

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Одеська політехніка»
Інститут комп'ютерних систем
Кафедра прикладної математики та інформаційних технологій

КОНСПЕКТ ЛЕКЦІЙ
з дисципліни «**ОБЧИСЛЮВАЛЬНІ МЕТОДИ**»
для здобувачів першого (бакалаврського) рівня вищої освіти за спеціальністю
F3 Комп'ютерні науки
освітня програма «Інтелектуальний аналіз даних»

Одеса – 2025

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Одеська політехніка»
Інститут комп'ютерних систем
Кафедра прикладної математики та інформаційних технологій

КОНСПЕКТ ЛЕКЦІЙ
з дисципліни «ОБЧИСЛЮВАЛЬНІ МЕТОДИ»
для здобувачів першого (бакалаврського) рівня вищої освіти за спеціальністю
F3 Комп'ютерні науки
освітня програма «Інтелектуальний аналіз даних»

Затверджено на засіданні
кафедри ПМІТ
Протокол № 1 від 29.08.2025 р.

Одеса – 2025

Конспект лекцій по курсу з дисципліни «Обчислювальні методи» для здобувачів першого (бакалаврського) рівня вищої освіти за спеціальністю F3 Комп'ютерні науки освітня програма «Інтелектуальний аналіз даних» / Укладачі: Н.А.Гуляєва. - Одеса: Національний університет «Одеська політехніка», 2025. – 74 с.

Укладач: Гуляєва Н.А., ст.викладач

Пропонований конспект лекцій призначений для студентів технічних спеціальностей. Його мета в доступній та стислій формі ознайомити студентів з основними положеннями по даній дисципліні. Викладений матеріал ілюструється прикладами розв'язання задач. Наведено список рекомендованої літератури.

Призначається для студентів денної форми навчання.

Зміст

ЛЕКЦІЯ 1. Предмет та об'єкт обчислювальних методів. Характеристики методів та чим вони відрізняються.	6
ЛЕКЦІЯ 2. Наближенні обчислювання та похибки. Абсолютна та відносна похибки	10
ЛЕКЦІЯ 3. Значущі, вірні й сумнівні цифри. Похибки при арифметичних діях з наближеними числами	13
ЛЕКЦІЯ 4. Розв'язання нелінійних рівнянь. Алгебраїчні та трансцендентні функції. Рівняння з одним невідомим.	18
ЛЕКЦІЯ 5. Розв'язання нелінійних рівнянь .Метод поділу навіпіл (дихотомії), комбінований метод хорд і дотичних та ітераційний метод (послідовних наближень)	22
ЛЕКЦІЯ 6. Розв'язання систем рівнянь. Системи лінійних рівнянь. Метод виключення Гауса та метод Жордана -Гаусса.	28
ЛЕКЦІЯ 7. Ітераційні методи розв'язання систем лінійних рівнянь. Метод простих ітерацій.	33
ЛЕКЦІЯ 8. Системи нелінійних рівнянь. Метод Ньютонa та модифікований метод Ньютонa.	35
ЛЕКЦІЯ 9. Обчислювальне інтегрування. Метод прямокутників. Метод трапецій. Метод Симпсона.	38
ЛЕКЦІЯ 10. Обчислювальні методи оптимізації. Постановка задачі і класифікація методів	44
ЛЕКЦІЯ 11. Лінійне програмування. Геометричний метод розв'язання задачі ЛП	51
ЛЕКЦІЯ 12 Розв'язання диференціальних рівнянь. Постановка задачі. Метод Ейлера. Метод Рунге Кутта.	55
ЛЕКЦІЯ 13. Апроксимація функцій. Метод найменших квадратів.	59
ЛЕКЦІЯ 14. Інтерполяція функції. Види інтерполяцій функції.	64
ЛЕКЦІЯ 15. Лінійні емпіричні формули. Лінеаризація емпіричної формули.	70
Список літератури	74

Вступ

Курс «Обчислювальні методи» викладається студентам спеціальності F3 Комп'ютерні науки освітня програма «Інтелектуальний аналіз даних» у п'ятому навчальному семестрі. Робоча навчальна програма з дисципліни передбачає наприкінці семестру складання екзамену.

Конспект лекцій призначений для підготовки студентів до занять з дисципліни «Обчислювальні методи», поглиблення теоретичних та практичних навичок.

Загальні положення

Відповідно до навчальної програми з дисципліни «Обчислювальні методи» передбачається п'ятнадцять лекційних занять. Кожне лекційне заняття - перелік основних понять та категорій, якими має володіти студент після роботи на відповідному занятті.

Мета лекційних занять – придбання студентами навиків самостійної роботи при розгляді задач. Проведення кожного лекційного заняття починається з короткого опитування з теорії, що привчає до систематичного вивчення матеріалу курсу за конспектами лекцій або навальними посібниками.

ЛЕКЦІЯ 1. Предмет та об'єкт обчислювальних методів. Характеристики методів та чим вони відрізняються.

Розвиток техніки і науки тісно пов'язано з використанням комп'ютерів. Їх застосування дозволяє від простих розрахунків і оцінок перейти до нової стадії роботи: детальному математичному моделюванню і дослідженню складних реальних процесів і об'єктів. Дослідження моделі проводиться за допомогою спеціально розроблених чисельних методів, призначених для розв'язання більш простих математичних проблем обчислення інтегралів, розв'язання скалярних і диференціальних рівнянь і їхніх систем, апроксимації функцій, оптимізації різного роду функцій.

Для розв'язання кожної задачі існує кілька методів. Навчитися правильно відбирати, використовувати потрібний метод це є задачею чисельних методів. Для розв'язання перерахованих проблем потрібна розробка докладного алгоритму, що зводить всі обчислення до послідовності арифметичних і логічних дій. Вибір того чи іншого алгоритму здійснюється з урахуванням пам'яті і швидкодії комп'ютера, так як занадто складний метод приводить до нереальних витрат часу роботи комп'ютера, а занадто простий найчастіше не дає необхідної точності.

Вивчення чисельних методів сприяє також переосмисленню і більш глибокому розумінню математики в цілому, оскільки однією із задач чисельних методів є сведення методів вищої математики до виконання простих арифметичних операцій.

Хоча існує безліч чисельних методів, усі вони мають багато спільних властивостей і характеристик. Чисельні методи:

- передбачають проведення великої кількості рутинних арифметичних обчислень за допомогою рекурсивних співвідношень, що використовуються для організації *ітерацій*, тобто повторюваних циклів обчислень зі зміненими початковими умовами для поліпшення результату;
- направлені на локальне спрощення задачі, коли, наприклад, використовувані нелінійні залежності лінеаризуються за допомогою своїх обчислених похідних або похідні замінюються різницеви́ми апроксимаціями;
- значно залежать від близькості початкового наближення, необхідного для початку обчислень до розв'язку, від властивостей нелінійних функцій, які використовуються в математичних моделях, що накладає обмеження (для забезпечення єдиного розв'язку) на їх диференційованість, на швидкість зміни функції та ін.

Чисельні методи характеризуються:

- різною *швидкістю збіжності*, тобто числом ітерацій, виконання яких необхідне для отримання заданої точності розв'язку;
- різною *стійкістю*, тобто збереженням достовірності розв'язку під час подальших ітерацій;
- різною *точністю* отриманого розв'язку в разі виконання однакового числа ітерацій або циклів обчислень.

Чисельні методи розрізняються:

- за широтою і легкістю застосування, тобто за ступенем своєї *універсальності* та *інваріантності* для розв'язання різних математичних задач;
- за *складністю* їх програмування;
- за *можливостями використання* у разі їх реалізації наявних бібліотек функцій і процедур, створених для підтримки різних алгоритмічних мов;
- за *ступенем чутливості* до погано обумовлених (або некоректних) математичних задач, коли малим змінам вхідних даних можуть відповідати великі зміни розв'язку.

Тобто *об'єктом* вивчення даної навчальної дисципліни є типові математичні задачі, до яких зводиться рішення практичних проблем, що виникають у ході розробки інформаційних систем та систем моделювання. *Предметом* вивчення даної навчальної дисципліни є чисельні методи розв'язання типових математичних задач.

Одним із способів розв'язання практичної задачі є експеримент. Наприклад, будують ракету, запускають і перевіряють характеристики, які цікавлять. Якщо вони не задовольняють, то будують нову ракету і т. д. Зрозуміло, що потрібний результат (при необмежених ресурсах) буде, врешті решт, досягнутий, проте занадто дорогою ціною.

Інший спосіб - побудова математичної (комп'ютерної) моделі об'єкта або явища, що вивчається, і проведення всіх розрахунків на комп'ютері. Для фізичних процесів математична модель зазвичай записується у вигляді набору рівнянь, в які як коефіцієнти входять характеристики тіл або речовин, що беруть участь у процесі.

Для розв'язання прикладної задачі потрібно використовувати математичну модель. Під математичною моделлю фізичної системи, об'єкта або процесу розуміють сукупність математичних співвідношень (формул, рівнянь, логічних виразів), які визначають характеристики стану і властивості системи, об'єкта і процесу та їх функціонування залежно від

параметрів їх компонентів, початкових умов, вхідних збуджень і часу. Тобто, математична модель описує функціональну залежність між вихідними залежними змінними, через які відображається функціонування системи, незалежними (такими, як час) і змінюваними змінними (такими, як параметри компонентів, геометричні розміри), а також вхідними збудженнями, прикладеними до системи. Ця функціональна залежність, що відображається математичною моделлю, може бути явною, чи неявною, тобто може бути зображена або як просте алгебраїчне співвідношення, або як велика за розміром сумісна система диференціально-алгебраїчних рівнянь. Сучасні комп'ютери дозволяють у багатьох випадках відмовитися від макетування проєктованих виробів, замінивши його математичним моделюванням (обчислювальним експериментом), що дуже важливо, коли натурне макетування або дуже складне, або взагалі неможливе (наприклад: макетування прориву дамби або дії смерчу). Але при цьому має бути підвищена точність математичної моделі об'єктів та систем. В результаті розмірність і складність математичної моделі істотно зростають, а їх розв'язання в аналітичному вигляді стає неможливим. Таким чином, одержується нова якість одного параметру (велика точність обчислювального експерименту і відмова від натурального макетування) за рахунок зменшення чи ускладнення іншого параметра (відмова від звичних для вищої математики аналітичних рішень). Для кожної математичної моделі формулюється математична задача. У загальному випадку, коли функціональні залежності для множини вхідних даних (значення незалежних та змінюваних змінних і вхідних збуджень), що виступають як множина аргументів, задано неявно, за допомогою математичної моделі необхідно визначити множину вихідних залежних змінних, що виступають як множина значень функції. При цьому відповідно до виду математичної моделі розрізняють такі базові типи математичних задач:

- розв'язання системи лінійних рівнянь;
- розв'язання нелінійних алгебраїчних рівнянь;
- апроксимація масиву даних або складної функції набором стандартних, більш простих функцій;
- чисельне інтегрування і диференціювання;
- розв'язання систем звичайних диференціальних рівнянь;
- розв'язання диференціальних рівнянь в частинних похідних;
- розв'язання інтегральних рівнянь.

Чисельні методи – це математичний інструментарій, за допомогою якого математична задача формулюється у вигляді, зручному для розв'язання на

комп'ютері. В цьому випадку кажуть про перетворення математичної задачі в обчислювальну.

Для розв'язання математичних задач в основному існує три групи методів:

1. *Аналітичні методи*, в яких розв'язок задачі подається у вигляді аналітичних виразів. Їх *перевагами* є: запис розв'язку у загальному вигляді; висока точність і малий об'єм комп'ютерної пам'яті для зберігання розв'язку. Основний *недолік* - неуніверсальність, бо тільки невелика частина математичних задач може бути розв'язана аналітично.

2. *Графічні методи*, в яких розв'язок задачі знаходиться візуально. Їх *перевагою* є наочність. *Недоліками* графічних методів є: велика трудомісткість; низька точність (залежить від точності побудови графіків); неуніверсальність (графіки можна побудувати тільки для невеликої розмірності та ін.).

3. *Чисельні методи*, що дозволяють звести розв'язування задачі до виконання скінченного числа арифметичних і логічних дій з числами. При цьому розв'язок визначається як набір чисел, які надалі можуть бути інтерпретовані різним способом (наприклад, подані у вигляді таблиць, графіків, анімації тощо). Їх *перевагами* є: абсолютна універсальність, бо теоретично можуть бути застосовані для розв'язання будь-яких задач; добре пристосовані для реалізації на комп'ютері. *Недоліком* є велика трудомісткість у ході ручного рахунку, що, зазвичай, не є проблемою, оскільки вони призначені для використання на комп'ютері.

Таким чином, чисельні методи є основним апаратом розв'язання математичних задач, а їх значущість тільки збільшуватиметься у міру вдосконалення комп'ютерної техніки.

Запитання для самоконтролю.

1. Назвіть спільні властивості чисельних методів.
2. Назвіть чим чисельні методи характеризуються та розрізняються.
3. Що є об'єктом та предметом чисельних методів?
4. Назвіть групи на які поділяються чисельні методи для розв'язання математичних задач.

ЛЕКЦІЯ 2. Наближенні обчислювання та похибки. Абсолютна та відносна похибки

Розв'язок прикладних і математичних задач, як правило, пов'язане з наближеними значеннями величин та обчисленнями.

Математична модель задачі - це наближене подання реального об'єкта. Вихідні дані, одержувані з експерименту, можливо в основному визначити лише приблизно. Навіть точні числа, такі як π , e , $\sqrt{3}$, $\frac{6}{7}$ і т.п., при обчисленнях заміняють десятковими дробами, залишаючи певне число знаків після коми.

Обчислювальні методи в основному також є наближеними. Навіть при використанні найпростішої формули результат, як правило, одержують наближеним.

Тому, перш ніж розпочати вивчення обчислювальних методів, варто ознайомитися із загальними правилами дій над наближеними числами й оцінки похибок.

Абсолютна та відносна похибки

Абсолютна похибка наближеного числа.

Якщо a_0 деяке число, а a - число, прийняте за наближене значення числа a_0 , то помилкою наближеного числа a називають різницю $a_0 - a$. Звичайно точне число a_0 не відомо, тому помилку наближеного числа a визначити не можна. Однак майже завжди можливо вказати число, що оцінює цю помилку.

Число $\Delta(a) > 0$, що задовольняє нерівності

$$|a_0 - a| < \Delta(a) \quad (1)$$

називається абсолютною похибкою (граничною абсолютною похибкою) наближеного числа a . Таке визначення абсолютної похибки не є однозначним. Так, якщо $a = \pi$, а за наближене значення взяти $a = 3.14$, то з огляду на те, що $3.140 < \pi < 3.142$, можна записати: $|\pi - a| < 0.002$, $|\pi - a| < 0.01$, $|\pi - a| < 0.1$. Кожне із чисел 0.002 , 0.01 , 0.1 буде абсолютною похибкою числа a . Але чим ближче між собою числа $|a_0 - a|$ і $\Delta(a)$, тим точніше абсолютна похибка оцінює фактичну помилку. Як абсолютну похибку наближеного числа a беруть по можливості найменше із чисел, що задовольняють нерівності (1).

Нерівність (1) рівносильна нерівностям

$$a - \Delta(a) \leq a_0 \leq a + \Delta(a)$$

які умовно записують у такий спосіб: $a_0 = a \pm \Delta(a)$, тобто a_0 приблизно дорівнює a з абсолютною похибкою $\Delta(a)$.

Так, наприклад, у попередньому прикладі можна записати $\pi = 3.14 \pm 0.002$

Абсолютну похибкою іноді також називають оцінкою точності наближеного числа.

Відносна похибка наближеного числа.

Крім величини абсолютної похибки, необхідно ще знати її відношення до вимірюваної величині, в основному виражене у відсотках.

Відносною похибкою $\delta(a)$ наближеного числа a називається відношення абсолютної похибки $\Delta(a)$ до модуля цього числа:

$$\delta(a) = \frac{\Delta(a)}{|a|} \quad (2)$$

або у відсотках

$$\delta(a) = \frac{\Delta(a)}{|a|} \cdot 100\%$$

Так, відносна похибка числа 3.14 прийнятого за наближене значення числа π , при $\Delta(3.14) = 0.002$ дорівнює $\delta(3.14) = \frac{0.002}{3.14} = 0.00064$ або 0.064%.

У технічних розрахунках точність вимірів звичайно характеризують відносною похибкою. Результат вважають хорошим, якщо відносна похибка не перевищує 0.1%. З формули (2) слідує, що $\Delta(a) = |a| \cdot \delta(a)$

Правило округлення чисел

Округлення числа полягає у відкиданні в ньому всіх цифр, що розташовані за деяким розрядом. При цьому якщо округлене число ціле, то відкинуті цифри цілої частини заміняють нулями. Округлення звичайно роблять за наступним правилом: Якщо перша відкинута цифра, менше п'яти, то попередня цифра не міняється. Якщо перша відкинута цифра, більше п'яти, то попередня цифра збільшується на одиницю. Якщо перша відкинута цифра, дорівнює п'яти, то додатне кожне із зазначених правил, але найчастіше округляють так, щоб остання цифра, була парною. Якщо вона непарна, то до неї додають одиницю, якщо ж парна, то залишають без змін (правило парної цифри).

Приклад 1. Округлити число $\pi = 3.14159\dots$ до а) одного; б) трьох; в) чотирьох десяткових знаків.

Розв'язок:

- а) $3.14159 \approx 3.1$ (округлення до 0.1);
- б) $3.14159 \approx 3.142$ (округлення до 0.001);
- в) $3.14159 \approx 3.1416$ (округлення до 0.0001).

При округленні цілих чисел замість відкинутих цифр записують не нулі, а число 10 у відповідному степені.

