МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

М. І. Романюк, О. А. Батіна

ОБЧИСЛЮВАЛЬНА МАТЕМАТИКА

КОНСПЕКТ ЛЕКЦІЙ

*Рекомендовано Методичною радою КПІ ім. Ігоря Сікорського
як навчальний посібник для студентів,
які навчаються за спеціальністю 171 «*Електроніка*»
спеціалізації «Електронні та інформаційні системи і технології
телебачення, кінематографії та звукотехніки»*

Київ
КПІ ім. Ігоря Сікорського
2019

Обчислювальна математика. Конспект лекцій: [Електронний ресурс]: навч. посіб. для студ. спеціальності 171 «Електроніка», спеціалізації «Електронні та інформаційні технології кінематографії та аудіовізуальних систем»/ М.І. Романюк; О. А. Батіна; КПІ ім. Ігоря Сікорського. – Електронні текстові данні (1 файл: 3,85 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2019. –193 с.

Гриф надано Методичною радою КПІ ім. Ігоря Сікорського (протокол № 9 від 30.05.2019 р.) за поданням Вченої ради факультету електроніки (протокол № 04/2019 від 26.04.2019 р.)

Електронне мережне навчальне видання

ОБЧИСЛЮВАЛЬНА МАТЕМАТИКА

КОНСПЕКТ ЛЕКЦІЙ

Укладачі: *Романюк Маргарита Ігорівна*, канд. техн. наук.
Батіна Олена Анатоліївна, старший викладач

Відповідальний редактор *Попович П.В.*, доцент, канд. техн. наук, доцент

Рецензенти: *Дрозденко О. І.,* доцент, канд. техн. наук, доцент

Навчальний посібник містить матеріали дисципліни «Обчислювальна математика», що викладається студентам спеціальності 171 «Електроніка», спеціалізації «Електронні та інформаційні технології кінематографії та аудіовізуальних систем». Посібник написано відповідно до навчальної програми вказаної дисципліни та складається з 16 лекцій, кожна з яких розглянута у відповідності з виділеними питаннями плану. Для закріплення теоретичного матеріалу у кінці кожної лекції представлені контрольні питання.

У посібнику викладені базові поняття, що стосуються числових методів, обчислювального експерименту, розв'язку нелінійних рівнянь та систем лінійних та нелінійних алгебраїчних рівнянь, інтерполяція та апроксимація табличних залежностей та функцій, чисельного інтегрування функції.

Матеріали посібника будуть корисними студентам, магістрантам, викладачам, а також усім зацікавленим.

ЗМІСТ

РОЗДІЛ 1. ВСТУП

Тема 1.1 Загальні питання.

Тема 1.2 Математична модель та обчислювальний експеримент

ЛЕКЦІЯ 1. Мета і задачі дисципліни. Поняття обчислювального експерименту. Вимоги до обчислювальних методів

Навчальні питання:

1.1 Предмет обчислювальної математики.

1.2 Місце обчислень у сучасному світі. Стисла історична довідка.

1.3 Математичне моделювання та обчислювальний експеримент

1.4 Класифікація та вимоги до обчислювальних методів

1.1 Предмет обчислювальної математики

Математика – це необхідний засіб пізнавання людиною навколишнього світу, без математики абсолютно неможливо проводити серйозні наукові та інженерні розрахунки. Інструментом, завдяки якому можливо прискорити математичні розрахунки є обчислювальна техніка, що постійно вдосконалюються.

Сучасна обчислювальна техніка представляє потужні засоби для фактичного виконання лічильної роботи. Завдяки цьому в багатьох випадках стало можливим відмовитися від наближеного трактування прикладних питань і перейти до вирішення завдань в точній постановці. Це передбачає використання більш глибоких спеціальних розділів математики (нелінійні диференціальні рівняння, функціональний аналіз, теоретико-імовірнісні методи і ін.). Розумне використання сучасної обчислювальної техніки неможливо без вмілого застосування *методів наближеного і чисельного аналізу*. Цим і пояснюється надзвичайно зрослий інтерес *до методів обчислювальної математики*. При виконанні масових обчислень важливо дотримуватися певних простих правил, дотримання яких економить працю обчислювача та дозволяє раціонально використовувати наявну обчислювальну техніку та допоміжні засоби. Перш за все обчислювач повинен розробити детальну обчислювальну схему, що точно вказує порядок дій і дає

можливість отримати шуканий результат найбільш простим і швидким шляхом. Це особливо необхідно при однотипних обчисленнях, так як така схема, автоматизуючи обчислення, дозволяє виконувати їх більш швидко і надійно, що з користю окупає час, витрачений на складання схеми. Крім того, маючи детальну обчислювальну схему для вирішення завдання, можна використовувати працю менш кваліфікованих обчислювачів.

Обчислювальна математика (ОМ) – прикладна галузь математики, в якій розробляються *методи чисельного розв'язання* типових математичних задач. У обчислювальної математики можна виділити наступні розділи (перші чотири розділи будуть розглянуті у курсі):

1) чисельні методи лінійної алгебри;

2) аналітичне наближення табличних функцій;

3) чисельне диференціювання та інтегрування;

4) чисельні методи рішення диференціальних рівнянь;

5) чисельні методи рішення рівнянь математичної фізики;

6) чисельні методи дослідження операцій і теорії ігор;

7) методи математичного програмування.

Чисельне рішення означає отримання *числового результату*, а не формули, теореми або алгоритму. У цьому перша особливість обчислювальної математики, її відмінність від математики фундаментальної. Друга особливість полягає в тому, що всі сучасні методи обчислювальної математики орієнтовані на *застосування комп'ютерів*.

Протягом всієї історії обчислювальної математики її розвиток в основному визначався потребами практики (природознавство, техніка, економіка, тощо), на сучасному ж етапі абсолютно всі методи обчислень розробляють для комп'ютерного застосування. Багато чисельних методів запрограмовані, до цих програм можна звернутися як до процедур або функцій в різних прикладних математичних пакетах програм. Для практичних цілей використовують такі сучасні системи автоматизованого проектування як Mathcad, MATLAB, Mathematica, Wolfram Alpha та ін..

Предметом обчислювальної математики є методи чисельного рішення задач, орієнтованих на комп'ютерну реалізацію, а *метою курсу* є вивчення цих чисельних методів, починаючи з найдавніших до сучасних.

1.2 Місце обчислень у сучасному світі. Стисла історична довідка.

Методи обчислень розроблялися фактично з самого зародження математики. Наприклад, ще з античності відомі методи Герона наближеного обчислення квадратних і кубічних коренів.

Математичні моделі для опису досліджуваних явищ у механіці, фізиці та інших точних науках природознавства використовувалися здавна. Три-чотири тисячі років тому вирішували завдання прикладної математики, пов'язані з обчисленням площ і об'ємів, розрахунками найпростіших механізмів, тобто з нескладними завданнями арифметики, алгебри і геометрії. Обчислювальними засобами служили власні пальці, а потім – рахівниці. Перше застосування обчислювальних методів належить стародавнім єгиптянам, які вміли обчислювати діагональ квадрата за кінцеву кількість дій. Вони також могли знаходити квадратний корінь з 2, за допомогою алгоритму, що в подальшому отримав назву формули Герона, а ще пізніше – методу Ньютона: $u_{k+1} = \frac{1}{2}\left(u_k + \frac{a}{u_k} \right)$, $a = u_0$, де a – додатне число, корінь з якого шукають, u_k – наближенні значення до кореня, k – номер наближення. Більшість обчислень виконувалось точно, без округлень.

Половина типового семестрового або річного курсу обчислювальної математики присвячена вивченню основних методів, запропонованих ще в XVII – XVIII століттях. Розробкою чисельних методів рішення прикладних задач займалися найбільші вчені свого часу: *Ньютон, Эйлер, Лобачевский, Коши, Лагранж, Лежандр, Лаплас, Пуанкаре, Гаусс, Эрмит, Чебишев* та багато інших відомих математиків. У XVII столітті Ньютон повністю описав закономірності руху планет навколо Сонця, вирішував *завдання геодезії*, проводив розрахунки *механічних конструкцій*. Такі завдання зводилися до звичайних диференціальних рівнянь, або до алгебраїчних систем з великим числом невідомих, обчислення

проводилися з досить високою точністю до 8 значущих цифр. При обчисленнях використовувалися таблиці елементарних функцій, арифмометр, логарифмічна лінійка. До кінця цього періоду з'явилися непогані клавішні машини з електромотором. В цей час були розроблені *алгоритми чисельних методів*, які до сих пір займають почесне місце в арсеналі обчислювальної математики. Так Ньютон запропонував ефективний чисельний метод рішення алгебраїчних рівнянь, а Ейлер – чисельний метод розв'язання звичайних диференціальних рівнянь.

Класичним прикладом *застосування чисельних методів є* відкриття планети Нептун. Уран планета, наступна за Сатурном, вважалася самою далекою планетою до 40-х років XIX ст. Проте точні спостереження показали, що Уран ледь помітно відхиляється від того шляху, по якому він повинен слідувати з урахуванням збурень з боку усіх відомих на той час планет. Левер'є (у Франції) і Адамс (в Англії) висловили припущення, що, якщо збурення з боку відомих планет не пояснюють відхилення в русі Урана, значить, на нього діє притяжіння ще невідомого тіла. Для розрахунку траєкторії Нептуна Левер'є знадобилося півроку. Суттєву роль у розвитку ОМ як самостійної науки зіграли такі вчені, як Карл Рунге (німецький фізик та математик 1864 р.н.), та Олексій Крилов (російський математик, механік та кораблебудувальник 1863 р.н.).

У воєнні часи виникає потреба у розробці новітніх засобів озброєння та захисту. Використовувались різні обчислювальні методи для розрахунку польоту бойових головок балістичних ракет, та іншої військової техніки.

ОМ отримала значний імпульс в 1950-і роки, що було пов'язано з розвитком ядерної фізики, механіки польоту, аеродинаміки космічних апаратів (космічна гонка). Надалі вирішувалися завдання, пов'язані не тільки з розрахунками дії ядерного вибуху і обтіканням боєголовок стратегічних ракет (друга світова війна закінчилась). Чисельні методи знайшли своє застосування в таких областях як динаміка атмосфери, фізика плазми, механіка гірських порід і льодовиків, синергетика, біомеханіка, теорія оптимізації, математична економіка та ін.

Найбільш наукомісткі та вимагають максимальних обчислювальних ресурсів завдання фізики, механіки та електродинаміки суцільних середовищ. До них

відносяться системи рівнянь в часткових похідних Ейлера, Лагранжа, Максвелла та ін., кінетичної теорії газів, а також завдання багатовимірної оптимізації.

Отже, як окремий розділ математичної науки обчислювальна математика сформувалася задовго до винаходу комп'ютерів. І методи обчислень, створені до комп'ютерної ери, природно, до сих пір використовуються і становлять «класику» обчислювальної математики. Але застосовуючи комп'ютер, можна вирішувати завдання для більшої кількості змінних, параметрів і великих розмірів. При комп'ютерному вирішенні обчислювального завдання легше знайти і виправити помилку, збільшити точність обчислень. Обчислення тепер відбуваються швидко. Але, тим не менше, однією з проблем чисельних методів є зменшення кількості обчислювальних операцій і часу розрахунку. Незважаючи на наявність і використання прикладних комп'ютерних засобів обчислень, необхідно знати, *як* побудовані методи вирішення завдань. Інакше неминуче неправильне використання і трактування отриманих автоматизованими методами результатів. Тому знання теоретичних основ і методів обчислювальної математики необхідно кожному інженеру та вченому.

У наші дні, жоден великий технічний проект не обходиться без різних розрахунків і обчислень, починаючи з дуже простих алгебраїчних моделей, закінчуючи найскладнішими науковими розрахунками, створенням алгоритмів, методів вирішення і т. ін. У будь-якому випадку, необхідно знати ступінь точності отриманого результату. Класична математика, як відомо, в основному націлена на вивчення явищ, що мають лінійний характер. Зміна причини призводить до пропорційної зміни слідства, тобто класичні рівняння розглядають не градієнтні середовища (вони вивчають малі відхилення маятника, дрібні хвилі і диференціал і т. д.).

1.3 Математичне моделювання та обчислювальний експеримент

Після Другої Світової Війни наука впритул наблизилась до вивчення явищ, що не є лінійними, де причина і наслідок не співмірні, саме завдяки таким явищам виникли: електронні лампи, транзистори, лазери, з'явилися високоточні прилади

здатні обирати потрібний сигнал. В більшості випадків такі явища дуже погано піддаються традиційним методам аналізу. Рівняння, що описують такі ситуації в багатьох випадках є звичайними диференціальними рівняннями, які можна вивчати і досліджувати за допомогою комп'ютера – електронної обчислювальної машини (ЕОМ).

Створення ЕОМ дало новий поштовх розвитку математики, виникло поняття «математичне моделювання». Слово «модель» походить від латинського modus (копія, образ). Моделювання – це заміщення деякого об'єкта А (оригіналу) іншим об'єктом Б (моделлю).

Предметною областю обчислювальної математики є математичне моделювання та обчислювальний експеримент.

Математична модель – це наближений опис будь-якого класу явищ зовнішнього світу, виражений за допомогою математичних понять.

Математичне моделювання – процес побудови і вивчення математичних моделей реальних процесів і явищ, тобто метод дослідження об'єктів і процесів реального світу за допомогою їх наближених описів на мові математики.

Класичним засобом вивчення математичних моделей і досліджень на їх основі властивостей реальних об'єктів є аналітичні методи, що дозволяють отримувати точні рішення у вигляді математичних формул. Ці методи дають найбільш повну інформацію про рішення задачі, і вони до теперішнього часу не втратили свого значення. Проте, клас задач, для якого вони можуть використовуватися, вельми обмежений. Тому рішення, як правило, здійснюється чисельними методами.

Чисельний метод – така інтерпретація математичної моделі («дискретна модель»), яка доступна для реалізації на ЕОМ. Чисельні методи – це методи, що зводять вирішення завдань до арифметичних і логічних дій над числами. Результат реалізації чисельного методу – число або таблиця чисел. Рішення, отримане чисельним методом, як правило, є *наближеним*.

У сучасних наукових дослідженнях значення обчислювальних методів особливо зросло завдяки широкому застосуванню *обчислювальних експериментів*.

У науці завжди був та залишається важливим експеримент. Але експерименти з реальними об’єктами можуть бути дуже складні, дорогі, небезпечні, можуть вимагати тривалого спостереження і очікування результату. Тому часто реальний об’єкт або процес замінюють математичною моделлю.

Обчислювальний експеримент (ОЕ) – метод дослідження складних проблем, заснований на побудові математичних моделей для досліджуваних об’єктів і аналізі цих моделей за допомогою ЕОМ.

В останні роки ряд Нобелівських премій по хімії, медицині, економіці, фізиці елементарних частинок були присуджені роботам, методологічну основу яких становило саме математичне моделювання.

ОЕ в порівнянні з натурним експериментом має ряд *переваг*:

- економічність, так як не витрачаються ресурси реальної системи;

- можливість моделювання гіпотетичних, тобто не реалізованих в природі об’єктів;

- доступність тестування режимів, небезпечних і важких у виконанні в реальності (критичний режим ядерного реактора, техногенні катастрофи);

- можливість зміни масштабу часу;

- велика прогностична сила внаслідок можливості виявлення загальних закономірностей.

Отже, основою обчислювального експерименту є математичне моделювання, теоретичною базою – прикладна математика, а технічною базою – потужні електронно-обчислювальні машини Сутність застосування чисельних методів розглянемо на схемі обчислювального експерименту (рис.1.1).

Основу ОЕ становить тріада, виражена академіком А.А. Самарським як *«модель – метод (алгоритм) – програма».* На основі цієї тріади зображений цикл розв’язання прикладних задач на рис. 1.1.

1. Аналіз (постановка завдання – технічне завдання) – найважливіша частина – 50% успіху реалізації проекту залежить від даного блоку (люди, які беруть участь в аналізі повинні бути фахівцями в даній області – фізик, хімік, біолог).

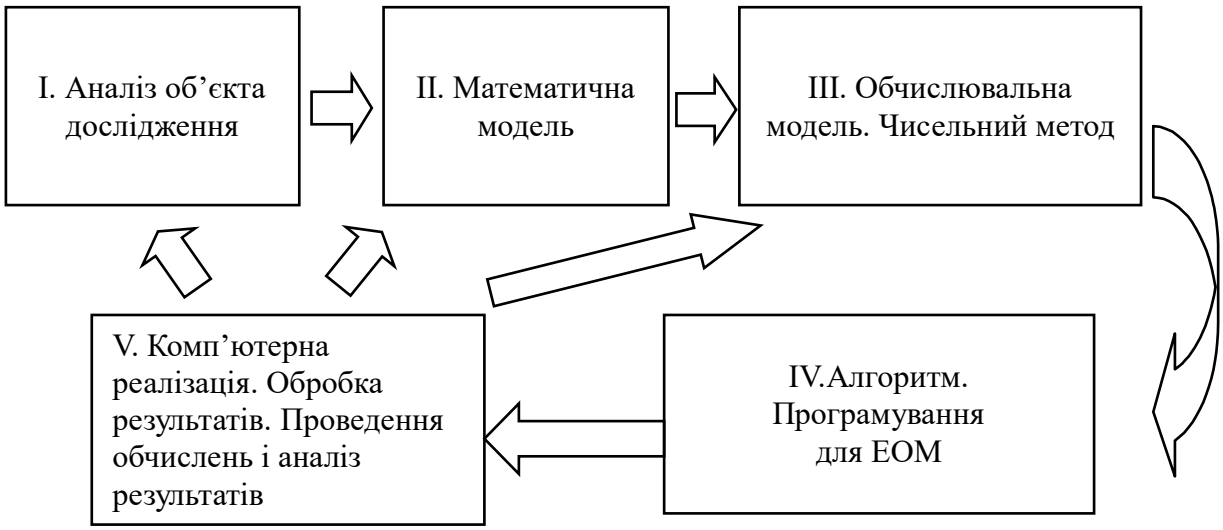


Рисунок 1.1 – Схема обчислювального експерименту

2. Математична модель. Вибирають або будують модель досліджуваного об'єкту, яка в математичній формі відображає його найважливіші властивості. Зазвичай математичні моделі реальних процесів досить складні і включають в себе системи нелінійних функціонально-диференціальних рівнянь. Ядром математичної моделі, як правило, є рівняння з частинними похідними. Для отримання попередніх знань про об'єкт побудована модель досліджується традиційними аналітичними засобами прикладної математики.

Для дослідження моделі, завдання переписують на зрозумілій для машини мові. Математична модель повинна якомога точніше описувати властивості фізичного явища, яке в загальному випадку нескінченно складне, що призводить до громіздкості формалізованого представлення моделі. Тому при побудові математичної моделі враховуються лише найважливіші (для даного завдання) сторони явища, що істотно спрощує вигляд моделі, однак призводить до появи *похибки невідповідності математичної моделі* досліджуваному фізичному явищу. Похибка невідповідності математичної моделі *є неусувною* і позначається на кінцевому результаті незалежно від способу вирішення завдання. На даному етапі виникає ***δ_1 - похибка математичної моделі.***

3. Обчислювальна модель (вибір обчислювального алгоритму). Вибирають або розробляють *обчислювальний алгоритм* для реалізації побудованої моделі на

комп’ютері, який не повинен спотворювати основні властивості моделі, повинен бути адаптованим до особливостей вирішуваних завдань і використовуваних обчислювальних засобів. Далі проводиться вивчення побудованої математичної моделі методами обчислювальної математики.

Відомо, що одну і ту ж задачу можна вирішити різними методами. Наприклад, для вирішення системи лінійних алгебраїчних рівнянь можна використовувати правило Крамера, метод Гауса або так звані ітераційні алгоритми, які реалізуються за допомогою багаторазового повторення однотипних обчислювальних операцій уточнюючих грубе апріорі задане початкове наближення рішення. Точне аналітичне рішення вдається визначити досить рідко, зазвичай в тих випадках, коли математична модель представлена в досить простому вигляді. Для більш складних моделей використовуються чисельні методи, які дозволяють отримати лише *наближене рішення* при цьому похибки виникають через те, що чисельним методом вирішується не вихідна задача, а деяка інша, близька у тій чи іншій мірі до вихідної. Таке наближення зумовлює *методичну похибку*, або **δ_2 - похибку числового методу**

4. Алгоритм. Програма. Створюють програмне забезпечення для реалізації моделі і алгоритму на комп’ютері. Створюваний програмний продукт повинен враховувати найважливішу специфіку математичного моделювання, пов’язану необхідністю використання набору математичних моделей і багатоваріантністю розрахунків. Для реалізації чисельного методу необхідно розробити програму на одній з мов програмування або застосувати готовий пакет прикладних програм, які дозволяють вирішувати більшість практичних завдань. Само по собі програмування алгоритму з використанням мов високого рівня не представляє особливих труднощів, проте досить трудомістким є процес налагодження програми на ЕОМ – усунення всіляких помилок і програмних збоїв з метою доведення програми до робочого стану (складні програми можуть бути налагоджені окремими блоками).

Після налагодження програми проводять розрахунки на ЕОМ – здійснюють введення початкових числових даних і знаходять кінцевий результат, який при

необхідності може бути представлений в зручній для користувача формі (таблиці, графіки).

Необхідно зазначити: задача, що розв'язується може характеризуватися параметрами будь-якої фізичного природи (температура, тиск, механічне переміщення і т.д), отже вихідні дані, що вводяться в пам'ять ЕОМ: повинні бути перетворені відповідними датчиками в електричні величини. Датчики – це прилади певного класу точності і перетворення завжди виконуються з деякою похибкою. Така похибка називається *похибкою завдання початкових даних*.

5. Обробка результатів (аналіз отриманих результатів) Наявність перерахованих вище похибок призводить до того, що отриманий результат може виявитися досить далеким від справжнього і його аналіз потребує уточнення рішення. Виникає **δ_3 - похибка обчислень.** До похибок обчислень належать:

1) Похибки пов'язані з наявністю нескінченних процесів. Наприклад деякі функції задаються у вигляді нескінченних рядів. Коли виникає потреба у обмежені нескінченної послідовності то виникає зменшення точності;
2) Похибки, пов'язані з наявністю у розрахункових формулах числових параметрів , які можуть бути визначені лише наближено (наприклад константи);
3) Похибки пов'язані зі системою числення, наприклад подання числа 1/3 у десятковій системі числення – 0,33(3);
4) Похибки пов'язані з виконанням дій над наближеними числами (це переносить похибку на результат).

Оцінювання похибки, яке виконується перед обчисленням, називається апріорним, а після – апостеріорним. Проведення оцінювання **δ_2** – апріорного або апостеріорного – одна з основних задач обчислювальної математики.

Існує два основні шляхи підвищення точності. Один з них – *уточнення математичної моделі*, що вимагає додаткового вивчення фізичної проблеми і призводить до суттєвого ускладнення постановки задачі, а тому видається менш доцільним, ніж другий – *вибір* іншого, *більш точного алгоритму розв'язання*

задачі. Іноді вдаються до розрахунків на ЕОМ з подвійною точністю, що зменшує обчислювальну похибку.

Після проведення першого циклу обчислювального експерименту може виникнути необхідність в уточненні моделі (новий цикл на рис.1.1). На другому циклі враховуються додаткові ефекти і зв'язки в досліджуваному явищі, або виникає необхідність знехтувати деякими закономірностями і зв'язками. Потім цей процес повторюють до тих пір, поки не переконуються, що модель адекватна об'єкту, що вивчається. У процесі роботи алгоритму похибки округлення зазвичай накопичуються, і в результаті рішення, отримане на ЕОМ, буде відрізнятися від точного рішення дискретизованої задачі. Результуюча похибка називається похибкою округлення (іноді її називають *обчислювальною похибкою*) вона має тенденцію *накопичення* зі збільшенням кількості обчислювальних операцій, тому чим простіше алгоритм тим в меншій мірі буде позначатися обчислювальна похибка на кінцевому результаті. Величина цієї похибки визначається двома факторами: точністю подання дійсних чисел у ЕОМ і чутливістю даного алгоритму до похибок округлення.

Розглянемо *приклад*, що проілюструє описані вище види похибок на прикладі задачі описання руху маятника (рис. 1.2), у якому потрібно передбачити кут відхилення маятника від вертикалі θ, починаючого рух у момент часу $t = t_0$. Рух маятника може бути описаний диференціальним рівнянням другого порядку:

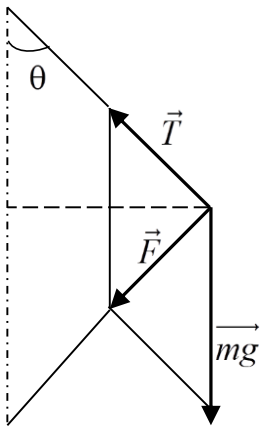


Рисунок 1.2 – Рух маятника. До пояснення виникнення похибок обчислювального експерименту

$$l\frac{d^2\theta}{dt^2} + g\,sin\theta + \mu\frac{d\theta}{dt} = 0$$

де l – довжина маятника, g – прискорення вільного падіння; μ – коефіцієнт тертя.

Похибки можуть бути викликані через наступні фактори:

- реальна сила тертя залежить від швидкості руху маятника по нелінійному закону;

- значення величин l, g, μ, t_0, $\theta(t_0)$, $\theta'(t_0)$ відомі з деякими похибками;
- для розв'язку рівняння, що описує рух маятника, яке не має аналітичного рішення, потрібно використовувати чисельний метод, у наслідок чого виникає похибка методу;
- обчислювальна похибка, виникає через кінцевість точності представлення чисел на комп'ютері.

Отже, слід розрізняти ***похибки моделі, методу і обчислювальну***. Яка ж з цих трьох похибок є переважаючою? Типовою є ситуація, що виникає при вирішенні задач математичної фізики, коли похибка моделі значно перевищує похибку методу, а похибкою округлення у випадку стійких алгоритмів можна знехтувати в порівнянні з похибкою методу. З іншого боку, при вирішенні, наприклад, систем звичайних диференціальних рівнянь можливе застосування настільки точних методів, що їх похибка буде порівнянна з похибкою округлення. У загальному випадку потрібно прагнути, щоб всі зазначені похибки мали один і той же порядок. *Похибка рішення всієї задачі повинна бути порівнянна з похибкою вихідних даних.*

Методи математичного моделювання та обчислювальний експеримент поєднують у собі переваги традиційних теоретичних і експериментальних методів дослідження. В рамках чинного курсу основним є виклад питань, що відображають лише один з етапів обчислювального експерименту, а саме етап побудови і дослідження чисельного методу (обчислювальний алгоритм).

Зазвичай складні обчислювальні задачі, що виникають при дослідженні фізичних і технічних проблем, розбиваються на ряд елементарних. Багато елементарних задач є нескладними, вони добре вивчені, для них вже розроблені методи чисельного рішення і є стандартні програми вирішення їх на ЕОМ.

1.4 Класифікація та вимоги до обчислювальних методів

Методи чисельного розв'язку задач поділяють на наступні класи:

- метод еквівалентних перетворень;
- методи апроксимації;
- прямі методи;

- ітераційні методи.

У методі *еквівалентних перетворень* вихідну задачу заміняють еквівалентною, що має такий самий розв'язок. Це потрібно у двох випадках, коли остання задача простіша або для її розв'язку існує готовий ефективний метод розв'язку.

У *методах апроксимації* вихідне завдання заміняють іншим, рішення якого є близьким до рішення вихідного в деякому сенсі. При цьому велике значення має оцінка похибки апроксимації, тобто різниця між рішенням вихідної та апроксимованої задач.

Прямі методи дозволяють отримати теоретично точне рішення за кінцеву кількість кроків. Ці методи складають досить невелику групу.

У *ітераційних методах* будують послідовність наближень (ітерацій) до точного розв'язку – ітераційну послідовність. Кожна наступна ітерація отримується застосуванням до однієї чи декількох попередніх однотипного набору операцій – *ітераційний крок*. Для запуску ітераційного процесу задають початкове наближення.

Виділяють дві групи вимог до чисельних методів. Перша група пов'язана з *адекватністю дискретної моделі вихідної математичної задачі*, друга – *з реалізованістю чисельного методу на наявній обчислювальній техніці*. До першої групи належать такі вимоги, як збіжність чисельного методу, виконання дискретних аналогів законів збереження, якісно правильна поведінка рішення дискретної задачі.

Числовий метод називається ***збіжним,*** якщо його результат наближається до точного розв'язку задачі за наближення параметрів методу до граничних значень.

Наприклад збіжність ітераційного процесу. Як сказано вище, ітераційний процес полягає в тому, що для вирішення деякої задачі будується послідовні наближення x_1, x_2, x_3, ..., x_n, ... Кажуть, що ця послідовність збігається, якщо існує $\lim\limits_{n\to\infty} x_n = a$.

Збіжність методу дискретизації. Припустимо, що дискретна модель математичної задачі являє собою систему великого числа алгебраїчних рівнянь.

Зазвичай, чим точніше ми хочемо отримати рішення, тим більше рівнянь доводиться брати. У цьому випадку кажуть, що чисельний метод збігається, якщо при необмеженому збільшенні числа рівнянь, рішення дискретної задачі прямує до вирішення початкової задачі при прямуванні до нуля параметра дискретизації.

Виконання дискретних аналогів законів збереження. Оскільки реальний комп'ютер може оперувати лише кінцевою кількістю рівнянь. Так при дискретизації задач математичної фізики приходять до різницевих схем, які представляють собою системи лінійних або нелінійних алгебраїчних рівнянь. Диференціальні рівняння математичної фізики є наслідками інтегральних законів збереження. Тому природно вимагати, щоб для різницевої схеми виконувалися аналоги таких законів збереження. Різницеві схеми, що задовольняють цим вимогам, називаються *консервативними*. Виявилося, що при одному і тому ж числі рівнянь в дискретній задачі консервативні різницеві схеми більш правильно відображають поведінку рішення вихідної різницевої задачі, ніж неконсервативні схеми.

Збіжність чисельного методу тісно пов'язана з *коректністю обчислювального алгоритму.*

Обчислювальний алгоритм – це точне розподілення дій над вихідними даними, що задає обчислювальний процес, спрямований на перетворення довільних вихідних даних в повністю визначений цими даними результат. Обчислювальний алгоритм називається *коректним*, якщо:

1. Він дозволяє, після виконання кінцевого числа елементарних для обчислювальної машини операцій, перетворити будь-яке вихідне допустиме дане $x \in X$ у єдиний результат $y \in Y$ (X – безліч допустимих вихідних даних, Y – безліч рішень).

2. Отримане за його допомогою рішення є стійким по відношенню до малих збурень вихідних даних.

3. Результат володіє обчислювальної стійкістю.

Задача називається коректно поставленою, якщо для будь-яких значень початкових даних з деякого класу її рішення існує, єдино і стійко за вихідними

даними. Отже, в поняття коректності чисельного методу включаються властивості однозначної розв'язності відповідної системи рівнянь і її стійкості.

Стійкість за вхідними даними означає те, що результат безперервним чином залежить від вихідних даних за умови, що відсутня обчислювальна похибка. Обчислювальна стійкість означає прямування до нуля обчислювальної похибки при прямуванні до нуля машинної похибки.

Алгоритм називається обчислювально стійким, якщо такий результат його роботи при будь-яких допустимих вихідних даних. *Алгоритм називається стійким*, якщо в процесі його роботи обчислювальні похибки зростають незначно, і нестійким – в протилежному випадку. Задача називається стійкою по параметру *x*, якщо її рішення *y* безперервно від нього залежить, тобто мале збільшення вихідної величини *x* викликає малий приріст *y***.**

Друга група вимог, що пред'являють до чисельних методів, пов'язана з можливістю *реалізації* даної дискретної моделі на даному комп'ютері, тобто з можливістю отримати чисельний розв'язок за прийнятний час. Основною перешкодою для реалізації коректно поставленого алгоритму є обмеження об'єму оперативної пам'яті ЕОМ та обмежений час відліку.

Головною метою вивчення дисципліни «обчислювальна математика» є знайомство з методологією побудови і дослідження основних чисельних методів алгебри і математичного аналізу і проблемами, що виникають при чисельному рішенні задач.

Будемо вивчати наступні теми і розділи:

- Особливості математичних розрахунків, що реалізуються на ЕОМ;
- Теоретичні основи чисельних методів, похибки обчислень,
- Чисельні методи лінійної алгебри;
- Рішення нелінійних рівнянь і систем;
- Чисельне інтегрування і диференціювання;
- Методи наближення функції;
- Методі розв'язку диференціальних рівнянь;
- Методи розв'язку інтегральних рівнянь.

Питання для самоперевірки:

1. Що є предметом обчислювальної математики?
2. У чому відмінність обчислювальної математики від фундаментальної?
3. У чому полягають особливості сучасних чисельних методів?
4. Дайте визначення та поясніть необхідність застосування чисельних методів.
5. Зобразіть схему та дайте визначення обчислювального експерименту.
6. Що таке обчислювальний експеримент? У чому його переваги перед натурним?
7. Надайте перелік вимог до числових методів.
8. Дайте визначення збіжності, коректності та стійкості числового методу.
9. Перерахуйте етапи розв’язання прикладної обчислювальної задачі, опишіть коротко кожен з них. Якими з них займається обчислювальна математика?

ЛЕКЦІЯ 2. Похибки обчислення (абсолютна, відносна похибки). Похибка визначення значення функції

Навчальні питання:

2.1 Джерела виникнення похибок чисельного рішення задачі та їх класифікація

2.2 Абсолютна та відносна похибки

2.3 Значущі вірні цифри у десятковому записі числа. Правила округлення та запису числа

2.4 Особливості машинної арифметики

2.5 Похибка визначення функції однієї та декількох змінних

2.1 Джерела виникнення похибок чисельного рішення задачі та їх класифікація

Як було зазначено у лекції 1, процес дослідження вихідного об’єкта методом математичного моделювання та обчислювального експерименту неминуче носить *наближений характер*, тому що на кожному етапі обчислювального експерименту вносяться ті чи інші похибки. Джерелами виникнення похибок є наступні фактори:

- неточність математичного опису, зокрема, неточність завдання початкових даних (будь-яка навіть дуже складна модель є лише наближенням реальності, тому у ній самій закладена похибка);
- неточність (наближеність) чисельного методу розв’язку задачі;
- похибки вихідних даних;
- кінцева точність машинної арифметики.

Похибки результатів обчислень можна поділити на *неусувні* (зумовлені причинами 1, 3) та *усувні* (причини 2, 4). Якщо розглядати усувні похибки, то мова не йде про повне їх обнуління. Такі похибки можна зменшити, у чому і полягає одне з головних завдань обчислювальної математики. Наприклад, похибки зумовлені причиною 4, усуваються (зменшуються) програмними та апаратними засобами, тобто застосуванням більш довершеної обчислювальної техніки та

організацією більш раціонального рахунку. З іншого боку, неусувність похибки не означає, що їх неможливо зменшити. Далі від загальних характеристик перейдемо до строгих визначень.

2.2 Абсолютна та відносна похибки

Позначимо точні значення числових скалярних величин латинськими літерами A, B і т. д., зазвичай, вони невідомі. Замість них при розрахунках використовують обчислені або виміряні *наближені значення*, які позначимо a, b і т.д.. Тепер визначимо числові величини, що характеризують ступінь близькості наближених і точних значень. Розглянемо тільки скалярні величини. Мірою точності подання величини A числом a є *похибка.*

Похибкою наближеного значення a називається величина

$$\varepsilon a = A - a, \tag{2.1}$$

тобто різниця між точним (істинним) і наближеним числами. Похибка незручна для обчислювальних операцій і для оцінки, оскільки може бути як додатною, так і від'ємною. Тому визначимо іншу, завжди додатну характеристику. *Абсолютною похибкою* a називається різниця між точним та наближеним значенням:

$$\Delta a = |\varepsilon a| = |A - a|. \tag{2.2}$$

Оскільки невідомо істинне значення величини, обчислити абсолютну похибку неможливо. З цієї причини в наближених обчисленнях завжди оперують з верхніми оцінками абсолютних похибок. Верхні оцінки похибок або граничні похибки цілком піддаються обчисленням. *Гранична абсолютна похибка* Δa – число, модуль якого є рівним або меншим за модуль абсолютної похибки:

$$\left|\overline{\Delta a}\right| \geq |\Delta a|. \tag{2.3}$$

Верхня оцінка абсолютної похибки дає величину відхилення наближеного значення від невідомого точного числа. За однією лише абсолютною похибкою важко судити про точність наближеного значення, так як вона залежить від величин A та a, і абсолютна похибка має ту ж розмірність, що і оцінювана.

Наприклад, при вимірюванні довжини олівця і відстані від Землі до Марсу отримані значення абсолютних похибок, що дорівнюють 1 см. Очевидно, що *точність вимірювання* у даному випадку більша для відстані між планетами. Тому потрібна безрозмірна характеристика, яка не залежить від масштабу. Такою є відносна похибка наближеного значення a.

Відносною похибкою наближеного значення величини A числом a називають модуль відношення абсолютної похибки до істинного значення цієї величини:

$$\delta = \left| \frac{\Delta a}{A} \right|. \tag{2.4}$$

Оскільки замість абсолютної похибки розглядають граничну абсолютну похибку, то відносну похибку замінюють *граничною відносною похибкою*:

$$\overline{\delta a} = \left| \frac{\overline{\Delta a}}{A} \right|. \tag{2.5}$$

Зазвичай відносну похибку записують у відсотках. Формули для обчислень треба намагатися перетворювати до такого виду, щоб в них не було віднімання близьких величин, тому що це може привести до великої втрати точності і до великих відносних похибок.

2.3 Значущі вірні цифри у десятковому записі числа. Правила округлення та запису числа

Зрозуміло, що наближене число не може бути нескінченним дробом, тому всі наближені числа записуються у вигляді кінцевих десяткових дробів з порядком. Тобто у вигляді:

$$\underbrace{\alpha_n \alpha_{n-1} ... \alpha_1 \alpha_0 , \beta_{-1} ... \beta_{-m}}_{M} \cdot 10^p, \tag{2.6}$$

де $M = \alpha_n \alpha_{n-1} ... \alpha_1 \alpha_0 , \beta_{-1} ... \beta_{-m}$ – мантиса. *Мантиса* – це число фіксованої довжини, яке представляє старші розряди дійсного числа. Її значення розраховують за формулою представлення числа у десятковій позиційній системі числення:

$$M = \alpha_n \cdot 10^n + \alpha_{n-1} \cdot 10^{n-1} + ... + \alpha_1 \cdot 10 + \alpha_0 + \beta_{-1} \cdot 10^{-1} + ... + \beta_{-m} \cdot 10^{-m}. \tag{2.7}$$

де α_i, β_{-j} – десяткові цифри, (α_i - цифри цілої частини числа, β – цифри дробової частини). Ціле число p – порядок числа.

При такому записі розрізняють значущі та незначущі цифри.

Значущими цифрами в десятковому записі числа називаються всі цифри мантиси, починаючи з першої ненульової зліва. Інші цифри (нулі) називають *незначущими.*

Наприклад, у наступних наближених числах значущі цифри підкреслено: $a = 0{,}00\underline{98800}; b = 00\underline{120}; c = -0{,}00\underline{730} \cdot 10^6$.

Нулі зліва можна прибрати без втрати значень, при необхідності змінивши порядок, тому вони і називаються незначущими. У прикладах вище числа можна записати без незначущих нулів так: $a = 9{,}8800 \cdot 10^{-3}; b = 120; c = -7{,}30 \cdot 10^3$.

Зауваження 1. Значущі нулі справа прибирати не можна, оскільки вони означають розряди числа, наприклад: 0,28 і 0,280 – різні наближені числа, так як перше дано з двома знаками після коми, а друге – з трьома.

Зауваження 2. Особливий випадок являє собою *нульова величина*. Як би не записували нуль, в записі єдиною значущою цифрою вважають останній нуль справа: 0,000$\underline{0}$, 0,$\underline{0}\cdot 10^{-2}$, 00$\underline{0}$.

При обчисленнях для отримання наближених значень деякі значущі цифри числа відкидаються. Така операція називається ***округленням*.** Є два правила округлення: ***відсіканням і по доповненню***.

При *округленні відсіканням* відкидаються всі значущі цифри правіше тієї, до якої здійснюється округлення. У разі потреби змінюється порядок числа. *Наприклад:* число 72,01396, округлене до другого знаку після коми (розряд 10^{-2}), дорівнює 72,01, до третього (розряд 10^{-3}) – 72,013.

Округлення відсіканням є округленням в меншу сторону для додатних чисел і в більшу для від'ємних чисел. Абсолютна похибка при цьому (величина відкинутої частини) не перевищує одиниці розряду, до якого виконувалось округлення. Більш точним, а отже, найбільш прийнятним на практиці є *округлення по доповненню*, коли також відкидаються всі значущі цифри правіше тієї, до якої

здійснюється округлення, при необхідності змінюється порядок. *Правила, за якими здійснюється округлення по доповненню* наступні:

1. якщо першою зліва відкидається цифра менше п’яти, то цифри, що залишаються, не змінюються, як і при відсіканні;

2. якщо ця цифра більше п’яти або вона дорівнює п’яти та серед інших цифр, що відкидаються є ненульові, то остання цифра, що зберігається, збільшується на одиницю;

3. якщо ця цифра дорівнює п’яти і всі інші цифри, що відкидаються є нулями, то остання цифра, що зберігається не змінюється, якщо вона парна, і збільшується на одиницю, якщо вона непарна.

Наприклад, число 0,03261, округлене до сотих, дорівнює 0,03, до тисячних – 0,033. Число 99500, округлене до тисяч, дорівнює $100 \cdot 10^3$.

Округлення по доповненню забезпечує величину абсолютної похибки, що не перевищує половини одиниці розряду, в якому перебуває остання цифра що залишається; округлення йде в найближчу сторону. Надалі будемо вважати округлення по доповненню таким, що застосовується за замовчуванням.

Значущі цифри за допомогою абсолютних похибок діляться на *вірні і сумнівні*. Нехай дано представлення наближеного числа (2.7), в якому всі цифри значущі.

Вірними в широкому сенсі називаються ті значущі цифри, для яких абсолютна похибка не перевищує одиниці розряду, в якому вони знаходяться. Значущі цифри, для яких ця умова не виконується, називаються ***сумнівними в широкому сенсі***.

Очевидно, що це визначення пов’язано з округленням відсіканням. З визначень вірної цифри і абсолютної похибки слідує, що вірні цифри з точністю до абсолютної похибки збігаються з відповідними цифрами точного значення, тому *тільки вони несуть інформацію про нього*. Сумнівні цифри є некорисними, «зайвими», у записі числа при даній абсолютній похибці. Тепер визначимо вірність і сумнівність у вузькому сенсі.

Значуща цифра називається ***вірною в вузькому сенсі***, якщо абсолютна похибка не перевищує половини одиниці розряду, в якому вона знаходиться. В іншому випадку вона називається сумнівною у вузькому сенсі. Тут також очевидно, що це визначення пов'язано з округленням по доповненню, оскільки, при такому округленні вся решта цифр вірні у вузькому сенсі за умови врахування тільки похибки округлення.

Кожне додатне десяткове число може бути представлено у вигляді нескінченного десяткового дробу

$$a = a_m \cdot 10^m + a_{m-1} \cdot 10^{m-1} + ... + a_{m-n+1} \cdot 10^{m-n+1} + ... \ ,$$

де a_i – цифра числа a в i – му разряді, m – старший десятковий розряд числа.

Наприклад $3{,}66 = 3 \cdot 10^0 + 6 \cdot 10^{-1} + 6 \cdot 10^{-2}$.

Враховуючи запис десяткового числа у десятковій позиційній системі, можна записати *правила визначення кількості вірних цифр (знаків)*.

Правило 1: якщо для наближеного числа a відомо (абсолютна похибка менше половини останнього десяткового розряду):

$$\Delta a = |A - a| \le \frac{1}{2} \cdot 10^{m-n+1}, \qquad (2.8)$$

то перші n цифр цього числа є вірними. m - старший десятковий розряд.

Приклад 1: Знайти кількість вірних цифр та записати округлений результат числа a, якщо дано $a = 36{,}00,\ \Delta a = 0{,}02$.

Розв'язок: користуємось правилом 1 та записуємо:

$\Delta a = 0{,}02 \le 0{,}05 = \frac{1}{2} \cdot 10^{-1} = \frac{1}{2} \cdot 10^{m-n+1}$, тобто $m - n + 1 = -1$, у даному прикладі – $m = 1$ (старший десятковий розряд), звідси $n = 3$. Отже, наближене число a має 3 вірні цифри, та його слід округлити наступним чином: $a = 36{,}0$.

Зв'язок *відносної похибки* з кількістю вірних знаків числа.

Правило 2. Якщо додатне наближене число a має відносну похибку δ, то кількість вірних знаків n даного числа можна визначити як:

$$\tilde{n} = 1 - lg(a_m \delta), \qquad (2.9)$$

та в якості n взяти найближче ціле до $\tilde{n}$ число.

Приклад 2. Округлити сумнівні цифри наближеного числа x з відносною похибкою δ, залишивши в його записи тільки вірні цифри, якщо дано: $x = 43{,}221,\ \delta = 0{,}5\%$.

Розв'язок: Знайдемо кількість вірних цифр числа x:

$\tilde{n} = 1 - lg\left(4 \times 0{,}005\right) = 2{,}699$. Звідси $n = 3$, отже округлюємо x до трьох цифр $x = 43{,}2$.

З викладеного вище зрозуміло, що тільки вірні цифри мають цінність, як ті що несуть інформацію про обчислювальні величини. Точність відповіді визначається не кількістю значущих цифр, а кількістю вірних цифр, саме їх треба залишати в проміжних і остаточних результатах. Однак округляти тільки до вірних цифр все-таки недоцільно з наступних причин:

1. оцінки похибок, за якими визначаються вірні цифри, завищені, деякі сумнівні цифри можуть виявитися вірними і тому можна втратити вірні цифри через завищення похибок;

2. похибки округлень можуть привести до того, що остання вірна цифра стане сумнівною, це означає, що бажано мати хоча б одну сумнівну цифру в запасі.

Тому *при наближених обчисленнях керуються такими правилами*:

1. В проміжних результатах залишають, крім вірних, одну-дві сумнівні цифри;

2. Остаточний результат округлюють зі збереженням не більше однієї сумнівної цифри. Для таких округлень необхідно вміти оцінювати похибки при арифметичних операціях і обчисленнях функцій. Цим питанням присвячені наступні пункти.

Зауваження: при поданні числа необхідно записувати його наближене значення зі значенням похибки записуючи при цьому однакову кількість знаків після коми, наприклад:

$$\begin{cases} a = 2{,}258 \pm 0{,}003 \\ a = 2{,}258 \pm 3 \cdot 10^{-3} \\ 2{,}258 - 0{,}003 \le a \le 2{,}258 + 0{,}003 \end{cases}$$

Похибка округлення безпосередньо пов'язана з представленням машинного числа. Тому слід розбиратися в особливостях машинної арифметики.

2.4 Особливості машинної арифметики

Основна відмінність обчислювальної математики полягає в тому, що при вирішенні обчислювальних задач людина *оперує машинними числами, які є дискретною проекцією дійсних* чисел на конкретну архітектуру комп'ютера, тому важливу роль у обчислювальній математиці грають оцінки точності алгоритмів і їх стійкість до представлення машинних чисел у комп'ютері. Знання основних особливостей машинної арифметики необхідно для грамотного використання ЕОМ при чисельному рішенні задач.

Будемо вважати, що всі обчислювальні машини працюють у двійковій системі числення. Для зберігання числа в пам'яті ЕОМ відводиться поле стандартної довжини (*машинне слово*), в якому число записується у вигляді послідовності двійкових цифр (0 або 1).

Ціле число n подається у вигляді:

$$n = \pm\left(a_l 2^l + ... + a_1 2^1 + a_0 2^0\right), \tag{2.10}$$

де l – деяке стандартне для ЕОМ число, a_i – двійкові числа (0 або 1).

Максимальним по модулю цілим числом, що представляють у ЕОМ, є

$$n_{max} = 2^l + ... + 2^1 + 2^0 = 2^{l+1} - 1. \tag{2.11}$$

Операції додавання, віднімання і множення над цілими числами реалізовані в ЕОМ так, що якщо результат не перевищує по модулю число n_{max}, то він виходить точним. Однак, якщо модуль результату перевищує n_{max}, то на багатьох обчислювальних машинах ця ситуація не доводиться до відома користувача, відбувається присвоєння результату деякого значення, меншого n_{max} по модулю, і обчислення тривають далі.

У сучасних комп'ютерах реалізований стандарт двійкової арифметики IEEE 754. В IEEE арифметиці дійсні числа діляться на два класи: *числа з одинарною точністю,* для запису яких використовуються 32 біта (4 байти) і числа *з подвійною точністю*, для запису яких використовуються 64 біта (8 байтів). Для *дійсних чисел* прийнята форма подання з плаваючою крапкою (комою)

$$x = \pm\left(\gamma_1 2^{-1} + \gamma_2 2^{-2} + ... + \gamma_t 2^{-t}\right)2^p, \tag{2.12}$$

де γ_1, γ_2,… γ_t – двійкові цифри (0, 1). Число *x* нормується так, щоб $\gamma_1 = 1$ і тому у ЕОМ зберігаються лише значущі цифри. Число $M = \gamma_1 2^{-1} + \gamma_2 2^{-2} + ... + \gamma_t 2^{-t}$ – мантиса числа *x* у двійковій позиційній системі числення.

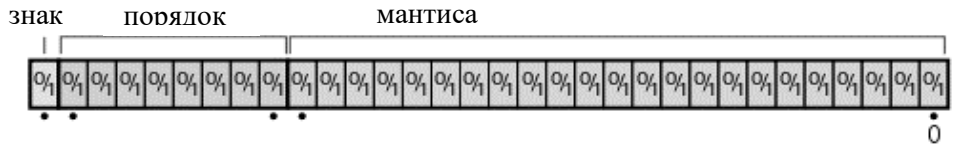
Кількість *t* цифр, яка відводиться для запису мантиси, називається *розрядністю* мантиси, *p* – ціле число, що називається *двійковим порядком*. Порядок також записується як двійкове ціле число, для зберігання якого у ЕОМ відводиться *l* + 2 двійкових розрядів. Так як нуль – ненормоване число, то для його зберігання передбачають особливий спосіб запису.

Представлення чисел з плаваючою комою дещо відрізняється від класичного представлення.

Під ***знак*** відводиться один біт: 0 – додатне, 1 – від'ємне число. *Мантиса* нормалізованих двійкових чисел задовольняє умові: $1 \le M \le 2$, тобто мантиса завжди містить одиницю у цілій частині. І так як ціла частина усіх нормалізованих чисел дорівнює одиниці, то ця одиниця не зберігається, а зберігається лише дробова частина мантиси. Число отримується додаванням одиниці до дробової частини. Зекономлений розряд використовують для збереження ще одного двійкового розряду.

