k[a,b]$, де $k=1,2$, то похибка $\left\| R_1 \right\| \le \frac{\bar{h}^l}{2} \cdot \left\| f^{(l)}(x) \right\| = O(\bar{h}^l)$. Маємо збіжність інтерполяції порядку $O(\bar{h}^l)$.

Кубічний інтерполяційний сплайн

Кубічні сплайн-функції моделюють дуже старий механічний пристрій, яким користувалися креслярі. Вони брали гнучкі рейки, виготовлені з досить пружного матеріалу, наприклад з дерева. Ці рейки закріпляли, підвішуючи важки в точках інтерполяції, що відповідають

інтерполяційним вузлам. Рейка або механічний сплайн набирали форму з найменшою потенційною енергією. Остання умова має свій математичний вираз $f^{(IV)}(x) \equiv 0$. Якщо при цьому сплайн не руйнується, то тоді функція та її похідні повинні бути неперервними на $[x_0,x_n]$. З теорії балок відомо, що функція $f(x)$ між кожною парою заданих точок може бути представлена поліномом 3-го степеня

$$f(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3,$$

де $x_{i-1} < x < x_i$. При цьому між кожною парою сусідніх вузлів поліноми з'єднуються неперервно (так само, як їх перші та другі похідні).

Інтерполяція кубічними сплайнами - це швидкий, ефективний і стійкий спосіб інтерполяції функцій, що є основним конкурентом поліноміальної інтерполяції. У його основу покладена така ідея - інтервал інтерполяції розбивається на невеликі відрізки, на кожному з яких функція задається поліномом третього степеня. Коефіцієнти полінома підбираються так, що на границях інтервалів забезпечується неперервність функції, її першої та другої похідних. Також є можливість задати граничні умови - значення першої або другої похідних на границях інтервалу. Якщо значення однієї з похідних на границі відомі, то задавши їх, ми одержуємо вкрай точну інтерполяційну схему. Якщо значення невідомі, то можна задати другу похідну на границі, що дорівнює нулю, й одержати досить гарні результати.

Кубічну сплайн-функцію, що задовольняє умови $f''(x_1) = f''(x_n) = 0$, називають природним кубічним сплайном. З математичної точки зору було доведено [Алберг, 1972], що вона є єдиною функцією з мінімальною кривизною серед усіх функцій, що інтерполюють функцію в заданих точках та мають квадратично інтегровану другу похідну. У цьому змісті кубічний сплайн буде самою гладкою функцією, що інтерполює задані точки.

Побудова кубічного сплайна - простий і чисельно стійкий процес. Для $S_3(x)$ треба визначити *4* коефіцієнти для кожного проміжку $[x_i, x_{i+1}]$, тобто $4n$ параметрів. Вимагається щоб у внутрішніх вузлах сплайн і його похідні до 2-го порядку були неперервними

$$S^{(r)}(x_i - 0) = S^{(r)}(x_i + 0),\ i=1,\ldots,n-1;\ r=0,1,2.$$

Це дає *3n-3* умови для визначення параметрів, ще *n+1* умова міститься у вимозі $S_3(x_i) = y_i,\ i=0,1,\ldots,n.$. Разом маємо *4n-2* умови. Ще 2 умови, необхідні для однозначного визначення коефіцієнтів сплайна, як правило, задаються у вигляді граничних умов, тобто умов у точках *a* й *b*. Розглянемо природні граничні умови $S''(a) = S''(b) = 0$.

Позначивши $S''(x_i) = M_i$ та враховуючи її лінійність, одержуємо

$$S''(x) = M_i \frac{x_{i+1} - x}{h_i} + M_{i+1} \frac{x - x_i}{h_i},\ x \in [x_i,\ x_{i+1}]. \qquad (5.27)$$

Двічі інтегруючи (5.27), одержуємо

$$S'(x) = -M_i \cdot \frac{(x_{i+1} - x)^2}{2 \cdot h_i} + M_{i+1} \cdot \frac{(x - x_i)^2}{2 \cdot h_i} + A, \qquad (5.28)$$

$$S(x) = M_i \frac{(x_{i+1} - x)^3}{6h_i} + M_{i+1} \frac{(x - x_i)^3}{6h_i} + Ax + B, \qquad (5.29)$$

де *A* та *B*- постійні інтегрування. Вищезгадані умови дають

$$\begin{aligned} S(x_i) &= \frac{M_i \cdot h_i^2}{6} + A \cdot x_i + B = y_i \\ S(x_{i+1}) &= \frac{M_{i+1} \cdot h_i^2}{6} + A \cdot x_{i+1} + B = y_{i+1}. \end{aligned} \qquad (5.30)$$

З них одержуємо

$$A = \frac{y_{i+1} - y_i}{h_i} + \frac{M_i - M_{i+1}}{6 \cdot h_i} \cdot h,$$

$$B = y_i - \frac{M_i \cdot h_i^2}{6} - \frac{y_{i+1} - y_i}{h_i} \cdot x - \frac{M_{i+1} - M_i}{6} \cdot h_i \cdot x_i.$$

Підставляючи *A* та *B* в (5.29), одержуємо

$$S(x) = M_i \cdot \frac{(x_{i+1} - x)^3}{6 \cdot h_i} + M_{i+1} \cdot \frac{(x - x_i)^3}{6 \cdot h_i} + \qquad (5.31)$$
$$+ (y_{i+1} - \frac{M_{i+1}}{6} \cdot h_i^2) \cdot \frac{x - x_i}{h_i} + (y_i - \frac{M_i}{6} \cdot h_i^2) \cdot \frac{x_{i+1} - x_i}{h_i}$$

$$S'(x) = -M_i \cdot \frac{(x_{i+1} - x_i)^2}{2 \cdot h_i} + M_{i+1} \cdot \frac{(x - x_i)^2}{2 \cdot h_i} + \frac{y_{i+1} - y_i}{h_i} - \frac{M_{i+1} - M_i}{6} \cdot h_i. \qquad (5.32)$$

З (5.28) знаходимо значення однобічних похідних для вузла x_i, $i=1,2,\ldots,n-1$

$$\begin{aligned} S'(x_i - 0) &= M_{i-1} \cdot \frac{h_{i-1}}{6} + M_i \cdot \frac{h_{i-1}}{3} + \frac{y_i - y_{i-1}}{h_{i-1}} \\ S'(x_i + 0) &= -M_i \cdot \frac{h_i}{3} - M_{i+1} \cdot \frac{h_i}{6} + \frac{y_{i+1} - y_i}{h_i} \end{aligned} \qquad (5.33)$$

Вимагаючи неперервності $S'(x)$ у вузлі x_i одержуємо

$$\frac{h_{i-1}}{6} M_{i-1} + \frac{h_{i-1} + h_i}{3} M_i + \frac{h_i}{6} M_{i+1} = \frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}},\ \text{де } i=1,\ldots,n-1. \qquad (5.34)$$

Отже, отримуємо систему рівнянь відносно M_i вигляду

$$A \cdot M = H \cdot F \qquad (5.35)$$

із квадратною матрицею *A*

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & \frac{h_0 + h_1}{3} & \frac{h_1}{6} & 0 & \ldots & 0 & 0 \\ 0 & \frac{h_1}{6} & \frac{h_1 + h_2}{3} & \frac{h_2}{6} & \ldots & 0 & 0 \\ 0 & 0 & \frac{h_2}{6} & \frac{h_2 + h_3}{3} & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & \frac{h_{n-2} + h_{n-1}}{3} & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}$$

і квадратною матрицею H

$$H=\begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \frac{1}{h_0} & -\left(\frac{1}{h_0}+\frac{1}{h_1}\right) & \frac{1}{h_1} & 0 & \dots & 0 & 0 \\ 0 & \frac{1}{h_1} & -\left(\frac{1}{h_1}+\frac{1}{h_2}\right) & \frac{1}{h_2} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -\left(\frac{1}{h_{n-2}}+\frac{1}{h_{n-1}}\right) & \frac{1}{h_{n-1}} \\ & 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

Координатами вектора F є значення $y_0, y_1, \dots, y_n$.

Для матриці A ненульові елементи розміщені на головній діагоналі й двох сусідніх з нею. Такі матриці називаються тридіагональними. Для невиродженої матриці A виконана умова діагональної переваги $|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|$. Отже, система (5.35) однозначно розв'язувана, тобто існує єдиний кубічний інтерполяційний сплайн.

Вигляд граничних умов змінює деякі елементи матриці А, але в кожному разі вона залишається матрицею з діагональною перевагою. Розв'язок системи (5.35) із тридіагональною матрицею A може бути знайдений методом прогонки.

Випадки використання кубічного сплайна

При побудові інтерполяційного кубічного сплайна найчастіше використовуються граничні (крайові) умови трьох типів. Вибір граничних (крайових) умов є однією з центральних проблем при інтерполяції функцій. Він особливо важливий за необхідності забезпечити високу точність апроксимації функції $f(x)$ сплайном $S(x)$ поблизу кінців відрізка $[a,b]$. Граничні значення суттєво

впливають на поведінку сплайна $S(x)$ поблизу точок a та b. Цей вплив швидко слабшає при відході від них.

Отже, якщо $h=\frac{b-a}{n}$, $x_i = x_0 + ih$ $(i=\overline{0,n-1})$, $x \in [x_i, x_{i+1}]$, то кубічний сплайн на цьому відрізку можна представити формулою

$$S_3(x)=\frac{(x_{i+1}-x)^2(2(x-x_i)+h)}{h^3}f(x_i)+\frac{(x-x_i)^2(2(x_{i+1}-x)+h)}{h^3}f(x_{i+1})+$$
$$+\frac{(x_{i+1}-x)^2(x-x_i)}{h^2}m_i+\frac{(x-x_i)^2(x-x_{i+1})}{h^2}m_{i+1}.$$

Тут $m_i = S_3'(x_i)$; $m_{i+1} = S_3'(x_{i+1})$. Для їх визначення накладають умови неперервності другої похідної в точках x_i та обмеження на значення сплайна і його похідних на кінцях проміжку $[a,b]$ - крайові умови. Тобто потрібна додаткова інформація про функцію, для якої є потреба в інтерполюванні.

Якщо на кінцях відрізка $[a,b]$ відомі значення 1-ї похідної $f'(x)$, то природно скористатися граничними (крайовими) умовами 1-го типу.

1 Граничні (крайові) умови 1-го типу. Якщо відомо, що $S_3'(a) = f'(a); S_3'(b) = f'(b)$, то для визначення m_i маємо систему рівнянь

$$\begin{cases} m_0 = f_0', \\ m_n = f_n', \\ m_{i-1} + 4m_i + m_{i+1} = \frac{3}{h}(f(x_{i+1}) - f(x_i)), i = \overline{1,n-1}. \end{cases} \quad (5.36)$$

Якщо на кінцях відрізка $[a,b]$ відомі значення 2-ї похідної $f''(x)$, то природно скористатися граничними (крайовими) умовами 2-го типу.

2 Граничні (крайові) умови 2-го типу. Якщо відомо

$S_3''(a) = f''(a), S_3''(b) = f''(b)$, то є система рівнянь

$$\begin{cases} 2m_0 + m_1 = \dfrac{3}{h}(f(x_1) - f(x_0)) - \dfrac{h}{2}f''(x_0) \\ 2m_n + m_{n-1} = \dfrac{3}{h}(f(x_n) - f(x_{n-1})) + \dfrac{h}{2}f''(x_n) \\ m_{i-1} + 4m_i + m_{i+1} = \dfrac{3}{h}(f(x_{i+1}) - f(x_{i-1})), i = \overline{1, n-1} \end{cases} \quad (5.37)$$

Якщо є можливість вибору між граничними (крайовими) умовами 1-го та 2-го типів, то перевагу потрібно надати умовам 1-го типу.

У випадку, коли ніякої додаткової інформації про поведінку апроксимованої функції немає, часто використовують так звані природні граничні (крайові) умови $S''(a) = 0, S''(b) = 0$.

Однак варто мати на увазі, що при такому виборі граничних (крайових) умов точність апроксимації функції $f(x)$ сплайном $S(x)$ поблизу кінців відрізка $[a,b]$ різко знижується. Іноді користуються граничними (крайовими) умовами 1-го або 2-го типу, але не з точними значеннями відповідних похідних, а з їх різницевими апроксимаціями. Точність такого підходу невисока.

Практичний досвід розрахунків показує, що в такій ситуації найбільш доцільним є вибір природних граничних (крайових) умов. Якщо $f(x)$ - періодична функція, то потрібно зупинитися на граничних (крайових) умовах 3-го типу.

3 Граничні (крайові) умови 3-го типу. Якщо $f(x)$ - періодична функція $f(x) = f(x+T)$, то $f(x_0) = f(x_n), f(x_1) = f(x_{n+1}) \Rightarrow m_0 = m_n, m_1 = m_{n+1}$ і система рівнянь має вигляд

$$\begin{cases} 4m_1 + m_2 + m_n = \dfrac{3}{h}(f(x_2) - f(x_0)), \\ m_{i-1} + 4m_i + m_{i+1} = \dfrac{3}{h}(f(x_{i+1}) - f(x_{i-1})), i = 2,3,\ldots,n-1, \\ m_1 + m_{n-1} + 4m_n = \dfrac{3}{h}(f(x_1) - f(x_{n-1})). \end{cases} \quad (5.38)$$

Приклад реалізації алгоритмів інтерполяції функцій на псевдокоді

```
//розв'язує СЛАР методом Гауса
//M – матриця системи і вільні члени
//X – вектор відповідей
//n – кількість невідомих
    linequ(M,n,e,X):
    //доступна в модулі naz.pas
end
//X – значення аргумента
//Y – значення функції
//xi – значення аргумента, для якого потрібно знайти значення
//    функції
//h – крок
//змінні DY, D2Y і т.д. – скінченні різниці
//змінні SDY, SD2Y і т.д. – розділені різниці
```

//відшукуємо значення інтерполюючого многочлена при x=xi

```
Find_PX(X,Y,n,h):
        1       n:=X.length;
        2       for i:=1 to n do
        3         DY[i]:=Y[i+1]-Y[i]
        4       done
        5       for i:=1 to (n-1) do
        6         D2Y[i]:=DY[i+1]-DY[i]
        7       done
        8       for i:=1 to (n-2) do
        9         D3Y[i]:=D2Y[i+1]-D2Y[i]
        10      done
        11      for i:=1 to (n-3) do
        12        D4Y[i]:=D3Y[i+1]-D3Y[i]
        13      done
        14      for i:=1 to n do
        15        SDY[i]:=Y[i+1]-Y[i]/(X[i+1]-X[i])
        16      done
        17      for i:=1 to (n-1) do
18  D2Y[i]:=SDY[i+1]-SDY[i]/(X[i+2]-X[i])
        19      done
        20      for i:=1 to (n-2) do
21  D3Y[i]:=SD2Y[i+1]-SD2Y[i]/(X[i+3]-X[i])
```



```
        22      done
        23      for i:=1 to (n-3) do
24  D4Y[i]:=SD3Y[i+1]-SD3Y[i]/(X[i+4]-X[i])
        25      done
        26      t:=(xi-X[1])/h
        27      return
    X[1]+(t/factorial(1))*DY[1]+((t*(t-
    1))/factorial(2))*D2Y[1]+((t*(t-1)*(t-
    2))/factotrial(3))*D3Y[1]+((t*(t-1)*(t-2)*(t-
    3))/factorial(4))*D4Y[1];
end
```

//R – дані сплайна

//eps – точність обчислень

//xi – значення аргумента, при якому потрібно знайти значення функції //

//відшукуємо значення кубічного сплайна в точці xi

```
Find_Spline(R,eps,xi)
        1       linequ(R,R.length,1E-6,M)
        2       k:=2
        3       s31:=(X[k+1]-xi)*(X[k+1]-xi)*(2*(xi-
    X[k])+1)*Y[k]+(xi-X[k])*(xi-X[k])*(2*(X[k+1]-
    xi)+1)*Y[k+1]
        4       s32:=(X[k+1]-xi)*(X[k+1]-xi)*(xi-
    X[k])*M[k]+(xi-X[k])*(xi-X[k])*(xi-X[k+1])*M[k+1]
        5       return s31+s32
end
```

Оцінка похибки та збіжності при інтерполяції кубічними сплайнами

Якщо $f(x) \in C[a,b]$, то похибка інтерполяції кубічним сплайном

$$\|R_3\| \le c\omega(f), \text{ де } c = 1 + \frac{3}{4}\beta,\ \beta = \frac{\overline{h}}{h} = \frac{\max h_i}{\min h_i}.$$

Якщо $f(x) \in C^r[a,b]$, $r=1,2,3,4$, то оцінка має вигляд для $\|R_3\| \le c\overline{h}^2\omega(f^{(r)})$.

Із цих оцінок треба встановити збіжність інтерполяційного процесу на послідовності сіток Δ_k.

Для простоти обчислень або при труднощах у пошуку першої та другої похідних заданої функції можна застосувати таку оцінку похибок:

$$\gamma^{\text{відн}}{}_{\max} = \frac{\max|f(x_i) - S(x_i)|}{\max\limits_{x\in[a,b]}|f(x)|}.$$

Приклад. Апроксимувати функції $y = x^2$ на відрізку [-2;2], використовуючи лінійний сплайн і природний кубічний сплайн.

Дослідження проведемо на рівномірних сітках з кількістю вузлів інтерполяції : 5, 7, 9, 15 , 51 , 101 відповідно. Визначимо відносну похибку інтерполяції сплайнами на різних сітках. Результати занесемо до таблиці.

Аналіз результатів показує, що точність апроксимації істотно залежить від кількості вузлових точок.

Число вузлів	Лінійний сплайн	Кубічний сплайн	Куб. сплайн по другій похідній
5	0,0625	0,024	0,0092
7	0,0278	0,011	0,0038
9	0,0156	0,0061	0,002
15	0,0051	0,002	0,0007
51	0,0004	0,00016	0,000055
101	0,0001	0,000036	0,000014

Апроксимаційні властивості кубічного сплайна

Апроксимаційні властивості кубічного сплайна залежать від гладкості функції $f(x)$ - чим вище гладкість інтерпольованої функції, тим вище порядок апроксимації при подрібленні сітки і тим швидшою є збіжність.

Якщо інтерпольована функція $f(x)$ неперервна на відрізку $[a,b]$, тобто $f(x) \in C^0[a,b]$, то

$\|f(x) - S(x)\|_C = \max\limits_{x\in[a,b]}|f(x) - S(x)| \to 0$ при $h = \max\limits_{0\le i\le N-1} h_i \to 0$.

Якщо інтерпольована функція $f(x)$ має на відрізку $[a,b]$ неперервну першу похідну, тобто $f(x) \in C^1[a,b]$, а $S(x)$ - інтерполяційний сплайн, що задовольняє граничні умови 1-го або 3-го типу, то при $h \to 0$ $\|f(x) - S(x)\|_C = o(h)$, $\|f'(x) - S'(x)\|_C = o(1)$.

У цьому випадку не тільки сплайн збігається до інтерпольованої функції, але і похідна сплайна збігається до похідної цієї функції.

На випадок, якщо $f(x) \in C^4[a,b]$, сплайн $S(x)$ апроксимує на відрізку $[a,b]$ функцію $f(x)$, а його 1-а та 2-а похідні апроксимують відповідно функції $f'(x)$ та $f''(x)$: $\|f(x) - S(x)\|_C = o(h^4)$, $\|f'(x) - S'(x)\|_C = o(h^3)$, $\|f''(x) - S''(x)\|_C = o(h^2)$.

5.2.7 Застосування інтерполяції для складання таблиць

Теорія інтерполяції має застосування при складанні таблиць функцій. Одержавши завдання на складання таблиць тих чи інших функцій, математик повинен вирішити перед початком обчислень ряд питань. Повинна бути обрана формула, за якою будуть проводитися обчислення. Ця формула може змінюватися від ділянки до ділянки. Як правило, формули для обчислення значень функції бувають громіздкими і тому їх використовують для одержання деяких опорних значень і потім, шляхом субтабулювання, згущують таблицю. Формула, що дає опорні значення функції, повинна забезпечувати потрібну точність таблиць із врахуванням наступного субтабулювання . Якщо передбачається скласти таблиці з постійним кроком, то спочатку необхідно визначити крок таблиці.

Найчастіше таблиці функцій складаються так, щоб була можлива лінійна інтерполяція (тобто інтерполяція з використанням перших двох членів формули Тейлора). У цьому випадку залишковий член буде мати вигляд $R_1(x)=\frac{f''(\xi)}{2!}h^2t(t-1)$. Тут ξ належить інтервалу між двома сусідніми табличними значеннями аргумента, у якому лежить x, а $t\in(0,1)$. Добуток $t(t-1)$ набуває найбільшого за модулем значення при $t=1/2$. Це значення дорівнює $\frac{1}{4}$. Отже, $|R_1(x)|\le\frac{M_2h^2}{8}$, де $M_2=\max|f''(\xi)|$.

Щоб помилка інтерполяції не перевищувала за абсолютною величиною деяке a, необхідно вибрати h, яке задовольняло б умову $h\le\sqrt{\frac{8a}{M_2}}$.

5.3 Метод найменших квадратів

Аналізуючи попереднє, можна зазначити, що інтерполювання може бути здійснене лише на невеликому інтервалі по кількох вузлах інтерполяції, процес обчислення скінченних різниць є нестійким. Окрім того, якщо значення x_i подають значення функції, яка наближується, зі значними похибками, інтерполювати ці значення недоцільно.

За таких умов застосовують середньоквадратичне наближення. Найбільш ефективним методом побудови середньоквадратичного наближення функції є метод найменших квадратів (МНК).

Нехай є відомими n значень x_i $(i=1,2,...,n)$ деякої фізичної величини $x(t)$, виміряної у моменти часу t_i. Припустимо, що ці значення подають істинні значення функції $x(t)$ у відповідні моменти часу зі значними похибками, значення яких невідомі, але припускається, що ці похибки є випадковими з математичним сподіванням, що дорівнює нулю. Будемо наближати невідому функцію $x(t)$ за допомогою лінійної комбінації деяких відомих m функцій $\varphi_k(t)$

$$X(t)=\sum_{k=0}^{m-1}c_k\varphi_k(t), \qquad (5.39)$$

де функції $\varphi_0(t),\varphi_1(t),\ldots,\varphi_{m-1}(t)$ називатимемо базовими функціями. Потрібно визначити m невідомих коефіцієнтів c_k $(k=0,...,m-1)$ з умови, щоб квадрат середньоквадратичного відхилення (СКВ) апроксимуючої функції $X(t)$ від апроксимованої $x(t)$ (обчисленого для заданих значень t_i аргумента t)

$$\sigma^2(c,\varphi)=\sum_{i=1}^{n}(x_i-X(t_i))^2=\sum_{i=1}^{n}(x_i-\sum_{k=0}^{m-1}c_k\cdot\varphi_k(t_i))^2 \qquad (5.40)$$

був мінімальним (саме тому відповідний метод називається МНК). Квадрат СКВ (5.40) є функцією m невідомих коефіцієнтів c_k $(k=0,1,...,m-1)$. Тому для пошуку його мінімуму необхідно знайти m частинних похідних за окремими коефіцієнтами

$$\frac{\partial\sigma^2(c,\varphi)}{\partial c_k}=-2\sum_{i=1}^{n}(x_i-\sum_{k=0}^{m-1}c_k\cdot\varphi_k(t_i))\varphi_k(t_i) \qquad (5.41)$$

$$(k=0,1,...,m-1)$$

і прирівняти їх до нуля. В результаті одержується система з m лінійних алгебричних рівнянь з m невідомими $c_0, c_1, ..., c_{m-1}$:

$$\begin{cases} c_0\sum_{i=1}^{n}\varphi_0^2(t_i)+c_1\sum_{i=1}^{n}\varphi_0(t_i)\cdot\varphi_1(t_i)+...+c_{m-1}\sum_{i=1}^{n}\varphi_0(t_i)\cdot\varphi_{m-1}(t_i)=\sum_{i=1}^{n}x_i\cdot\varphi_0(t_i) \\ c_0\sum_{i=1}^{n}\varphi_0(t_i)\cdot\varphi_1(t_i)+c_1\sum_{i=1}^{n}\varphi_1^2(t_i)+...+c_{m-1}\sum_{i=1}^{n}\varphi_1(t_i)\cdot\varphi_{m-1}(t_i)=\sum_{i=1}^{n}x_i\cdot\varphi_1(t_i) \\ ... \\ c_0\sum_{i=1}^{n}\varphi_0(t_i)\cdot\varphi_{m-1}(t_i)+c_1\sum_{i=1}^{n}\varphi_1(t_i)\cdot\varphi_{m-1}(t_i)+...+c_{m-1}\sum_{i=1}^{n}\varphi_{m-1}^2(t_i)=\sum_{i=1}^{n}x_i\cdot\varphi_{m-1}(t_i) \end{cases} \qquad (5.42)$$

Система (5.42) називається нормальною системою для методу найменших квадратів. Визначником цієї системи є визначник Грама сукупності функцій $\varphi_k(t)$:

$$\Delta=\begin{vmatrix} \sum_{i=1}^{n}\varphi_0^2(t_i) & \sum_{i=1}^{n}\varphi_0(t_i)\cdot\varphi_1(t_i) & ... & \sum_{i=1}^{n}\varphi_0(t_i)\cdot\varphi_{m-1}(t_i) \\ \sum_{i=1}^{n}\varphi_0(t_i)\cdot\varphi_1(t_i) & \sum_{i=1}^{n}\varphi_1^2(t_i) & ... & \sum_{i=1}^{n}\varphi_1(t_i)\cdot\varphi_{m-1}(t_i) \\ ... & ... & ... & ... \\ \sum_{i=1}^{n}\varphi_0(t_i)\cdot\varphi_{m-1}(t_i) & \sum_{i=1}^{n}\varphi_1(t_i)\cdot\varphi_{m-1}(t_i) & ... & \sum_{i=1}^{n}\varphi_{m-1}^2(t_i) \end{vmatrix} \qquad (5.43)$$

Як відомо, якщо функції $\varphi_k(t)$ $(k=0,1,...,m-1)$ складають сукупність взаємонезалежних функцій (тобто ніяку з цих функцій неможливо подати як лінійну комбінацію решти з них), то визначник Грама цих функцій

не дорівнює нулю. Це означає, що за базові функції при апроксимуванні потрібно обирати сукупності лінійно незалежних функцій. Тоді СЛАР (5.42) має єдиний розв'язок - значення коефіцієнтів c_k $(k=0,1,...,m-1)$, що забезпечують мінімум квадрата середньоквадратичного відхилення апроксимуючої та апроксимованої функцій.

Ортогональними на деякому інтервалі $[t_1,t_n]$ функціями називається сукупність таких функцій, що

$$\sum_{i=1}^{n}\varphi_k(t_i)\varphi_j(t_i)=\begin{cases}1, & k=j\\ 0, & k\neq j\end{cases}.$$

Матриця Грама для ортогональних функцій є одиничною.

У випадку, коли за базові при апроксимуванні обрані ортогональні функції, обчислення коефіцієнтів апроксимації значно спрощується. У цьому випадку значення їх можна визначити співвідношенням

$$c_k=\sum_{i=1}^{n}x_i\varphi_k(t_i). \qquad (5.44)$$

Тому при апроксимуванні бажано обирати за базові системи ортогональних функцій.

Класичними прикладами ортогональних функцій-поліномів є поліноми Якобі, Лежандра, Лагерра, Чебишева, Ерміта. Наприклад, поліноми Лежандра $P_n(t)$ є ортогональними на відрізку $[-1;1]$ і мають вигляд:

$$P_0(t)=1;\ P_1(t)=t;\ P_2(t)=\frac{1}{2}(3t^2-1);\ P_3(t)=\frac{1}{2}(5t^3-t);$$

$$P_4(t)=\frac{1}{8}(35t^4-30t^2+3);\ P_5(t)=\frac{1}{8}(63t^5-70t^3+15t);$$

$$P_6(t)=\frac{1}{16}(231t^6-315t^4+105t^2-5);\ P_7(t)=\frac{1}{16}(429t^7-693t^5+315t^3-35t).$$

Поліноми Чебишева першого роду $T_n(t)$ є ортогональними також на інтервалі $[-1;1]$. Їх можна задати співвідношенням

$$T_n(t) = \cos(n \cdot \arccos(t)); \qquad (n = 0,1,2,...),$$

а рекурентна формула їх визначення має такий вигляд:

$$T_{n+1}(t) = 2t \cdot T_n(t) \cdot T_{n-1}(t); \qquad (n = 1,2,...).$$

Наведемо приклади поліномів Чебишева першого роду:

$$T_0(t) = 1;\ T_1(t) = t;\ T_2(t) = 2t^2 - 1;\ T_3(t) = 4t^3 - 3t;$$

$$T_4(t) = 8t^4 - 8t^2 + 1; \quad T_5(t) = 16t^5 - 20t^3 + 5t;$$

$$T_6(t) = 32t^6 - 48t^4 + 18t^2 - 1;\ T_7(t) = 64t^7 - 112t^5 + 56t^3 - 7t.$$

Поліноми Чебишева другого роду $U_n(t)$ також ортогональні на тому самому інтервалі і мають такий вигляд:

$$U_0(t) = 1;\ U_1(t) = 2t;\ U_2(t) = 4t^2 - 1;\ U_3(t) = 8t^3 - 4t;$$

$$U_4(t) = 16t^4 - 12t^2 + 1; \qquad U_5(t) = 32t^5 - 32t^3 + 6t.$$

Поліноми Ерміта $H_n(t)$ ортогональні на всій числовій осі $(-\infty,+\infty)$ і мають вигляд

$$H_0(t) = 1;\ H_1(t) = 2t;\ H_2(t) = 4t^2 - 2;\ H_3(t) = 8t^3 - 12t;$$

$$H_4(t) = 16t^4 - 48t^2 + 12;\ H_5(t) = 32t^5 - 160t^3 + 120t.$$

Наведені системи ортогональних поліномів стають у нагоді, коли за апроксимуючу функцію обирається поліном певного степеня, тобто для здійснення так званої поліноміальної апроксимації.

Прикладом системи неортогональних базових поліномів може бути така система:

$$P_0(t) = 1; \qquad P_1(t) = t;\ ...,\ P_k = t^k;\ ....$$

Вона часто використовується на практиці. Тоді $X(t) = a_0 + a_1 t + ... + a_m t^m$ - многочлен степеня m. В цьому разі до розв'язку пропонується система вигляду

$$\begin{cases} (n+1)\cdot a_0 + a_1 \cdot \sum_{i=0}^{n} x_i + ... + a_m \cdot \sum_{i=0}^{n} x_i^m = \sum_{i=0}^{n} y_i \\ a_0 \cdot \sum_{i=0}^{n} x_i + a_1 \cdot \sum_{i=0}^{n} x_i^2 + ... + a_m \cdot \sum_{i=0}^{n} x_i^{m+1} = \sum_{i=0}^{n} y_i \cdot x_i \\ \qquad\qquad\qquad\qquad ..... \\ a_0 \cdot \sum_{i=0}^{n} x_i^m + a_1 \cdot \sum_{i=0}^{n} x_i^{m+1} + ... + a_m \cdot \sum_{i=0}^{n} x_i^{2\cdot m} = \sum_{i=0}^{n} y_i x_i^m \end{cases}$$

При $m = n$ отриманий многочлен збігається з інтерполяційним многочленом Лагранжа.

Приклад. Найпростіша емпірична формула $X = at + b$.

Про придатність цієї формули можна робити висновки за величинами $a_i = \frac{x_{i+1} - x_i}{t_{i+1} - t_i}$. Якщо $a_i \approx const$, то формула підходить. Невідомі коефіцієнти a, b знайдемо з необхідної умови екстремуму функції

$$\Phi(a,b) = \sum_{i=1}^{n} (x_i - at_i - b)^2.$$

У результаті одержимо систему лінійних рівнянь

$$\begin{cases} \sum_{i=1}^{n} [x_i - (at_i + b)] = 0 \\ \sum_{i=1}^{n} [x_i - (at_i + b)] t_i = 0. \end{cases}$$

Розв'язуючи систему ,знаходимо a і b , що при заданому вигляді рівняння регресії забезпечують мінімум $\Phi(a,b)$.

$$a=\frac{\sum_{i=1}^{n} x_i t_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} t_i^2 - \frac{1}{n}(\sum_{i=1}^{n} t_i)^2}; \quad b=\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{a}{n}\sum_{i=1}^{n} t_i.$$

При цьому, природно, у результаті апроксимування певної сукупності даних в усіх випадках одержується однаковий поліном. Різниця полягає лише у зручності, простоті отримання коефіцієнтів цього полінома.

Якщо при поліноміальній апроксимації кількість базових функцій-поліномів дорівнює 2, тобто $m=2$, апроксимація називається лінійною. В результаті лінійного апроксимування одержують так звану лінію регресії (пряму). При $m=3$ апроксимування називають квадратичним, а при $m=4$ -кубічним.

Звичайно, апроксимування не обов'язково має бути поліноміальним. Наприклад, якщо відомо, що вимірювана функція є періодичною з відомим періодом $T=\frac{2\pi}{\omega}$, де ω - кругова частота, то за базові функції зручно використовувати таку сукупність:

$$\varphi_0(t)=1;\ \varphi_1(t)=\sin(\omega t);\ \varphi_2(t)=\cos(\omega t);\ \ldots,$$

$$\varphi_{2k-1}(t)=\sin(k\omega t);\quad \varphi_{2k}(t)=\cos(k\omega t);\ \ldots,$$

тобто використовувати апроксимацію у вигляді ряду Фур'є. Тут k є цілим додатним числом, яке дорівнює номеру гармоніки у розкладі Фур'є.

Наведена сукупність функцій є ортогональною на інтервалі, кратному періодові T. Тому застосування її є вельми ефективним (потребує мінімуму обчислень), якщо інтервал вимірювання обрати кратним періодові.

Опис результатів спостережень методом найменших квадратів ускладнюється, якщо невідомі коефіцієнти в рівняння регресії входять нелінійно. Однак у багатьох

випадках задачу вдається спростити, застосовуючи деякі прості перетворення вихідного рівняння регресії.

Приклад. У ряді випадків до лінійної залежності можуть бути зведені експериментальні дані, коли їхній графік у декартовій системі координат не є пряма. Цього можна досягти шляхом уведення нових змінних $\xi=\varphi(t,x)$, $\eta=\psi(t,x)$, які вибираються так, щоб точки (ξ_i,η_i) лежали на прямій. Таке перетворення називається вирівнюванням даних. Наприклад, рівняння регресії має вигляд $x=ce^{kt}$. Прологарифмуємо функцію $lnx=lnc+kt$. Позначимо $lnx=z$, $lnc=a$. В результаті одержуємо лінійне рівняння $z=a+kt$. Методом найменших квадратів знаходимо значення a і k (див. приклад вище), після чого визначимо так само $c=e^{a}$.

Вибір вигляду регресійної залежності можна здійснити за таблицею. Для цього за вихідними даними обчислюють середні значення x_{cp} та y_{cp}

$$t(ap)=\frac{\sum_i t_i}{n},\ t(ga)=\frac{n}{\sum_i \frac{1}{t_i}},\ t(ge)=\sqrt[n]{\prod t_i},$$

$$x(ap)=\frac{\sum_i x_i}{n},\ x(ga)=\frac{n}{\sum_i x(i)},\ x(ge)=\sqrt[n]{\prod x_i}.$$

Величина $\hat{x}$ обчислюється в такий спосіб:

1) якщо t_{cp} збігається з одним із вихідних $t_i, i=1,..,n$, то $\hat{x}=x_i$;

2) якщо t_{cp} знаходиться між t_i і $t_{i+1}, i=1,..,n$, то $\hat{x}$ знаходимо як ординату відповідної точки на відрізку прямої, що з'єднує вузли (t_i, x_i) і (t_{i+1}, x_{i+1}), за формулою

$$\hat{x} = \frac{t_{cp} - t_i}{t_{i+1} - t_i}(x_{i+1} - x_i) + x_i .$$

Вибір рівняння регресії здійснюється шляхом пошуку мінімального значення виразу $\left|\frac{x_{cp} - \hat{x}}{\hat{x}}\right|$ і відповідної йому функції, використовуючи таблицю.

Таблиця 5.1 Вибір залежності

N	t_{cp}	x_{cp}	$\hat{x}$	$\left\|\frac{x_{cp} - \hat{x}}{\hat{x}}\right\|$	Вигляд функції
1	*t(ap)*	*x(ap)*			$x=a_0+a_1*t$
2	*t(ra)*	*x(ap)*			$x=a_0+a_1/t$
3	*t(re)*	*x(ap)*			$x=a_0+a_1 \lg t$
4	*t(ap)*	*x(re)*			$x=a_0*a_1^t$
5	*t(re)*	*x(re)*			$x=a_0*t^{a1}$
6	*t(ra)*	*x(re)*			$x=exp(a_0+a_1/t)$
7	*t(ap)*	*x(ra)*			$x=1/(a_0+a_1*t)$
8	*t(re)*	*x(ra)*			$x=1/(a_0+a_1 \lg t)$
9	*t(ra)*	*x(ra)*			$x=t/(a_0+a_1*t)$

Таблицею доречно користуватися, якщо значення нашої функції носять монотонний характер.

Приклад. Функція $y=f(x)$ задана таблицею значень $y_0, y_1, \ldots y_n$ у точках $x_0, x_1, \ldots x_n$. Використовуючи метод найменших квадратів (МНК), знайти многочлен $Pm(x) = a_0 + a_1 x + \ldots + a_m x^m$ найкращого середньоквадратичного наближення оптимального степеня $m=m^*$. За оптимальне значення m^* прийняти той степінь многочлена, починаючи з якого $\sigma_m = \sqrt{\frac{1}{n-m}\sum_{k=0}^{n}(P_m(x_k) - y_k)^2}$ стабілізується або починає зростати.

Порядок розв'язання задачі:

1 Задати вектори ***x*** та ***y*** вихідних даних.

2 Використовуючи функцію **mnk** , знайти многочлени Pm, m=0,1,2,..., за методом найменших квадратів. Обчислити відповідні їм значення σ_m.

3 Побудувати гістограму залежності σ_m від m, на підставі якої вибрати оптимальний степінь m^* многочлена найкращого середньоквадратичного наближення.

4 На одному кресленні побудувати графіки многочленів Pm, m=0,1,2,..., m* і точковий графік вихідної функції.

Вектори вихідних даних:

$$x := \begin{pmatrix} -2.75 \\ -2 \\ -1 \\ 0.5 \\ 1 \end{pmatrix} \qquad y := \begin{pmatrix} 0.2 \\ -1.1 \\ 2.3 \\ 0.1 \\ 1.1 \end{pmatrix}$$

Функція mnk, що будує многочлен степеня m за методом найменших квадратів, повертає вектор **a** коефіцієнтів многочлена:

```
mnk(x, y, n, m) :=
    for j ∈ 0.. m
        b_j ← Σ_{i=0}^{n} y_i · (x_i)^j
        for k ∈ 0.. m
            Γ_{j,k} ← Σ_{i=0}^{n} (x_i)^{k+j}
    a ← lsolve(Γ, b)
    a
```

- формуються вектор правих частин та матриця пормальної системи Га=b методу найменших квадратів (базисні функції - 1, x, $x^2 \ldots x^m$);

- lsolve(Г,b) – вбудована функція MATHCAD, що розв'язує систему лінійних алгебраїчних рівнянь .

Вхідні параметри:

x, y - вектори вихідних даних; *n+1* - розмірність *x,y*.

Обчислення коефіцієнтів многочленів степеня 0,1,2,3 за методом найменших квадратів: $n := 4$

$a0 := mnk(x, y, n, 0)$ $\quad a0 = -0.48$ $\quad a1 = \begin{bmatrix} -0.133 \\ 0.408 \end{bmatrix}$

$a1 := mnk(x, y, n, 1)$

$a2 := mnk(x, y, n, 2)$ $\quad a2 = \begin{bmatrix} -1.102 \\ 1.598 \\ 0.717 \end{bmatrix}$ $\quad a3 = \begin{bmatrix} -1.164 \\ 1.591 \\ 0.792 \\ 0.026 \end{bmatrix}$

$a3 := mnk(x, y, n, 3)$

Функція P повертає значення многочлена степеня m у точці t; многочлен задасться за допомогою вектора коефіцієнтів a:

$$P(a, m, t) := \sum_{j=0}^{m} a_j \cdot t^j$$

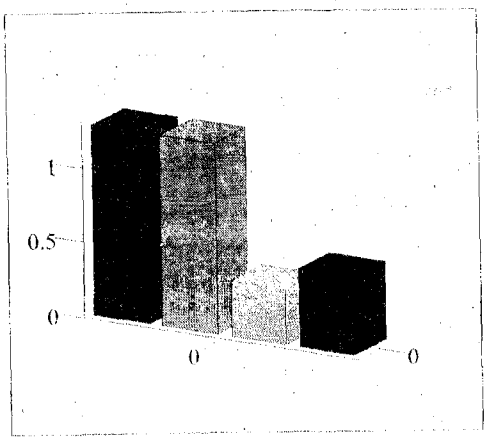
Функція $\sigma 0$ повертає значення середньоквадратичного відхилення многочлена P(a,m,t):

$$\sigma 0(a, m) := \sqrt{\frac{1}{n - m} \cdot \sum_{k=0}^{n} (P(a, m, x_k) - y_k)^2}$$

Обчислення значень σ_m,m=0,1,2,3:

$\sigma_0 := \sigma 0(a0, 0)$ $\quad \sigma_1 := \sigma 0(a1, 1)$ $\quad \sigma_2 := \sigma 0(a2, 2)$ $\quad \sigma_3 := \sigma 0(a3, 3)$

Гістограма



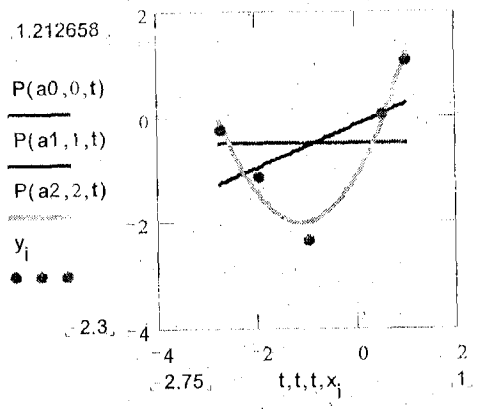
$$\sigma = \begin{bmatrix} 1.285 \\ 1.28 \\ 0.38 \\ 0.532 \end{bmatrix}$$

σ

Висновок: оптимальний степінь m*=2; многочлен найкращого середньоквадратичного наближення: *P2(x)=-1.102+1.598x+0.717x²*

Графіки многочленів степеня 0,1,2 і точковий графік вихідної функції:

$t := x_0, x_0 + 0.05 .. x_n$ $\quad i := 0 .. n$



Приклад реалізації методу найменших квадратів на псевдокоді.