Приклад 2. Округлити число 246250 до а) сотень тисяч; б) десятків тисяч; в) сотень.

Розв'язок:

- а) $246250 \approx 2 \cdot 10^5$ (округлення до 10^5);
- б) $246250 \approx 25 \cdot 10^4$ (округлення до 10^4);
- в) $246250 \approx 2462 \cdot 10^2$ (округлення до 10^2).

Якщо при округленні числа останні зберігають нулі, то їх варто записувати. Так, число 1.2997 округлене до 0.001, приймає вид 1.300.

При округленні числа виникає додаткова похибка - похибка округлення $\Delta_0(a)$, що не перевершує половини одиниці розряду останньої цифри, що зберігається (або п'яти одиниць першого відкинутого розряду).

Так, у прикладі 1 маємо наступні похибки округлення:

$$\Delta_0(3.1) = 0.04159 \leq 0.05 \quad \Delta_0(3.142) = 0.00041 \leq 0.0005$$

Абсолютна похибка округленого числа є сума його первісної абсолютної похибки й похибки округлення. Наприклад, нехай $\Delta_0(3.142) = 0.2$, при округленні 3.142 до 0.1 одержуємо 3.1 і $\Delta_0(3.1) = 0.042$, тоді $\Delta_0(3.1) = 0.2 + 0.042 = 0.242$

Звичайно абсолютну й відносну похибки округляють до однієї, рідше до двох цифр, відмінних від нуля. Округлення похибок роблять тільки у бік збільшення до тих же розрядів, що й наближене число. Так, при $\Delta_0(3.1) = 0.242$ розуміють $\Delta_0(3.1) = 0.3$

Запитання для самоконтролю.

1. Що називають абсолютною похибкою?
2. Що називають відносною похибкою?
3. Назвіть правила округлення чисел.

ЛЕКЦІЯ 3. Значущі, вірні й сумнівні цифри. Похибки при арифметичних діях з наближеними числами

Значущі, вірні й сумнівні цифри

Значущою цифрою наближеного числа називається будь-яка його цифра починаючи з першої ненульової (рахуючи з ліва на право). Наприклад, у числі 0.00030900 перші чотири нулі не є значущими цифрами. Всі інші цифри (включаючи й наступні три нулі) значущі

У цілому в округленому числі значущими звичайно вважають всі збережені цифри. Так, в округленому числі $120 \cdot 10^3$ цифри 1,2,0 - значущі.

Вірною цифрою (вірним знаком) наближеного числа a називається будь-яка його значуща цифра, для якої абсолютна похибка $\Delta(a)$ не перевершує половини розряду цієї цифри. Інші значущі цифри числа a називаються сумнівними.

Таким чином, якщо $\Delta(a) \leq 0.5 \cdot 10^{-n}$, то цифри числа a , починаючи з першої значущої й кінчаючи стоячою в n -м розряді після коми, - вірні, а розташовані далі—сумнівні. Наприклад, число 647,326 при

$\Delta(a) = 0.03 \leq 0.5 \cdot 10^{-1}$ має чотири вірні цифри 6,4,7,3 і дві сумнівні 2,6.

У математичних таблицях значень функцій приводяться тільки вірні цифри. Абсолютна похибка табличних значень, наприклад, у тризначних таблицях не перевершує $0.5 \cdot 10^{-3}$, у семизначних $0.5 \cdot 10^{-7}$. Точність наближеного числа залежить не від кількості значущих цифр, а від кількості вірних цифр. Остаточний наближений результат звичайно округляють до його вірних цифр, залишаючи одну сумнівну. При розрахунках з наближеними числами в проміжних результатах зберігають одну, дві, а іноді й три сумнівні цифри.

Приклад 3. Задані числа при зазначених абсолютних похибках округлити до вірних цифр. Визначити абсолютну похибку результату:

- 1) $a_1 = 2.6219$, $\Delta(a_1) = 0.024$; 2) $a_2 = 47.35$, $\Delta(a_2) = 1.3$;
3) $a_3 = 6.9971$, $\Delta(a_3) = 0.0009$.

Розв'язання:

- 1) Оскільки $0.024 < 0.05$, то число a_1 , варто округлити до 0.1.

Одержимо число $a_1 \approx 2.6$ з двома вірними цифрами. Визначимо абсолютну похибку результату

$$\Delta(2.6) = 0.024 + |2.6219 - 2.6| = 0.0459 \approx 0.05$$

2) Оскільки $1.3 < 0.5 \cdot 10^1$ то число a_2 варто округлити до десятків. Одержимо число $a_2 \approx 5 \cdot 10$ з однією вірною цифрою. Абсолютна похибка результату

$$\Delta(5 \cdot 10) = 1.3 + 2.65 = 3.95 \approx 4$$

3) Оскільки $0.0009 < 0.005$, то округлення числа робимо до 0.01. Одержимо число $a_3 \approx 7.00$ з трьома вірними цифрами. Абсолютна похибка результату

$$\Delta(7.00) = 0.0009 + |7.00 - 6.9971| = 0.0038 \approx 0.004$$

У числі 7.00 записані нулі свідчать про три його вірні знаки й цей запис відрізняється від 7 або 7.0.

Відносна похибка наближеного числа безпосередньо залежить від числа його вірних знаків. Є спеціальні таблиці, що визначають відносні похибки залежно від числа вірних знаків.

Похибки при арифметичних діях з наближеними числами

Похибки при арифметичних діях з наближеними числами виражаються через похибки первісних величин на підставі властивостей, які приведемо без доказу.

1) Абсолютна похибка алгебраїчної суми декількох чисел дорівнює сумі абсолютних похибок доданків:

$$\Delta(a_1 + \dots + a_n) = \Delta(a_1) + \dots + \Delta(a_n) \quad (4)$$

З (4) слідує, що якщо всі доданки $a_1, \dots, a_n$ (незалежно від їх знаків, мають ту саму абсолютну похибку $\Delta(a)$, то

$$\Delta(a_1 + \dots + a_n) = n\Delta(a) \quad (5)$$

Однак при великій кількості доданків ця формула дає завищені результати, оскільки відхилення доданків від їхніх точних значень можуть мати різні знаки й у сумі великої кількості доданків частково компенсуватися. У теорії ймовірностей доводиться, що при великому n можна прийняти

$$\Delta(a_1 + \dots + a_n) = \sqrt{3n}\Delta(a), \quad n > 10 \quad (6)$$

Приклад 4. У трикутнику дані сторони $a=17.3$ см, $b=23.6$ см, $c=14.2$ см, причому $\Delta(a) = \Delta(b) = \Delta(c) = 0.1$ см. Визначити його периметр p й $\Delta(p)$

Розв'язання: $p=17.3+23.6+14.2=55.1$ см. По формулі (5) маємо

$\Delta(55.1) = 3 \cdot 0.1 = 0.3$ см. У числі 55.1 остання цифра сумнівна. Результат можна записати у вигляді $p = 55.1 \pm 0.3$ см.

Ясно, що абсолютна похибка суми наближених чисел не менше найбільшої з абсолютних похибок доданків. Тому при обчисленні суми наближених чисел всі доданки варто округляти до кількості десяткових знаків числа з найбільшою абсолютною похибкою, залишаючи один сумнівний знак (а при великій кількості доданків – два). Отриманий результат округляється на один знак.

2) Відносна похибка суми декількох чисел визначається по формулі

$$\delta(a_1 + \dots + a_n) = \frac{\Delta(a_1) + \dots + \Delta(a_n)}{|a_1 + \dots + a_n|}$$

Якщо $a_1, \dots, a_n$ — числа одного знаку, то відносна похибка $\delta(a_1 + \dots + a_n)$ укладена між найменшою й найбільшою з відносних похибок доданків:

$$\min\{\delta(a_1), \dots, \delta(a_n)\} \leq \delta(a_1 + \dots + a_n) \leq \max\{\delta(a_1), \dots, \delta(a_n)\}$$

Так, у попередньому прикладі

$$\delta(a) = \frac{\Delta(a)}{|a|} \cdot 100 = \frac{0.1}{17.3} \cdot 100 = 0.58\%$$

$$\delta(b) = 0.42\%, \quad \delta(c) = 0.70\%$$

$$0.42\% \leq \delta(p) \leq 0.70\%$$

У дійсності $\delta(p) = \frac{0.3}{55.1} \cdot 100\% = 0.54\%.$

При вирахуванні двох чисел одного знака відносна похибка різниці

$$\delta(a_1 - a_2) = \frac{\Delta(a_1) + \Delta(a_2)}{|a_1 - a_2|}$$

може виявитися значно більше відносних похибок кожного з даних чисел. Це в основному буває, якщо $|a_1 - a_2|$ – мале число.

Приклад 5. Обчислити абсолютну й відносну похибки різниці чисел

$$a_1 = 9.78, \quad a_2 = 9.22, \quad \Delta(a_1) = \Delta(a_2) = 0.01.$$

Розв'язання: $a_1 - a_2 = 0.56, \quad \Delta(a_1 - a_2) = 0.01 + 0.01 = 0.02,$

$$\delta(a_1 - a_2) = \frac{0.02}{0.56} \cdot 100\% \approx 3.57\%$$

хоч $\delta(a_1) = \frac{0.01}{9.78} \cdot 100\% \approx 0.102\%, \quad \delta(a_2) = \frac{0.01}{9.22} \cdot 100\% \approx 0.108\%$

Звичайно вирахування близьких чисел намагаються уникати, замінюючи його по можливості іншими діями.

- 3) Відносна похибка добутку декількох чисел дорівнює сумі відносних похибок співмножників:

$$\delta(a_1 \cdot \dots \cdot a_n) = \delta(a_1) + \dots + \delta(a_n) \quad (7)$$

Абсолютна похибка добутку обчислюється по формулі (3)

$$\Delta(a_1 \cdot \dots \cdot a_n) = |a_1 \cdot \dots \cdot a_n| \cdot \delta(a_1 \cdot \dots \cdot a_n)$$

Зокрема, якщо в добутку са число c точне, то $\delta(c) = 0$ і

$$\delta(ca) = \delta(a), \quad \Delta(ca) = |ca| \cdot \delta(a) = |c| \Delta(a)$$

Звідси й з формули (4) одержуємо

$$\Delta(c_1 a_1 + \dots + c_n a_n) = |c_1| \Delta(a_1) + \dots + |c_n| \Delta(a_n) \quad (8)$$

- 4) Відносна похибка частки дорівнює сумі відносних похибок діленого й дільника:

$$\delta\left(\frac{a_1}{a_2}\right) = \delta(a_1) + \delta(a_2) \quad (9)$$

$$\text{Зокрема} \quad \delta\left(\frac{1}{a}\right) = \delta(a)$$

Абсолютну похибку частки визначають по формулі (3).

Якщо кількість чисел у добутку або відношенні велика, а відносна похибка кожного числа приблизно однакова (дорівнює $\delta(a)$) то аналогічно формулі (6) відносна похибка результату обчислюється по формулах

$$\delta(a_1 a_2 \dots a_n) = \sqrt{3n} \delta(a) \quad n > 10,$$

$$\delta\left(\frac{a_1 a_2 \dots a_n}{b_1 b_2 \dots b_m}\right) = \sqrt{3(n+m)} \delta(a) \quad (n+m) > 10$$

При наявності декількох співмножників, в одного із яких відносна похибка у багато разів більше, ніж в інших (він обчислений найменш точно), відносна похибка добутку буде визначатися саме по цій похибці. Тому число вірних знаків в інших співмножників треба вибирати по найменш точному числу, залишаючи один сумнівний. Аналогічно надходимо й при діленні.

*Похибка обчислення наближених значень функції
однієї змінної*

Нехай задана деяка диференціальна функція $y = f(x)$ і $\bar{x}$ - наближене значення аргументу. Наближеним значенням $\bar{y}$ функції y вважають те значення, що вона приймає при наближеному значенні аргументу, тобто $\bar{y} = f(\bar{x})$. Виникає питання про похибки цього наближення. Як відомо з курсу математичного аналізу, при досить малому приросту аргументу Δx приріст функції Δy приблизно дорівнює її диференціалу:

$$\Delta y = f(x) - f(\bar{x}) \approx dy = f'(\bar{x})\Delta x$$

Звідси, якщо відомо абсолютну похибку аргументу

$\Delta(\bar{x}) > |x - \bar{x}|$ абсолютну похибку функції можна визначити по формулі

$$\Delta(\bar{y}) = |f'(\bar{x})|\Delta(\bar{x}) \quad (10)$$

Відносна похибка функції, відповідно до формули (2)

$$\delta(\bar{y}) = \frac{\Delta(\bar{y})}{|\bar{y}|} = \left| \frac{f'(\bar{x})}{f(\bar{x})} \right| \Delta(\bar{x}) \quad (11)$$

Як приклади визначимо абсолютні й відносні похибки деяких основних елементарних функцій.

1) Логарифмічна функція $f(x) = \log_a x$. Абсолютна похибка логарифмічної функції, відповідно до формули (10) має вигляд

$$\Delta(\log_a \bar{x}) = \frac{1}{|\bar{x}|} |\log_a e| \Delta(\bar{x}) = |\log_a e| \frac{\Delta(\bar{x})}{|\bar{x}|} = |\log_a e| \delta(\bar{x}),$$

де $\delta(\bar{x}) = \frac{\Delta(\bar{x})}{|\bar{x}|}$

Зокрема, якщо $a=10$ то $lg_e \approx 0.5$ і $\Delta(lg \bar{x}) \approx 0.5\delta(\bar{x})$, якщо ж $a=e$, то $\Delta(\ln \bar{x}) = \delta(\bar{x})$ ($\ln e = 1$). Абсолютна похибка логарифмічної функції з основою e дорівнює відносній похибці аргументу. Відносну похибку функції знаходимо по формулі (1)

$$\delta(\log_a \bar{x}) = \left| \frac{\log_a e}{\log_a \bar{x}} \right| \delta(\bar{x})$$

2) Степенева функція $f(x) = x^\alpha$, де α -будь-яке дійсне число.

$$\Delta(\bar{x}^{\alpha-1}) = |\alpha \bar{x}^{\alpha-1}| \Delta(\bar{x}), \quad \delta(\bar{x}^\alpha) = \left| \frac{\alpha \bar{x}^{\alpha-1}}{\bar{x}^\alpha} \right| \Delta(\bar{x}) = |\alpha| \frac{\Delta(\bar{x})}{|\bar{x}|} = |\alpha| \delta(\bar{x})$$

Відносна похибка степеневі функції пропорційна відносній похибці аргументу. Зокрема

$$\delta(\bar{x}^2) = 2\delta(\bar{x}), \quad \delta(\bar{x}^3) = 3\delta(\bar{x}), \quad \delta(\sqrt{\bar{x}}) = \frac{1}{2}\delta(\bar{x})$$

3) Показникова функція $f(x) = a^x$

$$\Delta(a^{\bar{x}}) = a^{\bar{x}} |\ln a| \Delta(\bar{x}), \quad \delta(a^{\bar{x}}) = |\ln a| \Delta(\bar{x})$$

4) Тригонометричні функції $f(x) = \sin x$, $f(x) = \cos x$

$$\Delta(\sin \bar{x}) = |\cos \bar{x}| \Delta(\bar{x}) \leq \Delta(\bar{x}), \quad \Delta(\cos \bar{x}) = |\sin \bar{x}| \Delta(\bar{x}) \leq \Delta(\bar{x})$$

Абсолютні похибки функцій синус і косинус не перевершують абсолютної похибки аргументу

$$\delta(\sin \bar{x}) = |\operatorname{ctg} \bar{x}| \Delta(\bar{x}), \quad \delta(\cos \bar{x}) = |\operatorname{tg} \bar{x}| \Delta(\bar{x})$$

Похибки, що виникають при рішенні математичних задач чисельними методами, грубо можна розділити на дві групи. Першу групу становлять похибки, що не залежать від конкретного змісту задачі, а також похибки, викликані діями над наближеними числами. До другої групи відносяться похибки, що виникають за рахунок того, що математична задача як правило, замінюється спрощеною, близькою за результатом наближеною

задачі. Ці похибки (похибки методу) визначаються залежно від характеру задачі.

Запитання для самоконтролю.

1. Охарактеризуйте поняття значущі, вірні й сумнівні цифри.
2. Назвіть та надайте формули обчислювання похибок при арифметичних діях з наближеними числами.
3. Укажіть по яким формулам обчислюються похибки до основних арифметичних функцій.

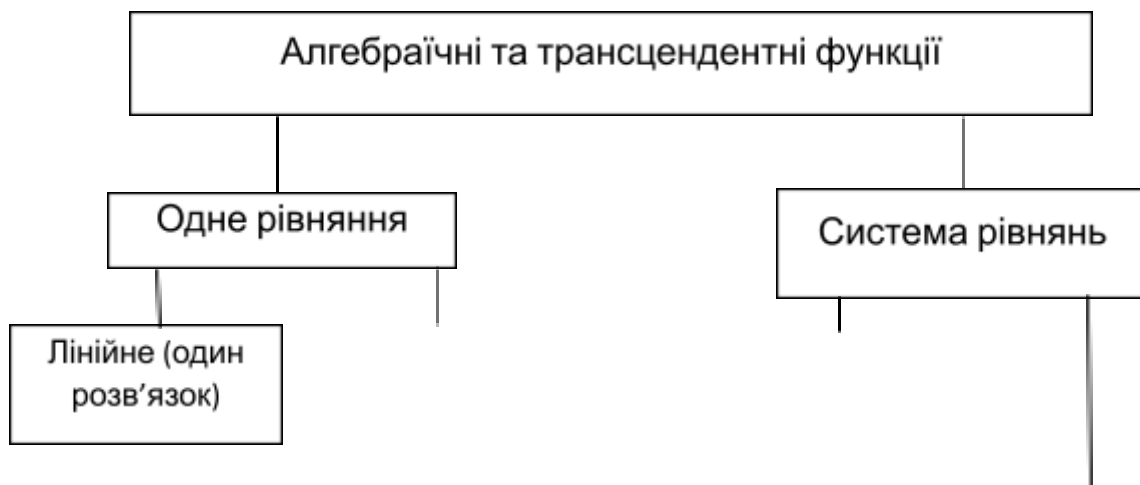
ЛЕКЦІЯ 4. Розв'язання нелінійних рівнянь. Алгебраїчні та трансцендентні функції. Рівняння з одним невідомим.