Порядок може бути як додатнім так і від'ємним. Для того щоб не вводити біт знаку порядку, використовують зміщений порядок, додаючи до порядку зміщення, що дорівнює $2^{l-1} - 1$, де *l* – число розрядів, що відведено під зміщений порядок.

Порядок і мантиса – цілі числа, які разом зі знаком дають *представлення числа з плаваючою комою* у наступному вигляді:

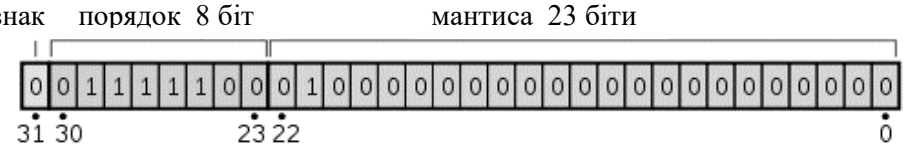


Математично це записується так:

$$(-1)^s \times M \times B^E,$$

де s – знак, B – основа, E – порядок, а M – мантиса. В числах одинарної точності (float/single) порядок складається з 8 біт, а мантиса – з 23. Ефективний порядок записується із зсувом і визначається як E-127 ($2^{8-1} - 1$).

Наприклад, число 0,140625 буде записано в пам’ять як



Знак s=0 – додатне число; $E=01111100_2-127_{10} = -3$, порядок дорівнює -3. Мантиса $M = 1.001_2$ (перша одиниця не явна), $1.001_2 = 1\cdot 2^0 + 0\cdot 2^{-1} + 0\cdot 2^{-2} + 1\cdot 2^{-3} = 1\cdot 2^0 + 1\cdot 2^{-3}$. Враховуючи порядок, зсуваємо кому на три позиції вліво та отримаємо:

$$1.001\,e-3 = 1\cdot 2^{-3} + 1\cdot 2^{-6} = 0{,}125 + 0{,}015625 = 0{,}140625$$

Потрібно зазначити, що у ЕОМ представлені не всі числа, а лише кінцевий набір раціональних чисел спеціального виду. Ці числа утворюють *представлену безліч обчислювальної машини.* Для всіх інших чисел можливо лише їх наближене представлення з похибкою (помилкою), яку прийнято називати *похибкою представлення* (або похибкою округлення). Зазвичай приблизне подання числа x в ЕОМ позначають як $\tilde{x} = fl(x)$.

Якщо округлення здійснюють по доповненню, то межа *відносної похибки подання* дорівнює одиниці першого відкинутого розряду мантиси тобто: $\overline{\delta}(\tilde{x}) = \varepsilon_M = 2^{-t}$. Якщо ж округлення здійснюють відсіканням, то $\overline{\delta}(\tilde{x}) = \varepsilon_M = 2^{1-t}$

Величина ε_M грає в обчисленнях на ЕОМ фундаментальну роль, її називають ***відносною точністю ЕОМ***, а також машинною точністю (або машинний епсилон). Значення цієї величини визначається розрядністю мантиси і способом округлення. Точне число x та відповідне йому округлене число $\tilde{x}$ пов’язані рівністю

$$\tilde{x} = x(1+\varepsilon_M)$$

Діапазон зміни чисел в ЕОМ обмежений. Для всіх поданих на ЕОМ чисел x (за винятком нуля) маємо

$$0 < X_0 \le |x| < X_\infty,$$

Всі числа по модулю більші X_∞ розглядаються, як *машинна нескінченність.* Спроба отримати таке число призводить до аварійного останову (авосту) ЕОМ по

переповненню. Такі числа визначають як: *OFL – (Over Flow Limit)* *поріг переповнення* максимальне число, яке може бути записано в арифметику. $X_{\infty} = 2^{p_{max}}$, $p_{max} = 2^{l+1} - 1$. У системі з одинарною точністю $OFL = 2^{256-127}(1+1) \approx 10^{38}$

Всі числа по модулю менші X_0 для ЕОМ не помітні і представляються, як нуль (*машинний нуль*). Отримання числа x, такого, що $x < X_0$, називають *зникненням порядку*. Вказані числа визначають як: *UFL (Under Flow Limit) – поріг машинного нуля* мінімальне число, яке може бути записано в арифметику: $X_0 = 2^{-(p_{max}+1)}$. У системі з одинарною точністю $UFL = 2^{-127} \approx 10^{-38}$.

Зазвичай, при зникненні порядку, автоматично покладається *fl (x) = 0* і обчислення тривають. Якщо наближене число $\tilde{x}$ наводиться в якості результату без вказівки величини похибки, то прийнято вважати, що всі його значущі цифри є вірними. Користувач часто довіряє виведеним з ЕОМ цифрам. Однак, число може бути виведено з такою кількістю значущих цифр, скільки програміст задає завданням відповідного формату. Подання числової інформації в комп'ютері, як правило, тягне за собою появу похибок, величина яких залежить від форми подання числа і від довжини розрядної сітки комп'ютера.

Для подання числа в комп'ютері також визначаються абсолютна і відносна похибки. Формули їх розрахунку аналогічні до (2.2) та (2.4), відповідно абсолютна та відносні похибки числа. У цьому випадку за наближене значення a приймають машинне представлення числа.

Основним джерелом похибки є обмежена розрядна сітка. При обчисленнях часто виникає ситуація, коли отриманий результат, а саме, його дробова частина, має більше розрядів, ніж є у розрядній сітці. У цьому випадку доводиться округляти результат до потрібного числа розрядів. Так, якщо розрядна сітка має довжину n розрядів, то максимальне значення абсолютної похибки дорівнюватиме 2^{-n}, а мінімальне значення дорівнює 0. Часто при оцінці підсумкової похибки використовують усереднену абсолютну похибку:

$$\Delta_{ср} = \frac{0 + 2^{-n}}{2} = 0{,}5 \cdot 2^{-n}. \qquad (2.13)$$

Абсолютне значення подання дробового числа в формі *з фіксованою комою* знаходиться в діапазоні від 2^{-n} до $1-2^{-n}$. Відносна похибка подання для максимального значення числа дорівнює

$$\delta A_{max} = \frac{\Delta_{ср}}{|A_{max}|} = \frac{0{,}5 \cdot 2^{-n}}{1-2^{-n}}. \qquad (2.14)$$

де A_{max} – максимальне значення подання дробового числа з фіксованою комою. Зазвичай довжина розрядної сітки $n = 16 \div 64$, тоді $2^{-n} \ll 1$ та $\delta A_{max} \approx 0{,}5 \cdot 2^{-n}$. Відносна похибка подання для мінімального значення числа дорівнює

$$\delta A_{min} = \frac{\Delta_{ср}}{|A_{min}|} = \frac{0{,}5 \cdot 2^{-n}}{2^{-n}} = 0{,}5, \qquad (2.14)$$

де A_{min} – мінімальне значення подання дробового числа з фіксованою комою. Видно, що похибки подання малих чисел у формі з фіксованою комою можуть бути дуже значними, тобто сумірними з самими числами.

Для знаходження відносної похибки подання числа в формі *з плаваючою комою* необхідно похибку мантиси помножити на величину порядку числа p:

$$\delta A_{min} = \frac{0{,}5 \cdot 2^{-n} \cdot p}{2^{-1} \cdot p} = 2^{-n}, \qquad (2.15)$$

$$\delta A_{max} = \frac{0{,}5 \cdot 2^{-n} \cdot p}{\left(1-2^{-n}\right) \cdot p} \approx 0{,}5 \cdot 2^{-n}. \qquad (2.16)$$

З (2.15), (2.16) слідує, що відносна похибка подання чисел у формі з плаваючою комою майже не залежить від величини числа. Тому всі математичні обчислення над дробовими числами проводять, коли ці числа представлені в формі з плаваючою комою. Обчислення над цілими числами можна проводити, коли ці числа представлені в формі з фіксованою комою. Але слід контролювати, щоб отриманий результат не перевищив порогові значення, які визначаються довжиною розрядної сітки.

2.5 Похибка обчислення функції однієї та декількох змінних

Досить часто на практиці необхідно визначити як впливає точність знаходження незалежного аргументу x на точність знаходження значення функції

$f(x)$, або навпаки якщо відомо з якою точністю має бути знайдено значення функції $f(x)$ та необхідно визначити якою має бути точність незалежної змінної x Для цієї мети може бути застосована наступна теорема:

Теорема: Гранична абсолютна похибка функції $f(x)$ дорівнює добутку модулів похідної цієї функції та граничної абсолютної похибки змінної x.

Доведення: Нехай x - наближене значення величини X, визначене з абсолютною похибкою Δx, тобто: $X = x + \Delta x$. Абсолютна похибка функції:

$$\left|\Delta f(x)\right| = \left|f(X) - f(x)\right| = \left|f(x + \Delta x) - f(x)\right|$$

Величина Δx має прагнути до нуля, через це його можна замінити диференціалом $f(x + \Delta x) \approx f(x) + \frac{df(x)}{dx}\Delta x$, звідси

$$\left|\Delta f(x)\right| = \left|\frac{df(x)}{dx}\Delta x\right| = \left|f'(x)\right| \cdot \left|\Delta x\right|$$

Позначимо граничну похибку аргументу як α ($\left|\Delta x\right| \le \alpha$), а граничну абсолютну похибку функції позначимо як β ($\left|\Delta f(x)\right| \le \beta$), тоді остаточно можна записати:

$$\beta = \alpha \cdot \left|f'(x)\right|, \tag{2.17}$$

теорему доведено.

Тепер потрібно з'ясувати залежності між відносними похибками функцій та аргументу. Позначимо відносну граничну похибку аргументу як $\overline{\delta_x} = \frac{\alpha}{|x|}$ та відносну граничну похибку функції як $\overline{\delta_f}$, тоді:

$$\overline{\delta_f} = \frac{\beta}{\left|f(x)\right|} = \frac{\alpha\left|f'(x)\right|}{\left|f(x)\right|} = \left|x\frac{f'(x)}{f(x)}\right|\overline{\delta_x}. \tag{2.18}$$

Можливі наступні часткові випадки:

1. Відносна гранична похибка степеневої функції $f(x) = x^n$, через відоме значення граничної відносної похибки аргументу $\overline{\delta_x}$

$$\overline{\delta_f} = \left|x\frac{f'(x)}{f(x)}\right|\overline{\delta_x} = \left|x\frac{nx^{n-1}}{x^n}\right|\overline{\delta_x} = |n|\overline{\delta_x} \tag{2.19}$$

2. Абсолютна гранична похибка логарифмічної функції $f(x)=ln(x)$ за відомою граничною відносною похибкою аргументу $\overline{\delta_x}$.

$$\beta=\alpha\cdot\left|f'(x)\right|=\frac{\alpha}{|x|}=\overline{\delta_x} \tag{2.20}$$

Розглянемо далі функцію ***двох змінних***: $Z=f(x,y)$.

Для кожного з аргументів функції існують абсолютні похибки:

$$X=x+\Delta x,\ Y=y+\Delta y$$

При цьому абсолютна похибка буде дорівнювати:

$$|\Delta Z|=\left|f(X,Y)-f(x,y)\right|=\left|f(x+\Delta x,y+\Delta y)-f(x,y)\right|.$$

Так як і для функції одного аргументу вважаємо величини Δx та Δy малими величинами та замінюємо їх диференціалами dx та dy, тоді прирощення ΔZ - заміняємо повним диференціалом dZ:

$$\Delta Z\approx dZ=\frac{\partial Z}{\partial x}\Delta x+\frac{\partial Z}{\partial y}\Delta y,$$

звідки абсолютне значення прирощення функції становитиме:

$$|\Delta Z|\le dZ=\left|\frac{\partial Z}{\partial x}\right||\Delta x|+\left|\frac{\partial Z}{\partial y}\right||\Delta y|.$$

За аналогією для одномірного випадку позначимо граничні абсолютні похибки величини x та y через величини α_x та α_y ($|\Delta x|\le\alpha_x$, $|\Delta y|\le\alpha_y$), тоді отримаємо:

$$|\Delta Z|\le\left|\frac{\partial Z}{\partial x}\right|\alpha_x+\left|\frac{\partial Z}{\partial y}\right|\alpha_y.$$

Гранична абсолютна похибка функції двох змінних дорівнює:

$$\beta=\left|\frac{\partial Z}{\partial x}\right|\alpha_x+\left|\frac{\partial Z}{\partial y}\right|\alpha_y. \tag{2.21}$$

Рівність (2.21) є справедливою для знаходження граничної абсолютної похибки функції будь-якої кількості аргументів. Користуючись правилами (2.18) та (2.21) можливо довести наступні правила:

1. Для функції $Z=x\pm y\pm\ldots\pm t$. Гранична абсолютна похибка суми та різниці дорівнює сумі граничних абсолютних похибок доданків:

$$\overline{\Delta Z} = \beta = \alpha_x + \alpha_y + ... + \alpha_t \qquad (2.22)$$

Значення відносної похибки суми доданків одного знака перебуває в межах від найменшої граничної відносної похибки доданків до найбільшого:

$$\overline{\delta_{min}} \le \overline{\delta_Z} \le \overline{\delta_{max}} \qquad (2.23)$$

де $\overline{\delta_Z} = \beta / Z$.

2. Для функції $Z = x \cdot y \cdot t$. Гранична відносна похибка добутку дорівнює сумі граничних відносних похибок співмножників:

$$\overline{\delta_Z} = \frac{\beta}{xyt} = \frac{yt\alpha_x + xt\alpha_y + xy\alpha_t}{xyt} = \frac{\alpha_x}{x} + \frac{\alpha_y}{y} + \frac{\alpha_t}{t} = \overline{\delta_x} + \overline{\delta_y} + \overline{\delta_t} \qquad (2.24)$$

3. Для функції $Z = \frac{x}{y}$. Гранична відносна похибка частки дорівнює сумі граничних відносних похибок діленого та дільника:

$$\overline{\delta_Z} = \frac{\beta}{xyt} = \frac{\frac{\alpha_x}{y} + \frac{x\alpha_y}{y^2}}{\frac{x}{y}} = \frac{\alpha_x y}{x} + \frac{x\alpha_y y}{xy^2} = \overline{\delta_x} + \overline{\delta_y} \qquad (2.25)$$

При обчисленні значень функції абсолютна похибка може істотно залежати від того яким чином записана розрахункова формула та яка послідовність операцій у цій формулі. Обчислювальні вирази потрібно перетворювати таким чином, щоб у них були відсутнє віднімання близьких за величиною значень, що може привести до втрати точності та до великих значень відносної похибки.

*Приклад 3:*Знайти значення функції $z(x,y)$ та оцінити абсолютну та відносну похибку результату, якщо задані наступні наближені значення $\overline{x} = 0{,}452 \pm 0{,}0005$, $\overline{y} = 2{,}156 \pm 0{,}0003$ для функції $z = \sqrt{\frac{y^3}{x}}$.

Розв'язок: для оцінки абсолютної граничної похибки β функції $z(x,y)$ скористуємось виразом (2.21)

$$\beta = \left|\frac{\partial Z}{\partial x}\right| \alpha_x + \left|\frac{\partial Z}{\partial y}\right| \alpha_y,$$

З форми запису умови слідує, що $\alpha_x = 0{,}0005$, $\alpha_y = 0{,}0003$, знаходимо наближене значення $\overline{z}$:

$$\bar{z} = \sqrt{\frac{2{,}156^3}{0{,}452}} = 11{,}79037$$

Розраховуємо часткові похідні:

$$\frac{\partial Z}{\partial x} = \frac{1}{2\sqrt{\frac{y^3}{x}}} \cdot \frac{x - y^3}{x^2} = \frac{\sqrt{x}\left(x - y^3\right)}{2\sqrt{y^3} \cdot x^2} \Rightarrow \left|\frac{\partial Z}{\partial x}\right| = \left|\frac{\sqrt{0{,}452}\left(0{,}452 - 2{,}156^3\right)}{2\sqrt{2{,}156^3} \cdot 0{,}452^2}\right| = 2{,}2482;$$

$$\frac{\partial Z}{\partial y} = \frac{1}{2\sqrt{\frac{y^3}{x}}} \cdot \frac{3y^2 x}{x^2} = \frac{3\sqrt{x}y^2}{2\sqrt{y^3 x}} \Rightarrow \left|\frac{\partial Z}{\partial y}\right| = \frac{3\sqrt{0{,}452} \cdot 2{,}156^2}{2\sqrt{2{,}156^3} \cdot 0{,}452} = 3{,}2760.$$

Визначаємо абсолютну граничну похибку функції $z(x,y)$:

$$\beta = 2{,}2482 \cdot 0{,}0005 + 3{,}2760 \cdot 0{,}0003 = 0{,}00211 \le 0{,}005,$$

звідси можемо округлити наближене значення $\bar{z}$, враховуючи тільки вірні значущі цифри: $\bar{z} = 11{,}79 \pm 0{,}002$

Отже відносна похибка функції $z(x,y)$:

$$\delta_z = \frac{\beta}{|\bar{z}|} = \frac{0{,}00211}{11{,}79037} \approx 0{,}00018 \approx 0{,}018\%.$$

Питання для самоперевірки:

1. Які основні джерела похибок чисельного рішення задачі?
2. Як класифікують похибки за походженням?
3. Що означають усувні та неусувні похибки?
4. Дайте визначення абсолютній та відносній похибкам обчислень. Що значить гранична абсолютна та відносна похибки? Чим викликана необхідність їх визначення?
5. Що таке значущі і незначущі цифри у записі наближеного числа? Що таке вірні і сумнівні значущі цифри? Наведіть приклад.
6. Сформулюйте правила округлення. Як вони пов’язані з вірними цифрами?
7. Наведіть правила за якими можна визначити вірні цифри числа без перевірки кожної цифри?
8. Як можна оцінити граничну похибку по вірним цифрам?

9. За якими правилами записуються результати наближених розрахунків?

10. Яким чином дійсні числа представлені у ЕОМ? Яка різниця між числами поданими з одинарною та подвійною точністю?

11. Що називають відносною точність ЕОМ?

12. Що таке абсолютна і відносна похибки подання числа в комп'ютері?

13. Як оцінюється похибка подання числа з фіксованою та плаваючою комою?

14. Як оцінюється похибка функції? При яких умовах ця оцінка є вірною?

15. Сформулюйте та доведіть теорему про граничну абсолютну похибку визначення значення функції.

16. Чому дорівнює відносна гранична похибка степеневої функції та гранична абсолютна похибка логарифма?

17. Виведіть вираз для оцінки похибки граничної абсолютної похибки функції суми та різниці двох змінних.

18. Чому дорівнює відносна гранична похибка функцій $Z = \frac{x}{y}$ та $Z = x \cdot y \cdot t$?

РОЗДІЛ 2 ЧИСЛОВІ МЕТОДИ ЛІНІЙНОЇ АЛГЕБРИ

Тема 2.1 Матриці. Методи і похибки розв’язання СЛАР

ЛЕКЦІЯ 3. Задача чисельного рішення лінійних систем. Поняття норми вектора та норм матриць. Число обумовленості

Навчальні питання:

3.1 Постановка задачі чисельного рішення систем лінійних рівнянь

3.2 Числові характеристики наближеного рішення. Норми векторів та норми матриць.

3.3 Число обумовленості системи. Властивості числа обумовленості.

До чисельного розв’язку систем лінійних алгебраїчних рівнянь (СЛАР) зводяться багато прикладних задач. Математичні моделі, що представляють собою СЛАР великої розмірності, зустрічаються в математичній фізиці, математичній економіці, біології, тощо. Теорія отримання наближених рішень СЛАР – частина обчислювальної лінійної алгебри.

Лінійна алгебра – це розділ алгебри, в якому досліджуються лінійні об’єкти: векторні (або лінійні) простори, лінійні відображення, системи лінійних алгебраїчних рівнянь. Основний математичний апарат, який використовується у лінійній алгебрі – це матриці. Відповідно, чисельні методи лінійної алгебри – це методи чисельного рішення матричних задач. При чисельному рішенні диференціальних і інтегральних рівнянь і в ряді інших випадків алгоритми побудови вирішення багатьох завдань призводять до обчислювальних завдань лінійної алгебри. Тому дуже важливо вміти добре вирішувати обчислювальні задачі лінійної алгебри, до яких відносять:

- задачі вирішення систем лінійних алгебраїчних рівнянь (СЛАР);
- обчислення обернених матриць A^{-1}, обчислення визначників A;
- завдання обчислення власних чисел і власних векторів матриць.

У практичних додатках найбільш часто доводиться мати справу з першою з перелічених задач. З приводу інших потрібно відзначити лише, що рішення другої

задачі, в кінцевому рахунку, зводиться до першої; остання ж пов'язана з проблемою пошуку коренів алгебраїчного полінома n-го ступеня (n – порядок матриці).

Труднощі вирішення зазначених задач, як правило, пов'язані з великою розмірністю матриць. Варто зауважити, що сучасні суперкомп'ютери витрачають приблизно 80% свого робочого часу саме на чисельний розв'язок СЛАР, що ще раз підкреслює важливість даної теми.

3.1 Постановка задачі чисельного рішення систем лінійних рівнянь

Нехай задано систему лінійних алгебраїчних рівнянь:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = f_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = f_2 \\ \quad \ldots \quad \ldots \quad \ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = f_n \end{cases}, \tag{3.1}$$

Перепишемо СЛАР у матричному вигляді:

$$A\bar{x} = \bar{f}\,, \tag{3.2}$$

де $A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$ – матриця системи, $\begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix}$ – вектор невідомих, $\begin{pmatrix} f_1 \\ f_2 \\ \ldots \\ f_n \end{pmatrix}$ – вектор правої частини.

За умови невиродженої матриці ($\det \neq 0$) система має єдине рішення $\bar{x}$. Задача полягає в тому, щоб обчислити наближене рішення $\bar{x}^*$. Існують точні та наближенні методи розв'язку системи (3.1). Наприклад з курсу лінійної алгебри відомо точний метод рішення СЛАР, що називається правило Крамера, яке дає однозначний кінцевий алгоритм за умови, що матриця невироджена, ($\det \neq 0$). Цей алгоритм досить зручний для теоретичних досліджень, однак, на практиці СЛАР вирішуються абсолютно іншими методами.

Нехай задача характеризується числом N – розмірністю простору, в якому присутні елементи x і f. У сучасних задачах N становить кілька мільйонів. Значить, для того, щоб порахувати всі компоненти вектора потрібно порахувати N

визначників матриці. Для того, щоб розрахувати визначник матриці розміром $N \times N$, потрібно N^3 операцій. Якщо $N=10^6$, то $N^3 = 10^{18}$ та загалом потрібно буде 10^{24} арифметичних операцій. Жоден суперкомп'ютер не вирішить таку задачу за розумний час. З іншого боку, для систем 2-х, 3-х рівнянь правило Крамера не викликає ніяких труднощів. Однак існують набагато більш економічні методи розв'язування СЛАР великої розмірності, у яких характерний час вирішення становить від декількох хвилин до декількох годин. Вивчення таких методів є нашою подальшою метою.

3.2 Числові характеристики наближеного рішення. Норми векторів та норми матриць

Як і для одномірного випадку, чисельне рішення задачі СЛАР має на меті не тільки знаходження наближеного рішення, а й оцінку похибки. Якість наближеного рішення $\bar{x}^*$ визначається властивостями різниці векторів $\bar{x} - \bar{x}^*$. Цей вектор будемо називати похибкою $\varepsilon\bar{x}^*$ рішення $\bar{x}^*$.

$$\varepsilon\bar{x}^* = \bar{x} - \bar{x}^*. \tag{3.3}$$

Ще одною характеристикою наближеного рішення є *нев'язка* $r\bar{x}^*$:

$$r\bar{x}^* = \bar{f} - A\bar{x}^* \tag{3.4}$$

Зв'язок між нев'язкою та похибкою наступний:

$$r\bar{x}^* = \bar{f} - A\bar{x}^* = A\bar{x} - A\bar{x}^* = A\left(\bar{x} - \bar{x}^*\right) = A\varepsilon\bar{x}^*$$

Далі визначимо скалярну величину, що характеризує ступінь близькості векторів, тобто похибку. Як описано вище для одновимірних числових величин абсолютна похибка визначається як модуль. У багатовимірному випадку *аналогами модуля є норми* векторів і матриць.

При вирішенні багатьох практичних завдань необхідно якось "вимірювати" матриці, щоб говорити, що одна матриця більша за іншу. Правило, за яким матриці (зокрема, матриці-стовпці) ставиться у відповідність деяке невід'ємне число, яке має сенс міри, і визначає поняття *норма матриці.*

3.2.1 Норма вектору

Введемо деякі додаткові поняття, що мають математичну цінність.

Визначення 1. Нормою вектора (матриці-стовпчика) $\bar{x} = \begin{pmatrix} x_1 \\ \dots \\ x_n \end{pmatrix}$, називається функція, яка позначається як $\|\bar{x}\|$ та задовольняє наступним аксіомам:

1. $\|\bar{x}\| \geq 0$ для будь-якого стовпця *x*, причому $\|\bar{x}\| = 0$, тоді і лише тоді, коли $\bar{x}$ –нульовий вектор-стовпчик;

2. $\|\alpha\bar{x}\| = |\alpha| \cdot \|\bar{x}\|$ для довільного вектора $\bar{x}$ та для будь-якого дійсного числа α;

3. $\|x + y\| \leq \|x\| + \|y\|$ для довільних векторів $\bar{x}$ та $\bar{y}$ розмірів $(n \times 1)$ (нерівність трикутника).

Визначення 2: Норма вектора визначається наступним чином:

$$\|x\| = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \qquad (3.5)$$

де при будь-якому цілому додатному *p* визначається функція, що задовольняє аксіомам 1-3. Неважко переконатися, що це визначення відповідає всім трьом правилам.

Норми векторів, що найбільш застосовані у практиці.

Підставляючи цілі додатні значення *p* до визначення (3.5), можливо отримати різні види векторних норм. У обчислювальній математиці найбільш вживаними є наступні норми: $\|\bar{x}\|_1$, $\|\bar{x}\|_2$, $\|\bar{x}\|_\infty$.

1. *перша норма* (читають «норма-один», інша назва – октаедрична норма), обчислюється за формулою:

$\|\bar{x}\|_1 = \sum_{i=1}^{n} |x_i|$ – сума модулів елементів стовпчика.

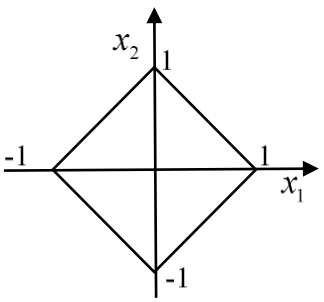


Рисунок 3.1 – до пояснення октаедричної норми

У разі введення такої норми безліч всіх одиничних векторів, норма яких дорівнює одиниці, тобто $|x_1| + |x_2| = 1$ буде виглядати на площині, як це

показано на рис. 3.1. У тривимірному випадку безліч всіх одиничних векторів, норма яких дорівнює одиниці, матиме форму октаедра. Тому перша норма також називається *октаедричною.*

2 *друга норма* (евклідова норма) знаходиться як:

$\|\bar{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$ – квадратний корінь з суми квадратів елементів.

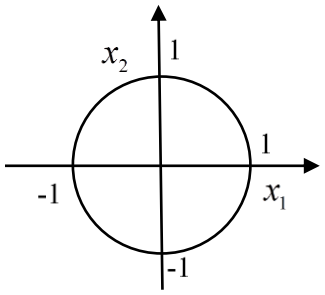


Рисунок 3.2 – до пояснення евклідової норми

У цьому випадку геометричне місце точок для двомірного простору буде задано рівнянням: $x_1^2 + x_2^2 = 1$, графік якого показано на рис. 3.2. Дана норма ще називається *евклідової нормою* через те, що співпадає з модулем стовпчика (довжиною вектору) тобто $\|\bar{x}\|_2 = |x| = \sqrt{x^{\mathrm{T}} x}$

3. кубічна норма («норма-нескінченність») обчислюється формулою:

$\|\bar{x}\|_\infty = \max_i |x_i|$ – максимум серед модулів елементів стовпчика.

На площині безліч всіх одиничних векторів, кубічна норма яких дорівнює одиниці, буде виглядати як це представлено на рис 3.3.

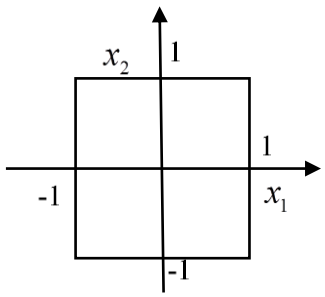


Рисунок 3.3 – до пояснення кубічної норми

Вище перелічені норми вектору пов’язані наступними нерівностями:

$\|\bar{x}\|_\infty \le \|\bar{x}\|_2 \le \|\bar{x}\|_1 \le n\|\bar{x}\|_\infty$ (n – дійсне додатне число).

За властивостями норми видно, що вона чисельно характеризує величину, або «довжину», вектора. Чим вона більше, тим більше відрізняється вектор (за сукупністю координат) від нульового. Тому норму є логічним взяти в якості характеристики величини похибки. Отже, абсолютною похибкою наближеного рішення називається норма вектора:

$$\Delta\bar{x}^* = \|\varepsilon\bar{x}^*\| = \|\bar{x} - \bar{x}^*\|$$

Як і у випадку з абсолютною похибкою скалярної величини $\Delta\bar{x}^*$ точно встановити неможливо. Замість неї знаходиться її верхня оцінка $\overline{\Delta\bar{x}^*}$. Безрозмірною характеристикою $\varepsilon\bar{x}^*$ є відносна похибка:

$$\delta\bar{x}^* = \frac{\bar{x}^*}{\|\bar{x}\|} = \frac{\|\bar{x} - \bar{x}^*\|}{\|\bar{x}\|}.$$

3.2.2 Норми матриць

Після того як ввели норму вектору необхідно ввести більш складну норму – *норму матриці.* Матричну норму можна визначити нескінченною безліччю способів. Так як в обчислювальних задачах лінійної алгебри потрібно оцінювати і вектори, і матриці, то доцільно вводити норму так, щоб вона розумним чином була пов'язана з застосованою векторною нормою. Зв'язок між нормою вектором та нормою матриці будемо визначати через погодженість. Нехай матриця С являє собою добуток матриць *A* та *B*. *C = AB* (кількість стовпчиків повинна дорівнювати кількості строк).

Визначення 3: Якщо $\|C\| \le \|A\| \cdot \|B\|$ то такі норми називають *погодженими* нормами матриць.

Будемо говорити, що матрична норма погоджена з даною векторною, якщо

$$\|A\bar{x}\| \le \|A\| \cdot \|\bar{x}\|. \qquad (3.6)$$

Матричні норми зручно визначати через норми векторів. Для цього задаючись нормою для матриць-стовбців, розглядають значення $\|A\bar{x}\|$ при будь-яких *x*, що задовільнять умові $\|\bar{x}\| = 1$. Особливе значення при оцінках має найменша норма матриці, узгоджена з даною векторною. З (3.6) очевидно випливає, що вона дорівнює максимуму відношення норм $\|A\bar{x}\|$ та $\|\bar{x}\|$. Виходячи з вищеописаних міркувань записуємо наступне визначення.

Визначення 4: ***Нормою матриці A***, погодженою з нормою вектору *x*, є величина, що дорівнює

$$\|A\| = \max_{x \ne 0} \frac{\|A\bar{x}\|}{\|\bar{x}\|}. \qquad (3.7)$$

Введена норма має наступні властивості, аналогічні властивостям норми вектора.

1. $\|A\| \geq 0$ для будь-якої матриці A, причому $\|A\| = 0$, тоді і лише тоді, коли A – нульова матриця;

2. $\|\alpha A\| = |\alpha| \cdot \|A\|$ для довільної матриці A та довільного числа α;

3. $\|A + B\| \leq \|A\| + \|B\|$ для будь-яких двох матриць розміру $n \times m$ (нерівність трикутника).

4. $\|A \cdot B\| \leq \|A\| \cdot \|B\|$ для будь-яких двох матриць, у яких визначений добуток

Кожній із векторних норм відповідає своя підлегла норма матриці. Найбільш вживаними є такі формули для обчислення значень норм матриць з дійсними елементами (приведемо їх без доведення):

$\|A\|_1 = \max\limits_{i \in N} \sum\limits_{j=1}^{n} |a_{ij}|$ – максимум суми модулів елементів у строчці;

$\|A\|_2 = \max\limits_{j \in N} \sum\limits_{i=1}^{n} |a_{ij}|$ – максимум суми модулів елементів у стовпчику;

$\|A\|_3 = \sqrt{\max\limits_{i \in N} \lambda_i \left(A^{\mathrm{T}} A\right)}$ – квадратний корінь з максимального власного значення λ_i матриці $A^{\mathrm{T}} A$;

$\|A\|_4 = \sqrt{\sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} a^2{}_{ij}}$ – квадратний корінь з суми квадратів елементів матриці A.

Потрібно зауважити, що обчислення норми $\|A\|_3$ пов’язане з трудомісткою операцією обчислення власних значень матриць. Справедливою є нерівність $\|A\|_3 = \sqrt{\max\limits_{i \in N} \lambda_i \left(A^{\mathrm{T}} A\right)} \leq \|A\|_4 = \sqrt{\sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} a^2{}_{ij}}$, тому для оцінок точності рішень норму $\|A\|_3$ замінюють нормою $\|A\|_4$.

Норма $\|A\|_4$ виникає, якщо матриці A поставити у відповідність «довгий стовпчик»: $\left(a_{11}, a_{21}, ..., a_{m1}, a_{12}, a_{22}, ..., a_{m2}, ..., a_{nn}\right)^T$ та застосувати норму вектору $\|x\|_3$.

За допомогою норм матриці оцінюють похибки розв’язку СЛАР. Відносна похибка вводиться аналогічно до похибки вектора за допомогою формули:

$$\delta A^* = \frac{\Delta A}{\|A\|} = \frac{\|A - A^*\|}{\|A\|},$$

де A та A^* – відповідно точні та наближенні значення матриці A.

Приклад 3.1 Визначити норми матриці $A = \begin{pmatrix} 4 & -2 & 1 \\ 1 & 3 & -2 \\ 3 & -1 & 4 \end{pmatrix}$.

Розв'язок:

$\|A\|_1 = max\{|4|+|-2|+|1|;|1|+|3|+|-2|;|3|+|-1|+|4|\} = 8;$

$\|A\|_2 = max\{|4|+|1|+|3|;|-2|+|3|+|-1|;|1|+|-2|+|4|\} = 8;$

$\|A\|_4 = \sqrt{4^2 + (-2)^2 + 1^2 + 1^2 + 3^2 + (-2^2) + 3^2 + (-1^2) + 4^2} = \sqrt{61} \approx 7{,}8.$

Матричні норми часто використовуються при аналізі обчислювальних методів лінійної алгебри. Наприклад, програма рішення систем лінійних алгебраїчних рівнянь може давати неточний результат, якщо *матриця коефіцієнтів* погано обумовлена («майже вироджена»).Для кількісної характеристики близькості до виродженості, потрібно вміти вимірювати відстань в просторі матриць. Таку можливість і дають матричні норми.

Отже, сенс визначення норм матриць полягає у тому, норми дозволяють оцінити вплив похибок правої частині та коефіцієнтів матриці на рішення системи лінійних алгебраїчних рівнянь. Щоб вміти визначати вказані похибки розглянемо таке поняття як *число обумовленості системи* (системи рівнянь).

3.3 Число обумовленості системи. Властивості числа обумовленості

Характер завдання та точність одержуваного рішення великою мірою залежать від його *обумовленості,* що є найважливішим математичним поняттям, що впливає на вибір методу рішення цього завдання. Оцінити таку обумовленість дозволяє число обумовленості, для виведення формули розрахунку числа обумовленості введемо деякі необхідні поняття.

Лема про еквівалентність норм: Нехай $\bar{x} \in R^N$. Обираємо дві різні норми одного й того ж вектору $\bar{x}$: $\|x\|_p$ та $\|x\|_q$ тоді:

$$\forall x \in R^N, \forall\, p,q \exists\, C_1 > 0, C_2 > 0 :\; C_1 \|x\|_p \le \|x\|_q \le C_2 \|x\|_p$$

де числа C_1 ,C_2 – константи еквівалентності.

Тепер перейдемо до визначення ***числа обумовленості системи.***

Записуємо СЛАР у матричному вигляді (3.2):

$$A\bar{x} = f$$

Нехай A – невироджена (неособлива) матриця, $\bar{x}$ – розв'язок задачі (3.1). Будемо вважати, що коефіцієнти матриці A точно відомі, а значення f було отримано з деякою похибкою Δf , тоді:

$$A(x + \Delta x) = f + \Delta f ,$$

де Δx – похибка розв'язку.

Нехай відома міра відносної похибки f у сенсі будь-якої норми:

$$\frac{\|\Delta f\|}{\|f\|} \le \delta_0 .$$

Знайдемо співвідношення, яке буде пов'язувати відносну похибку розв'язку з відносною похибкою правої частини. Тобто таке співвідношення, що дозволить оцінити на кільки сильно збуриться рішення системи (3.1)

$$\frac{\|\Delta x\|}{\|x\|} \le \mu \frac{\|\Delta f\|}{\|f\|},$$

де μ – число обумовленості. Якщо μ маленьке, то неточність рішення буде невеликою. Проте при великих μ навіть при маленьких похибках правої частини можливо допустити дуже велику похибку у рішенні. Знайдемо значення μ.

У силу лінійності задачі (3.2) можна переписати

$$A\Delta x = \Delta f ,$$

так як A – невироджена, то існує обернена матриця A^{-1}. Тоді:

$$\Delta x = A^{-1}\Delta f .$$

Через *погодженість* норми матриці та норми вектору отримаємо:

$$\|\Delta x\| = \|A^{-1}\| \cdot \|\Delta f\|,$$

та для погодженої задачі (3.1) маємо

$$\|f\| \le \|A\| \cdot \|x\| \Rightarrow \|x\| = \frac{\|f\|}{\|A\|},$$

$$\frac{\|\Delta x\|}{\|x\|} \le \|A^{-1}\| \cdot \|\Delta f\| \div \frac{\|f\|}{\|A\|} \Rightarrow \frac{\|\Delta x\|}{\|x\|} \le \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\Delta f\|}{\|f\|}.$$

Таким чином, число обумовленості системи дорівнює добутку норми матиці та оберненої до неї:

$$\mu(A) = \|A^{-1}\| \cdot \|A\|. \qquad (3.8)$$

Отже тепер можливо дати оцінку впливу похибок правої частини на розв'язок СЛАР. Тепер можливо дати визначення числу обумовленості:

Визначення: число обумовленості це - чисельна міра ступеня стійкості рішення, що дорівнює коефіцієнту можливого зростання похибки рішення по відношенню до похибок вихідних даних.

Для кращого розуміння поняття числа обумовленості розглянемо задачу розв'язку системи, що складається з двох лінійних рівнянь:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = f_1 \\ a_{21}x_1 + a_{22}x_2 = f_2 \end{cases},$$

точним розв'язком задачі є вектор, компоненти якого визначаються координатами точки перетину двох прямих, що відповідають рівнянням системи (рис. 3.4).

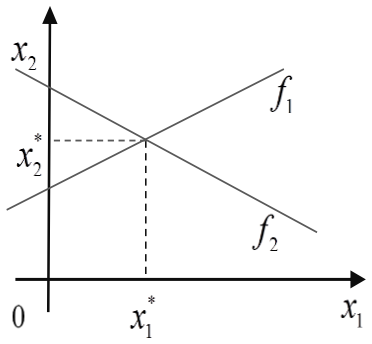


Рисунок 3.4 – графічне представлення розв'язку СЛАР

Проілюструємо характер обумовленості системи для різних випадків (рис 3.5):

- визначник *A суттєво відрізняється від нуля*, то точка перетину пунктирних прямих, що зміщуються відносно суцільни х прямих через похибки завдання *a* та *b*, несильно зміщається, тобто система добре обумовлена;

- визначник *A приблизно дорівнює нулю* – невеликі похибки у коефіцієнтах призводять до великих похибок у розв'язку – прямі близькі до паралельних;

- визначник *A дорівнює нулю* – прямі паралельні або співпадають тоді розв'язку не існує або їх нескінченна множина.

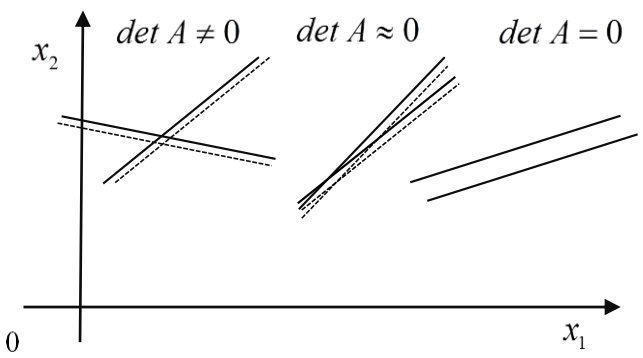


Рисунок 3.5 – Ілюстрація характеру обумовленості розв'язку СЛАР

Властивості числа обумовленості

1 $E = 1 = \left\| A \cdot A^{-1} \right\| \le \mu(A)$ якщо $\mu = 1$, то похибка у правій частині така ж сама як і похибка розв'язку СЛАР. μ завжди додатне.

2 $\mu(A) = \dfrac{max\left|\lambda_i(A)\right|}{min\left|\lambda_i(A)\right|}$, відношенню модулів максимальних до мінімальних власних чисел матриці А;

3. $\mu(AB) \le \mu(A)\mu(B)$.

Незважаючи на те, що число обумовленості матриці *залежить від вибору норми,* якщо матриця *добре обумовлена*, то її число обумовленості *буде малим* при будь-якому виборі норми а якщо вона погано обумовлена, то її число обумовленості буде великим при будь-якому виборі норми. Якщо $\mu(A) = 1-10$, матриця вважається добре обумовленою, а коли $\mu(A) \gg 10^2 - 10^3$.

Теоретично оцінювати похибку розв'язку системи $A\bar{x} = f$ за похибками вхідних даних можна за формулою $\delta x^* \le \mu(A) \cdot \left(\delta\left(f^*\right) + \delta\left(A^*\right)\right)$, де x^* розв'язок системи $A^*\bar{x}^* = f^*$.

Приклад 3.2

Розв'язком системи $\begin{cases} 100x + 99y = 199 \\ 99x + 98y = 197 \end{cases}$,

є пара значень $x = y = 1,$ вносимо у праві частини збурення

$$\begin{cases} 100x + 99y = 199{,}99 \\ 99x + 98y = 197{,}01 \end{cases},$$

при цьому рішення суттєво зміниться $x = 2{,}97,\ y = -0{,}99.$

Скористаємось погодженими першими нормами та для вектора правої частини та вектора внесеної абсолютної похибки відповідно отримаємо: $\|f\|_1 = 199,$ $\|\Delta f\|_1 = 10^{-2}.$ Відносна похибка $\delta_f = \dfrac{\|\Delta f\|_1}{\|f\|_1} \approx 0{,}5 \cdot 10^{-4},$ норма матриці коефіцієнтів оберненої до неї матриці дорівнюють $\|A\|_1 = \|A^{-1}\|_1 = 199,$ звідси число обумовленості системи $\mu = \|A\|_1 \cdot \|A^{-1}\|_1 = 199 \cdot 199 = 4 \cdot 10^4 \gg 1,$ що вказує на погано обумовлену систему. Тому невеликі збурення у правій частині викликають суттєві похибки у розв’язку.

Питання для самоперевірки:

1. Сформулюйте задачу чисельного рішення лінійної системи. Як розуміється оцінка похибки наближеного рішення?
2. Що таке похибка і нев’язка наближеного рішення? Як вони пов'язані та що характеризують?
3. Що таке векторна норма? Які характеристики вектора вона чисельно виражає?
4. Запишіть формули знаходження відомих вам векторних норм.
5. Що таке абсолютна і відносна похибки наближеного рішення лінійної системи?
6. Які векторні норми застосовуються в обчислювальній математиці? Як визначається вибір норми для оцінки похибки рішення?
7. Які властивості норми один, два і нескінченність.
8. Як пов'язані норми один, два і нескінченність?
9. Що таке матрична норма? Для чого вони застосовується в обчислювальній математиці?

10. Що таке погоджена матрична норма? Чому саме вона застосовується для оцінок матриць в обчислювальної математики?
11. Як обчислюється погоджена матрична норма в загальному випадку і для норм один, два і нескінченність?
12. Надайте визначення норми матриці. Які ви знаєте властивості норми матриці? Як розраховується відносна похибка матриці?
13. Надайте визначення відносного числа обумовленості матриці. Що таке погано обумовлена система рівнянь?
14. Надайте геометричну інтерпретацію погано обумовленої системи

Тема 2.2 Числові методи розв’язку систем лінійних алгебраїчних рівнянь (СЛАР)

ЛЕКЦІЯ 4 Методи розв’язку СЛАР. Прямі методи.

Навчальні питання:

4.1 Класифікація методів розв’язку СЛАР.

4.2 Прямі методи: правило Крамера, знаходження оберненої матриці

4.3 Метод Гауса. Модифікації методу Гауса

4.1 Класифікація методів розв’язку СЛАР

Методи розв’язку СЛАР виду

$$Ax = f \tag{4.1}$$

поділяють на два класи прямі (або точні) та ітераційні. Точні методи дозволяють отримати рішення шляхом виконання визначеної та точної кількості арифметичних операцій. При цьому похибка рішення визначається лише точністю представлення вихідних даних і точністю обчислювальних операцій.

Визначення 1: метод називається точним, якщо в припущенні відсутності похибок округлень, виходить точне рішення за *кінцеве число кроків*.

Ітераційні методи дають деяку послідовність наближень до рішення. Межею цієї послідовності є рішення системи, яке можливо визначити лише з деяким заданим ступенем точності ε. Кількість ітерацій, для досягнення необхідної точності рішення, визначається величиною ε, вибором початкового наближення та видом системи.

Визначення 2: метод називається ітераційним, якщо дозволяє обчислювати послідовність векторів, що при нескінченному збільшені кількості ітерацій *сходиться* до вірного рішення задачі.

На практиці при використанні ітераційних методів обмежуються обчисленням *кінцевого* числа наближень в залежності від допустимого рівня похибки.

До прямих методів відносять: правило Крамера; знаходження оберненої матриці; метод Гауса.

До ітераційних – метод Якобі, метод Зейделя, метод простої ітерації

Ітераційних методів існує дуже багато, тому щоб вивчити їх усі знадобився би окремий курс. У даній лекції йтиметься про прямі методи. У таких методах матрицю вихідної системи рівнянь еквівалентними перетвореннями призводять до більш простої матриці або розкладають на добуток простіших матриць. Це, як правило, різні варіанти методу послідовного виключення невідомих. Більшість точних методів відносяться до так званих методів виключення. У цих методах, послідовно виключаючи невідомі, вихідну систему призводять до системи з трикутною або діагональною матрицею.

Зважаючи на велике різноманіття методів вирішення задач, у подальшому важливу роль буде грати такий критерій відбору методу такий, як *трудомісткість* методу, що виражена в кількості необхідних арифметичних операцій потрібних для отримання розв'язку.

4.2 Прямі методи: правило Крамера, знаходження оберненої матриці

Як зазначено у визначені 1 у прямих методах точне (без урахування обчислювальних похибок) рішення досягається за кінцеве число кроків. Число необхідних арифметичних операцій в них залежить тільки від обчислювальної схеми і порядку матриці системи. Розглянемо перший прямий метод.

4.2.1 Правило Крамера. Будемо вважати, що $\Delta = det\, A \neq 0$ тобто рішення (4.1) існує та єдине.

Для знаходження розв'язку $x_i = \frac{\Delta_i}{\Delta}$ за правилом Крамера необхідно розрахувати *n+1* визначник. Розрахунок кожного визначника виконується за *n!* операцій. Якщо один доданок обчислюється, скажімо, за 10^{-6} с, що цілком допустимо для сучасних машин, та якщо взяти приблизно *n = 100,* $100!\ 10^{90}$, то час розрахунку складе абсолютно фантастичну цифру:

$$T \geq 10^{90} \cdot 10^{-6} c = \frac{10^{84}}{86400} діб = 3 \cdot 10^{76}\ років$$

Фактично ж в даний час з використанням відповідних методів вирішуються системи набагато більш високого порядку (до n ~10^4), отже можна стверджувати що метод Крамера придатний для розв'язку СЛАР не вище порядку 2 -4.

4.2.2 Знаходження оберненої матриці

Нехай задана система рівнянь (4.1), розв'язком такої системи буде вектор

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

Проте такий метод не завжди зручний. Нехай потрібно розв'язати просте рівняння $14x = 28$, скориставшись методом пошуку оберненої матриці знайдемо:

$$x = 14^{-1} \cdot 28 = 0{,}071428571428 \cdot 28 = 1{,}999999999984.$$

Крім втрати точності інший недолік пов'язаний з використанням методу оберненої матриці полягає у тому, що він потребує обчислення двох операцій – знаходження оберненої матриці – ділення та множення. Якщо би використовували метод Гауса, то це була б одна операція – ділення та точність при цьому більша.

4.3 Метод Гауса. Модифікації методу Гауса

Розглянемо метод виключень Гауса (схему єдиного поділу) розв'язку системи рівнянь. Метод Гауса складається з прямого та зворотного ходу.

При прямому ході матрицю приводять до верхньотрикутного вигляду. *Прямий хід* складається з $m-1$ кроків виключень.

1 крок: виключимо невідоме x_1 з рівнянь з номерами $i = 2,3,..m$. Припустимо, що $a_{11} \neq 0$, будемо називати його *провідним* (головним) елементом 1-го кроку. Знайдемо величини $\mu_{i1} = \frac{a_{1i}}{a_{11}}$ $(i = 2,3,..m)$, що звуться множниками 1-го кроку. Виключимо послідовно з другого, третього, ... m-го рівнянь системи перше рівняння, помножене відповідно на μ_{21}, μ_{31}, …, μ_{m1}. В результаті 1-го кроку отримаємо еквівалентну систему рівнянь:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + ... + a_{1m}x_m = f_1 \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + ... + a_{2m}^{(1)}x_m = f_2^{(1)} \\ \quad ... \qquad ... \qquad ... \\ a_{m2}^{(1)}x_2 + a_{m3}^{(1)}x_3 ... + a_{mm}x_m = f_m^{(1)} \end{cases}$$

Аналогічно проводяться інші кроки. Опишемо черговий *k-ий крок*. Припустимо, що провідний елемент $a_{kk} \neq 0$. Обчислимо множники k -го кроку: $\mu_{n1} \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}$, $i = k+1,...,m$ віднімемо послідовно з (k+1)-го, ...m -го рівнянь системи k -рівняння, помножене відповідно на $\mu_{k+1,k}$, $\mu_{k+2,k}$, …, $\mu_{m,k}$. Після (*m-1*)-го кроку виключення отримаємо систему рівнянь

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + ... + a_{1m}x_m = f_1 \\ \quad a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + ... + a_{2m}^{(1)}x_m = f_2^{(1)} \\ \quad\quad ... \quad\quad ... \quad\quad ... \\ \quad\quad\quad\quad\quad\quad a_{mm}^{(m-1)}x_m = f_m^{(1)} \end{cases}$$

матриця якої є верхньою трикутною. На цьому обчислення прямого ходу закінчуються. Визначимо трудомісткість прямого ходу методу Гауса. Для того щоб занулити елементи першого стовпчика потрібно $2n^2$ операцій. Для того щоб занулити елементи другого стовпчика потрібно $2(n-1)^2$ операцій і так далі. Всього потрібно буде обчислювальних операцій:

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

Отже трудомісткість операцій значно зменшується у порівнянні з методом оберненої матриці та тим паче з методом Крамера.