Нехай за допомогою зазначеного вище методу ми знайшли вигляд рівняння регресії

$y = a_0 + a_1/x$, отже $S(a_0, a_1) = \sum_{i=1}^{n} (y_i - a_0 - a_1 / x_i)^2$.

Значення невідомих коефіцієнтів

$$\begin{cases} a_1 = \dfrac{\sum_{i=1}^{n} \frac{y_i}{x_i} - \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} \frac{1}{x_i}}{\sum_{i=1}^{n} \frac{1}{x_i^2} - \frac{1}{n}\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^2} \\ a_0 = \dfrac{1}{n}\left(\sum_{i=1}^{n} y_i - a_1 \sum_{i=1}^{n} \frac{1}{x_i}\right) \end{cases}.$$

Очевидно, що процедура, яка знайде розв'язки, буде простішою, якщо ми домовимося, що функції, що обчислюють відповідні суми, нами вже реалізовані:

```
//обчислення коефіцієнтів регресійної формули.

//X,Y – задані в умові масиви; a1,a0 – шукані
коефіцієнти

//n – кількість заданих пар x,y в умові

Metod_Kvadr(n,X,Y,a1,a0):

        1       a1:=(yixi(X,Y,n)-
 1/N)*yi(Y,n)*yi_na_1(X,n))/(yi_na_1_kw(X,n)-
 (1/n)*pow(yi_na_1(X,n),2));

        2       a0:=(1./n)*(yi(Y,n)-a1*yi_na_1(X,n));

end
```


Питання і завдання до розділу 5

1 Постановка задач наближення функцій.
2 Метод найменших квадратів. Виведення нормальної системи методу найменших квадратів.
3 Обумовленість нормальної системи.
4 Вибір оптимального степеня апроксимуючого многочлена.
5 Поліноміальна інтерполяція. Многочлен у формі Лагранжа.
6 Многочлен у формі Ньютона.
7 Похибка інтерполяції.
8 Глобальна інтерполяція. Кусочно-поліноміальна інтерполяція. Вибір вузлів інтерполяції.
9 Інтерполяція із кратними вузлами.
10 Мінімізація оцінки похибки інтерполяції.
11 Інтерполяція сплайнами. Визначення сплайна. Лінійний сплайн.
12 Побудова кубічного сплайна.
13 Види граничних умов при побудові сплайнів.
14 обудова параболічного сплайна.
15 нтерполяція функції двох змінних.
16 Вивести нормальну систему методу найменших квадратів для визначення коефіцієнтів a, b, c функції:

a) $f(x) = a\sin x + bx + c$; b) $f(x) = a\sin x + bx^2 + \dfrac{c}{x}$.

17 Використовуючи метод найменших квадратів, апроксимувати на відрізку [10,12] функцію $f(x) = \ln x$ многочленом першого степеня. Обчислити величину середньоквадратичного відхилення.

18 Побудувати інтерполяційний многочлен у формі Лагранжа й у формі Ньютона для функції $y = f(x)$, заданої таблицею значень.

a)	x	-1	0	1	b)	x	1	2	4
	y	3	2	5		y	3	4	6

19 Обчислити $\ln 5.2$, знаючи значення $\ln 5 (\approx 1.609)$ та $\ln 6 (\approx 1.792)$.

20 Побудувати кусково-лінійну інтерполяцію функції $f(x) = |x|$ за вузлами –1, 0, 1.

21 Функція $\sin x$ наближається на відрізку $[0, \frac{\pi}{4}]$ інтерполяційним многочленом за значеннями в точках $0, \frac{\pi}{8}, \frac{\pi}{4}$. Оцінити похибку інтерполяції на цьому відрізку.

22 З яким постійним кроком h потрібно скласти таблицю функції $\sin x$ на відрізку $[0, \frac{\pi}{4}]$, щоб похибка лінійної інтерполяції не перевищувала 10^{-5}?

23 Для таблично заданих функцій

a)	x	-1	0	1	2	b)	x	1	2	4	5
	y	3	2	5	1		y	3	4	6	5

побудувати лінійний і параболічний сплайни.

24 Функція $y=y(x)$ задана таблицею своїх значень:

x	-1	0	1	2
y	1.8	2.4	2.2	2

Побудувати многочлени нульового й першого степенів, що наближають функцію за методом найменших квадратів. Обчислити величину середньоквадратичного відхилення.

Побудувати на одному кресленні точковий графік функції й графіки многочленів.

25 Побудувати інтерполяційні многочлени у формі Лагранжа й Ньютона, що наближають функцію $y=y(x)$, задану таблицею своїх значень. Порівняти результати.

x	0	2	4
y	3	0	2

26 Функція $y=y(x)$ задана таблицею своїх значень:

x	0	0.2	0.4	0.6	0.8	1
y	0.75	1.1	1.35	1.25	1.05	0.8

Запропонувати способи інтерполяції для знаходження значень функції в точках x=0.24,0.5, 0.96.

27 Відновити многочлен $P_2(x)$ за його значеннями:

x	1	2	3
$P_2(x)$	2	8	16

28 Функція $y = 2^x$ задана таблицею своїх значень:

x	0	1	2	3
y	1	2	4	8

Обчислити наближено значення функції в точці $x = 1.9$ за допомогою інтерполяційного многочлена другого ступеня: а) у формі Лагранжа; б) у формі Ньютона (зі скінченними різницями). Оцінити похибку інтерполяції.

29 Функція $y = \int_0^x \sqrt{arctg(t)}\,dt$ задана таблицею своїх значень:

x	0.6	0.8	1.0	1.2	1.4
y	0.302	0.458	0.629	0.811	1.002

Обчислити приблизно значення функції в точці $x = 1.36$ за допомогою інтерполяційного

многочлена у формі Ньютона (з розділеними різницями). Обчислити похибку інтерполяції.

30 Функція $y = y(x)$ задана таблицею своїх значень:

x	0	1	2
y	1	4	6

Побудувати природний інтерполяційний кубічний сплайн дефекту 1.

31 Відомо, що апроксимуюча функція має вигляд $y = \frac{a}{x} + bx$, де a й b – невідомі параметри. Використовуючи метод найменших квадратів, визначити a й b, якщо відомо таблицю значень функції:

x	0.1	0.2	0.5
y	10.22	5.14	2.76

32 Вивести систему рівнянь для визначення коефіцієнтів a й b функції $g(x) = a\sin(x) + be^x$, що здійснює середньоквадратичну апроксимацію таблично заданої функції $y(x)$ в $n+1$ точках.

33 Функція $y(x) = \sin(x)$ наближається інтерполяційним многочленом за значеннями в точках $x = 0, \frac{\pi}{8}, \frac{\pi}{4}$. Оцінити похибку інтерполяції на відрізку $\left[0, \frac{\pi}{4}\right]$.

34 З яким кроком варто задати таблицю логарифмів на відрізку [1,10], щоб при квадратичній інтерполяції значення в проміжній точці відновлювалося з похибкою 0.001?

Розділ 6
Чисельне диференціювання

Чисельне диференціювання застосовується, якщо функцію $f(x)$ важко чи неможливо продиференціювати, наприклад, якщо вона задана таблично. Воно також необхідне при розв'язанні диференціальних рівнянь за допомогою різницевих методів. Якщо маємо явний вигляд функції, то вираз для похідної часто виявляється досить складним і бажано його замінити більш простим. Якщо ж функція задана тільки в деяких точках (таблично), то одержати явний вигляд її похідних взагалі неможливо. У цих ситуаціях виникає необхідність наближеного (чисельного) диференціювання.

При чисельному диференціюванні функцію $f(x)$ апроксимують функцією $\varphi(x, a)$, що легко обчислюється, і вважають, що $y'(x) = \varphi'(x, a)$. При цьому можна використовувати різні способи апроксимації. Розглянемо найпростіший спосіб – апроксимацію інтерполяційним многочленом Ньютона.

Щоб побудувати формули чисельного диференціювання, задану на відрізку $[a, b]$ функцію $f(x)$ замінюють відповідним інтерполяційним многочленом $P(x)$. Тоді

$$f(x) = P(x) + R(x;f), \qquad (6.1)$$

де $R(x;f)$ — залишковий член інтерполяційної формули. Якщо функція $f(x) \in C^k_{[a,b]}$, то, диференціюючи (6.1), знаходимо

$$f'(x) = P'(x) + R'(x,f),$$
$$f''(x) = P''(x) + R''(x,f),$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$f^{(k)}(x) = P^{(k)}(x) + R^{(k)}(x;f).$$

Звідси отримуємо наближення

$$f'(x) \approx P'(x),\ f''(x) \approx P''(x), \ldots, f^{(k)}(x) = P^{(k)}(x), \qquad (6.2)$$

Тоді залишкові члени $r_i(x)$ (i=1,2, ... , k) формул чисельного диференціювання (6.2) дорівнюватимуть похідним від залишкового члена інтерполяційної формули (6.1), тобто

$$r_i(x) = f^{(i)}(x) - P^{(i)}(x). \qquad (6.3)$$

Варто зазначити, що з малості залишкового члена інтерполяційної формули $R(x;f)$ зовсім не випливає малість залишкових членів похідних (похибки чисельного диференціювання) $r_i(x)$, бо похідні від малих функцій можуть бути досить великими. Наприклад, функції $y_1(x)=f(x)$ і $y_2(x)= f(x) + \frac{\cos n^3 x}{n^2}$ для великих значень n можуть відрізнятися між собою як завгодно мало $|y_1 - y_2| \le \frac{|\cos n^3 x|}{n^2} \le \frac{1}{n^2}$.

Але похідні від них для деяких значень x і великих значень n можуть значно відрізнятися між собою:

$$|y_1' - y_2'| \le \frac{n^3 |-\sin n^3 x|}{n^2} = n\left|-\sin n^3 x\right| \le n,$$

$$|y_i'' - y_2''| \le n^4 \left|-\cos n^3 x\right| \le n^4.$$

Звідси бачимо, що не існує неперервної залежності значень похідної від значень функції. Тому задача чисельного диференціювання, загалом кажучи, - менш точна операція порівняно з інтерполюванням і є некоректною задачею.

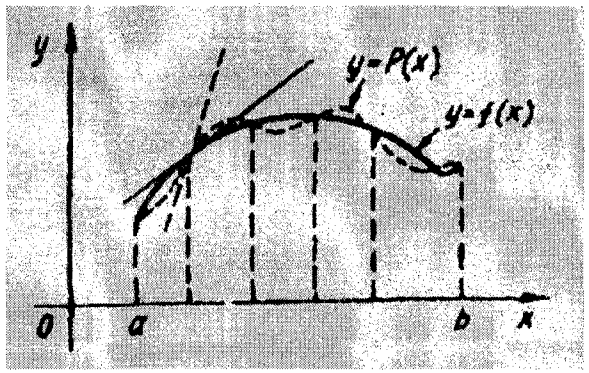


Рис. – 6.1.

На рис. 6.1 в точці x_i ординати функції f і многочлена P однакові, проте кутові коефіцієнти дотичних значно відрізняються.

Якщо інтерполяційний многочлен P на певній ділянці з достатньою точністю наближає функцію f, а сама функція f досить гладка і змінюється плавно на цій ділянці, то можна сподіватися, що при досить малому кроці інтерполювання похідні інтерполяційного многочлена також мало відрізнятимуться від похідних функції f. Проте не варто забувати, що зі зростанням порядку похідної точність чисельного диференціювання здебільшого різко спадає. Тому на практиці формули чисельного диференціювання для похідних, вищих від другого порядку, застосовують досить рідко.

Отже, функцію $f(x)$ треба продиференціювати кілька разів і знайти ці похідні в деякій точці.

6.1 Формули чисельного диференціювання

Розглянемо найпростіші формули чисельного диференціювання, що виводяться зазначеним способом.

Зупинимося на функції, що задана в рівновіддалених вузлах $x_i = x_0 + ih, h > 0, i = 0, \pm 1, \pm 2, \ldots$. Її значення і значення похідних у вузлах будемо позначати $f(x_i) = f_i$, $f'(x_i) = f_i'$, $f''(x_i) = f_i''$.

Нехай функція задана в двох точках x_0 і $x_1 = x_0 + h$, її значеннями є f_0, f_1.

Побудуємо інтерполяційний многочлен першого степеня $P_1(x) = f_0 + (x - x_0) f(x_0; x_1)$. Похідна $P_1'(x) = f(x_0; x_1) = \dfrac{f_1 - f_0}{h}$.

Похідну функції $f(x)$ в точці x_0 приблизно заміняємо похідною інтерполяційного многочлена

$$f_0'(x) \approx \frac{f_1 - f_0}{h}. \qquad (6.4)$$

Величина $\dfrac{f_1 - f_0}{h}$ називається першою різницевою похідною.

Нехай тепер $f(x)$ задана в трьох точках x_0, $x_1 = x_0 + h$, $x_{-1} = x_0 - h$. Інтерполяційний многочлен Ньютона другого степеня має вигляд

$$P_2(x) = f(x_0) + (x - x_0) f(x_0; x_1) + (x - x_0)(x - x_1) f(x_0; x_1; x_{-1}).$$

Тоді $P_2'(x) = f(x_0; x_1) + (2x - x_0 - x_1) f(x_0; x_1; x_{-1})$.

$$P_2'(x_0) = \frac{f_1 - f_0}{x_1 - x_0} + (x_0 - x_1) \times$$

$$\times \left[\frac{f_0}{(x_0 - x_1)(x_0 - x_{-1})} + \frac{f_1}{(x_1 - x_0)(x_1 - x_{-1})} + \frac{f_{-1}}{(x_{-1} - x_0)(x_{-1} - x_1)} \right] =$$

$$= \frac{f_1 - f_{-1}}{2h}.$$

Одержуємо наближену формулу

$$f_0' \approx \frac{f_1 - f_{-1}}{2h}. \qquad (6.5)$$

Величина $\dfrac{f_1 - f_{-1}}{2h}$ називається центральною різницевою похідною. Нарешті, для другої похідної

$$P_2''(x) = 2 f(x_0; x_1; x_{-1}) =$$

$$= 2 \left[\frac{f_0}{(x_0 - x_1)(x_0 - x_{-1})} + \frac{f_1}{(x_1 - x_0)(x_1 - x_{-1})} + \frac{f_{-1}}{(x_{-1} - x_0)(x_{-1} - x_1)} \right] =$$

$$= \frac{f_1 - 2f_0 + f_{-1}}{h^2},$$

одержуємо наближену формулу

$$f_0'' \approx \frac{f_1 - 2f_0 + f_{-1}}{h^2}. \qquad (6.6)$$

Величина $\dfrac{f_1 - 2f_0 + f_{-1}}{h^2}$ називається другою різницевою похідною.

Формули (6.4)-(6.6) називаються формулами чисельного диференціювання.

Вважаючи функцію f достатнє число разів неперервно диференційованою, одержимо похибки наближених формул (6.4)-(6.6). Надалі нам знадобляться такі леми.

Лема 1 Нехай $f \in C[a,b], \xi_i \in [a,b]$ – довільні точки, $i = \overline{1,n}$. Тоді існує така точка $\xi \in [a,b]$, що

$$\frac{f(\xi_1) + f(\xi_2) + \ldots + f(\xi_n)}{n} = f(\xi).$$

Доведення. Очевидна нерівність

$$\min_{[a,b]} f(x) \le \frac{f(\xi_1) + f(\xi_2) + \ldots + f(\xi_n)}{n} \le \max_{[a,b]} f(x).$$

За теоремою Больцано-Коши про проміжні значення неперервної функції на замкненому відрізку, вона набуває всіх значень між $\min\limits_{[a,b]} f(x)$ і $\max\limits_{[a,b]} f(x)$. Значить, існує така точка $\xi \in [a,b]$, що стверджує зазначену в лемі рівність.

Лема 2

1 Припустимо, що $f \in C^{(2)}[x_0, x_1]$. Тоді існує така точка ξ, що

$$f_0' = \frac{f_1 - f_0}{h} - \frac{h}{2} f'(\xi), \quad x_0 < \xi < x_1. \qquad (6.7)$$

2 Якщо $f \in C^{(3)}[x_{-1}, x_1]$, то є така точка ξ, що

$$f_0' = \frac{f_1 - f_{-1}}{2h} - \frac{h^2}{6} f'''(\xi), \quad x_{-1} < \xi < x_1. \qquad (6.8)$$

3 Коли $f \in C^{(4)}[x_{-1}, x_1]$, то є точка ξ така, що

$$f_0'' = \frac{f_{-1} - 2f_0 + f_1}{h^2} - \frac{h^2}{12} f^{(4)}(\xi), \quad x_{-1} < \xi < x_1. \qquad (6.9)$$

Доведення. Розглянемо розкладання

$$f_1 = f_0 + hf_0' + \frac{h^2}{2} f''(\xi), \qquad \xi \in (x_0, x_1),$$

звідки випливає (6.4). Якщо $f \in C^{(4)}[x_{-1}, x_1]$, то за формулою Тейлора

$$f_{\pm 1} = f_0 \pm hf_0' + \frac{h^2}{2} f_0'' \pm \frac{h^3}{6} f_0''' + \frac{h^4}{24} f^{(4)}(\xi_\pm), \qquad (6.10)$$

де $x_{-1} < \xi_- < \xi_+ < x_1$.

Підставимо (6.10) у вираз $\dfrac{f_{-1} - 2f_0 + f_1}{h^2}$. Одержимо

$$\frac{f_{-1} - 2f_0 + f_1}{h^2} = f_0'' + \frac{h^2}{24}\left[f^{(4)}(\xi_-) + f^{(4)}(\xi_+)\right].$$

Замінюючи відповідно до леми 1

$$f^{(4)}(\xi_-) + f^{(4)}(\xi_+) = 2f^{(4)}(\xi), \quad x_{-1} < \xi < x_1,$$

одержуємо $\dfrac{f_{-1} - 2f_0 + f_1}{h^2} = f_0'' + \dfrac{h^2}{12} f^{(4)}(\xi)$.

Звідки і випливає (6.9). Рівність (6.8) доводиться аналогічно. Формули (6.4)-(6.6) називаються формулами чисельного диференціювання із залишковими членами.

Похибки формул (6.4)-(6.6) оцінюються за допомогою наступних нерівностей, що випливають із співвідношень (6.7)-(6.9):

$$\left| f_0' - \frac{f_1 - f_0}{h} \right| \le \frac{h}{2} \max_{[x_0, x_1]} |f''(x)|,$$

$$\left| f_0' - \frac{f_1 - f_{-1}}{2h} \right| \le \frac{h^2}{6} \max_{[x_{-1}, x_1]} |f'''(x)|,$$

$$\left| f_0'' - \frac{f_{-1} - 2f_0 + f_1}{h^2} \right| \le \frac{h^2}{12} \max_{[x_{-1}, x_1]} |f^{(4)}(x)|.$$

Вважають, що похибка формули (6.4) має перший порядок відносно h, а похибка формул (6.5) і (6.6) другого порядку відносно h (чи порядку h^2). Також говорять, що формула чисельного диференціювання (6.4) першого порядку точності (відносно h), а формули (6.5) і (6.6) мають другий порядок точності.

Зазначеним способом можна одержувати формули чисельного диференціювання для старших похідних і для більшої кількості вузлів інтерполяції.

Для вибору оптимального кроку припустимо, що границя абсолютної похибки при обчисленні функції f в кожній точці задовольняє нерівність

$$\Delta f_i \le \bar{\Delta}. \qquad (6.11)$$

Нехай у деякому околі точки x_0 похідні, через які виражаються залишкові члени у формулах (6.8), (6.9), неперервні і задовольняють нерівності

$$|f'''(x)| \le \overline{M}_3, \quad |f^{(4)}(x)| \le \overline{M}_4, \qquad (6.12)$$

де $\overline{M}_3$, $\overline{M}_4$ - деякі числа. Тоді повна похибка формул (6.5), (6.6) (без урахування похибок округлення) відповідно до (6.8), (6.9), (6.11), (6.12) не перевершує відповідно величин

$$\varepsilon_1 = \frac{\overline{\Delta} + \overline{\Delta}}{2h} + \frac{h^2}{6}\overline{M}_3,$$

$$\varepsilon_2 = \frac{\overline{\Delta} + 2\overline{\Delta} + \overline{\Delta}}{h^2} + \frac{h^2}{12}\overline{M}_4.$$

Мінімізація за h цих величин приводить до таких значень кроків: $h_1 = \left(\frac{3\Delta}{\overline{M}_3}\right)^{1/3}, h_2 = 2\left(\frac{3\Delta}{\overline{M}_4}\right)^{1/4}$, при цьому

$$\varepsilon_1 = \frac{3}{2}\left(\frac{\overline{M}_3\Delta^{-2}}{3}\right)^{1/3}, \varepsilon_2 = 2\left(\frac{\overline{M}_4\overline{\Delta}}{3}\right)^{1/2}. \qquad (6.13)$$

Якщо при обраному для будь-якої з формул (6.5), (6.6) значенні h відрізок $[x_{-1}, x_1]$ не виходить за окіл точки x_0, у якому виконується нерівність (6.12), то знайдене h є оптимальним і повна похибка чисельного диференціювання оцінюється відповідною величиною (6.13).

6.2 Дослідження точності чисельного диференціювання

Дослідження точності отриманих виразів при чисельних розрахунках зручно робити за допомогою апостеріорної оцінки, за швидкістю спадання членів відповідного ряду Тейлора. Якщо крок сітки досить малий, то похибка близька до першого відкинутого члена.

У такий спосіб порядок точності результату відносно кроку сітки дорівнює числу залишених членів ряду, або, іншими словами, він дорівнює числу вузлів інтерполяції мінус порядок похідної. Тому мінімальне число вузлів необхідне для обчислення m-ої похідної, дорівнює $m+1$, воно забезпечує перший порядок точності.

Ці висновки відповідають принципу: *при почленному диференціюванні ряду швидкість його* збіжності зменшується.

Якщо врахувати погіршення збіжності ряду при диференціюванні, то можна зробити висновок: навіть якщо функція задана добре складеною таблицею на досить докладній сітці, то практично чисельним диференціюванням можна визначити першу і другу похідні, а третю і четверту – лише з великою похибкою. Похідні більш високого порядку рідко вдається обчислити із задовільною точністю.

6.2.1 Метод Рунге-Ромберга

Загальна ідея методу така: маємо деяку наближену формулу $\xi(x,\kappa)$ для обчислення величини $z(x)$ за її значеннями на рівномірній сітці з кроком h, а залишковий член цієї формули

$$z(x) - \xi(x,h) = \varphi(x)h^p + O(h^{p+1}). \qquad (6.14)$$

Наприклад. $z(x) = f'(x)$, $f(x)$ —задана функція. Нехай $f \in C^5[a,b]$, $\xi(x,h) = (f(x+h) - f(x-h))/(2h) \equiv f_x^0(x)$,

$$z(x) - \xi(x,h) = f'(x) - f_x^0(x) = -\frac{h^2}{6}f^{(3)}(x) - \frac{h^4}{60}f^{(5)}(\xi),$$

$\xi \in [x-h, x+h]$. Тут $p=2$. Якщо скористатися тією самою наближеною формулою для обчислення значення z в точці x, але використовуючи сітку з кроком rh, дістанемо

$$z(x)-\xi(x,rh)=\varphi(x)(rh)^p+O\left((rh)^{p+1}\right)=\varphi(x)(rh)^p+O\left(h^{p+1}\right) \quad (6.15)$$

Віднявши (6.14) від (6.15), дістанемо **першу формулу Рунге** для оцінки похибки

$$R\approx\varphi(h)h^p=\frac{\xi(x,h)-\xi(x,rh)}{r^p-1}+O\left(h^{p+1}\right). \quad (6.16)$$

Перший доданок у (6.16) є головним членом похибки, тобто розрахунок на другій сітці дає змогу оцінити похибки на першій сітці з точністю до членів вищого порядку. Виключаючи за допомогою (6.16) величину $\varphi(x)h^p$ з (6.14), дістанемо **другу формулу Рунге**

$$z(x)=\xi(x,h)+\frac{\xi(x,h)-\xi(x,rh)}{r^p-1}+O\left(h^{p+1}\right), \quad (6.17)$$

яка дає результат з вищим порядком точності, ніж (6.14). Іноді уточнення результату за формулою (6.17) називають **уточненням за Річардсоном**. Розглянемо приклади застосування описаного вище процесу для підвищення точності в задачі чисельного диференціювання.

Приклад 1 Нехай функція $y(x)=\lg x$ задана таблицею. Обчислити $y'(3)$.

x	$y=\lg x$
1	0,000
2	0,301
3	0,478
4	0,602
5	0,699

Розв'язання. Скориставшись формулою $f'(x_1)=\frac{f(x_2)-f(x_0)}{2h}+O(h^2)$ при $h=1$, дістанемо $y'(3)=[y(4)-y(2)]/2\approx 0{,}151$. Збільшуючи крок удвічі ($r=2$), дістанемо $y'(3)=[y(5)-y(1)]/(2\cdot 2)\approx 0{,}175$.

За формулою (6.16) при $p=2$ $y'(3)\approx 0{,}143$, що лише на 2% відрізняється від шуканого значення $y'(3)=0{,}145$.

Приклад 2 За допомогою методу Рунге вивести формулу чисельного диференціювання порядку $O(h^4)$ з формули більш низького порядку $O(h^2)$.

Розв'язання. Маємо

$$f'_{3/2}(h)\approx(f_2-f_1)/h, \quad f'_{3/2}(3h)\approx(f_3-f_0)/(3h).$$

Порядок точності цих формул $p=2$, а коефіцієнт збільшення кроку $r=3$, тому уточнення за методом Рунге дає формулу

$$f'_{3/2}\approx f'_{3/2}(h)+\frac{1}{8}[f'_{3/2}(h)-f'_{3/2}(3h)]=\frac{1}{24h}(f_0-27f_1+27f_2-f_3).$$

Як бачимо, для обчислення результату більш високого порядку точності не обов'язково використовувати безпосередньо формули високого порядку точності; можна виконати обчислення за простими формулами низької точності на різних сітках і потім уточнити результат за методом Рунге. Такий спосіб має перевагу ще й тому, що величина поправки (6.16) дає апостеріорну оцінку точності.

Метод Рунге узагальнюється на довільну кількість сіток.

Приклад 3 За допомогою розвинення в ряд Тейлора для функції $f(x)\in C^{2m+3}[x-a,x+a]$ і $|h|\le a$ дістаємо

$$z(x)-\xi(x,h)\equiv f'(x)-f_{\circ}(x)\equiv\psi_1(x)h^2+\psi_2(x)h^4+\ldots+$$
$$+\psi_m(x)h^{2m}+h^{2m+2}\tilde{\psi}_{m+1}(x,h).$$
$$\psi_k(x)=f^{(2k+1)}(x)/(2k+1)!$$
$$\tilde{\psi}_{m+1}(x,h)=\psi_{m+1}(x)+O(1),\quad k=\overline{1,m+1}. \qquad (6.18)$$

Приклад 4 Для односторонньої різницевої похідної $f_x(x)=(f(x+h)-f(x))/h\equiv\xi(x,h)$ при $f(x)\in C^{m+3}[x,\ x+a]$, $|h|\le a$ маємо

$$z(x)-\xi(x,h)\equiv f'(x)-f_x(x)=\psi_1(x)h+\psi_2(x)h^2+\ldots+$$
$$+\psi_m(x)h^m+h^{m+1}(\psi_{m+1}(x)+O(1)),$$
$$\psi_k(x)=-f^{(k+1)}(x)/(k+1)!,k=\overline{0,\ m+1}.$$

Нехай розрахунки виконано на q різних сітках h_j, $1\le j\le q$. Тоді із залишкового члена (6.18) можна вилучити $q-1$ складових. Для цього перепишемо (6.18) у вигляді

$$z(x)-\sum_{m=p}^{p+q-2}\psi_m(x)h_j^m=\xi(x,h_j)+O(h^{p+q-1}),\quad h_j\le h,\ 1\le j\le q.$$

Це система лінійних рівнянь відносно величин $z(x)$ і $\psi_m(x)$, $m=p,\ p+1,\ \ldots,\ p+q-2$. Використавши формули Крамера, дістанемо уточнений розв'язок за формулою Ромберга

$$z(x)=\Delta^{-1}\begin{vmatrix}\xi(x,h_1) & h_1^p & h_1^{p+1} & \ldots & h_1^{p+q-2}\\ \xi(x,h_2) & h_2^p & h_2^{p+1} & \ldots & h_2^{p+q-2}\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ \xi(x,h_q) & h_q^p & h_q^{p+1} & \ldots & h_q^{p+q-2}\end{vmatrix}+O(h^{p+q-1}), \qquad (6.19)$$

де $\Delta=\begin{vmatrix}1 & h_1^p & h_1^{p+1} & \ldots & h_1^{p+q-2}\\ 1 & h_2^p & h_2^{p+1} & \ldots & h_2^{p+q-2}\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ 1 & h_q^p & h_q^{p+1} & \ldots & h_q^{p+q-2}\end{vmatrix}$.

Ця формула виражає $z(x)$ через обчислені з точністю до $O(h^p)$ величини $\xi(x,h_i)$ з більш високою точністю $O(h^{p+q-1})$ (тобто розрахунок на кожній новій сітці дає змогу підвищити порядок точності на одиницю). Розкладаючи визначник за першим стовпчиком, формулу для $z(x)$ можна записати також у вигляді $z(x)=\sum_{i=1}^{q}\xi(x,h_i)L_{i,q}(0)+O(h^{p+q-1})$,

де $L_{i,q}(h)=\Delta^{-1}\begin{vmatrix}1 & h_1^p & h_1^{p+1} & \ldots & h_1^{p+q-2}\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ 1 & h_{i-1}^p & h_{i-1}^{p+1} & \ldots & h_{i-1}^{p+q-2}\\ 1 & h^p & h^{p+1} & \ldots & h^{p+q-2}\\ 1 & h_{i+1}^p & h_{i+1}^{p+1} & \ldots & h_{i+1}^{p+q-2}\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ 1 & h_q^p & h_q^{p+1} & \ldots & h_q^{p+q-2}\end{vmatrix}$.

Функції $L_{i,q}(h)$ мають, очевидно, такі дві властивості:

а) $L_{i,q}(h_j)=\delta_{i,j}$, $\delta_{i,j}$ - символ Кронекера;

б) $L_{i,q}(h)=a_0^{(i)}+a_1^{(i)}h^p+a_2^{(i)}h^{p+1}+\ldots+a_{q-1}^{(i)}h^{p+q-2}$ - дійсні коефіцієнти, тобто $L_{i,q}(h)$ є многочленами від h. Тому функція

$$P(h) \equiv P(h;\xi) = \sum_{i=1}^{q} \xi(x,h_i) L_{i,q}(h) \equiv \beta_0 + \beta_1 h^p + \ldots + \beta_{q-1} h^{p+q-2} \quad (6.20)$$

є інтерполюючою функцією для $\xi(x,h)$ (β_i - дійсні коефіцієнти), а величина $\sum_{i=1}^{q} \xi(x,h_i) L_{i,q}(0) \approx z(x)$

є значенням цієї функції при $h=0$, причому $h=0$ не належить найменшому інтервалу $l[h_1,h_2,\ldots,h_q]$, що охоплює всі точки $h_1,h_2,\ldots,h_q$. З цієї причини у випадку методу Рунге - Ромберга говорять також про екстраполяцію. Вживають також терміни «екстраполяція за Річардсоном», «екстраполяція до нуля», «екстраполяція до кроку нуль».

Оскільки система функцій $\varphi_0(h)=1, \varphi_1(h)=h^p, \ldots, \varphi_{q-1}(h)=h^{p+q-2}$ не при всіх p і не на довільному інтервалі буде системою Чебишева, то інтерполяційна функція (6.20) існує не для будь-якої послідовності $h_1,\ldots,h_q$. Але для послідовностей, які найчастіше трапляються на практиці, а саме:

а) $h_1,\ h_2=\alpha h_1,\ h_3=\alpha h_2=\alpha^2 h_1,\ \ldots$ (послідовність Рунге - Ромберга);

б) $h_1,\ h_2=h_1/2,\ h_3=h_1/3,\ \ldots$ можна довести, що $\Delta \neq 0$, і тим самим існування многочлена $P(x)$ гарантується.

Зауваження

1 Формула Рунге - Ромберга має ту перевагу, що вона може бути застосована для довільних кроків h_j та числа сіток q (за умови $\Delta \neq 0$). Недоліком її є те, що потрібно розв'язувати систему лінійних алгебраїчних

рівнянь і в проміжних розрахунках не контролюється точність.

2 Метод Ромберга можна застосувати не лише для розкладання вигляду $z(x)-\xi(x,h)=\sum_{m\geq 0}\psi_m(x)\varphi_m(h)$ з функціями $\varphi_m(h)=h^{p+m}$, як в (6.18), але й для довільних функцій $\varphi_m(h)$, $\varphi_m(0)=0$. Коли $\varphi_m(h)=h^{\gamma(m+1)}$ $(\gamma=const)$, що часто трапляється на практиці, тоді інтерполяційний многочлен $P(h)$ має вигляд

$$P(h) \equiv P(s) = \beta_0 + \beta_1 s + \beta_2 s^2 + \ldots + \beta_{q-1} s^{q-1}, \quad s = h^{\gamma}$$

і може бути обчислений в точці $s=0$ $(h=0)$ за допомогою, наприклад, алгоритму Ньютона без розв'язування системи.

3 Якщо сітки такі, що $h_j = r h_{j-1} = \ldots = r^{j-1} h_1$, тобто згущення їх відбувається за одну і ту саму кількість разів, то зручніше застосувати рекурентно метод Рунге. Це робиться таким чином. Спочатку на кожній парі сіток $(h_1,h_2),(h_2,h_3),\ldots,(h_{q-1},h_q)$ методом Рунге вилучають головний член похибки $\psi_q(x)h^p$. Уточнені значення, таким чином, групуються в пари і далі вилучається похибка наступного порядку $O(h^{p+1})$. Всього можна виконати $q-1$ уточнень. При кожному уточненні за формулою (6.16) обчислюється апостеріорна оцінка точності.

4 Якщо формула для обчислення $\xi(x,h)$ має симетричний вигляд, то на рівномірній сітці часто всі непарні члени ряду (6.18) перетворюються на нуль (див. приклад 3). У такій ситуації користуватися формулою (6.19) невигідно. Потрібно залишити в сумі (6.18) члени $\psi_p h^p;\ \psi_{p+2}h^{p+2},\ \psi_{p+4}h^{p+4},\ldots$ і відповідно змінити формулу

Ромберга. Те саме стосується і рекурентної процедури Рунге - при черговому вилученні порядок точності підвищується на 2, а не на 1.

5 Число членів суми (6.18) пов'язане з кількістю неперервних похідних у функції, для якої обчислюються $z(x)$ і $\xi(x,h)$ (див. приклади 4, 5). Для не досить гладких функцій недоцільно брати велике число сіток для уточнення. Практично навіть для гладких функцій використовують не більше 3 – 5 сіток, причому, як правило, беруть відношення r кроків сіток, що дорівнює 2.

6 Метод Рунге - Ромберга можна застосувати лише тоді, коли правильне (6.18), причому коефіцієнти $\psi_m(x)$ однакові для всіх сіток. Для формул чисельного диференціювання ці коефіцієнти залежать від положення вузлів сітки. Але якщо вибрані конфігурації вузлів на всіх сітках подібні відносно точки x, то залежність від вузлів однакова. У такому разі метод Рунге - Ромберга можна застосувати, в інших випадках його застосувати неможливо. Тому при чисельному диференціюванні метод Рунге - Ромберга застосовується лише для знаходження похідних у вузлах або в середніх точках інтервалів рівномірних і на деяких «близьких» до них сітках. Це так звані квазірівномірні сітки, які добирають так, щоб «найкращим чином» передати поведінку конкретної функції. Сітка (у змінних x) називається *квазірівномірною,* якщо існує двічі неперервно диференційована функція $x=\xi(t)$, яка переводить відрізок $0\le t\le 1$ у відрізок $a\le x\le b$ так, що кожній сітці $x_i^{(N)}$ відповідає рівномірна сітка $t_i^{(N)}=i/N$, причому на цьому відрізку $\xi'(t)\ge\varepsilon>0$, а $\xi''(t)$ обмежена. Якщо ці умови виконано, то крок сітки $h_i\approx\xi'(t_i)/N$, а

різниця двох сусідніх кроків $h_i-h_{i-1}\approx\xi''(t_i)/N^2$, тобто при значній кількості вузлів різниця сусідніх кроків є величина порядку $O(h^2)$ і сусідні інтервали майже рівні (хоча відношення довжин далеких один від одного інтервалів $h_i/h_j\approx\xi'(t_i)/\xi'(t_j)$ може бути великим).

6.2.2 Процес Ейткена

Метод розрахунків на декількох сітках застосовується для підвищення порядку точності і в тому випадку, коли невідомий порядок головного члена похибки. Він має назву *процесу Ейткена.* Нехай

$$z(x)-\xi(x,h)=\psi_p(x)h^p+\psi_q(x)h^q+\ldots,\quad q>p,$$

але p - невідоме. Проводяться обчислення на трьох сітках з кроками

$$h_1=h,\ h_2=\rho h,\ h_3=\rho^2 h\ \ (0<\rho<1).$$

Нехтуючи членами порядку $O(h^q)$, дістаємо

$$A=\frac{\xi(x,h_1)-\xi(x,h_2)}{\xi(x,h_2)-\xi(x,h_3)}\approx\frac{h_1^p-h_2^p}{h_2^p-h_3^p}=\frac{1-\rho^p}{\rho^p(1-\rho^p)}=\left(\frac{1}{\rho}\right)^p.$$

Звідси знаходимо $p\approx\dfrac{\ln A}{\ln(1/\rho)}$. Далі можна скористатися вже відомим методом Рунге, який можна трактувати таким чином. Утворимо комбінацію $\tilde{\xi}(x,h)=\sigma\xi(x,h_1)+(1-\sigma)\xi(x,h_2)$ і виберемо σ так, щоб

$$\sigma h_1^p+(1-\sigma)h_2^p=(\sigma+(1-\sigma)\rho^p)h^p=0.$$

Дістанемо $\sigma=\rho^p/(\rho^p-1)=1/(1-A)$, причому

$$\tilde{\xi}(x,h)-z(x)=O(h^q).$$

Питання і завдання до розділу 6

1 Формули , які можна використати для чисельного диференціювання функції, що задана таблично.

2 Апріорна оцінка похибки чисельного диференціювання.

3 Апостеріорна оцінка похибки чисельного диференціювання.

4 Екстраполяція за Річардсоном в чисельному диференціюванні.

5 Скласти таблицю наближених значень похідної функції $y = 2^x$ за таблицею її значень

x		1.1	1.2	1.3
y		2. 144	2. 297	2. 462

6 Переконатися в тім, що формула чисельного диференціювання $f''(x) \approx (f(x-h) - 2f(x) + f(x+h))/h^2$ має другий порядок точності.

7 Вивести формули чисельного диференціювання на основі лінійної інтерполяції.

8 Функція $y = f(x)$ задана таблично

x	1	2	3	4	5	6
y	1.34	3.05	5.21	7.22	9.11	12.34

Обчислити $f'(4.5)$.

Розділ 7

Чисельне інтегрування функцій

Задача чисельного інтегрування функції полягає в обчисленні наближеного значення визначеного інтеграла

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

з використанням значень підінтегральної функції $\{f(x)|_{x=x_k} = f(x_k) = y_k\}$ у вузлах сітки $\{a = x_0 < x_1 < \ldots < x_n = b\}$. Визначений інтеграл I представляє площу криволінійної трапеції, обмеженої кривою $y = f(x)$, віссю Ox та прямими $x = a$ та $x = b$.

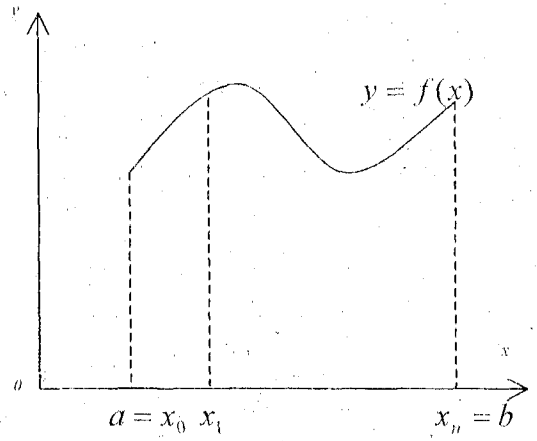


Рис. – 7.1

У практичних розрахунках нерідко виникає потреба в обчисленні визначених інтегралів вигляду

$$F = \int_a^b p(x)f(x)dx \quad (p(x) > 0),$$

де функція $f(x)$ та вагова функція $p(x)$ неперервні на відрізку $[a,b]$.

До чисельного інтегрування вдаються тоді, коли інтеграл неможливо виразити через елементарні функції або ж функція $f(x)$ задана таблично, а також коли внаслідок інтегрування одержано незручний для використання вираз. Тоді $f(x)$ наближають більш зручною функцією $g(x;C)$.

Найчастіше підінтегральну функцію $f(x)$ заміняють на деякий узагальнений поліном. Тоді внаслідок лінійності такої апроксимації функцію $f(x)$ можна записати так:

$$f(x)=\sum_{i=1}^{n} f(x_i)g_i(x)+r(x),$$

де $r(x)$- залишковий член апроксимації. Підставивши вираз $f(x)=\sum_{i=1}^{n} f(x_i)g_i(x)+r(x)$ у формулу $F=\int_a^b p(x)f(x)dx$, одержимо загальну формулу чисельного інтегрування – квадратурну формулу

$$F=\sum_{i=1}^{n} C_i f(x_i)+R;\quad C_i=\int_a^b p(x)g_i(x)dx;\qquad R=\int_a^b p(x)r(x)dx,$$

де x_i – вузли; C_i- ваги; R – похибка або залишковий член квадратурної формули.

Отже, інтеграл наближено замінено на суму, подібну до інтегральної, причому як вузли, так і коефіцієнти (ваги) квадратурної формули не залежать від функції $f(x)$.