Розв'язання нелінійних рівнянь і систем є не тільки важливою самостійною задачею, але і частиною інших задач обчислювальної математики. З ними пов'язана побудова різноманітних моделей пристроїв і систем автоматики й інформаційно-вимірювальної техніки.

Задачі, що зводяться до розв'язання алгебраїчних і трансцендентних рівнянь, можна класифікувати по числу рівнянь і в залежності від передбачуваного характеру і числа рішень (рис.1)

Трансцендентними називаються нелінійні рівняння, що містять тригонометричні або інші спеціальні функції, наприклад, логарифмічну або експонентну. Рівняння, що не містять спеціальних функцій, а тільки степені аргументу з відповідними коефіцієнтами, є *нелінійними алгебраїчними*.

Застосування прямих методів розв'язання таких рівнянь можливо лише для алгебраїчних рівнянь, причому практично доцільно при порядку не більш третього. Тому на перший план виходять ітераційні методи, особливо при наявності ефективних алгоритмів.



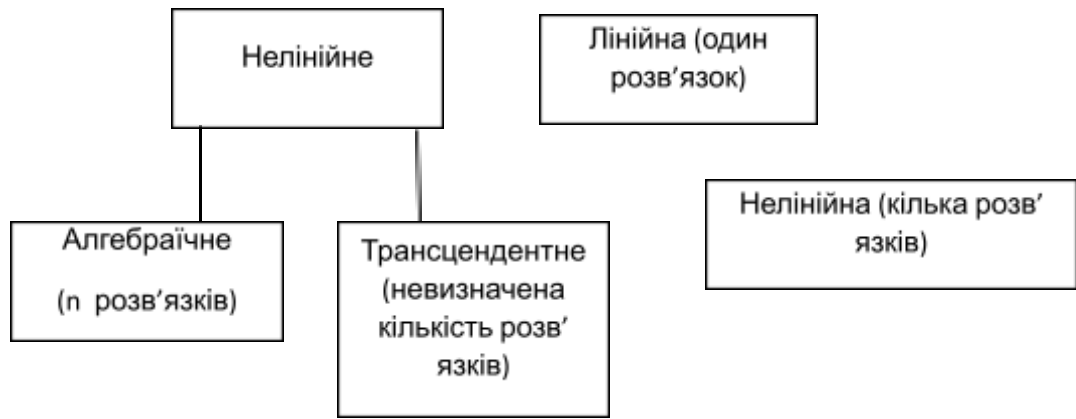


Рис. 1 - Класифікація задач, що зводяться до розв'язання алгебраїчних та трансцендентних рівнянь.

Стосовно алгебраїчних рівнянь можна вказати ряд властивостей, використання яких дозволить прискорити процес розв'язання задачі.

Загальний вигляд алгебраїчного рівняння:

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 = 0 \quad (1)$$

Властивості алгебраїчних рівнянь:

1. Основна теорема алгебри. Алгебраїчне рівняння порядку n має n коренів, що можуть бути як дійсними, так і комплексними. Кожен корінь рахується стільки разів, яка його кратність.
2. Якщо всі коефіцієнти a_i рівняння (1) дійсні, то всі комплексні корені утворюють комплексно сполучені пари.
3. Теорема Декарта. Кількість позитивних дійсних коренів дорівнює або менше кількості змін знаків у послідовності коефіцієнтів a_i , (це твердження справедливе відносно кількості від'ємних дійсних коренів при заміні в (1) x на $(-x)$).
4. Теорема Лагранжа. Верхня границя додатних дійсних коренів визначається як

$$R = 1 + \sqrt[k]{\frac{B}{a_0}} ; a_0 > 0$$

де a_k - перший від'ємний коефіцієнт; B - найбільша абсолютна величина від'ємного коефіцієнта.

5. Теорема Гюа. Якщо рівняння (1) має дійсні корені і дійсні коефіцієнти, то

$$a_k^2 > a_{k-1} a_{k+1}$$

У суворому математичному змісті, знайшовши один який-небудь корінь x_1 можна розділити рівняння на $(x - x_1)$, тим самим понизивши його степінь. Однак ця процедура може призвести до нагромадження помилок у коефіцієнтах одержуваних рівнянь, а отже, і в коренях. Відомі дві теореми, одна з яких дозволяє відокремити корені, тобто установити можливо тісні проміжки $[\alpha, \beta]$, у яких міститься один і тільки один корінь рівняння, а друга - оцінити степінь наближення.

Теорема 1. Якщо безперервна функція $f(x)$ приймає значення різних знаків на кінцях відрізка $[\alpha, \beta]$, тобто $f(\alpha)f(\beta) < 0$, то усередині цього відрізка міститься щонайменше один корінь рівняння $f(x) = 0$, тобто знайдеться хоча б одне число $\xi \in (\alpha, \beta)$ таке що $f(\xi) = 0$.

Теорема 2. Нехай ξ – точний, а $\bar{x}$ - наближений корені рівняння $f(x) = 0$, що знаходяться на тому самому відрізку $[\alpha, \beta]$, причому

$f'(\bar{x}) \geq m, \alpha \leq \bar{x} \leq \beta$. Тоді

$$|\bar{x} - \xi| \leq \frac{|f(\bar{x})|}{m}$$

Рівняння з одним невідомим

Знаходження коренів рівняння - це одна з найдавніших математичних проблем, вона часто зустрічається в найрізноманітніших областях науки і техніки.

У практичних розрахунках часто треба розв'язувати рівняння типу

$$x = f(x)$$

де $f(x)$ - деяка функція, що вважається неперервною.

Коренем або розв'язком цього рівняння називають таке число x при якому дане рівняння перетворюється в тотожність. Точне розв'язання поставленої проблеми далеко не завжди можливо, тому доводиться використовувати наближені методи розв'язання. Однак точне розв'язання рівнянь у більшості випадків не є необхідним, тому задачу відшукування коренів рівняння можна вважати практично вирішеною, якщо можливо визначити корені з потрібним ступенем точності. Слід зазначити, що більшість наближених методів розв'язання рівнянь розраховано на знаходження тільки одного кореня, хоча в дійсності таких коренів може бути кілька. Тому процедура розв'язання рівняння розкладається на два етапи:

1) Визначення інтервалу, у якому міститься один корінь.

2) Уточнення значення кореня до деякого заданого степеня точності.

Відрізком (інтервалом) $[a, b]$ відділення (ізоляції) кореня рівняння називають відрізок осі абсцис, на якому міститься тільки, один корінь даного рівняння. Іншими словами, якщо зазначена нижня границя a і верхня границя b для кореня x і відомо, що при $a < x < b$ корінь єдиний, то корінь x відділений.

Визначення інтервалу відділення кореня

Якщо приблизно відомо відрізок $[\alpha, \beta]$, на якому необхідно відокремити корінь, то розбиваємо його на n рівних по довжині частин точками x_i , причому $\alpha = x_0 < x_1 < x_2 < \dots < x_n = \beta$. Обчислюємо значення функції $f(x)$ у цих точках, тобто $y_i = f(x_i)$, $i = 0, 1, 2, \dots, n$. Для кожного i -го інтервалу $[x_i, x_{i-1}]$ аналізуємо знаки y_{i-1} і y_i . Можливі два випадки:

1) Значення y_i та y_{i-1} мають різні знаки. Тоді в інтервалі $[x_i, x_{i-1}]$ є хоча б один корінь. Для уточнення числа коренів досліджуємо знак першої похідної $f'(x)$ на знайденому інтервалі. Якщо в інтервалі $[x_i, x_{i-1}]$ похідна знакопостійна, то в ньому міститься лише один корінь; якщо похідна змінює знак, потрібні додаткові дослідження.

2) Значення y_i та y_{i-1} мають однакові знаки. У цьому випадку на інтервалі $[x_i, x_{i-1}]$ або немає коренів, або їх деяке (швидше всього парне число). Якщо похідна $f'(x)$ на інтервалі не змінює знак, то коренів у ньому немає. Знайдений інтервал $[x_i, x_{i-1}]$, у якому похідна не змінює знак і функція на кінцях інтервалу має різні знаки, приймається як інтервал відділення кореня.

У принципі, інтервал відділення кореня можна також визначити, провівши аналіз поведінки функції $f(x)$ за допомогою обчислення першої похідної і знаходженню точок перегину функції (що дозволяє побудувати приблизний графік функції)

Графічний метод розв'язання рівнянь

У найпростішому варіанті цей метод полягає в тому, що будується графік функції $y = f(x)$ і як корінь приймаються абсциси точок перетину графіка функції з віссю OX . Такий спосіб є досить грубим у силу неточностей побудови графіків функцій, та й графіки можна побудувати лише для вузького класу математичних функцій. Іноді вихідне рівняння доцільно представити у

вигляді $\varphi_1(x) = \varphi_2(x)$, тоді розв'язанням рівняння будуть точки перетину графіків функції $y = \varphi_1(x)$ і $y = \varphi_2(x)$.

На практиці графічний метод розв'язання рівнянь в основному застосовується для наближеної локалізації інтервалу відділення.

Запитання для самоконтролю.

1. Як можна класифікувати задачі, що зводяться до розв'язання алгебраїчних і трансцендентних рівнянь?
2. Які рівняння називають трансцендентними?
3. Назвіть властивості алгебраїчних рівнянь.
4. В чому полягає процедура розв'язування рівняння з одним невідомим?

ЛЕКЦІЯ 5. Розв'язання нелінійних рівнянь .Метод поділу навпіл (дихотомії), комбінований метод хорд і дотичних та ітераційний метод (послідовних наближень)

Метод поділу навпіл (дихотомії)

Задача ставиться у вигляді: потрібно знайти корінь рівняння $f(x) = 0$ на інтервалі відділення $[a, b]$ із заданою точністю ε (при цьому вважається, що функція $f(x)$ на цьому відрізку неперервна). Метод поділу навпіл заснований на наступній *теоремі з курсу математичного аналізу*: якщо функція неперервна на відрізку $[a, b]$ і приймає на кінцях цього відрізка значення різних знаків, то всередині відрізка $[a, b]$ існує принаймні один корінь рівняння $f(x) = 0$. Такі міркування дають можливість побудови послідовності відрізків, що входять друг у друга, причому кожний з них буде містити корінь рівняння.

Візьмемо середину відрізка $[a, b]$, позначимо її як c , $c = (a + b)/2$, і обчислимо в ній значення функції $f(c)$. Якщо $f(c) = 0$, то c - шуканий корінь рівняння. Якщо $f(c) \neq 0$, тоді порівнюємо знаки функції в точках a і c . Таке порівняння зручно робити шляхом множення функцій $f(a)$ і $f(c)$, якщо цей добуток більше 0, то знаки функції однакові, у протилежному випадку - різні. Якщо $f(a) \cdot f(c) < 0$, то відкидається відрізок $[c, b]$, тому що на ньому коренів не має, у випадку $f(a) \cdot f(c) > 0$ відкидається відрізок $[a, c]$. Відкидання відрізка відбувається шляхом відповідного присвоєння $b=c$ чи $a=c$. У результаті такої операції ми знову одержали відрізок $[a, b]$, що містить корінь, але його розмір менше вихідного. Подібну процедуру повторюють доти, поки не стане вірним умова $|b - a| < 2\varepsilon$. Після цього, як наближене

розв'язання рівняння знайдено з точністю ε , то приймається середина отриманого відрізка $x = (a + b)/2$.

Переваги методу поділу навпіл:

1) Універсальність, тому що процедура пошуку передбачає перебування одного кореня для будь-яких неперервних функцій $f(x)$, у тому числі не диференційовних.

2) Метод стійкий до помилок округлення, що виникають при обчисленнях. Точність відповіді гарантована.

Недоліки методу поділу навпіл:

1) Мала швидкість збіжності, за одну ітерацію точність збільшується приблизно вдвічі.

2) Для застосування його необхідно заздалегідь знайти інтервал відділення (якщо ж він попередньо не знайдений і в пропонуваному відрізку міститься декілька коренів, то заздалегідь неможливо сказати, до якого з них зійдеться процес)

На початку алгоритму можна також ввести перевірку на знакозмінність функції $f(x)$ на кінцях вихідного відрізка $[a, b]$ (якщо він заданий невірно, то алгоритм не приведе до розв'язання).

Розглянемо застосування методу поділу навпіл на прикладі розв'язання рівняння $x^3 + 3x^2 - 1 = 0$ на інтервалі відділення $[0, 1]$ з точністю $\varepsilon = 0,001$.

Для початку перевіримо знаки на кінці інтервала: $f(x) = x^3 + 3x^2 - 1$, $f(0) = -1$, $f(1) = 1$. Знаки змінюються, продовжуємо розв'язання.

Середина вихідного відрізка $c = 0,5$; $f(0,5) = -0,125$; $f(0) \cdot f(0,5) > 0$, тому з відрізків $[0; 0,5]$ і $[0,5; 1]$ вибираємо другий, переносимо точку a в точку c шляхом присвоєння $a = 0,5$, поділяємо новий відрізок $[0,5; 1]$ навпіл і так далі.

Обчислення наводяться в таблиці

Відрізок $[a, b]$	Середин а відрізка с	Знак $f(a) \cdot f(c)$	Перевірка $ b - a < 2\varepsilon = 0,002$
$[0, 1]$	0,5	$f(0) \cdot f(0,5) > 0$	$ 1 - 0 > 2\varepsilon = 0,002$
$[0,5; 1]$	0,75	$f(0,5) \cdot f(0,75) < 0$	$ 1 - 0,5 > 2\varepsilon = 0,002$
$[0,5; 0,75]$	0,625	$f(0,5) \cdot f(0,625) < 0$	$ 0,75 - 0,5 > 0,002$

$[0,5; 0,625]$	0,5625	$f(0,5) \cdot f(0,5625) < 0$	$ 0,625 - 0,5 > 0,002$
$[0,5; 0,5625]$	0,53125	$f(0,5) \cdot f(0,53125) < 0$	$ 0,53125 - 0,5 > 0,002$
$[0,5; 0,53125]$	0,51562	$f(0,5) \cdot f(0,51562) > 0$	$ 0,51562 - 0,5 > 0,002$
$[0,51562; 0,53125]$	0,52342	+	$ b - a > 0,002$
$[0,52342; 0,53125]$	0,52733	+	$ b - a > 0,002$
$[0,52733; 0,53125]$	0,52929	+	$ b - a > 0,002$
$[0,52929; 0,53125]$	0,53027		

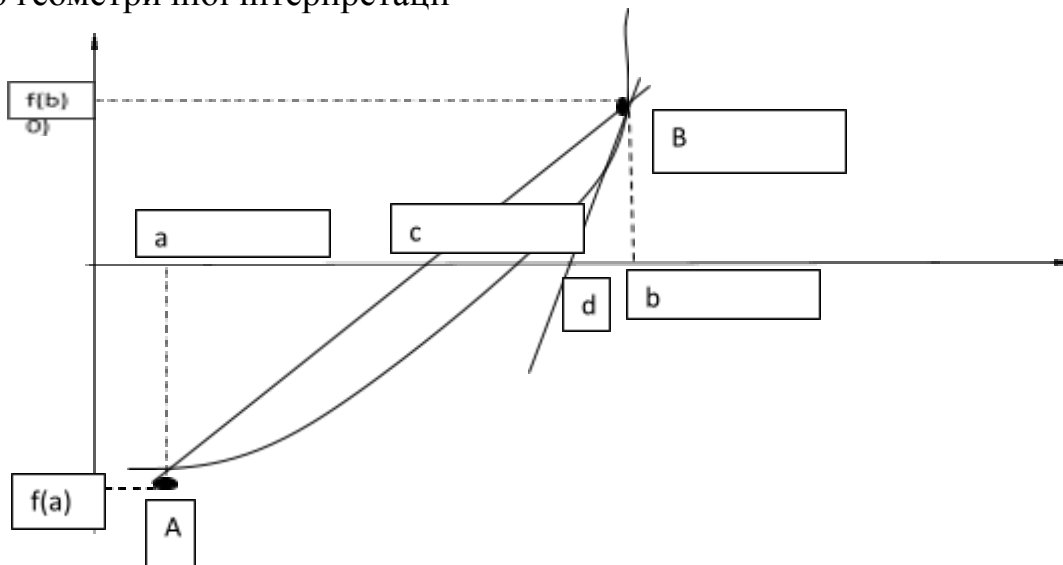
Довжина отриманого відрізка $|b - a| = |0,53125 - 0,52929| = 0,0019 < 0,002$,

тобто шукана точність досягнута. Розв'язання рівняння $x \approx 0,5302 \pm 0,001$

Комбінований метод хорд і дотичних

Застосування цього методу припускає, що на відрізку $[a, b]$ функція $f(x)$ відділена (тобто на кінцях відрізка має різні знаки), а $f'(x)$ і $f''(x)$ неперервні і знакопостійні у середині цього відрізка.