Після виконаних перетворень система має верхнєтрикутну матрицю. У цьому випадку розв'язок може бути отриманий у явному вигляді. Далі виконується з*воротний хід*: з останнього рівняння системи визначається x_m. Підставляючи знайдене значення x_m в передостаннє рівняння, отримуємо x_{m-1}. Далі послідовно знаходимо невідомі x_{m-2} ,...,x_2 ,x_1. Зворотній хід потребує приблизно $2n^2$ операцій.

Приклад 4.1

Дана система лінійних алгебраїчних рівнянь

$$\begin{cases} 2x_1 + x_2 + 4x_3 = 16 \\ 3x_1 + 2x_2 + x_3 = 10 \\ x_1 + 3x_2 + 3x_3 = 16 \end{cases},$$

знайти її розв'язок, використовуючи метод Гауса зі схемою єдиного поділу.

Розв'язок:

Виконуємо прямий хід. Поділимо перше рівняння системи на $a_{11} = 2 \neq 0$, отримаємо $x_1 + 0{,}5x_2 + 2x_3 = 8$.

Далі множимо отримане рівняння послідовно на a_{21}, a_{31} та віднімаємо, відповідно, з 2-го та 3-го рівнянь. У результаті отримаємо:

$$\begin{cases} x_1 + 0{,}5x_2 + 2x_3 = 8 \\ 0{,}5x_2 - 5x_3 = -14. \\ 2{,}5x_2 + x_3 = 8 \end{cases}$$

Таким чином виключили x_1 з усіх рівнянь, починаючи з другого. Далі перше рівняння вихідної системи залишаємо без змін та ділимо на $a_{22}^{(1)} = 0{,}5$ друге рівняння і аналогічно виключаємо x_2 з третього рівняння системи. У результаті приходимо до системи:

$$\begin{cases} x_1 + 0{,}5x_2 + 2x_3 = 8 \\ x_2 - 10x_3 = -28, \\ 26x_3 = 78 \end{cases}$$ та після ділення на $a_{33}^{(2)} = 26$ третього рівняння, запишемо

матрицю цієї системи, що має нулі нижче головної діагоналі (верхнє трикутна):

$$\begin{pmatrix} 1 & 0{,}5 & 2 & 8 \\ 0 & 1 & -10 & -28 \\ 0 & 0 & 1 & 3 \end{pmatrix}.$$

Виконаємо зворотній хід. Визначаємо невідомі x_i $(i = 1,2,3)$ з відповідних рівнянь починаючи з останнього та отримаємо розв'язок системи:

$$\begin{cases} x_1 + 0{,}5x_2 + 2x_3 = 8 \\ x_2 - 10x_3 = -28, \\ x_3 = 3 \end{cases} \quad \begin{cases} x_1 = 1 \\ x_2 = 2. \\ x_3 = 3 \end{cases}$$

Визначник матриці *A*, можна розрахувати як добуток провідних елементів:

$$det\, A = 2 \cdot 0.5 \cdot 26 = 26.$$

Зауваження

1. Схема єдиного поділу має обмеження, пов'язане з тим, що провідні елементи повинні бути *відмінні від нуля.* Одночасно бажано, щоб вони не були

малими по модулю, оскільки тоді похибки при відповідному розподілі будуть великими. З цієї точки зору використовується схема з вибором провідного елемента, що є більш кращою.

2. Після закінчення прямого ходу може бути обчислений визначник матриці А шляхом перемноження провідних елементів.

Виникає питання, що слід робити, якщо значення діагонального елемента близько до нуля, і коли метод Гауса є стійким. Відповіддю на це питання є наступна теорема.

Теорема про достатню умову стійкості методу Гауса.

За умови діагональної переваги:

$$|a_{ii}| > \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=1+1}^{n} |a_{ij}| + \varepsilon,\ \varepsilon > 0, \qquad (4.2)$$

метод Гауса є стійким. Тобто для $i = 1, ..., n$, при виконанні (4.2) проблеми, що з'являються в методі Гауса, не виникають. Якщо для всіх рядків матриці виконуються строгі нерівності, то говорять про суворе діагональне переважання.

Метод Гауса з вибором головного (провідного) елементу

Існують приклади, коли умова (4.2) за замовченням не виконується. У цьому випадку використовують модифікації методу Гауса, що мають покращені обчислювальні властивості. Таких модифікацій існує кілька.

1. ***Вибір головного елемента в рядку***. Виберемо в першому рядку найбільший елемент поміняємо місцями стовпчики з першим елементом і максимальним. На першому етапі будемо вирішувати таку систему. У модифікованому другому рівнянні вчинимо так само - поміняємо місцями стовпчики з другим елементом і максимальним і т. ін.

2. ***Вибір головного елемента в стовпці.*** Ідея методу така ж, як і в першому випадку, тільки місцями змінюються не стовпці, а рядки.

3. ***Вибір головного елемента по всій матриці.*** Такий алгоритм вимагає досить багато ресурсів і реалізується дуже рідко.

У схемах часткового вибору на *k*-му кроці прямого ходу в якості ведучого елемента вибирають максимальний по модулю коефіцієнт при невідомій в

рівняннях з номерами $i = k$ 1, ... , m. Потім рівняння, відповідне обраному коефіцієнту з номером, міняють місцями з k -им рівнянням системи для того, щоб головний елемент зайняв місце коефіцієнта $a_{kk}^{(k-1)}$. Після цієї перестановки рішення проводять як в схемі єдиного ділення. У цьому випадку всі масштабуючі множники по модулю менше одиниці і схема володіє обчислювальною стійкістю.

Метод Гауса для невироджених матриць еквівалентний *LU розкладанню матриці.* LU розкладання матриці представляє матрицю **A** у вигляді добутку нижньої трикутної матриці L і верхньої трикутної U.

Будь-яку квадратну матрицю, що має відмінні від нуля кутові мінори можна представити у вигляді *LU* розкладання, причому таке розкладання буде єдиним. Нехай матриця A представлена у наступному вигляді:

$$A = L \cdot U,\ L = \begin{pmatrix} 1 & 0 & 0 & ... & 0 \\ l_{21} & 1 & 0 & ... & 0 \\ l_{31} & l_{32} & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ l_{n1} & l_{n2} & l_{n3} & ... & 1 \end{pmatrix},\quad U = \begin{pmatrix} u_{11} & u_{12} & u_{13} & ... & u_{1n} \\ 0 & u_{22} & u_{23} & ... & u_{2n} \\ 0 & 0 & u_{33} & ... & u_{3n} \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & u_{nn} \end{pmatrix}. \qquad (4.3)$$

Тоді добуток матриць дорівнює:

$$\begin{pmatrix} 1 & 0 & 0 & ... & 0 \\ l_{21} & 1 & 0 & ... & 0 \\ l_{31} & l_{32} & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ l_{n1} & l_{n2} & l_{n3} & ... & 1 \end{pmatrix} \cdot \begin{pmatrix} u_{11} & u_{12} & u_{13} & ... & u_{1n} \\ 0 & u_{22} & u_{23} & ... & u_{2n} \\ 0 & 0 & u_{33} & ... & u_{3n} \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & u_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n} \\ a_{21} & a_{22} & a_{23} & ... & a_{2n} \\ a_{31} & a_{32} & a_{33} & ... & a_{3n} \\ ... & ... & ... & ... & ... \\ a_{n1} & a_{n2} & a_{n3} & ... & a_{nn} \end{pmatrix}$$

При перемноженні матриць отримаємо:

$$\begin{cases} u_{11} \cdot 1 = a_{11} \\ u_{12} \cdot 1 = a_{12} \\ ... \ .... \ .... \\ u_{1n} \cdot 1 = a_{1n} \end{cases}, \begin{cases} l_{21} \cdot u_{11} = a_{21} \\ l_{21} \cdot u_{12} + u_{22} = a_{22} \\ ... \ .... \ .... \\ l_{21} \cdot u_{1n} + u_{2n} = a_{2n} \end{cases}, \ ... \begin{cases} l_{n1} \cdot u_{11} = a_{n1} \\ l_{n1} \cdot u_{12} + l_{n2} \cdot u_{22} = a_{22} \\ ... \ .... \ .... \\ ... \ .... \ .... \end{cases}.$$

Для розв'язку такої системи з n^2 рівнянь існує прямий метод. Спочатку з першої системи визначають u_{11}, u_{12}, …, u_{1n}. Коли ці коефіцієнти відомі, знаходять l_{21}, l_{31}, …, l_{n1}. Далі коефіцієнти u_{22}, …., u_{2n} та l_{32}, l_{42}, …, l_{n2} і так далі. Кінцевий розв'язок має вигляд:

$$\begin{cases} u_{1j} = a_{1j}, \quad j = 1...n, \\ l_{i1} = \dfrac{a_{i1}}{u_{11}}, \quad i = 1...n, \\ u_{2j} = a_{2j} - l_{2j}u_{1j} \;\; j = 2...n, \\ l_{i2} = \dfrac{\left(a_{i2} - l_{i1}u_{12}\right)}{u_{22}}, \quad i = 2...n, \cdot \\ u_{ij} = a_{ij} - \sum\limits_{k=1}^{j-1} l_{ik}u_{kj}, \;\; i \le j \\ l_{ij} = d_{ij}^{-1}\left(a_{ij} - \sum\limits_{k=1}^{j-1} l_{ik}u_{kj}\right), i > j \end{cases} \tag{4.4}$$

Визначник матриці A може бути визначений як:

$$det\, A = det\, L \cdot det\, U = l_{11} \cdot l_{22} \cdot .... \cdot l_{nn}$$

З урахуванням (4.3) систему (4.1) можна записати як $L \cdot U \cdot x = f,$ тоді алгоритм метода LU-розкладання буде наступним:

1. Виконати факторизацію (розкладання на множники) вихідної матриці за формулами (4.4),
2. Розв'язати систему $L \cdot y = f,$ де $y = U \cdot x$,
3. Розв'язати систему $U \cdot x = y,$ x - вектор рішень СЛАР.

Приклад 4.2

Знайти розв'язок СЛАР методом LU-розкладання

$$\begin{cases} 3x_1 - x_2 = 5 \\ -2x_1 + x_2 + x_3 = 0 \\ 2x_1 - x_2 + 4x_3 = 1 \end{cases}$$

Розв'язок:

Виконаємо факторизацію вихідної матриці *A:*

$$\begin{pmatrix} 3 & -1 & 0 \\ -2 & 1 & 1 \\ 2 & -1 & 4 \end{pmatrix} \Rightarrow \begin{pmatrix} l_{11} & u_{12} & u_{13} \\ l_{21} & l_{22} & u_{23} \\ l_{31} & l_{32} & u_{33} \end{pmatrix} \Rightarrow \begin{pmatrix} 3 & -1/3 & 0 \\ -2 & 1/3 & 3 \\ 2 & -1/3 & 5 \end{pmatrix}$$

$l_{11} = a_{11} = 3;$ $\qquad u_{12} = \dfrac{a_{12}}{l_{11}} = -\dfrac{1}{3}$

$l_{21} = a_{21} = -2;$ $\qquad u_{13} = \dfrac{a_{13}}{l_{11}} = 0$

$l_{31} = a_{31} = 2;$

$l_{22} = a_{22} - l_{21} \cdot u_{12} = 1 - \dfrac{2}{3} = \dfrac{1}{3};$ $\qquad u_{23} = \dfrac{1}{l_{22}}\left(a_{23} - l_{21} \cdot u_{13}\right) = 1 \div \dfrac{1}{3}\left(1 - \left(-2\right) \cdot 0\right) = 3;$

$l_{32} = a_{32} - l_{31} \cdot u_{12} = -1 + \dfrac{2}{3} = -\dfrac{1}{3};$

$l_{33} = a_{33} - l_{31} \cdot u_{13} - l_{32} \cdot u_{23} = 4 - 2 \cdot 0 - 3 \cdot \dfrac{-1}{3} = 5.$

Перейдемо до розв'язку системи рівнянь $L \cdot y = f$

$$\begin{pmatrix} 3 & 0 & 0 \\ -2 & 1/3 & 0 \\ 2 & -1/3 & 5 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \\ 15 \end{pmatrix} \Rightarrow \begin{cases} 3y_1 = 5 \\ -2y_1 + \dfrac{y_2}{3} = 0 \\ 2y_1 - \dfrac{y_2}{3} + 5y_3 = 15 \end{cases} \Rightarrow \begin{cases} y_1 = 5/3 \\ y_2 = 10 \\ y_3 = 3 \end{cases}$$

Далі розв'язуємо систему $U \cdot x = y$ та записуємо розв'язок вихідної СЛАР:

$$\begin{pmatrix} 1 & -1/3 & 0 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5/3 \\ 10 \\ 3 \end{pmatrix} \Rightarrow \begin{cases} x_1 - x_2 \,/\, 3 = 5/3 \\ x_2 - 3x_3 = 10 \\ x_3 = 3 \end{cases} \Rightarrow \begin{cases} x_1 = 2 \\ x_2 = 1 \\ x_3 = 3 \end{cases}$$

Відповідь: $x_1 = 2,\ x_2 = 1,\ x_3 = 3.$

Слід зауважити, що крім перелічених вище прямих методів розв'язку СЛАР існують і інші, пристосовані для різного типу практичних завдань, методи розв'язку систем з матрицями спеціального виду. Так, наприклад, для матриць діагонального виду існує не розглянутий у лекції метод прогонки. До прямих методів також відносять методи обертань та квадратного кореня.

Питання для самоперевірки:

1. Що таке прямий метод рішення лінійної системи? Які прямі методи вам відомі?

2. Що таке трудомісткість методу? У якого з відомих вам прямих методів найбільша та найменша трудомісткість?
3. У чому полягає загальна ідея методу Гауса?
4. Що таке схема єдиного поділу методу Гауса? Опишіть її прямий хід.
5. Як вибираються провідні елементи кроків на прямому ході? Чому і за якої умови гарантується їх відмінність від нуля?
6. Опишіть зворотний хід методу Гауса.
7. Які складнощі прямого і зворотного ходів? Який з них більш трудомісткий?
8. Опишіть переваги та недоліки методу Гауса. Коли найдоцільніше його застосування?
9. Які модифікації методу Гауса вам відомі?
10. Що таке умови діагонального переважання? Що вони гарантують?
11. Що таке LU-розкладання?
12. Який алгоритм розв’язку СЛАР методом LU-розкладання?

ЛЕКЦІЯ 5 Ітераційні методи розв’язку СЛАР

Навчальні питання:

5.1 Метод простої ітерації. Теорема про збіжність МПІ

5.2 Канонічний вигляд запису ітераційних методів

5.3 Метод Якобі

5.4 Метод Зейделя

Альтернативою прямим методам розв’язку СЛАР є ітераційні методи, засновані на багаторазовому уточненні заданого наближеного рішення системи. Ітераційні методи (методи послідовних наближень) рішення систем є нескінченними методами і обчислюють тільки *наближені відповіді*. Їх ефективно застосовують для розв’язання задач великої розмірності, коли використання прямих методів неможливо через обмеження в доступній оперативній пам’яті ЕОМ або через необхідність виконання надмірно великої кількості арифметичних операцій. Великі системи рівнянь, що виникають в прикладних задачах, як правило, розріджені.

Ітераційні методи більш пристосовані для розріджених матриць, оскільки вони вимагають набагато менше оперативної пам’яті, ніж прямі методи, і можуть бути використані, незважаючи на те, що вимагають більше часу на виконання. У той же час ітераційні методи для погано обумовлених задач анітрохи не краще, ніж метод виключення Гауса. Застосовані до погано обумовлених задач вони дають, як правило, такий же невірний результат, що і прямі методи.

Отже, при використанні ітераційних методів будується послідовність наближених рішень (ітерацій), що збігається до точного. При досягненні заданої точності обчислення припиняються, і остання ітерація видається за рішення завдання. При цьому потрібно так вибрати початкову ітерацію, щоб послідовність збігалась до точного рішення. Тому для ітераційних методів, крім власне формули методу, потрібні: множина початкових ітерацій, для яких метод збігається (область збіжності) та оцінки похибки рішення на кожному кроці.

5. 1 Метод простої ітерації. Теорема про збіжність МПІ

Метод простої ітерації означає наступне: нехай знову дано систему рівнянь:

$$Ax = f \tag{5.1}$$

Проведемо кілька рівносильних перетворень. Помножимо обидві частини системи на один і той же скалярний множник τ, потім додамо до правої і лівої частин системи вектор x:

$$x + \tau Ax = \tau f + x$$
$$x = \tau f + x - \tau Ax$$
$$x = x\underbrace{\left(E - \tau A\right)}_{R} + \underbrace{\tau f}_{F}$$

Таким чином систему рівнянь можна записати у вигляді, зручному для ітерацій у наступного виді:

$$x = Rx + F \tag{5.2}$$

де R – матриця переходу. R – квадратна матриця порядку n, F – стовпчик – модифікована права частина. Якщо $R = E - A$, $F = f$, де E - одинична квадратна матриця порядку n, то у такому випадку метод простої ітерації буде називатись методом послідовних наближень.

Для перетворення (5.2) завжди можна влаштувати ітераційний процес виду:

$$x^{(k+1)} = Rx^{(k)} + F \tag{5.3}$$

Верхнім індексом в дужках тут і далі по тексту позначається номер ітерації (сукупності повторюваних дій). Стовпчик F приймається в якості початкового наближення $x^{(0)} = F$ і далі багаторазово виконуються дії щодо уточнення рішення, згідно рекурентному співвідношення (5.3). У розгорнутому виді (5.3) виглядає наступним чином:

$$\begin{cases} x_1^{(k+1)} = r_{11}x_1^{(k)} + r_{12}x_2^{(k)} + \ldots + r_{1n}x_n^{(k)} + f_1 \\ x_2^{(k+1)} = r_{21}x_1^{(k)} + r_{22}x_2^{(k)} + \ldots + r_{2n}x_n^{(k)} + f_2 \\ \quad \ldots \qquad \ldots \qquad \ldots \\ x_2^{(k+1)} = r_{n1}x_1^{(k)} + r_{n2}x_2^{(k)} + \ldots + r_{nn}x_n^{(k)} + f_n \end{cases}$$

Зауваження

1. Початкове наближення $x^{(0)}$ може вибиратися довільно або з деяких міркувань. При цьому може використовуватися завжди апріорна інформація про

рішення або просто "груба" прикидка.

При виконанні ітерацій виникають наступні питання:

а) чи сходиться процес (3), тобто чи має місце $x^{(k)} \to x^*$, при k → ∞, де x* – точне рішення?

б) якщо збіжність є, то яка її швидкість?

в) яка похибка знайденого рішення $x^{(k+1)}$, тобто чому дорівнює норма різниці $\|x^{(k)} - x^*\|$

Відповідь на питання збіжності методу простої ітерації та оцінки похибок дають наступні теореми.

ТЕОРЕМА 1 про достатню умову збіжності методу простих ітерацій. Ітераційна послідовність (5.3) збігається до точного розв'язку системи x (5.2) якщо виконується умова

$$\|R\| < 1$$

при будь-якій початковій ітерації $x^{(0)}$ і при цьому оцінка похибки буде наступною:

$$\Delta^{(k+1)} = \left\|x - x^{(k+1)}\right\| < \|R\|^{k+1} \cdot \left\|x - x^{(0)}\right\| = \|R\|^{k+1} \Delta^{(0)}, \qquad (5.4)$$

$k = 0,1,\ldots$.

Доведення: нехай x - точний розв'язок, $x^{(k)}$ - k-те наближення до розв'язку. З (5.2) віднімемо (5.3):

$$x - x^{(k+1)} = R\left(x - x^{(k)}\right), \qquad (5.5)$$

Застосовуємо норму матриць та враховуємо властивість норми (аксіома №4):

$$\left\|x - x^{(k+1)}\right\| = \left\|R\left(x - x^{(k)}\right)\right\| \le \|R\| \cdot \left\|\left(x - x^{(k)}\right)\right\| \qquad (5.6)$$

Рівність (5.5) є справедливою для будь-яких k, тому його можна переписати як:

$$x - x^{(k)} = R\left(x - x^{(k-1)}\right) \Rightarrow \left\|x - x^{(k)}\right\| = \left\|R\left(x - x^{(k-1)}\right)\right\| \le \|R\| \cdot \left\|\left(x - x^{(k-1)}\right)\right\|$$

Підставляючи замість $\left\|x - x^{(k)}\right\|$ в (5.6) вище отриману оцінку отримаємо:

$$\left\|x - x^{(k+1)}\right\| \le \|R\|^2 \cdot \left\|\left(x - x^{(k-1)}\right)\right\|,$$

застосовуючи таким самим чином (5.5) k разів, приходимо до (5.4). За умовою

$\|R\| < 1$, отже $\lim\limits_{k\to\infty}\left\|x - x^{(k+1)}\right\| = 0$. Тобто при наближенні кількості ітерацій до нескінченності, рішення наближається до точного, що вказує на збіжність ітераційної послідовності.

Оцінка (5.4) показує, що на кожному кроці похибка зменшується у $\|R\|$ разів. У такому випадку говорять, що метод сходиться зі швидкістю геометричної прогресії зі знаменником $q \le \|R\| < 1$. Таку швидкість називають лінійною.

Зауваження до теореми

1. Процес, що є збіжним має властивість "самовиправленості", тобто окрема помилка в обчисленнях не відіб'ється на остаточному результаті, так як помилкове наближення можна розглядати, як нове початкове.

2. Умови збіжності виконуються, якщо в матриці A діагональні елементи переважають, тобто

$$|a_{ii}| \ge |a_{i1}| + ... + |a_{i,i-1}| + |a_{i,i+1}| + ... + |a_{in}|, \qquad i = 1,2,...,n, \tag{5.7}$$

і хоча б для одного і нерівність строга. Іншими словами, модулі діагональних коефіцієнтів в кожному рівнянні системи більше суми модулів недіагональних коефіцієнтів (вільні члени не розглядаються).

3. Чим менше величина норми $\|R\|$, тим швидше збіжність методу.

Застосовувати (5.4) у якості *критерію зупину ітерацій* у практичних завданнях не можливо через відсутність інформації про точне значення рішень x, тому для цієї мети використовують оцінку, яку доводять з аналогічних міркувань Теореми 1:

$$\Delta^{(k+1)} = \left\|x - x^{(k+1)}\right\| \le \frac{\|R\|}{1-\|R\|}\left\|x^{(k+1)} - x^{(k)}\right\|. \tag{5.8}$$

Оцінка (5.8) дозволяє сформулювати умову зупинки ітераційного процесу, а саме: якщо потрібно знайти рішення з заданою точністю ε, то потрібно обчислювати $x^{(k+1)}$ доки не буде виконуватись нерівність

$$\frac{\|R\|}{1-\|R\|} \cdot \left\|x^{(k+1)} - x^{(k)}\right\| \le \varepsilon,$$

або $\left\|x^{(k+1)}-x^{(k)}\right\|\le\varepsilon_1$, де $\varepsilon_1=\dfrac{1-\|R\|}{\|R\|}\cdot\varepsilon$.

Також для при відомих власних значеннях матриці R справедливою є наступна теорема.

<u>ТЕОРЕМА 2</u> про необхідну і достатню умову збіжності методу простих ітерацій. (без доведення).

Для збіжності методу простих ітерацій при будь-яких $x^{(0)}$ і F необхідно і достатньо, щоб власні значення матриці R були по модулю менше одиниці, тобто $\left\|\lambda_i(R)\right\|<1,\ i=1,\ ...,\ n$

Перетворення системи до виду $x=Rx+F$ з матрицею R, що задовольняє умовам збіжності, може бути виконано кількома способами. Наведемо способи, використовувані найбільш часто.

1. Рівняння, що входять до системи $Ax=f$ переставляються таким чином, щоб виконувалась умова (6) переваги діагональних елементів. Далі перше рівняння представляють x_1, з другого x_2 і так далі. При цьому отримуємо матрицю R з нульовими діагональними елементами.

Наприклад, система

$\begin{cases}3x_1-x_2+4x_3=8\\10x_1-x_2+8x_3=10\\-x_1+5x_2-0{,}4x_3=20\end{cases}$ за допомогою перестановки рівнянь може бути приведена до виду

$\begin{cases}10x_1-x_2+3x_3=10,\\-x_1+5x_2-0{,}4x_3=20,\\2x_1-x_2+4x_3=8,\end{cases}$ звідси отримаємо $|10|>|-1|+|2|;\ \ |5|>|1|+|1|;\ \ |4|>|-0{,}4|+|3|;$

тобто діагональні елементи переважають. Виражаємо з відповідного рівняння невідомі x_1, x_2, x_3 та отримуємо систему виду $x=Rx+F$:

$\begin{cases}x_1=0\cdot x_1+0{,}1x_2-0{,}3x_3+1,\\x_2=0{,}2x_1+0\cdot x_2+0{,}08x_3+4,\\x_3=-0{,}5x_1+0{,}25x_2+0\cdot x_3+2,\end{cases}$, де $R=\begin{pmatrix}0&0{,}1&-0{,}3\\0{,}2&0&0{,}08\\-0{,}5&0{,}25&0\end{pmatrix}$, $F=\begin{pmatrix}1\\4\\2\end{pmatrix}$,

при цьому виконується умова теореми про достатню умову збіжності МПІ, тобто

$\|R\|_1 = max\{0,4;0,28;0,75\} < 1$.

2. Рівняння перетворюють таким чином, щоб виконувалась умова переваги діагональних елементів, но при цьому елементи головної діагоналі матриці *R* не обов'язково дорівнювали нулю

Наприклад систему

$$\begin{cases} 1,04x_1 - 0,2x_2 + 0,3x_3 = 2,7, \\ -0,1x_1 + 1,07x_2 - 0,1x_3 = 4 \\ 0,2x_1 - 0,1x_2 + 1,4x_3 = 3,5, \end{cases}$$, можливо записати у вигляді:

$$\begin{cases} x_1 = 0,4x_1 + 0,2x_2 - 0,3x_3 + 2,7, \\ x_2 = 0,1x_1 - 0,7x_2 + 0,1x_3 + 4, \\ x_3 = -0,2x_1 + 0,1x_2 - 0,4x_3 + 3,5, \end{cases}$$, для якого $\|R\|_1 = max\{0,9;0,9;0,7\} < 1$.

Алгоритм методу простих ітерацій:

1. Перетворити систему $Ax = f$ до виду $x = Rx + F$, одним з вищеописаних способів

2. Задати початкове наближення $x^{(0)}$ виходячи з аналітичних міркувань довільно, або покласти $x^{(0)} = F$

3. Розрахувати наступне наближення за формулою $x^{(k+1)} = Rx^{(k)} + F$

4. Якщо виконується умова $\left\|x^{(k+1)} - x^{(k)}\right\| < \varepsilon$, процес завершити та у якості наближеного рішення прийняти $x^* \approx x^{(k+1)}$, інакше покласти $k = k + 1$ та повернутися до пункту №3 алгоритму.

Приклад 5.1 Виконати розв'язок СЛАР методом ітерацій

$$\begin{cases} 2x_1 + 10x_2 - 6x_3 = 72, \\ -3x_1 + x_2 + 25x_3 = -92, \\ 20x_1 - 4x_2 - 2x_3 = 32. \end{cases}$$

Розв'язок:

Приведемо систему до виду $x = Rx + F$:

$$\begin{cases} x_1 = -5x_2 + 3x_3 + 36, \\ x_2 = 3x_1 - 25x_3 - 92, \\ x_3 = 10x_1 - 2x_2 + 16. \end{cases}$$ та перевіримо умову збіжності:

$R = \begin{pmatrix} 0 & -5 & 3 \\ 3 & 0 & -25 \\ 10 & -2 & 0 \end{pmatrix}$, норма матриці $\|R\|_1 = 28 > 1$, отже застосування методу ітерацій не гарантує збіжності. Потрібно врахувати умову діагональної переваги та знов привести вихідну систему до виду $x = Rx + F$:

$$\begin{cases} 20x_1 - 4x_2 - 2x_3 = 32, \\ 2x_1 + 10x_2 - 6x_3 = 72, \\ -3x_1 + x_2 + 25x_3 = -92, \end{cases} \Rightarrow \begin{cases} x_1 = 0{,}2x_2 + 0{,}1x_3 - 1{,}6 \\ x_2 = -0{,}2x_1 + 0{,}6x_3 + 7{,}2 \\ x_3 = 0{,}12x_1 - 0{,}04x_2 - 3{,}68 \end{cases} \Rightarrow R = \begin{pmatrix} 0 & 0{,}2 & 0{,}1 \\ -0{,}2 & 0 & 0{,}6 \\ 0{,}12 & -0{,}04 & 0 \end{pmatrix}$$

норма матриці $\|R\|_1 = 0{,}8 < 1$, тепер можна переходити до ітераційного процесу та шукати розв'язок СЛАР. Обираємо у якості початкового наближення $x^{(0)}$ вектор $F = \begin{pmatrix} -1{,}6 \\ 7{,}2 \\ -3{,}68 \end{pmatrix}$, звідси:

$$\begin{cases} x_1^{(1)} = 0{,}2x_2^{(0)} + 0{,}1x_3^{(0)} - 1{,}6 \\ x_2^{(1)} = -0{,}2x_1^{(0)} + 0{,}6x_3^{(0)} + 7{,}2 \\ x_3^{(1)} = 0{,}12x_1^{(0)} - 0{,}04x_2^{(0)} - 3{,}68 \end{cases} \Rightarrow \begin{cases} x_1^{(1)} = 0{,}2 \cdot 7{,}2 + 0{,}1 \cdot (-3{,}68) - 1{,}6 = -0{,}5280, \\ x_2^{(1)} = -0{,}2 \cdot (-1{,}6) + 0{,}6 \cdot (-3{,}68) + 7{,}2 = 5{,}3120, \\ x_3^{(1)} = 0{,}12 \cdot (-1{,}6) - 0{,}04 \cdot 7{,}2 - 3{,}68 = -4{,}1600 \end{cases}$$

Зведемо послідовність векторів наближень на першому та на наступних кроках у таблицю:

k	0	1	2	3	4	5	6
x_1	-1,6	-0,5280	-0,9536	-1,0337	-0,9952	-0,9975	-1,0008
x_2	7,2	5,3120	4,8096	5,0172	5,0146	4,9962	4,9995
x_3	-3,68	-4,1600	-3,9558	-3,9868	-4,0047	-4,0000	-3,9996

Послідовність ітерацій приводить до наближеного рішення з точністю $\varepsilon = \left\| x^{(6)} - x^{(5)} \right\| = 1{,}7 \cdot 10^{-2}$.

Відповідь: $x_1 = -1{,}0008$, $x_2 = 4{,}9995$, $x_3 = -3{,}9996$.

5.2 Канонічний вигляд запису ітераційних методів.

Слід зауважити, що існує велика кількість ітераційних методів розв'язку

СЛАР. У першій частині лекції розглянуті МПІ у широкому сенсі, тобто такі методи можуть застосовуватись до будь-яких матриць. Проте, матриці можна приводити до більш зручного для розв'язки виду. Це здебільшого буде впливати на вже знайомі параметри τ, R (матриця переходу).

Такі типи ітераційних процесів як $x^{(k+1)} = Rx^{(k)} + F$, коли для розрахунку $k+1$ -го наближення використовується попереднє наближення k, ще називають *двошаровими*. Для відповідних двошарових ітераційних методів передбачений *канонічний вид запису*

$$B_k \frac{x^{(k+1)} - x^{(k)}}{\tau_k} + Ax^{(k)} = f \tag{5.7}$$

Якщо порівняти з виразом $x^{(k+1)} = x^{(k)} \underbrace{(E - \tau A)}_{R} + \underbrace{\tau f}_{F}$ то зрозуміло що для МПІ $B_k = E$, $\tau_k = \tau \ \forall k$.

У тому випадку коли $B_k = E$, методи називаються явними. $\tau_k = 1$ – методи без параметру. Яка основна вимога до B_k? На це питання потрібно відповідати виходячи з оцінки кількості кроків для отримання точного рішення СЛАР. Згадаємо, що прямими методами потрібно виконати $\approx \frac{2}{3}n^3$ операцій. Звідси витікає основна вимога B_k – оберненість за $< \frac{2}{3}n^3$ ітерацій за скільки обертається A. Це можливо, коли матриця B_k діагональна, верхнє-, нижнє трикутна. Такі матриці обертаються за $\approx n^2$ операцій.

5.3 Метод Якобі

Перейдемо до розгляду МПІ у вузькому сенсі. Тобто до таких методів, які можна застосовувати до певного типу заданих розріджених матриць. До таких методів відноситься метод Якобі. Також метод Якобі називають загальним методом простої ітерації.

Запишемо систему рівнянь в координатному вигляді:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = f_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = f_2 \\ \quad ... \quad\quad ... \quad\quad ... \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = f_n \end{cases}$$

З першого рівняння будемо знаходити нові наближення для x_1 використовуючи значення $x_2 ... x_n$ отримані на попередніх ітераціях. З другого рівняння будемо знаходити нові наближення для x_2, залишаючи старі значення для всіх інших змінних. Тоді отримаємо:

$$\begin{cases} a_{11}x_1^{(k+1)} + a_{12}x_2^{(k)} + ... + a_{1n}x_n^{(k)} = f_1 \\ a_{21}x_1^{(k)} + a_{22}x_2^{(k)} + ... + a_{2n}x_n^{(k)} = f_2 \\ \quad ... \quad\quad ... \quad\quad ... \\ a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + ... + a_{nn}x_n^{(k)} = f_n \end{cases}$$

Розв'язок буде записаний наступним чином:

$$\begin{cases} x_1^{(k+1)} = \dfrac{1}{a_{11}}\left(f_1 - \sum\limits_{j=2}^{n} a_{1j}x_j^{(k)} \right) \\ x_i^{(k+1)} = \dfrac{1}{a_{ii}}\left(f_i - \sum\limits_{j=1}^{i-1} a_{1j}x_j^{(k)} - \sum\limits_{j=1+1}^{n} a_{1j}x_j^{(k)} \right) \end{cases} \tag{5.8}$$

Запишемо всі операції у матричному вигляді. Представимо матрицю A наступним чином:

$$A = L + D + U$$

$$A = \underbrace{\begin{pmatrix} 0 & 0 & ... & 0 \\ a_{21} & 0 & ... & 0 \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & 0 \end{pmatrix}}_{L} + \underbrace{\begin{pmatrix} a_{11} & 0 & ... & 0 \\ 0 & a_{22} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & a_{nn} \end{pmatrix}}_{D} + \underbrace{\begin{pmatrix} 0 & a_{12} & ... & a_{1n} \\ 0 & 0 & ... & a_{2n} \\ ... & ... & ... & ... \\ 0 & 0 & ... & 0 \end{pmatrix}}_{U}$$

Матриця D – діагональна, матриці L та U – нижнє та верхнє трикутні відповідно. Тоді отримаємо: $Lx^{(k)} + Dx^{(k+1)} + Ux^{(k)} = f$, $\quad Dx^{(k+1)} = -\left(L + U\right)x^{(k)} + f$, звідси

$$x^{(k+1)} = -D^{-1}\left(L + U\right)x^{(k)} + D^{-1}f \tag{5.9}$$

Отримали метод простої ітерації з іншою матрицею переходу. Природно, для цього методу справедливі твердження про збіжність і оцінки похибки попереднього пункту.

ТЕОРЕМА 3 (без доведення) про достатню умову збіжності методу Якобі

Метод Якобі збігається тоді і лише тоді, коли всі значення λ_i, що визначаються рівнянням

$$det\begin{vmatrix} \lambda a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & \lambda a_{22} & ... & a_{2n} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & \lambda a_{nn} \end{vmatrix}$$

по модулю менші одиниці $\forall i \quad |\lambda_i| < 1.$

Приклад 5.2 При яких a, b сходиться метод простої ітерації $x^{(k+1)} = x^{(k)} R + F$, де

$$R = \begin{pmatrix} a & b & 0 \\ b & a & b \\ 0 & b & a \end{pmatrix}$$

Розв'язок:

Для того, щоб метод простої ітерації сходився до розв'язку відповідної *СЛАР*, за *Теоремою 3* необхідно і достатньо, щоб всі власні значення матриці R по модулю були менше одиниці: $\forall i \quad |\lambda_i| < 1$. Розв'язуємо характеристичне рівняння:

$$det(R - \lambda E) = \begin{vmatrix} a-\lambda & b & 0 \\ b & a-\lambda & b \\ 0 & b & a-\lambda \end{vmatrix} = (a-\lambda)\begin{vmatrix} a-\lambda & b \\ b & a-\lambda \end{vmatrix} - b\begin{vmatrix} b & a \\ 0 & a-\lambda \end{vmatrix} =$$

$$= (a-\lambda)\left[(a-\lambda)^2 - b^2\right] - b^2(a-\lambda) = (a-\lambda)\left(a-\lambda-\sqrt{2b}\right)\left(a-\lambda+\sqrt{2b}\right) = 0$$

звідки отримаємо умову збіжності ітераційного методу: $|a| < 1, \left|a \pm \sqrt{2b}\right| < 1.$

5.4 Метод Зейделя

Метод Зейделя являє собою модифікацію методу Якобі, яка полягає в наступному. В методі Зейделя враховуються компоненти рішення, обчислені раніше.

Перше рівняння записуємо таким самим чином як і для методу Якобі. Для другого виписуємо компоненти, які вже знаємо за розрахунком з минулого кроку.

$$\begin{cases} a_{11}x_1^{(k+1)} + a_{12}x_2^{(k)} + \ldots + a_{1n}x_n^{(k)} = f_1 \\ a_{21}x_1^{(k+1)} + a_{22}x_2^{(k+1)} + \ldots + a_{2n}x_n^{(k)} = f_2 \\ \ldots \quad \ldots \quad \ldots \\ a_{n1}x_1^{(k+1)} + a_{n2}x_2^{(k+1)} + \ldots + a_{nn}x_n^{(k+1)} = f_n \end{cases}$$

Розв'язок буде записаний наступним чином:

$$\begin{cases} x_1^{(k+1)} = \dfrac{1}{a_{11}}\left(f_1 - \sum_{j=2}^{n} a_{1j}x_j^{(k)} \right) \\ x_i^{(k+1)} = \dfrac{1}{a_{ii}}\left(f_i - \sum_{j=1}^{i-1} a_{1j}x_j^{(k+1)} - \sum_{j=1+1}^{n} a_{1j}x_j^{(k)} \right) \end{cases}$$

Такий метод буде збігатися швидше, ніж метод Якобі.

<u>ТЕОРЕМА 4</u> (без доведення) про необхідну та достатню умову збіжності методу Зейделя

Необхідною і достатньою умовою збіжності методу Зейделя є вимога, щоб всі значення λ_i, що визначаються рівнянням

$$det\begin{vmatrix} \lambda a_{11} & a_{12} & \ldots & a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \lambda a_{n1} & \lambda a_{n2} & \ldots & \lambda a_{nn} \end{vmatrix},$$

були по модулю менші одиниці $\forall i \quad |\lambda_i| < 1.$

Розглянемо, що означає метод Зейделя з точки зору МПІ. Знову представимо *A* наступним чином:

$$A = L + D + U$$

Тоді отримаємо:

$$Lx^{(k+1)} + Dx^{(k+1)} + Ux^{(k)} = f, \quad (L+D)x^{(k+1)} = -Ux^{(k)} + f$$

звідси $$x^{(k+1)} = \underbrace{-(L+D)^{-1}U}_{R}x^{(k)} + \underbrace{(L+D)^{-1}f}_{F}, \qquad (5.10)$$

У розширеному вигляді метод Зейделя буде виглядати таким чином:

$$\begin{cases} x_1^{(k+1)} = a_{11}x_1^{(k)} + a_{12}x_2^{(k)} + \ldots + a_{1n}x_n^{(k)} + f_1 \\ x_2^{(k+1)} = a_{21}x_1^{(k+1)} + a_{22}x_2^{(k)} + \ldots + a_{2n}x_n^{(k)} + f_2 \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ x_n^{(k+1)} = a_{n1}x_1^{(k+1)} + a_{n2}x_2^{(k+1)} + \ldots + a_{nn}x_{n-1}^{(k+1)} + f_n \end{cases}, \tag{5.11}$$

Приклад 5.3 Знайти умову збіжності ітераційного *методу Зейделя* для *СЛАР* $Ax = f$ з матрицею A виду

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

Розв'язок:

Для методу Зейделя ітераційний процес $x^{(k+1)} = Rx^{(k)} + F$, де $R = -(L+D)^{-1}U$.

Має місце рівняння: $B\omega = \lambda\omega$, де λ та ω – відповідно, власне число і власний вектор. В такому випадку $-(L+D)^{-1}U\omega = \lambda\omega$, звідки (припускаємо наявність нетривіальних розв'язків в заданій *СЛАР*):

$$det(\lambda L + \lambda D + U) = 0.$$

Вирахувавши визначник, прийдемо до алгебраїчного рівняння

$$det\begin{vmatrix} \lambda a & b & 0 \\ \lambda b & \lambda a & b \\ 0 & \lambda b & \lambda a \end{vmatrix} = a\lambda^2\left(a^2\lambda - 2b^2\right) = 0.$$

В такому випадку, оскільки $\lambda_{1,2} = 0$, $\lambda_3 = 2\frac{b^2}{a^2}$, отримаємо умову збіжності *методу Зейделя*: $\frac{b}{a} < 2^{-\frac{1}{2}}$.

Зауваження:

1. Для забезпечення збіжності методу Зейделя необхідно перевести систему $Ax = f$ до виду $x = Rx + F$ з переважанням діагональних елементів в матриці A

2. Як правило, метод Зейделя забезпечує кращу збіжність, ніж метод простих ітерацій (за рахунок накопичення інформації, отриманої при вирішенні попередніх

рівнянь). Метод Зейделя може збігатися, якщо розходиться метод простих ітерацій, і навпаки.

3. При розрахунках на ЕОМ зручніше користуватися формулою (5.11).

4. Перевагою методу Зейделя, як і МПІ, є його "самовиправленність".

Алгоритм метода Зейделя:

1. Перетворити систему $Ax = f$ до виду $x = Rx + F$.
2. Задати початкове наближення $x^{(0)}$ довільно або покласти $x^{(0)} = F$, а також мале додатне число ε (точність). Покласти $k = 0$.
3. Розрахувати наступне наближення за формулою (5.10) або (5.11).
4. Якщо виконується умова $\left\| x^{(k+1)} - x^{(k)} \right\| < \varepsilon$, процес завершити та у якості наближеного рішення прийняти $x^* \approx x^{(k+1)}$, інакше покласти $k = k + 1$ та повернутися до пункту №3 алгоритму.

Розвитком методу Зейделя є ***метод релаксації***. В цьому методі вводять ітераційний параметр τ, що називають *параметром релаксації*.

Представимо метод релаксації в матричній формі:

$$\left(\tau L x^{(k+1)} + D x^{(k+1)}\right) + (\tau - 1) D x^{(k)} + \tau U x^{(k)} = \tau f,$$

Вибираючи τ, можна значно змінювати швидкість збіжності ітераційного методу. Виразимо $x^{(k+1)}$

$$x^{(k+1)} = -(D + \tau L)^{-1} + \left[(\tau - 1) D + \tau L\right] x^{(k)} + \tau (D + \tau L)^{-1} f,$$

В загальному випадку задача пошуку $\tau_{опт}$ (оптимального ітераційного параметру) не вирішена, однак відомо, що $1 < \tau_{опт} < 2$. В цьому випадку ітераційний метод називається методом *послідовної верхньої релаксації* або SOR – Successive Over Relaxation. Іноді зустрічається термін "надрелаксація" при $1 < \tau_{опт} < 2$, а при $0 < \tau < 1$ маємо метод нижньої релаксації.

Приклад 5.4 Виконати розв'язок СЛАР методом Зейделя, з точністю $\varepsilon = 0{,}005$

$$\begin{cases} 4x_1 - x_2 + x_3 = 4, \\ x_1 + 6x_2 + 2x_3 = 9, \\ -x_1 - 2x_2 + 5x_3 = 2. \end{cases}$$

Розв'язок:

Перевіримо умову діагональної переваги $|4|>|-1|+|1|$, $|6|>|1|+|2|$, $|5|>|-1|+|-2|$ – умова виконується. Приведемо систему до виду $x = Rx + F$.

$$\begin{cases} x_1 = \frac{1}{4}x_2 - \frac{1}{4}x_3 + 1, \\ x_2 = -\frac{1}{6}x_1 - \frac{1}{3}x_3 - \frac{2}{3}, \\ x_3 = \frac{1}{5}x_1 - \frac{2}{5}x_2 + \frac{2}{5}. \end{cases}$$ обираємо у якості початкового наближення $x^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.

$k = 0$, знаходимо $k+1$-наближення:

$$\begin{cases} x_1^{(1)} = 0{,}25x_2^{(0)} - 0{,}25x_3^{(0)} + 1 \\ x_2^{(1)} = -0{,}167x_1^{(1)} - 0{,}333x_3^{(0)} + 1{,}5 \\ x_3^{(1)} = 0{,}2x_1^{(1)} + 0{,}4x_2^{(1)} + 0{,}4 \end{cases} \Rightarrow \begin{cases} x_1^{(1)} = 1, \\ x_2^{(1)} = 1{,}333; \\ x_3^{(1)} = 1{,}133. \end{cases}$$

Зведемо послідовність векторів наближень на першому та на наступних кроках у таблицю:

k	**0**	**1**	**2**	**3**	**4**	**5**
x_1	0	1	1,0501	0,9896	1,0010	1,0000
x_2	0	1,3333	0,9473	1,0051	0,9999	1,0000
x_3	0	1,1333	0,9889	0,9999	1,0000	1,0000
$\varepsilon^* = \left\|x^{(k)} - x^{(k-1)}\right\|$	--	1,3333	0,3860	0,0604	0,0114	0,001

Розв'язок системи методом Зейделя збігається до точного на п'ятому ітераційному кроці, при цьому $\varepsilon^* < \varepsilon = 0{,}005$.

Відповідь: $x_1 = x_2 = x_3 = 1$.

Питання для самоперевірки:

1. Що таке ітераційний метод? Що потрібно для його вичерпного визначення?
2. Що означає збіжність ітераційного методу?
3. Як готується система для методу простої ітерації? Які особливості та значення переходу до еквівалентної системи?
4. Дайте визначення методу послідовних наближень?

5. Сформулюйте і доведіть умову збіжності і апріорну оцінку похибки методу простої ітерації.
6. Сформулюйте критерій зупинки ітераційного процесу.
7. Чи залежить збіжність методу від обраної норми? Чому?
8. Як готується система для методу Якобі? За яких умов можливий такий перехід?
9. Опишіть ітераційний процес методу Якобі? У чому відмінність від методу простої ітерації?
10. Які умова збіжності і оцінки похибки методу Якобі?
11. Опишіть ітераційний процес методу Зейделя? У чому відмінність від методу Якобі?
12. Які умова збіжності і оцінки похибки методу Зейделя?
13. З якою швидкістю сходиться метод Зейделя? Від чого вона залежить?
14. Сформулюйте критерій зупинки ітераційного процесу методу Зейделя.
15. Метод Зейделя є поліпшенням методу Якобі. Чи завжди він сходиться швидше методу Якобі? Чому?

РОЗДІЛ 3 ЧИСЛОВІ МЕТОДИ РОЗВ'ЯЗКУ НЕЛІНІЙНИХ РІВНЯНЬ

Тема 3.1 Найбільш поширені методи розв'язання нелінійних рівнянь.

ЛЕКЦІЯ 6-7 Локалізація кореня. Метод половинного поділу, метод ітерацій. Методи хорд, дотичних, комбінований метод.

Навчальні питання:

6.1 Формулювання задачі

6.2 Локалізація коренів

6.3 Метод половинного поділу

6.4 Метод простої ітерації (послідовних наближень)

6.5 Метод хорд

6.6 Метод дотичних (метод Ньютона)

6.7 Комбінований метод хорд і дотичних

Рішення рівнянь – це важлива прикладна задача. Практично у всіх інженерних, наукових розрахунках доводиться розв'язувати рівняння або їх системи. Для деяких видів рівнянь відомі формули точних коренів (квадратні, кубічні, деякі тригонометричні, тощо). Але, по-перше, вони охоплюють дуже вузьке коло рівнянь, в той час як в обчислювальній практиці доводиться вирішувати найрізноманітніші рівняння. По-друге, формули точного рішення можуть бути громіздкі, а тому важкі для практичного застосування. Тому виникає задача чисельного, або наближеного, рішення рівнянь і систем. У цій лекції визначені та описані методи вирішення нелінійних рівнянь.

6.1 Формулювання задачі

Нехай розглядається рівняння

$$f(x) = 0 \tag{6.1}$$

Таке рівняння називають нелінійним, якщо функція f – нелінійна. Коренем рівняння називається таке число, що його підстановка замість змінної повертає рівняння у тотожність. Точне (теоретичне) значення кореня будемо позначати так

само, як змінну *x*. Завдання полягає у знаходженні наближеного значення кореня x^* тобто такого числа, що $f(x) \approx 0$. Вказана наближена рівність означає що $\left| f(x^*) \right| < \varepsilon$.

Корінь x^* називається *простим*, якщо $f'(x^*) \neq 0$, в іншому випадку корінь називається *кратним*. Ціле число m називається *кратністю кореня* x^*, якщо $f^{(k)}(x^*) = 0$ для k = 1,2,3 -, m-1 та $f^{(m)}(x^*) \neq 0$.

$f(x)$ – функція, визначена та неперервна на деякому проміжку. У деяких випадках на функцію $f(x)$ можуть бути накладені додаткові обмеження, наприклад, безперервність першої і другої похідних, що спеціально оговорюється. Функція $f(x)$ може бути задана у вигляді *багаточлена або трансцендентною функцією* (тоді їй відповідає відповідно алгебраїчне або трансцендентне рівняння). Обмежимося обговоренням проблеми пошуку дійсних коренів.

6.2 Локалізація коренів

Всі методи чисельного рішення нелінійних рівнянь – ітераційні (прямі методи розробляються в фундаментальній математиці). Вони складаються з двох етапів.