Будемо будувати формулу чисельного інтегрування за правилом $\int_a^b f(x)dx\approx\sum_{i=1}^{n} C_i f(x_i)$.

Це відношення називається квадратурною формулою. При цьому: права частина виразу $\int_a^b f(x)dx\approx\sum_{i=1}^{n} C_i f(x_i)$ називається квадратурною сумою. Тут $\{C_i,x_i,n\}$ – параметри квадратурної

формули: C_i – квадратурні (вагові) коефіцієнти; x_i – квадратурні вузли.

Якщо межі інтегрування a,b являються квадратурними вузлами, то отримуємо формулу замкненого типу. Інакше маємо квадратурну формулу відкритого типу.

Величина $R_n(f)=\int_a^b f(x)dx-\sum_{i=1}^{n} C_i f(x_i)$ називається похибкою квадратурної формули $\int_a^b f(x)dx\approx\sum_{i=1}^{n} C_i f(x_i)$.

Якщо для деякої функції $f(x)$ маємо $R_n(f)=0$, то квадратурна формула являється для даної функції точною.

Квадратурна формула $\int_a^b f(x)dx\approx\sum_{i=1}^{n} C_i f(x_i)$ має алгебраїчний степінь точності m, якщо вона є точною при $f(x)=x^k$, $k=1,...,m$ і не точною при $f(x)=x^{m+1}$ $(R_n(x^k)=0,\ k=1,...,m,\ R_n(x^{m+1})\neq 0)$.

Звідси очевидно, що квадратурна формула степеня точності m є точною для всіх алгебраїчних многочленів степеня не вище за m, причому число m - максимальний степінь таких многочленів.

Визначимо верхню оцінку точності для формули $\int_a^b f(x)dx\approx\sum_{i=1}^{n} C_i f(x_i)$ при фіксованому n.

Лема. Степінь точності формули $\int_a^b f(x)dx\approx\sum_{i=1}^{n} C_i f(x_i)$ не може бути вище за $2n-1$ при будь-якому виборі параметрів $\{C_i,x_i,\},\ i=1,...,n$.

Доведення Розглянемо довільну квадратурну формулу $\int_a^b f(x)dx\approx\sum_{i=1}^{n} C_i f(x_i)$. Нехай $f(x)=(x-x_1)^2...(x-x_n)^2$. Це

многочлен степеня $2n$. Оскільки $f(x) \geq 0,\ x \in [a,b]$, то $\int f(x)dx > 0$. З іншого боку, $\sum_{i=1}^{n} C_i f(x_i) = 0$, тобто формула $\int_a^b f(x)dx \approx \sum_{i=1}^{n} C_i f(x_i)$ в даному випадку не є точною.

Формули чисельного обчислення однократного інтеграла називаються квадратурними формулами, подвійного й більшої кратності - кубатурними.

Наближеним значенням інтеграла будемо вважати вираз $I_n = \sum_{i=1}^{n} l_i(f)$, де $l_i(f)$ – наближене значення інтеграла на частковому відрізку $[x_{i-1}, x_i]$. При цьому формула для обчислення $l_i(f)$ називається *найпростішою квадратурною формулою*, а формула для обчислення I_n – *складеною квадратурною формулою*.

7.1 Квадратурні формули Ньютона-Котеса

Розглянемо формули для наближеного обчислення інтегралів

$$F = \int_a^b p(x) f(x)dx. \qquad (7.1)$$

Обмежимося випадком, коли $p(x) \equiv 1$. Цей метод заснований на заміні підінтегральної функції інтерполяційним многочленом Лагранжа з вузлами, що розбивають відрізок $[a,b]$ на рівні частини. Такі формули називаються формулами Ньютона-Котеса.

Отже, нехай задана рівномірна сітка $\{a = x_0 < x_1 < \ldots < x_n = b\}$, $x_i = x_0 + ih$, $i = \overline{0,n}$. Тобто крок $h = \dfrac{b-a}{n}$ – величина постійна й розбиває відрізок $[a,b]$ на n

рівних інтервалів. Формули Ньютона-Котеса - формули замкненого типу. Позначимо $y_i = f(x_i)$. За наближену функцію $g(x)$ оберемо інтерполяційний поліном Лагранжа

$$g(x) = L_n(x) = \sum_{i=0}^{n} y_i(x)\varphi_i(x) = y_0(x)\varphi_0(x) + y_1(x)\varphi_1(x) + \ldots + y_n(x)\varphi_n(x),$$

де $\varphi_i(x) = \dfrac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}$.

Отже, заданий інтеграл може бути поданий у вигляді

$$F = \int_a^b f(x)dx = \int_a^b L_n(x)dx + R_n(f) = \int_a^b \varphi_i(x) f(x_i)dx + R_n(f) =$$

$$= \sum_{i=0}^{n} C_i f(x_i) + R_n(f).$$

Таку квадратурну формулу називають квадратурною формулою інтерполяційного типу.

Нехай $\dfrac{x-x_0}{h} = q$ – виражена в сіткових кроках довжина $x - x_0$. Тоді

$$(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n) =$$

$$= h^{n+1}\left(\frac{x-x_0}{h}\right)\left(\frac{x-(x_0+h)}{h}\right)\ldots\left(\frac{x-(x_0+nh)}{h}\right) = h^{n+1}q(q-1)\ldots(q-n);$$

$$(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n) =$$

$$= h^n\left(\frac{x_i-x_0}{h}\right)\left(\frac{x_i-(x_0+h)}{h}\right)\ldots\left(\frac{x_i-x_{i-1}}{h}\right)\left(\frac{x_i-x_{i+1}}{h}\right)\ldots\left(\frac{x_i-(x_0+nh)}{h}\right) =$$

$$= h^n i(i-1)\cdot\ldots\cdot 1\cdot(-1)\cdot(-2)\cdot\ldots(-(n-i)) = (-1)^{n-i}\cdot h^n\cdot i!\cdot(n-i)!.$$

У такому випадку ваги можна розрахувати так:

$$C_i = \int_a^b \varphi_i(x)dx = \frac{(-1)^{n-i}}{i!(n-i)!}\cdot\frac{h^{n+1}}{h^{n+1}}\int_a^b \frac{q(q-1)...(q-n)}{q-i}dx = \left\{\begin{array}{l}\frac{x-x_0}{h}=q\\ \frac{dx}{h}=dq\end{array}\right\} =$$

$$= \frac{(-1)^{n-i}}{i!(n-i)!}\cdot h\cdot\int_0^n \frac{q(q-1)...(q-n)}{q-i}dq. \qquad (7.2)$$

Формула (7.2) остаточно визначає ваги квадратурної формули Ньютона-Котеса. Замінимо в ній $h=\frac{b-a}{n}$ і введемо позначення $C_i=(b-a)K_i$. Тоді коефіцієнти

$$K_i = \frac{(-1)^{n-i}}{i!(n-i)!}\cdot\frac{1}{n}\int_0^n \frac{q(q-1)...(q-n)}{q-i}dq \qquad (7.3)$$

називаються коефіцієнтами Котеса. А сама квадратурна формула Ньютона-Котеса набирає вигляду

$$\int_a^b f(x)dx = (b-a)\sum_{i=0}^n K_i f(x_i) + R_n(f). \qquad (7.4)$$

Для коефіцієнтів Котеса мають місце співвідношення:

$$1 \quad \sum_{i=0}^n K_i = 1.$$

$$2 \quad K_i = K_{n-i}.$$

З'ясуємо питання про степінь точності квадратурної формули $\int_a^b f(x)dx = \sum_{i=1}^n C_i f(x_i) + R_n(f)$.

Нехай $f(x)$ – алгебраїчний многочлен степеня не вищого за $n-1$. Тоді, згідно з властивостями інтерполяції $L_{n-1}(x)\equiv f(x)$, тобто $R_n(f)=0$. Таким чином, інтерполяційна

квадратурна формула $\int_a^b f(x)dx = \sum_{i=1}^n C_i f(x_i) + R_n(f)$ має степінь точності не нижчий за $n-1$.

Звідси можна зробити висновок, що квадратурні коефіцієнти C_i формули $\int_a^b f(x)dx = \sum_{i=1}^n C_i f(x_i) + R_n(f)$ є єдиним розв'язком лінійної системи рівнянь

$$\sum_{i=1}^n C_i x_i^k = \int_a^b x^k dx, \quad k=0,...,n-1,$$

яка отримана із $\int_a^b f(x)dx = \sum_{i=1}^n C_i f(x_i) + R_n(f)$ при $f(x)=x^k$.

Розглянемо окремі випадки квадратурних формул Ньютона-Котеса з рівновіддаленими вузлами, в яких підінтегральна функція $f(x)$ замінена на інтерполяційний поліном Лагранжа різного степеня.

7.1.1 Формула середніх (формула прямокутників)

Якщо на відрізку $[a,b]$ взяти єдиний вузол квадратурної формули x_0, то підінтегральна функція $f(x)$ апроксимується поліномом нульового степеня – сталою $f(x_0)$. У зв'язку з тим, що симетричне розміщення вузлів у чисельному диференціюванні привело до підвищення точності, за вузол x_0 візьмемо середину відрізка інтегрування $\bar{x}=\frac{a+b}{2}$. Замінивши наближено площу криволінійної трапеції на площу прямокутника з висотою $f(x)$ та основою $(b-a)$, одержимо формулу середніх

$$F = \int_a^b f(x)dx \approx f(\bar{x})(b-a) \quad (\bar{x}=\frac{a+b}{2}).$$

Це найпростіша квадратурна формула. Розклавши $f(x)$ у ряд Тейлора довкола точки $\bar{x}$

$$f(x) = f(\bar{x}) + (x-\bar{x})f'(\bar{x}) + \frac{(x-\bar{x})^2 f''(\bar{x})}{2} + \dots$$

і підставивши цей ряд в інтеграл, одержимо значення похибки формули середніх

$$R = \int_a^b f(x)dx - f(\bar{x})(b-a) \approx \frac{(b-a)^3 f''(\bar{x})}{24}.$$

Формула середніх є точною для лінійної підінтегральної функції $f(x)$, оскільки тоді $f''(x) = 0$.

Природно, що точність формули для довільної $f(x)$ можна підвищити, якщо скористатися докладнішою сіткою $x_i (i = \overline{0,N})$:

$$F \approx \sum_{i=1}^{N} (x_i - x_{i-1}) f(\bar{x}_i);\ R \approx \frac{1}{24}\sum (x_i - x_{i-1})^3 f''(\bar{x}_i)\ \ (\bar{x}_i = \frac{x_i + x_{i-1}}{2}).$$

Це так звана складена формула середніх або формула прямокутників. У разі рівномірної сітки, тобто якщо $x_i - x_{i-1} = h = const$, формула виглядатиме так:

$$F_h^C \approx h\sum_{i=1}^{N} f_{i-1/2};\quad f_{i-1/2} = f\left(x_i - \frac{h}{2}\right);$$

$$R \approx \frac{h^2}{24}\int_a^b f''(x)dx \approx \frac{h^3}{24}\sum_{i=1}^{N} f''(\bar{x}_i) \approx O(h^2).$$

Наведені оцінки R справедливі, якщо існує неперервна $f''(x)$; якщо ж $f''(x)$ кусково-неперервна, має місце лише мажорантна оцінка

$$R \le \frac{(b-a)h^2 M_2}{24}\ \ (M_2 = \max_{[a,b]} |f''(x)|).$$

7.1.2 Формула трапецій

Замінимо функцію $f(x)$ на відрізку $[a,b]$ інтерполяційним поліномом Лагранжа першого степеня з

вузлами $x = a,\ x = b$, що відповідає заміні кривої $f(x)$ на січну. Тоді значення шуканого інтеграла (площу криволінійної трапеції) можна наближено замінити на площу трапеції з висотою $(b-a)$ та основами $f(a), f(b)$. Отже, формула трапеції матиме вигляд

$$F = \int_a^b f(x)dx \approx (b-a)\frac{f(a)+f(b)}{2}.$$

Формула трапеції буде точною для лінійної підінтегральної функції з тієї самої причини, що й формула середніх.

На докладнішій сітці $x_i (i = \overline{0,N})$ одержимо складену формулу трапецій:

$$F \approx \frac{1}{2}\sum_{i=1}^{N} h_i (f_i + f_{i-1});\quad [h_i = x_i - x_{i-1},\ \ f_i = f(x_i)].$$

На рівномірній сітці вона стає такою:

$$F_h^T \approx h\left(\frac{f_0 + f_N}{2} + \sum_{i=1}^{N-1} f_i\right).$$

Зазначимо, що для одержання залишкового члена формули трапеції потрібно замінити чисельний коефіцієнт (1/24) у залишковому члені формули середніх на (-1/24).

7.1.3 Формула Симпсона

Квадратурна формула Симпсона є частковим випадком квадратурних формул Ньютона-Котеса при $n = 2$. Тут підінтегральна функція заміняється інтерполяційним поліномом Лагранжа 2-го степеня (рисунок 7.2). Із цієї причини формулу Симпсона ще називають формулою парабол.

Розіб'ємо відрізок $[a,b]$ на 2 рівних відрізки і одержимо сітку $\{a = x_0, x_1, x_2 = b\}$, що містить три вузли. Формула Симпсона містить три коефіцієнти Котеса:

$$K_0 = K_2;\quad K_0 + K_1 + K_2 = 1.$$

$$K_0 = \frac{1}{2}\cdot\frac{(-1)^{2-0}}{0!(2-0)!}\int_0^2 \frac{q(q-1)(q-2)}{q}dq = \frac{1}{4}\int_0^2 (q^2-3q+2)dq =$$

$$= \frac{1}{4}\left(\frac{q^3}{3}-\frac{3q^2}{2}+2q\right)\Big|_0^2 = \frac{1}{6}$$

$$K_0 = K_2 = \frac{1}{6};\ K_1 = 1-\frac{2}{6} = \frac{4}{6}.$$

$$\int_{x_0}^{x_2} f(x)dx = (x_2-x_0)\left(\frac{1}{6}y_0+\frac{4}{6}y_1+\frac{1}{6}y_2\right)+R_{симп} = \tag{7.5}$$

$$= \frac{h}{3}(y_0+4y_1+y_2)+R_{симп}$$

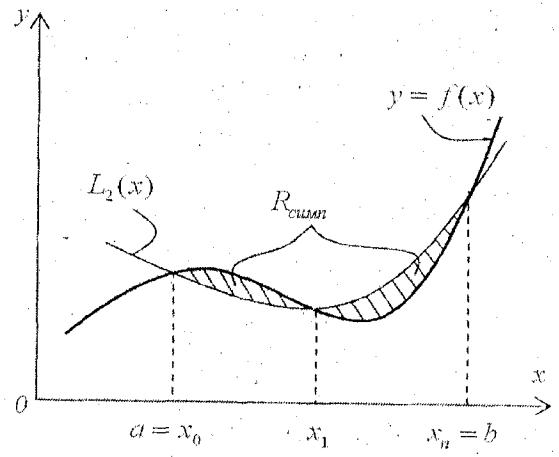


Рис. – 7.2

Формула (7.5) є трьохточковою квадратурною формулою Симпсона.

Якщо $[a,b]$ розбити на парну кількість відрізків, що дорівнює $n=2m$, і до кожного часткового здвоєного проміжку

$[x_0,x_2],[x_2,x_4],\ldots,[x_{2m-2},x_{2m}]$ застосувати формулу Симпсона, то одержимо складену формулу Симпсона

$$\int_a^b f(x)dx = \sum_{i=1}^{m}\int_{x_{2i-2}}^{x_{2i}} f(x)dx = \frac{h}{3}(y_0+4y_1+y_2)+\frac{h}{3}(y_2+4y_3+y_4)+\ldots$$

$$+\frac{h}{3}(y_{2m-2}+4y_{2m-1}+y_{2m})+R_{симп} =$$

$$= \frac{h}{3}\left(f(a)+2\sum_{i=1}^{m-1}f(x_{2i})+4\sum_{i=0}^{m-1}f(x_{2i+1})+f(b)\right)+R_{симп}. \tag{7.6}$$

Урахувавши, що головні члени похибок у формулі середніх та формулі трапеції одного порядку, але різних знаків, можна одержати точнішу квадратурну формулу. Для цього скомбінуємо ці формули так, щоб головний член сумарної похибки цих квадратурних формул перетворився на нуль, тобто

$$F \approx \frac{2F^c+F^T}{3} = \frac{2f(x)(b-a)+(b-a)[f(a)+f(b)]/2}{3}.$$

Отже, дійдемо формули парабол

$$F \approx \frac{(b-a)}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right].$$

Формула парабол є точною для кубічної підінтегральної функції $f(x)$, оскільки в похибку R входитиме $f^{IV}(x)$, а вона для такої підінтегральної функції дорівнює нулю.

Ця формула має такий залишковий член:

$$R \le -\frac{h^4(b-a)}{180}M_4 \approx O(h^4),\ M_4 = \max_{[a,b]}|f^{IV}(x)|,$$

тобто формула парабол має 4-й порядок похибки, а чисельний коефіцієнт досить малий. Через ці обставини формула парабол дає добру точність за відносно невеликого числа вузлів, якщо $f^{IV}(x)$ не дуже велика.

Приклад реалізації алгоритму чисельного інтегрування функції одного аргумента за формулою Симпсона на псевдокоді

```
f(x):
```

```
    //Повертає значення підінтегральної функції
end
//Повертає одну суму з формули Симпсона
//a - лівий кінець відрізка інтегрування
//h - крок
sum1(n,h,a):
    1    temp:=0
    2    for i:=1 to (an div 2) do
    3        temp+=f(a+(2*k-1)*ah)
    4    done
    5    return temp
end
//Повертає одну суму з формули Симпсона
//a - лівий кінець відрізка інтегрування
//h - крок
sum2(n,h,a):
    1    temp:=0;
    2    for k:=2 to (an div 2) do
    3        temp+=f(a+(2*k-2)*ah)
    4    done
    5    return temp
end
Calc_Integrate_Simpson(a,b,
    1    n:=4
    2    h:=(b-a)/n
```



```
    3    repeat
    4    I1:=(h*(f(a)+f(b)+4*sum1(n,h,a)
+2*sum2(n,h,a)))/3;
    5        n:=2*n;
    6        h:=(b-a)/n;
    7        I2:=(h*(f(a)+f(b)+4*sum1(n,
h,a)+2*sum2(n,h,a)))/3
    8        m:=abs(I1-I2)
    9    until (m<eps)
    10   return I2
end
```

7.2 Квадратурна формула Гауса

Загальний підхід для побудови квадратурної формули для інтегралів $F=\int_a^b \rho(x)f(x)dx \quad (\rho)>0)$ полягає у виборі параметрів $\{A_i,x_i\}$, $i=1,...,n$ (n – фіксоване) так, щоб забезпечити максимально можливий степінь точності. Квадратурна формула з такою властивістю носить назву формули Гауса. У розглянутих квадратурних формулах вибирали і знаходили вузли та ваги, а отже, тим самим не було використано всі можливості загальної квадратурної формули.

К.Ф.Гаус звернув увагу, що квадратурна формула має $2n$ невідомих параметрів C_i та x_i, тобто саме стільки, скільки параметрів має алгебраїчний поліном степеня $m=2n-1$. Він запропонував підбирати ці параметри так, щоб квадратурна формула була точною для підінтегральної функції $f(x)$ у вигляді полінома степеня, не вищого за m.

Спочатку для спрощення розглянемо відрізок $[-1;1]$, тобто інтеграл вигляду

$$\Phi = \int_{-1}^{1} \rho(\xi) f(\xi) d\xi \approx \sum_{i=1}^{N} \gamma_i f(\xi_i) \qquad (7.7)$$

Отже, знайдемо параметри γ_i, ξ_i з таких умов:

$$\sum_{i=1}^{n} \gamma_i \xi_i^j = \int_{-1}^{1} \rho(\xi) \xi^j d\xi \qquad (j = \overline{0,m}). \qquad (7.8)$$

Це система $2n$ нелінійних алгебраїчних рівнянь відносно γ_i, ξ_i $(i = \overline{1,n})$.

Для подальшого спрощення вважатимемо, що $\rho(\xi) \equiv 1$.

Якщо $n = 1$ одержимо $m = 1$ і система $\sum_{i=1}^{n} \gamma_i \xi_i^j = \int_{-1}^{1} \rho(\xi) \xi^j d\xi$ $(j = \overline{0,m})$ набере вигляду

$$j = 0: \ \gamma_1 = \int_{-1}^{1} d\xi = 2;$$

$$j = 1: \ \gamma_1 \xi_1 = \int_{-1}^{1} \xi d\xi = 0;$$

із другого рівняння випливає, що $\xi_1 = 0$, тобто дійшли відомої формули середніх для відрізка $[-1,1]$

$$\Phi = \int_{-1}^{1} f(\xi) d\xi \approx 2 f(0),$$

яка є точною для будь-якого полінома 1-го степеня.

Якщо $n = 2$, система $\sum_{i=1}^{n} \gamma_i \xi_i^j = \int_{-1}^{1} \rho(\xi) \xi^j d\xi$ $(j = \overline{0,m})$ матиме такий вигляд ($m = 3$):

$$j = 0: \ \gamma_1 + \gamma_2 = 2;$$

$$j = 1: \ \gamma_1 \xi_1 + \gamma_2 \xi_2 = 0;$$

$$j = 2: \ \gamma_1 \xi_1^2 + \gamma_2 \xi_2^2 = \frac{2}{3};$$

$$j = 3: \ \gamma_1 \xi_1^3 + \gamma_2 \xi_2^3 = 0.$$

Розв'язавши цю систему, знайдемо:

$$\gamma_1 = \gamma_2 = 1; \ \xi_2 = -\xi_1 = \sqrt{3}/3 = 0{,}5773502692,$$

тобто маємо квадратурну формулу

$$\Phi = \int_{-1}^{1} f(\xi) d\xi \approx f\left(\frac{-1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right),$$

яка є точною для будь-якого полінома 3-го степеня.

За довільного n як вузли квадратурної формули Гауса беруть нулі поліномів Лежандра

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

а ваги цієї квадратурної формули визначають за таким виразом:

$$\gamma_i = \frac{2}{(1 - \xi_i^2)[P_n'(\xi_i)]^2} \qquad (i = \overline{1,n}). \qquad (7.9)$$

Маючи значення γ_i та вузлів ξ_i на відрізку $[-1,1]$, значення інтеграла на довільному відрізку $[a,b]$ обчислюється за такою квадратурною формулою Гауса:

$$F \approx \frac{b-a}{2} \sum_{i=0}^{n} \gamma_i f(x_i); \qquad x_i = \frac{a+b}{2} + \frac{(b-a)\xi_i}{2} \qquad (i = \overline{1,n}).$$

Похибка квадратурної формули Гауса має вигляд

$$\max |R| \approx \frac{b-a}{2{,}5\sqrt{n}} \left(\frac{b-a}{3n}\right)^{2n} M_{2n}, \quad (M_{2n} = \max_{[a,b]} |f^{(2n)}(x)|.$$

Зауважимо, що, починаючи з $n = 4$, і вузли, і ваги є ірраціональними числами, а кінці a і b ніколи не входять до вузлів.

Іншими прикладами квадратурних формул типу Гауса є формули Чебишева, Ерміта.

7.2.1 Квадратурна формула Чебишева

Візьмемо за основу формулу $\int_a^b f(x) dx \approx \sum_{i=1}^{n} A_i f(x_i)$ і будемо вважати всі квадратурні коефіцієнти однаковими: $A_1 = A_2 = \ldots = A_n = A$. Тоді

$$\int_a^b f(x)dx \approx A\sum_{i-1}^{n} f(x_i). \qquad (7.10)$$

Параметри $A, x_i \ \ i = 1,...,n$ виберемо так, щоб формула $\int_a^b f(x)dx \approx A\sum_{i-1}^{n} f(x_i)$ була точною для всіх поліномів степеня не вище за n. При цьому достатньо розглянути функції $f(x) = x^k, \ k = 0,...,n.$

Для $f(x) = 1$ маємо $b - a = An, \quad A = \dfrac{b-a}{n}.$

Покладаючи в $\int_a^b f(x)dx \approx \sum_{i=1}^{n} A_i f(x_i) \quad f(x) = x^k, \ k = 1,...,n,$ приходимо до системи нелінійних рівнянь для визначення квадратурних вузлів $x_i, \ i = 1,...,n$

$$\frac{b-a}{n}\sum_{i=1}^{n} x_i^k = \frac{b^{k+1} - a^{k+1}}{k+1}, \ k = 1,...,n.$$

Отримана квадратурна формула називається формулою Чебишева. Зокрема при $n = 1$

$$x_1 = \frac{a+b}{2} \Rightarrow \int_a^b f(x)dx \approx f(\frac{a+b}{2})(b-a).$$

7.3 Стійкість квадратурного процесу. Оцінки похибки

Стійкість квадратурних формул характеризує їх чутливість до різного роду похибок. Вона безпосередньо пов'язана з поняттям збіжності квадратурних формул.

Квадратурна формула буде збіжною за умови, що залишок $R_n(f) \to 0$ при $n \to \infty$.

Крім похибки, що виникає внаслідок відкидання залишкового члена (похибки методу), виникає похибка, зумовлена виконанням дій з наближеними числами (у процесі обчислень

майже завжди доводиться мати справу з наближеними значеннями $f(x_k)$, в яких правильні тільки кілька значущих цифр). Нехай, наприклад, всі значення $f(x_k)$ обчислені наближено, причому абсолютні похибки їх не перевищують числа Δf. Обчисливши за допомогою наближених значень $f(x_k)$ квадратурну суму $\sum_{k=0}^{n} C_k f(x_k)$ при точних значеннях C_k, дістанемо похибку

$$\Delta_\Sigma = \Delta f \sum_{k=0}^{n} |C_k|. \qquad (7.11)$$

Це неусувна похибка квадратурної формули.

Отже, якщо сума $\sum_{k=0}^{n} |C_k|$ велика, то навіть незначні похибки в значеннях $f(x_k)$ можуть призвести до великої похибки в наближеному значенні інтеграла. Тому практичну цінність мають лише такі квадратурні формули, для яких сума $\sum_{k=0}^{n} |C_k|$ невелика. Якщо квадратурна формула точна для $f(x) = 1$, то неважко встановити умову, за якої сума $\sum_{k=0}^{n} |C_k|$ набуває найменших значень. Справді, формула точна для $f(x) = 1$ тоді й тільки тоді, коли $\int_a^b 1 \cdot dx = b - a = \sum_{k=0}^{n} C_k$. З останнього випливає, що $\sum_{k=0}^{n} |C_k|$ матиме найменше значення, коли всі C_k будуть додатні. Тому квадратурні формули з додатними коефіцієнтами використовуються найчастіше.

Отже, *повна похибка чисельного інтегрування дорівнює сумі трьох похибок: похибки методу, неусувної похибки* Δ_Σ *та заключної похибки округлення результату* Δ_0.

Існує можливість оцінити похибку квадратурної формули до початку розв'язання задачі. Така оцінка називається апріорною. Оцінка похибки після розв'язання задачі називається апостеріорною.

Розглянемо апріорну оцінку похибки квадратурної формули Симпсона. Ця похибка графічно визначається сумою площ між кривою $y = f(x)$ та інтерполяційним поліномом Лагранжа $L_2(x)$ (дивись рисунок 7.2). Залишок найпростішої формули Симпсона $R_{симп} = \int_{x_0}^{x_2} f(x)dx - \frac{h}{3}\left(f(x_0) + 4f(x_1) + f(x_2)\right)$.

Його можна розглядати як функцію від кроку h

$$R(h) = \int_{x_1-h}^{x_1+h} f(x)dx - \frac{h}{3}\left(f(x_1-h) + 4f(x_1) + f(x_1+h)\right).$$

Оскільки функція $f(x) \in C^4[a,b]$, то:

$$R'(h) = f(x_1+h) + f(x_1-h) - \frac{1}{3}\left(f(x_1-h) + 4f(x_1) + f(x_1+h)\right) -$$

$$-\frac{h}{3}\left(-f'(x_1-h) + f'(x_1+h)\right) =$$

$$= \frac{2}{3}\left(f(x_1-h) + f(x_1+h)\right) - \frac{4}{3}f(x_1) - \frac{h}{3}\left(-f'(x_1-h) + f'(x_1+h)\right);$$

$$R''(h) = \frac{2}{3}\left(-f'(x_1-h) + f'(x_1+h)\right) - \frac{1}{3}\left(-f'(x_1-h) + f'(x_1+h)\right) -$$

$$-\frac{h}{3}\left(f''(x_1-h) + f''(x_1+h)\right) =$$

$$= \frac{1}{3}\left(-f'(x_1-h) + f'(x_1+h)\right) - \frac{h}{3}\left(f''(x_1-h) + f''(x_1+h)\right);$$

$$R'''(h) = \frac{1}{3}\left(f''(x_1-h) + f''(x_1+h)\right) - \frac{1}{3}\left(f''(x_1-h) + f''(x_1+h)\right) -$$

$$-\frac{h}{3}\left(-f'''(x_1-h) + f'''(x_1+h)\right) =$$

$$= -\frac{h}{3}\left(-f'''(x_1-h) + f'''(x_1+h)\right) = -\frac{2}{3}h^2 f^{(IV)}(\xi_3), \xi_3 \in (x_1-h, x_1+h);$$

$$R''(h) = R''(0) + \int_0^h R'''(t)dt = -\frac{2}{3}\int_0^h t^2 f^{(IV)}(\xi_3)dt =$$

$$= -\frac{2}{3}f^{(IV)}(\xi_2)\int_0^h t^2 dt = -\frac{2}{9}h^3 f^{(IV)}(\xi_2), \quad \xi_2 \in (x_1-h, x_1+h);$$

$$R'(h) = R'(0) + \int_0^h R''(t)dt = -\frac{2}{9}\int_0^h t^3 f^{(IV)}(\xi_2)dt =$$

$$= -\frac{2}{9}f^{(IV)}(\xi_1)\int_0^h t^3 dt = -\frac{1}{18}h^4 f^{(IV)}(\xi_1), \quad \xi_1 \in (x_1-h, x_1+h);$$

$$R(h) = R(0) + \int_0^h R'(t)dt = -\frac{2}{9}\int_0^h t^3 f^{(IV)}(\xi_2)dt = -\frac{1}{18}f^{(IV)}(\xi)\cdot\int_0^h t^4 dt =$$

$$= -\frac{h^5}{90}f^{(IV)}(\xi).$$

Залишковий член загальної формули Симпсона

$$R_{симп} = -\frac{h^5}{90}\sum_{i=1}^{n} f^{(IV)}(\xi_i). \qquad (7.12)$$

Оскільки $f^{(IV)}(x)$ – неперервна на $[a,b]$ функція, то знайдеться така точка $\xi \in [a,b]$, що $f^{(IV)}(\xi) = \frac{1}{n}\sum_{i=1}^{n} f^{(IV)}(\xi_i)$

$$R_{симп} = -\frac{n\cdot h^5}{90}f^{(IV)}(\xi) = -\frac{(b-a)\cdot h^4}{180}f^{(IV)}(\xi).$$

Оцінка похибки квадратурних формул часто виявляється малоефективною через труднощі, що виникають при знаходженні похідної підінтегральної функції $f(x): f^{(IV)}(\xi) = \max\limits_{\xi\in[a,b]} f^{(IV)}(\xi) = M$.

У зв'язку з цим широкого застосування набуло правило Рунге апостеріорної оцінки похибки, суть якого полягає в тому, щоб, організувавши обчислення двох значень інтеграла на двох множинах вузлів, їх порівняти й одержати оцінку похибки. Найпоширеніше обчислення інтеграла двічі - із кроками h та $h/2$.

Якщо $I = \int\limits_a^b f(x)dx$ – точне значення інтеграла, I_h – його наближене значення, обчислене з кроком h, а $I_{h/2}$ – наближене значення інтеграла, обчислене із кроком $h/2$, то похибки кожної квадратурної формули із кроком h і $h/2$ можна записати відповідно у вигляді $R_h = h^p \cdot M$, $R_{h/2} = \left(\frac{h}{2}\right)^p \cdot M$, де p – порядок точності формул. Обчислимо наближене значення інтеграла за однією квадратурною формулою спочатку із кроком h, а потім із кроком $h/2$. Одержимо

$$I = I_h + h^p \cdot M, \qquad I = I_{h/2} + \left(\frac{h}{2}\right)^p \cdot M.$$

$$I_{h/2} - I_h = M \cdot \left(\frac{h}{2}\right)^p (2^p - 1);\ M \cdot \left(\frac{h}{2}\right)^p = \frac{I_{h/2} - I_h}{2^p - 1}.$$

Одержали оцінку похибки методом Рунге

$$|R_{h/2}| = \frac{|I_{h/2} - I_h|}{2^p - 1}. \qquad (7.13)$$

Користуючись цією формулою можна уточнити наближене значення інтеграла

$$I = I_{h/2} + R_{h/2} = I_{h/2} + \frac{I_{h/2} - I_h}{2^p - 1}. \qquad (7.14)$$

Цю формулу називають формулою екстраполяції за Річардсоном.

Для формули Симпсона $R_{h/2} = \frac{I_{h/2} - I_h}{15}$.

7.4 Вибір квадратурних формул чисельного інтегрування

Ми одержали ряд формул чисельного інтегрування. Виникають запитання: яку формулу потрібно застосовувати в тому або іншому випадку, які формули більш вигідні і які менш вигідні. На ці питання не можна відповісти однозначно. Усе залежить від того, яким способом задана підінтегральна функція, які обчислювальні засоби використовуються, яка необхідна точність і таке інше.

У такій загальній постановці питання відповісти можна лише так: та формула краща, що у даному випадку дає відповідь з потрібною нам точністю при якомога менших витратах праці і часу.

Якщо обчислення ведуться вручну або за допомогою малих обчислювальних машин, то мають значення формули, що містять різниці. Менш вживані формули Гауса і Чебишева, тому що обчислення з багатозначними коефіцієнтами й абсцисами в цьому випадку є громіздкими. З формул, що не містять різниці, найчастіше застосовується формула Симпсона.

При обчисленнях з використанням сучасних комп'ютерів найуживанішими є безрізницеві формули. Особливо вигідні найбільш точні формули Гауса, тому що вони вимагають найменшого числа операцій для одержання інтеграла з потрібною точністю.

Тут необхідно зробити деякі зауваження щодо більш точних і менш точних формул. Ці терміни були введені нами при виведенні формул чисельного інтегрування й у них вкладався певний зміст. Потрібно чітко розуміти, що більш

точна в цьому змісті формула не завжди дає практично більш точний результат. Справді, візьмемо найбільш точну з формул - формулу Гауса . Вона має вигляд

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} C_i f(x_i), \qquad (7.15)$$

де коефіцієнти C_i і абсциси x_i фіксовані і залежать тільки від n і $[a,b]$. Може трапитися, що підінтегральна функція набуває нульового значення у кожній із точок x_i, а абсолютна величина інтеграла від неї велика. Тоді різниця між точним значенням інтеграла і наближеним, отриманим за формулою Гауса, буде також дуже велика. У зв'язку з цим потрібно відзначити, що при виборі тієї або іншої формули чисельного інтегрування буває доцільно вивчити поводження підінтегральної функції і порівняти його з поводженням інтерполяційного многочлена, інтегруванням якого і знаходиться формула чисельного інтегрування. Іноді виникає необхідність розбивати відрізок інтегрування на окремі ділянки так, щоб краще описати поводження функції інтерполяційними многочленами.

Стосовно вибору квадратурних формул, то, очевидно, що за існування четвертої неперервної похідної від підінтегральної функції $f(x)$ краще користуватися формулою парабол, за існування лише другої похідної та аналітичного задання $f(x)$ - формулою середніх, а за табличного–формулою трапецій. Для функції $f(x)$ високої гладкості найзручнішою є формула Гауса.

Оскільки узагальнені формула середніх і формула трапецій є інтегральними сумами, вони мають збігатися до точного значення інтеграла для довільної неперервної функції. Сказане має місце і для формули парабол (оскільки формула парабол – комбінація формули середніх і формули трапеції). Можна довести збіжність і квадратурної формули Гауса.

Швидкість збіжності квадратурної формули визначається оцінкою залишкового члена. Тобто, якщо $R = O(h^p)$, то квадратурна формула називається збіжною з p – порядком збіжності (нагадаємо, що для формули середніх, формули трапеції $p=2$, а для формули парабол $p=4$). Як правило,

підінтегральна функція при цьому повинна мати похідну, яка входить до залишкового члена. Доведено, що чисельне інтегрування стійке за вхідними даними, хоча квадратурні формули нестійкі відносно похибок округлення, але ця нестійкість слабка і виявляється лише за розрахунків з малою кількістю правильних цифр.

Приклад. Обчислення інтеграла $F = \int_0^1 \sqrt{x}dx$. У підінтегральної функції перша похідна необмежена, тому використаємо метод Ейткена для оцінки похибки. Ефективний порядок точності невідомий. Складемо таблицю значень функції й обчислимо інтеграл за формулами трапецій і Сімпсона при різних кроках.

x	$f(x)=\sqrt{x}$	h	Трапецій	Сімпсона	Ейткена
0,00	0,0000	1,00	0.5000	–0,6381	–
0,25	0,5000	0,50	0,6036	0,6565	–
0,50	0,7071	0,25	0,6433		0,6446
0,75	0,8660				
1,00	1,0000				

Як бачимо, дві формули дають результати невисокої точності. Погана точність формули Сімпсона означає, що формула трапецій має фактично не другий порядок точності й уточнення методом Рунге тут недоцільне. А уточнення першого стовпця таблиці процесом Ейткена значно покращує результат; водночас з'ясовується, що в даному прикладі ефективний порядок точності формули трапецій $p \approx 1{,}38$.

Ефективний порядок точності виявився не цілим числом! З цим доводиться стикатися, якщо функція має особливості, а формула інтегрування явно цього не враховує, або якщо особливості має сама формула (це можливо в нелінійних формулах інтегрування).

Якщо ніяких особливостей немає, то ефективний порядок точності може лише трішки відрізнятися від теоретичного завдяки наявності у похибці не лише головного члена, але й

членів більш високого порядку малості. У такому випадку при $h \to 0$ ефективний порядок прямує до теоретичного.

Для обчислень можна використати пакет Maple:

```
> F1:=0.5;
                              F1 := 0.5
> F2:=0.6036;
                              F2 := 0.6036
> F3:=0.6433;
                              F3 := 0.6433
> F:=F1+(F1-F2)^2/(2*F2-F1-F3);
                              F := 0.6446490566
> q:=0.5;
                              q := 0.5
> beta=F-F1;
                              b = 0.1446490566
> p=ln(q)^(-1)*ln((F3-F2)/(F2-F1));
                              p = 1.383813091
```

Отже, якщо необхідно оцінити доцільність використання певної квадратурної формули до безпосередніх розрахунків, то необхідно скористатися апріорними оцінками похибки квадратурних формул, оскільки саме вони дають можливість оцінити точність даної формули, без виконання обчислень. Але, якщо апріорно неможливо оцінити похибку квадратурної формули, або необхідно виконати обчислення і додатково оцінити похибку квадратурної формули, то можна скористатися апостеріорними оцінками квадратурних формул, які дозволяють зробити оцінку похибки безпосередньо під час розрахунків.

Використавши один із методів оцінки похибки квадратурних формул, можна оцінити точність проведення обчислень для поставленого завдання, якщо необхідно – вибрати оптимальний варіант обчислень, тобто найбільш точну формулу для даного випадку.

7.5 Чисельне інтегрування кратних інтегралів

Розглянемо K-вимірний інтеграл вигляду

$$I = \underbrace{\int \dots \int}_{k \text{ разів}} f(x) \cdot dx_1 \dots dx_k, \qquad (7.16)$$

де $x = (x_1, x_2, \dots, x_k)$ - деяка K-вимірна точка. Далі розглянемо подвійні інтеграли (K=2), оскільки їх можна інтерпретувати графічно.

Кубатурні формули, або *формули чисельних кубатур,* призначені для чисельного визначення кратних інтегралів.

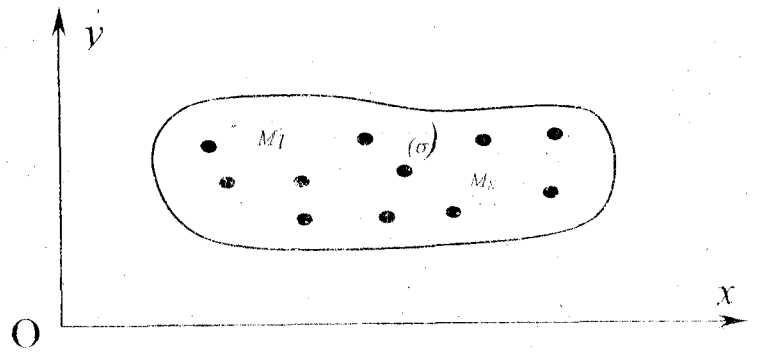


Рис. – 7.3

Нехай функція $z = f(x_1, x_2, \dots, x_k)$ визначена й неперервна в деякій обмеженій області σ. У цій області σ вибирається система точок (вузлів) $M_i(x_{1_i}, x_{2_i}, \dots, x_{k_i})$ $(i = \overline{1, N})$. Для обчислення інтеграла $\underbrace{\int \dots \int}_{k \text{ разів}} f(x_1, x_2, \dots, x_k) \cdot dx_1 \dots dx_k$ приблизно покладемо

$$\underbrace{\int \dots \int}_{k \text{ разів}} f(x_1, x_2, \dots, x_k) \cdot dx_1 \dots dx_k \approx \sum_{i=1}^{N} C_i f(x_{1_i}, x_{2_i}, \dots, x_{k_i}). \qquad (7.17)$$

Щоб знайти коефіцієнти C_i, зажадаємо точного виконання кубатурної формули (7.17) для всіх поліномів

$$P_n(x_1, x_2, \ldots, x_k) = \sum_{l_1+l_2+\ldots l_k \le n} c_{l_1 l_2 \ldots l_k} x_1^{l_1} \cdot x_2^{l_2} \cdot \ldots \cdot x_k^{l_k}, \qquad (7.18)$$

степінь яких не перевищує заданого числа n. Для цього необхідно й достатньо, щоб формула (7.17) була точною для добутку степенів

$$x_1^{l_1} \cdot x_2^{l_2} \cdot \ldots \cdot x_k^{l_k} \quad (l_1, l_2, \ldots, l_k = 0,1,2,\ldots,n; l_1 + l_2 + \ldots + l_k \le n).$$

Покладаючи в (7.17) $f(x_1, x_2, \ldots, x_k) = x_1^{l_1} \cdot x_2^{l_2} \cdot \ldots \cdot x_k^{l_k}$, маємо:

$$I_{l_1 l_2 \ldots l_k} = \underbrace{\int \ldots \int}_{k \text{ разів}} x_1^{l_1} \cdot x_2^{l_2} \cdot \ldots \cdot x_k^{l_k} \cdot dx_1 \ldots dx_k = \sum_{i=1}^{N} C_i x_{1_i}^{l_1} \cdot x_{2_i}^{l_2} \cdot \ldots \cdot x_{k_i}^{l_k} \qquad (7.19)$$

$(l_1, l_2, \ldots, l_k = 0,1,2,\ldots,n;\ l_1 + l_2 + \ldots + l_k \le n)$.