Розглянемо комбінований метод хорд і дотичних розв'язання рівнянь, по якому одночасно застосовується і метод хорд і метод дотичних, що дозволяє організувати більш швидкий обчислювальний процес. Звернемося до геометричної інтерпретації



У точці В проведемо дотичну до кривий $y = f(x)$ до перетину з віссю ОХ. Абсцису точки перетину позначимо через d. Проведемо хорду АВ і абсцису точки перетину позначимо с. Отриманий відрізок $[c, d]$ містить

шуканий корінь рівняння і його довжина значно менше вихідного відрізка. Продовжуючи проводити аналогічні побудови, можна звузити відрізок до необхідної точності ε .

Слід зазначити, що дотична повинна проводитися з тієї точки на границі інтервалу відділення, у якій знак функції збігається зі знаком другої похідної.

Алгоритм методу складається в наступному (будемо виходити з того, що дотичну потрібно проводити з точки b):

1. Знаходимо точки c і d , для чого робимо обчислення по формулах

$$c = a - f(a) \frac{b-a}{f(b)-f(a)} \quad (2)$$

$$d = b - \frac{f(b)}{f'(b)} \quad (3)$$

2. Порівнюємо довжину отриманого відрізка з точністю ε . Якщо $|d - c| \leq 2\varepsilon$, то кінець алгоритму, а як корінь береться значення $x = (b + a)/2$.

Якщо ж ця умова не виконується, робимо присвоєння $a=c$, $b=d$ і повертаємося до пункту 1. Відзначимо, що питання вибору точки проведення дотичної зважується один раз. Якщо дотичну потрібно проводити не з точки b , а з точки a , то в алгоритмі ці точки у формулах треба поміняти місцями.

Для ілюстрації уточнимо значення кореня попереднього рівняння

$f(x) = x^3 + 3x^2 - 1$ на відрізку $[0, 5; 0, 55]$ з точністю $\varepsilon = 0,0001$.

Знаходимо похідні $f'(x) = 3x^2 + 6x$ і $f''(x) = 6x + 6$. Знаходимо

$f'(0,55) = 0,738 > 0$ і $f''(0,55) = 4,2075 > 0$. Звідси випливає, що дотичну необхідно проводити з точки $b=0,55$. Використовуючи формули (2) і (3) знаходимо

$$d = 0,55 - \frac{0,07387}{3 \cdot 0,55^2 + 6 \cdot 0,55} = 0,53244$$

$$c = 0,5 - (-0,125) \cdot \frac{0,5-0,55}{-0,125-0,7387} = 0,531427$$

Отримано новий інтервал, що містить корінь $[0,531427; 0,53244]$.

Наступна ітерація:

$$d = 0,53244 - \frac{0,001419}{4,045117} = 0,5320988$$

$$c = 0,531427 - (-0,0026742) \cdot \frac{(-0,00997)}{-0,001433-0,0026742} = 0,5320761$$

дає значення інтервалу $[0, 5320761; 0, 5320988]$ довжина якого $0, 0000227 < 2\varepsilon = 0, 002$, звідси, як наближене розв'язання рівняння, приймаємо значення $0,5320 \pm 0,0001$.

Ітераційний метод (послідовних наближень)

Нехай дано рівняння (1) Щоб вирішувати його ітераційним методом, необхідно записати його в наступному вигляді:

$$x = \varphi(x) \quad (4)$$

Таке перетворення можна зробити декількома способами. Наприклад, до обох частин рівняння (1) можна додати x .

Для розв'язання рівняння (4) ітераційним методом виберемо початкове наближене x_0 і обчислимо подальше наближення по формулах

$$x_1 = \varphi(x_0)$$

$$x_2 = \varphi(x_1)$$

.....

$$x_{k+1} = \varphi(x_k), \quad (k = 0, 1, 2, \dots)$$

Остання формула називається *рекурентною* і визначає закон побудови послідовності. Послідовність $\{x_n\}$, члени якої обчислюються по

приведеній формулі, називається ітераційною. Має місце наступний факт: якщо ітераційна послідовність збігається і $x_n = \xi$, то ξ є коренем рівняння

(1). Послідовність $\{x_n\}$ не завжди збігається, тому приведемо достатню

умову збіжності ітераційної послідовності: якщо $|\varphi'(x)| < 1$,

то ітераційний процес збігається, якщо ж $|\varphi'(x)| > 1$, то процес розбігається. Для застосування ітераційного методу доцільно будь-яким грубим способом (наприклад графічним) знайти наближене значення кореня, оцінити значення похідної в цій точці і порівняти її з 1. У випадку

виконання умови $|\varphi'(x)| < 1$ можливе застосування ітераційного алгоритму для уточнення кореня, причому як перше наближення береться знайдена точка. Як наближене значення кореня при реалізації алгоритму доцільно вибирати такий член ітераційної послідовності $\{x_n\}$, для якого виконується умова $|x_n - x_{n-1}| < \varepsilon$.

Іноді як критерій зупинки обчислювального процесу вибирають перевищення числа ітерацій (повторень числа кроків алгоритму) над заздалегідь заданою величиною. Такий критерій обумовлений тією обставиною, що якщо $x_n = \xi$, то для даного ξ знайдеться таке n , що

$$|x_n - \xi| < \varepsilon$$

До недоліків ітераційних методів варто віднести:

1. Обмежена область застосування (не для всіх рівнянь).
2. Необхідність попередньої перевірки послідовності $\{x_n\}$ на збіжність.

Перевагами є:

1. Простота реалізації на комп'ютері.
2. Відсутність необхідності попереднього знаходження інтервалу відділення $[a, b]$.
3. Випадковий характер вибору вихідної точки x_0 .

Розглянемо застосування ітераційного методу для розв'язання рівняння

$x^3 + 10x - 10 = 0$ з точністю $\varepsilon = 0,0001$. Представимо рівняння у вигляді

$x = 1 - 0,1x^3$, тобто у вигляді (4), причому $\varphi(x) = 1 - 0,1x^3$. Знайдемо першу похідну функції: $\varphi'(x) = -0,3x^2$. Значення цієї похідної на відрізьку $[0, 1]$ не перевищує за абсолютною величиною одиницю. $\varphi'(0) = 0 < 1$ і $\varphi'(1) = -0,3 < 1$. Тоді покладемо початкове наближення $x_0 = 1$. Наступні обчислення виглядають у такий спосіб (точністю до четвертого десяткового знака):

$$x_1 = 1 - 0,1x_0^3 = 1 - 0,1 \cdot 1^3 = 0,9; \quad |x_1 - x_0| = 0,1 > \varepsilon = 0,0001$$

$$x_2 = 1 - 0,1x_1^3 = 1 - 0,1 \cdot 0,9^3 = 0,9271; \quad |x_2 - x_1| = 0,027 > \varepsilon = 0,0001$$

$$x_3 = 1 - 0,1x_2^3 = 1 - 0,1 \cdot 0,9271^3 = 0,9203; \quad |x_3 - x_2| = 0,007 > \varepsilon = 0,0001$$

$$x_4 = 1 - 0,1x_3^3 = 1 - 0,1 \cdot 0,9203^3 = 0,9221; \quad |x_4 - x_3| = 0,0018 > \varepsilon = 0,0001$$

$$x_5 = 1 - 0,1x_4^3 = 1 - 0,1 \cdot 0,9221^3 = 0,9216; \quad |x_5 - x_4| = 0,0005 > \varepsilon = 0,0001$$

$$x_6 = 1 - 0,1x_5^3 = 1 - 0,1 \cdot 0,9216^3 = 0,9217; \quad |x_6 - x_5| = 0,0001 = \varepsilon = 0,0001$$

$$x_7 = 1 - 0,1x_6^3 = 1 - 0,1 \cdot 0,9217^3 = 0,9217; \quad |x_7 - x_6| = 0 < \varepsilon = 0,0001$$

Так як $|x_7 - x_6| < \varepsilon$, то отримане значення кореня $x=0,9217$ можна вважати наближеним розв'язанням заданого рівняння.

Запитання для самоконтролю.

1. У чому заключаються переваги та недоліки методу поділу навпіл (дихотомії)?
2. Надайте геометричну інтерпретацію методу хорд і дотичних.
3. У чому заключаються переваги та недоліки ітераційних методів?

ЛЕКЦІЯ 6. Розв'язання систем рівнянь. Системи лінійних рівнянь. Метод виключення Гауса та метод Жордана -Гаусса.

Задачі лінійної алгебри (у тому числі і розв'язання систем лінійних алгебраїчних рівнянь) мають важливе прикладне значення при розв'язанні різних проблем науки і техніки, у першу чергу при розв'язанні різних економічних задач, а також при реалізації алгоритмів математичної фізики, обчислювальної математики, обробці результатів експериментальних досліджень тощо.

Система лінійних рівнянь (СЛР) щодо невідомих x_j виглядає таким чином:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = a_{10} \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = a_{20} \\ \dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = a_{n0} \end{cases} \quad (1)$$

де x_j - невідомі величини, a_{ij} - задані елементи матриці системи лінійних рівнянь. Дану систему можна записати й у вигляді:

$$\sum_{j=1}^n a_{ij}x_j = a_{i0}, \quad i = 1, 2, \dots, n$$

Також СЛР можливо записати у матричній формі:

$$Ax = A_0$$

де A - квадратна матриця розміром $n \times n$, x і A_0 – вектори стовбці довжиною n .

У загальному випадку кількість рівнянь може не збігатися з числом змінних (тоді деякі змінні використовують як параметр і відносно цього параметра знаходять значення інших змінних), але ми будемо розглядати тільки ті СЛР, у яких ця умова виконується. Кожне з j рівнянь СЛР визначає в j -мірному просторі гіперплощину, перетинання їх визначає розв'язання. Відомо, якщо визначник матриці A нульовий, тобто $\det A = 0$, то СЛР не має розв'язання або має безліч рішень. При $\det A \neq 0$ система має одне розв'язання або має безліч рішень.

Методи розв'язання лінійних систем поділяють на прямі й ітераційні. Прямі методи дають розв'язання за кінцеве число кроків, вони прості й універсальні (методи Гауса, Жордана-Гауса). Їх застосовують для розв'язання систем невеликого порядку $n \leq 200$. Ітераційні методи застосовуються для систем спеціального вигляду зі слабо заповненою матрицею великого порядку $n \approx 10^3 - 10^5$. За допомогою ітераційних методів розв'язання одержують як межу послідовних наближень, що обчислюються деяким однаковим процесом (метод простих ітерацій, метод Зейделя). При застосуванні ітераційних методів істотним є не тільки збіжність побудованих послідовних наближень, але і швидкість збіжності. У цьому розумінні кожен ітераційний метод не є універсальним: даючи швидку збіжність для одних матриць СЛР, він може повільно збігатись або навіть зовсім не збігатись для інших.

Метод виключення Гауса

Цей метод, заснований на послідовному виключенні невідомих, має безліч варіантів. Розходження між ними - у порядку виключення невідомих і в способі збереження проміжних результатів. У загальному випадку метод полягає в приведенні системи (1) до системи трикутного вигляду, у якій останнє рівняння містить тільки одну змінну (коефіцієнти при інших дорівнюють 0), а всі інші рівняння знизу нагору будуть містити на одну змінну більше, ніж попереднє рівняння.

Будь-яку сумісну СЛР можна привести до описаного вигляду за допомогою наступних елементарних перетворень:

- 1) множення рівнянь на постійний множник;
- 2) додавання двох рівнянь.

Для того, щоб привести СЛР до зазначеного вигляду, потрібно виконати наступні кроки:

а) вибрати провідне рівняння і провідну змінну. Поділити всі коефіцієнти у провідному рівнянні на коефіцієнт при провідній змінній;

б) виключити провідну змінну з рівнянь, розташованих нижче в системі шляхом множення провідного рівняння на придатний множник і додавання до інших рівнянь;

в) повторити крок а) вибираючи провідне рівняння з усіх рівнянь системи, що не були провідними на попередніх кроках.

Процедура закінчується тоді, коли вихідна система буде приведена до трикутного вигляду або коли буде встановлено несумісність системи. Усі розрахунки краще вести в матричному вигляді, переходячи від запису СЛР (1) до матриці в такий спосіб:

$$(2) \quad \begin{pmatrix} a_{11} & a_{12} & \dots & \dots & \dots & a_{n1} & a_{n2} & \dots & a_{1n} & a_{10} & \dots & \dots & a_{nn} & a_{n0} \end{pmatrix}$$

Тоді шукана форма буде мати вигляд (з верхньою діагональною матрицею, у якій всі елементи нижче головної діагоналі дорівнюють 0)

$$(3) \quad \begin{pmatrix} 1 & b_{12} & \dots & \dots & \dots & 0 & 0 & \dots & b_{1n} & b_{10} & \dots & \dots & 1 & b_{n0} \end{pmatrix}$$

Дія алгоритму знаходження трикутного вигляду матриці починається з першого рядка і першого стовпця: розділимо перший рядок матриці (2) на коефіцієнт a_{11} потім для кожного k-го рядка ($k=2,3,\dots,n$) робимо наступні операції: перший рядок множимо на $(-a_{k1})$ і складаємо з k-м рядком, при цьому елементи першого стовпця a_{21} по a_{n1} , приймають нульові значення. Надалі алгоритм повторюється для всіх рядків.

У загальному випадку для кожного m-го рядка, починаючи з першого, проводяться наступні дії:

а) усі коефіцієнти m-го рядка поділяються на елемент a_{mm} ;

б) для всіх рядків $k > m$ повторюємо операції: множимо m-й рядок на коефіцієнт $-a_{mk}$ і складаємо з k-м рядком.

Якщо коефіцієнт a_{mm} дорівнює 0, то рядок $k > m$ міняється місцями з обраним, щоб a_{mm} не дорівнював нулю. Дії повторюються до одержання матриці вигляду (3). Після цього значення останньої змінної $x_n = b_{n0}$. Значення інших змінних одержуємо за формулою:

$$x_k = b_{k0} - \sum_{j=k+1}^n b_{kj} x_j, \quad k = n-1, n-2, \dots, 1 \quad (4)$$

У процесі розв'язання методом Гауса можливі 3 випадки:

1. На останньому кроці одержуємо шукану трикутну матрицю, у цьому випадку СЛР має єдине розв'язання.
2. На деякому кроці k у системі виходить рядок s , у якому усі коефіцієнти при невідомих і вільний член рівні 0. У цьому випадку СЛР має нескінченну множину розв'язань.
3. На кроці k виходить рядок, у якому усі коефіцієнти при змінних дорівнюють нулю, а вільний член не дорівнює нулю. Вихідна система не має розв'язання (несумісна).

Таким чином, обчислення по методу Гауса можна розкласти на два етапи: побудова системи (3) прямий хід методу і обчислення змінних по формулах (4) зворотний хід методу.

Метод Жордана-Гауса

За методом Жордана-Гауса розрахунки виконуються аналогічно методу Гауса, але розв'язання системи знаходиться відразу без зворотного ходу. Для цього за допомогою уже наведених елементарних перетворень вихідна матриця вигляду (2) зводиться до матриці

$$\begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & 0 & 0 & \dots & 0 & b_{10} & \dots & \dots & 1 & b_{n0} \end{pmatrix} \quad (5)$$

у якій на головній діагоналі одиниці, а всі інші елементи, крім стовпця вільних членів, дорівнюють нулю. Метод вимагає більше обчислень для одержання матриці (5), ніж за методом Гауса, але вииграш у швидкодії виходить у результаті відсутності зворотного ходу алгоритму. Метод Жордана-Гауса застосовується для розв'язання СЛР, у матриці яких багато нулів.

У загальному випадку методи Гаусса і Жордана-Гаусса є точними, але в силу похибок, що виникають при обчисленнях, розв'язання СЛР, отримане з їх допомогою, є наближеним. Контроль отриманих результатів можна провести шляхом обчислення нев'язок (чим вони чисельно менше, тим більше точно розв'язана СЛР)

$$r_k = a_{k0} - \sum_{j=1}^n a_{kj} x_j, \quad k = 1, 2, \dots, n$$

Приклад: Розв'язати систему методом Гаусса з точністю 10^{-4}

$$\{6x_1 - x_2 - x_3 = 11,33 \quad -x_1 + 6x_2 - x_3 = 32 \quad -x_1 - x_2 + 6x_3 = 42\}$$

Розв'язок: Приведемо систему в матричний вигляд

$$\begin{pmatrix} 6 & -1 & -1 & -1 & 6 & -1 & -1 & -1 & 6 & 11,33 & 32 & 42 \end{pmatrix}$$

Приведемо отриману систему до трикутного виду за допомогою методу Гауса. Для зручності поміняємо містами першу та другу строки.

$$\begin{pmatrix} -1 & 6 & -1 & 6 & -1 & -1 & -1 & -1 & 6 & 32 & 11,33 & 42 \end{pmatrix} \rightarrow \begin{pmatrix} -1 & 6 & -1 & 0 & 35 & -7 & 0 & -7 & 32 & 10 & 203,33 & 42 \end{pmatrix}$$

Далі до другої строки додаємо першу строку помножену на 6, а з третьої строки витянемо першу. Таким чином отримаємо нулі у першому стовбці. Далі для зручності поміняємо містами другу та третю строки.

$$\begin{pmatrix} -1 & 6 & -1 & 0 & -7 & 7 & 0 & 35 & -7 & 32 & 10 & 203,33 \end{pmatrix} \rightarrow \begin{pmatrix} -1 & 6 & -1 & 0 & -7 & 7 & 0 & 0 & 28 & 32 & 10 & 203,33 \end{pmatrix}$$

Далі до третьої строки додаємо другу строку помножену на 5. Таким чином отримаємо нулі у другому стовбці. Получили трикутну матрицю. Закінчили прямий хід метода Гауса.