Перший, попередній – це локалізація (відділення) кореня. Другий, основний, - це ітераційне уточнення кореня, тобто власне рішення рівняння.

- **локалізація коренів**, тобто попередній аналіз розташування коренів на осі x, в результаті якого виявляються такі відрізки осі x, кожному з яких належить *не більше одного кореня;*

- **обчислення з необхідною точністю кореня** (коренів), що належать заданому відрізку (заданим відрізкам).

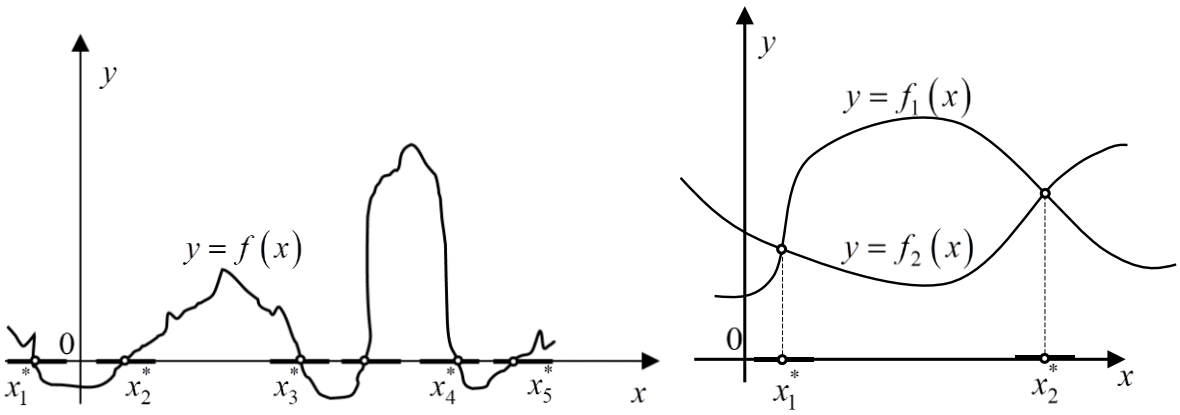
Локалізація здійснюється через дослідження функції. Для цього застосовуються найрізноманітніші методи, які залежать від виду рівняння, тому неможливо дати загальний універсальний алгоритм. Але є деякі *загальні рекомендації.*

Перша – використовувати графік. У цьому випадку будують графік (або ескіз графіка) функції та визначають проміжки (див. рис. 6.1*а*), на яких він перетинає вісь ОХ. Абсциси точок перетину ($x_1^*,...,x_n^*$) – це і є корені рівняння (6.1).

Друга рекомендація – перетворити рівняння (6.1). Наприклад, можна привести його до виду

$$f_1(x) = f_2(x),$$

де функції $f_1(x)$ і $f_2(x)$ вибираються так, щоб це рівняння було простіше для дослідження, ніж вихідне. Положення кореня можна визначити також графічно, побудувавши графіки функцій $y = f_1(x)$, $y = f_2(x)$ (див. рис. 6.1 *б*). Корінь рівняння – абсциса точки перетину цих графіків.



а) *б*)

Рисунок 6.1 – Графічне представлення способу локалізації кореня

Іноді корисно привести рівняння до виду:

$$x = \varphi(x), \tag{6.2}$$

і тоді корінь шукається як абсциса точки перетину графіка функції $y = \varphi(x)$ з прямою $y = x$. Також однією з рекомендацій є за можливістю побудувати *таблицю значень* функції і по ній визначити проміжки, на яких функція змінює знак. Якщо функція неперервна, то на такому проміжку буде хоча б один корінь. Зрозуміло, що для кращої локалізації потрібно, щоб вузли таблиці йшли з невеликим кроком, а

також слід застосовувати інші методи дослідження функцій. Зокрема, визначити проміжки монотонності функції.

При дослідженні властивостей рівняння можна застосовувати такі теореми з курсу математичного аналізу.

ТЕОРЕМА 1 (Больцано-Коші). Якщо безперервна на відрізку [*a, b*] функція має на його кінцях протилежні знаки, тобто $f(a)\,f(b) < 0$, то на інтервалі (*a, b*) вона хоча б один раз обертається в нуль.

ТЕОРЕМА 2. Безперервна строго монотонна функція має на відрізку єдиний нуль тоді і тільки тоді, коли на його кінцях вона приймає значення різних знаків.

Також при локалізації застосовуються всі методи дослідження функції з математичного аналізу. Часто замість відрізка локалізації достатньо вказати початкове наближення до кореня.

Потрібно зазначити, що локалізація коренів виходить за рамки власне обчислювальної математики та є у першу чергу предметом математичного аналізу. Втім, на етапі локалізації можна скористатися ЕОМ для того, щоб обчислити таблицю значень функції *f*(*x*) або побудувати графік цієї функції.

Вважаючи, що корні локалізовані в обумовленому вище сенсі, будемо займатися таким конкретним завданням.

Обчислити із заданою точністю ε *корінь рівняння (1), що належить відрізку* [*a, b*]. Передбачається, що на відрізку [*a, b*] знаходиться єдиний корінь. Крім того, будемо вважати, що шуканий корінь є простим (некратними).

Зауваження: при вирішенні завдань *фізичного змісту* часто відрізок [*a,b*], який містить необхідний корінь, відомий з фізичних міркувань. Що стосується малого параметра ε, що характеризує необхідну точність, або, іншими словами, *допустимий рівень похибки*, то сенс його полягає в тому, що обчислене значення кореня *x* повинно відрізнятися від точного не більше, ніж на $\varepsilon : \left|x - x^*\right| < \varepsilon$.

Найбільш поширеними методами уточнення коренів є наступні: *метод половинного поділу (дихотомії, бісекцій), метод простої ітерації, метод хорд, метод дотичних (Ньютона), існують ще комбіновані методи (хорд і дотичних).*

Після локалізації кореня вступає в дію його ітераційне уточнення: вибирається початкова ітерація $x^{(0)}$ з проміжку локалізації і будується послідовність $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$, що сходиться до точного рішення. Зрозуміло, що обчислення триває не безкінечно, а зупиняється при досягненні заданої точності. Тому для повноти методу треба визначити:

1. розрахункову формулу для довільної k -й ітерації;

2. оцінку похибки ітерації (для критерію зупину);

3. умови збіжності (зрозуміло, що інтервал локалізації сам по собі не забезпечує збіжність).

Все перераховане залежить від конкретного методу. Почнемо з найпростішого – *методу половинного поділу*.

6.3 Метод половинного поділу

Існує група методів, що базуються на ділені відрізка навпіл. Нехай функція, яку досліджуємо неперервна та існує, (a,b) – інтервал локалізації; функція неперервна на відрізку, $[a,b]$ та на його кінцях приймає значення різних знаків $f(a)\,f(b)<0$.

Ділимо відрізок, $[a,b]$ навпіл, $c=\dfrac{a+b}{2}$ – його середина (рис. 6.2),

- перевіряємо умову $f(c)=0$, тобто чи є c коренем рівняння. Якщо так, то рішення знайдено, тобто c – корінь. Якщо ні, то
- визначаємо, на якому з відрізків тепер корінь, на $[a,c]$ чи на $[c,b]$. Для цього достатньо перевірити знак добутку $f(a)\,f(c)$. Якщо він від'ємний – то корінь на відрізку $[a,c]$, якщо додатній то корінь на відрізку $[c,b]$. Відрізок, що містить корінь зменшиться у двічі. Як наступну ітерацію кореня приймаємо середину нового відрізка і т. д.

Зрозуміло, що виконання точної рівності (6.1) малоймовірно, а при наближеному обчисленні це практично неможливо, тому насправді перевіряють виконання цієї рівності із заданою точністю. На кожному кроці довжина відрізка,

що містить корінь, зменшується вдвічі; і можна локалізувати корінь з якою завгодно точністю.

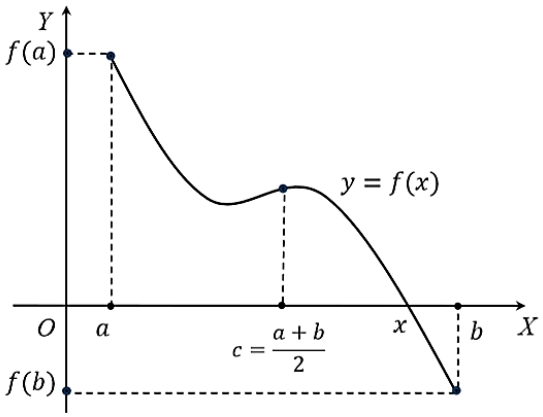


Рисунок 6. 2 – Геометрична інтерпретація розв’язку рівнянь методом половинного поділу

Після кожної ітерації відрізок, на якому розташований корінь, зменшується вдвічі, тобто після *n* ітерацій він скорочується в 2^n разів. Процес закінчується якщо $\left|x^{(k)} - x^{(k-1)}\right| < \varepsilon$.

Алгоритм методу половинного поділу

Алгоритм методу половинного поділу полягає в побудові послідовності відкладених відрізків, на кінцях яких функція приймає значення різних знаків. Кожен наступний відрізок отримують діленням навпіл попереднього. Опишемо кроки ітерації методу

1. Знайти початковий інтервал невизначеності $L_0 = [a_0, b_0]$ одним зі способів локалізації коренів. Задати мале додатне ε та покласти $k = 0$;

2. Знайти середину чинного інтервалу невизначеності $c_k = \dfrac{a_k + b_k}{2}$;

3. Якщо $f(a_k) f(c_k) < 0$, то покласти $a_{k+1} = a_k$, $b_{k+1} = c_k$, а якщо $f(c_k) f(b_k) < 0$, то прийняти $a_{k+1} = c_k$, $b_{k+1} = b_k$ в результаті знаходиться новий інтервал невизначеності $L_{k+1} = [a_{k+1}, b_{k+1}]$;

4. Якщо $b_{k+1} - a_{k+1} > \varepsilon$, то покласти $k = k + 1$ та перейти до пункту №2.

Критерій закінчення ітераційного процесу: якщо довжина відрізку локалізації менше 2ε, то ітерації припиняються та у якості значення кореня з заданою точністю приймають середину відрізку.

Зауваження:

Метод половинного поділу має лінійну, але безумовну збіжність та його похибка за кожну ітерацію зменшується у два рази:

$$|b_1 - a_1| = \frac{|b_0 - a_0|}{2}, \quad |b_2 - a_2| = \frac{|b_1 - a_1|}{2} = \frac{|b_0 - a_0|}{2^2}, \ldots, |b_k - a_k| = |b_0 - a_0| 2^{-k}.$$

Останній вираз дозволяє оцінити кількість ітерацій для досягнення заданої точності ε

$$|b_0 - a_0| 2^{-k} < \varepsilon \Rightarrow k \geq log_2 \frac{|b_0 - a_0|}{\varepsilon}. \tag{6.3}$$

Звідси витікає що, якщо наприклад, потрібно виконати обчислення з точністю $\varepsilon \approx 10^{-3}$, то задавшись $(b_0 - a_0) \approx 1$, необхідно буде здійснити приблизно 10 ітерацій.

Очевидно, що рано чи пізно потрібна локалізація буде досягнута при будь-якій початковій ітерації, тому метод поділу відрізка навпіл є гарантовано збіжним. Проте він має очевидний та істотний *недолік:* мала швидкість збіжності. На кожному кроці похибка пошуку кореня зменшується всього у два рази, тобто метод половинного поділу має лінійну швидкість збіжності і збігається зі швидкістю геометричної прогресії зі знаменником 1/2. Також до недоліків методу слід віднести, що він не узагальнюється на системи нелінійних рівнянь та не може бути використаним для пошуку кратних коренів.

6.4 Метод простої ітерації (послідовних наближень)

Метод простої ітерації є узагальненням методу простої ітерації для лінійних систем на нелінійний випадок. Для застосування методу ітерацій вихідне рівняння $f(x) = 0$ (де $f(x)$ - неперервна функція) необхідно,

- замінити ріносильним (таким, що має ті самі корені) рівнянням $x = \varphi(x)$;
- виділити інтервал $[a, b]$. ізоляції кореня цього рівняння;

- обрати нульове наближення кореня x_0.

Перехід від $f(x)=0$ до $x=\varphi(x)$ можна зробити багатьма способами, і він важливий, так як від функції $\varphi(x)$ залежить збіжність ітераційної послідовності. Наприклад, можна поступити зовсім просто: помножити (6.1) на число *c*, відмінне від нуля і додати до обох частин *x*, тоді отримаємо рівняння:

$$x = x - c \cdot f(x).$$

Ітераційна послідовність будується так само, як й в лінійному випадку. Для одержання першого наближення x_1 в праву частину рівняння $x=\varphi(x)$ замість x підставляємо x_0, так що $x_1=\varphi(x_0)$. Наступні наближення утворюються за схемою:

$$x^{(2)}=\varphi\left(x^{(1)}\right);\quad x^{(3)}=\varphi\left(x^{(2)}\right);\ldots;\quad x^{(k)}=\varphi\left(x^{k-1}\right);\ldots$$

Таким чином, у результаті застосування деякого однакового процесу будуються послідовні наближення: $x^{(0)},x^{(1)},x^{(2)},\ldots..,x^{(k)};\ldots$

Розрахункова формула виглядає наступним чином:

$$x^{(k)}=\varphi\left(x^{(k-1)}\right). \qquad (6.4)$$

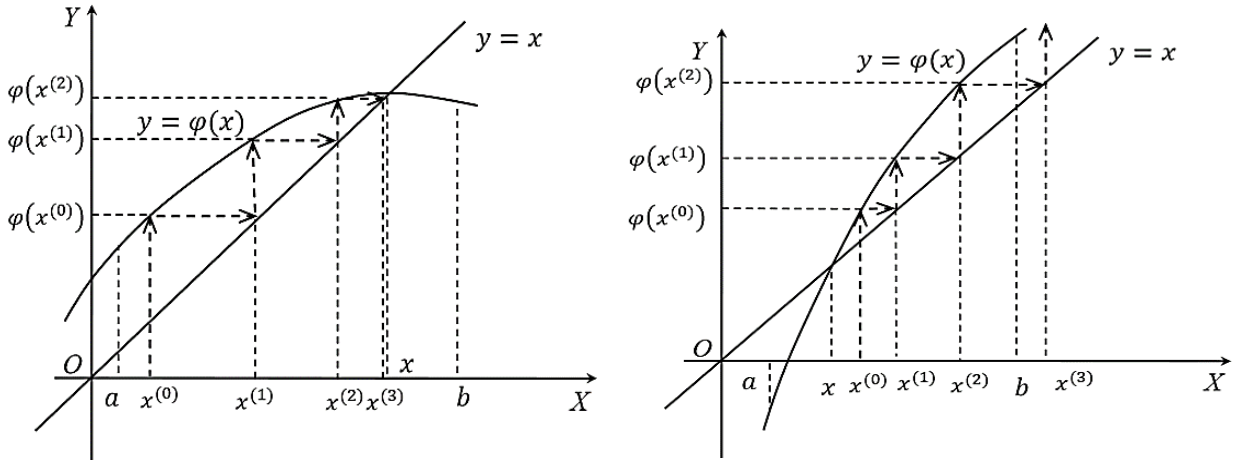
Якщо ітераційна послідовність $\{x^{(0)},x^{(1)},x^{(2)},\ldots,x^{(k)},\ldots\}$ має границю x^* та функція φ неперервна на проміжку локалізації, то ця границя є коренем рівняння. При цьому можуть бути два випадки:

1) процес може збігатися, тобто послідовні наближення прямують до деякої кінцевої границі x^*, що є коренем рівняння;

2) процес може розходиться, тобто кінцевої границі побудованих наближень існувати не буде; із цього не випливає, що розв'язок вихідного рівняння не існує, просто могло виявитися, що процес послідовних наближень *обраний невдало*.

На рис. 6.3 показана графічна ілюстрація методу простої ітерації. Корінь рівняння – абсциса точки перетину прямої $y=x$ та кривої $y=\varphi(x)$. Для початкової точки $x^{(0)}$ знаходиться точка $\left(x^{(0)},\varphi\left(x^{(0)}\right)\right)$ (рис. 6.3 *а*). Через неї проводиться пряма паралельно осі до перетину з прямою $y=x$. Абсциса точки перетину – нова ітерація

$x^{(1)}$. Далі проводяться аналогічні побудови. Послідовність ітерацій на рис. 6.3 *а* збігається до точного значення кореня: границя x послідовності $\{x^{(k)}\}$ існує і



а) б)

Рисунок 6.3 – Геометрична інтерпретація уточнення кореня рівнянь методом простої ітерації

збігається з коренем. Проте, як зазначено вище, існують випадки коли послідовність може розходитись (рис. 6.3 *б*), та це не означає що коренів не існує.

Природно, виникає питання про умови збіжності, а також оцінки похибки ітерації для критерію зупину. Збіжність процесу ітерації визначається наступною теоремою.

ТЕОРЕМА про збіжність методу простої ітерацій. Нехай інтервал $[a,b]$ є інтервалом кореня рівняння $x=\varphi(x)$, а функція $\varphi(x)$ визначена та диференційована на всьому інтервалі, причому усі її значення $\varphi(x)\in[a,b]$.

Тоді, якщо існує правильний дріб q, такий, що $|\varphi'(x)|\le q<1$, то:

1) процес ітерації $x^{(k)}=\varphi\left(x^{(k-1)}\right)$, $(k=1,2,...)$ збігається незалежно від початкового значення $x_0\in[a,b]$;

2) граничне значення $x^*=\lim\limits_{n\to\infty}x_n$ є єдиним коренем рівняння $x=\varphi(x)$ на відрізку $[a,b]$.

Критерій закінчення ітераційного процесу

Наближення $x^{(0)}, x^{(1)}, x^{(2)}, ..., x^{(k)}, ...$ слід обчислювати доти, поки не буде виконана нерівність:

$$\left|x^k - x^{k+1}\right| \le \frac{\varepsilon(1-q)}{q} = \varepsilon_1, \tag{6.4}$$

де ε - задана умовою гранична абсолютна похибка кореня x^* (точність, $\varepsilon > 0$). Якщо величина $0 < q < 0{,}5$, то можна використовувати більш простий критерій закінчення ітерацій: $\left|x^k - x^{k+1}\right| < \varepsilon$.

Алгоритм методу простої ітерації:

1. рівняння (6.1) рівносильними перетвореннями привести до вигляду $x = \varphi(x)$, це може бути виконано будь-яким способом але для збіжності має виконуватись умова теореми $\left|\varphi'(x)\right| \le q < 1$

2. задати початкове наближення $x_0 \in [a, b]$ та мале додатнє, покласти $k = 0$.

3. $x = \varphi(x)$ розрахувати наступне наближення $x^{(k)} = \varphi\left(x^{(k-1)}\right)$

4. якщо $\left|x^k - x^{k+1}\right| \le \varepsilon_1$, ітерації завершують, та $x^* \approx x^k$, якщо ні то повертаємось до п. №3.

Приклад 6.1 Методом ітерацій з $\varepsilon = 10^{-4}$ уточнити корінь x^* рівняння $5x^3 - 20x + 3 = 0$, ізольованийий на відрізку [0, 1].

Розв'язок:

Зведемо рівняння до вигляду $x = \varphi(x)$. Це можна зробити таким чином:

1. $x = x + \left(5x^3 - 20x + 3\right)$, тоді $\varphi_1(x) = 5x^3 - 19x + 3$;

2. $x = \dfrac{5x^3 + 3}{20}$, тоді $\varphi_2(x) = \dfrac{5x^3 + 3}{20}$.

Визначимо, яку з отриманих функцій $\varphi_1(x)$ чи $\varphi_2(x)$ слід використовувати. Знаходимо:

$$\left|\varphi_1'(x)\right| = \left|15x^2 - 19\right| > 1 \quad \textit{на } [0, 1];$$

$$\left|\varphi_2'(x)\right| = \left|\frac{15x^2}{20}\right| = \frac{3}{4}x^2 < 1 \quad на\ [0,\ 1].$$

Звідси, $\varphi_2(x)$ задовольняє теоремі про збіжність, тому $\varphi_2(x)$ можна використовувати для пошуку послідовних наближень за формулою:

$$x_{k+1} = \frac{5x_k^{\,3} + 3}{20}.$$

Визначимо критерій зупину ітераційного процесу:

$$\left|\varphi'(x)\right| \le q = \frac{15}{20} = 0{,}75 \quad на\ [0,\ 1], \quad \left|x_{k+1} - x_k\right| \le \frac{0{,}0001\cdot(1-0{,}75)}{0{,}75} = 0{,}00003 = 3\cdot 10^{-5}.$$

За початкове наближення обираємо ліву межу інтервалу $x^{(0)} = 0$.

Обчислення зводимо у таблицю нижче:

№ ітерації	x_k	$x_k^{\,3}$	$x_{k+1} = \varphi(x_k)$	$\left\|x_{k+1} - x_k\right\|$
0	0	0	0.15	0.15
1	0.15	0.003375	0.15084375	0.00084375
2	0.15084375	0.0034323	0.150858068	$0.000014 < \varepsilon_1$

Відповідь: $x^* \approx 0.150858068$, при цьому $f(x_2) \approx 4\cdot 10^{-6}$.

У зв'язку з розглядом питання про збіжність введемо наступну термінологію. Нехай деякий ітераційний процес генерує послідовність $\left\{x^{(k)}\right\}_{k=0}^{\infty}$, що має за границю точне значення x^*.

Збіжність послідовної ітерацій до x^* називається *лінійною,* якщо існує така постійна $C \in (0;1)$ і такий номер *K*, що

$$\left|\tilde{x} - x^{(k+1)}\right| \le C\left|\tilde{x} - x^{(k)}\right|, \tag{6.5}$$

при всіх $k \ge K$.

Збіжність називається *надлінійною*, якщо існує така додатна $\left\{C^{(k)}\right\}_{k=0}^{\infty}$ числова послідовність, що збігається до нуля і такий номер *K*, що

$$\left|\tilde{x} - x^{(k+1)}\right| \le C^{(k)}\left|\tilde{x} - x^{(k)}\right|, \tag{6.6}$$

при всіх $k \ge K$.

Послідовність $\left\{x^{(k)}\right\}_{k=0}^{\infty}$ сходиться до щонайменше з p-порядком, якщо, знайдуться такі константи $C>0,\ p\geq 1,\ K>0$, що

$$\left|\tilde{x}-x^{(k+1)}\right|\leq C\left|\tilde{x}-x^{(k)}\right|^{p}, \tag{6.7}$$

при всіх $k\geq K$. При $p=1$ виходить лінійна збіжність.

Метод простої ітерації має лінійну збіжність. У лекції про СЛАР така збіжність була названа збіжність зі швидкістю геометричної прогресії. Число q з теореми про збіжність методу простої ітерації – це знаменник прогресії, що характеризує швидкість збіжності: чим менше q, тим більше швидкість.

6.5 Метод хорд

Ідея методу хорд полягає в тому, що на достатньо малому проміжку $[a,b]$ *дуга кривої,* що описується рівнянням $y=f(x)$ *замінюється стягуючою її хордою.* Шуканий корінь рівняння $f(x)=0$ є абсциса точки перетину графіка функції $y=f(x)$ з віссю *OX*. Ця точка невідома, але замість неї можливо обрати точку x_1 перетину хорди із віссю *OX*. Показати хід ітераціїного процесу легше всього графічно.

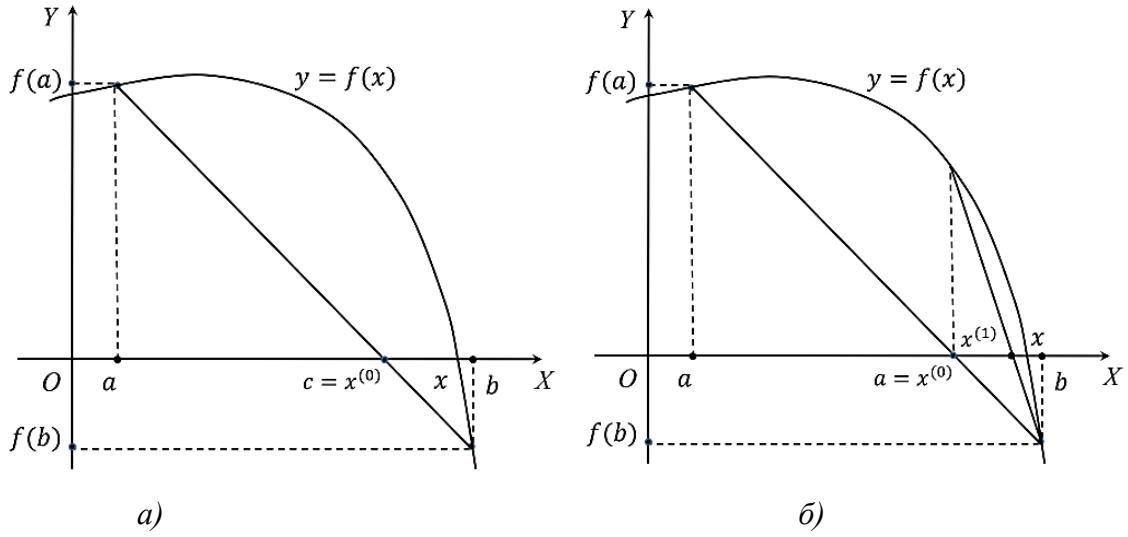
Розглянемо випадок, коли перша і друга похідні мають однакові знаки (рис. 6.4.), тобто $f'(x)\cdot f''(x)>0$ ($f'(x)<0$– функція спадає, $f''(x)<0$ – функція опукла вгору). Поєднаємо крайні точки дуги кривої $y=f(x)$ на $[a,b]$ хордою та візмемо у якості нульового наближення кореня в даному випадку абцису точки перетину хорди з *OX*, тобто $x^{(0)}=c$ (рис. 6.4 а). Далі потрібно визначити на якому з відрізків $[a,c]$ чи $[c,b]$ знаходиться шуканий корінь. За новий відрізок локалізації $[a,b]$ буде обране:

$$[a,b]=\begin{cases}[a,c], & \textit{якщо } f(a)f(c)<0,\\ [c,b], & \textit{якщо } f(a)f(c)>0.\end{cases}$$

Для нового відрізку локалізації кореня проводимо аналогічні дії: проводимо хорду, находимо точку перетину з віссю абцис та отримуємо нову ітерацію для даного випадку $x^{(1)}$ (рис. 6.4 *б*).

Виведемо розрахунковий вираз для методу хорд. Нехай $[a,b]$ – поточний відрізок локалізації на k-му кроці ($k=0,1,2,\ldots$). Записуємо рівняння прямої, що описує хорду:

$$\frac{y-f(a)}{f(b)-f(a)}=\frac{x-a}{b-a}.$$



а) *б)*

Рисунок 6. 4 – Геометрична інтерпретація розв'язку рівнянь методом хорд

Значення $x=c$, для якого $y=0$, тобто точка перетину хорди з віссю абсцис визначається з виразу:

$$c=a-f(a)\frac{b-a}{f(b)-f(a)}.$$

Обчислимо значення $f(c)$. Геометрично $f(c)$ – довжина перпендикуляра до осі *Ox*, проведеного з точки *c* до кривої $f(x)$. Якщо $f(c)<0$, то знайшли більш вузький інтервал існування кореня $[c,b]$, оскільки знаки $f(c)$ і $f(a)$ збігаються. Тепер корінь знаходиться у середині відрізка $[c,b]$. Враховуючи, що *c* – чергова ітерація, приходимо до розрахункової формули методу хорд:

$$x^{(k)}=a-f(a)\frac{b-a}{f(b)-f(a)}.$$

Для загального випадку

1. Якщо мають місце варіанти $f(b)\cdot f''(x)>0$ на відрізку $[a,b]$, то наближені значення коренів $x^{(1)},x^{(2)},\ldots,x^{(k)}$ будуть знаходитися усередині відрізків $[c,b]$, тобто

нерухомим кінцем відрізка $[a,b]$ буде кінець b, а наближені значення коренів будуть знаходитися за формулою

$$x^{(k+1)} = x^{(k)} - f\left(x^{(k)}\right)\frac{b - x^{(k)}}{f(b) - f\left(x^{(k)}\right)}, \tag{6.8}$$

при цьому початкове наближення $x^{(0)} = a$.

2. Якщо мають місце варіанти $f(a)\cdot f''(x) > 0$ на відрізку $[a,b]$, то наближені значення $x^{(1)}, x^{(2)}, \ldots, x^{(k)}$ коренів будуть знаходитися усередині відрізків $[a,c]$, тобто нерухомим кінцем відрізка $[a,b]$ буде кінець a, а наближені значення коренів будуть знаходитися за формулою

$$x^{(k+1)} = x^{(k)} - f\left(x^{(k)}\right)\frac{x^{(k)} - a}{f\left(x^{(k)}\right) - f(a)}, \tag{6.9}$$

при цьому початкове наближення $x^{(0)} = b$.

Вибір формул (6.8) чи (6.9) можна здійснити, користуючись простим *правилом*: *нерухомим кінцем відрізка* є той, для якого знак функції збігається зі знаком другої похідної:

$$f(x)\cdot f''(x) > 0. \tag{6.10}$$

Процес послідовного наближення до кореня слід продовжувати доти, поки не буде виконана умова $\left|x^{(k+1)} - x^{(k)}\right| \le \varepsilon$, де ε – задана точність; $x^{(k+1)}$ і $x^{(k)}$ – наближення, отримані на $k+1$-му та k-му кроках. При цьому уточнене значення кореня приймається $x^* = x^{(k+1)} \pm \varepsilon$. Або поки не буде досягнуто точності виконання наближеної рівності $f(c) \approx 0 : \left|f(c)\right| < \varepsilon$.

Приклад 6.2 Знайти корінь x^* рівняння $x^3 - 12.2x^2 + 7.45x + 42 = 0$ методом хорд на відрізку [10, 12].

Розв'язок:

Обчислюємо значення функції на кінцях відрізка:

$f(10) = -103{,}5;\ f(12) = 102{,}6.$ $f''(x) = 6x - 24{,}4 > 0$, отже $f''(x) > 0$, тому за нульове наближення приймаємо $x^{(0)} = 10$ та обчислення проводимо за формулою (6.8).

$$x^{(1)} = x^{(0)} - f\left(x^{(0)}\right)\frac{b - x^{(0)}}{f(b) - f\left(x^{(0)}\right)} = 10 + 103.5\frac{(12-10)}{102.6 - 103.5} \approx 11.$$

Перевіряємо умову (6.10), $f(11) = -21.25$, отже істинний корінь розташований в інтервалі [11, 12].

Повторюючи процес для визначення другого наближення кореня, одержимо $x^{(2)} = 11{,}17$, для якого значення функції $f(11.17) = -3.55$. Тепер корінь знаходиться в інтервалі [11,17; 12]. Нарешті, третє наближення дає нам $x^{(3)} = 11{,}2$, для якого $f(11.2) = 0$.

Відповідь: точне значення кореня x, знайдено на третьому кроці та становить $x^{(3)} = 11{,}2$.

6.6 Метод дотичних (метод Ньютона)

Метод Ньютона є найбільш популярним з чисельних методів розв'язку нелінійних рівнянь. Він швидко збігається (має квадратичну збіжність). Та припускає різні модифікації. Проте цей метод є ефективним за досить жорстких умов:

1) Існування другої похідної $f(x)$ на множині $G = \{a \le x \le b\}$

2) Перша похідна $f(x) \neq 0$ для всіх $x \in G$

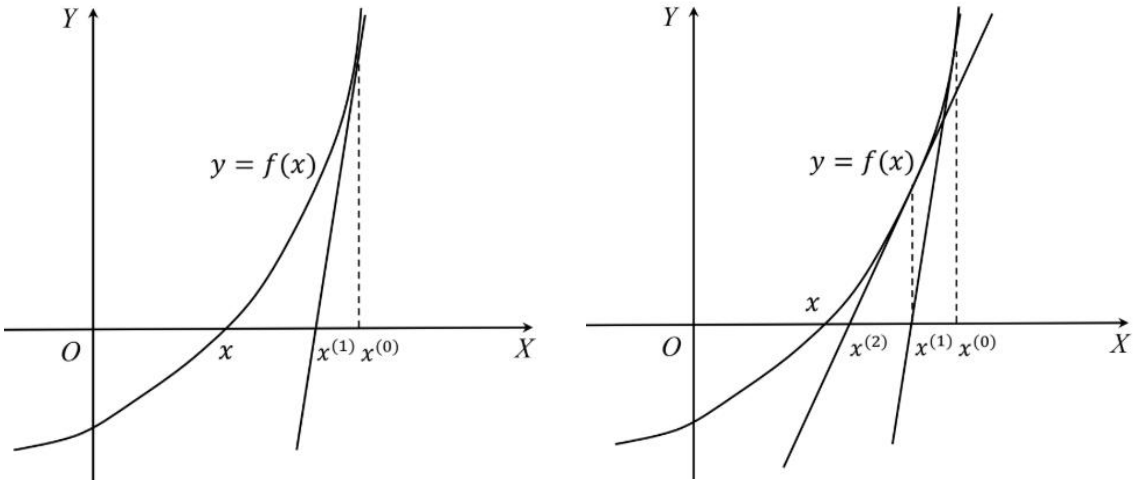
3) Знакопостійність першої та другої похідних для всіх $x \in G$

Тому цей метод бажано застосовувати разом з іншими методами, наприклад методом половинного поділу для досягнення діапазону, де вказані умови починають виконуватись.

Геометрично метод Ньютона еквівалентний заміні невеликої дуги кривої $y = f(x)$ дотичною, проведеною до деякої точки кривої з ацисою $x^{(0)}$ (рис. 6.5 *а*). Точка перетину цієї дотичної з віссю абсцис дає перше наближення $x^{(1)}$ кореня *x*. Далі у точці з абцисою $x^{(1)}$ проводимо ще одну дотичну та на перетині її з віссю абцис знаходимо нове наближення до корення $x^{(2)}$ (рис. 6.5 *б*), та повторюємо описані дії до виконання умов зупину або до моменту знайденя точного рішення.

Виведемо розрахунковий вираз методу дотичних. Нехай отримане значення k -го наближення $x^{(k)}$. Рівняння дотичної до кривої $y=f(x)$ у точці $\left(x^{(k)};f\left(x^{(k)}\right)\right)$ має вид:

$$y=f\left(x^{(k)}\right)+f'\left(x^{(k)}\right)\left(x-x^{(k)}\right)$$



а) *б)*

Рисунок 6. 5 – Геометрична інтерпретація розв'язку рівнянь методом дотичних

Знаходимо наступне наближення $x^{(k+1)}$. Вважаючи $y=0$, $x=x^{(k+1)}$ знаходимо абсцису $x^{(k+1)}$ точки перетину цієї дотичної з віссю *Ox*:

$$f\left(x^{(k)}\right)+f'\left(x^{(k)}\right)\left(x^{(k+1)}-x^{(k)}\right)=0\Rightarrow$$

$$\Rightarrow x^{(k+1)}=x^{(k)}-\frac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)},\quad k=1,2,3,... \qquad (6.11)$$

Для коректності методу необхідна нерівність ную похідної у деякому околі кореня. Вибір нульового наближення кореня $x^{(0)}$ здійснюється таким чином:

якщо $f(a)\cdot f''(x)>0$ на $[a,b]$, то $x^{(0)}=a$;

якщо $f(b)\cdot f''(x)>0$ на $[a,b]$, то $x^{(0)}=b$.

Послідовність $\left\{x^{(k)}\right\}$ не обов'язково збігається до кореня. Чим більше чисельне значення похідної $f'(x)$ в околі даного кореня, тим менша поправка, яку необхідно враховувати в k -му наближенні. Тому метод Ньютона особливо зручно застосовувати тоді, коли в околі даного кореня *графік функції має велику крутизну.*

Якщо чисельне значення похідної біля кореня мале, то поправки будуть великими і процес уточнення кореня може виявитися тривалим. Якщо крива поблизу точки перетину з віссю *Ox майже горизонтальна, то застосовувати метод Ньютона не рекомендується.* Наступна теорема вказує на достатні умови збіжності методу дотичних (Ньютона).

ТЕОРЕМА про збіжність методу Ньютона. Нехай x^* – простий корінь рівняння $f(x)=0$, в деякій околиці якого функція двічі безперервно диференційована. Тоді знайдеться така мала σ-околиця кореня x^*, що при довільному виборі початкового наближення $x^{(0)}$ з цієї околиці ітераційна послідовність $\left\{x^{(k)}\right\}$ методу Ньютона не виходить за межі околиці та справедлива оцінка $\left|x^{(k+1)}-x^*\right| \le C\left|x^{(k)}-x^*\right|^2$, де $k \ge 0$.

Доведення:

Перерисуємо вираз (6.11) для знаходження наступного наближення методом Ньютона $x^{(k+1)}=x^{(k)}-\dfrac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)}$ наступним чином:

$$x^{(k+1)}=x^{(k)}-\frac{f\left(x^{(k)}\right)-f\left(x^*\right)}{f'\left(x^{(k)}\right)}$$

Такий запис є правомірним, оскільки $f\left(x^*\right)=0$. Розкладаємо $f\left(x^*\right)$ в околиці попереднього наближення та підставляємо у (6.11):

$$f\left(x^*\right)=f\left(x^{(k)}\right)+f'\left(x^{(k)}\right)\left(x^*-x^{(k)}\right)+f''\left(x^{(k)}\right)\frac{1}{2}\left(x^*-x^{(k)}\right)^2$$

$$x^{(k+1)}=x^{(k)}+\frac{f\left(x^{(k)}\right)-f\left(x^{(k)}\right)-f'\left(x^{(k)}\right)\left(x^*-x^{(k)}\right)-f''\left(x^{(k)}\right)\frac{1}{2}\left(x^*-x^{(k)}\right)^2}{f'\left(x^{(k)}\right)},$$

$$x^{(k+1)}=x^{(k)}+\left(x^*-x^{(k)}\right)-\frac{1}{2}\frac{f''\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)}\left(x^*-x^{(k)}\right)^2,$$

звідси, кінцевий результат: $x^{(k+1)} - x^* = \frac{1}{2}\frac{f''\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)}\left(x^{(k)} - x^*\right)^2$.

Справедливими є наступні оцінки *C*: якщо функція задовольняє умовам $\left|f'(x)\right| \geq m_1 > 0,\ \left|f''(x)\right| \leq M_2 < \infty,\ x \in [a,b];$ тоді послідовність $\left\{x^{(k)}\right\}$ методу Ньютона повністю належить відрізку $[a,b]$ та сходиться до кореня x^*. При цьому справедливими є нерівності:

$$\left|x - x^{(k+1)}\right| \leq \frac{M_2}{2m_1}\left|x - x^{(k)}\right|^2, \qquad \left|x - x^{(k+1)}\right| \leq \frac{M_2}{2m_1}\left|x^{(k+1)} - x^{(k)}\right|^2.$$

З першої нерівності випливає, що метод Ньютона має другий порядок збіжності (про що й і вказано у теоремі), тобто він найшвидший з вивчених нами. А друга дозволяє оцінювати похибку нової ітерації. Тому *критерієм зупинки методу Ньютона* може служити виконання нерівності:

$$\left|x^{(k+1)} - x^{(k)}\right|^2 < \varepsilon_1, \quad \varepsilon_1 = \frac{2m_1}{M_2}\varepsilon. \tag{6.12}$$

Потрібно зауважити, якщо похідна $f'(x)$ мало змінюється на відрізку $[a,b]$, то для спрощення обчислень можна використовувати формулу

$$x^{(k+1)} = x^{(k)} - \frac{f\left(x^{(k+1)}\right)}{f'\left(x^{(0)}\right)},$$

тобто значення похідної в початковій точці достатньо обчислити один раз. Геометрично це означає, що дотичні в точках $\left(x^{(k)}; f\left(x^{(k)}\right)\right)$ замінюються прямими, паралельними дотичній, проведеної до кривої $y = f(x)$ у точці $\left(x^{(0)}; f\left(x^{(0)}\right)\right)$.

6.7 Комбінований метод хорд і дотичних

З розрахункової формули методу (6.11) видно, що у загальному випадку метод Ньютона потребує обчислення похідної при обчисленні кожного наступного наближення. Це може суттєво зменшити його реальну ефективність у сенсі затрат машинного часу. Тому існують інші комбіновані методи.

Методи хорд і дотичних дають наближення кореня з різних сторін (більше або менше істинного значення кореня), тому їх часто застосовують у сполученні один з одним, і уточнення кореня відбувається швидше.

Якщо $f'(x)\cdot f''(x)>0$, то метод хорд дає наближення кореня з недостачею, а метод дотичних – з надлишком. Якщо ж $f'(x)\cdot f''(x)<0$, то методом хорд одержуємо значення кореня з надлишком, а методом дотичних – із недостачею.

Проте в усіх випадках істинне значення кореня замкнене між наближеними значеннями коренів, що утворюються за методом хорд і методом дотичних. Проілюструємо можливі випадки застосування комбінованого методу на рис. 6.5.

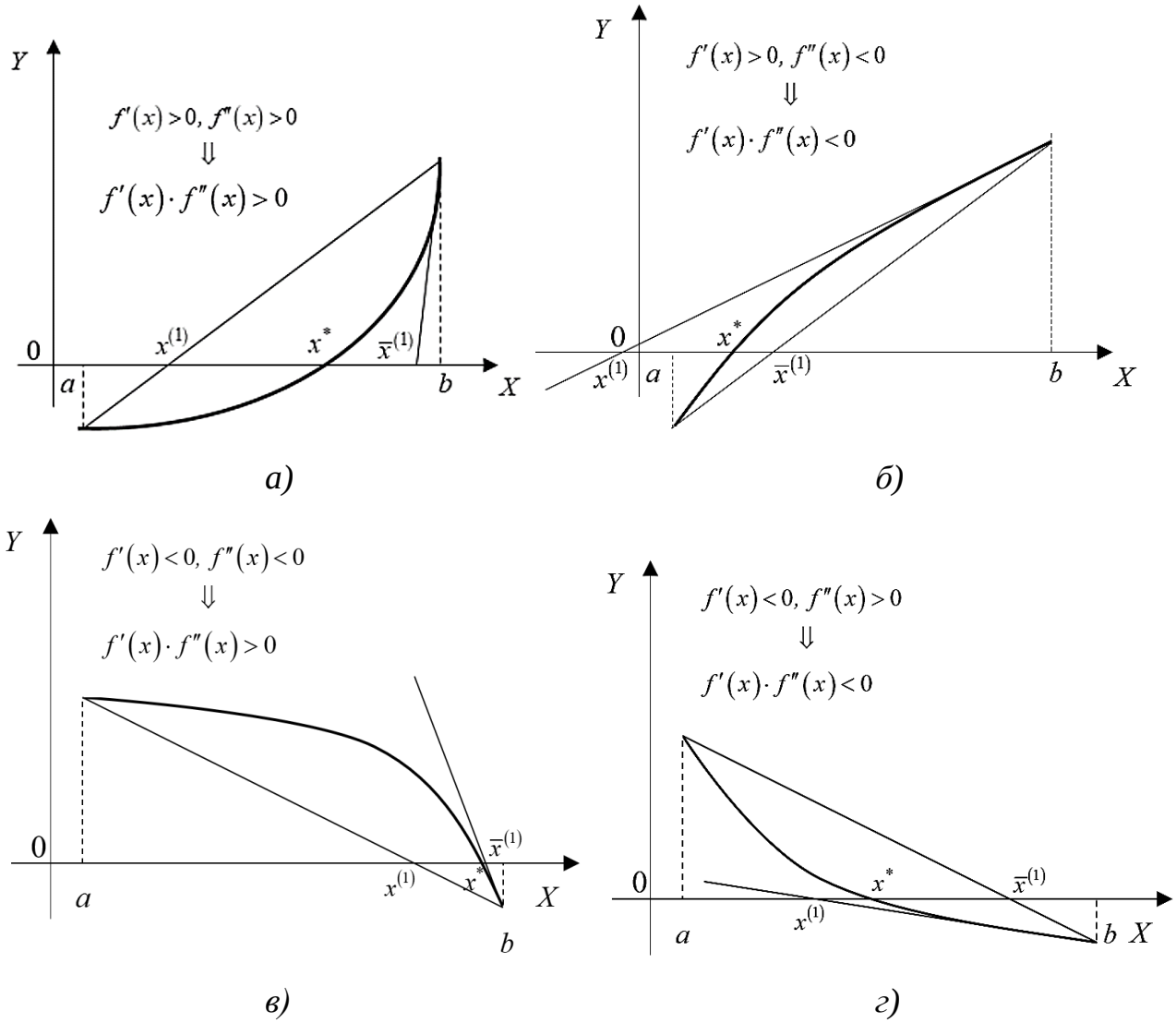


Рисунок 6. 5 – Геометрична інтерпретація розв’язку рівнянь комбінованими методами

Нехай $x^{(k+1)}$ і $\overline{x}^{(k+1)}$ - наближені значення кореня з недостачею та з надлишком відповідно.

1. Якщо $f'(x)\cdot f''(x)>0$ на $[a,b]$ (див. рис. 6.5 *а, в*), то

$$x^{(k+1)}=x^{(k)}-f\left(x^{(k)}\right)\frac{\overline{x}^{(k)}-x^{(k)}}{f\left(\overline{x}^{(k)}\right)-f\left(x^{(k)}\right)},\qquad \overline{x}^{(k+1)}=x^{(k)}-\frac{f\left(\overline{x}^{(k)}\right)}{f'\left(\overline{x}^{(k)}\right)}, \tag{6.13}$$

при цьому $x^{(0)}=a,\quad \overline{x}^{(0)}=b$.

2. Якщо $f'(x)\cdot f''(x)<0$ на $[a,b]$, (див. рис. 6.5 *б, г*) то

$$x^{(k+1)}=x^{(k)}-\frac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)},\quad \overline{x}^{(k+1)}=\overline{x}^{(k)}-f\left(x^{(k)}\right)\frac{\overline{x}^{(k)}-x^{(k)}}{f\left(\overline{x}^{(k)}\right)-f\left(x^{(k)}\right)}. \tag{6.14}$$

Процес обчислень припиняється, як тільки буде виконуватися нерівність

$$\left|x^{(k+1)}-\overline{x}^{(k+1)}\right|\le\varepsilon. \tag{6.15}$$

Значення кореня, який є уточненим, буде $x^{*}=\frac{1}{2}\left(x^{(k+1)}-\overline{x}^{(k+1)}\right)$.

Приклад 6.3 Методом дотичних уточнити до $\varepsilon=0{,}001$ корінь x^{*} рівняння $x^3-3x^2-3=0$, розташований на відрізку [-2. 75, -2. 5].

Розв'язок

За умовою $f(x)=x^3-3x^2-3$. Визначаємо другу похідну $f''(x)$: $f'(x)=3x^2+6x;\quad f''(x)=6x+6$. Таким чином, $f(-2.75)\cdot f''(x)>0$, тому $x^{(0)}=-2.75$. Визначаємо значення першої похідної у точці $x^{(0)}$: $f'\left(x^{(0)}\right)=f'(-2.75)=6.1875$.

Для зручності подальші обчислення зводимо в таблицю 3.2.

k	$x^{(k)}$	$x_{(k)}^2$	$x_{(k)}{}^3$	$3x_{(k)}{}^2$	$f\left(x^{(k)}\right)$	$-\frac{f\left(x^{(k)}\right)}{6.1875}$
0	-2.75	-20.797	7.5625	22.6875	-1.111	0.179
1	-2.571	-16.994	6.6100	19.8300	-0.164	0.026
2	-2.545	-16.484	6.4770	19.431	-0.053	0.008
3	-2.537	-16.329	6.4364	19.309	0.020	0.003
4	-2.534	-16.271	6.4212	19.2636	0.007	0.001
5	-2.533					

Відповідь: $x^{*}=-2.533\pm0.001$.

Питання для самоперевірки:

1. Яке рівняння називається нелінійним? З яких етапів складається чисельне рішення нелінійного рівняння?
2. Сформулюйте задачу чисельного рішення нелінійного рівняння
3. У чому полягає локалізація кореня? Чому вона необхідна?
4. Які методи можна застосовувати для локалізації кореня?
5. Що таке ітераційне уточнення кореня?
6. Опишіть алгоритм методу половинного поділу. Який критерій зупинки ітераційного процесу?
7. Як оцінюють кількість ітерацій для досягнення заданої точності кореня?
8. Які переваги і недоліки має метод половинного поділу?
9. Опишіть алгоритм методу простої ітерації
10. За яких умов ітераційна послідовність сходиться до кореня?
11. Сформулюйте і доведіть умови збіжності і оцінки похибок методу простої ітерації. Який критерій зупинки ітераційного процесу?
12. Що таке лінійна і надлінійна збіжності? Яку збіжність має метод простої ітерації?
13. Опишіть алгоритм методу хорд, виведіть розрахункову формулу
14. Виведіть оцінку похибки методу хорд. Який критерій зупинки? Яку швидкість збіжності має метод хорд?
15. Перерахуйте переваги і недоліки методу хорд
16. Опишіть алгоритм методу Ньютона (дотичних). Виведіть розрахункову формулу
17. Сформулюйте умови збіжності і оцінки похибок методу Ньютона. Які критерії зупину?
18. Яку швидкість збіжності має метод Ньютона?
19. Перерахуйте переваги і недоліки методу Ньютона.
20. Які комбіновані методи пошуку кореня нелінійного рівняння застосовують на практиці? В чому різниця від вже відомих методів?

Тема 3.2 Чисельне рішення систем нелінійних рівнянь

ЛЕКЦІЯ 8 Рішення систем нелінійних рівнянь числовими методами. Ітераційний метод рішення систем рівнянь. Теорема про достатню умову збіжності

Навчальні питання:

8.1 Формулювання задачі. Проблеми локалізації розв'язку

8.2 Метод Ньютона як метод лінеаризації вихідної задачі

8.3 Метод послідовних наближень (ітерацій) для системи нелінійних рівнянь

8.4 Вплив похибок округлення

На попередніх лекціях були розглянуті різні методи рішення *нелінійних рівнянь*. Їх можна поділити на двоточкові методи, які використовують локалізацію кореня, і одноточкові методи, які локалізацію кореня не використовують.

Двоточкові методи: метод поділу відрізка навпіл; метод січних (хорд).

Одноточкові методи: метод простої ітерації (МПІ); метод Ньютона.

Далі перейдемо до багатомірної ситуації та розглянемо *нелінійні системи*.

На відміну від *систем лінійних* рівнянь для систем *нелінійних* рівнянь не відомі прямі методи рішення. Лише в окремих випадках систему можна вирішити безпосередньо. Наприклад, для системи з двох рівнянь іноді вдається виразити одне невідоме через інше і таким чином звести задачу до вирішення одного нелінійного рівняння відносно одного невідомого. Тому ітераційні методи для нелінійних систем набувають особливої актуальності.