Таким чином, коефіцієнти C_i формули (7.17) можуть бути визначені із системи лінійних рівнянь (7.19).

Для того щоб система (7.19) мала розв'язок, необхідно, щоб число невідомих N дорівнювало числу рівнянь. У випадку подвійного інтеграла ($k = 2$) одержуємо

$$N = (n+1) + n + \ldots + 1 = \frac{(n+1)(n+2)}{2}.$$

7.6 Вибір кубатурних формул

Для одержання заданої точності ε у разі K-кратного інтеграла сітковим (різницевим) методом потрібно виконати близько $\left(\frac{1}{\varepsilon}\right)^{K/p}$ обчислень підінтегральної функції, де p – порядок точності сіткової формули.

Отже, якщо $\frac{K}{p} < 2$, вигідні сіткові методи, якщо ж $\frac{K}{p} > 2$, то вигідний метод Монте-Карло. Так, за p=2 тривимірний інтеграл обчислюють сітковими методами, а при K=5 – методом Монте-Карло.

Розглянемо інтеграл по k-вимірній області, яка розбита сіткою на комірки (Рис. 7.4). Його можна обчислити

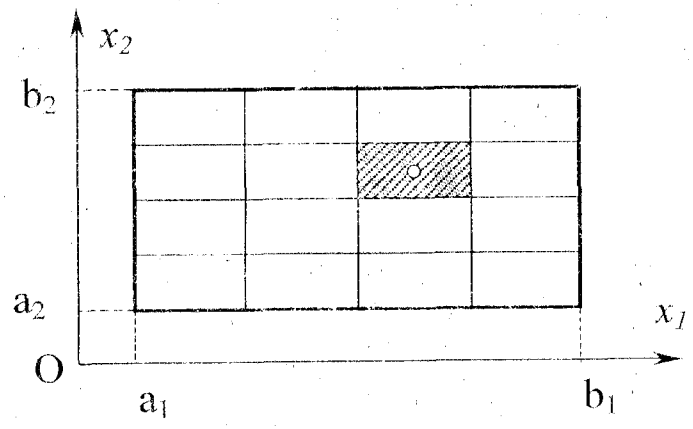


Рис. –7.4

послідовним інтегруванням:

$$I = \underbrace{\int_{a_1}^{b_1} \ldots \int_{a_k}^{b_k}}_{k \text{ разів}} f(x_1, x_2, \ldots, x_k) dx_1 \ldots dx_k = \int_{a_k}^{b_k} F_k(x_k) dx_k$$

$$F_k(x_k) = \underbrace{\int_{a_1}^{b_1} \ldots \int_{a_{k-1}}^{b_{k-1}}}_{k-1 \text{ разів}} f(x_1, x_2, \ldots, x_k) dx_1 \ldots dx_{k-1}$$

$$F_k(x_k) = \int_{a_{k-1}}^{b_{k-1}} F_{k-1}(x_{k-1}) dx_{k-1}$$

$$F_{k-1}(x_{k-1}) = \underbrace{\int_{a_1}^{b_1} \ldots \int_{a_{k-2}}^{b_{k-2}}}_{k-2 \text{ рази}} f(x_1, x_2, \ldots, x_k) dx_1 \ldots dx_{k-2} \ldots$$

Кожний однократний інтеграл легко обчислюється на даній сітці за квадратурними формулами типу

$$F = \sum_{i=0}^{n} c_i f(x_i) + R,$$

$$c_i = \int_a^b \varphi_i(x)\rho(x)dx, \quad R = \int_a^b r(x)\rho(x)dx\,.$$

Послідовне інтегрування в усіх напрямках приводить до кубатурних формул, які є *прямим добутком* одновимірних квадратурних формул:

$$I \approx \sum_{l_1 l_2 \ldots l_k} c_{l_1 l_2 \ldots l_k} f(x_1, x_2, \ldots, x_k,), \quad c_{l_1 l_2 \ldots l_k} = \prod_{i=1}^{k} c_{l_i}. \qquad (7.20)$$

Наприклад, при k=2, якщо по кожному напрямку обрана узагальнена формула трапецій, а сітка рівномірна, то ваги кубатурної формули дорівнюють $\frac{c_{l_1 l_2}}{h_{x_1} h_{x_2}} = 1, \frac{1}{2}$ та $\frac{1}{4}$ відповідно для внутрішніх, граничних і кутових вузлів сітки. Легко показати, що для двічі неперервно диференційованих функцій ця формула має другий порядок точності, і до неї застосуємо метод Рунге-Ромберга.

Взагалі для різних напрямків можна використати квадратурні формули різних порядків точності p_i $(i = \overline{1,k})$. Тоді головний член похибки має вигляд

$$R = O(h_{x_1}^{p_1} + h_{x_2}^{p_2} + \ldots + h_{x_k}^{p_k}).$$

Бажано для всіх напрямків використовувати квадратурні формули однакового порядку точності.

Можна підібрати ваги й положення ліній сітки так, щоб одновимірна квадратурна формула була точною для многочлена максимального степеня, тобто була б формулою Гауса. Тоді для випадку k=2:

$$c_{ij} = \frac{1}{4}(b_1 - a_1)(b_2 - a_2)\gamma_i\gamma_j, \quad x_{1_i} = \frac{1}{2}(a_1 + b_1) + \frac{1}{2}(b_1 - a_1)\xi_i,$$
$$x_{2_j} = \frac{1}{2}(a_2 + b_2) + \frac{1}{2}(b_2 - a_2)\xi_i, \quad 1 \le i, j \le n, \qquad (7.21)$$

де ξ_i, γ_i, $i = \overline{1,n}$-нулі многочленів Лежандра й відповідні ваги. Ці формули розраховані на функції високої гладкості й дають для них більшу економію за кількістю вузлів у порівнянні з простішими формулами.

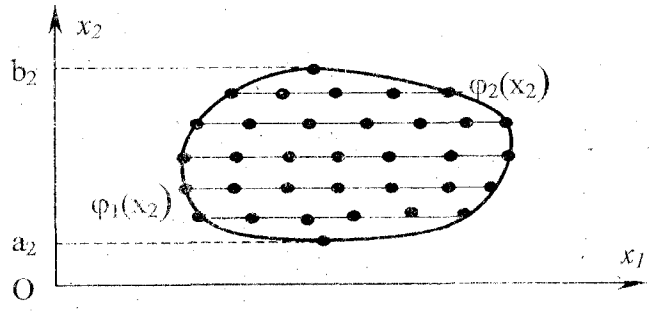


Рис. 7.5

Метод послідовного інтегрування можна застосовувати до області довільної форми, наприклад, із криволінійною границею. Розглянемо цей випадок при K=2. Для цього проведемо через область хорди, паралельні осі x_1, і на них уведемо вузли, розміщені на кожній хорді так, як нам потрібно (рис. 7.5). Представимо інтеграл у вигляді

$$I = \iint_G f(x_1, x_2)dx_1dx_2 = \int_{a_2}^{b_2} F(x_2)dx_2, \; F(x_2) = \int_{\varphi_1(x_2)}^{\varphi_2(x_2)} f(x_1, x_2)dx_1.$$

Спочатку обчислимо інтеграл по x_1 уздовж кожної хорди за будь-якою одномірною квадратурною формулою, використовуючи введені вузли. Потім обчислимо інтеграл по x_2; тут вузлами будуть служити проекції хорд на вісь ординат.

При обчисленні інтеграла по x_2 є одна особливість. Якщо область обмежена гладкою кривою, то при $x_2 \to a_2$ довжина хорди прямує до нуля не лінійно, а як $\sqrt{x_2 - a_2}$; виходить, поблизу цієї точки $F(x_2) \sim \sqrt{x_2 - a_2}$. Те саме буде при $x_2 \to b_2$. Тому інтегрувати безпосередньо $F(x_2)$ за формулами високого порядку точності не має сенсу. Доцільно виділити з $F(x_2)$ основну особливість у вигляді ваги $\rho(x_2) = \sqrt{(b_2 - x_2)(x_2 - a_2)}$, якій відповідають ортогональні многочлени Чебишева другого роду.

Тоді друге інтегрування виконується за формулами Гауса

$$I = \int_{a_1}^{b_2} F(x_2)dx_2 \approx \sum_{j=1}^{n} \frac{b_2 - a_2}{2} \gamma_j \psi\left(\frac{a_2 + b_2}{2} + \frac{a_2 - b_2}{2}\xi_j\right), \quad (7.22)$$

де $\psi(x_2) = \dfrac{F(x_2)}{\rho(x_2)}$, а $\xi_j = \cos\left[\dfrac{\pi j}{n+1}\right]$ й γ_j-нулі й ваги многочленів Чебишева другого роду.

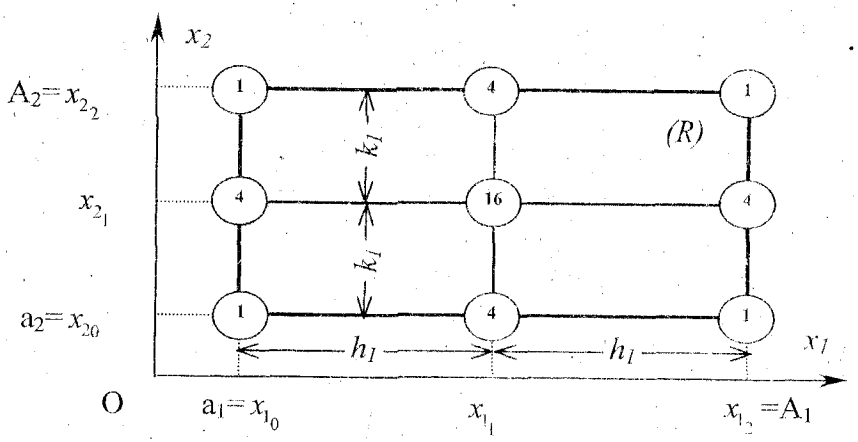


Рис. – 7.6

7.7 Кубатурна формула типу Симпсона

Нехай областю інтегрування є К-вимірний просторовий паралелепіпед $R\{a_1 \le x_1 \le A_1; a_2 \le x_2 \le A_2; \ldots; a_k \le x_k \le A_k;\}$ (рис.7.6), сторони якого паралельні осям координат. Кожний із проміжків $[a_i, A_i]$ $(i = \overline{1,k})$ розіб'ємо навпіл точками:

$$x_{i_0} = a_i, x_{i_1} = a_i + h_i, x_{i_2} = a_i + 2h_i = A_i \quad (i = \overline{1,k}), \quad \text{де}$$

$h_i = \dfrac{A_i - a_i}{2}$ $(i = \overline{1,k})$.

Усього, таким чином, одержимо 3^k точок сітки. Маємо

$$\underbrace{\int \ldots \int}_{k\ \text{разів}} f(x_1, x_2, \ldots, x_k)dx_1 \ldots dx_k = \underbrace{\int_{a_1}^{A_1} dx_1 \int_{a_2}^{A_2} dx_2 \ldots \int_{a_k}^{A_k}}_{k\ \text{разів}} f(x_1, x_2, \ldots, x_k)dx_k. \quad (7.23)$$

Знаходимо К-вимірний інтеграл, обчислюючи кожний внутрішній інтеграл за квадратурною формулою Симпсона на відповідному відрізку. Проведемо повністю всі обчислення для випадку К=2:

$$\iint f(x_1, x_2)dx_1dx_2 = \int_{a_1}^{A_1} dx_1 \cdot \frac{h_2}{3}\left[f(x_1, x_{2_0}) + 4f(x_1, x_{2_1}) + f(x_1, x_{2_2})\right] =$$

$$= \frac{h_2}{3}\left[\int_{a_1}^{A_1} f(x_1, x_{2_0})dx_1 + 4\int_{a_1}^{A_1} f(x_1, x_{2_1})dx_1 + \int_{a_1}^{A_1} f(x_1, x_{2_2})dx_1\right].$$

Застосовуючи до кожного інтеграла знову формулу Симпсона, одержимо:

$$\iint f(x_1, x_2)dx_1dx_2 = \frac{h_1h_2}{9}\left\{\left[f(x_{1_0}, x_{2_0}) + 4f(x_{1_1}, x_{2_0}) + f(x_{1_2}, x_{2_0})\right] + \right.$$
$$\left. + 4\left[f(x_{1_0}, x_{2_1}) + 4f(x_{1_1}, x_{2_1}) + f(x_{1_2}, x_{2_1})\right] + \left[f(x_{1_0}, x_{2_2}) + 4f(x_{1_1}, x_{2_2}) + f(x_{1_2}, x_{2_2})\right]\right\},$$

або

$$\iint f(x_1, x_2)dx_1dx_2 = \frac{h_1h_2}{9}\left\{\left[f(x_{1_0}, x_{2_0}) + f(x_{1_2}, x_{2_0}) + f(x_{1_0}, x_{2_2}) + f(x_{1_2}, x_{2_2})\right] + \right. \quad (7.24)$$
$$\left. + 4\left[f(x_{1_1}, x_{2_0}) + f(x_{1_0}, x_{2_1}) + f(x_{1_2}, x_{2_1}) + f(x_{1_1}, x_{2_2})\right] + 16f(x_{1_1}, x_{2_1})\right\}.$$

Формулу (7.24) будемо називати *кубатурною формулою Симпсона*. Отже,

$$\iint\limits_{(R)} f(x_1, x_2)dx_1dx_2 = \frac{h_1h_2}{9}(\sigma_0 + 4\sigma_1 + 16\sigma_2), \qquad (7.25)$$

де σ_0 – сума значень підінтегральної функції $f(x_1, x_2)$ у вершинах прямокутника R, σ_1 – сума значень $f(x_1, x_2)$ у серединах сторін прямокутника R, $\sigma_2 = f(x_{1_i}, x_{2_j})$ – значення функції $f(x_1, x_2)$ в центрі прямокутника R. Кратності цих значень позначені на рис. 7.6.

Якщо розміри просторового паралелепіпеда $R\{a_1 \le x_1 \le A_1; a_2 \le x_2 \le A_2; \ldots; a_k \le x_k \le A_k;\}$ великі, то для збільшення точності кубатурної формули область R

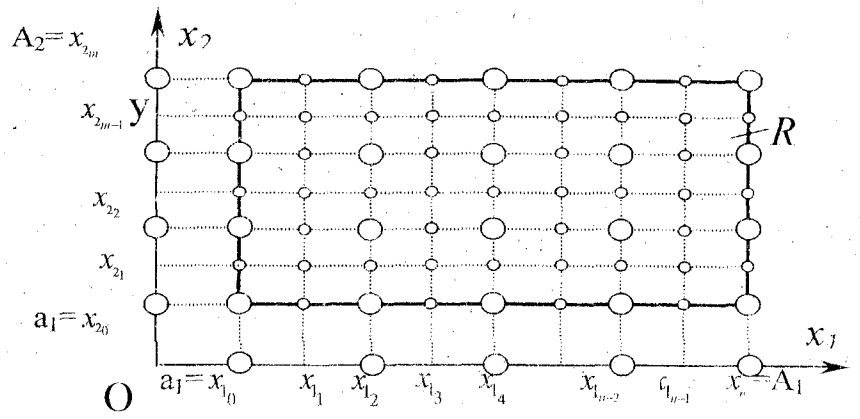


Рис. – 7.7

розбивають на систему паралелепіпедів, до кожного з яких застосовують кубатурну формулу Симпсона.

Знову розглянемо випадок К=2. Покладемо, що сторони прямокутника R ми розділили відповідно на n_1 й n_2 однакових частин; у результаті вийшла відносно велика мережа n_1n_2 прямокутників (на рис. 7.7 вершини цих прямокутників відзначені більшими кружками). Кожний із цих прямокутників, у свою чергу, розділимо на чотири однакові частини. Вершини

цієї останньої дрібної мережі прямокутників візьмемо за вузли M_{ij} кубатурної формули.

Нехай $h_1 = \frac{A_1 - a_1}{2n_1}$ і $h_2 = \frac{A_2 - a_2}{2n_2}$. Тоді мережа вузлів буде мати координати: $x_{1_i} = x_{1_0} + ih_1 \quad (x_{1_0} = a_1; \ i = \overline{0,2n_1})$;

$$x_{2_j} = x_{2_0} + jh_2 \quad (x_{2_0} = a_2; \ j = \overline{0,2n_2}).$$

Для скорочення введемо позначення $f(x_{1_i}, x_{2_j}) = f_{ij}$.

Застосовуючи формулу (7.24) до кожного із прямокутників великої мережі, будемо мати (рис.7.7):

$$\iint\limits_{(R)} f(x_1, x_2)dx_1dx_2 = \frac{h_1h_2}{9}\sum_{i=0}^{n}\sum_{j=0}^{m}[(f_{2i,2j} + f_{2i+2,2j} + f_{2i+2,2j+2} + f_{2i,2j+2}) +$$
$$+ 4(f_{2i+1,2j} + f_{2i+2,2j+1} + f_{2i+1,2j+2} + f_{2i,2j+1}) + 16f_{2i+1,2j+1}].$$

Звідси, виконавши зведення подібних членів, остаточно знаходимо:

$$\iint\limits_{(R)} f(x_1, x_2)dx_1dx_2 = \frac{h_1h_2}{9}\sum_{i=0}^{2n}\sum_{j=0}^{2m}\lambda_{ij}f_{ij}, \qquad (7.26)$$

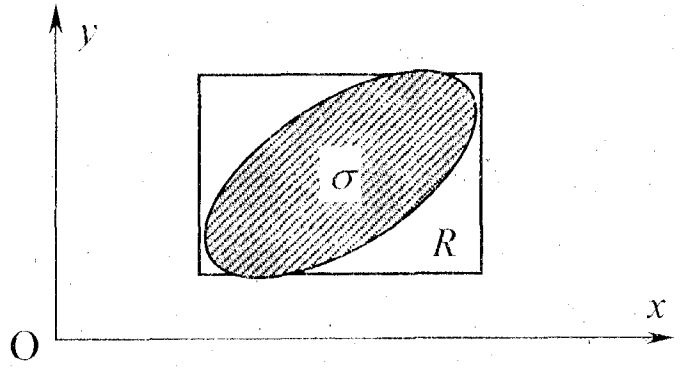


Рис. – 7.8

де коефіцієнти λ_{ij} є відповідними елементами матриці

$$\Lambda = \begin{bmatrix} 1 & 4 & 2 & 4 & 2 & \dots & 4 & 2 & 4 & 1 \\ 4 & 16 & 8 & 16 & 8 & \dots & 16 & 8 & 16 & 4 \\ 2 & 8 & 4 & 8 & 4 & \dots & 8 & 4 & 8 & 2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 2 & 8 & 4 & 8 & 4 & \dots & 8 & 4 & 8 & 2 \\ 4 & 16 & 8 & 16 & 8 & \dots & 16 & 8 & 16 & 4 \\ 1 & 4 & 2 & 4 & 2 & \dots & 4 & 2 & 4 & 1 \end{bmatrix}.$$

Якщо область інтегрування σ – довільна, то будуємо паралелепіпед $R \supset \sigma$, сторони якого паралельні осям координат (рис. 7.8). Розглянемо допоміжну функцію

$$f^*(x_1, x_2, \dots, x_k) = \begin{cases} f(x_1, x_2, \dots, x_k), & (x_1, x_2, \dots, x_k) \in \sigma; \\ 0, & (x_1, x_2, \dots, x_k) \in R - \sigma. \end{cases}$$

У такому випадку маємо

$$\underbrace{\int_{(\sigma)} \dots \int}_{\text{k раз}} f(x_1, x_2, \dots, x_k) dx_1 \dots dx_k = \underbrace{\int_{(R)} \dots \int}_{\text{k раз}} f^*(x_1, x_2, \dots, x_k) dx_1 \dots dx_k.$$

Останній інтеграл приблизно може бути обчислений за загальною кубатурною формулою (7.26).

Питання і завдання до розділу 7

1 Найпростіші квадратурні формули (прямокутників, трапецій, Симпсона), геометрична ілюстрація, оцінки похибки. Точність квадратурних формул.

2 Квадратурні формули інтерполяційного типу: виведення формул, оцінки похибки.

3 Квадратурні формули Гауса: виведення формул, точність формул.

4 Правило Рунге практичної оцінки похибки. Адаптивні процедури чисельного інтегрування.

5 Обчислити наближено з кроком h=1 інтеграл $\int_{-1}^{3} \frac{dx}{2+x}$ за формулами прямокутників, трапецій, Симпсона. Оцінити похибку теоретично.

6 Переконатися в тім, що формула прямокутників є точною для многочленів $1, t$, а формула Симпсона – для многочленів $1, t, t^2, t^3$.

7 Оцінити теоретично значення кроку інтегрування h для наближеного обчислення інтеграла $\int_0^1 \frac{dx}{1+x}$ за формулою трапецій з точністю $\varepsilon = 10^{-3}$.

8 Оцінити теоретично значення кроку інтегрування h для наближеного обчислення інтеграла $\int_0^1 \sin(x^2) dx$ по формулі Симпсона з точністю $\varepsilon = 10^{-4}$.

9 Одержати квадратурні формули прямокутників і трапецій із загальної формули інтерполяційного типу.

10 Переконатися, що квадратурна формула Гауса з одним вузлом точна для многочленів $1, t, t^2, t^3$.

11 Обчислити інтеграл $\int_0^1 \frac{dx}{1-x+x^2}$ за формулами трапецій і Симпсона з точністю $\varepsilon = 10^{-2}$, використовуючи правило Рунге оцінки похибки.

12 Знайти оцінку похибки обчислення інтеграла $\int_0^1 \frac{dx}{1+x^2}$ за складеною формулою $S = (f(0) + 2f(0.1) + 2f(0.2) + \dots + 2f(0.9) + f(1))/20$.

13 Оцінити мінімальне число розбиттів відрізка N інтегрування для наближеного обчислення інтеграла $\int_0^1 \sin(x^2) dx$ за складеною формулою трапецій, що забезпечує точність $\varepsilon = 10^{-4}$.

14 Обчислити інтеграли $I_k = \int_a^b P_k(x)dx$, де $P_k(x) = \sum_{i=0}^{k} c_i x^i$, k=0,1,...,5 аналітично й використовуючи квадратурну формулу Симпсона із кроком $h = (b-a)/2$. Для многочленів якого степеня використовувана квадратурна формула точна й чому? Оцінити похибку інтегрування за правилом Рунге.

15 Обчислити значення інтеграла $\int_a^b e^{2x-3}dx$ аналітично й, використовуючи формулу прямокутників із кроками h: $\frac{b-a}{2}, \frac{b-a}{3}, \ldots \frac{b-a}{20}$ ...$(a=0; b=1)$. При зазначених значеннях h знайти абсолютну похибку й оцінки теоретичної абсолютної похибки. На одному кресленні побудувати графіки знайдених похибок.

16 Побудувати графік функції $F(x) = \int_0^1 \frac{x}{1+t}dt, \quad 0 \le x \le 2$. Для обчислення інтеграла з точністю 10^{-8} використати квадратурну формулу трапецій і правило Рунге оцінки похибки.

17 Обчислити значення інтеграла I із задачі 14, використовуючи квадратурну формулу Гауса з одним, двома, трьома, чотирма вузлами. Визначити абсолютну похибку результату. Побудувати гістограму залежності похибки від числа вузлів. Переконатися, що квадратурні формули Гауса з N+1 (N=0,1,2,3) вузлами точні для многочленів $1, t, ..., t^m$, де $m=2N+1$.

18 Обчислити наближено площу фігури, обмеженої кривими $y = \sin x^2$, $y = 0$. Точки перетину кривих знайти графічно. Для обчислення інтегралів з точністю 10^{-8} використати квадратурну формулу Симпсона і правило Рунге оцінки похибки.

19 Наближено обчислити подвійний інтеграл по прямокутній області $D = \{(x,y), x \in [a,b], y \in [c,d]\}$ з точністю 0.001.

20 Функція $y=y(x)$ задана таблицею своїх значень:

x	0	0.1	0.2	0.3	0.4
y	1	1.2	1.24	0.76	0.6

Обчислити наближене значення інтеграла $\int_0^{0.4} y(x)dx$ за квадратурними формулами трапецій і Симпсона.

21 Побудувати квадратурну формулу $\int_{-1}^{1} f(x)dx \approx c_1 f(-1) + c_2 f(0) + c_3 f(1)$, точну для многочленів найбільш високого степеня, використовуючи метод невизначених коефіцієнтів.

22 Знайти наближене значення інтеграла $\int_1^{2.2} e^{-x^2}dx$, із кроком $h = 0.2$, використовуючи квадратурні формули прямокутників, трапецій, Симпсона. Оцінити похибку формули чисельного інтегрування двома способами: використовуючи теоретичну оцінку похибки та правило Рунге.

23 З яким кроком інтегрування потрібно обчислювати наближене значення інтеграла $\int_{1.5}^{2} \ln(x)dx$ за формулою трапецій для того, щоб забезпечити точність 0.00001.

Розділ 8

Чисельне розв'язання звичайних диференціальних рівнянь

Звичайними диференціальними рівняннями називаються рівняння, що пов'язують функцію та її похідні з однією незалежною змінною. Якщо незалежних змінних більше, ніж одна, то рівняння називається диференціальним рівнянням з частинними похідними.

За допомогою звичайних диференціальних рівнянь будуються моделі руху систем взаємодіючих часток, електротехнічних процесів у електричних ланцюгах, кінетики хімічних реакцій, процесів заселення рівнів енергії у високотемпературних середовищах і багатьох інших об'єктів і процесів.

До задач для звичайних диференціальних рівнянь зводяться деякі задачі для рівнянь у частинних похідних, коли рівняння дозволяє провести відокремлення змінних (наприклад, при обчисленні енергетичного спектра часток у полях визначеної симетрії).

Звичайне диференціальне рівняння будь-якого порядку за допомогою заміни змінних може бути зведене до системи рівнянь першого порядку.

У загальному вигляді перетворення є таким:
диференціальне рівняння n-го порядку

$$y^{(n)}(x) = f\left(x, y, y', y'', ..., y^{(n-1)}\right),$$

заміною змінних $y^{(k)} \equiv v_k$ зводяться до системи n рівнянь першого порядку

$$\begin{cases} v_k' = v_{k+1}, 0 \le k \le n-2, \\ v_{n-1}'(x) = f\left(x, v_0, v_1, v_2, ..., v_{n-1}\right), \end{cases}$$

де позначено $v_0 \equiv y$.

Відповідно до викладеного далі будуть розглядатися системи рівнянь першого порядку:

$$v_k'(x) = \varphi_k\left(x, v_1, v_2, ..., v_n\right), 1 \le k \le n.$$

Розв'язок системи n-го порядку залежить від n параметрів $c_1, c_2, ..., c_n$. Єдиний розв'язок визначається при використанні додаткових умов для шуканої функції. У залежності від того, яким чином ставляться такі умови, розрізняють три типи задач для звичайних диференціальних рівнянь: задача Коші, крайова задача і задача на власні значення.

У задачі Коші всі додаткові умови ставляться в одній точці $v_k(x_0) = v_{k,0}, 1 \le k \le n$. Розв'язок шукається на деякому інтервалі $x_0 \le x \le x_l$.

Якщо праві частини φ_k рівнянь неперервні в деякому околі початкової точки $\left(x_0, v_{1,0}, v_{2,0}, ..., v_{n,0}\right)$ і задовольняють умову Ліпшиця за змінними v_k, то розв'язок задачі Коші існує, єдиний і неперервно залежить від координат початкової точки, тобто задача є коректною. Умова Ліпшиця формулюється в такий спосіб:

$$\left|\varphi_k\left(x, v_{1,l}, v_{2,l}, ..., v_{n,l}\right) - \varphi_k\left(x, v_{1,m}, v_{2,m}, ..., v_{n,m}\right)\right| \le$$
$$\le L\left\{\left|v_{1,l} - v_{1,m}\right| + \left|v_{2,l} - v_{2,m}\right| + ... + \left|v_{n,l} - v_{n,m}\right|\right\},$$

для будь-яких точок $\left(x, v_{1,l}, v_{2,l}, ..., v_{n,l}\right), \left(x, v_{1,m}, v_{2,m}, ..., v_{n,m}\right)$, де L - деяка константа.

Можна виділити три класи методів розв'язання звичайних диференціальних рівнянь: ***точні, наближені та чисельні.***

Точні методи передбачають одержання розв'язку у вигляді комбінації елементарних функцій або у вигляді квадратур від останніх. Можливості точних методів обмежені.

Наближені методи зводяться до побудови послідовності функцій $w_n(x)$, що мають границею шукану функцію $v(x)$. Обриваючи цю послідовність на якомусь k, одержують наближений розв'язок.

Найбільш універсальними методами розв'язання є чисельні. Їхній основний недолік - можливість одержання тільки часткового розв'язку.

Варто зауважити, що успіх від застосування чисельного методу суттєво залежить від обумовленості задачі, тобто задача повинна бути добре обумовленою, а саме, малі зміни початкових умов повинні призводити до малих змін у розв'язку. У протилежному випадку (слабкої стійкості) малі похибки в початкових даних або похибки чисельного методу можуть призводити до великих похибок у розв'язку.

Приклад. Рівняння $\frac{dv}{dx} = \lambda \cdot v$ з початковою умовою $v(x_0) = v_0$ має розв'язок $v(x) = v_0 e^{\lambda(x-x_0)}$.

При $v_0 = 0$ виходить розв'язок $v(x) = 0$. Якщо припустити, що v_0 не дорівнює строго нулеві, а має невелике відхилення від нуля, наприклад, $v_0 = 10^{-6}$, тоді при великих x буде мати місце така ситуація.

Якщо $\lambda < 0$, то $v(x)$ при збільшенні x прямує до нуля, тобто до незбуреного розв'язку. У цьому випадку розв'язок називається асимптотично стійким за Ляпуновим.

Однак при $\lambda > 0$ зі збільшенням x $v(x)$ необмежено зростає, а саме, наприклад, при $x = 100, x_0 = 0, \lambda = 1$, $v(100) = 10^{-6} e^{1(100-0)} = 2{,}7 \cdot 10^{37}$.

Таким чином, розв'язок виявляється нестійким.

Далі будуть розглядатися алгоритми розв'язку задачі Коші на прикладі одного рівняння першого порядку $v'(x) = \varphi(x, v)$. Узагальнення на випадок системи n рівнянь здійснюється заміною $v(x)$ на $\bar{v}(x)$ і $\varphi(x, v)$ на $\bar{\varphi}(x, \bar{v})$, де

$$\bar{v}(x) = \left| v_1 \quad v_2 \quad \dots \quad v_n \right|, \qquad \bar{\varphi}(x, \bar{v}) = \begin{vmatrix} \varphi_1 \\ \varphi_2 \\ \cdot \\ \varphi_n \end{vmatrix}.$$

8.1 Різницева апроксимація диференціальних рівнянь однокроковими методами

Виберемо на відрізку деяку систему $\{x_n\}, n = \overline{0, N}$, значень аргумента так, щоб виконувалися співвідношення $x_0 < x_1 < x_2 < \dots < x_N = x_0 + H$. Множину $\{x_n\}$ називають сіткою, точки $x_0, x_1, \dots, x_N$ — вузлами сітки, величину $h_n = x_{n+1} - x_n$ - кроком сітки. Якщо $h_n = h$, сітка називається рівномірною, в іншому разі - нерівномірною. Сітковою функцією $y = y_j = y(x_j)$ називається функція, що задана у вузлах сітки. Будь-яку сіткову функцію $y_j = y(x_j)$ можна представити у вигляді вектора $Y = (y_0, y_1, \dots, y_{n-1}, y_n)$.

Нехай маємо диференціальне рівняння $Ly(x) = f(x,y)$ (наприклад, $\frac{dy}{dx} = f(x, y)$), де L — диференціальний оператор.

Замінимо Ly у вузлі сітки x_j лінійною комбінацією значень сіткової функції y_j на деякій множині вузлів сітки, яка називається шаблоном. Така заміна Ly на $L_h y_h$ називається апроксимацією на сітці диференціального оператора L різницевим оператором L_h. Заміна неперервної функції $f(x,y)$ у вузлах сітки на сіткову функцію $f(x_h, y_h)$ називається апроксимацією правої частини.

У такий спосіб диференціальне рівняння можна апроксимувати (замінити) на сітці різницевою схемою

$L_h y_h = f(x_h, y_h)$ (наприклад, $\frac{y_{i+1} - y_i}{h} = f(x_i, y_i)$).

Вивчення різницевих апроксимацій проводиться спочатку локально, тобто в будь-якому фіксованому вузлі сітки.

При розв'язуванні диференціальних рівнянь чисельним методом основним є питання про збіжність. Стосовно до різницевих методів традиційно більш уживане поняття збіжності при $h \to 0$. Позначимо за y_i значення сіткової функції, що відповідає значенню точного розв'язку диференціального

рівняння $\frac{dy}{dx} = f(x,y)$ у вузлі i - $\varphi(x_i)$ (y_i є наближеними значеннями $\varphi(x_i)$). Збіжність при $h \to 0$ означає таке. Фіксуємо точку x і будуємо сукупність сіток ω_h таким чином, що $h \to 0$ і $x_i = a + ih = x$ (при цьому $i \to \infty$). Тоді вважають, що чисельний метод збігається в точці x, якщо $|y_i - \varphi(x_i)| \to 0$ при $h \to 0$, $x_i = x$. Метод збігається на відрізку $[a,b]$, якщо він збігається в кожній точці $x \in [a,b]$. Вважають, що метод має p-й порядок точності, якщо можна знайти таке число $p > 0$, що $|y_i - \varphi(x_i)| = O(h^p)$ при $h \to 0$.

Уведемо далі поняття нев'язки, або похибки, апроксимації різницевого рівняння, що заміняє задане диференціальне рівняння, на розв'язку вихідного рівняння, тобто нев'язка ψ_i являє собою результат підстановки точного розв'язку рівняння $\varphi(x)$ у різницеве рівняння. Наприклад, рівняння $\frac{dy}{dx} = f(x,y)$ можна замінити таким найпростішим різницевим рівнянням

$$\frac{y_{i+1} - y_i}{h} - f(x_i, y_i) = 0, \quad i = 0,1,2,\ldots y_0 = \varphi_0.$$

Тоді нев'язка визначиться як $\psi_i = -\frac{\varphi_{i+1} - \varphi_i}{h} + f(x_i, y_i)$.

Наближений розв'язок не збігається з $\varphi_i = \varphi(x_i)$, тому нев'язка ψ_i в i-ій точці не дорівнює нулеві.

Чисельний метод апроксимує вихідне диференціальне рівняння, якщо $\psi_i \to 0$ при $h \to 0$, і має p-й порядок точності, якщо $\psi_i = O(h^p)$.

Доведено, що порядок точності чисельного методу розв'язання диференціального рівняння збігається з порядком апроксимації при досить загальних припущеннях.

8.1.1 Метод Ейлера

Ознайомлення з чисельними методами розв'язання звичайних диференціальних рівнянь першого порядку почнемо з вивчення методу Ейлера для задачі Коші

$$\begin{cases} y' = f(x,y), & (8.1) \\ y(x_0) = y_0. & (8.2) \end{cases}$$

Відзначимо, що на практиці цей метод використовується рідко через невисоку точність, однак він є найпростішим з чисельних методів і на його прикладі зручно пояснити їх суть, способи побудови і дослідження.

Для розв'язання задачі потрібно знайти наближені значення $y_1, y_2, \ldots, y_N$ точного розв'язку $y = \varphi(x)$ рівняння (8.1). Уведемо позначення $y_n \approx \varphi(x_n)$. Припустимо, що розв'язок y_n задачі (8.1) — (8.2) у вузлі x_n відомий. Знайдемо розв'язок у наступному вузлі $x_{n+1} = x_n + h$. Використовуючи формулу Тейлора, одержимо

$$y(x) = y(x_n) + y'(x_n)(x - x_n) + \frac{1}{2} y''(\xi)(x - x_n)^2, \xi \in (x_n; x) \quad (8.3)$$

Відзначимо, що похідну y'', що стоїть у правій частині, можна знайти, диференціюючи рівняння (8.1).

Підставимо у формулі (8.3) $x = x_{n+1}$, тоді

$$y(x_{n+1}) = y(x_n) + h_n y'(x_n) + \frac{1}{2} h_n^2 y''(\xi). \quad (8.4)$$

Припускаючи, що $y''(x)$ на відрізку $[x_n, x_{n+1}]$ обмежена, маємо $\frac{1}{2} y''(\xi) h_n^2 = O(h_n^2)$. Однак використовувати формулу (8.4) незручно з таких міркувань:

1) вираз y'' може виявитися громіздким; 2) якщо права частина рівняння (8.1) відома лише приблизно, що часто має місце при розв'язанні технічних задач, знаходити її похідні небажано.

Якщо $f(x,y)$ має q-і неперервні похідні по сукупності аргументів, то в розкладанні (8.3) можна враховувати значення членів аж до $O(h^{q+1})$

Відкидаючи в (8.4) величини другого порядку малості при $h \to 0$ в порівнянні з кроком сітки h_n, одержуємо формулу для обчислення наближеного значення $\varphi(x)$ у вузлі x_{n+1}: $y_{n+1} = y_n + h_n y_n'$ З огляду на те, що $y_n' = f(x_n, y_n)$, виводимо розрахункову ***формулу методу Ейлера***

$$y_{n+1} = y_n + h_n f(x_n, y_n). \qquad (8.5)$$

Для чисельного розрахунку за формулою (8.5) досить знати $y(x_0) = y_0$. Потім, використовуючи (8.5), можна послідовно знайти значення розв'язку $y_1, y_2, ..., y_N$ відповідно в точках $x_1, x_2, ..., x_N$

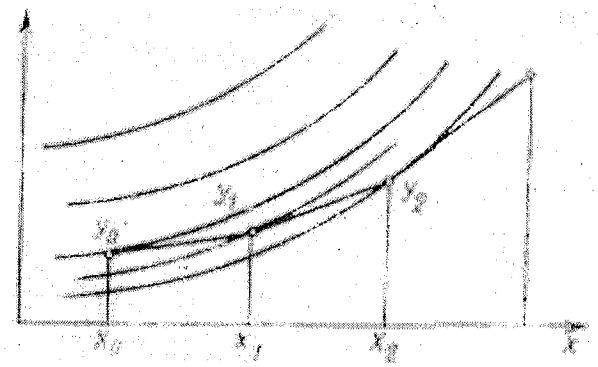


Рис. – 8.1

Геометрична інтерпретація методу Ейлера показана на рис. 8.1, де зображена множина інтегральних кривих рівняння (8.1). Використання тільки першого члена формули Тейлора рівносильне заміні інтегральної кривої на відрізку $[x_n, x_{n+1}]$ дотичною до неї в точці (x_x, y_n). На кожному кроці заново визначається дотична, і, отже, траєкторія буде ламаною лінією. Тому метод Ейлера називають також методом ламаних.

При визначенні наближеного розв'язку задачі надзвичайно важлива оцінка похибки використовуваного методу. Розглянемо таку оцінку для методу Ейлера.

Припустимо, що початкова умова $y(x_0)$ задана точно. При одержанні (8.5) у формулі Тейлора був відкинутий член, що містить h_n^2.

На першому кроці, при обчисленні y_1, отримана похибка $r_1 = y(x_1) - y_1 = O(h_0^2)$, яка називається локальною похибкою, або похибкою на кроці.

На другому кроці y_2 обчислюється за формулою $y_2 = y_1 + h_1 f(x_1, y_1)$. Величина y_1, знайдена раніше, визначена наближено. Тому сумарна похибка на другому кроці $r_2 = y(x_2) - y_2$ буде викликана не тільки заміною інтегральної кривої на відрізку $[x_1, x_2]$ дотичною до неї, але і помилкою, допущеною на першому кроці.

Аналогічно сумарна похибка n-го кроку залежить не тільки від заміни інтегральної кривої на відрізку $[x_{n-1}, x_n]$ дотичною, але і від помилок, допущених при обчисленні $y_1, y_2, ..., y_{n-1}$ (рис. 8.2). У випадку, коли початкова умова задана неточно, сумарна похибка на будь-якому кроці буде залежати і від похибки початкової умови (8.2).

Розглянемо похибку наближеного розв'язку $z_{n+1} = y(x_{n+1}) - y_{n+1}$, знайденого методом Ейлера (рис. 8.2).

Припустимо, що функція *f(x, y)* з (8.1) неперервна і має неперервні перші похідні в області зміни своїх аргументів. Віднімаючи (8.5) з (8.4), одержимо

$$z_{n+1} = z_n + h_n[f(x_n, y(x_n)) - f(x_n, y_n)] + \frac{1}{2} y''(\xi) h_n^2, \xi \in (x_n, x_{n+1}).$$

Використовуючи формулу Тейлора, з урахуванням того, що $z_n = O(h_{n-1}^2)$, одержуємо

$$f(x_n, y(x_n)) = f(x_n, y_n) + f'(x_n, y_n)z_n + \frac{1}{2} f''_{y^2}(x_n, \overset{\approx}{y_n}) z_n^2 =$$

$$= f(x_n, y_n) + z_n f'_y(x_n, y_n) + O(h_{n-1}^4), \overset{\approx}{y_n} \in (y_n; y(x_n)).$$

Звідси $f(x_n, y(x_n)) - f(x_n, y_n) = z_n f'_y(x_n, y_n) + O(h_{n-1}^4)$.