Розглянемо зворотній хід метода Гауса. Розглянемо отриману систему рівнянь

$$\begin{cases} 28x_3 = 153,33 \\ -7x_2 + 7x_3 = 10 \\ -x_1 + 6x_2 - x_3 = 32 \end{cases}$$

З першого рівняння отримаємо $x_3 = 9,05$. З другого рівняння, враховуючи отримане значення x_3 , $-7x_2 + 7 \cdot 9,05 = 10$ отримаємо

$x_2 = 7,62$. З останнього рівняння отримаємо $x_1 = 4,67$. Таким чином ми знайшли наближений розв'язок системи.

Знайдемо уточнення для отриманого розв'язку. Для цього знайдемо невязки нашої системи за формулою $\varepsilon = b - Ax_0$, а потім уточнення із розв'язку системи $A\delta = \varepsilon$.

$$\varepsilon_1 = b_1 - Ax_0 = 11,33 - (6 \cdot 4,67 - 7 \cdot 7,62 - 9,05) = -0,02$$

$$\varepsilon_2 = b_2 - Ax_0 = 32 - (-4,67 + 6 \cdot 7,62 - 9,05) = 0$$

$$\varepsilon_3 = b_3 - Ax_0 = 42 - (-4,67 - 7,62 + 6 \cdot 9,05) = -0,01$$

Побудуємо систему

$$\{6\delta_1 - \delta_2 - \delta_3 = -0,02 \quad -\delta_1 + 6\delta_2 - \delta_3 = 0 \quad -\delta_1 - \delta_2 + 6\delta_3 = -0,01$$

Розв'язавши її отримаємо

$$\delta_1 = -0,0025, \quad \delta_2 = -0,0039, \quad \delta_3 = -0,0011.$$

Уточнюємо розв'язок по формулі $x = x_0 + \delta$:

$$x_1 = 4,67 - 0,0025 = 4,6675$$

$$x_2 = 7,62 - 0,0039 = 7,6161$$

$$x_3 = 9,05 - 0,0011 = 9,0489$$

Запитання для самоконтролю

1. На які два класи поділяються методи розв'язання лінійних систем?
2. На чому заснований метод виключення Гауса?
3. Назвіть етапи обчислення по методу Гауса?
4. Чим відрізняються метод Гауса та метод Жордана-Гауса?

ЛЕКЦІЯ 7. Ітераційні методи розв'язання систем лінійних рівнянь. Метод простих ітерацій.

Для застосування ітераційних методів розв'язання СЛР необхідно від запису $Ax = x_0$ перейти до еквівалентного запису

$$x = Bx + b_0 \quad (6)$$

де B - квадратна матриця розміром $n \times n$, x і b_0 - вектори-стовпці довжиною n . Задамося початковим наближенням x^0 і на його основі побудуємо послідовність наближень до повного розв'язання за рекурентною формулою (верхній індекс відповідає номеру кроку алгоритму)

$$x^{k+1} = Bx^k + b_0 \quad (7)$$

У розширеному вигляді дана формула має вигляд:

$$x_i^{k+1} = \sum_{j=1}^n b_{ij} x_j^k + b_{0i}, \quad i = 1, 2, \dots, n$$

Обчислення продовжуються доти, поки не стане вірною умова:

$$|x_i^{k+1} - x_j^k| < \varepsilon, \quad i = 1, 2, \dots, n$$

де ε - задана похибка. Обчислення за представленою формою мають назву *методу простих ітерацій*. Для збіжності ітераційного процесу достатньо, щоб у будь-якому рядку матриці В сума відношень коефіцієнтів системи до діагональних коефіцієнтів, узятим з того ж рядка, була строго менше одиниці, тобто всі коефіцієнти системи повинні бути малі у порівнянні з діагональними. Метод простих ітерацій при виконанні даної умови збігається незалежно від вибору початкового вектора x^0 . Для представлення вихідної системи СЛР (1) вигляді, придатному для використання ітераційних методів, необхідно кожне рівняння розділити на діагональний елемент і розв'язати відносно невідомих:

$$\{x_1 = - \left(\frac{a_{12}}{a_{11}} x_2 - \left(\frac{a_{13}}{a_{11} x_3 - \dots - \left(\frac{a_{1n}}{a_{11} x_n + \frac{a_{10}}{a_{11}}} \right)} \right) \right) x_2 = - \left(\frac{a_{21}}{a_{22}} x_1 - \left(\frac{a_{23}}{a_{22}} \right) x_3 - \dots - \left(\frac{a_{2n}}{a_{22} x_n + \frac{a_{20}}{a_{22}}} \right) \dots \dots \dots \right.$$

Різновидом методу простих ітерацій є метод Зейделя, у якому для обчислення на k -й ітерації i -ї змінної використовуються значення вже обчислених перемінних $j < i$. Обчислювальна схема цього методу виглядає в такий спосіб (для k -ї ітерації):

Приклад. Розв'язати систему методом простих ітерацій прийнявши

$$\varepsilon = 0,005$$

$$\{3x_1 + x_2 - 5x_3 = 10 \quad 4x_1 + x_2 - 2x_3 = 8 \quad x_1 - 5x_2 + x_3 = -10$$

Розв'язок. Для того щоб виконувалися умови збіжності потрібно з першого рівняння системи виділити невідоме x_3 , з другого - x_1 , а з третього - x_2 . Для цього перше рівняння поділимо на 5, друге на 4, третє на 5. Отримаємо

$$\{x_3 = 0,6x_1 + 0,2x_2 - 2x_1 = -0,25x_2 + 0,5x_3 + 2 \quad x_2 = 0,2x_1 + 0,2x_3 + 2$$

Матриця С системи та умови збіжності мають вигляд:

$$C = (0,6 \ 0,2 \ 0 \ 0 \ - \ 0,25 \ 0,5 \ 0,2 \ 0 \ 0,2)$$

$$\max_i \sum_{j=1}^n |C_{ij}| = \max(0,6 + 0,2; 0,25 + 0,5; 0,2 + 0,2) = 0,8 < 1$$

За нульове наближення коренів системи приймаємо при $k=0$

$x_1^{(0)} = 2, x_2^{(0)} = 2, x_3^{(0)} = -2$. Підставимо отримані значення у систему

рівнянь та отримаємо нові значення $x_1^{(1)} = 0,5, x_2^{(1)} = 2, x_3^{(1)} = -0,4$

$$\{x_1^{(1)} = -0,25 \cdot 2 + 0,5 \cdot (-2) + 2 = 0,5 \quad x_2^{(1)} = 0,2 \cdot 2 + 0,2 \cdot (-2) + 2 = 2 \quad x_3^{(1)} =$$

Розрахуємо

$$\varepsilon = \max |x_i^{(k+1)} - x_i^{(k)}| = \max(|0,5 - 2|, |2 - 2|, |-0,4 - (-2)|) = 1,6 > \varepsilon$$

Продовжуємо наближення коренів, для цього отримані значення

$x_1^{(1)} = 0,5, x_2^{(1)} = 2, x_3^{(1)} = -0,4$ підставим до системи та отримаємо

$x_1^{(2)}, x_2^{(2)}, x_3^{(2)}$ і так далі доки не буде досягнуто задану точність.

Результати розв'язування наведено в таблиці.

K	x_1	x_2	x_3	$\varepsilon = \max x_i^{(k+1)} - x_i^{(k)} $
0	2	2	-2	
1	0,5	2	-0,4	$1,6 > \varepsilon$
2	1,3	2,02	-1,3	$0,9 > \varepsilon$
3	0,845	2	-0,816	$0,484 > \varepsilon$
4	1,092	2,0058	-1,093	$0,277 > \varepsilon$
5	0,9521	1,9998	-0,9436	$0,1494 > \varepsilon$
6	1,0282	2,0017	-1,0288	$0,0852 > \varepsilon$
7	0,9852	1,9999	-0,9827	$0,0461 > \varepsilon$
8	1,0087	2,0005	-1,0089	$0,0262 > \varepsilon$
9	0,9954	1,9999	-0,9947	$0,0142 > \varepsilon$
10	1,0027	2,0001	-1,0028	$0,0081 > \varepsilon$
11	0,9986	2,0000	-0,9984	$0,0044 < \varepsilon$

Отримали відповідь: $x_1^{(11)} = 0,9986$, $x_2^{(11)} = 2,0000$, $x_3^{(11)} = -0,9984$

Запитання для самоконтролю

1. Які преобразования системи необхідно провести для можливості застосування методу простих ітерацій?
2. Яка умова повинна виконуватися для збігання методу простих ітерацій?

ЛЕКЦІЯ 8. Системи нелінійних рівнянь. Метод Ньютона та модифікований метод Ньютона.

Системи нелінійних рівнянь

Система нелінійних рівнянь має вигляд:

$$(8) \quad \{f_1(x_1, \dots, x_n) = 0 \dots\dots\dots f_n(x_1, \dots, x_n) = 0$$

де $x_1, \dots, x_n$ - невідомі змінні, а система (8) називається звичайною системою порядку n якщо хоча б одна із функцій $f_i(x_1, \dots, x_n)$ нелінійна.

Розв'язання систем нелінійних рівнянь - одна із складних задач обчислювальної математики. Складність полягає у тому, щоб з'ясувати: чи має система розв'язок, і якщо – так, то скільки. Уточнення розв'язків у заданій області - більш проста задача. Нехай функції f_i визначені в областях Ω , $i = 1, \dots, n$. Тоді область $\Omega = \bigcap_{i=1}^n \Omega_i$ і буде тією областю, де можна знайти розв'язок. Найбільш відомими методами уточнення розв'язків є метод простих ітерацій та метод Ньютона

Метод Ньютона

Нехай A_k деяка послідовність невинроджених n -матриць. Тоді послідовність задач

$$x = x - A_k F(x), \quad k = 0, 1, 2, \dots$$

має ті ж розв'язки, що і вихідне рівняння $F(x) = 0$, і для приближеного знаходження цих розв'язків можна формально записати ітераційний процес

$$x^{(k+1)} = x^{(k)} - A_k F(x^{(k)}), \quad k = 0, 1, 2, \dots \quad (9)$$

який має вигляд метода простих ітерацій при

$\Phi(x) = \Phi_k(x) = x - A_k F(x)$. У випадку $A_k = A$ - це дійсно метод простих ітерацій з лінійною збіжністю послідовності $x^{(k)}$. Якщо же A_k різні за різних k то формула (9) визначає велику кількість ітераційних методів з матричними параметрами A_k . Розглянемо деякі з цих методів.

Припустимо $A_k = [F'(x^{(k)})]^{-1}$, де

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

- матриця Якобі функції $F(x)$. Підставимо це A_k в (9), отримаємо явну формулу метода Ньютона

$$x^{(k+1)} = x^{(k)} - [F'(x^{(k)})]^{-1} F(x^{(k)}) \quad (10)$$

Цю формулу, що вимагає перетворення матриць на кожній ітерації, можна переписати в неявному вигляді:

$$F'(x^{(k)})(x^{(k+1)} - x^{(k)}) = -F(x^{(k)}) \quad (11)$$

Використання (11) припускає при кожному $k=0,1,2,\dots$ розв'язок лінійної алгебраїчної системи

$$F'(x^{(k)})p^{(k)} = -F(x^{(k)})$$

відносно векторній поправці $p^{(k)} = (p_1^{(k)}, p_2^{(k)}, \dots, p_n^{(k)})$, а потім додавання цієї поправки до поточного наближення для отримання наступного:

$$x^{(k+1)} = x^{(k)} + p^{(k)}$$

До розв'язку таких лінійних систем можна використовувати найрізноманітніші методи як прямі, так і ітераційні в залежності від розмірності n розв'язуваної задачі і специфіки матриць Якобі $J(x^{(k)})$. Порівнюючи (11) з формальним розкладом $F(x)$ в ряд Тейлора

$$F(x) = F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)}) + \frac{1}{2!} F''(x^{(k)})(x - x^{(k)})^2 + \dots = 0$$

бачимо, що послідовність x_k в методі Ньютона отримується в результаті заміни при кожному $k=0,1,2,\dots$ нелінійного рівняння $F(x) = 0$ чи при допустимій гладкості $F(x)$ рівняння

$$F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)}) + \frac{1}{2!}F''(x^{(k)})(x - x^{(k)})^2 + \dots = 0$$

лінійним рівнянням

$$F(x^{(k)}) + F'(x^{(k)})(x - x^{(k)}) = 0$$

тобто з покроковою лінеаризацією. Як наслідок цього факту, може полягати, що при допустимій гладкості $F(x)$ і достатньо гарному початковому наближенні $x^{(0)}$ збіжність, яка виникає при методі Ньютона послідовності x_k до розв'язку x^* буде квадратичною і в багаторазовому випадку.

Новим фактором, який ускладнює використання метода Ньютона до розв'язання n вимірних систем, є необхідність розв'язання n -вимірних лінійних задач на кожній ітерації, обчислення яких збільшується зі збільшенням n , тобто кажучи, непропорційно швидко. Зменшення таких затрат є одним з напрямків модифікації метода Ньютона.

Модифікований метод Ньютона

Модифікований метод Ньютона полегшує першу задачу. Модифікація полягає в тому, що матриця обчислюється не в кожній точці, а лише в початковій.

Якщо матрицю Якобі $F'(x)$ обчислювати і перетворити лише один раз - в початковій точці $x^{(0)}$, то від метода Ньютона (11) перейдемо до модифікованого методу Ньютона

$$x^{(k+1)} = x^{(k)} - [F'(x^{(0)})]^{-1} F(x^{(k)}) \quad (12)$$

Цей метод потребує значно менших обчислень на один ітераційний крок, але ітерації при цьому може потребуватися значно більше для досягнення заданої точності в порівнянні з основним методом Ньютона (11) оскільки, будучи частковим випадком методу ітерацій $A_k = [F'(x^{(k)})]^{-1}$ він має лише швидкість збіжності геометричної прогресії.

Компромісний варіант - це обчислення і перетворення матриць Якобі не на кожному ітераційному кроці, а через декілька кроків (інколи такі методи називаються рекурсивними). Наприклад, просте чергування основного (11) і модифікованого (12) методів Ньютона приводить до ітераційної формули

$$x^{(k+1)} = x^{(k)} - [F'(x^{(k)})]^{-1} F(x^{(k)}) \quad (13)$$

де $A_k = [F'(x^{(k)})]^{-1}$ $k=0,1,2,\dots$ За $x^{(k)}$ тут приймається результат останнього прийому одного кроку основного, а потім одного кроку модифікованого метода, тобто

$$\{z^{(k)} = x^{(k)} - A_k F(x^{(k)})\} \quad x^{(k+1)} = z^{(k)} - A_k F(z^{(k)}) \quad (14)$$

Доведено, що такий процес при певних умовах призводить до збіжних послідовностей $x^{(k)}$.

Запитання для самоконтролю

1. В чьому полягає складність розв'язання систем нелінійних рівнянь?
2. Який фактор ускладнює використання методу Ньютона до розв'язання n вимірних систем?
3. Чим відрізняється метод Ньютона від модифікованого методу Ньютона?

ЛЕКЦІЯ 9. Обчислювальне інтегрування. Метод прямокутників. Метод трапецій. Метод Симпсона.

Задачі у яких потрібно обчислювати значення визначених інтегралів виникають майже в усіх галузях математики. Іноді вдається знайти аналітичну формулу, тобто виразити невизначений інтеграл у вигляді комбінації алгебраїчних трансцендентних функцій, визначений інтеграл при цьому обчислюється шляхом підстановки меж інтегрування за формулою Ньютона- Лейбніца

$$I = \int_a^b f(x)dx = F(b) - F(a) \quad (1)$$

- де a, b - нижня і верхня межі інтегрування;
 $f(x)$ - неперервна функція на відрізку $[a, b]$;
 $F(x)$ – первісна для підінтегральної функції $f(x)$.

Зазначена формула в практичних задачах застосовується рідко. Тому що тільки для вузького класу підінтегральних функцій можливо точно знайти первісну.

Таким чином, до чисельного інтегрування звертаються в тих випадках, коли не можна через елементарні функції аналітично записати первісну інтеграла (1) або коли подібний запис має складний вигляд. При цьому звичайно заміняють $f(x)$ на таку апроксимуючу функцію $\varphi(x) \approx f(x)$, щоб інтеграл від неї легко обчислювався в елементарних функціях.

При наближеному обчисленні інтеграла використовується той факт, що інтеграл чисельно дорівнює площі фігури обмеженою знизу – віссю абсцис, з боків – прямими $x = a$, $x = b$, зверху – графіком функції $y = f(x)$. Використання наближених методів припускає два способи розбивки вихідного інтервалу на менші:

1. Розбиття на підінтервали робиться заздалегідь, звичайно їхні розміри вибираються рівними.
2. Місце розташування і довжина підінтервалів визначається шляхом аналізу в процесі виконання алгоритму інтегрування.

Наведемо характеристику методів, що застосовується для чисельного інтегрування.

Методи *Ньютона-Котеса*: засновані на поліноміальній апроксимації підінтегральної функції (для конкретної функції підбирається поліном відповідного степеню). Алгоритми методів прості і легко піддаються програмній реалізації (методи прямокутників, трапеції, Симпсона).

Сплайнові методи: базуються на апроксимації підінтегральної функції сплайнами, що представляють собою кусковий поліном.

Методи *Гаусса-Кристоффеля*: для досягнення більшої точності при апроксимації підінтегральної функції використовують нерівновіддалені вузли (точки обчислення функції), обрані за алгоритмом, що забезпечує мінімальну похибку інтегрування для найбільш складних функцій.

Методи *Монте-Карло*: вузли вибираються за допомогою датчика випадкових чисел, відповідь носить імовірнісний характер.

Спеціальні методи: алгоритми розробляються на основі обліку особливостей конкретних підінтегральних функцій, що дозволяє істотно скоротити час і зменшити похибку обчислення інтегралів.