8.1Формулювання задачі. Проблеми локалізації розв'язку

Під час обговорення методів обчислення рішення *нелінійних систем*, в основному, будемо слідувати плану, реалізованому для відшукання коренів *нелінійних рівнянь*. Отже, розглядається система

$$f_i\left(x_1,\, x_2,\ldots,\, x_n\right)=0,\quad i=1,2,\ldots,n \tag{8.1}$$

де f_i – задані нелінійні функції. Серед них можуть бути й лінійні функції, проте нелінійність хоча б однієї призводить до нелінійної системи рівнянь.

Систему (8.1) також можна записати у векторній формі

$$\bar{F}(\bar{x}) = 0, \tag{8.2}$$

де $\bar{x}$ – вектор невідомих величин, а $\bar{F}(\bar{x})$ – вектор-функція, що визначає структуру рівнянь системи:

$$\bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \qquad \bar{F}(\bar{x}) = \begin{pmatrix} f_1(x_1,\dots,x_n) \\ f_2(x_1,\dots,x_n) \\ \dots \\ f_n(x_1,\dots,x_n) \end{pmatrix}$$

Рішенням системи (8.2) називається вектор $\bar{x}$, при підстановці якого у систему вона обертається у тотожність. Точне рішення у такому випадку позначається як $\bar{x}$, тоді наближене будемо позначати $\bar{x}^*$. Будемо вважати, що система (8.1) має принаймні одне рішення і що рішення $\bar{x}^*$, належить відомій околиці, що виявлена на стадії локалізації рішення.

Зауваження: Як і у випадку одного рівняння, *локалізація рішення* здійснюється або на основі фізичних міркувань (якщо задача має фізичний зміст), або із залученням методів математичного аналізу. Наприклад, для системи двох рівнянь можна приблизно оцінити місце розташування коренів, аналізуючи на площині (x_1, x_2) поведінку кривих, що задаються рівняннями $f_1(x_1, x_2) = 0$, $f_2(x_1, x_2) = 0$.

Нульове наближення $x^{(0)}$ у випадку двох змінних можна знайти графічно, побудувавши на площині криві $f_1(x_1, x_2) = 0$, $f_2(x_1, x_2) = 0$ і знайти точки їх перетину. Для трьох і більше змінних задовільних способів підбору нульових наближень немає. Отже двоточкові методи, розглянуті у попередніх лекціях, для пошуку коренів нелінійних систем застосувати неможливо.

Локалізація кореня у багатомірному випадку набагато складніша ніж на двомірній числовій осі. Локалізація дуже важлива: від неї багато в чому залежить успіх рішення, тобто збіжність ітераційного процесу та його швидкість.

Локалізація здійснюється дослідженням тепер уже багатовимірної функції $F(X)$. Методи дослідження найрізноманітніші, вони сильно залежать від функції, тому неможливо дати загальний універсальний алгоритм.

Формально *задача пошуку розв'язку системи рівнянь* може бути записана так само, як і задача пошуку кореня одного рівняння $f(x)=0$. Розглянемо найпоширеніші методи.

8.2 Метод Ньютона як метод лінеаризації вихідної задачі

Нехай вектор-рішення $\bar{x}$ ізольовано в деякій області локалізації, також в цій області присутні наближення $\bar{x}^{(k)}$ до $\bar{x}$. Припускаючи, що функції f_i безперервно диференційовані по всіх аргументах в деякій області, що містить $\bar{x}$ та $\bar{x}^{(k)}$, розкладемо f_i в ряди Тейлора в точці $\bar{x}$ в околиці $\bar{x}^{(k)}$. Зауважимо, що для других частинних похідних достатньо вимагати їхнього існування. Якщо значення $\bar{x}$ та $\bar{x}^{(k)}$ достатньо близькі одне до одного, то таку систему можна лінеаризувати (взяти лише перші два доданки розкладення у ряд Тейлора, нехтуючи доданками другого порядку та вище).

Отже, записуємо розклад (8.1) у ряд Тейлора:

$$f_i\left(x_1, x_2, ..., x_n\right) \approx f_i\left(x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\right) + \sum_{j=1}^{n} \left.\frac{\partial f_i}{\partial x_i}\right|^{(0)} \left(x_j - x_j^{(0)}\right) =$$

$$= f_i^{(0)} + \sum_{j=1}^{n} \left.\frac{\partial f_i}{\partial x_i}\right|^{(0)} \left(x_j - x_j^{(0)}\right) 0, \quad i = 1,2,...,n,$$

Напишемо ці рівності як точні (не забуваючи, що вони насправді наближені) і отримаємо систему рівнянь для знаходження компонент x_j точного вектора рішення $\bar{x}$. В матричній формі вона має вигляд:

$$\bar{F}\left(\bar{x}^{(k)}\right) + \bar{F}'\left(\bar{x}^{(k)}\right)\left(\bar{x} - \bar{x}^{(k)}\right) = \bar{0}\,, \tag{8.3}$$

де $\overline{F}'$ – функціональна матриця частинних похідних вектор-функції $\overline{F}$, яка називається матрицею Якобі, частіше її позначають як J – (оператор матриця Якобі $J_{i,j}=\frac{\partial f_i}{\partial x_j}$) з елементами:

$$\overline{F}' = \begin{pmatrix} \frac{\partial f_1(\vec{x})}{\partial x_1} & \frac{\partial f_1(\vec{x})}{\partial x_2} & \dots & \frac{\partial f_1(\vec{x})}{\partial x_n} \\ \frac{\partial f_2(\vec{x})}{\partial x_1} & \frac{\partial f_2(\vec{x})}{\partial x_2} & \dots & \frac{\partial f_2(\vec{x})}{\partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f_n(\vec{x})}{\partial x_1} & \frac{\partial f_n(\vec{x})}{\partial x_2} & \dots & \frac{\partial f_n(\vec{x})}{\partial x_n} \end{pmatrix}$$

Вирішуючи систему (8.3) відносно $\bar{x}$, отримуємо вектор $\bar{x}$:

$$\bar{x} = \bar{x}^{(k)} - \left(\overline{F}'\left(\bar{x}^{(k)}\right)\right)^{-1} \cdot \overline{F}\left(\bar{x}^{(k)}\right)$$

Тепер згадаємо, що $\bar{x}$ визначено насправді приблизно, і цей вектор приймаємо за наступне наближення до кореня $\bar{x}^{(k+1)}$:

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \left(\overline{F}'\left(\bar{x}^{(k)}\right)\right)^{-1} \cdot \overline{F}\left(\bar{x}^{(k)}\right) \qquad (8.4)$$

або $$\bar{x}^{(k+1)} = \bar{x}^{(k)} - J^{-1} \cdot \overline{F}\left(\bar{x}^{(k)}\right)$$

Таким чином, отримуємо схему для уточнення розв'язку системи рівнянь, аналогічну методу Ньютона для випадку одного рівняння. Зрозуміло, що для здійсненності методу необхідно, щоб всі матриці Якобі J були невиродженими.

Оскільки обчислювати *матрицю Якобі* J на кожному кроці досить складно, то зазвичай її елементи обчислюють *наближено* або використовують одні й ті ж значення на декількох кроках.

Формула (8.4) передбачає використання трудомісткої операції обернення матриці, тому безпосереднє її використання для обчислення $\bar{x}^{(k+1)}$ не завжди доцільно. Перетворимо (8.3) наступним чином.

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - J^{-}F\left(x^{(k)}\right) \;\Rightarrow\; \bar{x}^{(k+1)} - \bar{x}^{(k)} = -J^{-}F\left(x^{(k)}\right) \Rightarrow \left(\bar{x}^{(k+1)} - \bar{x}^{(k)}\right) \cdot J = -F\left(x^{(k)}\right)$$

$$\Delta\bar{x}^{(k+1)} \cdot J = -F\left(x^{(k)}\right) \tag{8.5}$$

де $\Delta\bar{x}^{(k+1)}$ – поправка поточного наближення.

Тобто отримуємо еквівалентну системі (8.4) систему лінійних алгебраїчних рівнянь, розв'язавши яку, будь-яким прийнятним методом, обчислюємо чергове наближення до кореня $\bar{x}^{(k+1)} = \bar{x}^{(k)} + \bar{\Delta}x^{(k+1)}, \quad k = 0,1,\ldots$.

Алгоритм методу Ньютона для розв'язку нелінійних систем:

1. Задати вектор початкових наближень $\bar{x}^{(0)}$ та мале додане число ε (точність). Покласти $k = 0$.

2. Розв'язати систему лінійних алгебраїчних рівнянь відносно поправки $\Delta x^{(\kappa)}$, а саме: $\Delta\bar{x}^{(k+1)} \cdot J = -F\left(x^{(k)}\right)$.

3. Обчислити наступне наближення $\bar{x}^{(k+1)} = \bar{x}^{(k)} + \Delta\bar{x}^{(k+1)}, \quad k = 0,1,\ldots$

4. Якщо $\Delta^{(k+1)} = max_i \left|x_i^{(k+1)} - x_i^{(k)}\right| \le \varepsilon$, завершити процес та покласти $\bar{x}^* \approx \bar{x}^{(k+1)}$, якщо $\Delta^{(k+1)} > \varepsilon$, то покласти $k = k+1$ та перейти до пункту №2.

Спрощений метод Ньютона для нелінійних систем

При реалізації методу за визначенням (8.5) необхідно обертати матрицю, замість цього тепер на кожному кроці розв'язують лінійну систему. Це теж трудомістка операція, але менш «ризикована» з точки зору похибки, ніж обернення матриці. Проте, все одно існує необхідність кожен раз рахувати матрицю Якобі. Ці недоліки усуваються різними модифікаціями. Розглянемо одну з них. Вона називається спрощеним методом Ньютона.

Існує велика кількість модифікацій методу Ньютона, що дозволяють в тих чи інших ситуаціях знизити його трудомісткість або уникнути необхідності обчислення похідних.

Розглянемо спрощений метод Ньютона. У цьому методі на відміну від методу Ньютона зворотну матрицю шукають тільки один раз в початковій точці $\bar{x}^{(0)}$. Замінимо в розрахункових формулах (8.5) методу Ньютона $\Delta x^{(k+1)} \cdot J = -F\left(x^{(k)}\right)$ матрицю *J*, що залежить від різних значень *k*, постійною матрицею $A = F'\left(\bar{x}^{(0)}\right)$.

В результаті отримуємо формули спрощеного методу Ньютона:

$$A\Delta\vec{x}^{(k+1)} = -F\left(\vec{x}^{(k)}\right),$$
$$\vec{x}^{(k+1)} = \vec{x}^{(k)} + \Delta\vec{x}^{(k+1)}, \quad k = 0,1,2... \tag{8.6}$$

Цей метод ще називають методом Бройдена (аналог методу січних для одного нелінійного рівняння). Метод сходиться зі швидкістю геометричної прогресії, якщо початкове наближення $\bar{x}^{(0)}$ вибрано досить близьким до рішення $\bar{x}$. Тобто метод забезпечує лінійну швидкість збіжності. Причому знаменник прогресії тим менше, чим ближче $\bar{x}^{(0)}$ до $\bar{x}$. Тому для досягнення потрібної точності за менше число кроків треба якомога якісніше локалізувати рішення.

Як *критерій зупинки* обирають виконання нерівностей $\left|x_i^{(k+1)} - x_i^{(k)}\right| \le \varepsilon$.

Обчислювальні витрати можуть виявитися меншими за рахунок того, що обчислення матриці Якобі проводиться тільки один раз і, згідно (8.6), багаторазово вирішуються системи лінійних алгебраїчних рівнянь з однією і тією ж фіксованою матрицею *A* і різними правими частинами. Це означає, що при вирішенні системи (8.6) методом Гауса можливе застосування *LU* -розкладення матриці *A*, яке різко зменшує число операцій, необхідних для обчислення $\Delta\bar{x}^{(k+1)}$.

Достатні умови збіжності методу Ньютона (для систем).

Як завжди постає питання швидкості збіжності до точного розв'язку застосованих методів. Метод Ньютона не виключення.

ТЕОРЕМА про збіжність для методу Ньютона для нелінійних систем:

Нехай в деякій околиці рішення $\bar{x}$ системи функції f_i (i = 1, ..., n) двічі безперервно диференційовані по всіх аргументах та матриця Якобі J не вироджена. Тоді знайдеться такий малий δ-окіл рішення $\bar{x}$, що при довільному виборі початкового наближення $\bar{x}^{(0)}$,у якому ітераційна послідовність методу Ньютона не виходить за межі цього околу, збігається до $\bar{x}$ та вірною є оцінка:

$$\bar{x}^{(k+1)} = \left\|\bar{x}^{(k+1)} - \bar{x}^{(k)}\right\| \le \frac{1}{\delta}\left\|\bar{x}^{(k)} - \bar{x}\right\|^2 \tag{8.7}$$

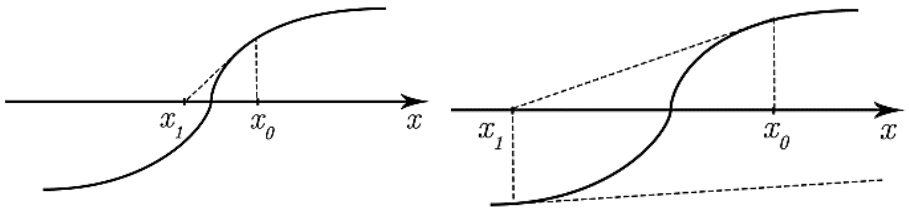
Під δ-околом рішення $\bar{x}$ тут розуміється або куля $S_{\bar{x}}(\delta)$ з центром в $\bar{x}$ радіуса δ, або куб $P_{\bar{x}}(\delta; ...; \delta)$ з центром в $\bar{x}$ розмірами 2δ по всіх осях. В теоремі не

зазначається яким має бути значення або оцінка δ. Тому вказаний окіл треба знаходити дослідженням конкретного рівняння.

Доведення квадратичної збіжності для систем нелінійних рівнянь аналогічне одномірному випадку (див. попередню лекцію).

Отже за аналогією з одновимірним випадком можна виписати відповідні умови збіжності методу Ньютона для систем. Ці умови досить складні через складність знаходження δ-околу, і на практиці перевірити їх не виявляється можливим. Тому частіше використовуються *емпіричні правила*. *Якщо метод не сходиться за декілька ітерацій (5-7), то це означає, що початкове наближення вибрано невдало.*

Можна зробити оцінку кількості необхідних ітерацій виходячи з вимоги, щоб точність була менше деякої наперед заданої величини ε. Якщо q істотно менше одиниці, то для збіжності потрібна невелика кількість ітерацій (наприклад різниця була 10^{-4} а стане 10^{-8}). Проте, метод Ньютона має також і суттєві недоліки. По-перше він досить чутливий до вибору початкової точки. Продемонструвати цю чутливість можна на прикладі графіка функції типу арктангенс



а) *б*)

Рисунок 8.1 – до пояснення недоліків методу Ньютона

Якщо точка для початкового наближення обрана вдало (рис. 8.2 *а*), то метод сходиться досить швидко. Однак, якщо точка вибрана так, як показано на рис.8.2 *б*, то процес буде розбіжним. Таким чином, видно, що метод Ньютона чутливий до вибору початкового наближення x^0.

Також до *недоліків* методу Ньютона слід віднести:

- відсутність глобальної збіжності для багатьох завдань (необхідно перевіряти збіжність для кожного рівняння системи);

- необхідність обчислення матриці Якобі на кожній ітерації;

- необхідність вирішення на кожній ітерації системи лінійних рівнянь, яка може бути погано обумовленою.

Перевагою методу є квадратична збіжність за хорошого початкового наближення та за умови невиродженості матриці Якобі.

Наступний метод – це МПІ для систем нелінійних рівнянь що по суті являє узагальнення однойменного методу для одного рівняння.

8.3 Метод послідовних наближень (ітерацій) для системи нелінійних рівнянь

Метод простої ітерації можна застосовувати до систем, що заздалегідь приведені до вигляду :

$$\begin{array}{l} x_1 = \varphi_1\left(x_1, x_2, ..., x_n\right) \\ \ldots\ldots\ldots \quad \ldots\ldots \\ x_n = \varphi_n\left(x_1, x_2, ..., x_n\right) \end{array}, \tag{8.7}$$

або у векторній формі $\bar{x} = \Phi\left(\bar{x}\right)$ (8.8)

З вектор-функцією $\Phi\left(\bar{x}\right) = \begin{pmatrix} \varphi_1\left(x_1, x_2, ..., x_n\right) \\ \varphi_2\left(x_1, x_2, ..., x_n\right) \\ \ldots\ldots\ldots \quad \ldots\ldots \\ \varphi_n\left(x_1, x_2, ..., x_n\right) \end{pmatrix}$

Припустимо, що $\bar{x}^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)}\right)$ – початкове наближення. У векторній формі наступні наближення в методі простої знаходять як:

$$\bar{x}^{(k+1)} = \Phi\left(\bar{x}^{(k)}\right), \quad k = 0,1,2,... \tag{8.9}$$

при заданому початковому наближенні $\bar{x}^{(0)}$.

Якщо послідовність векторів $\bar{x}^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}\right)$ збігається до вектора $\bar{x}^* = \left(x_1^*, x_2^*, ..., x_n^*\right)$, а функції $\varphi_i\left(x\right)$, $i = 1,2,...,n$ – безперервні, то вектор $\bar{x}^*$ є розв'язком системи.

При деяких умовах послідовність (8.9) може збігатися до вирішення (8.1). Ці умови формуються у наступній теоремі.

ТЕОРЕМА про достатню умову збіжності методу простих ітерацій:

Нехай $\bar{x}$ – шуканий розв'язок задачі (8.1) або, що теж саме, задачі (8.8). Якщо у околиці, якій належить розв'язок $\delta = \{\|x - \bar{x}\| \le r\}$ вектор-функція $\Phi(\bar{x})$ задовольняє умові $\|\Phi'(\bar{x})\| < q$ та, якщо при цьому $q < 1$, то послідовність векторів $\bar{x}^{(k)}$, обчислюваних згідно (8.9) з $\bar{x}^{(0)} \in \delta$, збігається до розв'язку, тобто $\|\bar{x}^{(k)} - \bar{x}\| \underset{k\to\infty}{\to} 0$

При цьому має місце оцінка $\|\bar{x}^{(k)} - \bar{x}\| \le q^k \|\bar{x}^{(0)} - \bar{x}\|$.

Зауваження. В умовах теореми вірною є також апостеріорна оцінка похибки

$$\|\bar{x}^{(k+1)} - \bar{x}^*\| \le \frac{q}{1-q} \|\bar{x}^{(k+1)} - \bar{x}^{(k)}\|, \tag{8.10}$$

яка зручна для формування критерію закінчення ітерацій, якщо відома величина q. У ряді випадків при наявності вдало обраного початкового наближення $x^{(0)}$ можна, вважаючи, що $q \approx q_0 = \|\Phi'(\bar{x}^{(0)})\|$, використовувати наступний практичний критерій закінчення ітераційного процесу: $\|\bar{x}^{(k+1)} - \bar{x}^{(k)}\| \le \varepsilon_1 = \frac{1-q_0}{q_0}\varepsilon$, де ε – задана точність.

У разі одного нелінійного рівняння (в скалярному випадку) у якості достатнього критерію збіжності ми перевіряли в розглянутій околиці рішення виконання нерівності $|\varphi'(x)| < 1$. Узагальнимо цей критерій стосовно багатовимірної ситуації. Запишемо вираз для різниці між компонентами k-го наближення, обчисленого за методом (8.9), і точним (шуканим) рішенням даної задачі:

$$\begin{aligned} x_i^{(k+1)} - x_i^* &= \varphi_i\left(x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}\right) - \varphi_i\left(x_1^*, x_2^*, ..., x_n^*\right) = \\ &= \sum_{j=1}^{n} \left.\frac{\partial \varphi_i}{\partial x_j}\right|^{(0)} \left(x_j^{(k)} - x_j^*\right), \quad i = 1, 2, ..., n, \end{aligned} \tag{8.11}$$

(за теоремою про середнє для функції багатьох змінних).

Введемо в розгляд матрицю Якобі для вектору-функції $\Phi(X)$:

$$M_{\varphi} = \begin{pmatrix} \dfrac{\partial \varphi_1}{\partial x_1} & \cdots & \dfrac{\partial \varphi_1}{\partial x_n} \\ & \cdots\cdots & \\ \dfrac{\partial \varphi_n}{\partial x_1} & \cdots & \dfrac{\partial \varphi_n}{\partial x_n} \end{pmatrix}$$

З її допомогою сукупність виписаних співвідношень можна записати у векторній формі: $\left\|\bar{x}^{(k)} - \bar{x}^*\right\| \le M \left\|\bar{x}^{(k+1)} - \bar{x}^*\right\|$. Нехай M – мажоріруюча матриця з елементами $m_{ij} = \underset{\delta^*}{max} \left|\partial \varphi_i / \partial x_j\right|$ такими, що $\left\|M_{\varphi}\right\| \le \left\|M\right\|$, тоді

$$\left\|\bar{x}^{(k)} - \bar{x}^*\right\| \le \left\|M\varphi\right\| \left\|\bar{x}^{(k-1)} - \bar{x}^*\right\| \le \left\|M\right\| \left\|\bar{x}^{(k-1)} - \bar{x}^*\right\|,$$

І якщо $$\left\|M\right\| \le q < 1, \qquad (8.12)$$

то $\left\|\bar{x}^{(k)} - \bar{x}^*\right\| \le q \left\|\bar{x}^{(k-1)} - \bar{x}^*\right\|$ і послідовні наближення збігаються до рішення $\bar{x}^*$.

Саме умова (8.12) залучається при аналізі збіжності конкретних ітераційних схем.

Приклад 8.1 Чи буде сходитися метод простої ітерації для системи:

$$\begin{cases} x^3 + y^2 - 6x + 3 = 0, \\ x^3 - y^2 - 6y + 2 = 0. \end{cases}$$

для кореня, який належить області $0 \le x \le 1,\ 0 \le y \le 1$.

Розв'язок:

Переписуємо задану систему у наступному виді:

$$\begin{cases} x = \dfrac{1}{6}\left(x^3 + y^2 + 3\right), \\ y = \left(x^3 - y^2 + 2\right). \end{cases}$$

Скористаємось для перевірки збіжності умовою (8.12). В даному випадку

$$M_{\varphi} = \begin{pmatrix} \dfrac{x^2}{2} & \dfrac{y}{3} \\ \dfrac{x^2}{2} & -\dfrac{y}{3} \end{pmatrix}, \qquad M = \begin{pmatrix} \dfrac{1}{2} & \dfrac{1}{3} \\ \dfrac{1}{2} & \dfrac{1}{3} \end{pmatrix}.$$

Звідси $\left\|M\right\| = \dfrac{1}{2} + \dfrac{1}{3} = \dfrac{5}{6} < 1$, отже умова збіжності виконана.

Відповідь: розв'язок системи методом простої ітерації буде сходитися до вірного.

Алгоритм МПІ для розв'язку систем нелінійних рівнянь

1. Задати початкове наближення $\bar{x}^{(0)}$ та мале додане число ε. Покласти $k = 0.$

2. Обчислити $\bar{x}^{(k+1)}$ за формулою (8.9).

3. Якщо $\Delta^{(k+1)} = max_i \left| x_i^{(k+1)} - x_i^{(k)} \right| \le \varepsilon$, завершити процес та покласти $\bar{x}^* \approx \bar{x}^{(k+1)}$, якщо $\Delta^{(k+1)} > \varepsilon$, то покласти $k = k+1$ та перейти до пункту 2 (для знаходження Δ можна використовувати різні норми векторів).

Ітераційний процес, який реалізується згідно (8.9), відповідає паралельному наближенню, тому що для обчислення $k+1$-го наближення всіх невідомих враховуються обчислені раніше їх *k*-те наближення.

Метод простої ітерації прямо веде до розв'язку та легко програмується, проте він має два суттєвих *недоліки*:

- жорстка умова збіжності – виконання нерівності (8.12). Та навіть якщо (8.12) виконується, це не гарантує збіжність у випадках, коли точка початкового наближення обрана невдало (далеко від істинного розв'язку);
- повільна збіжність (менше ніж у метода Ньютона).

У деяких випадках для прискорення збіжності корисно використовувати наступний *аналог методу Зейделя для нелінійних систем*:

$$
\begin{aligned}
x_1^{(k+1)} &= \varphi_1\left(x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, ..., x_n^{(k)}\right), \\
x_2^{(k+1)} &= \varphi_2\left(x_1^{(k+1)}, x_2^{(k)}, x_3^{(k)}, ..., x_n^{(k)}\right), \\
x_3^{(k+1)} &= \varphi_3\left(x_1^{(k+1)}, x_2^{(k+1)}, x_3^{(k)}, ..., x_n^{(k)}\right), \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
x_n^{(k+1)} &= \varphi_1\left(x_1^{(k+1)}, x_2^{(k+1)}, x_3^{(k+1)}, ..., x_n^{(k)}\right).
\end{aligned}
\qquad (8.13)
$$

Переваги цього підходу полягають в можливості використання методів розв'язання рівнянь з одним невідомим і під час відсутності необхідності приведення системи рівнянь до вигляду, зручного для ітерацій. Зазначений метод сходиться, якщо для матриці Якобі виконується умова діагональної переваги.

Метод Зейделя для систем призначений для вирішення систем, записаних у формі (8.13). Цей метод є модифікацією методу простих ітерацій, де після завдання

початкового наближення замість паралельної ітерації проводиться послідовна, причому на кожній ітерації в кожне наступне рівняння підставляються значення невідомих, отриманих з попередніх рівнянь.

8.4 Вплив неусувних похибок на обчислювані наближення

Вплив похибок округлення на обчислювані наближення аналізується так само, як для ітераційних схем рішення систем лінійних рівнянь.

Нехай

$\bar{x}^{(k)} = \Phi\left(\bar{x}^{(k-1)}\right)$ – ідеальний ітераційний процес, при реалізації якого не приймаються в розрахунок похибки округлень при виконанні елементарних арифметичних операцій.

$x^{(k)} = \Phi\left(x^{(k-1)}\right) + \delta^{(k)}$ – «реальний» обчислювальний процес, відповідний до даного методу ітерацій. Тут $x^{(k)}$ – реально обчислювані наближення, $\delta^{(k)}$ – сукупний вплив похибок округлень в межах одного кроку (пов'язаного з переходом від *k – 1*-го наближення до *k*-го). Віднімаючи з другого співвідношення перше, отримаємо

$$\tilde{x}^{(k)} - \bar{x}^{(k)} = \Phi\left(\tilde{x}^{(k-1)}\right) - \Phi\left(\bar{x}^{(k-1)}\right) + \delta^{(k)}$$

Припускаючи, що виконана достатня умова збіжності, оцінимо неусувну похибку *k*-го наближення по якій-небудь нормі:

$$\left\|x^{(k)} - \bar{x}^{(k)}\right\| = \left\|\Phi\left(x^{(k-1)}\right) - \Phi\left(\bar{x}^{(k-1)}\right)\right\| + \left\|\delta^{(k)}\right\| \le q\left\|x^{(k-1)} - \bar{x}^{(k-1)}\right\| + \left\|\delta^{(k)}\right\| \le$$

$$\le q\left(q\left\|x^{(k-2)} - \bar{x}^{(k-2)}\right\| + \left\|\delta^{(k-1)}\right\|\right) + \left\|\delta^{(k)}\right\| \le \ldots \le q^k\left\|x^{(0)} - \bar{x}^{(0)}\right\| + q^{k-1}\left\|\delta^{(1)}\right\| + q^{k-2}\left\|\delta^{(2)}\right\| + \left\|\delta^{(k)}\right\|,$$

$x^{(k)}(0) = \bar{x}^{(k)}(0)$, оскільки початкове наближення не обчислюється, а задається.

Вводячи в розгляд максимальну похибку, накопичену за рахунок округлення на одному кроці ітерацій $\delta = \max_i \left\|\delta^{(i)}\right\|$ отримаємо остаточну оцінку похибки на *k*-му кроці:

$$\left\|x^{(k)} - \bar{x}^{(k)}\right\| \le \delta\left(q^{k-1} + q^{k-2} + \ldots + 1\right) = \frac{\delta\left(1 - q^k\right)}{1 - q} \le \frac{\delta}{1 - q}$$

звідки випливає, що вплив похибки округлень помірний, якщо $q << 1$.

Питання для самоперевірки:

1. Що таке нелінійна система? Сформулюйте завдання наближеного рішення нелінійної системи
2. З яких етапів складається чисельне рішення нелінійної системи? У чому відмінності локалізації рішення в багатовимірному випадку від локалізації кореня в одновимірному?
3. Які методи можна використовувати для локалізації рішення?
4. Виведіть розрахункову формулу методу Ньютона. За яких умов вона вірна?
5. У чому полягають недоліки та переваги методу Ньютона? Чим викликана необхідність його модифікацій?
6. Сформулюйте умови збіжності і оцінку похибки методу Ньютона. Яку швидкість збіжності має метод? Який критерій можна застосовувати для зупинки?
7. Які переваги та недоліки має метод Ньютона для нелінійних систем?
8. Виведіть розрахункову формулу методу простої ітерації
9. Виведіть достатню умова збіжності і оцінку похибки методу. За яких умов на функцію вони вірні?
10. З якою швидкістю сходиться метод простої ітерації?
11. Який критерій зупинки можна використовувати для методу?
12. Опишіть аналог методу Зейделя для нелінійних систем.
13. Який вплив неусувних похибок на обчислювальні наближення методу простої ітерації?

РОЗДІЛ 4. НАБЛИЖЕННЯ ФУНКЦІЇ

Тема 4.1 Аналітичне наближення функцій. Інтерполяція.

ЛЕКЦІЯ 9 Задача наближеного обчислення функції. Задача інтерполяції. Поліноміальна (алгебраїчна) інтерполяція та її похибка.

Навчальні питання:

9.1 Задача наближеного обчислення функції

9.2 Інтерполяція

9.3 Наближення багаточленами Тейлора

9.4 Поліноміальна інтерполяція

9.5 Похибка поліноміальної інтерполяції

9.6 Інтерполяційний багаточлен Лагранжа

Задачі, що пов'язані з наближенням функції зустрічаються у багатьох прикладних задачах фізики та інших науках. Постановки задачі апроксимації різноманітні. Деякі розглянемо далі.

9.1 Задача наближеного обчислення функції

В даній темі розглядається задача наближеного обчислення скалярної функції $y = f(x)$ одного аргументу. Задачі, які вимагають наближення функції наступні:

1) *Прискорення часу обчислення функції*. Припустимо, що розв'язок деякої задачі пов'язаний з багаторазовим обчисленням функції $f(x)$, що задана громіздким виразом. Тоді природньо буде замінити таку функцію більш простою з меншим числом обчислення $g(x)$, щоб виконувалось співвідношення $|f(x)-g(x)|<\varepsilon$, де ε – точність наближення.

2) *Економія пам'яті ЕОМ*. Припускаємо, що функцію $f(x)$ задано своїми значеннями у точках x_i, $i=1, 2,...,n$, на інтервалі $a<x<b$. За великого числа n зберігання всієї таблиці у пам'яті ЕОМ може виявитись неможливим. Тоді постає

задача наближення точково заданої функції $f(x_i)$ близькою їй функцією, що залежить від невеликої кількості параметрів $g(a_1, a_2, ..., a_n)$. При цьому в пам'яті машини зберігаються значення цих параметрів, а значення функції $f(x_i)$ обчислюються наближено.

3) *Пошук закономірностей за експериментальними даними*. Результати експериментів зазвичай подаються у таблицях. Людина, що виконує експеримент припускає, що отримана залежність є реалізацією якогось фізичного закону $g(x,a)$ з невідомим параметром a. Тоді виникає задача визначення $g(x,a)$ та a, які найкраще наближають експериментальні дані.

У основі рішення перелічених завдань лежить підміна однієї функції іншою функцією. При реалізації цього процесу неминуче потрібно отримати відповіді на наступні питання.

1) Що відомо про функцію $f(x)$. Чи задана вона аналітично або у табличній формі, яка міра її гладкості і чи існують її похідні. Як розташовані точки в частині області визначення функції, де відомі її значення.

2) До якому класу функцій повинна належати функція $g(x)$.

3) Що розуміти під визначенням "близькість" між $f(x)$ і $g(x)$.

Загальний підхід до вирішення задач наближення полягає в наступному. Складається таблиця значень функції (коли її не вказано). Наприклад, функція, що важко обчислюється, розраховується в декількох точках, для яких можливе обчислення з прийнятною точністю. Далі по таблиці будується функція, що просто обчислюється (наприклад, поліном), яка вважається наближено рівною $f(x)$. Такий підхід називається аналітичним наближенням функції по табличних даних. Є два його різновиди: інтерполяція та апроксимація. Розглянемо спочатку інтерполяцію.

9.2 Інтерполяція

Одним з основних типів точкової апроксимації (наближення) є інтерполяція.

Нехай задана сукупність вузлів інтерполяції або сітка на деякому відрізку $[a,\ b]$. У простому випадку сітка - рівномірна, тобто відстань між сусідніми вузлами однакова.

Сукупність вузлів $\{x_i\}_{i=0}^{N}$, $x_i = a + n\tau$, $\tau = (b-a)/N$, $x \in [a,\ b]$

Сітковою проекцією функції $y(x)$ на $[a,\ b]$, є таблиця $y_i = \{y(x_i)\}_{i=0}^{N}$. Цю таблицю задає оператор обмеження на сітку або рестрикція R (від англійського restriction). Нехай функція $y = f(x)$ задана таблицею 9.1.

Таблиця 9.1 – значення функції у точках, вузли інтерполяції

i	x_i	y_i
0	x_0	y_0
1	x_1	y_1
….	….	….
n	x_n	y_n

Задача полягає в наближеному обчисленні її значення в точці x, яка не співпадає ні з одним з окремих значень $x_0, x_1, x_2, ..., x_n$. Її рішення методом інтерполяції передбачає побудову по табличних даних функції $y = g(x)$, значення якої в точках $x_0, x_1, x_2, ..., x_n$ в точності збігаються зі $y_0, y_1, y_2, ..., y_n$. Функція $g(x)$ називається інтерполюючою (або інтерполянтом), значення $x_0, x_1, x_2, ..., x_n$ – вузлами інтерполяції. Поза вузлів вважається, що $f(x) \approx g(x)$.

Таким чином, задача інтерполяції сформулюється в такий спосіб: по таблиці 1 побудувати функцію $y = g(x)$, що задовольняє умовам $g(x_i) = y_i$, $i = 1, 2, ..., n$, які називаються *умовами інтерполяції*.

Природно, бажано, щоб $g(x)$ володіла хорошими обчислювальними властивостями і до того ж поза вузлів приблизно повторювала значення $f(x)$. Тоді можна вважати, що $f(x) \approx g(x)$ при $x \neq x_i$.

Визначення: І*нтерполяція* – такий спосіб наближення при якому функція яку наближають $f(x)$ та функція, якою наближають $g(x)$ приймають однакові значення у заданих точках (вузлах).

У такій постановка задача інтерполяції не є коректною, оскільки вона має безліч рішень. Для усунення (або хоча б зменшення) цієї неоднозначності вводять клас функцій, в якому шукається рішення (клас інтерполюючих функцій). При належному виборі класу інтерполюючих функцій задача може мати єдине рішення.

Основна мета інтерполяції – знайти швидкий (економний) алгоритм обчислення функції $f(x)$ для значень x які не задано у таблиці.

9.3 Наближення багаточленами Тейлора

Для наближення функції широко використовуються наступні класи функцій: багаточлени (поліноми), тригонометричні функції, показові функції. Особливо часто використовуються багаточлени. Для того, щоб задати багаточлен потрібно лише задати кінцеву кількість його коефіцієнтів, яка дорівнює ступеню багаточлена плюс одиниця. Значення багаточлена швидко обчислюється, він легко диференціюється та інтегрується. Через вказані переваги алгебраїчні багаточлени широко використовують для наближення функцій. Функцію, яка визначена на відрізку $[a,b]$ та має на цьому відрізку неперервні похідні до *k*–го порядку включно, можна наблизити в околі точки x_0 рядом Тейлора. Обчислюють такі функції як подання їх багаточленами Тейлора або Маклорена з обмеженою кількістю членів ряду. При цьому кількість доданків, що беруть учать у розрахунку, визначають необхідною точністю знаходження відповідної функції.

Багаточленом Тейлора порядку n будемо називати функцію:

$$Q_n(x) = f(x_0) + \sum_{k=1}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k, \qquad (9.1)$$

Багаточлен Тейлора має таку важливу властивість: усі його похідні порядку n включно збігаються з відповідними похідними функції $f(x)$ в околі точки x_0.

Похибку, що виникає під час заміни багаточленом Тейлора, визначають залишковим членом ряду Тейлора

$$R(x) = f(x) - Q_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}, \tag{9.2}$$

де ξ – деяка точка, що лежить між x та x_0, $x \neq x_0$. У наслідок зробленого припущення, що похідна функції $f^{(n+1)}$ неперервна на відрізку, можна стверджувати, що вона обмежена на цьому відрізку, тобто існує таке кінцеве число

$$\left| f(x) - Q_n(x) \right| = \frac{M_{n+1}}{(n+1)!}\left| x - x_0 \right|^{n+1}. \tag{9.3}$$

Якщо у якості класу інтерполюючих функцій обрати сукупність алгебраїчних поліномів, то отримаємо задачу поліноміальної інтерполяції. Перейдемо до цього більш загального випадку.

9.4 Поліноміальна інтерполяція

Обираємо у якості класу інтерполюючих функцій сукупність алгебраїчних поліномів (багаточленів) ступеня m, тобто припустимо

$$g(x) = P_m(x) = a_0 + a_1 x + a_2 x^2 + ... + a_m x^m, \tag{9.4}$$

тоді отримуємо задачу поліноміальної інтерполяції. Вона має важливе практичне значення, так як алгебраїчні поліноми є досить просто обчислювані функції з хорошими властивостями. На алгебраїчній інтерполяції ґрунтуються багато методів чисельного диференціювання та інтегрування.

Багаточлен $P_m(x)$ повністю визначається своїми коефіцієнтами. Тому *задача поліноміальної інтерполяції* формулюється так:

По таблиці 1 знайти коефіцієнти $a_0, a_1, a_2, ..., a_m$ багаточлена, що задовольняє умовам інтерполяції

$$P_m(x_i) = y_i, \quad i = 0,1,...,n, \tag{9.5}$$

цей багаточлен називається інтерполяційним.

Вимагаючи, щоб у кожному вузлі таблиці 1 значення поліному співпадало з заданим значенням функції, отримаємо замкнену систему лінійних рівнянь відносно невідомих коефіцієнтів $a_0, a_1, a_2, ..., a_m$.

$$a_0 + a_1 x + a_2 x^2 + ... + a_n x^n = y_i, \quad i = 0,1,...,n. \tag{9.6}$$

Особливо цікавим є випадок, коли ступінь інтерполяційного багаточлена на одиницю менше числа вузлів $m = n$, оскільки задача поліноміальної інтерполяції за цієї умови має єдине рішення.

ТЕОРЕМА 1 про єдиний інтерполяційний багаточлен.

Якщо функція задана таблицею 1, вузли якої є попарно різні, то для неї існує єдиний інтерполяційний багаточлен P_n ступеня n.

Доведення: Інтерполяційний поліном $P_n(x) = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n$ повністю визначається своїми коефіцієнтами $a_0, a_1, a_2, ..., a_n$, що знаходять з умов інтерполяції (9.5). При $m = n$ вони являють собою систему лінійних рівнянь з $n+1$ невідомим:

$$\begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 + ... + a_n x_0^n = y_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + ... + a_n x_1^n = y_1 \\ \dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots \\ a_0 + a_1 x_n + a_2 x_n^2 + ... + a_n x_n^n = y_n \end{cases} \tag{9.7}$$

Визначником цієї системи є відомий з курсів лінійної алгебри визначник Вандермонда для системи неспівпадаючих точок:

$$\Delta = \begin{vmatrix} 1 & x_0 & x_0^2 & x_0^n \\ 1 & x_1 & x_1^2 & x_1^n \\ \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & x_n^n \end{vmatrix} = \prod_{0 \le j < i < n}^{n} \left(x_i - x_j \right) \neq 0$$

Отже шуканий поліном існує і єдиний, якщо координати табличних точок попарно різні. Звідси система (9.7) має єдиний розв'язок, який визначають коефіцієнти поліному P_n.

Зауваження:

- Ступінь інтерполяційного багаточлена може бути менше n, якщо кілька старших коефіцієнтів нульові.
- При $m > n$ задача має безліч рішень, а при $m < n$ може взагалі не мати рішень.

Далі будуть розглянуті методи побудови цього єдиного інтерполяційного багаточлена. Але спочатку досліджуємо дуже важливе питання, що стосується кожної обчислювальної задачі - оцінку похибки рішення.

9.5 Похибка поліноміальної інтерполяції

Для табличної функції $f(x)$ та інтерполюючого багаточлена $P_n(x)$ можна по аналогії з поліномами Тейлора записати рівність

$$f(x) = P_n(x) + R(x),$$

де $R(x)$ - залишковий член інтерполяції в точці x. Зрозуміло, що її похибка є модуль $R(x)$. Оцінка похибки ґрунтується на наступній теоремі про подання залишкового члена.

Теорема про подання залишкового члена інтерполяції (без доведення):
Нехай функція $f(x)$ $n+1$ разів неперервно диференційована на деякому відрізку $[a,b]$, що містить всі вузли $x_0, x_1, x_2, ..., x_n$. Тоді для будь-якого $x \in [a,b]$ існує таке ξ, що

$$R(x) = f(x) - P_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x), \qquad (9.8)$$

де $\omega_{n+1}(x) = (x-x_0)(x-x_1)...(x-x_n)$, $f^{(n+1)}$ - $n+1$-ша похідна. З формули (9.8) випливає оцінка похибки інтерполяції в точці x:

$$\Delta(P_n(x)) = |R(x)| = |f(x) - P_n(x)| = \left|\frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x)\right| \le \frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)| = \bar{\Delta}(P_n(x)), \qquad (9.9)$$

Оцінка $\bar{\Delta}(P_n(x))$, що залежить від x, називається оціночною функцією інтерполяції її точна верхня межа на відрізку $[a,b]$ є глобальною оцінкою похибки, тобто оцінка по всій таблиці. Величину $\bar{\Delta}(P_n(x))$ можливо застосовувати для оцінки похибки інтерполяції в точці або по всій таблиці. Однак її застосування ускладняється необхідністю обчислення константи M_{n+1}, адже про функції в загальному випадку невідомо нічого, крім таблиці 1. Для вирішення цієї проблеми доводиться залучати додаткові міркування, наприклад, геометричні або фізичні, все, що відомо про

вкожному конкретному випадку. На практиці найчастіше замість M_{n+1} доводиться використовувати її верхню оцінку, тобто таке число q_{n+1}, про яке відомо, що $M_{n+1} \le q_{n+1}$. Звичайно q_{n+1}, має бути якомога ближче до M_{n+1}, щоб оцінка похибки не була завищеною.

Приклад 9.1 Для функції $y = \sin 2x$, що задана таблицею

x_i	0,10	0,25	0,35	0,40	0,75
y_i	0,199	0,479	0,644	0,717	0,997

Знайти оціночну функцію похибки інтерполяції.

Розв'язок:

Скористаємось визначенням (9.9), для цього знаходимо $y^{(5)}(x) = 32\cos 2x$, $\left|y^{(5)}(x)\right| \le 32 = M_{(5)}$, звідси:

$$\overline{\Delta}\left(P_4(x)\right) = \frac{32}{5!}\left|(x-0,10)(x-0,25)(x-0,35)(x-0,40)(x-0,75)\right| =$$
$$= 0,267x^5 - 0,493x^4 + 0,334x^3 - 0,134x^2 + 0,0145x - 0,0007$$

Тепер для будь-якої проміжної точки, що входить до інтервалу визначення функції можливо оцінити похибку обчислення. Наприклад для $x = 0,15$ $\overline{\Delta}\left(P_4(0,15)\right) = 0.00004$.

9.6 Інтерполяційний багаточлен Лагранжа

Розглянемо метод побудови інтерполяційного багаточлена, існування і єдиність якого гарантує теорема 1. Функція $f(x)$ задана таблицею 1 з попарно незбіжними вузлами. Нехай шуканий многочлен має вигляд:

$$\begin{aligned} P_n(x) &= a_0(x-x_1)\ldots(x-x_n) + a_1(x-x_0)(x-x_2)\ldots(x-x_n) + \ldots \\ &+ a_i(x-x_0)\ldots(x-x_{i-1})\ldots(x-x_n) + \ldots + a_n(x-x_0)\ldots(x-x_{n-1}) \end{aligned} \qquad (9.10)$$

Таке представлення правомірне, оскільки (9.10) дійсно задає багаточлен ступеня n від x. Для знаходження коефіцієнтів $a_0, a_1, a_2, \ldots, a_n$, скористаємося умовами інтерполяції $P_m(x_i) = y_i$. Підстановка в (9.10) $x = x_0$ дає коефіцієнт a_0:

$$P_n(x_0) = y_0 \;\Rightarrow\; a_0(x_0 - x_1)(x_0 - x_2)\ldots(x_0 - x_n) = y_0 \Rightarrow$$
$$\Rightarrow a_0 = \frac{y_0}{(x_0 - x_1)(x_0 - x_2)\ldots(x_0 - x_n)},$$

Аналогічно отримуємо a_1:

$$P_n(x_1) = y_1 \;\Rightarrow\; a_1(x_1 - x_0)(x_1 - x_2)\ldots(x_1 - x_n) = y_1 \Rightarrow$$
$$\Rightarrow a_1 = \frac{y_1}{(x_1 - x_0)(x_1 - x_2)\ldots(x_1 - x_n)},$$

Загальна формула для a_i випливає з умови інтерполяції в точці x_i:

$$P_n(x_i) = y_i \;\Rightarrow\; a_i(x_i - x_0)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n) = y_i \Rightarrow$$
$$\Rightarrow a_i = \frac{y_i}{(x_i - x_0)\ldots(x_i - x_{i-1})(x_i - x_{i+1})\ldots(x_i - x_n)},$$

Підставляючи знайдені коефіцієнти і а в (9.10), маємо

$$P_n(x) = L_n(x) = y_0 \frac{(x - x_1)\ldots(x - x_n)}{(x_0 - x_1)\ldots(x_0 - x_n)} + y_1 \frac{(x - x_0)(x - x_2)\ldots(x - x_n)}{(x_1 - x_0)(x_1 - x_2)\ldots(x_1 - x_n)} + \ldots + \\ + y_n \frac{(x - x_0)\ldots(x - x_{n-1})}{(x_n - x_0)\ldots(x_n - x_{n-1})}. \qquad (9.11)$$

Це і є формула інтерполяційного багаточлена Лагранжа (його прийнято позначати буквою *L*). Її можна записати в скороченій формі:

$$L_n(x) = \sum_{i=0}^{n} y_i \cdot l_{i,i}(x), \qquad l_{i,i}(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}. \qquad (9.12)$$

Величини $l_{i,i}(x)$ називають множниками Лагранжа

Інтерполяційна формула Лагранжа просто виводиться, має зручну для обчислень структуру (наприклад, за допомогою таблиці), *недолік* ж її в тому, що при *додаванні нового вузла її всю треба перераховувати заново*. На закінчення наведемо формули Лагранжа для таблиць з двома і трьома вузлами.

При $n = 1$ (два вузла і перша ступінь багаточлена), то

$$L_1(x) = y_0 \frac{(x - x_1)}{(x_0 - x_1)} + y_1 \frac{(x - x_0)}{(x_1 - x_0)}. \qquad (9.13)$$

У цьому виразі неважко впізнати рівняння прямої, що проходить через дві точки. Також за допомогою (9.13) виконують лінійну інтерполяцію, коли криві

між двома точками замінюють прямими, що у сукупності створюють ламану, про цей вид інтерполяції будемо говорити у наступній лекції.

При $n=2$ (багаточлен Лагранжа для таблиці з трьома вузлами має другий ступінь) з (9.12) отримуємо рівняння параболи, що проходить через три точки:

$$L_2(x)=y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}+y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}+y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}. \qquad (9.14)$$

Приклад 9.2 Побудувати інтерполяційний поліном Лагранжа для функції $f(x)=\cos(2\pi x)$, по точкам (вузлам інтерполяції) $x_0=0$, $x_1=\frac{1}{6}$, $x_2=\frac{1}{3}$, $x_3=\frac{1}{2}$.

Розв'язок: За умовою, кількість вузлів складає 4, отже ступінь багаточлену буде - $n=3$. Знаходимо значення функції в вузлах:

$$f(x_0)=1,\ f(x_1)=0{,}5,\ f(x_2)=-0{,}5,\ f(x_3)=-1.$$

Користуючись виразом (9.12) отримуємо розрахункове співвідношення:

$$L_3(x)=y_0\frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}+y_1\frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}+$$
$$+y_2\frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}+y_3\frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}.$$

Звідси:

$$L_3(x)=\frac{\left(x-\frac{1}{6}\right)\left(x-\frac{1}{3}\right)\left(x-\frac{1}{2}\right)}{\left(-\frac{1}{6}\right)\cdot\left(-\frac{1}{3}\right)\cdot\left(-\frac{1}{2}\right)}+\frac{1}{2}\cdot\frac{x\cdot\left(x-\frac{1}{3}\right)\left(x-\frac{1}{2}\right)}{\left(\frac{1}{6}\right)\cdot\left(\frac{1}{6}\right)\cdot\left(-\frac{1}{3}\right)}-$$
$$-\frac{1}{2}\frac{x\cdot\left(x-\frac{1}{6}\right)\left(x-\frac{1}{2}\right)}{\frac{1}{3}\cdot\frac{1}{6}\cdot\left(-\frac{1}{2}\right)}-\frac{x\cdot\left(x-\frac{1}{6}\right)\left(x-\frac{1}{3}\right)}{\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{6}}=$$
$$=36x^3-27x^2-11x+1.$$

На рисунку 9.1 представлена побудова вихідної функції та знайденого наближаючого поліному Лагранжа 3-го ступеня. З представленої ілюстрації видно, що поліном Лагранжа (червона лінія на рис. 9.1) добре наближає вихідну функцію (блакитна лінія на рис. 9.1) лише на відрізку числової осі $[0;0{,}5]$, після відбувається суттєві відхилення наближаючого поліному від заданої функції, а отже і похибка

також суттєво збільшується. Тому використовувати знайдений наближаючий поліном Лагранжа на всій області існування вихідної функції не є доцільним.