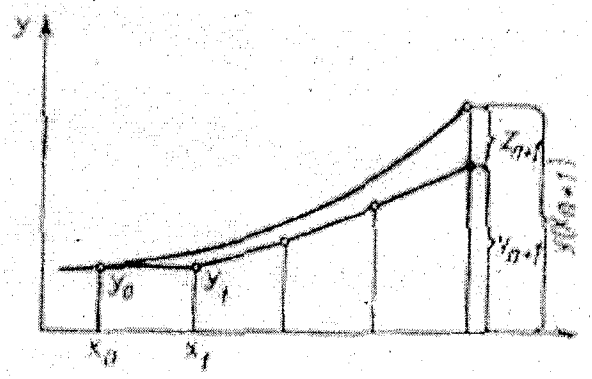


Рис. – 8.2

Отже, з точністю до величин більш високого порядку малості,

$$z_{n+1} = z_n[1 + h_n f'_y(x_n, y_n)] + \frac{1}{2} y''(\xi_n) h_n^2, \quad n = 0,1,2,....$$

Таким чином,

$$z_1 = z_0[1 + h_0 f'_y(x_0, y_0)] + \frac{1}{2} h_0^2 y''(\xi_0);$$

$$z_2 = z_1[1 + h1 f'_y(x_1, y_1)] + \frac{1}{2} h_1^2 y''(\xi_1) =$$

$$= z_0[1 + h_0 f'_y(x_0, y_0)][1 + h_1 f'_y(x_1, y_1)] +$$

$$+ \frac{1}{2} h_0^2 y''(\xi_0) z_0[1 + h_1 f'_y(x_1, y_1)] + \frac{1}{2} h_1^2 y''(\xi_1).$$

Аналогічно

$$z_3 = z_0 \prod_{k=0}^{2}[1 + h_k f'_y(x_k, y_k)] + \frac{1}{2} \sum_{k=0}^{2} h_k^2 y''(\xi_k) \prod_{\nu=k+1}^{2}[1 + h_\nu f'_y(x_\nu, y_\nu)].$$

Продовжуючи цей процес, одержимо

$$z_m = z_0 \prod_{k=0}^{m-1}[1 + h_k f'_y(x_k, y_k)] +$$

$$+ \frac{1}{2} \sum_{k=0}^{m-1} h_k^2 y''(\xi_k) \prod_{\nu=k+1}^{m-1}[1 + h_\nu f'_y(x_\nu, y_\nu)], \quad m = 1,2,... \tag{8.6}$$

Таким чином, похибка z_m на довільному кроці m виражається через похибку z_0.

При малих $h_k, k = \overline{0.m-1}$ має місце така оцінка:

$$\prod_{k=0}^{m-1}[1 + h_k f'_y(x_k, y_k)] \approx \prod_{k=0}^{m-1} \exp[h_k f'_y(x_k, y_k)] =$$

$$= \exp[\sum_{k=0}^{m-1} h_k f'_y(x_k, y_k)] \approx \exp[\int_{x_0}^{x_m} f'_y(t, y(t))dt] (\exp(\varphi) = e^{\varphi}).$$

Аналогічно

$$\prod_{\nu=k+1}^{m-1}[1 + h_\nu f'_y(x_\nu, y)] \approx \exp[\int_{t}^{x_m} f'_y(\tau, y(\tau))d\tau],$$

$$\frac{1}{2} \sum_{k=0}^{m-1} h_k^2 y''(\xi_k) \prod_{\nu=k+1}^{m-1}[1 + h_\nu f'_y(x_\nu, y_\nu)] \approx \frac{1}{2} \int_{x_0 t}^{x_m} h(t) y''(t) \exp[\int_{x_0}^{x_m} f'_y(\tau, y(\tau))d\tau]dt.$$

Тут $h(t)$ – кусково-лінійна функція, значення якої в кожному вузлі x_n дорівнює h_n.

Підставляючи ці вирази у формулу (8.6), одержимо оцінку похибки на довільному кроці m:

$$z_m = z_0 \exp[\int_{x_0}^{x_m} f'_y(t, y(t))dt] +$$

$$+ \frac{1}{2} \int_{x_0 t}^{x_m} h(t) y''(t) \exp[\int_{x_0}^{x_m} f'_y(\tau, y(\tau))d\tau]dt, \quad m = 1,2,... \tag{8.7}$$

Вона складається з двох доданків, перший з яких обумовлений похибкою z_0 початкових даних. Якщо вони точні, z_0 =0, що і будемо припускати надалі.

Поява другого доданка пов'язана з відкиданням у рівності (8.5) залишкового члена формули Тейлора. Оцінимо цей доданок зверху.

Припустимо, що на відрізку $[x_0; x_0 + H]$ $|f(x,y)| \le M_1$, $|f_x(x,y)| \le M_2$, $|f_3(x,y)| \le M_3$.

Тоді $|y''(x)| = |f'_x + f'_y f| \le M_2 + M_1 M_3$,

$$|z_m| \le \max h_k A(x_m) = 0(\max_{0 \le k \le m} h_k), \text{ де} \qquad (8.8)$$

$$A(x_m) = \frac{1}{2}\int_{x_0}^{x_m} |y''(t)| \exp[\int_t^{x_m} f'_y(\tau, y(\tau))d\tau]dt \le$$

$$\le M_3(x_m - x_0) \le \frac{1}{2}(M_2 + M_1 M_3)(x_m - x_0).$$

З нерівності (8.8) випливає твердження.

Якщо $f(x,y)$ неперервна й обмежена в системі зі своїми першими похідними в області зміни своїх аргументів, то наближений розв'язок задачі (8.1) – (8.2), знайдений методом Ейлера, при $\max h_n \to 0$ збігається до точного розв'язку рівномірно на обмеженому відрізку $[x_0, x_0 + H]$ із сумарною похибкою $O(\max h_n)$.

Отже, метод Ейлера має перший порядок точності.

8.1.2 Схеми Рунге-Кутта другого порядку

Невисокий ступінь точності методу Ейлера визначається перш за все тим, що залишковий член формули (8.4) $R_n = O(h_n^2)$. Зажадаємо, щоб $R_n = O(h_n^3)$. За формулою Тейлора

$$y(x_{n+1}) = y(x_n) + h_n y'(x_n) + \frac{1}{2}h_n^2 y''(x_n) + r,$$

де залишковий член

$$r = \frac{1}{6}h_n^3 y'''(\xi_n) = O(h_n^3), \xi_n \in (x_n, x_{n+1}). \qquad (8.9)$$

Рівність (8.9) справедлива, якщо $y'''(x)$ обмежена на $[x_n, x_{n+1}]$. З (8.1) випливає

$$y''(x) = df(x,y)/dx = f'_x(x,y) + f'_y(x,y)f(x,y).$$

Тут треба обчислювати частинні похідні функції $f(x,y)$, що з причин, зазначених раніше, небажано. Щоб уникнути диференціювання, замінимо $y''(x_n)$ виразом

$$y''(x_n) = \frac{d}{dx}f(x_n, y_n) = \frac{f(x_n + \alpha h_n, y_n + \beta h_n) - f(x_n, y_n)}{\theta h_n},$$

де α, β, θ - деякі параметри. Тоді, якщо у формулі (8.9) відкинути залишок r, одержимо

$$y(x_{n+1}) \approx y(x_n) + h_n f(x_n, y_n) + \frac{h_n}{2\theta}[f(x_n + \alpha h_n, y_n + \beta h_n) - f(x_n, y_n)],$$

або $y_{n+1} = y_n + h_n[\gamma f(x_n, y_n) + \delta f(x_n + \alpha h_n, y_n + \beta h_n)]$, (8.10)

де $\delta = 1/(2\theta)$; $\gamma = 1 - 1(2\theta)$.

Параметри $\alpha, \beta, \gamma, \delta$ виберемо так, щоб розкладання точного розв'язку $y(x_{n+1})$ задачі (8.1)- (8.2) у вузлі x_n і його наближення y_{n+1}, що обчислюється за формулою (8.10), у ряди за степенями h_n, збігалися з точністю до нескінченно малої найбільш високого порядку щодо h_n.

Для одержання точного розв'язку використовуємо формулу (8.9)

$$y(x_{n+1}) = y(x_n) + f(x_n, y_n)h_n + \frac{1}{2}[f'_x(x_n, y_n) + f'_y(x_n, y_n)f(x_n, y_n)]h_n^2 + O(h_n^3),$$

аналогічно для наближеного розв'язку

$$y_{n+1} = y_n + h_n\{\gamma f(x_n, y_n) + \delta[f(x_n, y_n) + \alpha f'_x(x_n, y_n)h_{n+1} +$$
$$+ \beta f'_y(x_n, y_n)h_n]\} + O(h_n^3) = y_n + (\gamma + \delta)f(x_n, y_n)h_n +$$
$$+ \delta[\alpha f'_x(x_n, y_n) + \beta f'_y(x_n, y_n)]h_n^2 + O(h_n^3).$$

Припускаючи, що $y(x_n) = y_x$, і порівнюючи члени при однакових степенях h_n, одержимо

$$\gamma + \delta = 1, \quad \alpha\delta = 0.5, \quad \beta\delta = 0{,}5 f(x_n, y_n).$$

Для визначення чотирьох невідомих параметрів маємо три рівняння. Виразимо δ через інші параметри:

$$\alpha = \frac{1}{2\delta},\ \beta = \frac{1}{2\delta} f(x_n, y_n),\ \gamma = 1 - \delta.$$

Підставляючи ці значення у (8.10), одержимо однопараметричне сімейство двочленних схем Рунге-Кутта:

$$y_{n+1} = y_n + h_n[(1-\delta) f(x_n, y_n) + \delta f(x_n + \frac{h_n}{2\delta}, y_n + \frac{h_n}{2\delta} f(x_n, y_n))] \quad (8.11)$$

Відзначимо, що вибрати параметр δ так, щоб збігалися коефіцієнти у формулі Тейлора при h_n^3, неможливо.

Формула (8.11) завдяки своїй досить великій точності широко використовується в чисельних розрахунках, при цьому найчастіше беруть або $\delta = 1$, або $\delta = 0{,}5$.

Підставляючи в (8.11) $\delta = 1$, одержимо розрахункову формулу

$$y_{n+1} = y_n + h_n f(x_n + 0{,}5h_n, y_n + 0{,}5h_n f(x_n, y_n)), \quad (8.12)$$

відому як ***формулу вдосконаленого методу Ейлера.***

При використанні даного методу спочатку за формулою (8.5) обчислюємо наближене значення розв'язку при $x_{n+0,5} = x_n + 0{,}5h_n, y_{n+0,5} = y_n + 0{,}5h_n f(x_n, y_n)$. Після цього в знайденій точці визначаємо нахил інтегральної кривої: $y'_{n+0,5} = f(x_{n+0,5}, y_{n+0,5})$, а потім знаходимо значення

$$y_{n+1} = y_n + h_n f(x_{n+0,5}, y_{n+0,5}).$$

Покладаючи в (8.11) δ =0,5, одержимо

$$y_{n+1} = y_n + 0{,}5h_n[f(x_n, y_n) + f(x_{n+1}, y_n + h_n f(x_n, y_n))] \quad (8.13)$$

При використанні формули (8.13) спочатку обчислюється за методом Ейлера наближене значення $y^*_{n+1} = y_n + h_n f(x_n, y_n)$, потім нахил інтегральної кривої в новій точці $y'_{n+1} = f(x_{n+1}, y^*_{n+1})$ (рис. 8.4) Після цього визначається уточнене значення $y(x_{n+1})$

$$y_{n+1} = y_n + 0{,}5h_n[f(x_n, y_n) + f(x_{n+1}, y^*_{n+1})]. \quad (8.14)$$

Розрахункова схема (8.13) або (8.14) називається ***методом Ейлера-Коші***, або обчислювальним правилом типу ***предиктор-коректор***.

Для схеми (8.14) можна довести, що якщо $f(x,y)$ неперервна й обмежена разом зі своїми другими похідними, то розв'язок, отриманий за схемою (8.11) при будь-якому $\delta (0 < \delta \le 1)$ і при $\max h_n \to 0$, рівномірно збігається до точного розв'язку із сумарною похибкою $O(\max h_n^2)$. Отже, схема (8.11) має другий порядок точності.

8.1.3 Схеми Рунге-Кутта четвертого порядку

Методом Рунге-Кутта можна будувати схеми різного порядку точності.

Так схема ламаних Ейлера (8.5) є схемою Рунге-Кутта першого порядку точності. Найбільш уживані схеми четвертого порядку точності. Наведемо без виведення найчастіше використовувані з них:

$$y_{n+1} = y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$
$$k_1 = f(x_n, y_n),\ k_2 = f(x_n + 0{,}5h, y_n + 0{,}5hk_1),$$
$$k_3 = f(x_n + 0{,}5h, y_n + 0{,}5hk_2),$$
$$k_4 = f(x_n + h, y_n + hk_3). \quad (8.15)$$

Відзначимо, що формули більш високого порядку точності практично не вживаються через громіздкість, що зростає значно швидше, ніж точність формули.

Схеми Рунге-Кутта мають ряд переваг:

1) мають досить високий ступінь точності (за винятком схеми ламаних);

2) є явними, тобто значення y_{n+1} обчислюється за раніше знайденими значеннями;

3) допускають використання змінного кроку, що дає можливість зменшити його там, де функція швидко змінюється, і збільшити в іншому випадку;

4) є легко застосовними, тому що для початку розрахунку досить вибрати сітку x_n і задати значення $y_0 = f_0(x_0)$.

Зазначені властивості схем досить корисні при розрахунках на ЕОМ.

Оцінки похибок різних схем Рунге-Кутта пов'язані з максимумами модулів відповідних похідних функції $f(x,y)$ досить складними формулами типу (8.8). У зв'язку з цим при розв'язанні конкретної задачі виникає питання, якою з формул Рунге-Кутта доцільно користуватися і як обирати крок сітки.

Якщо *f(x,y)* неперервна й обмежена разом зі своїми четвертими похідними, то гарні результати дає схема четвертого порядку (8.15). Якщо права частина рівняння (8.1) не має зазначених похідних, то граничний порядок точності схеми (8.15) не може бути реалізований. Тоді доцільно користуватися схемами меншого порядку точності, що дорівнює порядкові наявних похідних, наприклад для двічі неперервно диференційованої функції *f(x,y)* —схемами (8.12) і (8.13).

Крок сітки варто вибирати настільки малим, щоб забезпечити необхідну точність розрахунку. З огляду на складність виразів залишкових членів (типу 8.8) апріорною (від лат. *a priori* — з попередніх) оцінкою точності для вибору кроку при практичних розрахунках не користуються, а заміняють її, наприклад, розрахунками зі згущенням сітки і дають апостеріорну (від лат. a posteriori - з наступного) оцінку точності.

Зауваження. Якщо функція *f(x,y)* досить гладка, але швидко змінюється на $[x_0; x_0 + H]$, схеми Рунге-Кутта як низького, так і високого порядку точності вимагають неприйнятно малого кроку для отримання задовільного результату. Для таких задач використовуються спеціальні методи, орієнтовані на даний вузький клас задач.

Одним із найбільш простих, широко застосовуваних і досить ефективних методів оцінки похибки й уточнення отриманих результатів у наближених обчисленнях з використанням сіток є правило Рунге.

Нехай маємо наближену формулу $\tilde{y}(x,h)$ для обчислення величини *y(x)* за значеннями на рівномірній сітці $h_n = h$ і залишковий член цієї формули має такий вигляд:

$$y(x) - \tilde{y}(x,h) = \psi(x)h^p + O(h^{p+1}). \qquad (8.16)$$

Виконаємо тепер розрахунок за тією самою наближеною формулою для тієї самої точки *x*, але використовуючи рівномірну сітку з іншим кроком *rh*, $r < 1$. Тоді отримане значення $\tilde{y}(x,h)$ пов'язане з точним значенням співвідношенням

$$y(x) - \tilde{y}(x,rh) = \psi(x)(rh)^p + O(rh)^{p+1}. \qquad (8.17)$$

Зауважимо, що $O((rh)^{p+1}) \approx O(h^{p+1})$.

Маючи два розрахунки на різних сітках, неважко оцінити величину похибки. Для цього віднімемо (8.17) з (8.16) і одержимо першу формулу Рунге

$$R \approx \psi(x)h^p = \frac{\tilde{y}(x,h) - \tilde{y}(x,rh)}{r^p - 1} + O(h^{p+1}) \qquad (8.18)$$

Перший з доданків є головним членом похибки. Таким чином, розрахунок за другою сіткою дозволяють оцінити похибку розрахунку за першою (з точністю до членів більш високого порядку).

Відзначимо, що при користуванні правилом Рунге практично досить застосувати формулу оцінки похибки у вигляді

$$R \approx |\tilde{y}(x,h) - \tilde{y}(x,rh)|, \qquad (8.19)$$

де $\tilde{y}(x,h), \tilde{y}(x,rh)$ - наближені значення розв'язку рівняння в одній і тій самій точці, отримані з кроком *h* і *h/2*. *При цьому необхідна точність може вважатися досягнутою, якщо величина R не перевищує заданої похибки у всіх збіжних вузлах.*

Застосування правила Рунге дає можливість одержувати результати досить високої точності, використовуючи обчислення за формулами низького порядку точності, тобто дозволяє уточнити результати.

Відзначимо, що правило Рунге застосовується й у випадках, якщо сітки з різним числом вузлів нерівномірні, але їх можна описати функціями *h(x),* відношення яких $h_1(x)/h_2(x) = r = const$. Величина відношення кроків r у правилі Рунге може бути будь-якою, але використовується вона найчастіше для цілого r, при цьому всі вузли менш докладної сітки повинні бути вузлами більш докладної. Особливо зручно згущати сітки вдвічі. У цьому випадку у вузлах, що є загальними для декількох сіток, можна уточнювати *y(x)* безпосередньо за правилом Рунге (8.18). Якщо розв'язок не уточнюється, а лише оцінюється його похибка, то використовується формула (8.19). Можна уточнити значення *y(x,h)* у всіх вузлах найдетальнішої сітки.

Використовуємо збіжнії вузли сіток для визначення виправлень до значень фікції $\tilde{y}(x,h)$:

$$\Delta_m = \frac{\tilde{y}(x_m,h) - \tilde{y}(x_m,rh)}{r^p - 1}, \quad m = n, n+2.$$

Значення виправлень в інших вузлах знайдемо інтерполяцією. Для рівномірних сіток покладемо $\Delta_{n+1} = 0.5(\Delta_n + \Delta_{n+2})$. Потім обчислимо уточнені значення $\tilde{y}(x_m,h) = \tilde{y}(x_m,h) + \Delta m, \quad m = n, n+1, n+2.$ Цей спосіб узагальнюється на будь-яке число сіток.

Відзначимо ще раз, що виконати уточнення, як правило, простіше, ніж скласти і використовувати схему високого порядку точності.

Похибки наведених схем Рунге-Кутта визначаються максимальними значеннями відповідних похідних.

Оцінку похибки легко одержати для окремого випадку правої частини диференціального рівняння $f(x,y) \equiv f(x)$.

У цьому випадку розв'язок рівняння може бути зведений до квадратури й усі схеми різницевого розв'язку переходять у формули чисельного інтегрування. Наприклад, схема (8.14) набирає вигляду

$$y_{n+1} = y_n + \frac{h_n}{2}\left[\varphi(x_n) + \varphi(x_{n+1})\right],$$

тобто має вигляд формули трапецій, а схема (8.15) переходить у схему

$$y_{n+1} = y_n + \frac{h_n}{2}\left[\varphi(x_n) + 4\varphi(x_n + h_n/2) + \varphi(x_n + h_n)\right],$$

що являє собою формулу Симпсона з кроком $\frac{h_n}{2}$.

Мажорантні оцінки похибок формул трапецій і Симпсона відомі. З них видно, що точність схем Рунге-Кутта досить висока.

Приклад. Знайти наближений розв'язок задачі Коші для звичайного диференціального рівняння (ОДУ) 1 порядку

y'(t)=2ty, t0=0, T=1, y(0) =1 та оцінити його похибку.

Вихідні дані:

Права частина: $f(t,y) := 2 \cdot t \cdot y$. Початкове значення: $y_0 := 1$.

Кінці відрізка: $t0 := 0 \quad T := 1$ Крок сітки: $h := 0.2$.

Число вузлів сітки: $N := \frac{T - t0}{h} \quad N = 5$.Функція, що реалізує явний метод Ейлера, повертає вектор розв'язку:

$$\begin{aligned} &\text{eyler}(f, y0, t0, h, N) := \\ &\quad y_0 \leftarrow y0 \\ &\quad \text{for } i \in 0..N-1 \\ &\quad\quad y_{i+1} \leftarrow y_i + h \cdot f(t0 + i \cdot h, y_i) \\ &\quad y \end{aligned}$$

Вхідні параметри:

f - функція правої частини;

y0 - початкове значення;
t0 - початкова точка відрізка;
h - крок сітки;
N - число вузлів сітки.
Обчислення розв'язку за методом Ейлера:

$yE := eyler(f, y_0, t0, h, N)$

Обчислення розв'язку за методом Рунге-Кутта 4 порядку точності:

$yRK4 := rkfixed(y, t0, T, N, f)$;

вхідні параметри:
y - вектор початкових значень;
t0- початкова точка відрізка;
T - кінцева точка відрізка;
N - число вузлів сітки;
f - функція правої частини. Функція **rkfixed** повертає матрицю, перший стовпець якої містить вузли сітки, а другий - наближений розв'язок у цих вузлах.

Точний розв'язок: $Y(t) := e^{t^2}$.
Точний розв'язок у вузлах сітки:

$i := 0..N \quad t_i := t0 + i \cdot h \quad yt_i := Y(t_i)$

Розв'язок за методом Ейлера

$$yE = \begin{bmatrix} 1 \\ 1 \\ 1.08 \\ 1.253 \\ 1.553 \\ 2.051 \end{bmatrix}$$

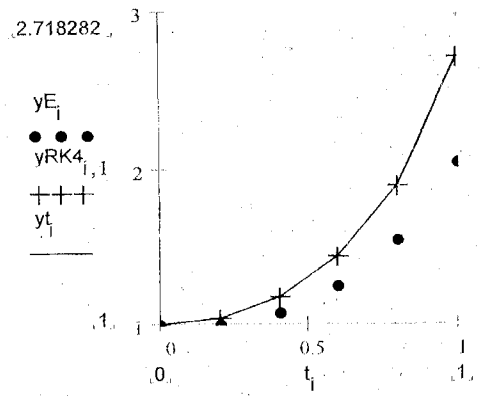
Розв'язок за методом Рунге-Кутта

$$yRK4 = \begin{bmatrix} 0 & 1 \\ 0.2 & 1.041 \\ 0.4 & 1.174 \\ 0.6 & 1.433 \\ 0.8 & 1.896 \\ 1 & 2.718 \end{bmatrix}$$

Точний розв'язок

$$yt = \begin{bmatrix} 1 \\ 1.040811 \\ 1.173511 \\ 1.433329 \\ 1.896481 \\ 2.718282 \end{bmatrix}$$

Графіки наближених і точних розв'язків



Обчислення похибки за правилом Рунге:
Обчислення наближених розв'язків із кроком *h*/2:

$h2 := \frac{h}{2} \qquad N2 := \frac{T - t0}{h2} \qquad N2 = 10$

$yEh2 := eyler(f, y_0, t0, h2, N2) \qquad yRK4h2 := rkfixed(y, t0, T, N2, f)$

Обчислення похибок: $i := 0..N$

$$zE_i := \left| yE_i - yEh2_{2 \cdot i} \right| \qquad zRK4_i := \frac{\left| (yRK4^{\langle 1 \rangle})_i - (yRK4h2^{\langle 1 \rangle})_{2 \cdot i} \right|}{15}$$

$max(zE) = 0.284 \qquad max(zRK4) = 1.088 \cdot 10^{-5}$

Приклад реалізації алгоритму методу Рунге-Кутта четвертого порядку з заданою точністю ε на псевдокоді.

```
f(x,y):
    //повертає значення заданої похідної
при заданих x та y
end
//BeginValue - початкові умови
//YF - відповідь - масив значень функції
//h - крок
//n - кількість точок розбиття
```

```
//eps - точність розрахунку
//Метод Рунге-Кутта 4-го порядку
SolveRungeKutt(BeginValue,YF,h,n,eps):
    repeat
      for j:=1 to 2 do
     Y[j,1]:=BeginValue
     for i:=1 to n do
      x:=a+(i-1)*h
      k1:=f(x,Y[j,i])
    k2:=f((x+h/2),(Y[j,i]+h*k1/2))
    k3:=f((x+h/2),(Y[j,i]+h*k2/2))
      k4:=f((x+h),(Y[j,i]+h*k3))
Y[j,i+1]:=Y[j,i]+h*(k1+2*k2+2*k3+k4)/6
      done
      if j=1 then
      h:=h/2
      n:=round((b-a)/h)
          fi
      done
      maxr:=0
      for i:=1 to 10 do
    r:=abs(Y[1][((n*i) div 20)+1]-
     -Y[2][((n*i) div 10)+1])/15;
          if r>maxr then
            maxr:=r
          fi
      done
      if maxr<eps then
          for i:=1 to 10 do
    YF[i]:=Y[2][((n*i) div 10)+1]
          done
      fi
  until maxr<eps
end
```


8.2 Багатокрокові методи

8.2.1 Метод прогнозу і корекції

Підправивши схему Ейлера в середній точці одержимо схему прогнозу

$$p_{n+1} = y_{n-1} + 2hy_n', \tag{8.20}$$

де p_{n+1} – наближене значення y_{n+1}. Використовувати формулу (8.20) не можна через те, що схема прогнозу нестійка. З цієї причини використовуємо схему корекції

$$c_{n+1} = y_n + \frac{h}{2}\left(y_n' + y_{n+1}'\right) \tag{8.21}$$

Для оцінки похибки корекції розкладемо в ряд Тейлора в околі точки x_n корекцію

$$c_{n+1} \approx y_n + \frac{h}{2}(y_n' + y_n' + hy_n'' + \frac{h^2}{2}y_n''') = y_n + hy_n' + \frac{h^2}{2}y_n'' + \frac{h^3}{4}y_n'''$$

та саму функцію $\quad y_{n+1} \approx y_n + hy_n' + \frac{h^2}{2}y_n'' + \frac{h^3}{3!}y_n'''$.

Віднявши ці два розкладання, отримаємо

$$y_{n+1} - c_{n+1} = -\frac{h^3}{12}y''' . \tag{8.22}$$

Для оцінки похибки прогнозу розкладемо в ряд Тейлора в околі точки x_n прогноз

$$p_{n+1} \approx y_n - hy'_n + \frac{h^2}{2}y''_n - \frac{h^3}{3!}y'''_n + 2hy'_n \text{ та саму функцію}$$

$$y_{n+1} \approx y_n + hy'_n + \frac{h^2}{2}y''_n + \frac{h^3}{3!}y'''_n .$$

Віднявши ці два розкладання , отримаємо похибку прогнозу

$$y_{n+1} - p_{n+1} = \frac{h^3}{3}y''' . \qquad (8.23)$$

З цих оцінок зрозуміло, що прогноз наближає розв'язок з недостачею, а корекція з надлишком. Отже, точне значення розв'язку лежить між прогнозом і корекцією. На будь-якому кроці можна оцінити точність розв'язку . Якщо задається її значення ε , то $|c_{n+1} - p_{n+1}| < \varepsilon$.

Віднімаючи рівності (8.23) та (8.22), маємо

$$c_{n+1} - p_{n+1} = \frac{5}{12}h^3 y''' .$$

Уточнюємо розв'язок, виходячи з формули (8.22)

$$y_{n+1} = c_{n+1} + \frac{p_{n+1} - c_{n+1}}{5} . \qquad (8.24)$$

Відтак формула (8.24) завершує схему прогнозу і корекції.

Приклад 1 Розв'язати задачу Коші для диференціального рівняння другого порядку:

$$y'' - y' - 2(1-t) = 0;$$
$$y(0) = 1; y'(0) = 1.$$

Розв'язання. 1 Введемо нові змінні :

U=y; V= y'. Тоді, обравши сітку $t_0 = 0, t_i = t_0 + ih,\ i = 1,2,..., h = 0{,}05$, визначимо на ній сіткові функції U, V :

$$V' - V - 2(1-t) = 0$$

$$\begin{cases} V' = V + 2(1-t); V = f_2(t,U,V) \\ U' = V = f_1(t,U,V) \end{cases}$$

$$\begin{vmatrix} U_0 = 1; \\ V_0 = 1; \end{vmatrix} \begin{vmatrix} U'_0 = f_1(x_0,U_0,V_0) \\ V'_0 = f_2(x_0,U_0,V_0) \end{vmatrix} \begin{vmatrix} V'_0 = V_0 + 2 - 2\times 0 = 3; \\ V''_0 = V'_0 - 2 = 1; \\ V'''_0 = V''_0 = 1; \end{vmatrix} \quad U'_0 = V_0 = 1;$$

$$U''_0 = V'_0 = 3;$$

$$U'''_0 = V''_0 = 1;$$

2 Знаходимо U_1 і V_1 ($t = t_1 = 0{,}05$) за допомогою розкладання в ряд Тейлора :

$$U_1 = U_0 - hU'_0 + \frac{h^2}{2}U''_0 + \frac{h^3}{6}U'''_0 = 1 + 0.05 * 1 + \frac{0.05^2}{2}3 + \frac{0.05^3}{6} = 1.0537708;$$

$$V_1 = V_0 + hV'_0 + \frac{h^2}{2}V''_0 + \frac{h^3}{6}V'''_0 = 1 + 0.05 * 3 + \frac{0.05^2}{2} + \frac{0.05^3}{6} = 1.1512708;$$

3 З диференціального рівняння знаходимо U'_1 і V'_1 :

$$V'_1 = V_1 + 2(1-t) = 1.1512708 + 2 - 2 * 0.1 = 3.0512708;$$

$$U'_1 = V_1.$$

Визначаємо прогноз розв'язку в точці t_2

$$P_2 = U_0 + 2hU'_1; P_2^* = V_0 + 2hV'_1 .$$

Тимчасово покладаючи $U_2 = P_2$; $V_2 = P_2^*$, можна отримати

$$U'_2 = f_1(x_2, P_2, P_2^*),$$
$$V'_2 = f_2(x_2, P_2, P_2^*), \quad t_2 = 2h,$$
$$V'_2 = P_2^* + 2(1-t_2).$$

Обчислюємо корекцію $C_2^* = V_1 + \frac{h}{2}(V'_1 + V'_2)$. Тепер можна визначити V2

$$V_2 = C_2^* + \frac{1}{5}(P_2^* - C_2^*);\ U'_2 = V_2;\ C_2 = U_1 + \frac{h}{2}(U'_1 + U'_2).$$

І нарешті на даному кроці уточнюємо розв'язок

$$U_2 = C_2 + \frac{1}{5}(P_2 - C_2).$$

Похибка обчислення $|P_2 - C_2| < \varepsilon$.

Приклад 2 Використовуючи метод прогнозу і корекції, реалізувати алгоритм розв'язку крайової задачі для звичайного диференціального рівняння з точністю $\varepsilon = 0{,}0001$ на псевдокоді.

$$\begin{cases} y'' + \dfrac{y'}{x} + 2y = x \\ y(0{,}7) = 0{,}5 \\ 2y(1) + 3y'(1) = 1{,}2 \end{cases}$$

Розв'язання

Після заміни

$y = u$

$y' = v$

отримаємо (користуючись системою символічної математики Derive)

```
vo1:=-0.3-vo/(0.7)
vo2:=1-2*vo-(vo1*(0.7)-vo)/(0.7)^2
vo3:=-2*vo1-(vo2*(0.7)^3-2*(0.7)*(vo1*(0.7)-
vo))/(0.7)^4
u1:=0.5+0.3*vo+vo1*(0.3)^2/2+vo2*(0.3)^3/6
(24018*vo+48307)/98000
v1:=vo+0.3*vo1+vo2*(0.3)^2/2+vo3*(0.3)^3/6
(2099500*vo-129339)/3430000

2*u1+3*v1=1.2
SOLVE(2*u1+3*v1=1.2,vo);
Simp(#9)
[vo=1122527/7979760]; Approx(#10)
[vo=5122/36411=0.140671]
```

тоді розв'яжемо систему:

$$\begin{cases} v' = x - 2u - \dfrac{v}{x} \\ u' = v \\ u_0 = 0{,}5 \\ v_0 = 0{,}140671 \end{cases}$$

```
//x0 - початкове значення
//xn - кінцеве значення
//h - початковий крок
//X,Y - масиви результатів
Prognose_Correction(x0,u0,v0,h,X,Y):
      1     u01:=v01;
      2     v01:=x0-2*u0-v0/x0
      3     u02:=v01
4     v02:=1-2*u01-((v01*x0-v0)/sqr(x0))
      5     u03:=v02
6     v03:=-2*u02-((v02*x0*sqr(x0)-
2*v01*sqr(x0)+2*v0*x0)/sqr(sqr(x0)))
7     u1:=u0+h*u01+h*h*u02/2+h*h*h*u03/6
8     v1:=v0+h*v01+h*h*v02/2+h*h*h*v03/6
      9     i:=0
      10    while (x1<=xN) do
      11        i++
      12        repeat
      13            x1:=x0+i*h
      14            x2:=x0+(i+1)*h
      15            v11:=x1-2*u1-v1/x1
```

```
16          u11:=v1
17          p2:=u0+2*h*u11
18          p2z:=v0+2*h*v11
19          u2:=p2
20          v2:=p2z
21          u21:=p2z
22          v21:=x2-2*p2-p2z/x2
23          c2z:=v1+h*(v11+v21)/2
24          v2:=c2z+(p2z-c2z)/5
25          u21:=v2
26          c2:=u1+h*(u11+u21)/2
27          u2:=c2+(p2-c2)/5
28          if abs(p2-c2)<eps then
29              X[i]:=x2;
30              Y[i]:=u2;
31          fi
32          if abs(p2-c2)>e then
33              h:=h/2;
34          fi
35      until abs(p2-c2)<eps;
36  done //while
end
```

8.2.2 Методи Адамса

На відміну від однокрокових методів, у яких числовий розв'язок одержують тільки з диференціального рівняння і початкової умови, алгоритми Адамса складаються з двох

частин: перша з них – стартова процедура для визначення $y_1,...,y_{k-1}$ (наближені значення точного розв'язку в точках $x_0+h,...,x_0+(k-1)h$), а друга – багатокрокова формула для одержання наближеного значення точного розв'язку $y(x_0+kh)$. Потім ця формула застосовується рекурсивно для того, щоб за числовим розв'язком на k послідовних кроках обчислити $y(x_0+(k+1)h)$ і т.д.

Стартові значення можна одержати декількома способами. Дж. К. Адамс обчислював їх за допомогою розкладання точного розв'язку в ряд Тейлора. Інший спосіб полягає у використанні якого-небудь однокрокового методу, наприклад, Рунге-Кутта. Стартові значення часто також обчислюють методами Адамса низького порядку з дуже малим кроком.

Розглянемо чисельні методи розв'язання задачі Коші (8.1)-(8.2), які можуть бути задані формулою

$$y_{n+k}=F(f;x_{n+k},x_{n+k-1},...,x_n;y_{n+k},y_{n+k-1},...,y_n). \qquad (8.25)$$

Тут значення розв'язку y_{n+k} в точці x_{n+k} визначається через значення розв'язку в k точках, що передують x_{n+k}. Такий метод називається k - кроковим.

З класу (8.25) виділимо багатокрокові методи вигляду

$$\sum_{i=0}^{k}\alpha_i y_{n+i}=h\sum_{i=0}^{k}\beta_i f(x_{n+i},y_{n+i}),\ \alpha_k\neq 0, \qquad (8.26)$$

застосовувані на сітці з постійним кроком

$$x_n=x_0+nh,\ n=0,1,2,...,N_h\quad N_h=[X/h] \qquad (8.27)$$

Різниця між найбільшим і найменшим значеннями індексу невідомої функції y_n, що входить у рівняння (8.26), дорівнює k. Тому співвідношення (8.26) є різницевим рівнянням k-го порядку, загальний розв'язок якого залежить від k параметрів. Щоб виділити єдиний розв'язок цього рівняння, необхідно задати k додаткових умов на функцію y_n. Цими додатковими умовами є значення функції y_n при $n=0,1,...,k$-1:

$$y_0=g_0,\ y_1=g_1,\ ....,\ y_{k-1}=g_{k-1}, \qquad (8.28)$$

які передбачаються відомими.

Використовуючи значення (8.28), з рівняння (8.26) при n=0 можна знайти y_k, потім, використовуючи значення $g_1,...,g_{k-1},y_k$ і покладаючи в (8.26) n =1, знайти y_{k+1} і т.д. Таким чином, даний метод чисельного розв'язання диференціального рівняння полягає в розв'язанні різницевої задачі Коші для різницевого рівняння (8.26) і початкових умов (8.28).

Якщо шуканий розв'язок y_{n+k} входить до правої частини цього рівняння, що буває, коли $\beta_k \neq 0$, то формула (8.26) визначає неявний метод. Якщо $\beta_k = 0$, то шуканий розв'язок до правої частини не входить і рівняння (8.26) може бути розв'язане відносно y_{n+k}. У цьому випадку формула (8.26) визначає явний метод.

Виведемо групу явних багатокрокових формул. Для точок сітки введемо позначення $x_i = x_0 + ih$ і припустимо, що нам відомі числові наближені значення $y_n, y_{n-1},..., y_{n-k+1}$ точного розв'язку $y(x_n),..., y(x_{n-k+1})$ задачі (8.1)-(8.2).

З диференціального рівняння випливає

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(t, y(t))dt . \qquad (8.29)$$

До правої частини (8.29) входить шуканий розв'язок $y(x)$. Але оскільки нам відомі його наближені значення $y_n, y_{n-1},..., y_{n-k+1}$, то ми маємо також і величини

$$f_i = f(x_i, y_i) \ \ при\ i = n-k+1,..., n, \qquad (8.30)$$

а тому природно замінити функцію $f(t, y(t))$ в (8.29) інтерполяційним многочленом, що проходить через точки $\{(x_i, f_i) \mid i = n-k+1,..., n\}$. Його можна виразити через скінченні різниці вигляду $\Delta^0 f_n = f_n$, $\Delta^{j+1} f_n = \Delta^j f_n - \Delta^j f_{n-1}$ у такий спосіб:

$$p(t) = p(x_n + sh) = \sum_{j=0}^{k-1} (-1)^j \binom{-s}{j} \nabla^j f_n \qquad (8.31)$$

(інтерполяційна формула Ньютона). Тоді чисельний аналог (8.29) задається формулою $y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} p(t)dt$, або після підстановки (8.31)

$$y_{n+1} = y_n + h\sum_{j=0}^{k-1} \gamma_j \Delta^j f_n , \qquad (8.32)$$

де коефіцієнти γ_j задовольняють рівність

$$\gamma_j = (-1)^j \int_0^1 \binom{-s}{j} ds . \qquad (8.33)$$

Числові значення цих коефіцієнтів наведені в таблиці 8.1.

Окремі випадки формули (8.32).

Для $k = 1,2,3,4$, виразивши різниці назад через f_{n-j}, одержимо такі формули:

$k=1$: $y_{n+1} = y_n + hf_n$

$k=2$: $y_{n+1} = y_n + h(\frac{3}{2}f_n - \frac{1}{2}f_{n-1})$

$k=3$: $y_{n+1} = y_n + h(\frac{23}{12}f_n - \frac{16}{12}f_{n-1} + \frac{5}{12}f_{n-2})$

$k=4$: $y_{n+1} = y_n + h(\frac{55}{24}f_n - \frac{59}{24}f_{n-1} + \frac{37}{24}f_{n-2} - \frac{9}{24}f_{n-3})$

Зауваження. Для $k = 1$ ми маємо явний метод Ейлера.

Таблиця 8.1

j	0	1	2	3	4	5	6	7	8
γ_j	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$	$\frac{19087}{60480}$	$\frac{5257}{17280}$	$\frac{1070017}{3628800}$

Похибка апроксимації явного двокрокового методу Адамса має другий порядок.

Неявний двокроковий метод Адамса виглядає так:

$$\frac{y_{n+1}-y_n}{h}=\frac{1}{6}[\frac{5}{2}f_{n+1}+4f_n-\frac{1}{2}f_{n-1}],\quad n=1,2,....$$

Похибка апроксимації має третій порядок .

8.2.3 Стійкість різницевих методів

Уведемо поняття стійкості різницевого методу. Для цього розглянемо різницеве рівняння багатокрокового методу

$$\sum_{k=0}^{m}\frac{a_k}{h}y_{n-k}=\sum_{k=0}^{m}b_k f(x_{n-k},y_{n-k})\ ,\quad n=m,m+1,...\ . \qquad (8.34)$$

Однорідне різницеве рівняння, що відповідає (8.34), має вигляд

$$\sum_{k=0}^{m}a_k y_{n-k}=0\ . \qquad (8.35)$$

Вважають, що *рівняння (8.35) є стійким за початковими даними, якщо існує постійна M, що не залежить від n, така, що при будь-яких початкових даних $y_0,y_1,..,y_{m-1}$ здійснюється нерівність*

$$|y_n|\le M\max_{0\le j\le m-1}|y_j|\ ,\ n=m,m+1,...\ .$$

Питання стійкості за початковими даними вирішується шляхом розгляду коренів так званого характеристичного рівняння, одержуваного з (8.35), якщо розв'язок цього рівняння шукати у вигляді $y_{n-k}=q^{n-k}$. Підставляючи таке y_{n-k} в (8.35) і скорочуючи на q^{n-m} , одержимо характеристичне рівняння для визначення q

$$a_0q^m+a_1q^{m-1}+...+a_{m-1}q+a_m=0\ . \qquad (8.36)$$

Теорема 1 Для стійкості рівняння (8.35) за початковими даними необхідно і достатньо, щоб виконувалася так звана умова коренів: *усі корені $q_1,q_2,..,q_m$ характеристичного рівняння знаходилися всередині або на границі одиничного кола*

комплексної площини, причому на границі не повинно бути кратних коренів.

Теорема 2 Нехай $0\le nh\le T$, умова коренів виконана, $|y_i-v(x_i)|\to 0$ при $h\to 0$, $i=0,1,2,..,m-1$, і різницеве рівняння (8.34) апроксимує вихідне диференціальне рівняння (8.1). Тоді розв'язок різницевої задачі (8.34) збігається при $h\to 0$ до розв'язку вихідної задачі (8.1).

Інакше кажучи, ***з апроксимації і стійкості за початковими даними випливає збіжність на обмеженому відрізку*** $[0,T]$.