Метод прямокутників

Розіб'ємо відрізок $[a, b]$ на n рівних по довжині підінтервалів у такий спосіб:

$$a = x_0 < x_1 < \dots < x_n = b$$

Довжина кожного підінтервалу дорівнює $h = (b - a)/n$ тобто

$$x_{i+1} = x_i + h$$

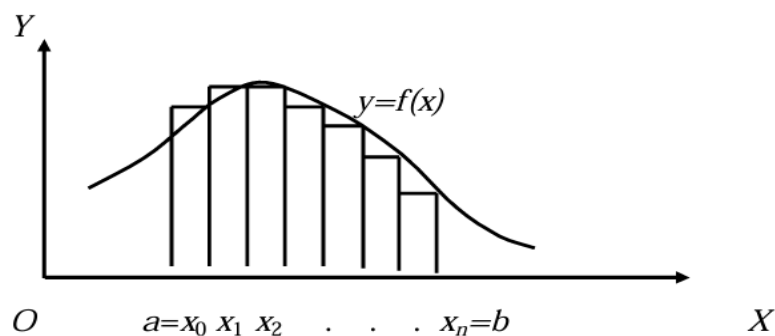
У кожному підінтервалі замінимо площу отриманої фігури площею прямокутника, безумовно, з якоюсь похибкою (тобто апроксимуємо підінтегральну функцію поліномом нульового ступеня)

$$S_i = f(\varepsilon_i)h, \quad x_i < \varepsilon_i < x_{i+1}$$

Тоді значення площі всіх прямокутників, що утворились (і наближене значення інтеграла) буде дорівнювати:

$$\int_a^b f(x)dx \approx (b - a)/n * [f(\varepsilon_0) + f(\varepsilon_1) + \dots + f(\varepsilon_n)]$$

У залежності від того, яку точку ми обираємо в якості ε_i ми можемо одержати метод лівих прямокутників ($\varepsilon_i = x_i$) або метод правих прямокутників ($\varepsilon_i = x_{i+1}$) і метод середніх прямокутників ($\varepsilon_i = (x_{i+1} + x_i)/2$). З усіх перерахованих методів найменшу похибку має останній.



Незалежно від обраного методу в процесі чисельного інтегрування похибка, викликана неточним описом функції $f(x)$ більш простою функцією $\varphi(x)$, буде зменшуватись при збільшенні числа розбиття n за рахунок більш точної апроксимації підінтегральної функції (помилка методу або помилка обмежень) але при цьому буде зростати похибка за рахунок підсумовування часткових інтегралів, викликана кінцівкою розрядної сітки комп'ютера (так звана помилка округлення).

Сумарна функція помилки буде мати явний мінімум, тобто оптимальне число розбиття. Для всіх методів помилка округлення приблизно дорівнює $\frac{1}{h}$. Для методу середніх прямокутників помилка обмежень дорівнює

$$R \approx \frac{h^2}{24} \sum_{i=1}^n h f''(x_i) \approx \frac{h^2}{24} \int_{x_0}^{x_n} f''(x) dx$$

Степінь кроку h , що пропорційна величина помилки R , називається порядком методу інтегрування. Метод середніх прямокутників має другий порядок.

Приведена вище формула дозволяє оцінити порядок отриманої помилки до обчислення інтеграла, але для практичного застосування вона малопригодна. На практиці для оцінки помилок застосовується наступна формула (формула Рунге)

$$R \approx (I_h - I_{2h}) / (2^p - 1)$$

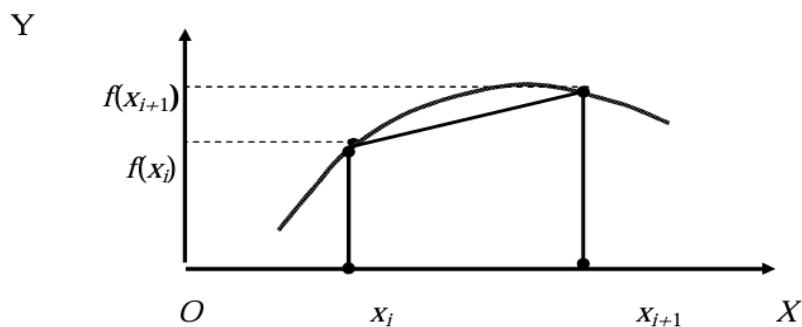
де I_h, I_{2h} – значення інтеграла, обчисленні відповідно з кроками h і $2h$

p – порядок методу, що застосовується.

Ця формула дозволяє оцінити похибку у ході обчислювального процесу при реалізації того чи іншого методу.

Метод трапецій

Як і в попередньому випадку, розбиваємо відрізок $[a, b]$ на n рівних частин. На кожному підінтервалі $[x_i, x_{i+1}]$ підінтегральну функцію $f(x)$ замінюємо поліномом першого ступеня, при цьому проведемо пряму через значення функції на межах підінтервалу



Таким чином, обчислення площі криволінійної фігури в підінтервалі замінюємо обчисленням площі трапеції, що дорівнює добутку висоти h на півсуму $(f(x_i) + f(x_{i+1}))/2$

Значення всього інтеграла при цьому дорівнює

$$I = \int_a^b f(x)dx \approx h/2 \cdot (f(x_0) + 2f(x_1) + \dots + 2f(x_{n-1}) + f(x_n))$$

Для методу трапецій помилка обмеження дорівнює

$$R \approx h^2/12 \cdot \int_{x_0}^{x_n} f''(x)dx$$

Приведена формула показує, що метод трапецій має більшу похибку ніж метод середніх прямокутників, хоча апроксимація підінтегральної функції проводиться поліномом першого, а не нульового ступеня.

Екстраполяційний перехід до межі

Для того щоб зменшити помилку при чисельному інтегруванні, можна порівняно просто удосконалити метод трапецій. Нехай I_h - значення інтеграла, отримане по формулі трапецій з кроком h , а I_k - значення інтеграла з кроком k , тоді більш точне значення інтеграла може бути отримано по формулі Ричардсона

$$I = I_h + \frac{I_h - I_k}{h^2 - k^2}$$

Цей метод називається екстраполяційним переходом до межі і має краще ніж у методі трапецій наближення. Якщо друга похідна $f''(x)$ постійна на $[a, b]$, то помилка обмежень у формулі Ричардсона дорівнює 0.

Метод Симпсона.

Припустимо, що ми обчислили значення інтеграла за формулою трапецій із кроками h і $k=2h$, тоді

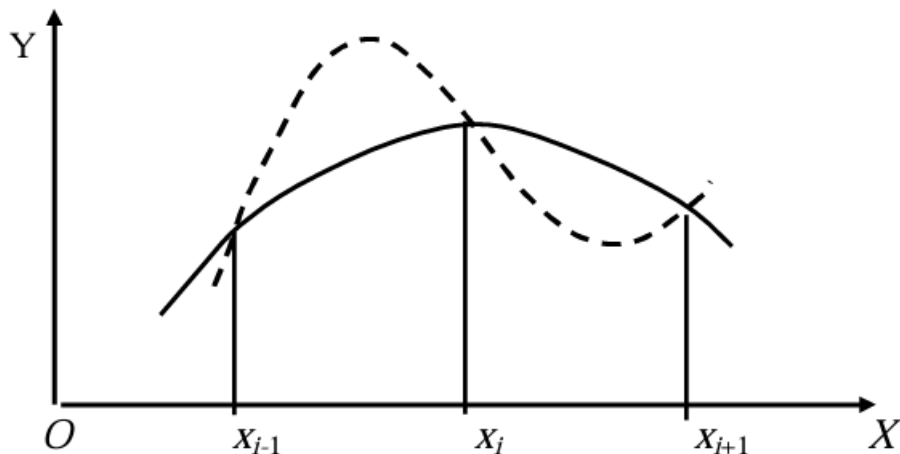
$$I_h = \frac{h}{2} \cdot (f(x_0) + 2f(x_1) + \dots + f(x_n))$$

$$I_k = \frac{k}{2} \cdot (f(x_0) + 2f(x_1) + \dots + f(x_n))$$

При підстановці цих значень у формулу Ричардсона, одержимо наступний вираз для обчислення інтеграла

$$I = \frac{h}{3} * (f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \dots + 4f(x_{n-1}) + f(x_n))$$

Метод обчислення визначених інтегралів по цій формулі має назву методу Симпсона, при цьому підінтегральна функція апроксимується параболою, що проходить через 3 точки.



Помилка обмеження для цього методу дорівнює

$$R \approx \frac{h^4}{180} \int_{x_0}^{x_n} f^{(IV)}(x) dx$$

З формули слідує що метод Симпсона має четвертий порядок точності, тобто його застосування дозволяє одержати гарні результати, якщо четверта похідна підінтегральної функції не занадто велика.

Приведемо приклад застосування методу середніх прямокутників для знаходження наступного інтеграла (кількість розбивок $n=10$)

$$I = \int_0^1 \frac{1}{(1+x^2)} dx$$

Обчислення інтеграла наведено у таблиці

i	x_i	ε_i	$F(\varepsilon_i)$
0	0	0.05	0.9995

1	0.1	0.15	0.9780
2	0.2	0.25	0.9412
3	0.3	0.35	0.8909
4	0.4	0.45	0.8316
5	0.5	0.55	0.7678
6	0.6	0.65	0.7029
7	0.7	0.75	0.6400
8	0.8	0.85	0.5806
9	0.9	0.95	0.5256
10	1.0		
Σ			7.8581

Тоді за формулою прямокутників значення визначеного інтеграла

$$I \approx \left(\frac{1-0}{10} \right) \cdot 7,8581 = 0,78581$$

Запитання для самоконтролю

1. В яких випадках звертаються до чисельного інтегрування?
2. Назвіть та охарактеризуйте методи чисельного інтегрування.
3. Вкажіть формулу оцінки похибки для методів?
4. У чому состоит метод Симпсона?

ЛЕКЦІЯ 10. Обчислювальні методи оптимізації. Постановка задачі і класифікація методів

Для дослідження багатьох інженерно-технічних і економічних процесів спочатку складають математичну модель у вигляді математичної задачі, а потім використовують чисельні методи для її розв'язування. Часто це приводить до необхідності вирішувати задачу нелінійного програмування (НЛП), що складається в знаходженні таких значень змінних $x = (x_1, x_2, \dots, x_n)$, при яких досягається мінімум цільової функції

$$f(x_1, x_2, \dots, x_n) \rightarrow \min \quad (1)$$

і виконуються обмеження

$$g_1(x_1, x_2, \dots, x_n) \leq 0 \quad g_2(x_1, x_2, \dots, x_n) \leq 0 \quad \dots \quad g_m(x_1, x_2, \dots, x_n) \leq 0 \quad (2)$$

де хоча б одна функція $f, g_1, g_2, \dots, g_m$ нелінійна, а обмеження можливі у вигляді рівностей. На змінні x можуть бути накладені обмеження не заперечності. Якщо всі функції лінійні то задача буде задачею лінійного програмування (ЛП), для розв'язання якої розроблено спеціальні методи. Функція (1) називається *цільовою* функцією. Вектор x називається *припустимим*, якщо він задовольняє обмеженням (2). Вектор x називається *оптимальним розв'язанням* задачі (1)-(2) якщо є припустимим і функція (1) досягає на ньому мінімуму. Задача оптимізації може бути поставлена і на пошук максимуму цільової функції але вона може бути легко зведена до функції максимізації.

З математичної точки зору оптимізаційна задача- це задача, у якій потрібно знайти оптимум (максимум або мінімум) деякої величини при виконанні яких-небудь умов. Розробкою методів розв'язання подібних задач займається галузь математики що називається математичним програмуванням. Однак до тепер загальних методів розв'язання довільної задачі НЛП не існує. Для різних класів задач, пов'язаних із визначеними видами цільової функції і структурою обмежень, розроблені різні класи методів:

- методи лінійного програмування;
- методи динамічного програмування (коли обмеження включають у себе як параметр час);
- методи стохастичного програмування (коли в задачі є імовірнісний параметр);
- методи геометричного програмування;
- комбінаторні та евристичні методи.

Умовно класифікацію оптимізаційних задач (і пов'язаних з ними методів) можна провести таким чином:

- 1) у залежності від вигляду функцій:
 - лінійні і нелінійні;
 - опуклі і неопуклі.
- 2) по використанню похідних у розв'язанні:
 - не використовують похідні (метод покоординатного спуска);
 - що використовують 1-у похідну (градієнтний метод);
 - що використовують 2-у похідну (метод Ньотона).
- 3) у залежності від наявності обмежень:
 - задачі з обмеженнями;
 - задачі без обмежень.

- 4) у залежності від кількості змінних:
 - з однією змінної;
 - з багатьма змінними.
- 5) у залежності від кількості екстремумів:
 - однокстремальні;
 - багатокстремальні.

Геометрично цільова функція (1) визначає собою поверхню, а обмеження (2)- припустима множина розв'язків n - мірного евклідового простору (припустима область). Розв'язання задачі НЛП зводиться до знаходження точки з припустимої множини у якій $f(x)$ досягає екстремуму. Розв'язання може знаходитися або у середині області або на її межі. Наступна теорема дає необхідні і достатні умови існування розв'язання задачі.

Теорема: Якщо цільова функція неперервна та припустима множина замкнута, непорожня та обмежена, то розв'язання задачі (1)-(2) існує.

Мінімізація функції однієї змінної.

Знаходження мінімуму функції однієї змінної є важливою самостійною задачею. Крім того, мінімізація функції однієї змінної складає істотну частину різних процедур мінімізації функції декількох змінних.

Точка x_0 називається точкою локального мінімуму функції $y = f(x)$ на $[a, b]$ якщо $x_0 \in [a, b]$ і існує таке $\delta > 0$, що при усіх $x \in [a, b]$, $|x - x_0| < \delta$ виконується нерівність $f(x_0) \leq f(x)$.

Точка x^* називається точкою глобального мінімуму функції $y = f(x)$ на $[a, b]$, якщо при всіх $x \in [a, b]$ виконується нерівність $f(x^*) \leq f(x)$.

Кожен глобальний мінімум є локальним. Зворотне не вірно.

Задача мінімізації полягає в відшукуванні точки глобального мінімуму x^* функції $y = f(x)$ на $[a, b]$ і обчисленні $f(x^*)$. Класичний метод розв'язання задачі мінімізації припускає, що функція $y = f(x)$ неперервна і диференційована. Тоді на $[a, b]$ існує точка глобального мінімуму, що або збігається з одним з кінців відрізка або є внутрішньою точкою у якій $f'(x) = 0$. Тому для відшукування точки глобального мінімуму потрібно знайти всі корені рівняння $f'(x) = 0$, що належать відрізку $[a, b]$, обчислити значення функції в цих точках а також $f(a)$ і $f(b)$. Та точка у якій досягається мінімум з усіх обчислених значень і є шуканою точкою глобального мінімуму функції.

Описаний метод має обмежене застосування, тому що не всі функції, що застосовуються на практиці є диференційованими. Крім того, розв'язання рівняння $f(x) = 0$ може виявитися нелегкою задачею. На практиці частіше використовують наближені методи.

Методи рівномірної сітки (повного перебору) та золотого перетину.

Суть методу рівномірної сітки полягає в тому, що на відрізку $[a, b]$ будується рівномірна сітка із кроком h і обчислюються значення функції $f(x)$ у вузлах сітки. За точку глобального мінімуму береться вузол сітки у якому значення функції мінімально. Основна проблема цього методу складається в раціональному виборі значення кроку h , тому що він повинен забезпечувати необхідну точність і не приводити до зайвих витрат машинного часу. Будемо рахувати що функція $y = f(x)$ задовольняє на відрізку $[a, b]$ умові Липшиця, якщо існує таке число L (константа Липшиця) що для будь яких $x', x'' \in [a, b]$ виконується нерівність $|f(x') - f(x'')| \leq L |x' - x''|$.

Таким чином, умова Липшиця накладає обмеження на швидкість або збільшення спадання функції. Можливо показати, що якщо вибрати крок $h = \frac{(b-a)}{n} \leq \frac{2\varepsilon}{L}$ то знайдене за допомогою методу повного перебору наближене значення точки глобального мінімуму буде задовільняти необхідній точності. Такий вибір гарантує мінімальну кількість обчислень функції при досягненні необхідної точності.

Задача методу золотого перетину полягає в тому, щоб одержати послідовність вкладених одне в одне відрізків $[x_i, x_{i+1}]$ що містять точку x^* , у якій функція приймає мінімальне значення. Нехай відрізок $[a, b]$ розбитий на дві частини точкою x . Говорять що точка x робить золотий перетин відрізка якщо відношення довжини усього відрізка до довжини більшої його частини дорівнює відношенню довжини більшої частини до меншої

$$\frac{b-a}{x-a} = \frac{x-a}{b-x}.$$

Опишемо стратегію пошуку в методі золотого перетину. Для заданого початкового відрізка $[a, b]$ по приведених формулах обчислюємо x_1 і x_2 та знаходимо значення функції $f(x_1)$ та $f(x_2)$. Із властивостей функції випливає що якщо $f(x_1) < f(x_2)$, то точка мінімуму належить відрізку $[x_2, b]$, а якщо $f(x_1) > f(x_2)$, то точка мінімуму належить відрізку $[a, x_1]$. Знову отриманий відрізок вкладений у вихідний і містить шукану точку мінімуму функції. Крім того він вже містить одну з точок (x_1 – у першому випадку і x_2 – у другому),

що робить його золотим перетином. Тоді, побудувавши по формулам другу точку, ми можемо продовжити процес одержання інтервалів, довжина яких спадає та всередині кожного з яких розташована точка мінімуму x^* . Обчислювальна процедура закінчується коли досягнута задана точність $|b - a| \leq \varepsilon$.