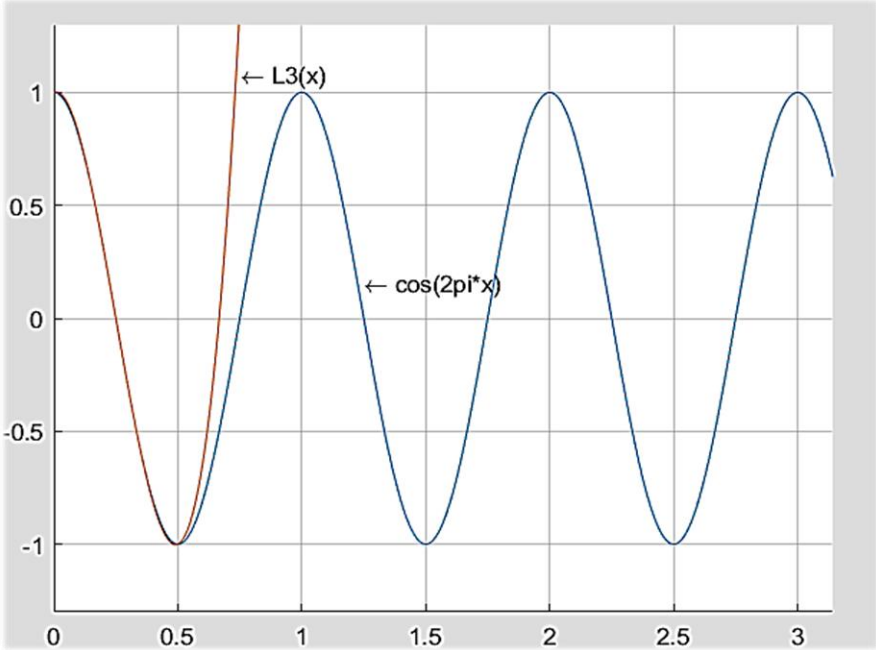


Рисунок 9.1 – ілюстрація розв’язку прикладу 9.2

Збільшуючи кількість точок, та тим самим ступінь інтерполюючого поліному можливо отримати менші похибки, проте доведено (феномен Рунге), що збільшення точок розбиття приводить до зворотної ситуації – до збільшення глобальної оцінки похибки. Також через те, що навіть для таблиці зі 100 вузлами можливо отримати поліноми 99-го ступеня, використання інтерполюючих поліномів Лагранжа на практиці для великих таблиць не є доцільним.

У подальшому будемо вивчати інші способи інтерполяції, які допомагають наближати функцію, не будуючи поліноми високого ступеня на всій області існування функції, а для кожної ділянки з певним кроком будувати свій поліном ступеня не вище 3-го.

Питання для самоперевірки:

1. Сформулюйте задачу наближеного обчислення функції. В яких випадках вона виникає як обчислювальна задача?

2. Що таке аналітичне наближення табличній функції? Які його різновиди вам відомі?
3. Сформулюйте задачу інтерполяції. Що таке інтерполююча функція, вузли інтерполяції?
4. Яким умовам повинна задовольняти інтерполююча функція?
5. Чому задача інтерполяції в загальній постановці некоректна?
6. Чому для задач наближення використовують багаточлени Тейлора? Яке значення має залишковий член багаточлену Тейлора? Як він визначається?
7. Сформулюйте задачу поліноміальної інтерполяції. За яких умов визначаються коефіцієнти інтерполяційного багаточлена?
8. У чому особливість задачі інтерполяції в окремому випадку, коли ступінь інтерполяційного багаточлена на одиницю менше числа вузлів?
9. За якими формулами можна оцінювати похибку інтерполяції в точці і на всьому відрізку між крайніми вузлами? У чому труднощі застосування цих формул?
10. Чому на практиці зазвичай не застосують інтерполяційні багаточлени високих ступенів? Як вирішують задачу аналітичного наближення великих таблиць?
11. Виведіть інтерполяційну формулу Лагранжа. Що таке множники Лагранжа?

Тема 4.2 Інтерполювання з різно-віддаленими вузлами

ЛЕКЦІЯ 10 Інтерполяція з застосуванням різниць. Інтерполяційні поліноми Ньютона

Навчальні питання:

10.1 Роздільні різниці та їх властивості

10.2 Інтерполяційні поліноми Ньютона з роздільними різницями

10.3 Кінцеві різниці та їх властивості

10.4 Інтерполяційні поліноми Ньютона з кінцевими різницями

10.5 Оцінка похибок інтерполяційних поліномів Ньютона

На минулій лекції почали вивчати питання пов’язані з наближенням таблично заданих функцій. Дано визначення апроксимації та інтерполяції. Розглянули приклад побудови інтерполяційного поліному Лагранжа.

Продовжуючи вивчення питань наближення табличнозаданої функції, виведемо інтерполяційні формули Ньютона *з розділеними і кінцевими різницями*. Їх вирішальна перевага в порівнянні з формулою Лагранжа полягає в "гнучкості" по відношенню до розширення таблиці. Також ці багаточлени мають локальну спрямованість, тобто вони дають краще наближення (при неповних таблицях різниць) в певних частинах таблиці, оскільки розроблялися саме для певних областей таблиць.

10.1 Роздільні різниці та їх властивості

Почнемо з визначення роздільних різниць. Нехай функція $y = f(x)$ задана таблицею з попарно незбіжними вузлами, які пронумеровані в довільному порядку (тобто не обов’язково у порядку зростання). Для такої таблиці можна обчислити величини, які називаються *розділеними різницями*. Вони визначаються рекурентно, починаючи з першого порядку.

Визначення. Розділеною різницею першого порядку функції f в вузлах x_i, x_{i+1} називається величина

$$f\left(x_i;x_{i+1}\right)=\frac{y_{i+1}-y_i}{x_{i+1}-x_i}=\frac{f\left(x_{i+1}\right)-f\left(x_i\right)}{x_{i+1}-x_i} \tag{10.1}$$

індекс i приймає значення від 0 до n-1. Таким чином, є n різниць першого порядку (по числу пар сусідніх вузлів). Формула (1) показує, що розділена різниця є дискретним аналогом першої похідної. Різниці порядків вище першого визначаються рекуррентно через попередні порядки.

Визначення. Розділеною різницею k-го порядку функції f в вузлах x_i, x_{i+1}, …, $x_{i+k,}$ називається величина

$$f\left(x_i;...;x_{i+k}\right)=\frac{f\left(x_{i+1};...;x_{i+k}\right)-f\left(x_i;...;x_{i+k-1}\right)}{x_{i+k}-x_i} \tag{10.2}$$

$i=0,1,\ldots,n-k$.

Якщо дано таблицю зі значеннями функцій у її вузлах, то можливо знайти роздільні різниці різного порядку. Наступне твердження дає формулу для безпосереднього обчислення розділених різниць по таблиці функції.

*Лема 1***.** Розділена різниця *k*-го порядку в вузлах x_i, x_{i+1}, …, $x_{i+k,}$ обчислюється за формулою

$$f\left(x_i;...;x_{i+k}\right)=\sum_{j=i}^{i+k}\frac{f\left(x_j\right)}{\left(x_j-x_i\right)\cdot\left(x_j-x_{i+1}\right)\cdots\left(x_j-x_{j-1}\right)\cdot\left(x_j-x_{j+1}\right)\cdots\left(x_j-x_{i+k}\right)} \tag{10.3}$$

Доведення: доведемо по індукції, тобто від часткового заключення перейдемо до загального. Нехай база індукції при ($k=1$):

$$f\left(x_i;x_{i+1}\right)=\frac{f\left(x_i\right)}{x_i-x_{i+1}}+\frac{f\left(x_{i+1}\right)}{x_{i+1}-x_i}=\frac{f\left(x_{i+1}\right)-f\left(x_i\right)}{x_{i+1}-x_i}$$

Прийшли до визначення роздільної різниці (10.1) першого порядку. Тобто для $k=1$ Лема доведена.

Якщо припустити, що (10.3) є вірною для будь-якого порядку $k\leq s$ при будь-яких $i=0,1,\ldots,n-k$. Далі прирівняти $k=s+1$ та виконати $s+1$ додавання, аналогічно отримаємо загальний випадок для (10.3).

Наслідок з Леми 1 є дуже важливим та полягає у наступному:

Розділені різниці є симетричними функціями своїх аргументів, тобто їх значення не залежать від перестановок вузлів, в яких вони визначені. Формула (10.3) не

застосовується для обчислення розділених різниць, оскільки набагато зручніше це робити за допомогою (10.1), (10.2) за таблицями. Вона має значення для теорії, зокрема, наслідок з неї буде використовуватися при виведенні інтерполяційних формул.

Для зручності введемо таблицю розділених різниць. Загальний вигляд трикутної таблиці розділених різниць наступний:

Таблиця 10.1 – таблиця розділених різниць

x_i	$f(x_i)$	$f(x_i;x_{i+1})$	$f(x_i;x_{i+1};x_{i+2})$	...	$f(x_0;\dots;x_n)$
x_0	$f(x_0)$				
		$f(x_0;x_1)$			
x_1	$f(x_1)$		$f(x_0;x_1;x_2)$		
		$f(x_1;x_2)$		$\ddots$	
x_2	$f(x_2)$				$f(x_0;\dots;x_n)$
				$\vdots$	
$\vdots$	$\vdots$		$\vdots$		
				$\iddots$	
x_{n-1}	$f(x_{n-1})$		$f(x_{n-2};x_{n-1};x_n)$		
		$f(x_{n-1};x_n)$			
x_n	$f(x_n)$				

По таблиці обчислювати різниці досить просто. На рис. 10.1 показано принцип знаходження різниці першого порядку. Рахуємо різницю кліток з x_{i+1} і x_i (показано стрілкою), це знаменник нашого дробу; рахуємо різницю значень у клітинках з y_{i+1} і y_i (показано стрілкою), пишемо в чисельник; частку пишемо в клітку різниці (темна клітка). Для наступної різниці спускаємося на два рядки і повторюємо ті ж операції.

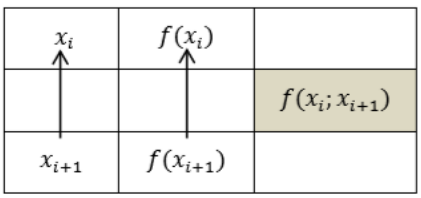


Рисунок 10.1 – до пояснення знаходження розділеної різниці першого порядку

Для розрахунку різниці *k*-го порядку $f(x_i,...,x_{i+k})$, виконують аналогічні дії.

Рахуємо різницю клітинок з x_{i+k} і x_i, пишемо в знаменник; рахуємо різницю кліток з сусідніми різницями попереднього (*k*-1) -го порядку, пишемо в чисельник; частку пишемо в клітинку різниці. Для наступної різниці спускаємося на два рядки і повторюємо ті ж операції.

10.2 Інтерполяційні поліноми Ньютона з роздільними різницями

З наслідку з Леми 1 витікає перевага використання роздільних різниць при інтерполяції. Отримаємо за допомогою роздільних різниць іншу, відмінну від поліномів Лагранжа форму запису інтерполяційного багаточлену.

$$f(x)-L_n(x)=f(x)-\sum_{i=0}^{n}y_i l_{n,i}(x)=f(x)-\sum_{i=0}^{n}f(x_i)\cdot\prod_{\substack{j=0\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}=\prod_{j=0}^{n}(x-x_j)\times$$

$$\times\left(\frac{f(x)}{\prod_{j=0}^{n}(x-x_j)}+\sum_{i=0}^{n}\frac{f(x_i)}{(x_i-x)\cdot\prod_{\substack{j=0\\ j\neq i}}^{n}(x_i-x_j)}\right).$$

Порівнюючи вираз в дужках з (10.3), переконуємося в тому, що він дорівнює $f(x,x_0,...,x_n)$. Тому:

$$f(x)-L_n(x)=f(x;x_0;...;x_n)\cdot\prod_{j=0}^{n}(x-x_j)=f(x;x_0;...;x_n)\omega_{n+1}(x). \qquad (10.4)$$

Представимо багаточлен Лагранжа у наступному вигляді:

$$L_n(x)=L_0(x)+(L_1(x)-L_0(x))+(L_2(x)-L_1(x))+\cdots+(L_n(x)-L_{n-1}(x)) \qquad (10.5)$$

Кожна різниця L_i-L_{i-1} являє собою багаточлен ступеня $i-1$, що обертається у нуль у точках $x_0,x_1,...,x_{i-1}$ (вузлах інтерполяції). Звідси поліном можна представити у вигляді

$$L_i(x)-L_{i-1}(x)=A_{i-1}(x-x_0)(x-x_1)\cdots(x-x_{i-1})=A_{i-1}\omega_i(x),$$

де A_{i-1} – постійний коефіцієнт. Якщо покласти у цій рівності $x=x_i$, то отримаємо:

$$L_i(x_i)-L_{i-1}(x_i)=f(x_i)-L_{i-1}(x_i)=A_{i-1}\omega_i(x_i).$$

З іншого боку з (10.4) при $n=i-1$, $x=x_i$ отримуємо f

$$f(x_i)-L_{i-1}(x_i)=f(x_i;x_0;...;x_{i-1})\omega_i(x_i)=f(x_0;...;x_{i-1};x_i)\omega_i(x_i).$$

Тут застосовано наслідок з Леми 1. Порівнюючи останні дві рівності, робимо висновок: $A_{i-1} = f(x_0,...,x_{i-1},x_i)$, тому

$$L_i(x) - L_{i-1}(x) = f(x_0;...;x_i)\omega_i(x) = f(x_0;...;x_i)(x-x_0)(x-x_1)\cdots(x-x_{i-1}).$$

Тепер підставимо останнє в представлення (10.5):

$$\begin{aligned} L_n(x) &= L_0(x) + f(x_0;x_1)\omega_1(x) + f(x_0;x_1;x_2)\omega_2(x) + \cdots + \\ &+ f(x_0;...;x_n)\omega_n(x) = f(x_0) + f(x_0;x_1)(x-x_0) + \\ &+ f(x_0;x_1;x_2)(x-x_0)(x-x_1) + \cdots + \\ &+ f(x_0;...;x_n)(x-x_0)\cdots(x-x_{n-1}). \end{aligned}$$

Отримана форма запису формули Лагранжа називається інтерполяційним багаточленом Ньютона з розділеними різницями першого виду:

$$\begin{aligned} &P_n^{(1)}(x) = f(x_0) + f(x_0;x_1)(x-x_0) + f(x_0;x_1;x_2)(x-x_0)(x-x_1) + ... \\ &+ f(x_0;...;x_n)(x-x_0)...(x-x_{n-1}) = f(x_0) + \sum_{i=1}^{n} f(x_0;...;x_i)(x-x_0)...(x-x_{i-1}). \end{aligned} \qquad (10.6)$$

По-іншому він називається багаточленом для *інтерполяції вперед*. Це пояснюється тим, що розділені різниці, які розраховують для нього (якщо рухатися від початку до кінця таблиці) утворюють верхню діагональ, номера захоплюваних ними вузлів зростають (тобто рух йде вперед по таблиці).

Для багаточлена Ньютона, як для інтерполяційного полінома ступеня n є справедливою теорема минулої лекції про залишковий член інтерполяції та витікаючі оцінки з неї.

За умови гладкості функції f на деякому відрізку $[a;b]$, що містить всі вузли інтерполяції, оціночна функція похибки $\bar{\Delta}\left(P_n^{(1)}(x)\right)$ багаточлена Ньютона наближено дорівнює:

$$\bar{\Delta}\left(P_n^{(1)}(x)\right) \approx \left| f(x;x_0;...;x_n)\omega_{n+1}(x) \right|, \qquad (10.7)$$

де $f(x,x_0,...,x_n)$ – розділена різниця (n+1)-го порядку, обчислена за таблицею 10.1, в яку точка інтерполяції x додана як вузол (вважається, що $f(x) \approx P_n^{(1)}(x)$). Згідно з наслідком з Леми 1 це можна робити в будь-якому місці таблиці. Зазвичай x ставлять до першого або після останнього вузла, так як в цьому випадку досить розрахувати тільки один рядок. Формула (10.7) дуже зручна для наближеної оцінки

похибки, оскільки немає необхідності знаходити значення похідної (як за умови теореми).

Інтерполяційним багаточленом Ньютона з розділеними різницями *другого виду* називається

$$P_n^{(1)}(x) = f(x_n) + f(x_{n-1}; x_n)(x - x_n) + f(x_{n-2}; x_{n-1}; x_n)(x - x_n)(x - x_{n-1}) + \ldots$$
$$+ f(x_0; \ldots; x_n)(x - x_n)\ldots(x - x_1) = f(x_n) + \sum_{i=1}^{n} f(x_{n-i}; \ldots; x_n)(x - x_n)\ldots(x - x_{n-i+1}) \quad (10.8)$$

Виводиться вона аналогічно (10.6). По-іншому ця формула називається багаточленом для *інтерполяції назад*. Це пояснюється тим, що залучені для отримання вказаного поліному розділені різниці утворюють нижній косий рядок таблиці 1, (тобто рух по таблиці йде назад). Для багаточлена (10.8) також справедлива наближена формула оцінки похибки (10.7).

При неповних таблицях розділених різниць для інтерполяції на початку таблиці зазвичай застосовується багаточлен (10.7), так як в цьому випадку він має меншу похибку в порівнянні з багаточленом другого виду. Останній застосовується для інтерполяції в кінці таблиці, так як там його похибка менше. В середині таблиці можна застосовувати будь-який з них, але краще використовувати спеціально пристосовані для цього поліноми Гауса, Стерлінга або Беселя (факультативно).

10.3 Кінцеві різниці та їх властивості

Тепер перейдемо до розгляду кінцевих різниць. Нехай функція $y = f(x)$, задана таблицею, в якій вузли йдуть у порядку зростання з постійним кроком

$h = x_i - x_{i-1}$, тобто $x_i = x_0 + ih$, $i = 0, 1, \ldots, n$.

Такі таблиці будемо називати *рівномірними*. Визначимо для f кінцеві різниці.

Визначення. Кінцевою різницею першого порядку функції f в i-му вузлі називається величина

$$\Delta y_i = f(x_{i+1}) - f(x_i) = y_{i+1} - y_i, \quad (10.9)$$

де індекс $i = 0, \ldots, n-1$.

Таким чином, є n кінцевих різниць першого порядку. Видно, що вони являють собою приріст функції при збільшенні h (кроку) аргументу.

Різниці порядків вище першого визначаються рекуррентно через попередні.

Визначення. Кінцевою різницею k-го порядку функції f в i-му вузлі називається

$$\Delta^k y_i = \Delta^{k-1} y_{i+1} - \Delta^{k-1} y_i, \tag{10.10}$$

$i = 0, ..., n-k$; $k = 2, ..., n$. Наприклад, при $k = 2$ з (10.9), (10.10) отримуємо формулу різниці другого порядку:

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

Для k вищих порядків формули отримують рекурентно. Коефіцієнти в цих формулах є *числа сполучень*, або *біноміальні коефіцієнти*. Загальне твердження про це формулюється з наступної леми.

Лема 2. Кінцева k-го порядку в вузлі x_i обчислюється за формулою

$$\Delta^2 y_i = \sum_{j=0}^{k} (-1)^j C_k^j y_{i+k-j}, \tag{10.11}$$

де $C_k^j = \dfrac{k!}{j!(k-j)!}$ – біноміальний коефіцієнт (біном Ньютона – кількість сполучень з k елементів за i); $i = 0, ..., n-k$; $k = 1, ..., n$.

Так само як і у випадку з роздільними різницями, кінцеві різниці легше знаходити за таблицями, а сама Лема 2 потрібна тільки для теоретичних результатів та аналітичних розрахунків. Правило обчислення дуже просте, воно показано на рис. 10.2.

Для заповнення стовпчика різниць k-го порядку віднімаємо сусідні клітини в попередньому стовпці (показано стрілкою), результат пишемо в клітинку різниці (темна клітинка). Для наступної різниці спускаємося на два рядки і повторюємо ті ж операції.

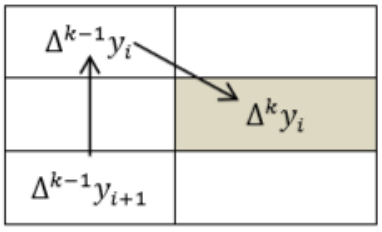


Рисунок 10.2 – до пояснення знаходження кінцевої різниці

У кінцевому випадку буде побудована наступна таблиця кінцевих різниць

Таблиця 2 – таблиця кінцевих різниць

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\cdots$	$\Delta^n y_i$
x_0	y_0				
		Δy_0			
x_1	y_1		$\Delta^2 y_0$		
		Δy_1		$\ddots$	
x_2	y_2				$\Delta^n y_0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
				$\iddots$	
x_{n-1}	y_{n-1}		$\Delta^2 y_{n-1}$		
		Δy_{n-1}			
x_n	y_n				

Для використання роздільних різниць при побудові поліномів слід зауважити, що розділені та кінцеві різниці пов'язані між собою наступною Лемою

Лема 3. Розділена різниця k-го порядку в вузлах x_i, $x_{i+1},\ldots,x_{i+k}$ виражається через кінцеву різницю k-го порядку в вузлі x_i формулою

$$f\left(x_i;\ldots;x_{i+k}\right)=\frac{\Delta^k y_i}{k!\,h^k}, \qquad (10.12)$$

$i = 0, \ldots, n\text{-}k$; $k = 1, \ldots, n$.

Доведення: Застосовуємо індукцію по k. При $k = 1$

$$f\left(x_i;\ldots;x_{i+1}\right)=\frac{\Delta y_i}{h}=\frac{y_{i+1}-y_i}{h}=\frac{y_{i+1}-y_i}{x_{i+1}-x_i},$$

що збігається з визначенням розділеної різниці першого порядку (10.1). База індукції доведена. Далі припускаємо, що формула (10.12) справедлива для всіх різниць k-го порядку.

Застосовуються розглянуті різниці для побудови поліномів Ньютона з кінцевими різницями, до розгляду яких і переходимо.

10.4 Інтерполяційні поліноми Ньютона з кінцевими різницями

Замінимо в інтерполяційному багаточлені Ньютона (10.6) розділені різниці за формулою (10.12) і отримаємо:

$$P_n^{(1)}(x) = f(x_0) + \frac{\Delta f_0}{1!\,h}(x - x_0) + \frac{\Delta^2 f_0}{2!\,h^2}(x - x_0)(x - x_1) + \ldots$$
$$+ \frac{\Delta^n f_0}{n!\,h^n}(x - x_0)\ldots(x - x_{n-1}) = f(x_0) + \sum_{i=1}^{n} \frac{\Delta^i f_0}{i!\,h^i}(x - x_0)\ldots(x - x_{i-1}). \qquad (10.13)$$

Отриманий вираз називається *інтерполяційним багаточленом Ньютона з кінцевими різницями першого виду* (або для інтерполяції вперед). Вхідні у (10.13) різниці утворюють верхній рядок таблиці 10.2.

На практиці частіше застосовується інша форма запису поліному Ньютона. Для отримання іншої форми запису багаточлену (10.13) зробимо заміну $q = \frac{x - x_0}{h}$, тоді $x = x_0 + qh$, звідси

$$x - x_0 = qh,$$
$$x - x_1 = x_0 + qh - x_1 = (q - 1)h,$$
$$x - x_{i-1} = x_0 + qh - x_{i-1} = (q - i + 1)h,$$

тоді доданки (10.13) перетворюються до вигляду

1-й $$\frac{\Delta y_0}{1!\,h}(x - x_0) = \frac{\Delta y_0}{1!}q$$

n-й $$\frac{\Delta^n y_i}{n!\,h^n}(x - x_0)\ldots(x - x_{n-1}) = \frac{\Delta^n y_0}{n!} \cdot \frac{x - x_0}{h} \cdot \frac{x - x_1}{h} \cdot \ldots \cdot \frac{x - x_{n-1}}{h} = \frac{\Delta^n y_0}{n!}q(q-1)\ldots(q-n+1).$$

Підставляючи ці різниці у (10.13) отримаємо

$$P_n^{(1)}(x) = P_n^{(1)}(x_0 + qh) = f(x_0) + \frac{\Delta y_0}{1!}q + \frac{\Delta^2 y_0}{2!}q(q-1) + \ldots +$$
$$+ \frac{\Delta^n y_0}{n!}q(q-1)\ldots(q-n+1) = f(x_0) + \sum_{i=1}^{n} \frac{\Delta^i y_0}{i!}q(q-1)\ldots(q-i+1) \qquad (10.14)$$

Такий запис зручний при застосуванні у тих випадках, коли потрібно вичислити інтерполяцію у конкретній точці і не потрібно при цьому виводити загальний вираз.

Аналогічно за допомогою (10.12) з (10.2) виводиться інтерполяційний багаточлен Ньютона з кінцевими різницями другого виду (або для інтерполяції назад):

$$P_n^{(2)}(x) = f(x_n) + \frac{\Delta f_{n-1}}{1!\,h}(x - x_n) + \frac{\Delta^2 f_{n-2}}{2!\,h^2}(x - x_n)(x - x_{n-1}) + \ldots$$
$$+ \frac{\Delta^n f_i}{n!\,h^n}(x - x_n)\ldots(x - x_1) = f(x_n) + \sum_{i=1}^{n} \frac{\Delta^i f_{n-i}}{i!\,h^i}(x - x_n)\ldots(x - x_1).$$

Вхідні в нього різниці знаходяться в нижньому косому рядку таблиці 10.2.

Заміною $q = \frac{x - x_n}{h}$,. за допомогою елементарних перетворень можна отримати іншу форму запису

$$P_n^{(2)}(x) = P_n^{(2)}(x_n + qh) = f(x_n) + \frac{\Delta y_{n-1}}{1!}q + \frac{\Delta^2 y_{n-2}}{2!}q(q+1) + \ldots +$$
$$+ \frac{\Delta^n y_0}{n!}q(q+1)\ldots(q+n-1) = f(x_n) + \sum_{i=1}^{n} \frac{\Delta^i y_{n-1}}{i!}q(q+1)\ldots(q+i-1) \qquad (10.15)$$

формула (10.15) застосовується для розрахунку значення в даній конкретній точці x. Для інтерполяційних багаточленів з кінцевими різницями залишаються в силі ті ж практичні правила застосування, що і для поліномів з розділеними різницями: *перший* з них краще підходить для інтерполяції *на початку*, *другий – в кінці* таблиць при неповних таблицях кінцевих різниць.

10.5 Оцінка похибок інтерполяційних поліномів Ньютона

Оцінка похибки розглянутих поліномів проводиться за загальними формулами попередньої лекції (лекція№9). Для їх представлень (10.14), (10.15) оціночні функції мають вигляд:

$$\overline{\Delta}\left(P_n^{(1)}(x_0 + qh)\right) = \frac{M_{n+1}}{(n+1)!}h^{n+1}\left|q(q-1)\ldots(q-n)\right|,$$
$$\overline{\Delta}\left(P_n^{(2)}(x_n + qh)\right) = \frac{M_{n+1}}{(n+1)!}h^{n+1}\left|q(q+1)\ldots(q+n)\right|, \qquad (10.16)$$

відповідно. З цих оцінок слід очікувати, що зменшення кроку таблиці (і відповідно, збільшення n) призведе до значного зниження похибки. Це певною мірою

справедливо для точності інтерполяції *в даній* точці. Але глобальна оцінка похибки (по всій таблиці) не обов'язково прагне до нуля при зменшенні h.

Питання для самоперевірки:

1. Дайте визначення розділених різниць першого і вищих порядків. За яких умов на таблицю можливе обчислення розділених різниць?
2. Як заповнюється таблиця розділених різниць? Якими властивостями володіють розділені різниці?
3. Що таке інтерполяційний багаточлен Ньютона з розділеними різницями першого виду? Чому він по-іншому називається багаточленом для інтерполяції вперед?
4. Що таке інтерполяційний багаточлен Ньютона з розділеними різницями другого виду?
5. Як найзручніше оцінювати похибку багаточлену Ньютона з розділеними різницями першого та другого виду?
6. У чому відмінність практичного застосування багаточленів Ньютона з розділеними різницями для інтерполяції вперед і назад?
7. Яка головна перевага поліномів Ньютона перед формулою Лагранжа?
8. За яких умов на таблицю можливе обчислення кінцевих різниць?
9. Дайте визначення кінцевих різниць першого і вищих порядків.
10. Як заповнюється таблиця кінцевих різниць? Якими властивостями володіють кінцеві різниці?
11. Як пов’язані кінцеві і розділені різниці?
12. Що таке інтерполяційний багаточлен Ньютона з кінцевими різницями першого виду? Які форми його запису ви знаєте?
13. Що таке інтерполяційний багаточлен Ньютона з кінцевими різницями другого виду? Які форми його записи ви знаєте?
14. Як оцінюються похибки багаточленів Ньютона з кінцевими різницями?

Тема 4.3 Кусково-багаточленна інтерполяція. Сплайни

ЛЕКЦІЯ 11 Кусково-лінійна інтерполяція. Інтерполяція сплайнами

Навчальні питання:

11.1 Кусково-лінійна інтерполяція

11.2 Інтерполяційний сплайн. Кубічний сплайн

11.3 Граничні умови

Усі розглянуті раніше методи інтерполяції припускають побудову однієї інтерполюючої функції для всієї таблиці. Така *інтерполяція* називається *глобальною.*

Для глобальної інтерполяції характерними є такі недоліки:

- По-перше, рішення матиме складний вид. Наприклад, інтерполяційний поліном при великому числі вузлів матиме високий ступінь (на одиницю менший кількості вузлів). Наслідками цього будуть обчислювальні складності при знаходженні чисельного рішення.
- По-друге, застосування глобальної інтерполяції на великих таблицях може давати значні похибки в деяких частинах таблиці.

Виникає питання, що робити, якщо необхідно наблизити функцію на великому інтервалі при наявності великої кількості проміжних точок. Як наблизити функцію, якщо побудова інтерполянта є некоректною задачею, що приводить до того, що з певною ймовірністю процес розійдеться?

Були розроблені вільні від цих недоліків методи *локальної інтерполяції***.** В цьому випадку дуже добрий результат дає спосіб наближення функції за допомогою сплайнів. Методи локальної інтерполяції припускають побудову декількох функцій, що інтерполюються по певних ділянках таблиці. Потім з них «склеюється» результуюча функція – це є рішення задачі. Найбільш популярним методом локальної інтерполяції є сплайн-інтерполяція, до вивчення якої переходимо.

11.1 Кусково-лінійна інтерполяція

Нехай є певні точки, які отримані під час розрахунку чи під час експерименту, по ним потрібно провести гладку криву. Або вже є функція, яка має розрив першого роду і є точки у яких ця функція має розрив.

Якщо виконувати інтерполяцію функції з розривом, то у околі розриву виникнуть осциляції, вони будуть тим більше, чим більше ступінь поліному, яким будемо наближувати вихідну функцію. Ці осциляції є не інформативними та від них потрібно позбавлятися або зробити амплітуди таких осциляцій найменшими.

Теорія сплайнів отримала свій розвиток на початку 50-х років XX ст. Та до того часу, сплайни використовували фізики та інженери, і вже потім математики. Найпростіший спосіб зменшити описані вище осциляції осциляції – замінити функцію кусково-лінійною функцією, та це є не завжди оптимальне рішення.

Інженери у такому випадку працювали досить просто: на міліметровому папері відмічали потрібні точки, вбивали у ці точки цвяхи та між цими цвяхами просовували гнучку лінійку (гнучке стальне лекало) що англійською називається – сплайн. Відповідно цей сплайн отримував форму, яка фізично намагалась мінімізувати потенціальну енергію. А мінімум потенціальної енергії – це і є мінімум можливих осциляцій.

Кусково-лінійна інтерполяція являє собою так звану інтерполяцію лінійними сплайнами. У цьому випадку найпростіший у використанні поліном (це поліном 1-го ступеня) створює багатокутний шлях з відрізків ліній, які проходять через точки. Щоб представити цю кусково-лінійну криву, використовується поліном Лагранжа (див. Лекцію №9 вирази (9.12), (9.13)):

$$S_k(x) = y_k \frac{(x - x_{k+1})}{(x_k - x_{k+1})} + y_{k+1} \frac{(x - x_k)}{(x_{k+1} - x_k)}, \qquad (11.1)$$

$$\forall \; x_k \le x \le x_{k+1}, \; k = 0, 1, \ldots, n$$

Результуюча крива виглядає подібно ламаній лінії (рис.11.1).

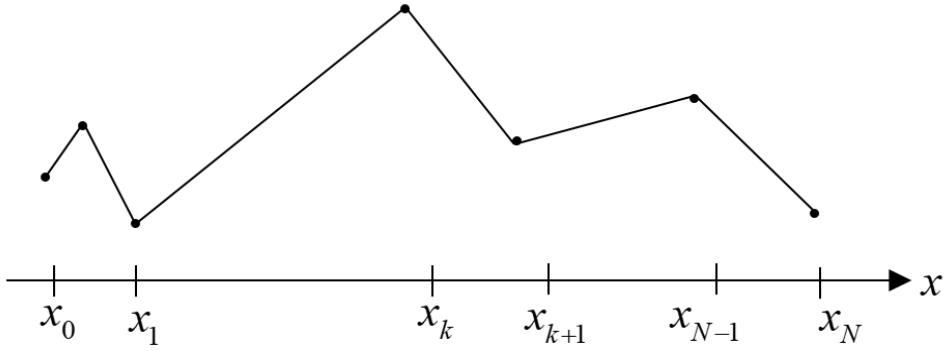


Рисунок 11.1 – геометрична інтерпретація кусково-лінійної інтерполяції

Еквівалентний вираз можна отримати, якщо використовувати формулу для тангенса кута нахилу відрізка лінії в точці:

$$S_k(x) = y_k + d_k(x - x_k),$$

де $d_k = \dfrac{y_{k+1} - y_k}{x_{k+1} - x_k}$, при $x_k \le x \le x_{k+1}$. Отриманий лінійний сплайн можна переписати наступним чином:

$$S(x) = \begin{cases} y_0 + d_0(x - x_0) & для\ x \in [x; x_0]; \\ y_1 + d_1(x - x_1) & для\ x \in [x_1; x_2]; \\ \dots & \dots \\ y_k + d_k(x - x_k) & для\ x \in [x_k; x_{k+1}]; \\ \dots & \dots \\ y_{N-1} + d_{N-1}(x - x_{N-1}) & для\ x \in [x_{N-1}; x_N]. \end{cases} \qquad (11.2)$$

Цю техніку можна узагальнити для поліномів вищого порядку. Наприклад, якщо задано непарне число вузлів то кусково-квадратичний поліном можна побудувати на кожному під інтервалі $[x_{2k}; x_{2k+2}]$

Недоліком отриманого в результаті квадратичного сплайну є те, що кривизна в парних вузлах різко змінюється, і це може викликати небажаний вигин або викривлення графіка. Друга похідна квадратичного сплайна має розрив в парних вузлах.

Якщо використовувати кусково-кубічні поліноми, то і першу, і другу похідні можна зробити безперервними.

11.2 Інтерполяційний сплайн. Кубічний сплайн

Нехай заданий відрізок $[a; b]$, і на ньому задано достатню кількість точок $a = x_0 < x_1 < \cdots < x_n = b$. Замість того, щоб побудувати єдиний багаточлен високого ступеня n, побудуємо на кожному інтервалі $[x_k; x_{k+1}]$ (див. рис. 11.2) свою функцію $S_m(x, k)$ невисокого ступіню m. При цьому є деяка свобода при побудові $S_m(x, k)$. Як правило, багаточлен вибирають максимально гладким.

Визначення. Сплайном називають кусково-поліноміальну функцію, яка разом із декількома похідними неперервна на всьому заданому відрізку $[a; b]$, а на частковому відрізку $[x_k; x_{k+1}]$ є деяким алгебраїчним багаточленом Це означає, що існує розбиття $[a; b]$ на підобласті, при якому всередині кожної з них сплайн збігається з деяким поліномом ступеня m. Число m називається ступенем сплайна.

Далі будемо розглядати тільки одновимірний випадок, тобто область існування – відрізок числової прямої.

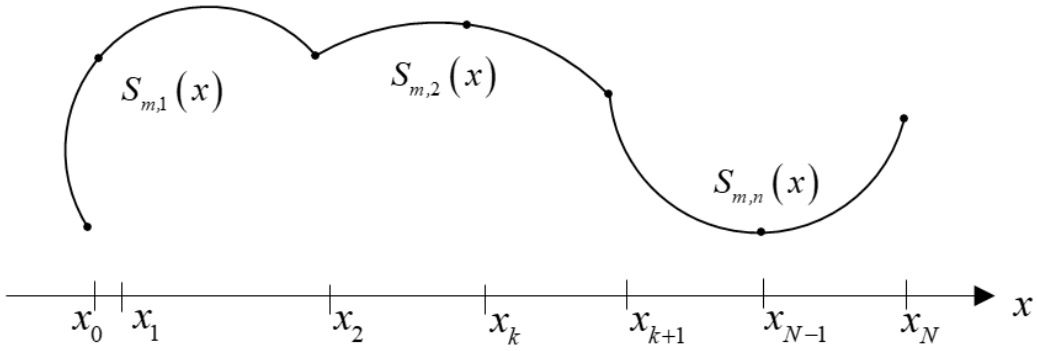


Рисунок 11.2 – кусково-поліноміальна інтерполяція

Визначення. Інтерполяційним сплайном ступеня m називається функція $S_m(x)$, яка задовольняє таким умовам:

1. на кожному відрізку $[x_k; x_{k+1}]$, збігається з деяким багаточленом $S_{m,k}$ ступеня m; ($k=0, 1, n$ – нумерація відрізку)
2. S_m задовольняє умовам інтерполяції по всій таблиці, тобто $S_m(x_k)=y_k$
3. S_m неперервна на $[a; b]$ з усіма своїми похідними до порядку p включно (число m-p називається *дефектом* сплайна).

Взагалі кажучи, вузли сплайна можуть не збігатися з вузлами інтерполяції. Для сплайнів непарного степеню стійкою виявляється ситуація, коли вузли

інтерполяції і вузли сплайна збігаються. Для сплайнів парного степеню - коли вони рознесені на півкроку.

Найбільш застосовними на практиці є кубічні сплайни, тобто сплайни третього ступеня.

Побудова кубічного сплайну

Нехай вузли інтерполяції збігаються з вузлами сплайна. Візьмемо наступне розбиття відрізку: $a = x_0 < x_1 < \cdots < x_n = b$. На кожному відрізку $[x_k; x_{k+1}]$, будемо будувати багаточлен $S_3(x,k)$ третього ступеня.

$$S_3(x,k) = a_k + b_k(x_k - x) + c_k(x_k - x)^2 + d_k(x_k - x)^3 \qquad (11.3)$$

Введемо величину $h_k = x_{k+1} - x_k$ – довжина відрізку, на якому будується сплайн. Кожен багаточлен містить чотири невідомих коефіцієнта, значить, разом отримаємо $4n$ невідомих коефіцієнтів. Побудуємо таку систему рівнянь, щоб сплайн був інтерполяційним, і функція була якомога більш гладкою.

1. Будемо вимагати, щоб виконувались умови інтерполяції у лівій межі інтервалу, тобто $S_3(x_k,k) = f_k$. І у цьому випадку підставляючи значення вузлів лівої границі інтервалу в (1), (цей багаточлен має відповідати умовам інтерполяції, а саме мають збігатися значення багаточлена в вузлах з відповідними значеннями функції) отримаємо

$$a_k = f_k, \quad k = 0,1,...,n-1. \qquad (11.4)$$

Таких умов буде *n.*

2. Аналогічним чином вимагаємо виконання умов інтерполяції для правої межі інтервалу $S_3(x_{k+1},k) = f_{k+1}$, і отримуємо:

$$a_k + b_k h_k + c_k h_k^2 + d_k h_k^3 = f_{k+1}, \qquad (11.5)$$

таких умов також *n.*

На цьому умови інтерполяції закінчились, переходимо до умов гладкості функції

3. Умова неперервності S_3' в точках x_k.

$$S_3'(x_{k+1},k) = S_3'(x_{k+1},k+1)$$

$$b_k + 2c_k h_k + 3d_k h_k^2 = b_{k+1}, \quad k = 0,1,...,n-1. \tag{11.6}$$

Отримаємо *n*-1 рівняння.

Пам’ятаємо що невідомих коефіцієнтів всього $4n$, тому залишається ще деяка свобода у заданні сплайну.

4. Будемо вимагати неперервності другої похідної

$$S_3''(x_{k+1}, k) = S_3''(x_{k+1}, k+1)$$

що дасть наступний вираз:

$$2c_k + 6d_k h_k = 2c_{k+1}, \quad k = 0,1,...,n-1. \tag{11.7}$$

Отримаємо *n*-1 рівняння.

Всього отримаємо $4n$-2 умов і $4n$ невідомих. Виникає питання, де взяти дві відсутні умови? В цьому випадку додаткові умови можна вибрати в деякій мірі довільно, в залежності від виду функції та від заданих умов задачі. Існує різні типи побудови сплайнів з додатковими умовами (крайовими умовами). Далі розглянемо найпоширеніші граничні (крайові) умови. Та для початку приведемо за допомогою елементарних перетворень рівняння (11.4)-(11.7) до загального виду з *n* невідомим. Нехай цим невідомим буде коефіцієнт *c.*

З умови (11.7) виразимо d_k:

$$d_k = \frac{c_{k+1} - c_k}{3h_k}. \tag{11.8}$$

Підставляємо d_k та умову (11.4) в умови (11.5) та (11.6):

$$f_k + b_k h_k + c_k h_k^2 + \frac{c_{k+1} - c_k}{3h_k} h_k^3 = f_{k+1}, \tag{11.9}$$

$$b_k + 2c_k h_k + 3\frac{c_{k+1} - c_k}{3h_k} h_k^2 = b_{k+1}, \tag{11.10}$$

Перетворюємо (11.9) відносно b_k

$$b_k = \frac{f_{k+1} - f_k}{h_k} - c_k h_k - \frac{1}{3}(c_{k+1} - c_k) h_k, \qquad b_k = \frac{f_{k+1} - f_k}{h_k} - \frac{2}{3} c_k h_k - \frac{1}{3} c_{k+1} h_k.$$

Тоді для b_{k+1} отримаємо:

$$b_{k+1} = \frac{f_{k+2} - f_{k+1}}{h_{k+1}} - \frac{2}{3} c_{k+1} h_{k+1} - \frac{1}{3} c_{k+2} h_{k+1}.$$

Підставляємо b_k та b_{k+1} у (11.10):

$$h_k c_k + 2\left(h_k + h_{k+1}\right) c_{k+1} + h_{k+1} c_{k+2} = 3\left(\frac{f_{k+2} - f_{k+1}}{h_{k+1}} + \frac{f_{k+1} - f_k}{h_k} \right)$$

Зсуваємо всі індекси на одиницю. Тоді отримаємо:

$$\frac{1}{h_k} c_{k-1} + 2c_k \left(\frac{1}{h_k} + \frac{1}{h_{k+1}} \right) + c_{k+1} \frac{1}{h_{k+1}} = 3\left(\frac{f_k - f_{k-1}}{h^2_k} + \frac{f_{k+1} - f_k}{h^2_{k+1}} \right) \qquad (11.11)$$

Таким чином була отримана система рівнянь відносно коефіцієнтів c_k, що володіє матрицею тридіагональної структури. Тобто кожне рівнянні містить коефіцієнти $k-1, k, k+1$, а інші коефіцієнти дорівнюють нулю. Для розв'язку таких матриць використовують модифікації методу Гауса для неповних систем. Після того, як система буде розв'язана відносно c_k, поступово повертаючись до заміни, знаходять інші невідомі b_k за формулою (11.10), далі значення d_k, а значення a_k відомі. І таким чином знаходять розв'язок задачі на всьому інтервалі від x_0 до x_n. Також потрібно зауважити, що на відміну від інтерполяції, сплайн наближає не тільки функцію, а і її похідні.

Вище було сказано, що система (11.11) недовизначена (для $4n$-2 умов і $4n$ невідомих). Існують стратегії, які дозволяють визначити невистачаючі два рівняння. Ці стратегії обирають з заданих відомостей про функцію, її початкових умов. Коли всі коефіцієнти c_k знайдені, складають кубічний сплайн на кожному відрізку $[x_k; x_{k+1}]$. Для цього можна користуватися інтерполяційною формулою Ерміта:

$$S_3(x) = P_{3,k}(x) = y_{k-1} \frac{\left(x - x_k\right)^2 \left(2\left(x - x_{k-1}\right) + h_k\right)}{h_k^3} + c_{k-1} \frac{\left(x - x_k\right)^2 \left(x - x_{k-1}\right)}{h_k^2} +$$
$$+ y_k \frac{\left(x - x_{k-1}\right)^2 \left(2\left(x_k - x\right) + h_k\right)}{h_k^3} + c_k \frac{\left(x - x_{k-1}\right)^2 \left(x - x_k\right)}{h_k^2}, \qquad (11.12)$$

$x \in [x_k; x_{k+1}]$, $k=1,\ldots,n$.

Безпосередньою перевіркою можна перевірити, що для (11.12) виконуються умови інтерполяції та забезпечується неперервність сплайну во внутрішніх вузлах.

Повертаємось до стратегій визначення двох невистачаючих рівнянь для системи (11.11).

11.3 Граничні умови

1) Перший тип граничних умов. Змикаючийся сплайн.

Якщо відомі значення похідних у граничних вузлах, то природньо покласти:

$$c_0 = f'(x_0),\ c_n = f'(x_n) \qquad (11.13)$$

Отримаємо два невистачаючі рівняння та разом з тим число невідомих також зменшується на два. Отриманий таким чином сплайн ***називають змикаючимся***. Він має визначений нахил у крайніх точках. Цей сплайн можна представити як криву, яка отримана коли гнучкий еластичний стрижень має проходити через задані точки та ту, що примикає до кожного краю з фіксованим нахилом.

2) Другий тип граничних умов. Умови вільного провисання. Природній сплайн.

У цьому випадку відомими вважають $f''(x_0)$ та $f''(x_n)$. Тоді потрібно прирівняти ці значення другим похідним сплайну у лівій та правій граничних точках відповідно

$$S_3''(x_0) = P_{3,k}''(x_0) = -\frac{4}{h_1}c_0 - \frac{2}{h_1}c_1 + \frac{6(f_1 - f_0)}{h_1^2} = f''(x_0),$$

$$S_3''(x_n) = P_{3,k}''(x_n) = \frac{2}{h_n}c_{n-1} + \frac{4}{h_n}c_n + \frac{6(f_n - f_{n-1})}{h_n^2} = f''(x_n),$$

Та отримаємо два невистачаючі рівняння рішення системи (11):

$$-\frac{4}{h_1}c_0 - \frac{2}{h_1}c_1 = f''(x_0) + \frac{6(f_0 - f_1)}{h_1^2}, \qquad \frac{2}{h_n}c_{n-1} + \frac{4}{h_n}c_n = f''(x_n) + \frac{6(f_n - f_{n-1})}{h_n^2}, \qquad (11.14)$$

Є важливим частковий випадок таких граничних умов, якщо $f''(x_0) = 0$, $f''(x_n) = 0$, то такі граничні умови називають ***умовами вільного провисання*** і відповідні рівняння (13) спростяться:

$$2c_0 + c_1 = 3\frac{f_1 - f_0}{h_1}; \qquad c_{n-1} + 2c_n = 3\frac{f_n - f_{n-1}}{h_n} \tag{11.15}$$

Побудований за таких умов сплайн називають ***природнім***.

Завдяки своїй простоті умови вільного провисання набули широкого поширення, вони застосовуються часто тоді, коли немає ніякої інформації про f'' на кінцях таблиці. Однак слід мати на увазі, що природний сплайн в таких випадках менш точно наближає функцію в порівнянні з іншими.

3) Третій тип граничних умов. Екстраполяційний сплайн.

При відсутності будь-якої інформації про поведення функції у крайній точках, можна покласти $P_{3,1}(x) \equiv P_{3,2}(x)$, $P_{3,n-1}(x) \equiv P_{3,n}(x)$. Фактично це означає злиття суміжних проміжків $[x_0 ; x_1]$, $[x_1 ; x_2]$ та $[x_{n-2}; x_{n-1}]$, $[x_{n-1}; x_n]$, тобто викреслення з таблиці вузлів x_1, x_{n-1}.

Тому ці умови називають умовами «відсутності вузла». Вони приводять до граничних рівнянь.

$$\frac{1}{h_1^2}c_0 + \left(\frac{1}{h_1^2} - \frac{1}{h_2^2}\right)c_1 - \frac{1}{h_2^2}c_2 = 2\left(\frac{1}{h_2^3}(f_1 - f_2) + \frac{1}{h_1^3}(f_1 - f_0)\right)$$

$$\frac{1}{h_{n-1}^2}c_{n-2} + \left(\frac{1}{h_{n-1}^2} - \frac{1}{h_n^2}\right)c_{n-1} - \frac{1}{h_2^2}c_n = 2\left(\frac{1}{h_n^3}(f_{n-1} - f_n) + \frac{1}{h_{n-1}^3}(n = f_{n-1} - f_{n-2})\right) \tag{11.16}$$

4) <u>Четвертий тип граничних умов</u>.

У цьому випадку функцію вважають *періодичною* та виконується умова $f_0 = f_n$, або $f_0 \approx f_n$. Тоді перше рівняння перше граничне рівняння має вид:

$$c_0 = c_n$$

Друге рівняння знаходять з умови $S_3''(x_0) = S_3''(x_n)$. З формули Ерміта отримаємо:

$$\frac{2}{h_1}c_0 + \frac{1}{h_1}c_1 - \frac{1}{h_n}c_{n-1} + \frac{2}{h_n}c_n = 3\left(\frac{1}{h_1^2}(f_1 - f_0) + \frac{1}{h_1^2}(f_n - f_{n-1})\right) \tag{11.17}$$

З вище описаного зробимо головний висновок:

Додаючи до (11.11) два рівняння, що відповідають певним граничним умовам, отримамаємо лінійну систему, яку можемо записати у матричному вигляді:

$$A \cdot \bar{c} = \bar{b}$$

A - матриця системи (відомі значення), $\bar{c}$ - вектор невідомих нахилів, $\bar{b}$ - вектор правих частин.

Для побудови кубічного сплайну отримаємо наступний *алгоритм:*

1. Розв'язуємо (11.11) з додаванням граничних умов
2. Знаходимо нахили $\bar{c}$
3. По формулі Ерміта (11.12) будуємо частинки сплайну $P_{3,k}$ на відрізках

та збираємо кубічний сплайн $S_3(x) = P_{3,k}$, $x \in [x_k, x_{k+1}]$, $k = 1,...,n$

Треба зауважити, що описані граничні умови можна комбінувати, в залежності від того що відомо про функцію.

Питання для самоперевірки:

1. У чому відмінність глобальної інтерполяції від локальної? Які вирішальні недоліки глобальної інтерполяції?
2. Дайте загальне визначення сплайна. Що таке ступінь, дефект сплайна?
3. Дайте визначення інтерполяційного сплайна. Сформулюйте задачу сплайн-інтерполяції.
4. Що таке нахили сплайна?
5. Що таке граничні умови? Для чого вони потрібні?
6. Виведіть граничні рівняння при заданих значеннях другої похідної в крайніх вузлах.
7. Що таке умови вільного провисання і природний кубічний сплайн?
8. Які переваги і недоліки природного сплайна?
9. Що таке граничні умови "відсутності вузла"? Виведіть граничні рівняння для цих умов. Чому треба для цього використовувати безперервність третьої похідної сплайна?
10. Виведіть граничні рівняння при періодичної функції з періодом, рівним розмаху таблиці.