Сформульована умова стійкості, що базується на аналізі розміщення коренів характеристичного рівняння (8.36), є досить загальною. Конкретизуємо питання про стійкість різницевого рівняння стосовно до асимптотично стійких розв'язків рівняння (8.1). Нехай $f(x,y(x))=\lambda\cdot y(x)$, $\lambda<0$, тобто

$$\frac{dy}{dx}=\lambda\cdot y(x)\,. \qquad (8.37)$$

Розв'язок цього рівняння асимптотично стійкий, тобто для будь-яких $x>0$ справедлива оцінка

$$|y(x+h)|\le|y(x)|\ . \qquad (8.38)$$

Логічно вимагати, щоб і різницеве рівняння давало розв'язок, що задовольняє властивість (8.38). Використовуючи явний метод Ейлера першого порядку апроксимації, одержимо різницевий аналог (8.37)

$$\frac{y_{n+1}-y_n}{h}=\lambda\cdot y_n\ ,\quad n=0,1,2,... \qquad ,(8.39)$$

або $y_{n+1}=(1+h\lambda)y_n$, тобто $q=1+h\lambda$.

Оцінка (8.38) буде виконана для (8.39) лише за умови $|q|\le 1$, оскільки тоді $|y_{n+1}|\le|y_n|$. З $|q|\le 1$ випливає обмеження на крок h: $0\le h\le\dfrac{2}{|\lambda|}$.

Різницевий метод (8.34) називається абсолютно стійким, якщо стійкість має місце при будь-яких $h>0$, й умовно

стійким, якщо вона може бути забезпечена тільки введенням обмежень на крок h.

Як приклад абсолютно стійкого методу традиційно розглядається неявний метод Ейлера, що має перший порядок апроксимації

$$\frac{y_{n+1}-y_n}{h}=\lambda\cdot y_{n+1}\ . \qquad (8.40)$$

З (8.40) випливає $y_{n+1}=\dfrac{y_n}{1-h\lambda}=\dfrac{y_n}{1+h|\lambda|}$, тобто $|q|=\dfrac{1}{1+h|\lambda|}<1$ завжди, при будь-яких $h>0$.

Умовна стійкість приводить до необхідності вибирати малі значення кроку h, що є недоліком явного методу. Неявний метод, позбавлений даного обмеження, має інший досить істотний недолік, обумовлений необхідністю розв'язувати на кожному кроці алгебраїчне рівняння (або систему рівнянь, у загальному випадку нелінійних).

Запишемо різницеві рівняння (8.34) для задачі (8.37)

$$\sum_{k=0}^{m}(a_k-\mu b_k)y_{n-k}=0\ ,\ n=m,m+1,...\ , \qquad (8.41)$$

де $\mu=h\lambda$ - у загальному випадку комплексний параметр.

Характеристичне рівняння для (8.41) має вигляд

$$\sum_{k=0}^{m}(a_k-\mu b_k)q^{m-k}=0\ . \qquad (8.42)$$

При малих μ корені (8.42) близькі до коренів (8.35).

Областю стійкості методу (8.34) називається множина точок комплексної площини $\mu=h\lambda$, для яких метод, що застосований до рівняння спеціального вигляду (8.37), є стійким.

Для явного методу Ейлера умова стійкості $|1+\mu|\le 1$ при комплексному $\mu=\mu_0+i\mu_1$ $(\mu_0=\mathrm{Re}\mu,\ \mu_1=\mathrm{Im}\mu)$ виглядає в такий спосіб: $(\mu_0+1)^2+\mu_1^2\le 1$, тобто областю стійкості є коло

одиничного радіуса, центр якого знаходиться в точці $(-1;0)$ комплексної площини.

Для неявного методу Ейлера умова $\dfrac{1}{|1-\mu|}\le 1$ відповідає нерівності $(1-\mu_0)^2+\mu_1^2\ge 1$, тобто областю стійкості є зовнішність кола одиничного радіуса з центром у точці $(1;0)$.

Різницевий метод називається $A-$стійким, якщо область його стійкості включає ліву півплощину $\mathrm{Re}\mu<0$ (або $h\cdot\mathrm{Re}\,\lambda<0$). Варто звернути увагу на те, що рівняння (8.37) асимптотично стійке при $\mathrm{Re}\,\lambda<0$. Отже, $A-$ стійкий різницевий метод є абсолютно стійким (тобто стійким при будь-яких $h>0$), якщо стійким є розв'язок вихідного диференціального рівняння.

З вищезазначеного видно, що неявний метод Ейлера має властивість $A-$стійкості, а явний метод – не має.

Розглянемо ще один неявний метод більш високого порядку апроксимації (другого):

$$\frac{y_{n+1}-y_n}{h}=\frac{1}{2}\left[f(x_{n+1},y_{n+1})+f(x_n,y_n)\right]\ . \qquad (8.43)$$

Цей метод виходить заміною інтеграла від правої частини (8.1) за формулою трапецій. Стосовно рівняння (8.37) метод (8.43) виглядає так: $y_{n+1}=\dfrac{1+0{,}5\mu}{1-0{,}5\mu}y_n$, тобто $\left|\dfrac{1+0{,}5\mu}{1-0{,}5\mu}\right|\le 1$, якщо $\mathrm{Re}\mu\le 0$. Отже, метод (8.43) належить до $A-$ стійких методів.

Доведеними є такі положення:

-серед методів (8.43) не існує явних $A-$стійких методів;

-серед неявних лінійних багатокрокових методів немає $A-$ стійких методів, що мають порядок точності вище другого.

$A-$стійкі різницеві схеми досить ефективні при розв'язанні так званих жорстких систем рівнянь, оскільки ці методи не накладають обмежень на крок h. Розглянемо докладніше це твердження.

8.2.4 Жорсткі диференціальні рівняння

Система звичайних диференціальних рівнянь

$$\frac{d\vec{y}}{dx} = A\vec{y} \qquad (8.44)$$

з незалежною від x матрицею $A(m\times m)$ називається жорсткою, якщо $\mathrm{Re}\,\lambda_k < 0$, $k = 1,2,..m$ і відношення $s = \dfrac{\max\limits_{1\le k\le m}\left|\mathrm{Re}\,\lambda_k\right|}{\min\limits_{1\le k\le m}\left|\mathrm{Re}\,\lambda_k\right|}$ велике, де λ_k - власні числа матриці A. Величина s називається числом жорсткості. Якщо матриця A залежить від x, то і λ_k - залежать від x, тоді вводиться змінне число жорсткості

$$s(x) = \frac{\max\limits_{1\le k\le m}\left|\mathrm{Re}\,\lambda_k(x)\right|}{\min\limits_{1\le k\le m}\left|\mathrm{Re}\,\lambda_k(x)\right|}$$

і оперують з величиною $\sup s(x)$ на відрізку інтегрування.

Відмінною рисою жорстких систем є наявність у їхньому розв'язку як швидко, так і повільно спадних компонентів. При $x > 0$ розв'язок системи практично визначається повільно спадним компонентом, однак, якщо скористатися явними різницевими методами, то швидко спадна складова буде негативно впливати на стійкість, і в результаті весь розрахунок необхідно вести з малим кроком інтегрування. При використанні ж неявних методів обмеження на крок зняті, і його величину визначають з умови досягнення потрібної точності, не хвилюючись особливо за стійкість.

При розв'язанні жорстких систем диференціальних рівнянь добре зарекомендував себе метод Гіра, що належать до чисто неявних багатокрокових різницевих методів, загальна формула яких виглядає так: $\sum\limits_{k=0}^{m} a_k y_{n-k} = hf(x_n, y_n)$,

тобто розглядається частковий варіант методу (8.43), коли $b_1 = b_2 = ... = b_m = 0$, а $b_0 = 1$.

При $m = 1$ і $a_0 = 1, a_1 = -1$ маємо $y_n - y_{n-1} = hf(x_n, y_n)$, тобто неявний метод Ейлера. При $m = 2$ і $m = 3$ методи виглядають так:

$$\frac{3}{2}y_n - 2y_{n-1} + \frac{1}{2}y_{n-2} = hf(x_n, y_n)\ ; \qquad (8.45)$$

$$\frac{11}{6}y_n - 3y_{n-1} + \frac{3}{2}y_{n-2} - \frac{1}{3}y_{n-3} = hf(x_n, y_n)\ . \qquad (8.46)$$

Різницеве рівняння (8.45) має другий порядок точності, а (8.46) - третій. Щоб знайти область стійкості методу, варто записати аналогічні рівняння для диференціального рівняння (8.46). Наприклад, (8.45) набере вигляду

$$\frac{3}{2}y_n - 2y_{n-1} + \frac{1}{2}y_{n-2} = \mu \cdot y_n\ .$$

Відповідне характеристичне рівняння запишеться в такий спосіб:

$$(\frac{3}{2} - \mu)q^2 - 2q + \frac{1}{2} = 0\ . \qquad (8.47)$$

Потрібно визначити область комплексної площини $\mu = \mu_0 + i\mu_1$, у точках якої обидва корені (8.47) за модулем менше одиниці. Виявляється, що ця область цілком розміщується у правій півплощині і метод (8.45) є A – стійким.

8.3 Метод скінченних різниць

Основний зміст методу можна легко пояснити на прикладі розв'язання задач в одновимірній області.

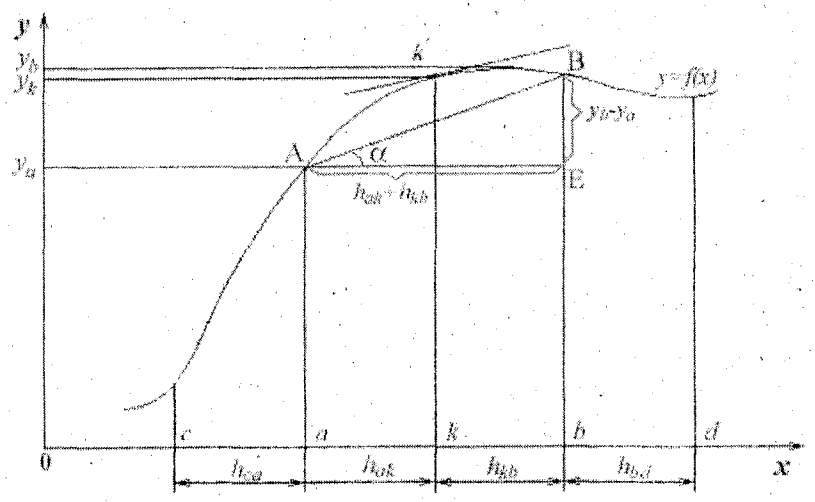


Рис. – 8.3

Виразимо похідну функції $y=f(x)$ лінійною комбінацією значень цієї функції у визначених точках розглянутого проміжку зміни незалежних змінних, які називаємо вузлами. Існує кілька способів вираження похідної подібним чином. Наприклад, першу похідну функції $f(x)$ у вузлі k (рис. 8.3) можна виразити такими скінченними різницями (дивись розділ 6):

$$\left(\frac{dy}{dx}\right)_k \approx \begin{cases} \dfrac{y_k - y_a}{h_{ak}}; & (8.48) \\ \dfrac{y_b - y_k}{h_{kb}}; & (8.49) \\ \dfrac{y_b - y_a}{h_{ak} + h_{kb}}. & (8.50) \end{cases}$$

Відстань (крок) між вузлами беруть однаковою $h_{ak} = h_{kb} = h$ і формула (8.50) записується у вигляді

$$\left(\frac{dy}{dx}\right)_k = \frac{y_b - y_a}{2h}. \qquad (8.51)$$

Другу похідну можна наближено виразити (мал. 8.3), застосовуючи формулу (8.51) при $h_{ca} = h_{ak} = h_{kb} = h_{bd} = h$ в такий спосіб:

$$\left(\frac{d^2y}{dx^2}\right)_k = \frac{d}{dx}\left(\frac{dy}{dx}\right)_k = \frac{\left(\frac{dy}{dx}\right)_b - \left(\frac{dy}{dx}\right)_k}{2h}; \qquad (8.52)$$

$$= \frac{\frac{y_d - y_k}{2h} - \frac{y_k - y_a}{2h}}{2h} = \frac{y_d - 2y_k + y_c}{4h^2}.$$

Застосовується також формула для другої похідної, отримана на основі виразів (8.48), (8.49) для однобічних різниць (при $h_{kb} = h_{ak} = h$):

$$\left(\frac{d^2y}{dx^2}\right)_k = \frac{\left(\frac{dy}{dx}\right)_b - \left(\frac{dy}{dx}\right)_k}{h} = \frac{\frac{y_b - y_k}{h} - \frac{y_k - y_a}{h}}{h}; \qquad (8.53)$$

$$= \frac{y_b - 2y_k + y_a}{h^2}.$$

Розв'язання крайової задачі методом скінченних різниць зводиться до обчислення значень шуканої функції в обраних вузлах шляхом розв'язання відповідної системи лінійних алгебраїчних рівнянь.

Докладно розглянемо різницевий метод на прикладі крайової задачі для лінійного рівняння другого порядку з крайовими умовами першого роду

$$u''(x) - p(x)u(x) = f(x), \qquad (8.54)$$

$$u(a) = \alpha \ , \ u(b) = \beta \, . \qquad (8.55)$$

Уведемо на $[a,b]$ сітку $a = x_0 < x_1 < x_2 < \ldots < x_N = b$, що для спрощення викладень будемо вважати рівномірною. Наближено виразимо другу похідну від розв'язку через значення

розв'язку у вузлах сітки $u_n = u(x_n)$; наприклад, скористаємося найпростішою апроксимацією

$$u''(x_n) \approx \frac{1}{h^2}(u_{n-2} - 2u_n + u_{n+1}),\ h = x_{n+1} - x_n = const.$$

Таку апроксимацію можна записати в будь-якому вузлі сітки $x_n, 1 \le n \le N-1$. Якщо підставити її в рівняння (8.54), то рівняння стане наближеним; точно задовольняти це рівняння буде вже не шуканий розв'язок $u(x)$, а деякий наближений розв'язок $y_n \approx u(x_n)$. Виконуючи цю підстановку і позначаючи $p_n = p(x_n)$ і $f_n = f(x_n)$, одержимо

$$y_{n-1} - (2 + h^2 p_n) y_n + y_{n+1} = h^2 f_n,\ \ 1 \le n \le N-1. \tag{8.56}$$

Ця формула складається з N-1 алгебраїчного рівняння, а невідомими в ній є наближені значення розв'язку у вузлах сітки. Число невідомих $y_n, 0 \le n \le N$ дорівнює N+1, тобто воно більше, ніж число рівнянь (8.56). Відсутні два рівняння легко одержати з крайових умов (8.55):

$$y_0 = \alpha, y_N = \beta. \tag{8.57}$$

У випадку використання граничних умов другого роду апроксимація проводиться за допомогою формул чисельного диференціювання першого порядку:

$$y_0' = \frac{y_1 - y_0}{h} + O(h);$$

$$y_N' = \frac{y_N - y_{N-1}}{h} + O(h).$$

Розв'язуючи алгебраїчну систему (8.56), (8.57), знайдемо наближений розв'язок.

Як ілюстрацію проведемо повне дослідження розглянутого вище прикладу, додатково вимагаючи $p(x) > 0$.

Спочатку розглянемо питання про існування різницевого розв'язку. Вихідна задача (8.54) була лінійною, різницева апроксимація (8.56)– теж лінійна. Завдяки цьому система (8.56,8.57) виявилася системою лінійних алгебраїчних рівнянь. Оскільки $p_n > 0$, то в матриці цієї системи діагональні елементи переважають: у кожному рядку модуль діагонального елемента більше суми модулів інших елементів, при цьому розв'язок лінійної системи існує і єдиний.

Обчислити розв'язок лінійної системи рівнянь завжди можна методом виключення Гауса. У даному випадку завдяки використанню триточкової апроксимації (8.54) система (8.56) має тридіагональну матрицю. Тому розв'язок доцільніше знаходити за допомогою різновиду методу Гауса – методом прогонки.

Щоб оцінити похибку наближеного розв'язку задачі, використовують інформацію, отриману в процесі чисельних розрахунків (такі оцінки називаються апостеріорними). Найефективнішими можна вважати оцінки з подвійним перерахунком.

Наявність наближених значень y_k і y_k^*, обчислених відповідно з кроками h і $h/2$, дає можливість зробити оцінку. Похибка методу – це $\varepsilon_k = y_k^* - u(x_k)$, визначена в точці x_k.

Отже, якщо $y_1 - u(x_1) = M \cdot h^{s+1}$, де M – невідомий коефіцієнт пропорційності, s – порядок точності методу, то

$$y_2 - u(x_2) = 2M \cdot h^{s+1},$$
$$y_3 - u(x_3) = 3M \cdot h^{s+1},$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$y_n - u(x_n) = nM \cdot h^{s+1}.$$

Виходить, для похибки в точці x_k при визначенні розв'язку з кроком h маємо рівність $y_k - u(x_k) = kM \cdot h^{s+1}$, а при розв'язку з кроком $h/2$ – рівність

$$y_k^* - u(x_k) = 2kM \cdot \left(\frac{h}{2}\right)^{s+1}. \qquad (8.58)$$

Знайшовши різницю між наведеними вище рівностями і розв'язавши отриману рівність відносно невідомого коефіцієнта M, визначимо

$$M = \frac{2^s\left(y_k - y_k^*\right)}{kh^{s+1}\left(2^s - 1\right)}.$$

Підставивши це значення M у формулу (8.58), одержимо $y_k^* - u(x_k) = \frac{\left(y_k - y_k^*\right)}{\left(2^s - 1\right)}$. Звідси для абсолютної похибки в точці x_k остаточно одержимо таку рівність:

$$\left|\varepsilon_k^*\right| = \left|y_k^* - u(x_k)\right| = \frac{\left|y_k - y_k^*\right|}{2^s - 1}.$$

Таку оцінку абсолютної похибки методу називають, як відомо, ***правилом Рунге***.

Зупинимося на стійкості розрахунку. Якщо $p(x) > 0$, то задача Коші для рівняння (8.54) погано обумовлена, причому, чим більше *p(x)*, тим гірша її стійкість. А з оцінки (8.58) видно, що похибка нашого різницевого розв'язку при великих *p(x)* мала. Звідси виходить, що добре побудовані різницеві схеми не чуттєві до нестійкості задачі Коші. У випадку, коли $p(x) < 0$, не виконується достатня умова збіжності ітераційного процесу для систем лінійних алгебраїчних рівнянь, однак у практичних обчисленнях дана обставина, як правило, виявляється несуттєвою і не викликає складностей в одержанні розв'язку.

8.4 Різницева задача на власні значення

Розглянемо диференціальну задачу Штурма-Ліувілля

$$\begin{cases} \dfrac{d^2u}{dx^2} + \lambda u = 0, \quad 0 < x < 1, \\ u(0) = 0, \\ u(1) = 0. \end{cases}$$

Числа λ і відповідні функції $u(x) \neq 0$, що задовольняють поставлену крайову задачу називаються власними числами і власними функціями відповідно. Для даної задачі

$$u_m(x) = \sin \pi m x, \quad \lambda_m = (\pi m)^2, \quad m = 0,1,....$$

Зауважимо, що функції $u_m(x)$ є лінійно незалежними і взаємно ортогональними й можуть бути нормовані.

Для різницевої задачі на власні значення

$$\begin{cases} \dfrac{y_{j+1} - 2y_j + y_{j-1}}{h^2} + \lambda y_j = 0, \; j = 1,..., n-1, \\ y_0 = 0, \\ y_n = 0 \end{cases}$$

відповідні власні функції і власні значення різницевої задачі мають вигляд

$$y_m(x_j) = \sin \pi m x_j, \quad \lambda_m = \frac{4}{h^2}\sin^2\frac{\pi m h}{2}, \quad m = 1,2,...,n-1.$$

Відмітимо, що функції $y_m(x)$ є лінійно незалежними і взаємно ортогональними, як і в диференціальному випадку, й можуть бути нормовані.

Питання і завдання до розділу 8

1 Постановка задачі Коші. Дискретна задача Коші: основні поняття і визначення (сітка, сіткові функції, чисельний метод, апроксимація, збіжність).

2 Виведення формули методу Ейлера, його геометрична інтерпретація, стійкість, оцінка похибки, вплив обчислювальної похибки.

3 Методи Рунге-Кутта. Виведення формул. Оцінка похибки.

4 Явні однокрокові методи. Оцінка похибки за правилом Рунге.

5 Чисельне розв'язання задачі Коші для систем диференціальних рівнянь.

6 Апроксимація, стійкість і збіжність чисельних методів розв'язання задачі Коші.

7 Багатокрокові методи Адамса.

8 Виведення формул методу прогнозу і корекції.

9 Жорсткі задачі і методи їхнього розв'язання.

10 Застосовуючи метод Ейлера , знайти розв'язок задачі Коші $\begin{cases} y' = 0.5xy \\ y(0) = 1 \end{cases}$ у трьох послідовних точках: $x_1 = 0.2, x_2 = 0.4, x_3 = 0.6$

11 Для задачі Коші $\begin{cases} y' = y - x \\ y(0) = 1.5 \end{cases}$ виконати один крок довжини 0.1 за методом Ейлера й оцінити похибку знайденого значення за правилом Рунге.

12 Методом Рунге-Кутта 2 порядку точності знайти розв'язок системи диференціальних рівнянь $\begin{cases} y' = z + 1 \\ z' = y - x \\ y(0) = 1,\ z(0) = 1 \end{cases}$ у двох послідовних точках $x_1 = 0.1, x_2 = 0.2$.

13 Оцінити похибку апроксимації похідної різницевим відношенням $y'(x_i) = \dfrac{y_{i-2} - 8y_{i-1} + 8y_{i+1} - y_{i+2}}{12h}$.

14 Звести рівняння другого порядку до системи рівнянь першого порядку і скласти розрахункові формули методу

прогнозу і корекції для розв'язку отриманої системи рівнянь $y'' + y' - xe^{-x}y = \cos(x)$ $y(1) = 1$, $y'(1) = 3$.

15 З'ясувати, чи апроксимують методи

a) $\dfrac{y_n - y_{n-3}}{3h} = f_{n-1}$, b) $\dfrac{y_n - 3y_{n-2} + 2y_{n-3}}{8h} = \dfrac{f_{n-1} + f_{n-2}}{2}$

перше рівняння задачі Коші $\begin{cases} y' = f(t, y), \\ y(t_0) = y0. \end{cases}$

16 Для розв'язання задачі Коші $\begin{cases} y' + y = t + 1, \\ u(0) = 0 \end{cases}$ застосовується метод вигляду $\begin{cases} \dfrac{y_{n+1} - y_{n-1}}{2h} + y_n = nh + 1, \\ y_0 = 0, y_1 = 0. \end{cases}$ Визначити порядок апроксимації.

17 Дано систему ОДУ першого порядку з постійними коефіцієнтами $y' = Ay$, причому відомі власні значення матриці A:

a) $\lambda_1 = 0.1 + i, \lambda_2 = -0.6 + 2i$,

b) $\lambda_1 = 0.1 + i, \lambda_2 = -2000 + i$,

c) $\lambda_1 = 10 - 0.5i, \lambda_2 = -1000 + 2i$.

У яких випадках систему можна вважати жорсткою?

Розділ 9
Чисельне розв'язання диференціальних рівнянь у частинних похідних

9.1 Класифікація диференціальних рівнянь у частинних похідних

Диференціальні рівняння в частинних похідних широко використовуються в математичній фізиці, гідродинаміці, акустиці й інших галузях знань. Здебільшого такі рівняння в явному вигляді не розв'язуються. Тому широкого поширення набули методи чисельного їх розв'язання, зокрема метод сіток.

Ми зосередимо увагу на розв'язання лінійних диференціальних рівнянь другого порядку. Їх загальний вигляд:

$$F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = 0. \qquad (9.1)$$

Розв'язком рівняння (9.1) називається функція $u=u(x,y)$, що перетворює це рівняння в тотожність.

Рівняння (9.1) називається лінійним, якщо воно може бути записане у вигляді

$$A\frac{\partial^2 u}{\partial x^2} + B\frac{\partial^2 u}{\partial y \partial .x} + C\frac{\partial^2 u}{\partial y^2} + a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} + cu = f(x, y) \qquad (9.2)$$

У загальному випадку A,B,C,a,b,c - це коефіцієнти, що можуть залежати тільки від x,y.

Уводиться дискримінант $D = AC - B^2$, залежно від знака якого лінійне диференціальне рівняння (9.2) відноситься до одного з таких типів:

1) якщо D>0, то рівняння еліптичне;

2) якщо D=0, то рівняння параболічне;

3) якщо D<0, то рівняння гіперболічне;

4) якщо D не зберігає знак – змішаний тип.

Диференціальні рівняння в частинних похідних мають незліченну множину розв'язків, тому для однозначності розв'язку необхідно до вихідного рівняння приєднати додаткові умови. Для диференціальних рівнянь у частинних похідних 2-го порядку вони можуть бути початковими і граничними. По суті,

розрізнити ці умови можна лише в тому випадку, коли одна з незалежних змінних диференціального рівняння відіграє роль часу, а інша – роль координати. Тоді, якщо умови задані для початкового моменту часу, то це початкові умови, а умови, що відносяться до фіксованих значень координат, є граничними, або крайовими.

Розглянемо загальну постановку задачі з початковими умовами. Нехай задане лінійне диференціальне рівняння

$$L[u] = f(x, y), \qquad (9.3)$$

де $$L[u] = A\frac{\partial^2 u}{\partial x^2} + B\frac{\partial^2 u}{\partial y \partial x} + C\frac{\partial^2 u}{\partial y^2} + a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} + cu.$$

Визначення розв'язку $u = u(x, y)$ такого рівняння, що задовольняє початкові умови, називається задачею Коші, а самі умови мають назву початкових даних Коші.

Задача Коші частіше розглядається для лінійного рівняння (9.2) параболічного і гіперболічного типів.

Для рівнянь еліптичного типу задача Коші зазвичай не розглядається. Це зумовлено тим, що, як правило, задача Коші для рівнянь еліптичного типу поставлена некоректно, тобто мізерно малі зміни початкових даних можуть спричинити істотні зміни розв'язку. Для цих рівнянь звичайно ставляться лише крайові задачі, що поділяються на три типи:

1 На контурі Γ, що обмежує область G, задана неперервна функція $\varphi(P) = \varphi(x, y)$. Потрібно знайти функцію $u(P)=u(x,y)$, що задовольняє всередині G рівняння

$$L[u] \equiv \Delta u + au_x + bu_y + cu = F(x, y) \qquad (9.4)$$

та набуває на границі заданих значень $\varphi(P)$, тобто повинні бути виконані умови $L[u(P)] = F(P)$ при $P \in G$; $u(P) = \varphi(P)$ при $P \in \Gamma$.

2 На контурі Γ, що обмежує область G, задана неперервна функція $\varphi_1(P)$. Потрібно знайти функцію $u(P)=u(x,y)$, що задовольняє всередині G рівняння (9.4), нормальна похідна якої на Γ набуває заданих значень $\varphi_1(P)$, тобто потрібно, щоб:

$L[u(P)] = F(P)$ при $P \in G$; $\dfrac{\partial u(P)}{\partial n} = \varphi_1(P)$ при $P \in \Gamma$.

3 На контурі Г, що обмежує область G, задана неперервна функція $\psi(P) = \psi(x, y)$. Потрібно знайти функцію $u(P)=u(x,y)$, таку, щоб $L[u(P)] = F(P)$ при $P \in G$; $a_0 u(P) + a_1 \dfrac{\partial u(P)}{\partial n} = \psi(P)$ при $P \in \Gamma$, де $|a_0| + |a_1| \neq 0$.

Третя задача – узагальнення перших двох.

Розглянемо широко відомі приклади диференціальних рівнянь у частинних похідних.

1 Хвильове рівняння

$$\frac{\partial^2 u(x,t)}{\partial t^2} = a^2 \frac{\partial^2 u(x,t)}{\partial x^2}. \qquad (9.5)$$

До нього приводить дослідження процесів поперечних коливань струни, поздовжніх коливань стрижня, електричних коливань у проводі, крутильних коливань валу, коливань газу і т.д. Це рівняння є найпростішим рівнянням гіперболічного типу.

2 Рівняння теплопровідності (дифузії)

$$\frac{\partial u(x,t)}{\partial t} = a^2 \frac{\partial^2 u(x,t)}{\partial x^2}. \qquad (9.6)$$

Воно описує процеси поширення тепла, дифузії рідини й газу, деякі питання теорії ймовірності і т.д. Це рівняння параболічного типу.

3 Рівняння Лапласа

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 0. \qquad (9.7)$$

До дослідження цього рівняння приводять задачі електродинаміки, про стаціонарний тепловий стан, задачі гідродинаміки, дифузії і т.д. Це рівняння є найпростішим рівнянням еліптичного типу.

Як правило, аналітичні методи розв'язання рівнянь у частинних похідних пов'язані з розділенням змінних (метод Фур'є). Використання цього методу викликає великі труднощі,

якщо область незалежних змінних, де знаходиться розв'язок, відрізняється від найпростіших фігур (прямокутник, коло). Іншою перешкодою для застосування методу Фур'є є залежність коефіцієнтів лінійного рівняння від часу і просторових змінних. Наприклад, залежність коефіцієнта $a=a(t,x)$ у хвильовому рівнянні з функцією $a(t,x)$ досить загального вигляду вже не дозволить розділити змінні.

Відзначені обмеження застосування аналітичних методів привели, особливо з розвитком обчислювальної техніки, до широкого поширення чисельних методів розв'язання рівнянь. Досвід розв'язання багатьох складних задач науки і техніки чисельними методами підтверджує їх ефективність.

Рівняння (9.5),(9.6), у яких одна з незалежних змінних t є часом, називаються нестаціонарними. Рівняння (9.7) для функції $u(x,y)$, що залежить тільки від просторових координат x і y, є стаціонарним.

9.2 Апроксимація частинних похідних

Першим етапом у чисельному розв'язанні диференціальних рівнянь із частинними похідними є перехід від неперервної задачі до дискретної. Дискретизація задачі – це основа чисельного методу.

Розглянемо деякі кінцево-різницеві аналогії частинних похідних від функції $u(x,y)$, використовуючи підходи для звичайних диференціальних рівнянь, з урахуванням тієї особливості, що тепер функція має дві незалежні змінні. Почнемо з того, що розглянемо різниці тільки в напрямку x. Згадаємо, що розкладання Тейлора функції $u(x+h,y)$ в околі точки (x,y) можна записати у вигляді

$$u(x+h, y) = u(x, y) + h\frac{\partial u(x, y)}{\partial x} + \frac{1}{2}h^2 \frac{\partial^2 u(\xi, y)}{\partial^2 x}, \qquad (9.8)$$

де ξ лежить між x і $x+h$. Звідси одержуємо

$$\left[\frac{\partial u(x,y)}{\partial x} - \frac{u(x+h,y)-u(x,y)}{h}\right] = \frac{h}{2}\frac{\partial^2 u(\xi,y)}{\partial x^2}.$$

Таким чином, якщо вважати

$$u_x = \frac{\partial u(x,y)}{\partial x} \approx \frac{u(x+h,y)-u(x,y)}{h}, \qquad (9.9)$$

то похибка апроксимації (дискретизації), або похибка обмеження, буде дорівнювати

$$r_1(h) = -\frac{h}{2}\frac{\partial^2 u(\xi,y)}{\partial x^2} = -\frac{h}{2}u_{xx}(\xi,y) = O(h), \text{ де } x<\xi<x+h. \quad (9.10)$$

У формулі (9.9) частинна похідна виражена через праву скінченну різницю *[u(x+ h, y)-u(x, y)]*. Виразимо її через ліву різницю *[u(x, y)-u(x-h, y)]*. Для цього в розкладання Тейлора замість *(x+h)* підставимо *(x-h)*

$$u(x-h,y) = u(x,y) - h\frac{\partial u(x,y)}{\partial x} + \frac{1}{2}h^2\frac{\partial^2 u(\xi,y)}{\partial x^2}, \quad \text{де } (x-h)<\xi<x. \quad (9.11)$$

При цьому одержуємо наближену рівність

$$\frac{\partial u(x,y)}{\partial x} \approx \frac{u(x,y)-u(x-h,y)}{h}, \qquad (9.12)$$

яка виконується з похибкою, що дорівнює:

$$r_2(h) = \frac{h}{2}\frac{\partial u(\xi,y)}{\partial x^2} = O(h), \quad \text{де } (x-h)<\xi<x. \quad (9.13)$$

Тепер одержимо різницеве наближення для $u_{xx} = \frac{\partial^2 u}{\partial x^2}$.

Якщо *u(x,y)* має частинні похідні до четвертого порядку включно, тоді розкладання функцій *u(x+h,y)*, *u(x-h,y)* в околі точки *(x,y)* можна записати у вигляді

$$u(x+h,y) = u(x,y) + h\frac{\partial u(x,y)}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{h^3}{6}\frac{\partial^3 u(x,y)}{\partial x^3} + \frac{h^4}{8}\frac{\partial^4 u(\xi,y)}{\partial x^4},$$

де (x-h)<ξ<x;

$$u(x-h,y) = u(x,y) - h\frac{\partial u(x,y)}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u(x,y)}{\partial x^2} - \frac{h^3}{6}\frac{\partial^3 u(x,y)}{\partial x^3} + \frac{h^4}{8}\frac{\partial^4 u(x,y)}{\partial x^4},$$

де (x-h)<ξ<x.

Склавши дві останні рівності, одержимо співвідношення

$$u_{xx}(x,y) = \frac{\partial^2 u(x,y)}{\partial x^2} \approx \frac{u(x+h,y)-2(x,y)+u(x-h,y)}{h^2}, \quad (9.14)$$

яке виконується з точністю

$$r_3(h) = -\frac{h^2}{12}\frac{\partial^2 u(x,\xi)}{\partial x^4} = O(h^2), \text{ де } (x-h)<\xi<(x+h). \quad (9.15)$$

Подібний аналіз можна провести для похідних у напрямку *y* й одержати формули, аналогічні до формул (9.8)-(9.15). Використовуючи ці вирази, можна представити диференціальні рівняння в частинних похідних (9.3) через різницеві наближення.

9.3 Метод сіток

Головна ідея чисельного розв'язання рівнянь у частинних похідних дуже схожа на метод розв'язку крайових задач для ЗДР, розглянутий нами в попередній главі. Основною відмінністю від ЗДР є необхідність дискретизації рівняння не за однією, а за кількома змінними (залежно від розмірності задачі). Таким чином, спочатку потрібно покрити розрахункову область *(x,y)* сіткою і використовувати потім вузли цієї сітки для різницевої апроксимації рівняння. У результаті, замість пошуку неперервних залежностей *u(x,y)*, досить буде відшукати значення функції у вузлах сітки (а її поводження в проміжках між вузлами може бути отримане за допомогою побудови якої-небудь інтерполяції). З цієї причини дискретне представлення функції часто називають ***сітковою, або різницевою, функцією***. Оскільки рівняння в частинних похідних за визначенням залежать від похідних невідомих функцій за кількома змінними, то способів дискретизації цих рівнянь може бути, як правило, багато. Конфігурацію вузлів, що використовується для різницевого запису рівнянь у частинних похідних на сітці, називають ***шаблоном***, а отриману систему різницевих рівнянь - **різницевою схемою**. Отже, різницевою схемою є сукупність різницевих рівнянь, що апроксимують первісне диференціальне рівняння в усіх внутрішніх вузлах сітки, та додаткові (початкові і граничні) умови – в граничних вузлах сітки. Різницеву схему за аналогією з диференціальною задачею назвемо різницевою задачею. Про принципи побудови різницевих схем і, зокрема, про класи явних і неявних схем ми вже докладно говорили на прикладі крайових задач для ЗДР (дивись розділ 8), тому перейдемо до розгляду типових особливостей рівнянь у

частинних похідних, що виникають при розробці і реалізації різницевих схем.

Метод сіток для наближеного розв'язання крайових задач двовимірних диференціальних рівнянь полягає в такому:

-у плоскій області G, у якій розшукується розв'язок, будується сіткова область G_h (рис. 9.1);

-задане диференціальне рівняння заміняється у вузлах побудованої сітки відповідними різницевими рівняннями;

-на підставі граничних умов визначаються значення шуканого розв'язку в граничних вузлах області Γ_h.

Розв'язавши отриману систему різницевих рівнянь, ми знайдемо значення шуканої функції у вузлах сітки, тобто будемо мати числовий розв'язок нашої задачі.

Вибір сіткової області здійснюється залежно від конкретної задачі, але у всіх випадках контур Γ_h сіткової області G_h варто обирати так, щоб він якнайкраще апроксимував контур Γ заданої області G.

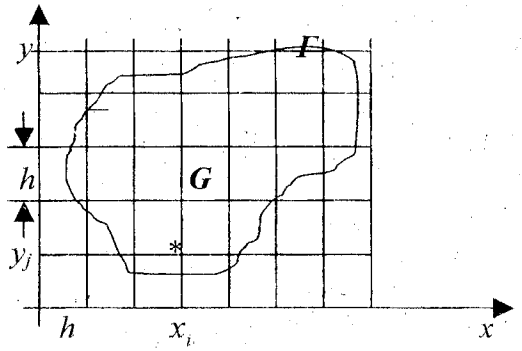


Рис. – 9.1

Сітка будується таким чином, щоб вузли (x_i,y_j) сітки G_h або належали області G, або відступали від її границі Γ на відстань меншу за h.

Точки (вузли) сітки G_h називаються сусідніми, якщо вони віддалені одна від одної в напрямку вісі Ox або вісі Oy на відстань, що дорівнює кроку сітки h. Вузол A_h сітки G_h називається внутрішнім, якщо він належить області G, а всі чотири сусідні з ним вузли – множині G_h; інакше він називається граничним.

Граничний вузол сітки Γ_h називається вузлом першого роду, якщо він має сусіднім внутрішній вузол цієї сітки, інакше – граничний вузол називається вузлом другого роду. Внутрішні вузли і граничні вузли першого роду сітки $S_h=G_h\cup\Gamma_h$ називаються розрахунковими точками. Граничні вузли другого роду не входять в обчислення і можуть бути вилучені із сітки.

На перший погляд процедура застосування методу сіток, що складається з трьох етапів, здається простою і такою, що прямо веде до розв'язку. Однак насправді це не так. Через велику розмаїтість типів і розмірів сіток, рівнянь у частинних похідних, граничних і початкових умов, можливих різницевих апроксимацій цих рівнянь і методів їхнього розв'язання, чисельне розв'язання рівнянь у частинних похідних вимагає модифікацій алгоритму при розгляді кожного конкретного випадку.

9.4 Апроксимація для диференціальних рівнянь

Розглянемо деяку задачу математичної фізики в операторній формі

$$\begin{aligned} Lu &= f \text{ в } G, \\ lu &= g \text{ на } \Gamma, \end{aligned} \tag{9.16}$$

де L, l—лінійні оператори, $u\in U$ и $f\in F$.

Тут U та F - гільбертові простори з областями визначення елементів у $G\cup\Gamma$ і G відповідно, l - лінійний оператор граничної умови, $g\in D$, D - гільбертів простір функцій з областю визначення Γ.

Поряд із рівнянням (9.16) розглянемо рівняння в скінченновимірному просторі сіткових функцій

$$\begin{aligned} L^h u^h &= f^h \text{в } G_h, \\ l^h u^h &= g^h \text{ на } \Gamma_h, \end{aligned} \tag{9.17}$$

де L^h — лінійний оператор, що залежить від кроку сітки h, $u^h\in U_h$, $f^h\in F_h$ — простору сіткових функцій. Тут G_h -

множина внутрішніх вузлових точок області G, а Γ_h - множина вузлових точок, на якій апроксимується гранична умова задачі, l^h - лінійний оператор, $g^h \in D_h$, D_h - евклідів простір векторів з областю визначення Γ_h.

Уведемо в сіткових просторах F_h, D_h, U_h відповідно норми $\|\cdot\|_{F_h}, \|\cdot\|_{D_h}, \|\cdot\|_{U_h}$. Нехай $(\cdot)_h$ — лінійний оператор, що елементові $u \in U$ ставить у відповідність елемент $u^h \in U_h$ так, що $\lim\limits_{h\to 0}\|u^h\|_{U_h} = \|u\|_U$.

Будемо вважати, що ***задача (9.17) апроксимує задачу (9.16) з порядком n на розв'язку u, якщо існують додатні константи*** $\bar{h}$, M_1, M_2 ***такі, що для усіх*** $h < \bar{h}$ ***здійснюються нерівності:***

$$\begin{aligned} \left\|L^h(u)_h - f^h\right\|_{F_h} &\le M_1 h^{n_1}, \\ \left\|l^h(u)_h - g^h\right\|_{D_h} &\le M_2 h^{n_2} \end{aligned} \qquad (9.18)$$

і $n = \min(n_1, n_2)$.

У тих випадках, коли розв'язок u задачі (9.4) має достатню гладкість, порядок апроксимації зручно знаходити за допомогою норми, природної для простору неперервних і диференційованих функцій. З цією метою, як правило, користуються розкладанням розв'язку й інших функцій, що беруть участь у задачі, у ряди Тейлора.

Надалі будемо вважати, що редукція задачі (9.16) до задачі (9.17) здійснена і, більше того, гранична умова з (9.17) використана для виключення значень розв'язку в граничних точках області $G_h \cup \Gamma_h$. У результаті приходимо до еквівалентної задачі

$$\tilde{L}^h \tilde{u}^h = \tilde{f}^h. \qquad (9.19)$$

При цьому значення розв'язку в граничних точках знайдеться з рівняння (9.17) після розв'язання рівняння (9.19). У

деяких випадках зручно користуватися записом апроксимаційної задачі у формі (9.19), а в інших випадках - у формі (9.17). Отже, у результаті проведеної редукції з урахуванням необхідної апроксимації задача з неперервним аргументом (9.16) зводиться до задачі лінійної алгебри (9.19). Подальше завдання полягає в розв'язанні системи алгебраїчних рівнянь.

9.5 Проблема збіжності методу сіток

Одним із найважливіших моментів у чисельному розв'язанні диференціальних задач із частинними похідними є дослідження отриманого розв'язку різницевої схеми на збіжність до точного. Розв'язок різницевої задачі u^h збігається до розв'язку u вихідної задачі, якщо $\lim\limits_{h\to 0}\left\|(u)_h - u^h\right\|_{U_h} = 0$.