Метод золотого перетину є найбільш економічним аналогом методу дихотомії стосовно задач мінімізації. Він застосований навіть до недиференційованих функцій і завжди збігається; збіжність його лінійна. Якщо на відрізку $[a, b]$ функція має кілька локальних мінімумів то процес зійдеться до одного з них, але не обов'язково до найменшого.

Для розв'язання одновимірної задачі нелінійного програмування використовують також метод Пауелла і метод парабол. У методі Пауелла проводяться зростаючі за величиною кроки доти поки не буде пройдений мінімум, а потім виконується одноразова квадратична інтерполяція. По методу парабол проводяться обчислення, що використовують першу та другу похідні функції що мінімізується.

Порівняльний аналіз мінімізації спеціально підібраної функції Розенброка методами послідовного перебору, золотого перетину і метода Пауелла показує, що найбільш переважним методом для розв'язання задач цього класу є метод золотого перетину.

Мінімізація функцій багатьох змінних.

Численні методи відшукування екстремума функції складаються з побудови в побудові послідовності векторів $x^{(k)}$ що задовольняють умові $f(x^0) > f(x^1) > \dots > f(x^n)$. методи побудови таких послідовностей називають методами спуску, найбільш розповсюджені розглянемо далі.

Метод покоординатного спуску.

Цей метод є найбільш простим із прямих методів пошуку мінімуму функції без обмежень. Розглянемо ідею метода на прикладі мінімізації функції двох змінних $f(x, y)$. Виберемо початкове наближення $M_0(x^0, y^0)$ зафіксуємо y^0 і знайдемо мінімум функції однієї змінної $f(x, y^0)$ будь яким відомим методом. Нехай він досягається при $x = x^1$. Уздовж прямої паралельній осі ОХ перемістимося в точку $M_1(x^1, y^0)$. Фіксуємо тепер точку x^1 і знайдемо мінімум функції однієї змінної $f(x^1, y)$. Нехай це буде y^1 .

Рухаємося вздовж прямої паралельній осі ОУ перемістимося в точку $M_2(x^1, y^1)$. Далі спускаємося по прямій паралельній осі ОХ і т.і. Продовжуємо цей процес доти поки функція продовжує спадати, тобто поки $f(x^{k+1}, y^{k+1}) < f(x^k, y^k)$. Як тільки функція $f(x, y)$ перестає спадати, спуск припиняється.

Існують багато модифікацій методу покоординатного спуску, наприклад метод покоординатного спуску з постійним кроком. У цьому випадку не потрібно на кожному кроці алгоритму використовувати який-небудь метод для знаходження мінімуму функцій однієї змінної. Спуск здійснюється в напрямку координатних вісей у бік спадання функції з постійним кроком h . У загальному випадку потрібно застосовувати різні кроки для різних змінних h_x, h_y але для спрощення можна обходитися одним кроком h , що трохи зменшує швидкість методу, при цьому наступні значення змінних обчислюються по формулах

$$x^{k+1} = x^k + h, \quad y^{k+1} = y^k + h$$

В цілому алгоритм покоординатного спуску повинен припинятись, коли функція $f(x, y)$ перестає убавати. Для цього на кожному кроці алгоритму запам'ятовується поточне значення функції, що порівнюється зі значенням функції, отриманому після одномірних мінімізацій при фіксації відповідних змінних. Приведені міркування легко переносяться на випадок функції багатьох змінних: при цьому на кожному кроці алгоритму фіксується усі змінні, крім однієї, по якій проводиться мінімізація. Порядок фіксації змінних при цьому не має значення.

Таким чином, точку глобального мінімуму за допомогою методів НЛП в загальному випадку достатньо важко знайти, тому, для підвищення імовірності її знаходження можна робити декілька спусків, використовуючи різні початкові точки вибираючи з отриманих значень локальних мінімумів найменше.

Метод можливо доповнити блоком у якому після неможливості зменшення функції можна зменшити крок алгоритму h що дозволить збільшити точність методу але збільшить і час рахування. У цьому випадку алгоритм припиняється при досягненні кроку деякого мінімального заздалегідь заданого значення. Зазвичай довжина кроку приблизно пропорційна числу вдаливих кроків, що мали місце раніше в цьому координатному напрямку під час декількох попередніх циклів.

У методі спірального покоординатного спуску крок h змінюється щоразу, по деякому алгоритму, при переході від пошуку мінімуму по одній

змінній до пошуку мінімуму по іншій змінній Y тривимірному просторі це нагадує спуск у западину по спіралі Звичайно цей метод дає деяке скорочення часу пошуку.

Алгоритм покоординатного спуску працює погано якщо має місце взаємодія між змінними (наприклад є добуток $x_1 x_2$ у цільовій функції).

Таким чином цей метод не можна рекомендувати, якщо користувач не має справи з цільовою функцією, у яких взаємодії між змінними не істотні.

Перевагою методу варто вважати його універсальність (застосовність до недиференційовних функцій), недоліком- повільна швидкодія в деяких випадках.

Метод градієнтного спуску.

Для простоти ознайомлення з методом обмежимося задачею мінімізації функції двох змінних $f(x, y)$, вважаючи її диференційованою.

Градієнтом функції в точці називають вектор $(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y})$ що вказує напрямом зростання функції. Напрямок убуття функції характеризується *антиградієнтом* $(-\frac{\partial f}{\partial x}, -\frac{\partial f}{\partial y})$.

Існують різні версії методу градієнтного спуску, об'єднані загальною ідеєю: кожне наступне наближення до точки мінімуму вибирають у напрямку, обумовленому антиградієнтом, проведеним із точки, що відповідає попередньому наближенню. Однак антиградієнт дає тільки напрям оптимізації, але не величину кроку на який можна пересунитися одночасно зі зменшенням функції. Тому численні модифікації методу в основному відрізняються правилами вибору кроку на кожній ітерації і правилами руху уздовж отриманого напрямку.

Один з різновидів методу покоординатного спуску- метод найшвидкішого спуску. Спуск починається з нульового наближення точки (x^0, y^0) , верхній індекс позначає номер кроку. Обчислимо компоненти антиградієнта $d_1 = -\frac{\partial f}{\partial x}(x^0, y^0)$, $d_2 = -\frac{\partial f}{\partial y}(x^0, y^0)$. Будемо вважати що числа d_1, d_2 не дорівнюють нулю одночасно (якщо ці два числа дорівнюють 0, то точка (x^0, y^0) є підозрілою на екстремум, у цьому випадку запам'ятовується значення функції в ній і вибирається інша точка як нульове наближення). Виберемо деякий крок $h_1 > 0$ і покладемо $x^1 = x^0 + h_1 \cdot d_1$, $y^1 = y^0 + h_1 \cdot d_2$. Будемо вважати, що крок обрано вірно, якщо

$f(x^1, y^1) < f(x^0, y^0)$, при не виконанні цієї нерівності крок зменшується. Після одержання нового наближення (x^1, y^1) обчислюється антиградієнт у новій точці, будується нове наближення (x^2, y^2) тощо. Обчислення по алгоритму продовжується доти поки значення кроку не стане менше деякого мінімального значення.

За допомогою викладеного алгоритму можна знаходити стаціонарні точки, що у загальному випадку можуть не бути точками мінімуму. Тому для підвищення імовірності знаходження мінімуму рекомендується робити кілька спусків з різних початкових точок.

Недоліками запропонованого алгоритму є його застосування тільки до диференційовних функцій, а також залежність від вибору масштабу змінних що оптимізуються. Якщо мінімізуєма функція дуже витягнута в просторі, то процедура найшвидкішого спуску збігається занадто повільно, щоб бути ефективною чи може взагалі не зійтися за розумний час. Іноді в цьому випадку допомагає масштабування змінних. Перевагою застосування методу є його висока швидкодія в більшості випадків.

Випадковий пошук.

Методи спуску при мінімізації функцій неповноцінні при завданні складних функцій мети. Якщо локальних екстремумів багато, то спуск з одного нульового наближення може зійтися тільки до одному з локальних мінімумів, не обов'язково абсолютному. Тоді для дослідження і рішення задачі можна застосувати випадковий пошук.

Для його застосування потрібно мати послідовність рівномірно розподілених величин. Їх можна одержати використовуючи який небудь пристрій на виході якого з імовірністю 0,5 можуть з'являтися цифри 0 чи 1; поява тієї або іншої цифри повинна бути випадковою. Якщо тепер записати число як $\gamma = \alpha_1 \alpha_2 \alpha_3 \dots$ де як розряди будуть виступати цифри, вироблені генератором, та перевести його з двійкової системи числення в десяткову, то одержимо послідовність випадкових чисел $0 \leq \gamma \leq 1$. У дійсності всі оператори алгоритмічних мов, що дають послідовність "випадкових" чисел, реалізовані як своєрідні генератори, що дають псевдовипадкові числа, що повторюються через визначену кількість ітерацій. Однак період їх повторень настільки великий, що вони з успіхом використовуються в численних алгоритмах випадкових пошуків. У більшості випадків у комп'ютерах для одержання псевдовипадкових чисел застосовується алгоритм, зв'язаний з виділенням дробової частини добутку

$\gamma_i = \{A \cdot \gamma_{i-1}\}$ де A - дуже велика константа, а фігурна дужка позначає дробову частину числа. Якість псевдовипадкових чисел дуже залежить від вибору величини A - у двійковій формі число A повинне мати досить "випадковий" вигляд.

Для розв'язання умовної задачі нелінійного програмування припускають, що мінімум належить деякій замкнутій області; лінійним перетворенням координат розміщують її усередині одиничного n -мірного куба. Вибирається N випадкових точок (координатами яких є послідовності псевдовипадкових чисел γ_i) у кожній з них обчислюються значення цільової функції з усіх вибирається найменше. При високій швидкодії сучасних комп'ютерів такий підхід безумовно себе виправдовує. Чим більше точок обирається для рішення, тим більш імовірно знаходження екстремума.

Великою перевагою запропонованого підходу є відсутність яких-небудь вимог до цільової функції, недоліком - повільна швидкість збіжності та імовірнісний характер знайденого рішення.

Запитання для самоконтролю

1. Від чого залежить класифікацію оптимізаційних задач?
2. Які обмеження має мінімізація функції однієї змінної?
3. Вкажіть суть методу рівномірної сітки?
4. Чим відрізняються друг від друга методи спуску?

ЛЕКЦІЯ 11. Лінійне програмування. Геометричний метод розв'язання задачі ЛП

Теорія ЛП займається створенням методів що дозволяють оптимізувати лінійну функцію, на яку накладено лінійні обмеження. В основному задачі ЛП ставляться в теорії планування і керування економічними процесами, а також у технічних задачах. З погляду класичної математики подібні задачі являють собою оптимізаційні задачі на умовний екстремум. Канонічна форма задачі ЛП виглядає в такий спосіб:

$$\text{Знайти } Z = \max(c_1 x_1 + c_2 x_2 + \dots + c_n x_n)$$

при виконанні обмежень

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2$$

.....

$$a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m$$

та обмежень на змінні (обмеження не заперечності)

$$x_1 \geq 0, x_2 \geq 0, \dots, x_n \geq 0$$

Різні постановки задач ЛП можуть приводити до різного математичного запису (в обмеженнях можуть бути як знаки рівностей так і нерівностей).Перехід від різних формулювань до канонічного запису потрібний для того щоб користатися стандартними програмами реалізованими на комп'ютерах(а вони у свою чергу орієнтовані на реалізацію відомого симплекс-методу розв'язання задачі ЛП). Однак для використання графічного методу розв'язання краще використовувати стандартну форму запису у якій обмеження представлені у формі нестрогих нерівностей .

Для аналізу задача ЛП уведемо деякі визначення . Сукупність значень $(x_1, x_2, \dots, x_n)$ що задовольняють обмеженням називається припустимим (можливим) розв'язанням. Множина усіх припустимих розв'язань називають областю припустимих розв'язань (ОПР). Припустиме розв'язання для якого цільова функція досягає максимуму чи мінімуму називається оптимальним. Використовуючи приведені визначення задачу ЛП можна сформулювати в такий спосіб: необхідно із сукупності припустимих розв'язань вибрати те що відповідає мінімуму цільової функції.

Для переходу від одного виду запису задачі ЛП до іншого необхідно використовувати наступні прийоми:

- 1) перехід від задачі максимізації до мінімізації та навпаки:

$$Z = \max(c_1x_1 + c_2x_2 + \dots + c_nx_n)$$

$$Z_1 = -Z = \min(-c_1x_1 - c_2x_2 - \dots - c_nx_n)$$

- 2) перехід від обмежень-нерівностей до обмежень рівностей. Для цього в задачу вводиться додаткова змінна x_{n+1} , що є фіктивною й у цільову функцію входить з коефіцієнтом 0.

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \leq b_1$$

перехід

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n + x_{n+1} = b_1$$

За допомогою таких прийомів можна звести задачу ЛП до канонічної або стандартної форми. Для точного розв'язання задачі ЛП застосовується симплекс-метод. Основна ідея методу складається з пошуку точки перетинання n лінійно незалежних гіперплощин, що задаються обмеженнями і послідовному переході до іншої точки зі зменшенням значення цільової функції. Такі точки можна представити вершинами деякого багатогранника, що є перетинанням кінцевого числа півпросторів- припустимих рози'зань. Головна перевага симплекс-метода універсальність і простота реалізації на комп'ютері.

Геометричний метод розв'язання задачі ЛП

Нехай задача ЛП задана в стандартній формі. Геометрично ОПР утвориться перетинанням m множин, кожне з них визначається нерівністю $a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n \leq b_i$ і являє собою півпростір, що розташовано по одну сторону від гіперплощини $a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n = b_i$.

Перетинання півпросторів є областю припустимих розв'язків (і частіше опуклим багатогранником якщо область обмежена і не порожня). Лінії рівня $(c_1x_1 + c_2x_2 + \dots + c_nx_n)$ утворюють сімейство гіперплощин. Вектор нормалі до цих площин $c = (c_1, c_2, \dots, c_n)$ перпендикулярний цим площинам і визначає напрямок зростання лінійної форми.

Множина точок заданих нерівностями може бути порожньою обмеженою і необмеженою відповідно задача ЛП може не мати розв'язків мати один розв'язок мати нескінченну множину розв'язків або бути необмеженою. Варто мати на увазі що оптимум задачі ЛП досягається тільки в обмеженій точці припустимої множини. На відміну від задач нелінійного програмування в задачі ЛП оптимум завжди є глобальним.

Геометричний метод застосовується для розв'язання задачі ЛП двох змінних тому що тільки в цьому випадку можна побудувати наочно ОПР на площині. Метод складається з двох етапів. На першому будується ОПР як перетинання півплощин кожна з яких задана нерівністю- обмеженням. Після другого будується пряма $c_1x_1 + c_2x_2 = a$, де a - довільне число.

Думкою зрушуємо цю пряму паралельно самої собі вправо- нагору, якщо ми максимізуємо цільову функцію і вліво -униз якщо мінімізуємо. Рух продовжується доти поки пряма ще буде належати ОПР.

Розглянемо приклад розв'язання задачі ЛП графічним методом.

Етапи розв'язання:

1. будують прямі, рівняння, рівняння яких знаходять унаслідок заміни в обмеженнях знаків нерівностей на знаки рівностей;
2. знаходимо напівплощини, які визначаються з обмежень задачі;
3. визначаємо багатогранник розв'язку;
4. будуємо вектор $\vec{c} = (c_1, c_2)$;
5. будуємо пряму перпендикулярну вектору $\vec{c} = (c_1, c_2)$;
6. пересуваємо пряму з пункту 5 у напрямку вектора з пункту 4;
7. визначаємо координати крайньої точки множини, якої доторкнеться побудована пряма.
8. визначаємо числовий вираз для цільової функції.

Задача

$$z = \max(30x_1 + 40x_2)$$

при обмеженнях

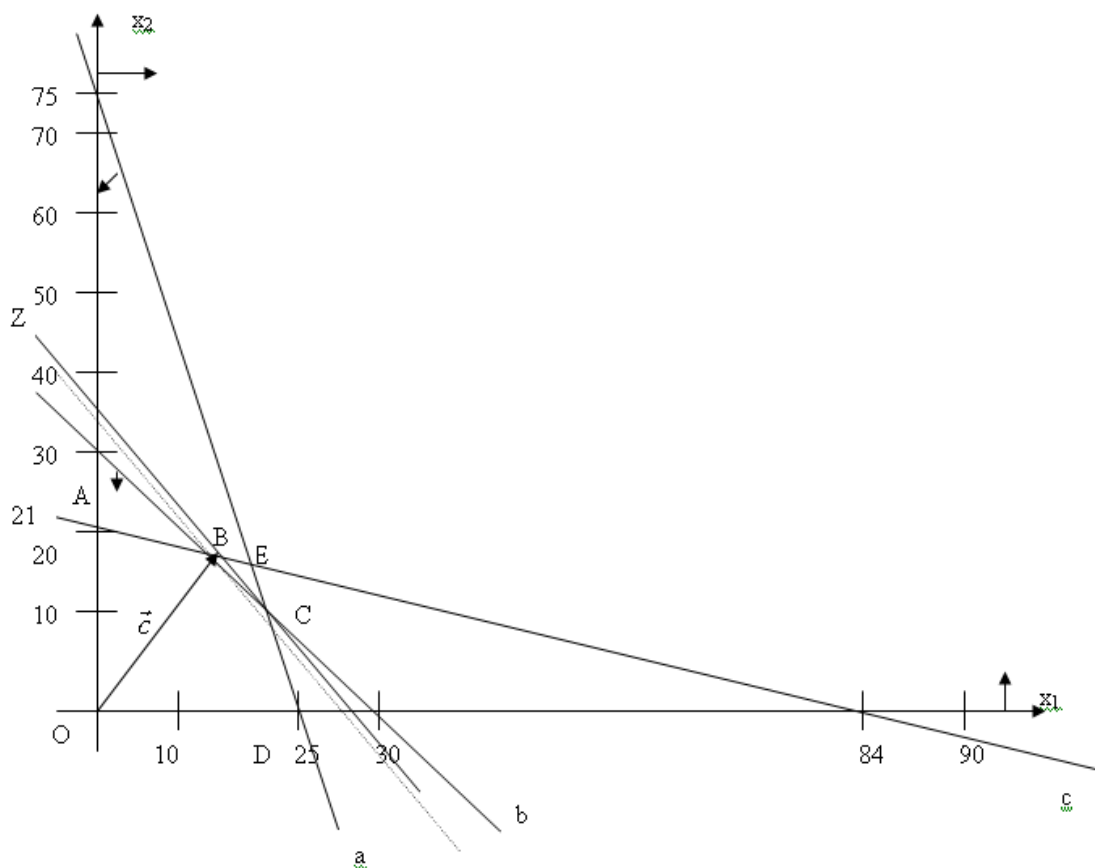
$$\begin{cases} 12x_1 + 4x_2 \leq 300 \\ 4x_1 + 4x_2 \leq 120 \\ 3x_1 + 12x_2 \leq 252 \\ x_1 \geq 0, \quad x_2 \geq 0 \end{cases}$$

Приведемо до рівностей систему

$$\begin{cases} 12x_1 + 4x_2 = 300 & (a) \\ 4x_1 + 4x_2 = 120 & (b) \\ 3x_1 + 12x_2 = 252 & (c) \\ x_1 = 0 & (d) \\ x_2 = 0 & (e) \end{cases}$$

Побудуємо прямі, розв'яжемо нерівності, визначимо область розв'язків задачі.. Показано на малюнку.