11. Які граничні умови використовуються при відсутності будь-якої інформації про поведінку функції на краях таблиці?
12. Яку структуру має матриця системи для знаходження нахилів кубічного сплайна?
13. Який порядок наближення функції і її похідних кубічним сплайном і його похідними?

Тема 4.4 Апроксимація. Метод найменших квадратів.

ЛЕКЦІЯ 12 Задача апроксимації. Поліноміальна та неполіноміальна апроксимація методом найменших квадратів

Навчальні питання:

12.1 Здача апроксимації табличної функції

12.2 Метод найменших квадратів

12.3 Поліноміальна апроксимація методом найменших квадратів

12.4 Неполіноміальна апроксимація

12.1 Здача апроксимації табличної функції

Як вже зазначалося у минулих лекціях, крім інтерполяції існує другий підхід до вирішення завдання аналітичного наближення табличних функцій - апроксимація. Чим викликана необхідність такого підходу?

Застосування інтерполяції не завжди виправдано, це особливо стосується *таблиць з великим числом вузлів*. Пояснюється це наступними причинами. По-перше, функції, що інтерполюють для них дуже складні; наприклад, інтерполяційні багаточлени мають *високій ступень* або обчислюються з великими похибками. По-друге, при наявності похибок табличних даних, наприклад, при обробці результатів вимірювань, статистичних випробувань, вони переходять в інтерпольовану функцію, яка за визначенням повинна точно відстежувати табличні значення (міряєте сигнал із завадами).

Таким чином, похибки вихідних даних повторюються в рішенні, а бажано, щоб вони *згладжувалися.* Саме це і відбувається при апроксимації. Ідею апроксимації можна сформулювати так: будується функція, яка приблизно повторює табличні дані, відстежуючи загальну тенденцію зміни невідомої функції, що апроксимується. Отже потрібно провести криву так, щоб вона в найменшій мірі залежала від випадкових помилок. Це завдання називається згладжуванням (апроксимацією) експериментальної залежності і часто вирішується *методом найменших квадратів.* Згладжуючу криву називають апроксимуючою.

Тепер поставимо задачу більш точно. Функція $f(x)$ задана таблицею. Потрібно побудувати її аналітичне наближення. Для цього за даними таблиці знаходиться функція $y = g(x)$, що апроксимує вихідну функцію $f(x)$, тобто задовольняє умову

$$g(x_i) \approx y_i, \qquad i = 1,...,n,. \tag{12.1}$$

Точки x_i називаються вузлами апроксимації. Нехай визначений клас функцій, в якому зшукається g, цей клас функцій називається *клас апроксимуючих функцій*.

У такій постановці задача некоректна. І справа не в тому, що вона може мати безліч рішень. Умови (12.1) не є математично точними, через не визначеність наближеної рівності. Також не відомим є прийнятний рівень похибки для умов (12.1). Ці та інші питання будуть розглянуті далі.

12.2 Метод найменших квадратів

Апроксимуюча функція $g \in \Omega$, що належить певному класу апроксимуючих функцій повинна бути близька до значень f у всіх вузлах, тому логічно будувати її *з умови мінімуму деякої величини*, що характеризує сукупну похибку наближених рівностей (12.1).

Виберемо в якості такої похибки середньоквадратичне відхилення g від f в вузлах $x_1,...,x_n$.

$$\delta_2(g) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - g(x_i)\right)^2} \tag{12.2}$$

(Квадратний корінь доданий для того, щоб розмірність $\delta_2(g)$ була та ж, що у f).

Тоді задача апроксимації сформулюється математично абсолютно точно: по таблиці 1 побудувати функцію $g \in \Omega$, для якої величина $\delta_2(g)$ (12.2) мінімальна. Це і є *задача методу найменших квадратів* (МНК).

Далі наведено її рішення в *класі узагальнених багаточленів*. Також такі багаточлени називають ще багаточленами найкращого середнє квадратичного наближення.

Узагальненим багаточленом ступеня m на системі $\{\varphi_0,\varphi_1,...,\varphi_m\}$ базисних функцій (ця система за визначенням лінійно незалежна) називається

$$\Phi_m(x)=a_0\varphi_0(x)+...+a_m\varphi_m(x)=\sum_{j=0}^{m}a_j\varphi_j(x) \tag{12.3}$$

Отже, нехай Ω - множина узагальнених багаточленів (3). Тоді

$$\begin{gathered} g(x)=\Phi_m(x)\Rightarrow\delta_2(g)=\delta_2(\Phi_m)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-g(x_i)\right)^2}= \\ =\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\Phi_m(x_i)\right)^2}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{m}a_j\varphi_j(x)\right)^2} \end{gathered} \tag{12.4}$$

Зрозуміло, що Φ_m повністю визначається своїми коефіцієнтами $a_0,a_1,...,a_m$ тому математично розглянута проблема являє собою завдання безумовної мінімізації функції $\delta_2(g)$, що залежить від $m+1$ аргументу $a_0,a_1,...,a_m$.

$$\delta_2(\Phi_m)=\delta_2(a_0,...,a_m)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{m}a_j\varphi_j(x_i)\right)^2}\to\min_{a_0,...,a_m}$$

Перейдемо від неї до еквівалентної задачі мінімізації функції. Те, що задачі еквівалентні, випливає з монотонного зростання квадратного кореня.

$$S(a_0,...,a_m)=\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{m}a_j\varphi_j(x_i)\right)^2$$

Метод рішення відомий з математичного аналізу: треба взяти часткові похідні, прирівняти їх до нуля і отримати систему рівнянь для знаходження критичних точок функції. Проробимо цю операцію:

$$\begin{gathered} \frac{\partial S}{\partial a_k}=2\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{m}a_j\varphi_j(x_i)\right)\varphi_k(x_i)=0\Leftrightarrow\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{m}a_j\varphi_j(x_i)\right)\varphi_k(x_i)=0\Leftrightarrow \\ \Leftrightarrow\sum_{i=0}^{n}\sum_{j=0}^{m}a_j\varphi_j(x_i)\varphi_k(x_i)=\sum_{i=0}^{n}y_i\varphi_k(x_i)\Leftrightarrow \\ \Leftrightarrow\sum_{j=0}^{m}a_j\sum_{i=0}^{n}\varphi_j(x_i)\varphi_k(x_i)=\sum_{i=0}^{n}y_i\varphi_k(x_i) \end{gathered} \tag{12.5}$$

$k=0,...,m$. Отримана система $m+1$ рівняння для знаходження $m+1$ невідомого, $a_0,a_1,...,a_m$.

У розгорнутому вигляді вона записується так:

$$\begin{cases} a_0\sum_{i=1}^{n}\varphi_0(x_i)^2 + a_1\sum_{i=1}^{n}\varphi_1(x_i)\varphi_0(x_i) + ... + a_m\sum_{i=1}^{n}\varphi_m(x_i)\varphi_0(x_i) = \sum_{i=1}^{n} y_i\varphi_0(x_i) \\ a_0\sum_{i=1}^{n}\varphi_0(x_i)\varphi_1(x_i) + a_1\sum_{i=1}^{n}\varphi_1(x_i)^2 + ... + a_m\sum_{i=1}^{n}\varphi_m(x_i)\varphi_1(x_i) = \sum_{i=1}^{n} y_i\varphi_1(x_i) \\ ... \quad .... \quad .... \quad .... \quad .... \\ a_0\sum_{i=1}^{n}\varphi_0(x_i)\varphi_m(x_i) + a_1\sum_{i=1}^{n}\varphi_1(x_i)\varphi_m(x_i) + ... + a_m\sum_{i=1}^{n}\varphi_m(x_i)^2 = \sum_{i=1}^{n} y_i\varphi_m(x_i) \end{cases}$$

Очевидно, що вона лінійна і її можна записати в матричному вигляді:

$$\Gamma\bar{a} = \bar{b} \tag{12.6}$$

де $\Gamma = P^T P$

$$P = \begin{pmatrix} \varphi_0(x_1) & \varphi_1(x_1) \ ... \ \varphi_m(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) \ ... \ \varphi_m(x_2) \\ & ... \\ \varphi_0(x_n) & \varphi_1(x_n) \ ... \ \varphi_m(x_n) \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} (\varphi_0,\varphi_0) & (\varphi_0,\varphi_1) \ ... \ (\varphi_0,\varphi_m) \\ (\varphi_1,\varphi_0) & (\varphi_1,\varphi_1) \ ... \ (\varphi_1,\varphi_m) \\ & ... \\ (\varphi_m,\varphi_0) & (\varphi_m,\varphi_1) \ ... \ (\varphi_m,\varphi_m) \end{pmatrix}$$

$$\bar{a} = \begin{pmatrix} a_0 \\ a_1 \\ ... \\ a_m \end{pmatrix}, \ \bar{b} = P^T\bar{y}, \ \bar{y} = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix}$$

(P^T - транспонована матриця P); Γ називається матрицею Грама, яка складається зі скалярних добутків функцій $\varphi_0,...,\varphi_m$. Систему функцій називають ортогональною, якщо

$$\begin{aligned} &(\varphi_i,\varphi_k) = 0, \ \ i \neq k; \\ &(\varphi_i,\varphi_k) > 0, \ \ i \geq 0, \ \ k \leq m. \end{aligned} \tag{12.7}$$

Систему функцій називають лінійно незалежною, якщо лінійна комбінація $a_0\varphi_0 + a_1\varphi_1 + ... + a_m\varphi_m = 0$, тоді і лише тоді, коли $a_0 = a_1 = ... = a_m = 0$

Система (12.6) – нормальна система метода найменших квадратів. Якщо матриця Грама неособлива, то система (12.6) має єдиний розв'язок. Можна показати, що це завжди має місце при попарно незбіжних вузлах.

Припустимо, що (12.6) має єдине рішення. Потрібно перевірити, чи буде воно точкою мінімуму функції S. У загальному випадку це робиться досить складно. Приймемо без доведення той факт, що якщо система (12.6) має єдине рішення, то

воно є точкою мінімуму. За знайденими коефіцієнтами,, далі будується багаточлен (12.3), який називається многочлен му найкращого середньоквадратичного відхилення. Мінімальне відхилення обчислюється за формулою (12.4). Таким чином, задача МНК повністю вирішена.

12.3 Поліноміальна апроксимація методом найменших квадратів

Тепер перейдемо до розгляду важливого часткового випадку задачі МНК. Нехай система базисних функцій $\{\varphi_0,\varphi_1,...,\varphi_m\}$ – степенева, тобто $\varphi_j(x)=x^j,\ j=0,1,...,m$. Тоді узагальнений багаточлен перетворюється в звичайний поліном $P_m=a_0+a_1x+...+a_mx^m$, і ми приходимо до задачі і ми приходимо до задачі поліноміальної апроксимації

$$\delta_2(P_m)=\delta_2(a_0,...,a_m)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{m}a_jx^j\right)^2}\to\min_{a_0,...,a_m} \tag{12.8}$$

Нормальна система для неї має вигляд:

$$\sum_{j=0}^{m}a_j\sum_{i=1}^{n}x_i^{j+k}=\sum_{i=1}^{n}y_ix_i^k \tag{12.9}$$

$k=0,...,m$.

Якщо серед вузлів немає співпадаючих, то система (12.9) має єдине рішення, за яким будується поліном найкращого середньоквадратичного відхилення.

Зауваження. Якщо $m=n-1$ (ступінь апроксимуючого багаточлена на одиницю менше числа вузлів), то побудований за рішенням (12.9) поліном збігається з інтерполяційним багаточленом ступеня, для якого відхилення $\delta_2(P_m)$ дорівнює нулю. Таким чином, завдання інтерполяції є окремим випадком апроксимації.

Запишемо систему (12.9) для двох окремих випадків. Якщо $m=1$, то отримуємо задачу *лінійної апроксимації*. Нормальна система для знаходження коефіцієнтів $P_m=a_0+a_1x$ має вигляд

$$S(a_0,a_1)=\sum_{i=1}^{n}\left(y_i-a_0-a_1x_i\right)\Leftrightarrow\begin{cases}\dfrac{\partial S}{\partial a_0}=0\\[2ex]\dfrac{\partial S}{\partial a_1}=0\end{cases}\Leftrightarrow\begin{cases}a_0n+a_1\sum\limits_{i=1}^{n}x_i=\sum\limits_{i=1}^{n}y_i,\\[2ex]a_0\sum\limits_{i=1}^{n}x_i+a_1\sum\limits_{i=1}^{n}x_i^2=\sum\limits_{i=1}^{n}y_ix_i.\end{cases}\qquad(12.10)$$

Найкраще середньоквадратичне відхилення розраховується за формулою:

$$\delta_2(P_1)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-a_0-a_1x_i\right)^2}$$

Приклад 12.1: Побудувати згладжуючу функцію для таблиці даних

x_i	0	-1
y_i	-1	2

Розв'язок: Складаємо нормальну систему методу найменших квадратів (МНК) для $n=2$:

$$\sum_{i=1}^{n}x_i=0+(-1)=-1,\ \sum_{i=1}^{n}x_i^2=0^2+(-1)^2=1,\ \sum_{i=1}^{n}y_i=(-1)+2=1,\ \sum_{i=1}^{n}x_iy_i=0\cdot(-1)+(-1)\cdot2=-2.$$

Звідси: $\begin{cases}2a_0-a_1=1,\\-a_0+a_1=-2.\end{cases}\Rightarrow$ Розв'язуючи систему та знаходимо найкращу лінійну апроксимацію $\begin{cases}a_0=-1,\\a_1=-3.\end{cases}\Rightarrow P_1=-1-3x$

Найкраще середньоквадратичне відхилення буде дорівнювати:

$$\delta_2(P_1)=\sqrt{\frac{1}{2}\left(\left(-1-P_1(0)\right)^2+\left(2-P_1(-1)\right)^2\right)}=\sqrt{\frac{1}{2}\left((-1+1)^2+(2-2)^2\right)}=0$$

Отриманий результат є цілком логічним, оскільки апроксимуюча крива точно проходить через 2 точки, а отже відхилення буде нульовим.

При $m=2$ приходимо до задачі *квадратичної апроксимації*. Нормальна система для знаходження коефіцієнтів найкращого багаточлена другого ступеня $P_m=a_0+a_1x+a_2x^2$, наступна:

$$\begin{cases} a_0 n + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i, \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} y_i x_i, \\ a_0 \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} y_i x_i^2. \end{cases} \quad (12.11)$$

Найкраще середньоквадратичне відхилення розраховується за формулою:

$$\delta_2(P_2) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_1 x_i^2 \right)^2}.$$

Геометрично лінійна і квадратична апроксимація являє собою найкраще згладжування (наближення) набору точок на площині *прямою або параболою* відповідно.

Приклад 12.2: Побудувати багаточлени найкращого середньоквадратичного відхилення ступеня 2 і 3 для таблиці

x_i	-4	-2	0	2
y_i	2	-1	0	1

Розв'язок:

Складаємо допоміжну таблицю та знаходимо коефіцієнти для квадратичної апроксимації

i	x_i	y_i	x_i^2	x_i^3	x_i^4	x_i^5	x_i^6	$x_i y_i$	$x_i^2 y_i$	$x_i^3 y_i$
1	-4	2	16	-64	256	-1024	4096	-8	32	-128
2	-2	-1	4	-8	16	-32	64	2	-4	8
3	0	0	0	0	0	0	0	0	0	0
4	2	1	4	8	16	32	64	2	4	8
Σ	-4	2	24	-64	288	-1024	4224	-4	32	-112

$$\begin{cases} 4a_0 - 4a_1 + 24a_2 = 2 \\ -4a_0 + 24a_1 - 64a_2 = -4 \\ 24a_0 - 64a_1 + 288a_2 = 32. \end{cases}$$ Розв'язуючи її будь-якими відомими методами розв'язку СЛАР знаходимо: $\begin{cases} a_0 = 0{,}04, \\ a_1 = 0{,}08, \\ a_2 = 0{,}09. \end{cases} \Rightarrow \; P_2(x) = 0{,}04 + 0{,}08x + 0{,}09x^2$

$$\delta_2(P_2) = \sqrt{\frac{1}{4}\left(\left(2 - P_2(-4)\right)^2 + \left(-1 - P_2(-2)\right)^2 + \left(0 - P_2(0)\right)^2 + \left(1 - P_2(2)\right)^2\right)} = 3{,}7 \cdot 10^{-9}.$$

Для поліному третього ступеню $P_m = a_0 + a_1 x + a_2 x^2 + a_2 x^3$, користуючись (12.6) та вже складеною вище таблицею, записуємо нормальну систему МНК:

$$\begin{cases} 4a_0 - 4a_1 + 24a_2 - 64a_3 = 2, \\ -4a_0 + 24a_1 - 64a_2 + 288a_3 = -4, \\ 24a_0 - 64a_1 + 288a_2 - 1024a_3 = 32, \\ 64a_0 + 288a_1 - 1024a_2 + 4224a_3 = -112. \end{cases}$$ Розв'язуючи яку знаходимо:

$$\begin{cases} a_0 = 0, \\ a_1 = \dfrac{5}{6}, \\ a_2 = 0, \\ a_3 = -\dfrac{1}{12}. \end{cases} \Rightarrow P_3 = \frac{5}{6}x - \frac{1}{12}x^3.$$

$$\delta_2(P_3) = \sqrt{\frac{1}{4}\left(\left(2 - P_3(-4)\right)^2 + \left(-1 - P_3(-2)\right)^2 + \left(0 - P_3(0)\right)^2 + \left(1 - P_3(2)\right)^2\right)} = 1{,}05 \cdot 10^{-8}.$$

Отже, МНК дозволяє знаходити функції які згладжують залежності представлені таблично. Але застосовуючи МНК можна побудувати не тільки поліноміальне наближеня, з чого і витікає наступне питання, що розглянемо.

12.4 Неполіноміальна апроксимація

Якщо задана таблична функція і є припущення щодо її вигляду, який визначається деяким набором параметрів, то ці параметри можна приблизно оцінити методом найменших квадратів так само, як і коефіцієнти полінома. Зазвичай таблична функція – це дані експерименту, зведені в таблицю. За ними можна приблизно припустити, яка функція апроксимує їх. Наприклад, по точкам

побудувати ламану, що їх сполучає. Потім оцінити, до якого класу функцій найближче схожа ця ламана, записати загальний вираз функцій цього класу з параметрами. А потім методом найменших квадратів знайти наближені значення цих параметрів. Часто використовуються такі сімейства функцій:

1) Гіперболічна $y = b + \frac{a}{x}$, $y = \frac{1}{ax+b}$, $y = \frac{x}{ax+b}$,

2) Логарифмічна $y = a + ln(x)$,

3) Степенева $y = ax^b$,

4) Показникова $y = ae^{bx}$.

Всі ці сімейства двопраметричні. Розглянемо для прикладу показникову функцію. Нехай апроксимуюча функція має вигляд $y = ae^{bx}$.

Таке наближення застосовується, наприклад, при оцінці параметрів експоненціального зростання експериментальних даних (або взагалі, перевірці адекватності експоненційної моделі зростання даних). Діємо за алгоритмом рішення задачі МНК:

$$S(a,b) = \sum_{i=1}^{n}\left(y_i - a_0 e^{bx_i}\right)^2 \Rightarrow \begin{cases} \frac{\partial S}{\partial a} = -2\sum_{i=1}^{n}\left(y_i - a_0 e^{bx_i}\right)e^{bx_i} = 0, \\ \frac{\partial S}{\partial b} = -2a\sum_{i=1}^{n}\left(y_i - a_0 e^{bx_i}\right)x_i\, e^{bx_i} = 0, \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n}\left(y_i - a_0 e^{bx_i}\right)e^{bx_i} = 0, \\ \sum_{i=1}^{n}\left(y_i - a_0 e^{bx_i}\right)x_i\, e^{bx_i} = 0. \end{cases}$$

Отримуємо нормальну систему МНК. Це вже нелінійна система, причому явні формули рішення не виводяться. Вона вирішується наближеними методами. Вирішуємо її і отримуємо значення параметрів, при яких апроксимуюча функція дає мінімальне середньоквадратичне відхилення.

Для того, щоб лінеаризувати систему необхідно підібрати таке перетворення вихідної залежності, в результаті якого вона набуває лінійний вид. Далі вирішується задача лінійної апроксимації для нової залежності і нові обчислені коефіцієнти у перераховуються у попередні коефіцієнти зворотною заміною.

Для ряду двопараметричних залежностей можливі заміни змінних, що наведені в табл. 12.1.

Таблиця 12.1 – заміна змінних для лінеаризації нелінійної залежності

Вид залежності	Заміна змінних		Обмеження
гіперболічна $y = a + \frac{b}{x}$	$v = y$	$u = \frac{1}{x}$	$x \neq 0$
логарифмічна $y = a + b\,ln(x)$	$v = y$	$u = ln\,x$	$x > 0$
показникова $y = ae^{bx}$	$v = ln\,y$	$u = x$	$y > 0$, $a > 0$
степенева $y = ax^b$	$v = ln\,y$	$u = ln\,x$	$x > 0$, $y > 0$ $a > 0$
комбінована $y = \frac{1}{a + be^{-x}}$	$v = \frac{1}{y}$	$u = e^{-x}$	$y \neq 0$

Побудова згладжуючих поліномів має чітку та зрозумілу послідовність дій. Складність неполіноміальной апроксимації полягає у тому, що МНК в загальному вигляді не є обґрунтованим. Це означає, що рішення нормальної системи може не існувати, а якщо й існує, то не повинно бути точкою мінімуму. У кожному разі треба встановлювати існування рішення і перевіряти його на мінімум. Це в тому числі стосується і приведених вище прикладів.

Питання для самоперевірки:

1. Чим викликана необхідність апроксимування табличних функцій?
2. У чому принципова відмінність апроксимації від інтерполяції?
3. Сформулюйте задачу апроксимації. Що таке вузли, умови апроксимації, клас апроксимуючих функцій?
4. Сформулюйте задачу апроксимації методом найменших квадратів (МНК). Поясніть походження (принцип побудови) формули середньоквадратичного відхилення.
5. Що таке узагальнений багаточлен? Яким умовам повинні задовольняти базисні функції?

6. Вирішіть задачу МНК в класі узагальнених багаточленів. Що таке нормальна система МНК, матриця Грама, багаточлен найкращого середньоквадратичного відхилення?
7. За яких умов задача МНК в класі узагальнених багаточленів має рішення?
8. Сформулюйте та вирішіть задачу поліноміальної апроксимації МНК.
9. Чому інтерполяція є окремим випадком апроксимації?
10. Що таке лінійна і та квадратична апроксимація?
11. Що таке неполіноміальна апроксимація? У чому її практичне значення?
12. Які сімейства функцій використовуються для апроксимування?
13. У чому головна складність неполіноміальної апроксимації?

Тема 4.5 Рівномірне наближення функцій. Багаточлени Чебишева

ЛЕКЦІЯ 13. Багаточлени Чебишева. Вузли, що мінімізують похибку інтерполяції

Навчальні питання:

13.1 Визначення багаточленів Чебишева

13.2 Властивості багаточленів Чебишева

13.3 Мінімізація похибки інтерполяції

Повернемось до питань, що пов’язані з інтерполяцією. При інтерполюванні великих таблиць багаточленами, виникає проблема необмеженого зростання глобальної оцінки інтерполяції зі збільшенням кількості вузлів. Це означає, що у деяких частинах таблиці з багаточисленими вузлами похибка може бути дуже великою.

Розберемося з вирішенням задачі мінімізації глобальної похибки. До розв’язання такої похибки потрібно підходити з боку вибору вузлів, тобто підлаштовувати таблицю таким чином, щоб похибка стала мінімальною. Спочатку згадаємо, що похибку інтерполяції можливо оцінити наступним чином:

$$\left\| f(x) - P_n(x) \right\| \le \frac{\max\limits_{x \in (x_0, x_n)} \left| f^{(n+1)}(x) \right|}{(n+1)!} \cdot \left\| \omega_{n+1}(x) \right\|$$

За яких умов можна зменшити похибку інтерполяції, якщо ступінь багаточлена обрати заздалегідь? У даному випадку єдиним чим можна керувати – це значення норми $\left\| \omega_{n+1}(x) \right\|$. Отже потрібно обирати точки інтерполяції таким чином, щоб значення цієї норми було мінімальним.

Отже по таблиці значень функції треба побудувати інтерполяційний багаточлен, для якого глобальна оцінка похибки буде мінімально можливою. При цьому мінімізація буде проводитися за рахунок вибору вузлів таблиці.

Виявляється, що рішення дають такі математичні об'єкти як багаточлени Чебишова, які взагалі відіграють фундаментальну роль при оптимізації

обчислювальних алгоритмів. У цій лекції вони будуть застосовані для вирішення задачі рівномірного наближення заданої функції інтерполяційними багаточленами.

13.1 Визначення багаточленів Чебишева

Існують кілька визначень многочленів Чебишева. Почнемо з рекуррентного визначення.

Багаточлени Чебишева нульового та першого ступенів T_0, T_1, рівні $T_0(x)=1$, $T_1(x)=x$. Багаточлени ступеня n визначаються рекурентним співвідношенням

$$T_n(x)=2xT_{n-1}(x)-T_{n-2}(x) \qquad (13.1)$$

З нього отримуємо, наприклад, багаточлени кількох перших ступенів:

$$T_2(x)=2x^2-1,$$
$$T_3(x)=4x^3-3x,$$
$$T_4(x)=8x^4-8x^2+1,$$
$$T_5(x)=16x^5-20x^3+5x,$$
$$\ldots\ldots$$

Для прикладу розглянемо яким чином виводять $T_2(x)$ $T_3(x)$:

$$T_2(x)=2T_1(x)-T_0(x)=2x\cdot x-1=2x^2-1;$$
$$T_3(x)=2xT_2(x)-T_1(x)=2x\cdot(2x^2-1)-x=4x^3-3x,$$

і так далі.

З розглянутого прикладу витікає така властивість багаточлена Чебишева: *старший коефіцієнт T_n дорівнює 2^{n-1}*

Довести цю властивість не важко: Старший член T_n отримується з старшого члена T_{n-1} шляхом множення на $2x$. Враховуючи, що $T_0(x)=1$, приходимо до висновку, що старший член $T_n=2^{n-1}x^n$.

Також можна довести, що *поліноми T_{2n} є парними функціями, а T_{2n+1} – непарними.*

Доведемо цю властивість індукцією по $n\geq 0$ База індукції $T_0(x)=1$ – функція парна, $T_1(x)=x$ – функція непарна. Припустимо, що властивість вірно для всіх

$k \le n$. Доводимо індукційний крок. Якщо $n+1$ парне, то $2xT_n$ - парна функція як добуток двох непарних (n непарне, T_n непарна за припущенням), і тоді з (13.1) випливає, що $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$ - парна (різниця двох парних, $n-1$ парно, $T_{n-1}(x)$ парна за припущенням). Якщо ж $n+1$ непарне, то $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$ - непарна (як різниця двох непарних, $2xT_n$ - непарна як добуток непарної і парної).

Окрім рекурентного запису (13.1) для багаточленів Чебишева справедливим є тригонометричний тип запису:

$$T_n(x) = cos(n\,arccos\,x) \quad \text{при} \quad x \in [-1;1] \tag{13.2}$$

Згадаємо тригонометричні тотожності:

$$cos(\alpha \pm \beta) = cos\,\alpha \cdot cos\,\beta \mp sin\,\alpha \cdot sin\,\beta$$

Тепер доведемо (13.2). Розглянемо $T_{n+1}(x)$

$$\begin{aligned} T_{n+1}(x) &= cos\big((n+1)\,arccos\,x\big) = cos\big(n\,arccos\,x + arccos\,x\big) = \\ &= cos\big(n\,arccos\,x\big)\cdot cos\big(arccos\,x\big) - sin\big(n\,arccos\,x\big)\cdot\big(arccos\,x\big) = \\ &= 2\,cos\big(n\,arccos\,x\big)\cdot cos\big(arccos\,x\big) - \\ &- \underbrace{\big(cos\big(n\,arccos\,x\big)\cdot cos\big(arccos\,x\big) + sin\big(n\,arccos\,x\big)\cdot\big(arccos\,x\big)\big)}_{cos((n-1)\,arccos\,x)} = \\ &= 2x\underbrace{cos\big(n\,arccos\,x\big)}_{T_n(x)} - \underbrace{cos\big((n-1)\,arccos\,x\big)}_{T_{n-1}(x)}, \end{aligned}$$

що повністю збігається з визначенням (13.1). Отже тригонометричне представлення (13.2) справедливе. Розглянемо, як будуть виглядати багаточлени Чебишева для різних n (рис. 13.1):

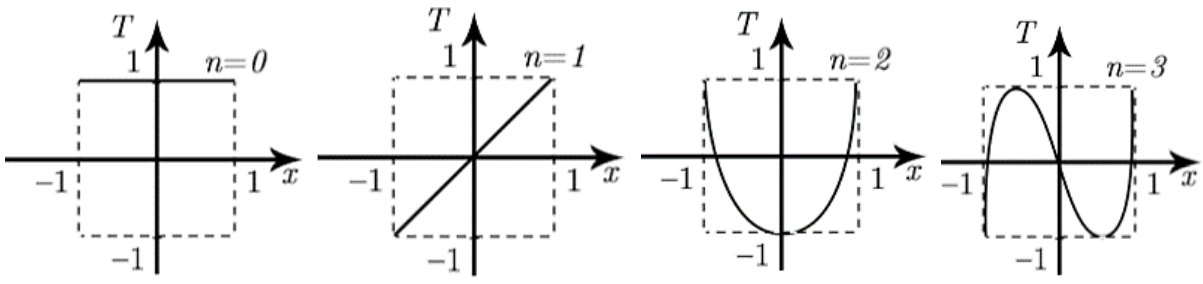


Рисунок 13.1 – Графік функції $T_n(x)$ для різних n

Усі багаточлени визначені на інтервалі $[-1;1]$ і для кожного виконується $\forall n \quad |T_n(x)| \le 1$.

13.2 Властивості багаточленів Чебишева

Для доведення властивостей багаточленів Чебишева будемо користуватись його тригонометричним записом (13.2). Визначимо *нулі багаточлену Чебишева* (корені):

$$T_n(x) = cos(n\,arccos\,x) = 0 \Rightarrow \; n\,arccos\,x = \frac{\pi}{2} + \pi m = \frac{\pi(1+2m)}{2},$$

$$arccos\,x = \frac{\pi(1+2m)}{2n}, \Rightarrow \quad x = cos\,\frac{\pi(1+2m)}{2n}, \tag{13.3}$$

де $m = 0,1,...,n-1$.

Рішення (13.3) має зрозумілу геометричну інтерпретацію (див. рис. 13.2). Згущення нулів відбувається біля країв інтервалів, у центрі нулі зустрічаються рідше.

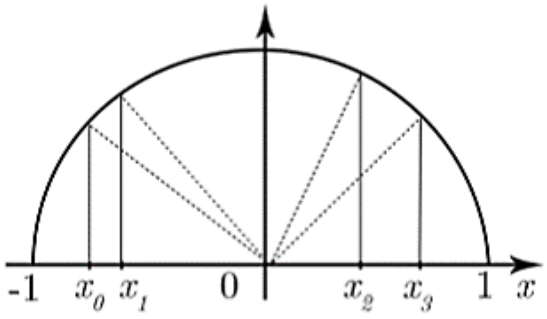


Рисунок 13.2 – геометрична інтерпретація пошуку нулів багаточлену Чебишева

Далі визначаємо *точки екстремуму багаточлену Чебишева*.

$$T_n(x) = cos(n\,arccos\,x) = \pm 1 \quad \Rightarrow \; n\,arccos\,x = \pi m$$

$$arccos\,x = \frac{\pi m}{n} \Rightarrow x = cos\,\frac{\pi m}{n} \tag{13.4}$$

де $m = 0,1,...,n-1$.

Визначення: нормованим (зведеним) багаточленом Чебишева будемо називати багаточлен, я якого старший коефіцієнт при x^n дорівнює одиниці:

$$\overline{T}_n(x) = 2^{1-n} T_n(x).$$

Виявляється, що $\overline{T}_n$ найменше ухиляється від нуля серед всіх нормованих багаточленів ступеня n на відрізку $[-1;1]$. Це можна строго сформулювати наступною лемою.

ЛЕМА 1. Нехай $P_n(x) = x^n + a_1 x^{n-1} + ... + a_n$ будь-який багаточлен зі старшим коефіцієнтом, що дорівнює одиниці. Тоді

$$\max_{-1\le x\le 1} |P_n(x)| \ge \max_{-1\le x\le 1} |\overline{T}_n(x)|$$

Серед усіх багаточленів, коефіцієнт при старшому ступеню яких дорівнює одиниці, багаточлен Чебишева має найменший максимум на відрізку [-1; 1]. Цей максимум дорівнює 2^{n-1}.

Доведення: Доведемо від протилежного. Нехай існує багаточлен $P_n(x)$ такий, що

$$\max_{-1\le x\le 1} |P_n(x)| < \max_{-1\le x\le 1} |\overline{T}_n(x)|$$

Розглянемо значення різниці: $T_n(x) - P_n(x)$ у точках екстремуму. Через те що $|T_n(x)| < |P_n(x)|$, то знак різниці буде співпадати зі знаком $T_n(x)$

$$sign\left(T_n(x) - P_n(x)\right) = (-1)^k, \quad k = 0,...,n.$$

Так як коефіцієнти при старших ступенях співпадають і дорівнюють одиниці, то

$$T_n(x) - P_n(x) = \tilde{P}_{n-1}(x)$$

де $\tilde{P}_{n-1}(x)$ - багаточлен ступеня n та має n нулів, що неможливо (адже ступінь багаточлену $n-1$. Отже отримали протиріччя, та умова $\max_{-1\le x\le 1} |P_n(x)| \ge \max_{-1\le x\le 1} |\overline{T}_n(x)|$ виконується.

Отже можна заключити що найважливіша властивість багаточленів Чебишева для вирішення задач полягає в тому, що нормовані багаточлени Чебишева є такими, що найменш відхиляється (по модулю) від нуля на даному відрізку серед всіх нормованих багаточленів даного ступеня.

13.3 Мінімізація похибки інтерполяції

Багаточлен Чебишева забезпечує мінімум функції $\omega_n(x)$, тобто:

$$min\left\|\omega_n(x)\right\| \rightarrow 2^{-n} T_{n+1}(x), \quad x \in [-1;1]$$

Якщо інтервал не одиничний, а є відрізком $[a;b]$, то відобразити $x \in [-1;1]$ на $\tilde{x} \in [a;b]$ можливо наступним чином:

$$\tilde{x} = \frac{a+b}{2} + \frac{b-a}{2} x.$$

Виразивши x через $\tilde{x}$, отримаємо:

$$x = \frac{2\tilde{x} - (a+b)}{b-a}.$$

Тоді багаточлен $\overline{T}_n(x)$, який є мінімальним на відрізку [-1;1], відобразиться в багаточлен

$$\overline{T}_n(x) \to \overline{T}_n\left(\frac{2\tilde{x} - (a+b)}{b-a}\right).$$

Старший коефіцієнт такого багаточлену дорівнює: $\frac{2^n}{(b-a)^n}$. Тоді на відрізку $[a;b]$:

$$\overline{T}_n^{(a+b)}(x) = \frac{2^n}{(b-a)^n} \overline{T}_n\left(\frac{2\tilde{x} - (a+b)}{b-a}\right). \qquad (13.5)$$

Цей багаточлен є таким, що найменш відхиляється від нуля на відрізку $[a; b]$. Було показано, що правильний вибір вузлів інтерполяції, відповідно з нулями багаточлена Чебишева, мінімізує норму $\|\omega(x)\|$. Таким чином, похибка інтерполяції зменшується.

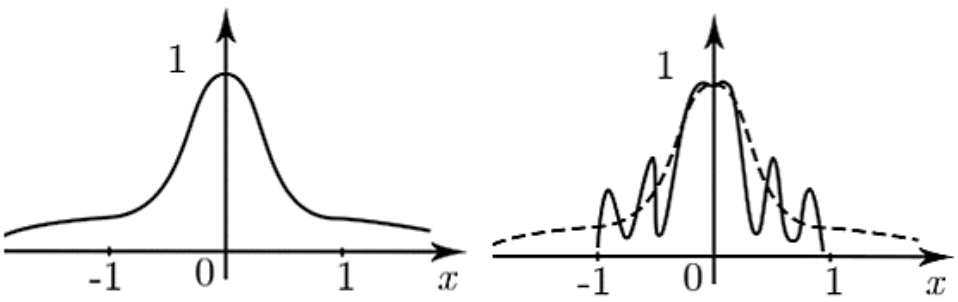
Далі постає задача найкращого наближення функції інтерполяційними багаточленами. Вона сформулюється в такий спосіб. Функція задана $y = f(x)$ формулою. Потрібно вибрати інтерполяційну сітку $\{x_j\}_{j=0}^n$ таким чином, щоб побудований за нею багаточлен $P_n(x)$ найменш ухилявся від $f(x)$ протягом всієї таблиці інтерполяції.

Для більш точного математичного формулювання введемо поняття *рівномірної норми* функції на відрізку. Нехай функція g визначена на відрізку $[a;b]$.

Величина $\|g\| = \sup\limits_{x\in[a;b]} |g(x)|$ називається рівномірною нормою на $[a;b]$. Тоді розглянута задача є проблемою мінімізації $\|f - P_n\|$ в класі поліномів P_n ступеня n на відрізку $[a;b]$, що містить всі вузли інтерполяції. Мінімізація проводиться за рахунок вибору інтерполяційної сітки $\{x_j\}_{j=0}^{n}$.

Важливо, що застосування рівномірної сітки з великим числом вузлів не дає рішення задачі. Більш того, при $n \to \infty$ величина $\|f - P_n\|$ може необмежено зростати. Це явище відоме як *феномен Рунге*.

Розглянемо функцію Рунге $f(x) = \dfrac{1}{1+25x^2}$, графік якої зображено на рис. 13.3 *а*. При виборі рівномірного розбиття на відрізку [-1;1] отримаємо незбіжний інтерполяційний процес рис. 13.3 *б.*



а) б)

Рисунок 13.3 – а) Графік функції Рунге, б) незбіжний інтерполяційний процес

Для цієї ж функція на відрізку [-1; 1] при розбитті відповідно до нулями багаточлена Чебишева значення функції та інтерполянта будуть збігатися. Такий вплив невеликих збурень на збіжність інтерполяційного процесу.

Питання для самоперевірки:

1. Запишіть рекурентне визначення багаточленів Чебишева.
2. Користуючись рекурентним визначенням, доведіть, що коефіцієнт при старшому члені багаточлена Чебишева ступеня n дорівнює 2^{n-1}.
3. Користуючись рекурентним визначенням, доведіть, що багаточлени Чебишева парних ступенів є парними функціями, а непарних - непарними.

4. Виведіть тригонометричне визначення багаточленів Чебишева. В яких межах укладені їх значення ?
5. Виведіть формулу багаточленів Чебишева, вирішивши рекурентне співвідношення для них. За яких існують багаточлени Чебишева, що визначаються цією формулою?
6. Переконайтеся безпосереднім обчисленням, що тригонометричний запис багаточленів Чебишева дає ці багаточлени при $n = 2,3,4.$
7. Користуючись тригонометричним визначенням, знайдіть корені багаточленів Чебишева на відрізку $[-1;1]$.
8. Користуючись тригонометричним визначенням, знайдіть точки екстремуму і екстремальні значення багаточленів Чебишева на відрізку $[-1;1]$.
9. Що таке нормований багаточлен?
10. Доведіть, що нормований багаточлен Чебишева мірою є найменшим, що ухиляється по модулю від нуля серед всіх нормованих багаточленів ступеня на відрізку.
11. Який багаточлен найменш ухиляється від нуля по модулю серед всіх нормованих багаточленів даного ступеня на довільному відрізку?

ЛЕКЦІЯ 14 Тригонометрична інтерполяція. Дискретне перетворення Фур’є

Навчальні питання:

14.1 Визначення перетворення Фур’є

14.2 Тригонометричний ряд Фур’є

14.3 Апроксимація і інтерполяція тригонометричними поліномами

Повернемося до методів наближення функцій. Але якщо раніше застосовували для наближення алгебраїчні поліноми, то тепер будемо використовувати для цього тригонометричні функції. Таким чином, вивчимо методи тригонометричної апроксимації та інтерполяції, які отримали широке застосування в цифровій обробці сигналів, автоматиці, радіотехніці, обробці зображень.

Для тригонометричної інтерполяції є добре розроблений математичний апарат - ряди Фур’є. Почнемо з викладу основ цієї теорії.

14.1 Визначення перетворення Фур’є

Перетворення Фур'є - це інтегральне перетворення, яке ставить у відповідність вихідної функції f дійсного аргументу t деяку іншу функцію дійсного аргументу ω за формулою:

$$f(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt. \qquad (14.1)$$

Функцію f називають Фур'є-образом функції f. Зворотне перетворення має вигляд:

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(\omega) e^{i\omega t} d\omega. \qquad (14.2)$$

Змінна t зазвичай позначає час, тобто f - це змінний у часі сигнал; ω - це частота гармонійної складової (гармоніки) сигналу f. Функція f описує коефіцієнти (амплітуди) розкладання f по гармонійкам.

Перетворення Фур'є має місце тільки для *безперервних функцій*. Ідея Фур'є полягала в тому, що безперервний періодичний сигнал може бути представлений

сумою обраних певним чином сигналів синусоїдальної форми. Зручність Фур'є-образів замість вихідних функцій в тому, що часто вихідні диференціальні рівняння після застосування перетворення Фур'є перетворюються в алгебраїчні.

Основні властивості перетворення Фур'є

1. $f^n(t) = (i\omega)^n f(\omega)$,

2. $f(t - t_0) = e^{-i\omega t_0} f(\omega)$,

3. $f(at) = |a|^{-1} f\left(\frac{\omega}{a}\right)$,

4. Фур'є-образ лінійної комбінації функцій є лінійною комбінацією Фур'є-образів цих функцій.

14.2 Тригонометричний ряд Фур'є

При аналізі періодичних процесів, що зустрічаються в радіотехніці та електроніці, періодичні функції розкладають в ряд Фур'є. Ряд Фур'є для функції f - це функціональний тригонометричний ряд

$$S(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nt + \sum_{n=1}^{\infty} b_n \sin nt, \qquad (14.3)$$

де $a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)dt$, $a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos ntdt$, $b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin ntdt$, $n = 1,2,...$

Його коефіцієнти обчислюються за функцією f. Вона повинна бути періодичною з періодом 2π. Ряд може бути побудований і для функцій з довільним періодом $T = 2l$ (l - напівперіод). В цьому випадку у ряді Фур'є проводиться заміна

$$u = t\frac{l}{\pi} \quad \Leftrightarrow \quad t = u\frac{\pi}{l},$$

тоді

$$S(u) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{\pi n}{l}u\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{\pi n}{l}u\right) \Rightarrow$$

$$\Rightarrow S(u) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2\pi n}{T}u\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{2\pi n}{T}u\right),$$

де $a_0 = \frac{1}{l}\int_{-l}^{l} f(u)du$, $a_n = \frac{1}{l}\int_{-l}^{l} f(u)\cos\left(\frac{\pi n}{l}u\right)du$, $b_n = \frac{1}{l}\int_{-l}^{l} f(u)\sin\left(\frac{\pi n}{l}u\right)du$, $n = 1,2,...$

Достатньою умовою існування ряду Фур'є функції f є її інтегрованість на відрізку [-π; π] ([-*l; l*] для функції з періодом $T = 2l$). *Умови збіжності ряду Фур'є* функції дає теорема Дирихле.

ТЕОРЕМА Дирихле. Якщо функція f кусково-безперервна і кусково-диференцьована на відрізку [-π; π], то її ряд Фур'є сходиться на всій числовій осі і його сума S є періодична функція з періодом 2π.

У точках безперервності t функції суми ряду $S(t)$ збігається з самою функцією: $S(t) = f(t)$. У точках розриву t_0 функції суми ряду дорівнює

$$S(t_0) = \frac{f(t_0 - 0) + f(t_0 + 0)}{2}.$$

У точках -π і π, тобто на кінцях відрізка [-π; π] сума ряду дорівнює:

$$S(-\pi) = S(\pi) = \frac{f(-\pi + 0) + f(\pi - 0)}{2}.$$

Ці результати сформульовані для функції з періодом 2π. Вони справедливі і для функцій з довільним періодом (з відповідними замінами величин).

14. 3 Апроксимація і інтерполяція тригонометричними поліномами

Далі поставимо і вирішимо задачу апроксимації табличної функції тригонометричними поліномами. Нехай дана таблична функція $y = f(x)$ (табл. 14.1).

Таблиця 14.1 – значення функції $y = f(x)$ у точках

i	x_i	y_i
0	x_0	y_0
1	x_1	y_1
….	….	….
n	x_n	y_n

Тригонометричним поліномом називається часткова сума тригонометричного ряду:

$$S_k(x) = \frac{a_0}{2} + \sum_{i=1}^{k} a_i \cos ix + \sum_{i=1}^{k} b_i \sin ix. \qquad (14.4)$$

Число *k* називається ступенем полінома, a_i, b_i - коефіцієнти поліному. Проте це часткова сума ряду Фур'є для функції з періодом 2π. Необхідно перевести її на період $T = x_n - x_0$.

Нехай $y_n = y_0$. Тоді функцію *f* можна вважати періодичної з періодом $T = x_n - x_0$. Поза таблиці вона довизначається по періодичності. Тригонометричний поліном підстановки $u = x\frac{l}{\pi} \Leftrightarrow x = u\frac{\pi}{l}$, де $l = \frac{T}{2} = \frac{x_n - x_0}{2}$, приводиться до виду:

$$S_k(x) = \frac{a_0}{2} + \sum_{i=1}^{k} a_i \cos\left(\frac{\pi i}{l} x\right) + \sum_{i=1}^{k} b_i \sin\left(\frac{\pi i}{l} x\right) =$$
$$= \frac{a_0}{2} + \sum_{i=1}^{k} a_i \cos\left(\frac{2\pi i}{T} x\right) + \sum_{i=1}^{k} b_i \sin\left(\frac{2\pi i}{T} x\right).$$

Задача полягає в побудові тригонометричного поліному S_k ступеня *k*, що наближає функцію *f*. У такій постановці вона некоректна: потрібно з точністю визначити наближення функції поліномом. Зробимо це так, як в методі найменших квадратів: побудувати тригонометричний поліном, для якого середньоквадратичне відхилення від табличних значень буде найменшим. далі введемо деякі умови на вихідні дані, щоб більш точно сформулювати завдання.

Нехай таблиця функції *f* рівномірна, тобто вузли йдуть з постійним кроком *h*. Робимо заміну $t = \frac{x - x_0}{h}$, тоді точки x_j, які можна назвати вузлами апроксимації, перейдуть у множину $\{0,1,...,n\}$:

$$t_j = \frac{x_j - x_0}{h} \Rightarrow t_j = \frac{x_0 + jh - x_0}{h} = j$$

$j = 0,1,...,n$; період функції T по новій змінній t дорівнює n, а напівперіод $\frac{n}{2}$.

Тоді введемо у розгляд багаточлен S_k від нової змінної:

$$S_k(t) = \frac{a_0}{2} + \sum_{i=1}^{k} a_i \cos\left(\frac{2\pi i}{T} t\right) + \sum_{i=1}^{k} b_i \sin\left(\frac{2\pi i}{T} t\right) =$$

$$= \frac{a_0}{2} + \sum_{i=1}^{k} \left(a_i \cos\left(\frac{2\pi i}{T} t \cdot \frac{x - x_0}{h}\right) + b_i \sin\left(\frac{2\pi i}{T} t \cdot \frac{x - x_0}{h}\right)\right).$$

Тоді вважається, що

$$f(x) \approx S_k(t) = S_k\left(\frac{x - x_0}{h}\right). \qquad (14.5)$$

Тригонометричний багаточлен визначений своїми коефіцієнтами $a_0, a_1, ..., a_k, b_1, ..., b_k$ тому величина, яку потрібно мінімізувати, залежить також від них та визначається формулою:

$$T(a_0, a_1, ..., a_k, b_1, ..., b_k) = \sum_{j=0}^{n} \left(S_k(j) - y_j\right)^2. \qquad (14.6)$$

Сума квадратів відхилень поліному від табличних значень.

Задача методу найменших квадратів тепер сформулюється так: побудувати тригонометричний поліном S_k, для якого значення T буде мінімальним. Потрібно знайти $2k+1$коефіцієнт $a_0, a_1, ..., a_k, b_1, ..., b_k$. Відомо, що задача має єдине рішення, якщо $2k+1 < n+1$, або $k \le \frac{n}{2}$.