А.Ф. Філіппов визначив стійкість для довільних різницевих задач

$$L^{h\tau} u^{h\tau} = f^{h\tau},$$

де h і τ - кроки дискретизації двовимірної області G як рівномірну обмеженість оператора $(L^{h\tau})^{-1}$ і довів, що з апроксимації та стійкості випливає збіжність розв'язку різницевої задачі до розв'язку диференціальної. П. Лакс запропонував для коректно поставлених еволюційних задач таку систему визначень апроксимації та стійкості, за якою стійкість має місце одночасно зі збіжністю, якщо має місце апроксимація. Ця теорема відома як теорема Лакса. Еволюційними називають рівняння, що явно можна розв'язати відносно першої похідної за часом і не містить у правій частині похідних за часом.

Дослідження збіжності різницевого розв'язку до розв'язку різних задач здійснюється на основі однакових принципів. Це дозволяє прослідкувати основну ідею доведення на прикладі стаціонарної задачі (9.16), що апроксимується різницевою схемою (9.17).

Теорема збіжності. Нехай

1) різницева схема (9.17) апроксимує вихідну задачу (9.16) на розв'язку u з порядком n;

2) L^h, l^h - лінійні оператори;

3) різницева схема (9.17) стійка , тобто існують додатні константи $\overline{h}, C_1, C_2$ такі, що для всіх $h<\overline{h}$, $f^h \in F_h$, $g^h \in \Gamma_h$ існує, і до того ж один, розв'язок u^h задачі (9.17), який задовольняє нерівність

$$\left\|u^h\right\|_{U_h} \le C_1\left\|f^h\right\|_{F_h} + C_2\left\|g^h\right\|_{\Gamma_h}. \qquad (9.20)$$

Тоді розв'язок різницевої задачі u^h збігається до розв'язку u вихідної задачі, тобто $\lim\limits_{h\to 0}\left\|(u)_h - u^h\right\|_{U_h} = 0$.

При цьому має місце така оцінка швидкості збіжності

$$\left\|(u)_h - u^h\right\|_{U_h} \le (C_1M_1 + C_2M_2)h^n, \qquad (9.21)$$

де M_1 та M_2 - константи з (9.18).

Доведення. Нехай h- мінімальне з $\overline{h}$, уведених у визначеннях апроксимації та стійкості. Тоді внаслідок стійкості для будь-яких правих частин f^h та g^h при $h<\overline{h}$ існує єдиний розв'язок, тобто можна розглядати різницю $(u)_h - u^h$. Завдяки лінійності оператора L^h отримаємо

$$L^h[(u)_h - u^h] = L^h(u)_h - L^hu^h = L^h(u)_h - f^h.$$

Аналогічно $l^h[(u)_h - u^h] = l^h(u)_h - g^h$.

Звідси, внаслідок стійкості та апроксимації,

$$\left\|(u)_h - u^h\right\|_{U_h} \le C_1\left\|L^h(u)_h - f^h\right\|_{F_h} + C_2\left\|l^h(u)_h - g^h\right\|_{\Gamma_h} \le$$

$$\le C_1M_1h^{n_1} + C_2M_2h^{n_2} \le (C_1M_1 + C_2M_2)h^n.$$

Не порушуючи загалу, можна вважати, що $h<1$.

Доведення завершено.

У випадку еволюційної задачі, коли, крім просторового кроку h, є ще крок за часом τ, з визначення стійкості та апроксимації аналогічно можна отримати оцінку

$$\left\|(u)_{h\tau} - u^{h\tau}\right\|_{U_{h\tau}} \le K_1h^n + K_2\tau^p.$$

Ця оцінка доводить збіжність різницевого розв'язку до точного залежно від обраних кроків сітки.

9.6 Рівняння параболічного типу

9.6.1 Явні різницеві схеми

Розглянемо мішану задачу для рівняння теплопровідності зі сталими коефіцієнтами. В області $\{0<x<1, 0<t\le T\}$ потрібно знайти розв'язок рівняння

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t), \qquad (9.22)$$

що задовольняє початкову умову

$$u(x,0) = u^0(x),\ 0 \le x \le 1 \qquad (9.23)$$

і граничні умови

$$u(0,t) = \mu_1(t),\ u(1,t) = \mu_2(t). \qquad (9.24)$$

Як відомо, при таких припущеннях щодо гладкості розв'язок задачі (9.22) — (9.24) існує і єдиний. При вивченні апроксимації різницевими схемами припустимо, що розв'язок $u(x,t)$ має необхідну кількість похідних по t і x. Розв'язок задачі (9.22) —(9.24) неперервно залежить від початкових та граничних даних.

При побудові різницевої схеми введемо сітку в області змінних і задамо шаблон, тобто множину точок сітки, що бере участь в апроксимації диференціального виразу. Сітка вводиться за змінною x із кроком $h=\Delta x$ так:

$$\omega_h = \{x_j = j\Delta x = jh,\ j = 0,1,...,N,\ hN = 1\},$$

а за змінною t із кроком $\tau = \Delta t$ (позначимо її $\omega_\tau = \{t_m = m\Delta t = m\tau,\ m = 0,1,...,K,\ K\tau = T\}$)

Точки (x_j, t_m), $j=0,1,...,N$, $m=0,1,...,K$ утворюють вузли просторово - часової сітки. Вузли (x_j, t_m), що належать відрізкам $I_0 = \{0 \le x \le 1, t = 0\}$, $I_1 = \{x = 0, 0 \le t \le T\}$,

$I_2 = \{x = 1, 0 \le t \le T\}$ належать до граничних вузлів сітки $\omega_{\tau,h}$, а всі інші вузли —внутрішні.

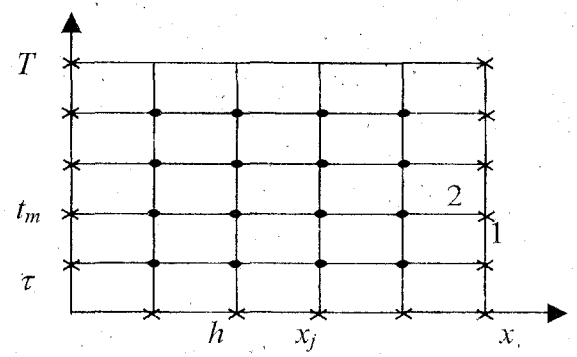


Рис. – 9.2

На рис. 9.2 позначено 1 — граничні, 2 — внутрішні вузли сітки. Варіанти шаблонів наведені на рис. 9.3.

Шаром називається множина всіх вузлів сітки $\omega_{\tau,h}$, що мають одну й ту саму часову координату. Таким чином, n-м шаром буде множина вузлів $(x_0,t_n),(x_1,t_n),\ldots,(x_N,t_n)$. Для функцій $u(x,t)$ та $f(x,t)$, визначених на сітці $\omega_{h,\tau}$, введемо позначення: $u_j^n = u(x_j,t_n)$, $f(x_j,t_n) = \varphi_j^n$. А для частинних похідних оберемо такі наближення:

$$\left.\frac{\partial u}{\partial t}\right|_{(x_j,t_n)} = \frac{u_j^{n+1} - u_j^n}{\tau}; \qquad (9.25)$$

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{(x_j,t_n)} = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}. \qquad (9.26)$$

Для апроксимації рівняння (9.22) у точці (x_j,t_m) введемо шаблон (рис.9.3,а). Отримаємо різницеву схему

$$\begin{cases} \dfrac{y_j^{n+1} - y_j^n}{\tau} = \dfrac{y_{j+1}^n - 2y_j^n + y_{j-1}^n}{h^2} + \varphi_j^n, (j = \overline{1, N-1}, n = \overline{0, K-1}); \\ y_j^0 = u^0(x_j),\ j = \overline{0, N}; \\ y_0^n = \mu_1(t_n),\ n = \overline{0, K}; \\ y_N^n = \mu_2(t_n),\ n = \overline{0, K}, \end{cases} \qquad (9.27)$$

де y_j^n - це значення сіткової функції, що наближає точне значення розв'язку u_j^n у вузлі (x_j,t_m).

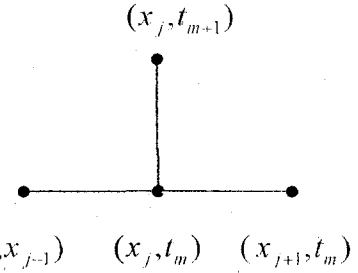


Рис. – 9.3 а)

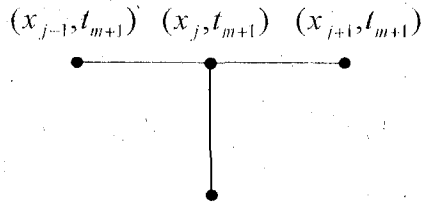


Рис. – 9.3 б)

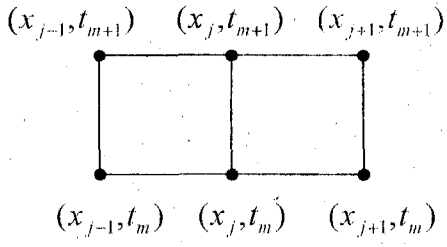


Рис. – 9.3 в)

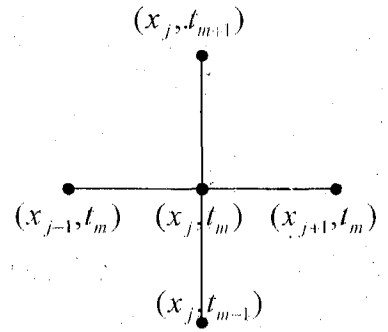


Рис. – 9.3 г)

Дослідження різницевих схем передбачає відповідь на такі питання:

1) існування та єдиність розв'язку;
2) методи отримання розв'язку різницевої схеми;
3) як співвідносяться різницева схема та вихідна диференціальна задача (проблема апроксимації);
4) чи збігається наближений розв'язок до точного.

Розглянемо перше з рівнянь (9.27) відносно y_j^{n+1}.

$$y_j^{n+1} = (1-\frac{2\tau}{h^2})y_j^n + \frac{\tau}{h^2}(y_{j+1}^n + y_{j-1}^n) + \tau\varphi_j^n,\ j=\overline{1,N-1},\ n=\overline{0,K-1}. \quad (9.28)$$

При $n=0$ отримаємо

$$y_j^1 = (1-\frac{2\tau}{h^2})y_j^0 + \frac{\tau}{h^2}(y_{j+1}^0 + y_{j-1}^0) + \tau\varphi_j^0,\ j=\overline{1,N-1}.$$

У правій частині всі значення відомі з крайових умов. Тому можна отримати шукану сіткову функцію на всьому першому часовому шарі. Аналогічно можна розрахувати другий і наступні шари. Звідси зрозуміло, що розв'язок існує і він єдиний. Оскільки всі формули явні, то і вся схема є явною різницевою схемою.

Похибка різницевої схеми (9.27) визначається як різниця між розв'язками вихідної задачі (9.22) – (9.24) та задачі (9.27). Якщо підставити точний розв'язок у різницеве рівняння (9.27), то воно задовольниться не повністю, а з деякою похибкою, яка називається локальною похибкою дискретизації, або нев'язкою. Визначимо її поведінку:

$$\psi_j^n = -\frac{u_j^{n+1}-u_j^n}{\tau} + \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} + \varphi_j^n.$$

Для цього розкладемо $u_{j\pm1}^n$ та u_j^{n+1} в ряд Тейлора через u_j^n

$$u_{j\pm1}^n = u_j^n \pm u_{x,j}^n h + u_{xx,j}^n \frac{h^2}{2} \pm u_{xxx,j}^n \frac{h^2}{6} + O(h^4);$$

$$u_j^{n+1} = u_j^n + u_{t,j}^n \tau + O(\tau^2).$$

Тоді отримаємо вираз для нев'язки $\psi_j^n = (-u_{t,j}^n + u_{xx,j}^n + f_j^n) + \varphi_j^n - f_j^n + O(\tau + h^2)$.

Вираз у дужках дорівнює нулю, оскільки u_j^n - точний розв'язок рівняння теплопровідності. Якщо взяти функції φ_j^n так, що $\varphi_j^n = f_j^n + O(\tau + h^2)$, то для нев'язки буде справедливою оцінка

$$\|\psi\|_h = O(\tau + h^2).$$

Тобто наша різницева схема апроксимує диференціальне рівняння з другим порядком апроксимації по h і з першим порядком апроксимації по τ.

Розглянемо питання про збіжність наближеного розв'язку до точного. Виразимо наближений розв'язок y_j^n через точний розв'язок і похибку: $y_j^n = u_j^n + z_j^n$. Тоді різницева схема набере вигляду

$$\begin{cases} \dfrac{z_j^{n+1}-z_j^n}{\tau} = \dfrac{z_{j+1}^n - 2z_j^n + z_{j-1}^n}{h^2} + \psi_j^n, (j=\overline{1,N-1}, n=\overline{0,K-1}); \\ z_j^0 = 0, \quad j=\overline{0,N}; \\ z_0^n = 0, \quad n=\overline{0,K}; \\ z_N^n = 0, \quad n=\overline{0,K}. \end{cases}$$

Виразимо з першого рівняння похибку на $n+1$ часовому шарі

$$z_j^{n+1} = (1-\frac{2\tau}{h^2})z_j^n + \frac{\tau}{h^2}(z_{j+1}^n + z_{j-1}^n) + \tau\psi_j^n.$$

Отримаємо оцінку на похибку

$$|z_j^{n+1}| \le \left|1-\frac{2\tau}{h^2}\right||z_j^n| + \frac{\tau}{h^2}(|z_{j+1}^n| + |z_{j-1}^n|) + |\tau\psi_j^n| \le (\left|1-\frac{2\tau}{h^2}\right| + 2\frac{\tau}{h^2})\max_j|z_j^n| + \tau\max_j|\psi_j^n|.$$

Норму похибки на n часовому шарі визначимо так: $\|z_j^n\|_{G_h} = \max_j|z_j^n|$. Тоді на $n+1$ шарі маємо

$$\|z_j^{n+1}\|_{G_h} \le (\left|1-\frac{2\tau}{h^2}\right| + 2\frac{\tau}{h^2})\|z_j^n\|_{G_h} + \tau\|\psi_j^n\|_{G_h}.$$

Якщо вважати $1-2\frac{\tau}{h^2} \ge 0$, що накладає обмеження на крок, то $\|z_j^{n+1}\|_{G_h} \le \|z_j^n\|_{G_h} + \tau\|\psi_j^n\|_{G_h}$. Ця оцінка справедлива для будь-якого n. Застосувавши її рекурсивно n разів, отримаємо $\|z_j^n\|_{G_h} \le \|z_j^0\|_{G_h} + \tau\sum_{k=0}^{n-1}\|\psi_j^k\|_{G_h}$. Але, згідно з умовою, перший

доданок справа дорівнює нулю, тому $\left\|z_j^n\right\|_{G_h} \le \tau\sum_{k=0}^{n-1}\left\|\psi_j^k\right\|_{G_h}$. Згадаємо про оцінку на нев'язку $\left\|\psi\right\|_{G_h} = O(\tau + h^2)$. Звідси

$$\left\|\psi_j^k\right\|_{G_h} = M_k(\tau + h^2),\ \max_{k=0,n-1} M_k = \overline{M},$$

$$\left\|z_j^n\right\|_{G_h} \le \tau(\tau + h^2)\sum_{k=0}^{n-1} M_k \le t_n\overline{M}(\tau + h^2).$$

Позначимо $t_n\overline{M} = \breve{M}$ - це константа, яка не залежить від кроку τ, оскільки $t_n = t_n$. Отримаємо остаточну оцінку

$$\left\|z_j^n\right\|_{G_h} \le \breve{M}(\tau + h^2).$$

Тобто отримана різницева схема має перший порядок точності по τ та другий – по h.

Структура різницевої схеми і задачі для похибки однакова, відмінність лише в тім, що в різницевій схемі права частина дорівнює f, а в задачі для похибки - ψ. Отже, за аналогією маємо $\left\|y_j^n\right\|_{G_h} \le \left\|y_j^0\right\|_{G_h} + \tau\sum_{k=0}^{n-1}\left\|\varphi_j^k\right\|_{G_h}$. Це дискретний аналог принципу максимуму для рівняння теплопровідності. Він свідчить про те, що розв'язок крайової задачі для рівняння теплопровідності є стійким за початковими даними та правою частиною. Але згадаємо, що ця оцінка отримана при обмеженні кроків дискретизації

$$\frac{\tau}{h^2} \le \frac{1}{2},$$

тобто побудована різницева схема (9.27) є ***умовно стійкою.*** В цьому є істотний недолік таких схем, адже, наприклад, взявши $h = 0.01$, треба обрати $\tau \le \frac{1}{2}10^{-4}$. Якщо $T = 1$, то кількість кроків за часом буде $N = \frac{1}{\tau} \approx 20000$. Для того, щоб виправити

цей недолік, підійдемо інакше до апроксимації крайової задачі, обравши інший шаблон для побудови різницевої схеми.

9.6.2 Неявна різницева схема

Оберемо шаблон, зображений на рис.9.3 б). Це досягається використанням інших формул чисельної апроксимації відповідних похідних. Рівняння різницевої схеми наберуть вигляду:

$$\begin{cases} \dfrac{y_j^{n+1} - y_j^n}{\tau} = \dfrac{y_{j+1}^{n+1} - 2y_j^{n+1} + y_{j-1}^{n+1}}{h^2} + \varphi_j^n, (j = \overline{1, N-1}, n = \overline{0, K-1}); \\ y_j^0 = u^0(x_j),\ j = \overline{0, N}; \\ y_0^{n+1} = \mu_1(t_{n+1}),\ n = \overline{0, K-1}; \\ y_N^{n+1} = \mu_2(t_{n+1}),\ n = \overline{0, K-1}. \end{cases} \quad (9.28)$$

Тоді нев'язка матиме вигляд

$$\psi_j^{n+1} = -\frac{u_j^{n+1} - u_j^n}{\tau} + \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} + \varphi_j^{n+1}. \quad (9.29)$$

Як і в попередньому випадку, розкладемо u_{j+1}^{n+1} та u_j^n в ряд Тейлора:

$$u_{j\pm1}^{n+1} = u_j^{n+1} \pm u_{x,j}^{n+1}h + u_{xx,j}^{n+1}\frac{h^2}{2} \pm u_{xxx,j}^{n+1}\frac{h^2}{6} + O(h^4);$$

$$u_j^n = u_j^{n+1} - u_{t,j}^{n+1}\tau + O(\tau^2).$$

Тепер підставимо ці розкладання в (9.29) і в праву частину додамо та віднімемо f_j^{n+1}

$$\psi_j^{n+1} = (-u_{t,j}^{n+1} + u_{xx,j}^{n+1} + f_j^{n+1}) + \varphi_j^{n+1} - f_j^{n+1} + O(\tau + h^2).$$

Якщо функцію φ_j^{n+1} взяти такою, що дорівнює f_j^{n+1} з точністю $O(\tau + h^2)$, то нев'язка набере вигляду

$$\psi_j^{n+1} = O(\tau + h^2).$$

Ця різницева схема має перший порядок апроксимації по τ і другий по h.

Для дослідження на стійкість використаємо метод гармонік. Зіставимо наше різницеве рівняння й: однорідне рівняння

$$\frac{y_j^{n+1}-y_j^n}{\tau}=\frac{y_{j+1}^{n+1}-2y_j^{n+1}+y_{j-1}^{n+1}}{h^2}. \qquad (9.30)$$

Дослідимо його на розв'язку вигляду $y_j^n=q^n e^{ijh\lambda}$, де q,λ - деякі параметри. Підставимо таке y_j^n в (9.30) і скоротимо:

$$\frac{q-1}{\tau}=q\frac{e^{ih\lambda}-2+e^{-ih\lambda}}{h^2} \Rightarrow \frac{q-1}{\tau}=-q\frac{4}{h^2}\sin^2\frac{h\lambda}{2}.$$

Звідси $q=\dfrac{1}{1+\dfrac{4\tau}{h^2}\sin^2\dfrac{h\lambda}{2}}$. Очевидно, оскільки знаменник завжди не менше ніж 1, що ця неявна різницева схема є ***абсолютно стійкою***, тобто є стійкою при будь-яких значеннях кроків сітки.

Розглянемо проблему розв'язання системи (9.29). Перепишемо перше рівняння з (9.29) так:

$$\frac{\tau}{h^2}y_{j-1}^{n+1}-(1+\frac{2\tau}{h^2})y_j^{n+1}+\frac{\tau}{h^2}y_{j+1}^{n+1}=-y_j^n+\tau\varphi_j^{n+1}.$$

Розглянемо його при $n=0$:

$$\frac{\tau}{h^2}y_{j-1}^1-(1+\frac{2\tau}{h^2})y_j^1+\frac{\tau}{h^2}y_{j+1}^1=-y_j^0+\tau\varphi_j^1,\ \ j=\overline{1,N-1}$$

- це система лінійних алгебраїчних рівнянь відносно $y^1=(y_0^1,y_1^1,...,y_N^1)$, права частина якої повністю визначається початковими умовами, а матриця коефіцієнтів при невідомих має тридіагональну структуру. Отже, можна застосувати метод прогонки. Таким чином, різницева функція може бути визначена у всіх вузлах першого часового шару. Далі застосуємо ті самі міркування для наступних часових шарів, що дозволить знайти значення різницевої функції у всій сітковій області.

9.6.3 Схема з вагами для рівняння теплопровідності

Розглянемо тепер не мінімально можливий шаблон, а «надлишковий» (рис.9.3в). «Надлишковість» схеми компенсуємо введенням деякого параметра – вагового множника. Різницева схема для крайової задачі така:

$$\begin{cases}\dfrac{y_j^{n+1}-y_j^n}{\tau}=\sigma\dfrac{y_{j+1}^{n+1}-2y_j^{n+1}+y_{j-1}^{n+1}}{h^2}+(1-\sigma)\dfrac{y_{j+1}^n-2y_j^n+y_{j-1}^n}{h^2}+\varphi_j^n,\\ (j=\overline{1,N-1},n=\overline{0,K-1});\\ y_j^0=u^0(x_j),\ j=\overline{0,N};\\ y_0^{n+1}=\mu_1(t_{n+1}),\ n=\overline{0,K-1};\\ y_N^{n+1}=\mu_2(t_{n+1}),\ n=\overline{0,K-1}.\end{cases} \qquad (9.31)$$

Вочевидь, при $\sigma=0$ маємо явну різницеву схему, а при $\sigma=1$ - неявну різницеву схему.

Визначимо порядок апроксимації .Розглянемо нев'язку

$$\psi_j^{n+1/2}=\frac{u_j^{n+1}-u_j^n}{\tau}+\sigma u_{\bar{x}x,j}^{n+1}+(1-\sigma)u_{\bar{x}x,j}^n+\varphi_j^n.$$

Розкладемо функції u_j^n і u_j^{n+1} в ряд Тейлора:

$$u_j^{n+1}=u_j^{n+1/2}+u_{t,j}^{n+1/2}\frac{\tau}{2}+\frac{1}{2}u_{tt,j}^{n+1/2}\frac{\tau^2}{4}+O(\tau^3);$$

$$u_j^n=u_j^{n+1/2}-u_{t,j}^{n+1/2}\frac{\tau}{2}+\frac{1}{2}u_{tt,j}^{n+1/2}\frac{\tau^2}{4}+O(\tau^3).$$

Тоді отримаємо

$$\psi_j^{n+1/2}=-u_{t,j}^{n+1/2}+O(\tau^2)+\sigma u_{\bar{x}x,j}^{n+1}+(1-\sigma)u_{\bar{x}x,j}^n+\varphi_j^n. \qquad (9.32)$$

Тепер у розкладанні другої різницевої похідної розкладемо всі функції в ряд Тейлора з членами до п'ятого порядку включно:

$$u_{\bar{x}x}(x_j,t)=\frac{1}{h^2}(u(x_{j+1},t)-2u(x_j,t)+u(x_{j-1},t))=$$

$$=\left\{u(x_{j\pm1},t)=u(x_j,t)\pm u_x(x_j,t)\frac{h^2}{2}\pm u_{xxx}(x_j,t)\frac{h^3}{6}+u_{xxxx}(x_j,t)\frac{h^4}{4!}\pm\right.$$

$$\left.\pm u_{xxxxx}(x_j,t)\frac{h^5}{5!}+O(h^6)\ \right\}=u_{xx}(x_j,t)+u_{xxxx}(x_j,t)\frac{h^2}{12}+O(h^4).$$

Скористаємося цим розкладанням для доданків у виразі (9.33) для розкладання в ряд Тейлора з центром у точці $x_j^{n+1/2}$:

$$u_{\bar{x}x,j}^{n+1}=u_{xx,j}^{n+1}+u_{xxxx,j}^{n+1}\frac{h^2}{12}+O(h^4)=u_{xx,j}^{n+1/2}+u_{xxt,j}^{n+1/2}\frac{\tau}{2}+u_{xxxx,j}^{n+1/2}\frac{h^2}{12}+u_{xxxxt,j}^{n+1/2}\frac{h^2}{12}\frac{\tau}{2}+O(\tau^2+h^4);.$$

$$u_{\bar{x}x,j}^{n}=u_{xx,j}^{n}+u_{xxxx,j}^{n}\frac{h^2}{12}+O(h^4)=u_{xx,j}^{n+1/2}-u_{xxt,j}^{n+1/2}\frac{\tau}{2}+u_{xxxx,j}^{n+1/2}\frac{h^2}{12}-u_{xxxxt,j}^{n+1/2}\frac{h^2}{12}\frac{\tau}{2}+O(\tau^2+h^4).$$

Таким чином, вираз для нев'язки набирає вигляду

$$\psi_j^{n+1/2}=-u_{t,j}^{n+1/2}+u_{xx,j}^{n+1/2}+u_{xxxx,j}^{n+1/2}\frac{h^2}{12}+\varphi_j^n+(\sigma-\frac{1}{2})u_{xxt,j}^{n+1/2}\tau+$$

$$+(\sigma-\frac{1}{2})u_{xxxxt,j}^{n+1/2}\frac{h^2}{12}\tau+O(\tau^2+h^4).$$

Додаючи та віднімаючи $f_j^{n+1/2}$, отримаємо

$$\psi_j^{n+1/2}=-u_{t,j}^{n+1/2}+u_{xx,j}^{n+1/2}+f_j^{n+1/2}+u_{xxxx,j}^{n+1/2}\frac{h^2}{12}+\varphi_j^n-f_j^{n+1/2}+$$

$$+(\sigma-\frac{1}{2})u_{xxt,j}^{n+1/2}\tau+(\sigma-\frac{1}{2})u_{xxxxt,j}^{n+1/2}\frac{h^2}{12}\tau+O(\tau^2+h^4).$$

Згідно з рівнянням теплопровідності перші три доданки обертаються на нуль

$$\psi_j^{n+1/2}=u_{xxxx,j}^{n+1/2}\frac{h^2}{12}+\varphi_j^n-f_j^{n+1/2}+(\sigma-\frac{1}{2})u_{xxt,j}^{n+1/2}\tau+(\sigma-\frac{1}{2})u_{xxxxt,j}^{n+1/2}\frac{h^2}{12}\tau+O(\tau^2+h^4).$$

При $\sigma=\frac{1}{2}$ схема (9.31) називається симетричною, або схемою Кранка-Ніколсона. Тоді в останній рівності останні доданки обнуляються, а за допомогою умови на параметр

$\varphi_j^n=f_j^{n+1/2}+O(\tau^2+h^2)$ ми можемо досягнути такого порядку апроксимації:

$$\psi_j^n=O(\tau^2+h^2).$$

Тепер повернемося на крок назад і двічі продиференціюємо рівняння теплопровідності по змінній x: $u_{xxt}=u_{xxxx}+f_{xx}$. Тоді формула для нев'язки буде дещо іншою:

$$\psi_j^{n+1/2}=\varphi_j^n-f_j^{n+1/2}-f_{xx,j}^{n+1/2}\frac{h^2}{12}+$$

$$+[(\sigma-\frac{1}{2})\tau+\frac{h^2}{12}]u_{xxt,j}^{n+1/2}+(\sigma-\frac{1}{2})u_{xxxxt,j}^{n+1/2}\frac{h^2}{12}\tau+O(\tau^2+h^4).$$

Узявши $\sigma=\frac{1}{2}-\frac{h^2}{12\tau}$, ми спростимо вираз. Якщо поставити вимогу, щоб $\varphi_j^n=f_j^{n+1/2}+f_{xx,j}^{n+1/2}\frac{h^2}{12}+O(\tau^2+h^4)$, тоді матимемо такий порядок апроксимації:

$$\psi_j^{n+1.2}=O(\tau^2+h^4).$$

При такому σ різницева схема (9.31) називається різницевою схемою підвищеного порядку апроксимації.

Тепер дослідимо схему на стійкість методом гармонік. Розглянемо однорідне різницеве рівняння

$$\frac{y_j^{n+1}-y_j^n}{\tau}=\sigma\frac{y_{j+1}^{n+1}-2y_j^{n+1}+y_{j-1}^{n+1}}{h^2}+(1-\sigma)\frac{y_{j+1}^n-2y_j^n+y_{j-1}^n}{h^2}.$$

Підставимо розв'язок $y_j^n=q^ne^{ijh\lambda}$ і скоротимо:

$$\frac{q-1}{\tau}=\sigma\frac{q}{h^2}(e^{ih\lambda}-2+e^{-ih\lambda})+(1-\sigma)\frac{1}{h^2}(e^{ih\lambda}-2+e^{-ih\lambda})\Rightarrow$$

$$\Rightarrow\frac{q-1}{\tau}=-4\sigma\frac{q}{h^2}\sin^2\frac{h\lambda}{2}-4(1-\sigma)\frac{1}{h^2}\sin^2\frac{h\lambda}{2}.$$

Звідси $q = \dfrac{1-(1-\sigma)4\dfrac{\tau}{h^2}\sin^2\dfrac{h\lambda}{2}}{1+4\sigma\dfrac{\tau}{h^2}\sin^2\dfrac{h\lambda}{2}}$.

Для отримання умов на стійкість вимагаємо, щоб $\|q\| \le 1$. Це дає дві нерівності:

$$\begin{cases} 1-(1-\sigma)4\dfrac{\tau}{h^2}\sin^2\dfrac{h\lambda}{2} \le 1+4\sigma\dfrac{\tau}{h^2}\sin^2\dfrac{h\lambda}{2}; \\ -1-4\sigma\dfrac{\tau}{h^2}\sin^2\dfrac{h\lambda}{2} \le 1-(1-\sigma)4\dfrac{\tau}{h^2}\sin^2\dfrac{h\lambda}{2}. \end{cases}$$

Перша виконується завжди, оскільки $\tau > 0$. З другої знайдемо $\sigma \ge \dfrac{1}{2} - \dfrac{h^2}{4\tau\sin^2\dfrac{h\lambda}{2}}$. Взявши максимум за правою частиною, приходимо до остаточної умови для σ

$$\sigma \ge \frac{1}{2} - \frac{h^2}{4\tau}.$$

Значення $\sigma = \dfrac{1}{2}$ задовольняє цю нерівність. Це значить, що метод Кранка-Ніколсона є абсолютно стійким.

Обґрунтуємо можливість обчислень за цією схемою. Для цього (9.31) перепишемо у вигляді

$$\sigma\frac{\tau}{h^2}y_{j+1}^{n+1} - (1+2\sigma\frac{\tau}{h^2})y_j^{n+1} + \sigma\frac{\tau}{h^2}y_{j-1}^{n+1} = -y_j^n - (1-\sigma)\tau y_{\bar{x}x,j}^n - \varphi_j^n.$$

Для кожного часового шару маємо, як і у випадку неявної схеми, систему лінійних алгебраїчних рівнянь з тридіагональною матрицею. До її розв'язку можна застосувати метод прогонки, тим самим визначити наближення у всіх вузлах сітки.

Приклад програмної реалізації на псевдокоді чисельного розв'язання першої крайової задачі для рівняння параболічного типу методом сіток по явній і неявній схемах.

```
f(x):
//повертає значення u(x,0)
end
f1(x):
//повертає значення u(0,t)
end
f2(x):
//повертає значення u(1,t)
end
//Явна схема
//за умови, що τ=h^2/6
//h - крок по x
//u - матриця з шуканими значеннями функції
Solve_Yavnaya(h,u):
    1  ta:=sqr(h)/6
    2  m:=10
    3  n:=trunc(1/h+0.5)
    4  for k:=0 to m do //Граничні умови
    5      U[1,k+1]:=f1(k*ta);
    6      U[n+1,k+1]:=f2(k*ta);
    7  done
    8  for i:=1 to (n-1) do
//початкові умови
```

```
        9       U[i+1,1]:=f(i*h)
        10 done
   11   for k:=0 to (m-1) do //Явна схема
        12      for i:=1 to (n-1) do
         13U[i+1,k+2]:=(U[i,k+1]+4*U[i+1,k
+1]+U[i+2,k+1])/6;
        14      done
        15 done
   end
   //неявна схема
   //за умови, що τ=h²/6
   //h - крок по x
   //P,Q - масиви прогоночних коефіцієнтів
//U - матриця з шуканими значеннями функції
   Solve_Neyavnaya(h,U):
        1    ta:=sqr(h)/6
        2    m:=10
        3    n:=trunc(1/h+0.5)
   4  for k:=0 to m do //граничні умови
        5       U[1,k+1]:=f1(k*ta)
        6       U[n+1,k+1]:=f2(k*ta)
        7    done
        8    for i:=1 to (n-1) do
     //початкові умови
        9       U[i+1,1]:=f(i*h)
        10   done
```



```
        11   la:=ta/sqr(h)
12      for k:=1 to m do //розв'язання СЛАР
      методом прогонки
        13      P[1]:=la/(1+2*la)
14      Q[1]:=(U[i+1,k]+la*f1(k*ta))/(1+2*la)
        15      for i:=2 to (n-1) do
        16   P[i]:=la/(1+2*la-la*P[i-1])
17   Q[i]:=(U[i+1,k]+la*Q[i-1])/(1+2*la-
la*P[i-1])
        18      done
19   U[n,k+1]:=(U[n,k]+la*f2(k*ta))/(1+2*la-
la*P[n-2])
        20      for i:=(n-2) downto 1 do
     21    U[i+1,k+1]:=Q[i]+P[i]*U[i+2,k+1]
        22      done
        23   done
   end
```

9.7 Різницева схема для рівняння гіперболічного типу

Розглянемо крайову задачу для рівняння коливань, що належить до гіперболічного типу:

$$
\begin{aligned}
&u_{tt} = u_{xx} + f(x,t), 0 < x < 1, 0 < t \le T;\\
&u(0,t) = \mu_1(t), \quad 0 \le t \le T;\\
&u(1,t) = \mu_2(t), \quad 0 \le t \le T;\\
&u(x,0) = u^0(x), \quad 0 \le x \le 1;\\
&u(x,0) = \psi(x), \quad 0 \le x \le 1.
\end{aligned}
\tag{9.32}
$$

Розв'язок такої задачі існує і єдиний.

Уведемо дискретну сітку на розглянутій області.

$$\omega_h = \left\{ x_j = jh, h = \frac{1}{N}, j = \overline{0, N} \right\}; \omega_\tau = \left\{ t_n = n\tau, \tau = \frac{T}{K}, n = \overline{0, k} \right\}.$$

Зауважимо, що початкову умову дають нам значення шуканої функції на границі прямокутника.

Будемо використовувати вже звичні нам позначення: $u(x_{j,t_n}) = u_j^n$ - точний розв'язок у вузлах сітки, y_j^n - шукане наближення.

Випишемо аналоги крайових і початкових умов

$$\begin{cases} y_0^n = \mu_1(t_n); \\ y_N^n = \mu_2(t_n); \\ y_j^0 = u^0(x_j). \end{cases}$$

Зіставимо рівняння й дискретний аналог у кожному вузлі сітки. Будемо використовувати шаблон з п'яти точок (рис.9.4г - це мінімально необхідний шаблон для апроксимації других похідних по t і по x).

Наближення побудуємо так:

$$u_{tt} \approx u_{\overline{tt},j}^n = \frac{1}{\tau^2}(u_j^{n+1} - 2u_j^n + u_j^{n-1});$$

$$u_{xx} \approx u_{\overline{xx},j}^n = \frac{1}{h^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n).$$

Тоді одержимо таку різницеву схему:

$$\begin{cases} y_{\overline{tt},j}^n = y_{\overline{xx},j}^n + \varphi_j^n, j = \overline{1, N-1}, n = \overline{1, K-1}; \\ y_0^n = \mu_1(t_n); \qquad n = \overline{0, K}; \\ y_N^n = \mu_2(t_n); \qquad n = \overline{0, K}; \\ y_j^0 = u^0(x_j); \qquad j = \overline{0, N}; \\ \dfrac{y_j^1 - y_j^0}{\tau} = \psi(x_j), j = \overline{0, N}. \end{cases} \tag{9.33}$$

Варто відзначити, що для того щоб різницева схема була збалансованою (тобто щоб не робити зайвих обчислень в одному місці, а потім загрубляти їх в іншому), необхідно, щоб порядки апроксимації в диференціальному рівнянні та у крайовій умові другого типу були узгоджені. Інакше, якщо використовувати низький (перший) порядок апроксимації в крайовій умові, то вся схема буде апроксимувати вихідну задачу з першим порядком апроксимації.

Постараємося всі дані в задачі наблизити з другим порядком точності. Для цього розкладемо u_j^1 в ряд Тейлора в точці $(x_j, 0)$ з залишковим членом другого порядку малості

$$\frac{u_j^1 - u_j^0}{\tau} = u_{t,j}^0 + u_{tt,j}^0 \frac{\tau}{2} + o(\tau^2).$$

Спробуємо позбутися від $u_{tt,j}^0$. Нехай на границі також виконується рівняння коливання, тоді $u_{tt} = u_{xx} + f(x,0)$, і крайова умова набере вигляду

$$\frac{u_j^1 - u_j^0}{\tau} = u_{t,j}^0 + u_{xx,j}^0 + f(x_j, 0))\frac{\tau}{2} + o(\tau^2),$$

де вираз $u_{xx,j}^0 + f(x_j, 0)$ точно визначимо із початкових умов.

Тоді різницеву схему можна записати у вигляді

$$\begin{cases} y^n{}_{\overline{tt},j} = y^n{}_{\overline{xx},j} + \varphi^n{}_j, j = \overline{1, N-1}, n = \overline{1, K-1}; \\ y^n{}_0 = \mu_1(t_n), n = \overline{1, K}; \\ y^n{}_N = \mu_2(t_n), n = \overline{1, K} \\ y^0{}_j = u^0(x_j), n = \overline{0, N} \\ \dfrac{y^1{}_j - y^0{}_j}{\tau} = \psi(x_j) + \dfrac{\tau}{2}(u^0{}_{xx}(x_j) + f^0{}_j), j = \overline{0, N}. \end{cases}$$

Порядок апроксимації крайової умови буде $O(\tau^2)$. Будемо вимагати в задачі другого порядку апроксимації і в першому рівнянні. Для цього достатньо взяти $\varphi_j^n = f(x_j, t_n) + O(h^2 + \tau^2)$.

Таким чином, різницевий аналог вихідної задачі буде її апроксимувати з другим порядком по τ і по h.

Розглянемо питання про визначення наближеного розв'язку. Знайдемо із побудованої різницевої схеми вираз для y_j^{n+1}

$$y_j^{n+1} = 2y_j^n - y_j^{n-1} + \tau^2 y_{\bar{x}x,j}^n + \tau^2 \varphi_j^n. \qquad (9.34)$$

Покажемо, як визначається сіткова функція на всій сітці. Розглянемо формулу(9.34) при $n=1$

$$y_j^2 = 2y_j^1 - y_j^0 + \frac{\tau^2}{h^2}(y_{j+1}^1 - 2y_j^1 + y_{j-1}^1)\tau^2\varphi_j^1.$$

Оскільки y_j^0 відоме з початкових умов, а y_j^1 можемо знайти із другої крайової умови, то y_j^2 знаходиться явно на усьому другому часовому шарі. Аналогічно можна знайти y_j^3 і так далі, тобто знайти значення шуканої сіткової функції на всій сітці (як і раніше роблячи це попарово).

Як видно із процесу пошуку розв'язку, ця схема – явна і побудований розв'язок буде єдиним.

Розглянемо питання стійкості. Будемо використовувати відомий нам метод гармонік. Запишемо відповідне однорідне рівняння

$$y_j^{n+1} - 2y_j^n + y_j^{n-1} = \frac{\tau^2}{h^2}(y_{j+1}^n - 2y_j^n + y_{j-1}^n).$$

Підставимо в нього розв'язок вигляду $y_j^n = q^n e^{ijh\varphi}$, тоді отримаємо

$$q^2 - 2q + 1 = \frac{\tau^2}{h^2} q(e^{ijh\varphi} - 2 + e^{-ih\varphi}) \Leftrightarrow q^2 - (2 - 4\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2})q + 1 = 0$$

Розв'язки цього квадратного рівняння будуть такими:

$$q_{1,2} = \left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right) \pm \sqrt{\left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right)^2 - 1}.$$

Проаналізуємо дискримінант (підкорінний вираз) квадратного рівняння:

1 $D>0$: оскільки вільний член дорівнює одиниці, то $|q_1||q_2|=1$. В цьому випадку (можна показати, що корені не можуть бути протилежними по знаку) $|q_1|$ або $|q_2|$ більше одиниці, тому стійкості не буде при даних значеннях параметрів.

2 $D\le 0$: аналогічно, $|q_1||q_2|=1$. Знайдемо значення абсолютної величини кореня

$$|q_1|^2 = \left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right)^2 + 1 - \left(1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right)^2 = 1.$$

Тоді $|q_1| = |q_2| = 1$, звідки випливає, що розв'язок буде стійким.

Розпишемо умову недодатності дискримінанту

$$\left|1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2}\right| \le 1,$$

або розписавши

$$-1 \le 1 - 2\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2} \le 1.$$

Права нерівність виконана завжди, перепишемо ліву

$$\frac{\tau^2}{h^2}\sin^2\frac{h\varphi}{2} \le 1.$$

У гіршому випадку отримаємо $\frac{\tau^2}{h^2} \le 1$, що еквівалентно $\frac{\tau}{h} \le 1$ (нагадаємо, що в рівнянні теплопровідності ми отримували обмеження вигляду $\frac{\tau^2}{h^2} \le \frac{1}{2}$). Отримуємо, що побудована схема є умовно стійкою.

Отже, можна виділити кілька особливостей для рівняння коливань:

1) додаткова умова на апроксимацію;

2) розв'язок будується пошарово;
3) умовна стійкість з простими обмеженнями.

Приклад програмної реалізації на псевдокоді чисельного розв'язання змішаної задачі для хвильового рівняння методом сіток

```
f(x):
//повертає значення u^0(x)
end
g(x):
//Повертає значення ψ(x)
end
f1(t):
//повертає значення μ1(t)
end
f2(t):
//повертає значення μ2(t)
end
//h - крок по x
//k - крок по t
//U - матриця з шуканими значеннями функції
Solve_WaveExpr(h,k,U):
1     n:=trunc(1/h+0.5)
2     la:=k/h
3   for j:=0 to m do //граничні умови
4       U[1,j+1]:=f1(j*k)
5       U[n+1,j+1]:=f2(j*k)
6   done
7   for i:=1 to (n-1) do
//початкові умови
8       U[i+1,1]:=f(i*h);
9       U[i+1,2]:=f(i*h)+k*g(i*h);
10  done
11  for j:=1 to (m-1) do
12      for i:=1 to (n-1) do
13  U[i+1,j+2]:=2*(1-
sqr(la))*U[i+1,j+1]+sqr(la)*(U[i+2,j+1]+U[i,j+
1])-U[i+1,j];
14      done
15  done
end
```


9.8 Рівняння еліптичного типу

Прикладом повної математичної постановки задачі для рівняння еліптичного типу є задача з крайовими умовами першого роду, яка має назву *задачі Діріхле для рівняння Пуассона*

$$a(x,y)\frac{\partial^2 u}{\partial x^2}+b(x,y)\frac{\partial^2 u}{\partial y^2}+c(x,y)\frac{\partial u}{\partial x}+d(x,y)\frac{\partial u}{\partial y}+g(x,y)=f(x,y), \quad (9.35)$$

яке задане в однозв'язній області G з границею Γ. Коефіцієнти в рівнянні, права частина та границя Γ є достатньо гладкими: $a(x,y)>0$, $b(x,y)>0$, $g(x,y)\le 0$ в G. Потрібно знайти функцію $u(x,y)$, яка задовольняє всередині деякої області G рівняння (9.35), а на границі Γ – умову

$$u(x,y)\big|_{\Gamma}=\varphi(x,y), \quad (9.36)$$

де $\varphi(x,y)$ — задана неперервна функція. Припускаємо, що $f(x,y)$, $\varphi(x,y)$ є такі, що розв'язок задачі (9.35), (9.36) існує, є єдиним і є достатньо гладкою функцією. При $f = 0$ отримуємо задачу Діріхле для рівняння Лапласа, однією з важливих властивостей якої є виконання **принципу максимуму**: ***неперервний в G і відмінний від константи розв'язок u(x,y) може досягати свого максимального за модулем значення тільки на границі Г.*** Звідси випливає, що є справедливою оцінка

$$\max_{x,y\in G}|u(x,y)| \le \max_{x,y\in \Gamma}|\varphi(x,y)|, \qquad (9.37)$$

яка означає стійкість задачі за граничними даними. Вкриємо область G площини (x,y) сіткою паралельних прямих $x_i = x_0 + ih, \quad y_k = y_0 + kh, \quad i,k = 0,\pm1,\pm2,\ldots$, утворивши сітчасту область вузлів (x_i,y_k) G_h з границею Γ_h (рис. 9.1).

Замінимо похідні рівняння (9.35) у внутрішніх вузлах різницевими співвідношеннями другого порядку точності апроксимації за формулами:

$$\frac{\partial u}{\partial x}(x_i,y_k) = \frac{u(x_{i+1},y_k)-(x_{i-1},y_k)}{2h} + O(h^2);$$

$$\frac{\partial u}{\partial y}(x_i,y_k) = \frac{u(x_i,y_{k+1})-(x_i,y_{k-1})}{2h} + O(h^2);$$

$$\frac{\partial^2 u}{\partial x^2}(x_i,y_k) = \frac{u(x_{i+1},y_k)-2u(x_i,y_k)+(x_{i-1},y_k)}{h^2} + O(h^2);$$

$$\frac{\partial^2 u}{\partial y^2}(x_i,y_k) = \frac{u(x_i,y_{k+1})-2u(x_i,y_k)+(x_i,y_{k-1})}{h^2} + O(h^2).$$

Підставляючи ці співвідношення в (9.35), відкинувши похибку апроксимації похідних, отримаємо різницеві рівняння для невідомих значень сіткової функції $u^h_{i,k}$

$$a_{i,k}\frac{u^h_{i+1,k}-2u^h_{i,k}+u^h_{i-1,k}}{h^2} + b_{i,k}\frac{u^h_{i,k+1}-2u^h_{i,k}+u^h_{i,k-1}}{h^2} + c_{i,k}\frac{u^h_{i+1,k}-u^h_{i-1,k}}{2h} + d_{i,k}\frac{u^h_{i,k+1}-u^h_{i,k-1}}{2h} + g_{i,k}u_{i,k} = f_{i,k}, \qquad (9.38)$$

де введені позначення коефіцієнтів і правої частини у вузлі (x_i, y_k): $a_{i,k}$, $b_{i,k}$, $c_{i,k}$, $d_{i,k}$, $g_{i,k}$, $f_{i,k}$ ($f_{i,k} = f(x_i, y_i)$, $(x_i,y_k) \in G_h$). Тут для внутрішніх вузлів використовувався п'ятиточковий шаблон (дивись рис. 9.3г).

Співвідношення (9.38) містять, крім невідомих $u^h_{i,k}$ у внутрішніх вузлах, ще й невідомі $u_{i,k}$ на границі сітчастої області. Для граничних вузлів запишемо співвідношення

$$u(x_i,y_k) = \frac{\theta u(x_{i+1},y_k)+\varphi(x_i \pm \theta h, y_k)}{\theta+1} + O(h^2), 0$$

або $$u(x_i,y_k) = \frac{\theta u(x_i,y_{k+1})+\varphi(x_i, y_k \pm \theta h)}{\theta+1} + O(h^2),$$

залежно від того, яка точка $(x_i \pm \theta h, y_k)$, або $(x_i, y_k \pm \theta h)$, $0 < \theta < 1$ перетину неперервної границі Γ з лініями сітки знаходиться ближче до граничного вузла. Ці співвідношення означають, що значення $u(x_b y_k)$ при $(x_b y_k) \in \Gamma_h$ отримуються лінійною інтерполяцією значень $u(x,y)$ у внутрішньому вузлі і в точці перетину Γ з сіткою. Відкинувши в останніх співвідношеннях похибку апроксимації, дістанемо вирази для невідомих $u^h_{i,k}$ в граничних вузлах

$$u^h_{i,k} = \frac{\theta}{\theta+1}u^h_{i\pm1,k\pm1} + \frac{1}{\theta+1}\varphi^h_{i\pm\theta,k\pm\theta}, \qquad (9.39)$$

де введено позначення $\varphi^h_{i\pm\theta,k} = \varphi(x_i \pm \theta h, y_k)$, $\varphi^h_{i,k\pm\theta} = \varphi(x_i, y_k \pm \theta h)$. Приєднуючи рівняння (9.39) до (9.38), отримаємо систему лінійних рівнянь відносно $u^h_{i,k}$. У цій системі число рівнянь дорівнює числу невідомих і числу вузлів в області $G_h \cup \Gamma_h$.

Система рівнянь (9.38), (9.39) — це різницева схема неперервної задачі (9.35), (9.36). Роз'язок цієї різницевої схеми — наближення до точного розв'язку у вузлах (x_i,y_k).

Для розв'язання системи лінійних рівнянь отриманої різницевої схеми можуть застосовуватися методи, викладені в розділі 2 (метод простої ітерації).

Якщо розв'язок задачі Діріхле для заданої області і при заданих граничних умовах існує і виражається функцією $u(x,y)$ і якщо у вузлі з координатами x_i, y_k обраної сітки з кроком h отримано значення $u^h_{i,k}$ шуканої функції, то похибка наближеного розв'язку може бути оцінена нерівністю $\left| u(x_i, y_k) - u^h_{i,k} \right| < Ah^2$, де A не залежить від h.

Таким чином, теоретично похибка розв'язку задачі Діріхле зменшується зі зменшенням кроку сітки як h^2. Варто, однак, мати на увазі, що зі зменшенням кроку сітки збільшується число невідомих і відповідно порядок n системи лінійних алгебраїчних рівнянь, а це призводить, як правило, при великих n до накопичення похибок округлення.

Припустимо, що $g_{i,k} < 0$. Розв'яжемо систему лінійних рівнянь відносно $u^h_{i,k}$ методом простої ітерації згідно з ітераційним процесом:

для внутрішніх вузлів

$$u^{(p+1)}_{i,k} = p_{i,k}u^{(p)}_{i,k-1} + r_{i,k}u^{(p)}_{i-1,k} + s_{i,k}u^{(p)}_{i+1,k} + t_{i,k}u^{(p)}_{i,k+1} + z_{i,k};$$

для граничних вузлів

$$u^{(p+1)}_{i,k} = \frac{\theta}{\theta+1}u^{(p)}_{i\pm1,k\pm1} + \frac{1}{\theta+1}\varphi_{i\pm\theta,k\pm\theta},\ \text{p=0,1,2,...},\ u^{(0)}_{i,k}\ \text{задане.}$$

Доведено, що в такому разі послідовні наближення $u^{(p)}_{i,k}$ збігаються до точного розв'язку різницевої схеми $u^h_{i,k}$ або системи рівнянь (9.38), (9.39) і має місце оцінка

$$\max_{i,k}\left|u^{(p)}_{i,k} - u^h{}_{i,k}\right| \le \frac{q^p}{1-q}\max_{i,k}\left|u^{(0)}_{i,k}\right|,$$

де $q_{i,k} = \max\left(\frac{\theta_{i,k}}{1+\theta_{i,k}},\ p_{i,k} + r_{i,k} + s_{i,k} + t_{i,k}\right)$.

Доведення цього твердження полягає в перевірці умови збіжності методу простої ітерації для системи лінійних рівнянь, при цьому мається на увазі, що невідомий вектор $\bar{x}$ утворюють елементи $u_{i,k}$. Наприклад, компоненти вектора $\bar{x}$ можна перенумерувати таким чином: нехай $1 \le i \le N_1, 1 \le k \le N_2$, тоді

$$x_1 = u_{1,1},\quad x_2 = u_{2,1},\quad \ldots,\ x_{N_1} = u_{N_1,1};$$
$$x_{N_1+1} = u_{1,2},\ x_{N_1+2} = u_{2,2},\ \ldots,\ x_{2N_1} = u_{N_1,2};$$
$$\ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad \ldots\quad x_{N_1N_2} = u_{N_1,N_2}.$$

Відносно вектора $\bar{x} = \{x_1, x_2, \ldots, x_N\}$ різницева схема є системою лінійних рівнянь у матричному записі $A\bar{x} = \bar{b}$, де матриця A має в кожному рядку не більше п'яти елементів

$$A = \begin{pmatrix} * & * & . & . & . & . & . & . & . & . & . & . \\ * & * & * & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & * & . & . & * & * & * & . & * & . \\ . & . & . & * & . & . & * & * & * & . & * & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & * & . & * & . \end{pmatrix}$$

Це пов'язано з тим, що похідні в кожному внутрішньому вузлі (i,k) апроксимувалися за п'ятьма сусідніми вузлами.

Розв'язання різницевих рівнянь при $h \to 0$ збігається до точного розв'язку крайової задачі зі швидкістю, яка визначається порядком апроксимації рівнянь та крайових умов. Таким чином, для точного розв'язку $(u(x,y) \in C^4[G])$ оцінки похибки

$$\max_{i,k}\left|u^h{}_{i,k} - u(x_i, y_k)\right| = O(h^2),\ h \to 0 \quad (9.40)$$

Оцінка похибки (9.40) є справедливою, якщо точний розв'язок неперервно диференційований чотири рази в області G. Для областей з кутовими точками, наприклад прямокутника, взагалі кажучи, $u(x,y) \notin C^4[G]$. Але якщо гранична функція, тобто $\varphi(x,y)$, задовольняє в кутах спеціальні умови узгодження, то точний розв'язок $u(x,y) \in C^4[G]$ і є правильною оцінка (9.40).

Для прямокутної області $G = \{x_0 \le x \le x_1, y_0 \le y \le y_1\}$ такими умовами узгодження можуть бути:

достатня гладкість $\varphi(x,y)$;

функція $\varphi(x,y)$ повинна задовольняти в кутах прямокутника диференціальне рівняння.

Оцінка похибки (9.40) має в основному теоретичне значення, оскільки містить константу C, яку практично важко визначити,

$$\max_{i,k}\left|u_{i,k}-u(x_i,y_k)\right| = Ch^2+O(h^2),\ h\to 0$$

Тому в реальних розрахунках використовується правило Рунге оцінки похибки, аналогічне до того, яке використовується в чисельному розв'язанні задачі Коші і розв'язанні звичайних диференціальних рівнянь. Робиться два варіанти розрахунку $u^h_{i,k}$ з кроком h та $u^{h/2}_{i,k}$; тоді похибка має вигляд

$$\max\left|u^{h/2}_{i,k}-u(x_i,y_k)\right| = \frac{1}{3}\max\left|u^{h/2}_{i,k}-u^h_{i,k}\right|+O(h^2)$$

і головна частина похибки визначається на вузлах вихідної сітки.

Необхідно зазначити, що рівномірними прямокутними сітками найбільш зручно користуватися при розв'язанні задач у прямокутних областях. Якщо область має форму паралелограма (скошена система),то користуються координатами, осі яких паралельні сторонам цього паралелограма. Декартові прямокутні координати пов'язані з косокутними координатами (ξ,η) співвідношеннями $x=\xi+\eta\cos\alpha;\ \ y=\eta\sin\alpha$, де α — кут між ξ та η. У диференціальних виразах похідні по x та y замінюються похідними по ξ та η. Усі похідні апроксимуються за допомогою центральних різниць. Якщо область має форму кола, зручно користуватися полярними координатами $x=\rho\cos\theta, y=\rho\sin\theta$.

Наведемо деякі загальні зауваження. При чисельному розв'язанні крайових задач для диференціальних рівнянь у частинних похідних методом сіток можуть бути використані тільки збіжні різницеві схеми, оскільки в цьому разі можна розраховувати на отримання наближеного розв'язку задачі,

достатньо близького до точного. Але й збіжні різницеві схеми не завжди можуть бути використані при практичному розв'язанні задачі, оскільки, використовуючи метод сіток, при обчисленні значень граничних функцій та правої частини виникають похибки. Щоб ці похибки не спотворили істинного розв'язку різницевої схеми, остання повинна бути стійкою за граничними умовами і за правою частиною. При використанні нестійкої різницевої схеми спотворення істинного розв'язку тим сильніше, чим дрібніша сітка; при використанні ж великої сітки не можна розраховувати на те, що розв'язок різницевої схеми буде близький до точного розв'язку крайової задачі для диференціального рівняння через погану різницеву апроксимацію рівняння.

Крім того, під час розв'язання різницевої задачі в процесі розрахунків нам обов'язково доведеться округляти значення розв'язків у вузлах сітки. Ці помилки можуть значно спотворити розв'язок, тому необхідною вимогою є стійкість різницевої схеми щодо помилок, які виникають у результаті округлення значень розв'язку в вузлах сітки. Оскільки помилки округлення значень розв'язку в вузлах сітки, принаймні в найпростіших випадках, можна компенсувати зміною правої частини різницевого рівняння, то особливо суттєвою є вимога до стійкості правої частини. Необхідно взяти до уваги й числовий алгоритм, який використовується для розв'язання різницевої схеми. Навіть у випадку, коли різницева схема стійка за граничними умовами і за правою частиною, при невдалому виборі алгоритму для розрахунку розв'язку цієї різницевої схеми може відбутися сильне накопичення обчислювальної похибки, у цьому разі нестійким буде сам процес розрахунку. Нестійкі алгоритми розрахунку практично непридатні у випадку дрібної сітки.

Приклад програмної реалізації на псевдокоді чисельного розв'язання задачі Діріхле для рівняння Лапласа методом сіток

```
//Функції, що реалізують обчислення граничних
умов
```

```
f1(x):
//повертає значення u(0,y)
end
f2(x):
// повертає значення u(1,y)
end
f3(x):
// повертає значення u(x,0)
end
f4(x):
// повертає значення u(x,1)
end
//h - крок по x
//k - крок по t
Solve_Dirihle(h,k):
 1   for i:=0 to m do
 2       for j:=0 to n do
 3           V[i+1,j+1]:=0
 4           U[i+1,j+1]:=0
 5       done
 6   done
 7   repeat
 8       for i:=0 to m do
 9           for j:=0 to n do
10   U[i+1,j+1]:=V[1+i div 2,1+j div 2]
11           done
```

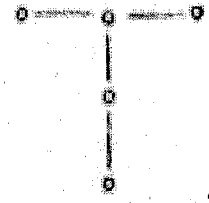


```
12       done
13       for j:=0 to n do
//граничні умови
14           U[1,j+1]:=f1(j*k)
15       done
16       for j:=0 to n do
17           U[m+1,j+1]:=f2(j*k)
18       done
19       for i:=1 to m do
20           U[i+1,1]:=f3(i*h)
21       done
22       for i:=1 to m do
23           U[i+1,n+1]:=f4(i*h)
24       done
25       p:=true
26       while p do //Метод ітерацій
27           for i:=1 to n-1 do
28               for j:=1 to m-1 do
29  unew := 0.25 * (U[i+2,j+1] +
U[i,j+1] +  U[i+1,j+2] + U[i+1,j])
30      d:=abs(unew - U[i+1,j+1])
31        if (d>e) then
32                          p:=true
33                    fi
34        U[i+1,j+1] := unew
35                done
```

```
36              done
37          iter++
38          if iter>1 then
39              max:=0
40              for i:=0 to 10 do
41                  for j := 0 to 10 do
42  if abs(U[((n*i) div 10)+1,((m*j) div 10)+1] - V[((i*n) div 20)+1,((j*m) div 20)+1])/3>max then
43  max:=abs(U[((n*i) div 10)+1,((m*j) div 10)+1] - V[((i*n) div 20)+1,((j*m) div 20)+1])/3
44                      fi
45              done
46          done
47          for i := 0 to m do
48              for j := 0 to n do
49   V[i+1,j+1] := U[i+1,j+1]
50              done
51          done
52          h:=h/2
53          k:=k/2;
54          n:=n*2;
55          m:=m*2;
56      until (max<e)
end
```


Питання і завдання до розділу 9

1 Дослідити на апроксимацію і стійкість неявну схему із

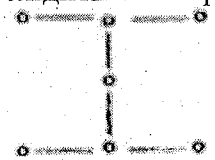


шаблоном для хвильового рівняння.

2 Розв'язати задачу $u_{tt} - a^2 u_{xx} = F(u,x,t)$, $0 \le x \le 10$, $0 \le t \le T$ ($T \ge 2$); $u(x,0) = u_0$ при $x < x_0$, $u(x,0) = 0$ при $x \ge x_0$; $u_t(x,0) = 0$; $u_x(0,t) = 0$, $u(10,t) = 0$ при $x_0 = 1, 2, 3, 4.5, 6, 7, 8, 9, 10$, $u_0 = 0.2, 0.4, 0.5, 0.6, 0.6, 0.8, 1$, $a = 0.5, 1, 1.5, 2$, використовуючи

схему «хрест» $(u_{n+1,m} - 2u_{n,m} + u_{n-1,m})/t^2 - a^2 (u_{n,m+1} - 2u_{n,m} + u_{n,m-1})/h^2 = F_{n,m}$ при $F(u,x,t) = \sin u$.

3 Дослідити на апроксимацію і стійкість неявну схему із



шаблоном для хвильового рівняння.

4 Розв'язати задачу $u_{tt} - a^2 u_{xx} = F(u,x,t)$, $0 \le x \le 10$, $0 \le t \le T$ ($T \ge 2$); $u(x,0) = u_0$ при $x < x_0$, $u(x,0) = 0$ при $x \ge x_0$; $u_t(x,0) = 0$; $u_x(0,t) = 0$, $u(10,t) = 0$ при $x_0 = 1, 2, 3, 4.5, 6, 7, 8, 9, 10$, $u_0 = 0.2, 0.4, 0.5, 0.6, 0.6, 0.8, 1$, $a = 0.5, 1, 1.5, 2$, використовуючи неявну схему $(u_{n+1,m} - 2u_{n,m} + u_{n-1,m})/t^2 - a^2 [(u_{n+1,m+1} - 2u_{n+1,m} + u_{n+1,m-1}) - (u_{n-1,m+1} - 2u_{n-1,m} + u_{n-1,m-1})]/2h^2 = F_{n,m}$ із

шаблоном при $F(u,x,t) = \sin u$.

5 Явна чотириточкова схема; шаблон, порядок апроксимації, стійкість для рівняння параболічного типу.

6 Нобудувати апроксимацію граничної умови $u_x(0,t) = g(t)$ із залученням значень сіткових функцій у точках $(0, t_n)$, (h, t_n), $(2h, t_n)$. Оцінити похибку апроксимації.

7 Дослідити за допомогою спектральної ознаки стійкість різницевої схеми при постійному коефіцієнті a: $(u_{n+1,m} - (u_{n,m+1} - u_{n,m-1})/2)/t + a(u_{n+1,m+1} - u_{n+1,m-1})/2h = 0$.

8 Дослідити за допомогою спектральної ознаки стійкість різницевої схеми при постійному коефіцієнті a: $(u_{n+1,m} - u_{n,m})/t + a(u_{n,m} - u_{n,m-1})/h = 0$.

9 Записати різницеву схему $[u_{i,k+1}+u_{i,k-1}+u_{i+1,k}+u_{i-1,k}-4u_{i,k}]/h^2=f_{i,k}$; $i, k = 1-5$; $u_{i,0}=u_{i,6}=u_{0,i}=u_{6,i}=0$, $i = 0-6$ у матричному вигляді $\mathbf{Ax} = \mathbf{b}$, використовуючи різні способи впорядкування при утворенні вектора $\mathbf{x}$ зі значень сіткової функції u^h (значення u^h на границі попередньо виключити).

10 Дослідити стійкість явної схеми для рівняння теплопровідності $\begin{cases} \dfrac{y_m^{n+1} - y_m^n}{\tau} = \dfrac{y_{m-1}^n - 2y_m^n + y_{m+1}^n}{h^2} + f_m^n \end{cases}$.

11 Дослідити стійкість неявної схеми для рівняння теплопровідності: $\begin{cases} \dfrac{y_m^{n+1} - y_m^n}{\tau} = \dfrac{y_{m-1}^{n+1} - 2y_m^{n+1} + y_{m+1}^{n+1}}{h^2} + f_m^{n+1} \end{cases}$.

12 Поняття явної й неявної різницевої схеми для рівняння теплопровідності.

13 Запропонуйте різницеву схему й алгоритм розв'язку задачі Діріхле для рівняння Пуассона в одиничному квадраті

$(0 \le x \le 1,\ 0 \le y \le 1)$ $u_{xx} + u_{yy} = f(x,y)$, $u(0,y) = X_0(y)$, $u(1,y) = X_1(y)$, $u(x,0) = \Phi_0(x)$, $u(x,1) = \Phi_1(x)$ для задачі: $f(x,y)=(5p^{2/8})\sin(px)[\sin(py/2)+\cos(py/2)]$, $X_0(y)=0$, $X_1(y)=0$, $\Phi_0(x)=[\sin(px)]^{1/2}$, $\Phi_1(x)=[\sin(px)]^{1/2}$. На основі алгоритму напишіть програму на одній з мов програмування. У програмі потрібно передбачити можливість здрібнювання розрахункової сітки. Здійсніть розрахунки при різних просторових кроках розрахункової сітки, послідовно подвоюючи їхнє число, аж до досягнення 1% точності за правилом Рунге. Результати представити у вигляді рівномірної таблиці значень сіткової функції з кроком 0.2 по координатах x і y.

14 Знайдіть похибку апроксимації різницевої схеми $u_{i+1,j+1}+u_{i-1,j+1}+u_{i+1,j-1}+u_{i-1,j-1}-2[(h_x^2-5h_y^2)/(h_x^2+h_y^2)](u_{i+1,j}+u_{i-1,j})+2[(5h_x^2-h_y^2)/(h_x^2+h_y^2)](u_{i,j+1}+u_{i,j-1})-20u_{i,j}=0$ розв'язку рівняння Лапласа при $h_x = h_y$.

15 Побудувати різницеву схему, що апроксимує на сітці (x_m,y_n), $x_m=mh$, $y_n=nh$ $(m=0,1,\ldots,M;\ n=0,1,\ldots,N)$, $h_x=1/M$, $h_y=1/N$ задачу $a(x,y)u_{xx}+b(x,y)u_{yy}+c(x,y)u_x+d(x,y)u_y+e(x,y)u=f(x,y)$, $0 \le x \le 1$, $0 \le y \le 1$; $u(x,0)=a_0(x)$, $u(x,1)=a_1(x)$, $0 \le x \le 1$; $u(0,y)=b_0(y)$, $u(1,y)=b_1(y)$, $0 \le y \le 1$, із другим порядком відносно h_x і h_y

16 Оцінити порядок апроксимації оператора Лапласа при $h_x = h_y$ за схемою $D^h u_i=[u_{i+1,j-1}+u_{i+1,j+1}+u_{i-1,j+1}+u_{i-1,j-1}-4u_{i,j}]/(2h^2)$.

17 Метод Зейделя для рівняння Пуассона.

18 Написати різницеву схему для рівняння $\Delta u = f$ з апроксимацією порядку $O(h^4)$.

19 Запропонуйте різницеву схему й алгоритм розв'язку задачі Діріхле для рівняння Пуассона в одиничному квадраті $(0 \le x \le 1,\ 0 \le y \le 1)$ $u_{xx}+u_{yy}=f(x,y)$, $u(0,y)= X_0(y)$, $u(1,y)=X_1(y)$, $u(x,0)=\Phi_0(x)$, $u(x,1) =\Phi_1(x)$ для задачі: $f(x,y)=(2p^{2/3})\cos(px)\sin(py)$,

$X_0(y)=[\sin(py)]^{0.333333}$, $X_1(y)=-[\sin(py)]^{0.333333}$, $\Phi_0(x)=0$, $\Phi_1(x)=0$. На основі алгоритму напишіть програму на одній з мов програмування. У програмі потрібно передбачити можливість здрібнювання розрахункової сітки. Виконайте розрахунки при різних просторових кроках розрахункової сітки, послідовно подвоюючи їхнє число, аж до досягнення 1% точності за правилом Рунге. Результати представити у вигляді рівномірної таблиці значень сіткової функції з кроком 0.2 по координатах x і y.

Додатки

1 Метричні простори

Оскільки розв'язками рівнянь математичної моделі можуть бути об'єкти різного походження: числа, упорядковані набори чисел, вектори, функції (при розгляді диференціальних і інтегральних рівнянь), то при обґрунтуванні методів реалізації математичних моделей різноманітних задач доцільно одержати аналоги відповідних теорем для відображень множин довільної природи. Це призводить до необхідності розгляду множин, на яких уведена відстань. Такі множини називаються метричними просторами. Дамо точні визначення.

Нехай X — непорожня множина і кожній парі елементів x, y з множини поставлено у відповідність невід'ємне число $\rho(x, y)$ з такими трьома властивостями:

1) $\rho(x, y) = 0$ тоді і тільки тоді, коли $x = y$;

2) $\rho(x, y) = \rho(y, x)$ (властивість симетрії);

3) $\rho(x, y) = \rho(x, z) + \rho(z, y)$ (нерівність трикутника).

Ці властивості виконуються для всіх x, y, z з X. ***Число $\rho(x, y)$ називають відстанню між елементами x та y. Множина X із заданою на ньому відстанню називається метричним простором.*** Розглянемо кілька прикладів метричних просторів.

Приклад 1 На множині R дійсних чисел відстань визначається, як правило,: $\rho(x, y)=|x-y|$, $x,y \in R$. На R можна розглядати й інші відстані. Наприклад, $d(x, y)=|arctgx-arctgy|$, $x, y \in R$.

Приклад 2 На множини R^n упорядкованих наборів n дійсних чисел можна розглянути такі три відстані:

$$\rho_1(x,y)=\left(\sum_{k=1}^{n}|x_k-y_k|^2.\right)^{1/2}$$ (евклідова відстань);

$$\rho_2(x,y)=\max_{1\le k\le n}|x_k-y_k|;$$

$$\rho_3(x,y)=\sum_{k=1}^{n}|x_k-y_k|,$$

де $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ — довільна пара наборів дійсних чисел.

Цей метричний простір, як правило, використовується при вивченні систем рівнянь з декількома невідомими.

Приклад 3 Нехай $f[a, b]$ — множина неперервних функцій на відрізку [a, b]. Тоді формула

$$\rho(f,\varphi) = \max_{[a,b]} |f(t) - \varphi(t)|$$

визначає відстань між функціями на $[a, b]$.

У будь-якому метричному просторі X з відстанню ρ можна визначити поняття збіжної фундаментальної послідовності.

Послідовність $\{x_n\} = (x_1, x_2, \ldots)$ елементів метричного простору X називається збіжною до елемента $x_0 \in X$, якщо для кожного $\varepsilon > 0$ можна знайти таке натуральне число N, що $\rho(x_n, x_0) < \varepsilon$ для всіх $n > N$. Елемент x_0 називається границею послідовності $\{x_n\}$, тобто $x_0 = \lim\limits_{n\to\infty} x_n$.

Послідовність $\{x_n\}$ елементів з X називається фундаментальною (чи послідовністю Коші), якщо для кожного $\varepsilon > 0$ існує таке натуральне N, що $\rho(x_n, x_m) < \varepsilon$ при всіх $n, m > N$.

Метричний простір X називається повним, якщо будь-яка фундаментальна послідовність з X збігається до деякого елемента з X.

В усіх розглянутих тут прикладах метричних просторів вони є повними, крім метричного простору R з відстанню d.

2 Абсолютна величина і норма матриці

Нерівність $A \le B$ між матрицями $A = [a_{ij}]$ й $B = [b_{ij}]$ одного типу означає, що $a_{ij} \le b_{ij}$.

У такому сенсі не всякі дві матриці можна порівняти між собою.

За *абсолютну величину (модуль)* матриці $A = [a_{ij}]$ будемо вважати матрицю

$$|A| = [|a_{ij}|],$$

де $|a_{ij}|$ – модулі елементів матриці A.

Якщо A і B – матриці, для яких операції $A+B$ і AB мають сенс, то:

а) $|A + B| \le |A| + |B|$;

б) $|AB| \le |A| \cdot |B|$;

в) $|\alpha A| = |\alpha| |A|$, (α - число).

За *норму матриці* $A = [a_{ij}]$ вважаємо дійсне число $\|A\|$, що задовольняє умови:

а) $\|A\| \ge 0$, причому $\|A\| = 0$ тоді і тільки тоді, коли $A = 0$;

б) $\|\alpha A\| = |\alpha| \|A\|$ (α - число) і, зокрема, $\|-A\| = \|A\|$;

в) $\|A + B\| \le \|A\| + \|B\|$;

г) $\|AB\| \le \|A\| \cdot \|B\|$

(A і B - матриці, для яких відповідні операції мають сенс).

Відзначимо ще одну важливу нерівність між нормами матриць A і B одного типу. Застосовуючи умову в), будемо мати

$$\|B\| = \|A + (B - A)\| \le \|A\| + \|B - A\|.$$

Звідси

$$\|A - B\| = \|B - A\| \ge \|B\| - \|A\|.$$

Аналогічно

$$\|A - B\| \ge \|A\| - \|B\|.$$

Отже,

$$\|A - B\| \ge \|A\| - \|B\|.$$

Назвемо *норму канонічною*, якщо додатково виконані умови:

д) якщо $A = [a_{ij}]$, то $|a_{ij}| \le \|A\|$,

причому для скалярної матриці $A=[a_{11}]$ маємо $\|A\| = |a_{11}|$;

е) з нерівності $|A| \le |B|$ (A і B – матриці) випливає нерівність

$$\|A\| \le \|B\|.$$

Зокрема, $\|A\| = \| \, |A| \, \|$.

Надалі для матриці $A = [a_{ij}]$ довільного типу ми будемо розглядати головним чином три канонічні норми, що легко обчислюються:

1) $\|A\|_m = \max\limits_i \sum\limits_j |a_{ij}|$ (m – норма);

2) $\|A\|_l = \max\limits_j \sum\limits_i |a_{ij}|$ (l - норма);

3) $\|A\|_k = \sqrt{\sum\limits_{i,j} |a_{i,j}|^2}$ (k - норма).

Приклад. Нехай

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}.$$

Маємо:

$\|A\|_m = \max(1+2+3,\ 4+5+6,\ 7+8+9) = \max(6,\ 15,\ 24) = 24;$

$\|A\|_l = \max(1+4+7,\ 2+5+8,\ 3+6+9) = \max(12,\ 15,\ 18) = 18;$

$\|A\|_k = \sqrt{1^2+2^2+3^2+4^2+5^2+6^2+7^2+8^2+9^2} =$

$= \sqrt{1+4+9+16+25+36+49+64+81} = \sqrt{285} \approx 16{,}9.$

Нехай маємо послідовність матриць $A_k = [a_{ij}^{(k)}](k = 1,2,...)$ одного типу $m \times n$ ($i = 1,\ 2,...,\ m;\ j = 1,\ 2,...,\ n$).

За *границю* послідовності матриць A_k вважається матриця

$$A = \lim_{k\to\infty} A_k = \left[\lim_{k\to\infty} a_{ij}^{(k)}\right]$$

Послідовність матриць, що має границю, є *збіжною.*

Лема 1 *Для збіжності послідовності матриць* A_k (k=1,2,…) *до матриці A необхідно і достатньо, щоб*

$$\|A - A_k\| \to 0 \quad \text{при } k \to \infty \quad ,$$

де $\|A\|$ *- будь-яка конічна норма матриці A. При цьому*

$$\lim_{k\to\infty} \|A_k\| = \|A\|.$$

$$\lim_{k\to\infty} \|A - A_k\| = 0.$$

Лема 2 *Для збіжності послідовності матриць* $A_k (k = 1,2,...)$ *необхідно і достатньо, щоб був виконаний узагальнений критерій Коші,* а саме: для будь-якого $\varepsilon > 0$ повинен існувати такий номер $N = N(\varepsilon)$, що при $k > N$

$\|A_{k+1} - A_k\| \le \varepsilon$, де $\| \ \|$ - будь-яка канонічна норма.

3 Псевдокод

Псевдокод нагадує існуючі мови програмування (такі, як Pascal або C) та на відміну від них є простішим і більш компактним. Крім того, у псевдокоді інколи можна записувати дії алгоритму своїми словами, якщо так ясніше.

На відміну від конкретної мови програмування псевдокод дозволяє опустити несуттєві технічні подробиці, такі як обробка помилок, що необхідно у реальній програмі але заважає вивченню алгоритмів.

Коментар у псевдокоді починається з // і йде до кінця рядка.

Програма на псевдокоді складається з процедур

```
Insertion_Sort(A):
...
end
```

Кожна процедура починається з імені, після якого у дужках ідуть імена аргументів процедури (якщо є), і двокрапка. Починаючи з наступного рядка, записується тіло процедури, яке завершується інструкцією end. Рядки у тілі процедури

нумеруються для полегшення посилань на окрему конструкцію в алгоритмі.

Змінні вважаються локальними стосовно процедури, якщо не зазначене інше.

Типи змінних не вказуються (якщо не сказано інакше, вважається, що прості змінні мають тип „дійсне число").

Присвоювання позначається операцією ':=' : x := y.

Часто використовуються масиви та об'єкти, які складаються з кількох полів, або атрибутів. Значення атрибута записується Object_Name.Attibute_Name. Наприклад, довжина масиву вважається його атрибутом і записується як A.length.

Індекс масиву записується у квадратних дужках '[' і ']'. Операція '..' виділяє частину масиву: A[i .. j] (A[i], [i+1], … , A[j]).

Змінна, що позначає масив або об'єкт, вважаться посиланням (покажчиком) на дані, що його складають. Наприклад, якщо ми маємо об'єкти x та y, то після виконання операції присвоювання y := x для будь-якого поля f виконується умова x.f = y.f.

Більше того, якщо ми далі виконаємо операцію y.f := 3, то і y.f = 3, і x.f = 3, тому що після виконання y := x обидві змінні вказують на один і той самий об'єкт.

Параметри до процедур передаються за значенням: процедура одержує власну копію параметрів і будь-яку зміну параметра всередині процедури ззовні не видно. Але масиви і об'єкти передаються через посилання, тобто процедура одержує копію покажчика, а не копію даних, на які він указує. Тобто, якщо всередині процедури виконується A[j] := 10 або y.f := 3, це помітно ззовні.

Приклад: неможливо написати процедуру, яка міняє місцями свої параметри.

```
Swap(x, y):
  t := x
  x := y
```



```
  y := t
end
a := 5
b := 7
Swap(a, b)
print a, b // 5 7
```

Але можна написати процедуру, яка міняє місцями два елементи масиву, якщо передати сам масив та індекси елементів, які необхідно поміняти місцями:

```
Swap(A, i, j):
  t := A[i]
  A[i] := A[j]
  A[j] := t
end
```

Кожна інструкція у тілі процедури записується на окремому рядку. Інколи бажано розмістити більше однієї інструкції в одному рядку, в такому разі між ними ставлять ;.

Інструкції записуються з відступами, для позначення рівня вкладеності однієї інструкції в іншу.

Перелік інструкцій:

```
if умова then
 інструкція
 інструкція
elseif умова then
 інструкція
 інструкція
else
```

інструкція

інструкція

fi

Частини elseif … та else — необов'язкові, але в будь-якому разі необхідно завершувати інструкцію командою fi.

while умова do

інструкція

інструкція

done

repeat

інструкція

інструкція

until умова

for змінна := початок to кінець do

інструкція

інструкція

done

for змінна := кінець downto початок do

інструкція

інструкція

done

Зазначимо, що немає потреби у спеціальних командах begin та end для позначення початку та кінця блоку, тому що кожна інструкція є також блоком, що охоплює інструкції всередині неї. Тому необхідно завершувати інструкції спеціальними командами на зразок fi у інструкції if then else fi.

Умови можуть складатися з порівнянь на зразок '=', '<>', '<', '<=', '>', '>=' та булевих операцій not, and, or. Порівняння мають більш високий пріоритет (тобто виконується раніше), ніж булеві операції, тому немає потреби писати дужки навколо порівнянь (на відміну від Pascal).

Розширення псевдокоду

- abs() – абсолютна величина аргумента
- return() – перериває виконання функції і повертає її значення
- sqrt() – квадратний корінь
- sqr() – квадрат числа
- factorial() – повертає факторіал аргумента
- div – цілочислове ділення
- round() – округлення до найближчого цілого
- Дійсні операції інкременту та декременту (наприклад, i++ та i--), а також операції +=, -=, *=.
- A.lengthI, A.lengthJ – кількість рядків і стовпців матриці відповідно. Але в деяких алгоритмах для цих цілей можуть використовуватися деякі змінні (n і m, наприклад) для зручності написання коду.
- Імена масивів бажано позначати великими літерами
- Алгоритми можуть використовувати функції з попередніх алгоритмів без їх попереднього опису
- Індекс масиву може починатися як з 0, так і з 1.

Список літератури

1. Ляшко И.И., Макаров В.Л., Скоробагатько А.А. Методы вычислений.–Киев: Вища школа, 1977,-406с.
2. Гаврилюк І.П., Макаров В.Л. Методи обчислень. — К.:Вища школа, 1995, ч.1, ч.2.
3. Бахвалов Н.С. Численные методы. – М.:Наука, 1973.- 632с.
4. Данилович В., Кутнів М. Чисельні методи. — Львів:Кальварія, 1998.
5. Форсайт Дж., Молер К. Численное решение систем линейных алгебраических уравнений.– М.: Мир, 1969.–168с.
6. Калиткин Н.Н. Численные методы,– М.: Наука, 1978. – 512с.
7. Волков Е.А. Численные методы,– М.: Наука, 1982. – 256с.
8. Форсайт Дж., Малькольм М., Моулер К. Машинные методы математических вычислений. – М.: Мир, 1980. –280с.
9. Копченова Н.В., Марон И.А. Вычислительная математика в примерах и задачах. – М.: Наука, 1972. – 368с.
10. Крылов В.И., Бобков В.В., Монастырный П.Н. Вычислительные методы высшей математики: В 2-х т. – Минск: Вышэйшая школа, 1972.- Т. 1. – 304с.; Т.2.-400с.
11. Березин И.С., Жидков Н.П. Методы вычислений: В 2–х т. – М., 1959. Т.1.– 464 с.;Т.2 – 602 с.
12. Марчук Г.И. Методы вычислительной математики. - М.: Наука, 1989. - 608 с.
13. Боглаев Ю.П. Вычислительная математика и программирование.- М.: Высш.школа,1990.-544с.
14. Григоренко Я.М., Панкратова Н.Д. Обчислювальні методи в задачах прикладної математики.-К.: Либідь,1995.-277с.
15. Мак-Кракен Д., Дорн У. Численные методы и программирование на Фортране.-М.:Мир,1977.-580с.
16. Самарский А.А. Теория разностных схем. — М.:Наука, 1989.
17. Самарский А.А., Гулин А.В. Численные методы. — М.:Наука, 1989.
18. Самарский А.А. Введение в численные методы. - М.: Наука, 1987. - 288 с.

Навчальне видання

Любчак Володимир Олександрович,
Назаренко Людмила Дмитрівна

МЕТОДИ ТА АЛГОРИТМИ ОБЧИСЛЕНЬ

Навчальний посібник

Дизайн обкладинки І.В. Шелехова
Редактор Н.А. Гавриленко
Комп'ютерне верстання Л.Д. Назаренко

НБ ПНУС

739143

Підп. до друку 2.07.2008.
Формат 60x84/16.Папір офс.Гарнітура Times New Roman Cyr.Друк офс.
Ум. друк. арк 18,37. Обл.-вид арк. 14,22.
Тираж 300 пр. Вид. № 277.
Зам. №1129.

Видавництво СумДУ при Сумському державному університеті
40007, Суми, вул. Римського-Корсакова,2
Свідоцтво про внесення суб'єкта видавничої справи до Державного реєстру ДК №3062 від 17.12.2007.
Надруковано у друкарні СумДУ
40007, Суми, вул. Римського-Корсакова, 2.