Приймемо це без доведення і будемо вирішувати задачу за цієї умови. Тригонометричний багаточлен є за формою узагальнений багаточлен по системі тригонометричних функцій

$$1, \cos\left(\frac{2\pi}{T} t\right), ..., \cos\left(\frac{2\pi(k-1)}{T} t\right), \cos\left(\frac{2\pi k}{T} t\right),$$

$$\sin\left(\frac{2\pi}{T} t\right), ..., \sin\left(\frac{2\pi(k-1)}{T} t\right), \sin\left(\frac{2\pi k}{T} t\right)$$

Відомо, що ця система є ортогональною на відрізку $[0; T]$ (або $[0; n]$:

$$\sum_{j=0}^{n} sin\left(\frac{2\pi mj}{T}\right) sin\left(\frac{2\pi pj}{T}\right) = \begin{cases} 0, \ якщо \ m \neq p, \\ \frac{n+1}{2}, якщо \ m = p \neq 0, \end{cases}$$

$$\sum_{j=0}^{n} cos\left(\frac{2\pi mj}{T}\right) cos\left(\frac{2\pi pj}{T}\right) = \begin{cases} 0, \ якщо \ m \neq p, \\ \frac{n+1}{2}, якщо \ m = p \neq 0, \\ n+1, \ якщо \ m = p = 0, \end{cases} \qquad (14.7)$$

$$\sum_{j=0}^{n} sin\left(\frac{2\pi mj}{T}\right) cos\left(\frac{2\pi pj}{T}\right) = 0.$$

Система функцій *називається ортогональною*, якщо скалярний добуток будь-яких попарно різних функцій з неї дорівнює нулю. Скалярний твір вводиться будь-яким чином в даному функціональному просторі. Тут скалярний добуток є сума попарних добутків значень функцій у вузлах:

$$\varphi\psi = \sum_{j=0}^{n} \varphi(j)\psi(j).$$

Скористаємося результатом, отриманим раніше в лекції 12 про апроксимацію. Тоді було виведено рішення: вектор коефіцієнтів узагальненого многочлена знаходиться з нормальної системи МНК

$$\Gamma \bar{a} = \bar{b}$$

де $\Gamma = P^T P$ - матриця Грама, $\bar{b} = P^T \bar{y}$, (P^T - транспонована матриця P)

$$P = \begin{pmatrix} \varphi_0(t_0) & \varphi_1(t_0) & \dots & \varphi_{2k}(t_0) \\ \varphi_0(t_1) & \varphi_1(t_1) & \dots & \varphi_{2k}(t_1) \\ & & \dots & \\ \varphi_0(t_n) & \varphi_1(t_n) & \dots & \varphi_{2k}(t_n) \end{pmatrix}, \qquad \bar{a} = \begin{pmatrix} \frac{a_0}{2} \\ a_1 \\ \dots \\ a_k \\ b_1 \\ \dots \\ b_k \end{pmatrix}, \quad \bar{y} = \begin{pmatrix} y_0 \\ y_1 \\ \dots \\ y_n \\ \end{pmatrix}.$$

При цьому у відповідності зі структурою вектору $\bar{a}$ базисні функції позначено наступним чином:

$$: \bar{a} = \begin{pmatrix} \frac{a_0}{2} \\ a_1 \\ \dots \\ a_k \\ b_1 \\ \dots \\ b_k \end{pmatrix}, \Rightarrow \begin{cases} \varphi_0(t) = 1, \\ \varphi_1(t) = cos\left(\frac{2\pi}{T}t\right), \\ \dots \\ \varphi_k(t) = cos\left(\frac{2\pi k}{T}t\right), \\ \varphi_{k+1}(t) = sin\left(\frac{2\pi}{T}t\right), \\ \dots \\ \varphi_{2k}(t) = sin\left(\frac{2\pi k}{T}t\right). \end{cases}$$

Обчислимо матрицю Грама для тригонометричної системи базисних функцій:

$$\|\Gamma\|_{mp} = \varphi_m(j)\varphi_p(j) = 0,\ m \neq p,$$
$$\|\Gamma\|_{11} = \varphi_0^2 = n+1,$$
$$\|\Gamma\|_{mm} = \varphi_{m-1}^2 = \frac{n+1}{2},\ m = 2,\dots,2k+1.$$

Позадіагональні елементи дорівнюють нулю через ортогональність, а діагональні обчислюються за формулами (14.7). У підсумку матриця Грама буде діагональною

$$\Gamma = P^T P = \begin{pmatrix} \varphi_0^2 & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots \\ \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & \varphi_{2k}^2 \end{pmatrix} = \begin{pmatrix} n+1 & 0 & \dots & 0 \\ 0 & \frac{n+1}{2} & \dots & \dots \\ \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & \frac{n+1}{2} \end{pmatrix}.$$

Обчислимо вектор правих частин

$$\bar{b} = P^T \bar{y} = \begin{pmatrix} \varphi_0(t_0) & \varphi_0(t_1) & \dots & \varphi_0(t_n) \\ \varphi_1(t_0) & \varphi_1(t_1) & \dots & \varphi_1(t_n) \\ & & \dots & \\ \varphi_{2k}(t_0) & \varphi_{2k}(t_1) & \dots & \varphi_{2k}(t_n) \end{pmatrix} \cdot \begin{pmatrix} y_0 \\ y_1 \\ \dots \\ y_n \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & cos\left(\frac{2\pi}{T}\right) & \dots & cos\left(\frac{2\pi n}{T}\right) \\ \dots & \dots & \dots & \dots \\ 1 & cos\left(\frac{2\pi k}{T}\right) & \dots & cos\left(\frac{2\pi kn}{T}\right) \\ 0 & sin\left(\frac{2\pi}{T}\right) & \dots & sin\left(\frac{2\pi n}{T}\right) \\ \dots & \dots & \dots & \dots \\ 0 & sin\left(\frac{2\pi k}{T}\right) & \dots & sin\left(\frac{2\pi kn}{T}\right) \end{pmatrix} \cdot \begin{pmatrix} y_0 \\ y_1 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{j=0}^{n} y_j \\ \sum_{j=0}^{n} y_j \, cos\left(\frac{2\pi j}{T}\right) \\ \dots \\ \sum_{j=0}^{n} y_j \, cos\left(\frac{2\pi kj}{T}\right) \\ \sum_{j=0}^{n} y_j \, sin\left(\frac{2\pi j}{T}\right) \\ \dots \\ \sum_{j=0}^{n} y_j \, sin\left(\frac{2\pi kj}{T}\right) \end{pmatrix}.$$

Отримуємо систему з діагональною матрицею, яка дуже легко розв'язується. У підсумку маємо коефіцієнти тригонометричного полінома найкращого середньоквадратичного відхилення:

$$a_0 = \frac{2}{n+1} \sum_{j=0}^{n} y_j \,, \qquad a_i = \frac{2}{n+1} \sum_{j=0}^{n} y_j \, cos\left(\frac{2\pi ij}{T}\right), \; i = 1,\dots,k$$

$$b_i = \frac{2}{n+1} \sum_{j=0}^{n} y_j \, sin\left(\frac{2\pi ij}{T}\right), \; i = 1,\dots,k$$

де $T = n$ - період функції. Будуємо багаточлен:

$$S_k(t) = \frac{a_0}{2} + \sum_{i=1}^{k} a_i \, cos\left(\frac{2\pi i}{T} t\right) + b_i \, sin\left(\frac{2\pi i}{T} t\right),$$

з найденими коефіцієнтами та обчислюємо наближено $f(x)$ як $S_k\left(\frac{x - x_0}{h}\right)$

$$f(x) \approx S_k(t) = S_k\left(\frac{x - x_0}{h}\right)$$

Як було сказано вище, при $k \le \frac{n}{2}$ задача має єдине рішення. Якщо $k = \frac{n}{2}$ (n - парне), то апроксимуючий поліном перетворюється в інтерполяційний, та

$S_k\left(x_j\right) = y_j$, $j = 0, 1, ..., n$. Це вже тригонометрична інтерполяція табличної функції. Якщо n непарне, то інтерполяція має місце для ступеня полінома $k = \frac{n+1}{2}$.

Нехай $n+1 = 2N$, тобто число вузлів парне. Однозначно визначається апроксимація, як і для алгебраїчного полінома, має місце при $k \le N$. Найбільший ступінь полінома може бути N. При $k = N$ апроксимуючий поліном перетворюється в інтерполяційний і $S_N\left(x_j\right) = y_j$. В цьому випадку $h = \frac{\pi}{N}$ (функція f періодична з періодом $T = 2\pi$, розглядається на відрізку [-π; π], вузли йдуть від - π до π з постійним кроком). Тоді f інтерполюється полиномом

$$f\left(x\right) \approx S_N\left(\frac{x - x_0}{h}\right) = \frac{a_0}{2} + \sum_{i=1}^{k} a_i \cos i\left(x - x_0\right) + b_i \sin i\left(x - x_0\right) + a_N \cos N\left(x - x_0\right),$$

$$a_i = \frac{1}{N}\sum_{j=0}^{2N-1} f\left(x_j\right)\cos\left(\frac{\pi ij}{N}\right),\ i = 1,...,N-1,\ \ b_i = \frac{1}{N}\sum_{j=0}^{2N-1} f\left(x_j\right)\sin\left(\frac{\pi ij}{N}\right),\ i = 1,...,N-1,$$

$$a_N = \frac{1}{2N}\sum_{j=0}^{2N-1} f\left(x_j\right)\cos \pi j.$$

Разом коефіцієнтів $2N$, що визначаються з $2N$ умов у точках $-\pi, -\pi + \frac{\pi}{N}, ..., \pi - \frac{\pi}{N}$. Отримуємо тригонометричний інтерполюючий поліном ступеня N.

Питання для самоперевірки:

1. Що таке перетворення Фур'є? У чому сенс перетворення функції з часової області в частотну? Що показують значення Фур'є-образу?
2. Для яких функцій можливо перетворення Фур'є?
3. Перерахуйте властивості перетворення Фур'є
4. Що таке тригонометричний ряд Фур'є? Який його зв'язок з перетворенням Фур'є?
5. Для яких функцій існує ряд Фур'є?
6. Як переводиться ряд Фур'є для функції з періодом на функцію з довільним періодом?

7. Які достатні умови збіжності ряду Фур’є для функції?

8. Чому дорівнює сума ряду Фур’є?

9. Що таке тригонометричний поліном?

10. Сформулюйте задачу апроксимації табличної функції тригонометричними поліномами. Поясніть ідею її виникнення

11. Як будується тригонометричний поліном найкращого середньоквадратичного відхилення?

12. За яких умов апроксимуючий поліном стає інтерполюючим?

13. Як оцінюється похибка поліноміальної інтерполяції?

14. У чому перевага тригонометричної інтерполяції перед алгебраїчною?

РОЗДІЛ 5. НАБЛИЖЕНЕ ІНТЕГРУВАННЯ

Тема 5.1 Формули чисельного інтегрування

ЛЕКЦІЯ 15. Чисельне інтегрування. Найпростіші квадратурні формули

Навчальні питання:

15.1 Постановка задачі чисельного інтегрування

15.2 Найпростіші формули чисельного інтегрування

15.2.1 Формули прямокутників

15.2.2 Формула трапецій

15.2.3 Формула парабол (Сімпсона)

Починаємо вивчення основ чисельного інтегрування – найпростіших формул. Звичайно, почнемо з постановки задачі чисельного інтегрування, усвідомимо причини її виникнення. Про практичне значення Вам, напевно, добре відомо: інтеграли мають численні геометричні, фізичні, технічні та інші додатки.

Виведемо найпростіші формули, виходячи з найпростіших геометричних міркувань, послідовно демонструючи загальні принципи побудови квадратурних формул чисельного інтегрування. До простих відносяться формули прямокутників, трапецій і парабол (Сімпсона). З них тільки формули лівих і правих прямокутників не застосовуються на практиці, так що ці найпростіші формули корисно вивчити не тільки в методичних цілях.

15.1Постановка задачі чисельного інтегрування

Інтегрування – це знаходження первісної функції; чисельно можна вирішувати тільки задачі розрахунку визначеного інтеграла. Відомо, що визначений інтеграл можна точно обчислити за формулою Ньютона-Лейбніца.

Для її застосування необхідно знайти первісну, тобто взяти невизначений інтеграл, що не завжди можливо. Наприклад, інтеграл може не виражатися в елементарних функціях. Крім того, інтегрована функція може бути задана таблично, і формула Ньютона-Лейбніца непридатна.

Отже, задача виникає в тих випадках, коли:

• інтеграл не можливо представити у вигляді елементарних функціях;

• підінтегральна функція задана таблично або обчислюється експериментально;

• застосування формули Ньютона-Лейбніца є важким;

• деякі параметри інтеграла задані наближено.

Перейдемо до задачі наближеного обчислення інтегралів.

Нехай функція $y = f(x)$ визначена і інтегрована на відрізку $[a;b]$.

Необхідно знайти значення визначеного інтеграла $I = \int_a^b f(x)dx$, коли первісна $F(x)$ ($F'(x) = f(x)$) невідома або її важко знайти, або $y = f(x)$ задана своїми значеннями $y_i = f(x_i)$, $i = \overline{0,n}$, $x_i \in [a;b]$.

Загальний підхід в чисельному інтегруванні полягає в наступному: для функції $y = f(x)$ будується апроксимуюча функція $F(x)$, так щоб $f(x) \approx F(x)$ на відрізку $[a;b]$, при цьому клас апроксимуючої функції $F(x)$ може залежати

- від властивостей функції $y = f(x)$,
- від необхідної точності обчислення інтеграла,
- від числа арифметичних дій,
- від часу роботи алгоритму і т. д.;

Функція $F(x)$ вибирається так, щоб інтеграл $\int_a^b F(x)dx$ легко рахувався, або замінюється відомими вам інтерполяційними поліномами

Функція $F(x)$ вибирається так, щоб

$$I = \int_a^b f(x)dx \approx \int_a^b F(x)dx \quad \text{або} \quad \left| I - \int_a^b F(x)dx \right| \le \varepsilon,$$

де ε - задана точність обчислення інтеграла. Потрібно розібратись який підхід потрібно обрати.

Як відомо, *визначений інтеграл* – це границя інтегральних сум при прагненні до нуля максимального розміру часткових відрізків розбивки. Тому очевидна ідея його наближення - заміна сумою.

Визначення будь-яка проста формула, що апроксимує окремий інтеграл I_i називається *квадратурною*. (тобто формула для числового обчислення однократного інтегралу). Формула для числового обчислення подвійного інтегралу називається *кубатурною*.

Для застосування методів числового інтегрування ділять відрізок $[a;b]$ системою рівновіддалених точок $x_k = x_0 + kh$, $h = \frac{b-a}{n}$, $k = \overline{0,n}$, $x_0 = a$, $x_n = b$ на відрізки $[x_k, x_{k+1}]$, $k = \overline{0,n-1}$ і розглядають суму інтегралів

$$I = \sum_{k=0}^{n-1} I_k = \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} f(x)dx . \qquad (15.1)$$

Квадратурна формула, наближено замінює інтеграл сумою, отже:

$$\int_a^b f(x)dx = \sum_{k=0}^{n} A_k \cdot f(x_k) + R, \qquad (15.2)$$

де x_k - вузли інтегрування, деякі точки на відрізку інтегрування [*a, b*] , n - їх кількість; A_k - коефіцієнти, залежні від вибору вузлів; називаються вагою, R – залишковий член або похибка квадратурної формули.

Складова квадратурна формула - це формула, яка дає наближення інтегралу $I(f)$ у вигляді суми наближень інтегралами I_i по даній квадратурній формулі.

15.2 Найпростіші формули чисельного інтегрування

Найпростіші формули засновані на геометричному змісті, що визначений інтеграл на відрізку від *а* до *в* від функції $f(x)$ є площа криволінійної трапеції, обмеженою зліва вертикальною прямою $x = a$, праворуч – прямою $x = в$, віссю абсцис знизу і зверху – кривою $f(x)$

15.2.1 Формули прямокутників

Перша найпростіша ідея - замінити криволінійну трапецію прямокутником (рис. 15.1). Нехай - площа k-ї криволінійної трапеції, побудованої на k-му відрізку розбиття $[x_k, x_{k+1}]$.

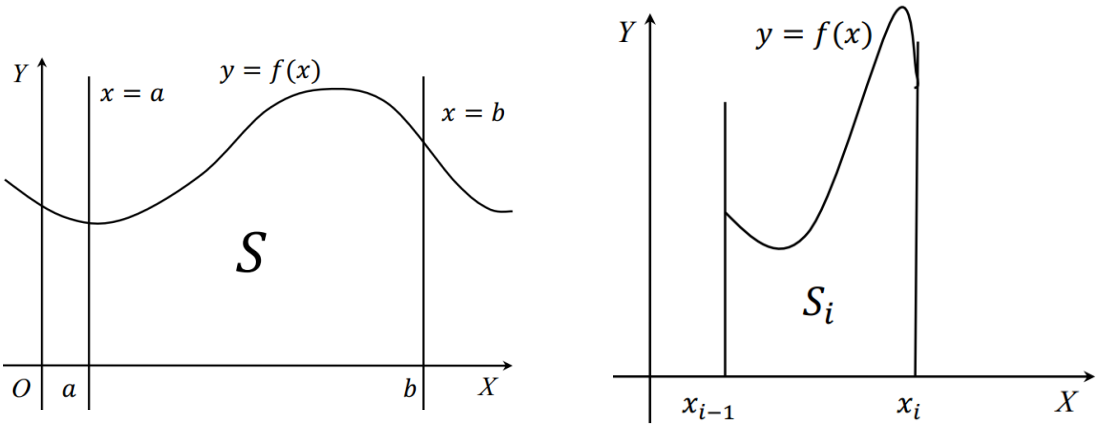


Рисунок 15.1 – геометричний зміст визначеного інтегралу

У цьому випадку довжина кожного з n відрізків розбивки постійна, дорівнює $\frac{b-a}{n}$ і називається *кроком інтегрування*.

Позначивши $h = \frac{b-a}{n}$. Підставляючи це значення в формулу для обчислення площі криволінійної трапеції через площі часткових трапецій, отримуємо ***квадратурну формулу лівих прямокутників:*** У цьому випадку $f(x)$ на відрізку $[x_k, x_{k+1}]$ замінюється функцією $F(x) = f(x_k)$, тоді

$$I_k \approx \int_{x_k}^{x_{k+1}} f(x_k)dx = f(x_k)\int_{x_k}^{x_{k+1}} dx = f(x_k)[x_{k+1} - x_k] = h \cdot f(x_k),$$

$$I = \sum_{k=0}^{n-1} h \cdot f(x_k) = h\sum_{k=0}^{n-1} f(x_k), \qquad (15.3)$$

оцінка похибки здійснюється у такому випадку як: $\Delta_I \le \frac{n \cdot h^2}{2} \max_{x\in[a,b]} |f'(x)|$.

Формула правих прямокутників. В цьому випадку $f(x)$ на відрізку $[x_k, x_{k+1}]$ замінюється функцією $F(x) = f(x_{k+1})$, тоді:

$$I_k \approx \int_{x_k}^{x_{k+1}} f(x_{k+1})dx = f(x_{k+1})\int_{x_k}^{x_{k+1}} dx = f(x_k)[x_{k+1} - x_k] = h \cdot f(x_{k+1}),$$

$$I = \sum_{k=0}^{n-1} h \cdot f\left(x_{k+1}\right) = h \sum_{k=0}^{n-1} f\left(x_{k+1}\right), \tag{15.4}$$

оцінка похибки $\Delta_I \le \frac{n \cdot h^2}{2} \max\limits_{x \in [a,b]} \left| f'(x) \right|$.

Формула центральних прямокутників. Нарешті, третя формула прямокутників виходить заміною k-ї криволінійної трапеції прямокутником з тою ж основою – відрізком $[x_k, x_{k+1}]$ та висотою $F(x) = f\left(\frac{x_k + x_{k+1}}{2}\right)$ (рис. 15.2).

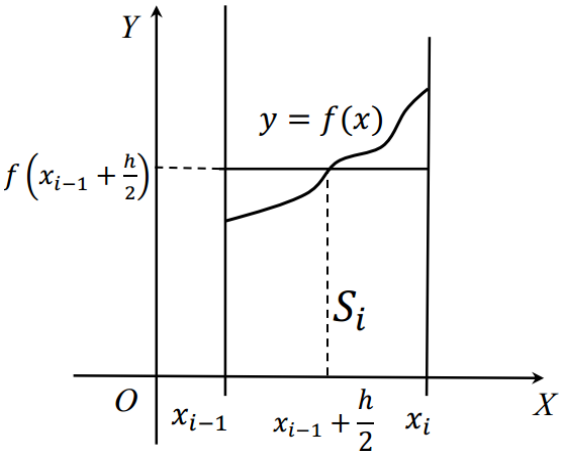


Рисунок 15.2 – до пояснення виводу формули центральних прямокутників

Тоді

$$I_k \approx \int_{x_k}^{x_{k+1}} f\left(\frac{x_k + x_{k+1}}{2}\right) dx = f\left(\frac{x_k + x_{k+1}}{2}\right) \int_{x_k}^{x_{k+1}} dx =$$

$$= f\left(\frac{x_k + x_{k+1}}{2}\right)\left[x_{k+1} - x_k\right] = h \cdot f\left(\frac{x_k + x_{k+1}}{2}\right)$$

$$I = \sum_{k=0}^{n-1} h \cdot f\left(\frac{x_k + x_{k+1}}{2}\right) = h \sum_{k=0}^{n-1} f\left(\frac{x_k + x_{k+1}}{2}\right), \tag{15.5}$$

оцінка похибки $\Delta_I \le \frac{n \cdot h^3}{24} \max\limits_{x \in [a,b]} \left| f''(x) \right|$.

15.2.2 Формула трапецій

Тепер замінимо k-ту криволінійну трапецію звичайною трапецією . Її паралельні основи – відрізки довжин $f(x_k)$, $f(x_{k+1})$, бокові сторони $[x_k, x_{k+1}]$ та хорда, що стягує дугу кривої, тобто відрізок, що поєднує точки з координатами $\left(x_k; f(x_k)\right)$, $\left(x_{k+1}; f(x_{k+1})\right)$ Висота цієї трапеції дорівнює довжині відрізка $[x_k, x_{k+1}]$, тобто h (рис. 15.3).

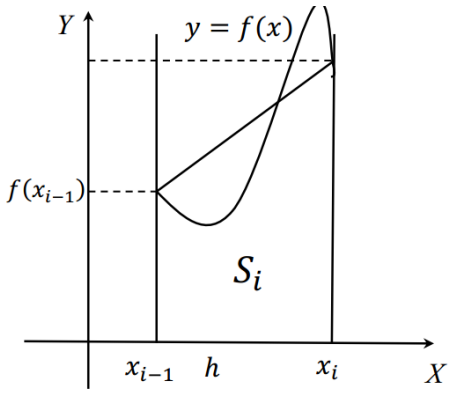


Рисунок 15.3 – до пояснення виводу формули трапецій

В цьому випадку $f(x)$ на відрізку $[x_k, x_{k+1}]$ замінюється функцією $F(x) = \frac{1}{2}\left[f(x_k) + f(x_{k+1}) \right]$,

тоді

$$I_k \approx \int_{x_k}^{x_{k+1}} \frac{1}{2}\left[f\left(x_k\right) + f\left(x_{k+1}\right)\right]dx = \frac{1}{2}\left[f\left(x_k\right) + f\left(x_{k+1}\right)\right] \int_{x_k}^{x_{k+1}} dx = \frac{h}{2}\left[f\left(x_k\right) + f\left(x_{k+1}\right)\right]$$

$$I = \sum_{k=0}^{n-1} \frac{h}{2}\left[f\left(x_k\right) + f\left(x_{k+1}\right)\right] = \frac{h}{2}\sum_{k=0}^{n-1}\left[f\left(x_k\right) + f\left(x_{k+1}\right)\right], \qquad (15.6)$$

оцінка похибки $\Delta_I \le \frac{n \cdot h^3}{12} \max\limits_{x\in[a,b]} \left| f''(x)\right|$.

15.2.3 Формула парабол (Сімпсона)

А тепер замінимо криволінійну трапецію іншою криволінійною трапецією, але такою, що її площу можна просто і точно обчислити за формулою Ньютона-Лейбніца. А саме, нехай крива з довільної перетвориться в параболу. Парабола задається квадратним тричленом ($F\left(x\right) = cx^2 + dx + e$), інтеграл від якого легко обчислюється.

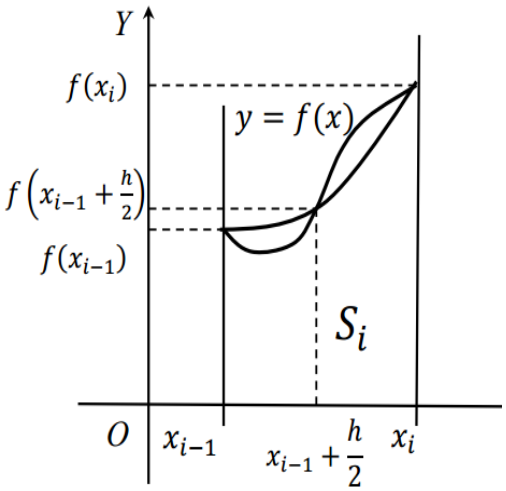


Рисунок 15.4 – до пояснення виводу формули парабол

Візьмемо на дузі кривої три точки $\left(x_{i-1}; f\left(x_{i-1}\right)\right), \left(x_{i-1}+\frac{h}{2}; f\left(x_{i-1}+\frac{h}{2}\right)\right), \quad \left(x_i; f\left(x_i\right)\right)$ (рис. 15.4).

Як відомо, через три різних точки площині можна провести єдину параболу. Її рівняння можна знайти за допомогою інтерполяційних формул. Наприклад, оскільки вузли, $x_{i-1}, x_{i-1}+\frac{h}{2}, x_i$, йдуть рівномірно, можна застосувати формулу Ньютона з кінцевими різницями:

$$P_{2,i}\left(x\right) = f\left(x_{i-1}\right) + \Delta f_{i-1} t + \frac{t\left(t-1\right)}{2}\Delta^2 f_{i-1},$$

$$t = 2\frac{x - x_{i-1}}{h}, \Delta f_{i-1} = f\left(x_{i-1}+\frac{h}{2}\right) - f\left(x_{i-1}\right),$$

$$\Delta^2 f_{i-1} = \Delta f_{i-\frac{1}{2}} - \Delta f_{i-1} = f\left(x_i\right) - 2f\left(x_{i-1}+\frac{h}{2}\right) + f\left(x_{i-1}\right).$$

Тоді площа криволінійної трапеції дорівнює

$$S_i^* = \int_{x_{i-1}}^{x_i} P_{2,i}(x)\,dx = \int_{x_{i-1}}^{x_i} \left(f(x_{i-1}) + \Delta f_{i-1} t + \frac{t(t-1)}{2} \Delta^2 f_{i-1} \right) dx =$$

$$= \frac{h}{2} \int_0^2 \left(f(x_{i-1}) + \Delta f_{i-1} t + \frac{t(t-1)}{2} \Delta^2 f_{i-1} \right) dt =$$

$$= \frac{h}{2} \left(2 f(x_{i-1}) + 2\Delta f_{i-1} + \frac{1}{2} \Delta^2 f_{i-1} \right) =$$

$$= \frac{h}{6} \left(6 f(x_{i-1}) + 6\left(f(x_i) - f(x_{i-1}) \right) + f(x_{i-1}) - 2 f\left(x_{i-1} + \frac{h}{2} \right) + f(x_i) \right) =$$

$$= \frac{h}{6} \left(f(x_{i-1}) + f\left(x_{i-1} + \frac{h}{2} \right) + f(x_i) \right).$$

Підставляючи в формулу для обчислення площі криволінійної трапеції через площі часткових трапецій, отримуємо квадратурну формулу парабол, або Сімпсона:

$$I \approx I^* = \frac{h}{6} \left(f(a) + f(b) + 4 \sum_{i=1}^{n} f\left(x_{i-1} + \frac{h}{2} \right) + 2 \sum_{i=1}^{n-1} f(x_i) \right). \quad (15.7)$$

Зауважимо, що якщо інтегрована функцій є поліном другого ступеня, то формула парабол дає точне значення інтегралу.

Питання для самоперевірки

1. Сформулюйте задачу наближеного обчислення визначеного інтеграла. У яких випадках вона виникає?
2. Що таке квадратурна формула? Які її параметри? Як Ви поясните її походження?
3. У чому полягає геометричний зміст визначеного інтеграла?
4. Поясніть геометричний принцип побудови квадратурних формул
5. Виведіть формули лівих, правих і центральних прямокутників
6. Виведіть формули трапецій і парабол. Чим наближається підінтегральна функція в цих випадках?

ЛЕКЦІЯ № 16. Квадратурні формули Ньютона-Котеса та формули Гауса

Навчальні питання:

16.1Виведення формул Ньютона-Котеса

16.2 Похибки квадратурних формул

16.3 Квадратурні формули Гауса

У попередній лекції вирішена задача чисельного інтегрування, використовуючи геометричний зміст визначеного інтегралу. Але існують інші методи розв'язку, більш складні але і більш точні. Виведемо формули, що мають зрозумілий зміст, достатньо легко виводяться та широко застосовуються у розрахунках. Спочатку виведемо формули Ньютона-Котеса. Ці формули виводяться інтерполюванням підінтегральної функції деяким багаточленом.

16.1 Виведення формул Ньютона-Котеса

Постановка задачі залишається попередньою. Потрібно обчислити наближене значення інтегралу

$$I = \int_a^b f(x)dx$$

та оцінити похибки наближення.

Згадуємо, що загальна ідея рішення полягає в наближенні інтеграла квадратурною сумою

$$I^* = \sum_{i=0}^{n} A_i f(x_i)$$

x_i – вузли, а A_i – ваги формули

Формула Ньютона-Котеса відносяться до формул інтерполяційного типу, це значить, що підінтегральну функцію f інтерполюють на проміжку $[a;b]$ деякою функцією, інтеграл від якої легко обчислюється. Тоді наближене значення шуканого інтеграла дорівнює інтегралу від інтерполюючої функції.

Нехай вузли інтерполяції x_i йдуть з постійним кроком h:

$$x_i = a + ih, \qquad i = 0,1,...,n, \qquad h = \frac{b-a}{n}$$

Замінюємо функцію f інтерполяційним поліномом Лагранжа

$$f(x) = L_n(x) + R_n(x)$$

де
$$L_n(x) = \sum_{i=0}^{n} y_i \frac{\omega_{n+1}(x)}{(x - x_i)\omega_n(x)}, \tag{16.1}$$

$$\omega_{n+1}(x) = \prod_{i=0}^{n}(x - x_i), \tag{16.2}$$

$$\omega_n(x) = \prod_{\substack{i=0 \\ j \neq i}}^{n}(x_i - x_j), \qquad y_i = f(x_i) \tag{16.3}$$

У виразі (16.1) – поліном Лагранжа $L_n(x)$ у скороченій формі. $R_n(x)$ – залишковий член наближення, що дорівнює похибці інтерполяції. Згадаємо що він визначається наступним чином

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\omega_{n+1}(x) \tag{16.4}$$

$\xi(x) \in (a;b)$ – деяка точка, що належить відрізку інтегрування та визначається для кожного *x.* Тепер можна записати:

$$\int_a^b f(x)dx = \int_a^b L_n(x)\,dx + \int_a^b R_n(x)\,dx,$$

Інтеграл від залишкового члена *є похибка наближеного інтегрування*.

Виконуємо заміну змінної $q = \frac{x - x_0}{h}$,

тоді:
$$x = x_0 + qh, \tag{16.5}$$

$x_0 = a$. Перейдемо від x до q у всіх різницях в (16.2), (16.3):

$$x - x_i = x_0 + qh - (x_0 + ih) = (q - i)h,$$

$$x_i - x_j = (i - j)h,$$

$$\omega_{n+1}(x) = q(q-1)...(q-n)h^{n+1},$$

$$\omega_n'(x_i) = i(i-1)...1 \cdot (-1)...(i-n)h^n = (-1)^{n-i}\, i!\,(n-i)!\,h^n. \tag{16.6}$$

Тоді багаточлен Лагранжа (16.1) можна записати у вигляді

$$L_n\left(x_0+qh\right)=\sum_{i=0}^{n}\frac{(-1)^{n-i}\,y_i}{i!\left(n-i\right)!}\cdot\frac{q\left(q-1\right)\ldots\left(q-n\right)}{q-i}. \tag{16.7}$$

Оскільки при $x = a$ $q = 0$, а при $x = b$ $q = n$, інтегрування за новою змінною q буде йти від 0 до n. Наближене значення інтеграла рахуємо, як інтеграл від багаточлена Лагранжа (16.7). Тоді, маючи на увазі, що $dx = hdq$, отримуємо

$$I\approx I^{*}=h\int_0^n\sum_{i=0}^{n}\frac{(-1)^{n-i}\,y_i}{i!\left(n-i\right)!}\cdot\frac{q\left(q-1\right)\ldots\left(q-n\right)}{q-i}dq.$$

Це і є формула Ньютона -Котеса. Її можна записати у вигляді

$$I\approx I^{*}=\left(b-a\right)\sum_{i=0}^{n}H_i y_i, \tag{16.8}$$

$$H_i=\frac{1}{n}\cdot\frac{(-1)^{n-i}}{i!\left(n-i\right)!}\int_0^n\frac{q\left(q-1\right)\ldots\left(q-n\right)}{q-i}dq- \tag{16.9}$$

це коефіцієнти Котеса

Розглянемо декілька часткових випадків.

Нехай $n=1$, $x_0 = a$, $x_1 = b$ $h=b-a$. Обчислюємо коефіцієнти Котеса:

$$H_0=\frac{-1}{1}\int_0^1\frac{q\left(q-1\right)}{q}dq=\frac{1}{2};\qquad H_1=\int_0^1\frac{q\left(q-1\right)}{q-1}dq=\frac{1}{2}.$$

Підставляючи їх у (16.8), одержуємо наближене значення інтеграла:

$$I^{*}=\left(b-a\right)\frac{y_0+y_1}{2}=\left(b-a\right)\frac{f\left(a\right)+f\left(b\right)}{2}. \tag{16.10}$$

Це знайома нам формула трапецій. Правда, вона тут застосована до всього відрізку інтегрування. Але якщо по ній обчислити інтеграл на частковому відрізку, а потім просумувати ми отримаємо рівно ту ж складову формулу з минулого лекції.

Тепер візьмемо інтерполяційний багаточлен Лагранжа другого ступеня, тобто $n = 2$ тоді

$$h=\frac{b-a}{2},\quad x_0=a,\quad x_1=\frac{a+b}{2},\quad x_2=b.$$

Рахуємо коефіцієнти Котеса (16.9):

$$H_0 = \frac{1}{2}\cdot\frac{1}{2!}\int_0^2 \frac{q(q-1)(q-2)}{q}dq = \frac{1}{6};\quad H_1 = \frac{1}{2}\cdot\frac{(-1)}{1}\int_0^2 \frac{q(q-1)(q-2)}{q-1}dq = \frac{2}{3};$$

$$H_2 = \frac{1}{2}\cdot\frac{1}{2!}\int_0^2 \frac{q(q-1)(q-2)}{q-2}dq = \frac{1}{6}.$$

Тоді за формулою (16.8) отримуємо наближення інтеграла:

$$I^* = (b-a)\left(\frac{1}{6}y_0 + \frac{2}{3}y_1 + \frac{1}{6}y_2\right) = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right). \qquad (16.11)$$

Це формула Сімпсона. Обчисливши по ній приблизно інтеграл на кожному частковому відрізку і склавши отримані значення, прийдемо до формули, отриманої в попередній лекції.

Тепер нехай $n = 3$, тоді $h = \frac{b-a}{3}$. Коефіцієнти Котеса при цьому дорівнюють:

$$H_0 = \frac{1}{3}\cdot\frac{(-1)}{3!}\int_0^3 (q-1)(q-2)(q-3)\,dq = \frac{1}{8};$$

$$H_1 = \frac{1}{3}\cdot\frac{1}{2!}\int_0^3 q(q-2)(q-3)\,dq = \frac{3}{8};\quad H_2 = \frac{1}{3}\cdot\frac{(-1)}{2!}\int_0^3 q(q-1)(q-3)\,dq = \frac{3}{8};$$

$$H_3 = \frac{1}{3}\cdot\frac{1}{3!}\int_0^3 q(q-1)(q-2)\,dq = \frac{1}{8};$$

звідси отримуємо наближений інтеграл:

$$I^* = \frac{b-a}{8}\left(f(a) + 3f(a+h) + 3f(a+2h) + f(b)\right). \qquad (16.12)$$

Формула (16.12) називається «формула три восьмих».

Отже, квадратурні формули Ньютона-Котеса утворюються інтегруванням інтерполяційного багаточлена, що наближає підінтегральну функцію.

Загальний вид квадратурних формул Ньютона-Котеса наступний:

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n} A_i^{(n)} y_i,$$

$$A_i^{(n)} = (b-a)B_i^{(n)},\qquad B_i^{(n)} = \frac{1}{n}\cdot\frac{(-1)^{n-i}}{i!(n-i)!}\int_0^n \frac{q(q-1)\ldots(q-n)}{q-i}dq,$$

$y_i = f(x_i) = f(a + ih)$. Коефіцієнти $B_i^{(n)}$, як неважко помітити, *не залежать від відрізка інтегрування*, тому їх можна заздалегідь обчислити. Деякі з них, отримані в прикладах вище, наведені в таблиці 16.1.

Таблиця 16.1 – значення коефіцієнтів $B_i^{(n)}$ для різних квадратурних формул

n	$B_0^{(n)}$	$B_1^{(n)}$	$B_2^{(n)}$	$B_3^{(n)}$	Назва формули
1	$\frac{1}{2}$	$\frac{1}{2}$			Трапецій
2	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$		Сімпсона
3	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	«три-восьмих»

16.2 Похибки квадратурних формул

За похибку квадратурної формули відповідає залишковий член $R_n(x)$. Для оцінки похибки зробимо ту ж заміну (16.5), (16.6) в залишковому члені (16.4):

$$R_n(x) = \frac{f^{(n+1)}\left(\xi(x)\right)}{(n+1)!}\omega_{n+1}(x) = \frac{f^{(n+1)}\left(\xi(q)\right)}{(n+1)!}h^{n+1}q(q-1)\cdots(q-n),$$

$\xi(q)$ в силу заміни буде вже залежати від q. Тоді

$$I - I^* = \int_a^b R_n(x)dx = \frac{h^{n+2}}{(n+1)!}\int_0^n f^{(n+1)}\left(\xi(q)\right)q(q-1)\cdots(q-n)\,dq. \qquad (16.13)$$

Для обчислення інтеграла в цьому виразі можна застосувати теореми про середнє значення інтеграла. При неможливості застосування цієї теореми можна отримати верхню оцінку абсолютної похибки, використовуючи властивості модуля і інтеграла. Сформулюємо теорему про середнє.

ТЕОРЕМА про середнє значення інтегралу (без доведення): Нехай функції f і g інтегровані на відрізку $[a;b]$, $m_1 \le f(x) \le m_2$, $x \in [a;b]$; g – знакостала на цьому відрізку. Тоді існує таке число γ, що $m_1 \le \gamma \le m_2$ та

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

Тепер для першого розглянутого прикладу, візьмемо в якості f функцію f'', а g нехай це $q(q-1)$. Нехай функція f двічі неперервно диференційована на $[a;b]$. Тоді f'' задовольняє умові теореми, крім того, для будь-якого γ, такого, що

$$\min_{x\in[a;b]} f''(x) \le \gamma \le \max_{x\in[a;b]} f''(x),$$

існує таке число $\xi \in [a;b]$, що $\gamma = f''(\xi)$. Добуток $q(q-1)$ є від'ємним на $[0;1]$. Тоді

$$I - I^* = \int_a^b R_n(x)dx = \frac{h^{n+2}}{(n+1)!}\int_0^n f^{(n+1)}\big(\xi(q)\big)q(q-1)\cdots(q-n)\,dq.$$

$$\int_0^1 f''\big(\xi(q)\big)q(q-1)\,dq = f''(\xi)\int_0^1 q(q-1)\,dq = -\frac{f''(\xi)}{6},$$

залишковий член інтегрування дорівнює:

$$I - I^* = -\frac{(b-a)^3}{12} f''(\xi),$$

$\xi \in [a; b]$. Тут зроблена підстановка $b\text{-}a = h$. Прийшли до оцінки похибки формули трапецій.

$$\Delta I^* = \left|I - I^*\right| \le \frac{(b-a)^3}{12} M_2 = \frac{M_2}{12}(b-a)h^2,$$

де $M_2 = \max_{x\in[a;b]}\left|f''(x)\right|$. Оцінки похибок цієї та інших найпростіших квадратурних формул зведені у табл. 16. 2.

Таблиця 16. 2 – Оцінки похибок квадратурних формул

Квадратурна формула	Оцінка похибки
Лівих, правих прямокутників	$\Delta I^* \le \frac{M_1}{2}(b-a)h. \qquad M_1 = \max_{x\in[a;b]}\left\|f'(x)\right\|$
Центральних прямокутників	$\Delta I^* = \left\|I - I^*\right\| \le \frac{M_2}{24}(b-a)h^2. \quad M_2 = \max_{x\in[a;b]}\left\|f''(x)\right\|$
Трапецій	$\Delta I^* = \left\|I - I^*\right\| \le \frac{M_2}{12}(b-a)h^2. \quad M_2 = \max_{x\in[a;b]}\left\|f''(x)\right\|$
Парабол (Сімпсона)	$\Delta I^* = \left\|I - I^*\right\| \le \frac{M_4}{2880}(b-a)h^4. \quad M_4 = \max_{x\in[a;b]}\left\|f^{(4)}(x)\right\|$

16.3 Квадратурні формули Гауса

Отже, ознайомились з двома способами виведення формул чисельного інтегрування. Перший з них – найпростіші формули. Цей підхід базувався на

геометричному змісті визначеного інтегралу. Другий тип – квадратурні формули Ньютона-Котеса, які отримують заміною підінтегральної функції інтерполяційним багаточленом.

Підхід який запропонував Гаус розглянемо далі. Гаус запропонував обирати ваги та вузли квадратурної формули таким чином, щоб формула була точною для багаточленів якомога більш високого ступеню. Формули, які побудовані на базі цього принципу і називають гаусовими квадратурами. Виводом таких формул далі й будемо займатися. Для початку розглянемо прості приклади. Нехай потрібно обчислити інтеграл:

$$I = \int_{-1}^{1} f(x)dx$$

Побудуємо квадратуру за двома вузлами:

$$I^* = A_1 f(x_1) + A_2 f(x_2)$$

Потрібно знайти чотири невідомих, A_1, A_2, x_1, x_2, тобто два вузла і дві ваги. Вимагатимемо, щоб для поліномів ступеня від 0 до m квадратура давала точні значення інтегралів. Очевидно, що формула буде точна для будь-якого багаточлена P_m ступеня m тоді і тільки тоді, коли вона буде точна для всіх $x^k, k = 0,...,m$. Обчислимо точні значення цих інтегралів:

$$I_k = \int_{-1}^{1} x^k dx = \frac{1}{k+1}\left(1-(-1)^{k+1}\right).$$

Отже, маємо такі рівняння для визначення невідомих:

$$A_1 x_1^k + A_2 x_2^k = \frac{1}{k+1}\left(1-(-1)^{k+1}\right) \qquad (16.14)$$

Невідомих чотири, значить, і рівнянь має бути чотири, для ступенів поліномів 0,1,2,3. Таким чином, ця квадратура з двома вузлами буде точна для багаточленів не вище третього ступеня, і це максимально можливий ступінь. Маємо систему рівнянь

$$A_1 x_1^k + A_2 x_2^k = \frac{1}{k+1}\left(1-(-1)^{k+1}\right) \quad k = 0,1,2,3$$

або у розгорнутому вигляді, розв'язуючи цю нелінійну систему, отримуємо

$$\begin{cases} A_1 + A_2 = 2, \\ A_1 x_1 + A_2 x_2 = 0, \\ A_1 x_1^2 + A_2 x_2^2 = \dfrac{2}{3}, \\ A_1 x_1^3 + A_2 x_2^3 = 0. \end{cases} \Rightarrow \begin{cases} A_1 = 1, \\ A_2 = 1, \\ x_1 = -\dfrac{1}{\sqrt{3}}, \\ x_1 = \dfrac{1}{\sqrt{3}}. \end{cases}$$

Це означає, що наближене значення інтеграла дорівнює

$$I^* = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right),$$

причому воно збігається з точним значенням для будь-якого багаточлена від нульового до третього ступеня.

Якщо ж кількість точок $n = 3$ то система рівнянь (16.14) для невідомих ваг і вузлів квадратури набуде вигляду

$$A_1 x_1^k + A_2 x_2^k + A_2 x_3^k = \frac{1}{k+1}\left(1 - (-1)^{k+1}\right)$$

$k = 0,1,2,3,4,5$. 6 невідомих та 6 рівнянь, після розв'язку системи, отримаємо, що інтеграли обчислюють за формулою:

$$I^* = \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right)$$

що є точною для багаточлена не вище п'ятого ступеня.

Вищенаведені приклади ілюструють те, на що у свій час вказував Гаус. Що *можна так підібрати вузли та ваги квадратури, що вона буде точною для багаточленів не вище вказаного (якогось заданого) ступеня.*

Так як в нашому розпорядженні є *2n* постійних x_i та A_i, а поліном ступеня *2n -1* визначається *2n* коефіцієнтами, то ця найвища ступінь в загальному випадку, очевидно, дорівнює *2n - 1*. Доведено, що Гаусові квадратури, дозволяють будувати формули на основі *n* точок, що будуть точні для поліномів ступені *2n-1*.

У зв'язку з тим, що отримані системи нелінійні, а розв'язок таких систем викликає певні труднощі, були розроблені певні прийоми, які використовують різного типу ортогональні поліноми у якості інтерполяційної (вагової) функції. Далі нам будуть потрібні деякі відомості про поліноми Лежандра.

Поліноми виду

$$P_n(x)=\frac{1}{2^n n!}\frac{d^n}{dx^n}\left[\left(x^2-1\right)^n\right]\ \ n=0,1,2,...,$$

Називають поліномами Лежандра. Для прикладу приведемо п’ять перших поліномів Лежандра

$$P_0(x)=1,\ P_1(x)=x,\ P_2(x)=\frac{1}{2}\left(3x^2-1\right),\ P_3(x)=\frac{1}{2}\left(5x^3-3x\right),\ P_4(x)=\frac{1}{8}\left(35x^4-30x^2+3\right)$$

Якщо у якості вагової функції обрати поліном Лежандра:

$$I=\int_a^b P(x)f(x)dx$$

тоді ваги для квадратури Гауса будуть обчислюватись як:

$$A_i=\int_a^b P(x)\frac{\omega_n(x)}{(x-x_i)\omega'_{n-1}(x_i)}dx$$

Взагалі, квадратури Гауса також відносяться до інтерполяційних. Вони виводяться з умови точності для поліномів якомога більш високого ступеня при даному числі вузлів інтерполяції. Досягається це вибором вузлів і ваг квадратури. *Основні властивості квадратурних формул Гауса наступні:*

1. всі ваги A_i додатні;

2. алгебраїчний ступінь точності дорівнює $2n-1$, тобто вони точні для всіх багаточленів ступеня не вище $2n-1$ і неточні для багаточленів ступеня $2n$ і вище;

3. мають найвищий ступінь точності, тобто ніяка квадратура по n вузлам не може мати алгебраїчну ступінь точності більшу, ніж $2n-1$;

4. при заданому n квадратура Гауса єдина.

Важливо, що всі ці властивості виконуються за таких умов на вагову функцію:

1. $P(x)\geq 0,\ x\in[a;b]$

2. інтегрованість на $[a;b]$, причому $\int_a^b P(x)dx>0$.

У таблиці 16.3 наведені вузли та ваги квадратур Гауса для інтегралу $I=\int_{-1}^{1} f(x)dx$.

Якщо необхідно обчислити інтеграл на відрізку $[a;b]$, то коефіцієнти залишаться попередніми, але вузли потрібно перерахувати за формулою: $t_i = \frac{a+b}{2} + \frac{b-a}{2} x_i$.

Таблиця 16.3 – Розраховані значення ваг та вузлів квадратур Гауса

n	i	x_i	A_i
2	1; 2	−0,577350; 0,577350	1
3	1; 3	−0,774597; 0,774597	5/9
	2	0	8/9
4	1; 4	−0,861136; 0,861136	0,347855
	2; 3	−0,339981; 0,339981	0,652145
5	1; 5	−0,906180; 0,906180	0,236927
	2; 4	−0,538469; 0,538469	0,478629
	3	0	0,568889

На останок зазначимо, що Гаусові квадратури ефективні тільки для гладких підінтегральних функцій *f*(*x*), інакше квадратурна формула втратить частину своєї високої точності.

Питання для самоперевірки:

1. У чому принцип побудови квадратурних формул Ньютона-Котеса?
2. Як оцінюється похибка формул Ньютона-Котеса?
3. Виведіть формулу трапецій як окремий випадок квадратури Ньютона-Котеса та оцінку її похибки. За яких умов на підінтегральну функцію вірна ця оцінка?
4. Виведіть формулу Сімпсона як окремий випадок квадратури Ньютона-Котеса.
5. Виведіть квадратурну формулу "три восьмих"
6. Які Ви знаєте приклади застосування інших інтерполяційних поліномів для отримання квадратур Ньютона-Котеса?

7. Чому формули Ньютона-Котеса не застосовуються для інтегрування на всьому відрізку при великих n?
8. У чому полягає принцип побудови квадратур Гауса?
9. Для поліномів якого ступеню можна забезпечити точність квадратури з вузлами n за рахунок вибору вузлів і ваг?
10. Сформулюйте задачу на побудову квадратур, поставлену Гаусом
11. Що таке інтерполяційна квадратурна формула?
12. Що таке вагова функція? Які умови накладаються на неї для вирішення задачі Гауса?
13. Як знаходяться вузли квадратури Гауса? При яких умовах?
14. Як знаходяться ваги квадратури Гауса?
15. Які Ви знаєте властивості квадратур Гауса?
16. Як оцінюється похибка квадратури Гауса?

СПИСОК ВИКОРИСТАНОЇ ТА РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

1. Бахвалов Н.С., Жидков Н. П., Кобельков Г.М. Численные методы. - М.: Наука, 1987.-600с.
2. Березин И.С., Жидков Н.П. Методы вычислений:В 2 т. – М.: Физматгиздат, 1962. – 640с.
3. Волков А.Е. Численные методы. – М.:Наука, 1982. -248с.
4. Воробьева Г.Н., Данилова А.Н. Практикум по вычислительной математике. – М.:Высш. шк.., 1990. – 308с.
5. Гловацкая А. П. Методы и алгоритмы вычислительной математики. - М.:Радио и связь, 1999.- 406 с.
6. Демидович Б.П., Марон И.А. Основы вычислительной математики. – М.:Наука. 1970. – 665с.
7. Демидович Б.П., Марон И.А., Шувалова Э.З. Численные методы анализа. – М.: Наука, 1962. – 367с.
8. Денис Дж., Шнабель Р. Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1988. – 249с.
9. Заварыкин В.М., Житомирский В.Г.,Лапчик М.П. Численные методы. – М.:Просвещение, 1991.-176 с.
10. Калиткин И. Н. Численне методы. – М.: Наука, 1978. – 500с.
11. Копченова Н. В., Марон И. А. Вычислительная математика в примерах и задачах. – М.: Наука, 1972. – 368с.
12. Мак-Кракен Д., Дорн У. Численные методы и программирование на ФОРТРАНЕ. – М.: Мир, 1977. – 293с.
13. Марчук Г.И. Методы вычислительной математики. – М.: Наука, 1977. – 456 с.
14. Мэтьюз Джон Д., Финк Куртин Д. Численные методы. Использование Matlab: Пер. с англ. – 3-е изд. – М.: Издат. Дом «Вильямс», 2001. – 720с.
15. Петергеря Ю.С., Соболєв О.В., Абакумова О.О. Обчислювальна математика: Навч. посібник / К.: НТУУ «КПІ», 2007. – Ч.1. – 92 с
16. Поршнев С.В. Вычислительная математика. Курс лекций. – СПб.: БХВ-Петербург, 2004. – 320 с.
17. Самарский А. А., Гулин А. В. Численные методы: Учеб, пособие для вузов,— М.: Наука. Гл. ред. физ-мат. лит., 1989.— 432 с.
18. Турчак Л. И. Основы численных методов. – М.: Наука, 1987. – 350с.
19. Форсайт Дж., Малькольм М., Моулер К. Машинные методы математических вычислений. – М.: Мир, 1980. – 279с.
20. Шуп Т. Е. Решение инженерных задач на ЭВМ. – М.: Мир, 1990. – 235с.