Многогранник розв'язку ОАВСД. Припустими розв'язки - вершини багатогранника. Вектор c - відображає напрямок зростання значень цільової функції. Максимальне значення цільової функції досягається в точці В(12,18).
Значення цільової функції $Z = 30 \cdot 12 + 40 \cdot 18 = 1080$



Запитання для самоконтролю.

1. Опишіть задачу лінійного програмування.
2. За якими правилами можливо привести задачу ЛП до задачі в канонічному вигляді?
3. Опишіть геометричний метод розв'язання задач ЛП.

ЛЕКЦІЯ 12 Розв'язання диференціальних рівнянь. Постановка задачі. Метод Ейлера. Метод Рунге Кутта.

Звичайними диференціальними рівняннями (ДР) можна описати рух системи взаємодіючих матеріальних точок, задачі хімічної кінетики, електричних ланцюгів, опору матеріалів і т.і. Взагалі говорячи, будь-яка фізична ситуація, де розглядається степінь зміни однієї змінної стосовно іншій змінній, описується ДР. Конкретна прикладна задача може приводити до будь-якого порядку чи до системи рівнянь будь-якого порядку. Однак звичайне ДР p -го порядку

$$y^p(x) = f(x, y, y', y'', y''', \dots, y^{p-1})$$

за допомогою заміни і підстановок можна звести до еквівалентної системи p рівнянь першого порядку

$$y'_k(x) = y_{k+1}(x), \quad 0 < k < p - 2$$

$$y'_{p-1}(x) = f(x, y_0, y_1, \dots, y_{p-1}), \quad y_0(x) = y(x)$$

Таким чином, довільну систему ДР будь-якого порядку можна звести до деякої еквівалентної системи рівнянь першого порядку. Тому надалі будемо розглядати ДР першого порядку. ДР в загальному випадку розділяються на звичайні ДР і ДР в частинних похідних, ми будемо розглядати перші. У свою чергу звичайні ДР поділяються на крайові задачі, задачі на власні значення і задачі Коші з початковими умовами.

Ми будемо вирішувати задачу Коші в наступній постановці

$$\frac{dy}{dx} = f(x, y) : \quad y = y_0, \quad x = x_0 \quad (1)$$

Методи розв'язання таких ДР розділяються на *точні* і *чисельні*/

Точні методи дозволяють виразити розв'язання ДР через елементарні функції тобто в аналітичній формі. Чисельні методи це алгоритми обчислення наближених значень шуканого розв'язання на деякій сітці значень аргументу x . Розв'язання при цьому виходить у вигляді таблиці. Взагалі розв'язанням ДР є кінцеве рівняння, що може бути виражене в аналітичній (у вигляді формули) у табличній чи графічній формах. Розв'язання ДР може бути загальним і частковим. Загальне розв'язання являє собою сімейство непрямих $y=y(x, C)$ відрізняються одним коефіцієнтом C , часткове розв'язання виходить із загального шляхом обліку початкових умов.

Методи розв'язання ДР поділяють на однокрокові багатокрокові. В однокрокових методах використовується інформація про саму інтегральну криву в попередній точці. До таких методів відносяться метод Ейлера, метод Рунге-Куты, розв'язання рівнянь за допомогою рядів Тейлора. Недолік їх у тому що важко оцінювати помилку, що допускається. У багатокрокових методах наступну точку інтегральної кривої можна одержати, не роблячи повторних обчислень функції, як в однокрокових методах, у них використовуються прийоми прогнозу і корекції. Труднощі реалізації таких методів полягають у необхідності одержання декількох початкових точок, у складності організації обчислювальної процедури, однак їхні переваги - у можливості оцінювання припущеної помилки.

Застосування ряду Тейлора.

Для визначення значення наступної точки інтегральної кривої (у h -околиці точки x), можна скористатися розкладанням функції в ряд Тейлора:

$$y(x + h) = y(x) + h \frac{\partial y}{\partial x} + \frac{h^2}{2!} \frac{\partial^2 y}{\partial x^2} + \frac{h^3}{3!} \frac{\partial^3 y}{\partial x^3} + \dots$$

Чим більше членів ряду ми візьмемо для обчислення, тим точніше буде знайдена наступна точка інтегральної кривої. З практичної точки зору труднощі застосування цього підходу полягає в необхідності обчислення похідних високих порядків, тому він не одержав поширення, хоча є основою для інших методів.

Метод Ейлера.

Необхідно розв'язати рівняння (1) на інтервалі $[x_0, b]$. Для цього задаймося постійним кроком h і обчислюємо значення $y_1, y_2, \dots, y_n$ у послідовності точок $x_i = x_0 + ih, i = 1, 2, \dots, n$. У початковий момент відома тільки одна точка, через яку проходить крива, а саме x_0, y_0 . Диференціальне рівняння задає нахил кривої в будь-якій точці як функцію від x та y . Тому, починаючи з початкової точки, обчислюємо нахил кривої і просуваємося на деяку малу відстань уздовж одержаної дотичної

$$x_i = x_{i-1} + h, \quad y_i = y_{i-1} + hf(x_{i-1}, y_{i-1})$$

Повторюючи процедуру одержуємо послідовність точок які є наближенням до шуканої функції. Якщо відкласти отримані результати на площині у вигляді відрізків прямих то одержимо ламану Ейлера (рис) що є графічним розв'язанням вихідного ДР

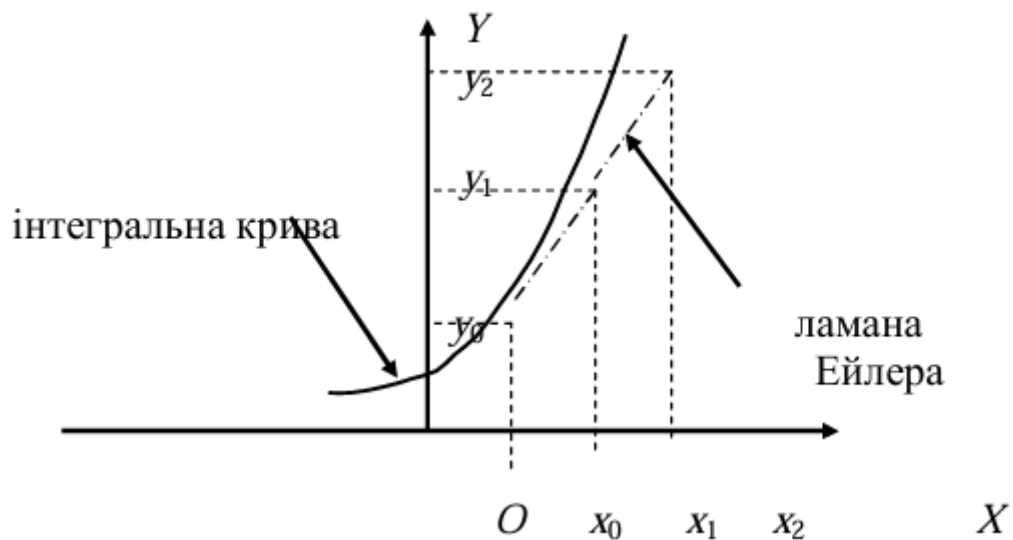


Схема методу Ейлера випливає з розкладання функції в ряд Тейлора, при використанні тільки першого його члена що означає рух не по інтегральній кривій а по дотичній до неї. Однак такий метод на практиці виявляється досить грубим, похибка на кожному кроці досить велика. При досить малому кроці $h \ll 1$ похибка методу близька до h^2 . Недолік методу Ейлера полягає у великій похибці, якщо ж для її зменшення вибирається менший крок h , то відбувається різке уповільнення обчислень/ Крім того він дуже часто виявляється нестійким - мала помилка збільшується із зростанням x .

Приклад: Використовуючи метод Ейлера чисельно розв'язати диференціальне рівняння $y' = \frac{xy}{2}$ на відрізку $[0; 0,9]$ із кроком $h=0,1$ початковою умовою $y(0)=1$

Результати обчислень значень x_i і y_i , приведені в таблиці, є розв'язанням вихідного диференціального рівняння, відклавши їх на площині, одержимо ломану Ейлера.

i	x_i	y_i	$\Delta y_i = hf(x_{i-1}, y_{i-1})$
0	0	1	0
1	0,1	1	0,005
2	0,2	1,005	0,0101
3	0,3	1,0151	0,0152
4	0,4	1,0303	0,0207
5	0,5	1,0509	0,263

6	0,6	1,0772	0,323
7	0,7	1,1095	0,0388
8	0,8	1,1483	0,0459
9	0,9	1,1942	0,0537

Метод Рунге-Кутта

В основі одержання обчислювальних схем методу Рунге-Кутта є розкладання функції $y(x)$ у ряд Тейлора з наступним перетворенням відрізка ряду до вигляду що не містить похідних. Обчислювальна схема буде схемою m -го порядку якщо вона отримана з ряду Тейлора, що містить похідні функції $y(x)$ до m -го порядку включно. Тому метод Ейлера є елементарною схемою методу Рунге-Кутта, однак у зв'язку з більш ранньою появою методу Ейлера за ним збережена колишня назва.

Таким чином на кожному кроці алгоритму Рунге-Кутта починаючи з початкової точки x у обчислюються

$$x_{i+1} = x_i + h, \quad y_{i+1} = y_i + \lambda, \quad \text{де } \lambda = \frac{1}{6} \cdot (k_1 + 2k_2 + 2k_3 + k_4)$$

а коефіцієнти k обчислюються по формулах (приклад схеми 4-го порядку)

$$k_1 = hf(x_i, y_i)$$

$$k_2 = hf(x_i + \frac{h}{2}, y_i + k_1)$$

$$k_3 = hf(x_i + \frac{h}{2}, y_i + k_2)$$

$$k_4 = hf(x_i + h, y_i + k_3)$$

При використанні такого методу четвертого порядку на кожному кроці алгоритму функцію приходится обчислювати 4 рази, що значно більше ніж у методі Ейлера. Але таке збільшення числа обчислень з лишком окупається більшою точністю розв'язання. Перевагами методу Рунге-Кутта є простота обчислень (у ньому немає необхідності обчислення похідних), досить висока точність і можливість вести обчислення зі змінним кроком (там де функція швидко міняється, неважко зменшити крок і збільшити його в зворотному випадку). Перевагою є також використання інформації тільки про чергову крапку (не використовуються дані про раніше знайдені точки). До недоліків методу необхідно віднести те, що на кожному кроці приходится обчислювати функцію $f(x,y)$ у декількох точках, а також труднощі при одержанні оцінки

помилки обмеження . Похибка викладеного методу приблизно дорівнює h^5 . При виборі кроку h у методі Рунге-Кутта можна як крок використовувати критерій

$$q = \left| \frac{k_2 - k_3}{k_1 - k_2} \right|.$$

Якщо $q < 10^{-2}$, то крок обрано вірно, в іншому випадку крок потрібно зменшити.

Запитання для самоконтролю.

1. У чому ідея методу Ейлера?
2. У чому ідея методу Рунге-Кутта?
3. Що спільного у методах Ейлера та Рунге-Кутта?
4. Чим відрізняються методи Ейлера та Рунге-Кутта?

ЛЕКЦІЯ 13. Апроксимація функцій. Метод найменших квадратів.

Задача наближення функцій виникає при розв'язанні багатьох практичних і теоретичних задач, а інколи і як самостійна. Так, наближення функцій є важливим допоміжним апаратом при розв'язанні інших задач чисельного аналізу: чисельного інтегрування і диференціювання, розв'язання диференціальних рівнянь, розв'язання систем нелінійних рівнянь, задач оптимізації та ін.

Постановка задачі

Постановка 1. Проста задача, що приводить до наближення функцій, полягає в наступному. Нехай є деяка функція $f(x)$, $f: R1 \rightarrow R1$, про яку відомо, що в n точках $x_1, x_2, \dots, x_n$ вона приймає, відповідно, значення $y_1, y_2, \dots, y_n$, тобто $y_i = f(x_i)$. Потрібно відновити її значення при інших значеннях $x \in (x_1, x_n)$. Функція $y = f(x)$ може бути як невідомою, тобто її значення тільки вимірюються (так званий "чорний ящик"), так і просто дуже складною для обчислень. Наприклад, вона може використовуватися в яких-небудь фізико-технічних або суто математичних розрахунках, де її доводиться багато разів обчислювати. Слід зазначити, що якщо функція $y = f(x)$ невідома, то набір точок (x_i, y_i) , називають табличним заданими функціями, оскільки ці значення подають у вигляді таблиці.

У цих випадках функцію $f(x)$ стараються замінити на простішу функцію $g(x; a)$, близьку до $f(x)$, тобто:

$$f(x) \approx g(x; a), \quad (1)$$

де $a \in R^k$, $a = (a_1, a_2, \dots, a_k)^T$ – деякі параметри функції g , вид якої відомий. Процес наближення однієї функції іншою ще називають *апроксимацією*, а функцію g при цьому називають *апроксимуючою* для f .

Варто зазначити, що параметри $a \in R^k$ є інструментом для наближення функції g до функції f .

Якщо параметри $a_1, a_2, \dots, a_n$ визначаються з умови збігу значень функції $f(x)$ і апроксимуючої функції в точках $x_1, x_2, \dots, x_n$, тобто:

$$g(x_i; a) = f(x_i), \quad \forall i = \overline{1, n}$$

то такий спосіб наближення називають *інтерполяцією* або *інтерполюванням*.

Постановка 2. У задачах планування експериментів виникає така проблема. Відомо вид гарного наближення функції, наприклад функція $f(x)$ добре наближається поліномом 2-го ступеня. В той же час вимірювані значення функції y_i мають великі помилки. Потрібно отримати найкраще в певному значенні наближення при мінімальній кількості вимірювань. Така задача виникає при плануванні експериментів в біології, хімії, фізиці, географії, військовій справі та ін.

Постановка 3. Задача наближення функцій розв'язується і при складанні стандартних процедур обчислення елементарних і спеціальних функцій, наприклад $\cos(x)$, $\sin(x)$ й інших.

$$\text{Так:} \quad e^x \approx \sum_{k=0}^n \frac{x^k}{k!}, \quad \sin(x) \approx x + \sum_{k=1}^n (-1)^k \frac{x^{2k-1}}{(2k-1)!}$$

де n підбирається так, щоб забезпечити достатню точність обчислень.

Метод найменших квадратів для апроксимації функцій

Найбільш відомим і ефективним із методів розв'язання задачі апроксимації функцій (1) є метод найменших квадратів (МНК). МНК призначений безпосередньо для знаходження параметрів a апроксимуючої функції $g(x; a)$, вид якої вже обраний. Суть метода полягає в такому.

Вводиться функція від a виду:

$$\Phi(a) \equiv \sum_{i=1}^n ((y_i - g(x_i, a))^2 \quad (2)$$

яку можна розглядати як міру відхилення функції $g(x; a)$ (по аргументу x) від функції $f(x)$ по сукупності точок $x_1, x_2, \dots, x_n$. Тоді параметри a функції

$g(x;a)$ можна визначити з умови найменшого відхилення $g(x;a)$ від $f(x)$, тобто параметри a знаходяться як точка, в якій функція $\Phi(a)$ досягає по $a \in R^k$ мінімального значення (точка мінімуму). Нехай a^* - шукані параметри. При цьому величини

$$r_i \equiv y_i - g(x_i, a^*), \quad i = \overline{1, n}$$

називаються залишковими відхиленнями (або залишками) і використовують-ся для аналізу отриманого розв'язку.

Вид графіка залишків дозволяє оцінити, наскільки правильно спочатку був вибраний вид апроксимуючої функції $g(x;a)$. Якщо в графіку залишків спостерігається деяка функціональна закономірність (наприклад, парабола), то це означає, що цю закономірність потрібно перенести в апроксимуючу функцію. Нормальною є ситуація, коли графік залишків має вид ламаної синусоїди, іншими словами, має випадковий характер.

Слід зазначити, що метод найменших квадратів зазвичай застосовують для розв'язання задачі наближення функцій у постановці 2.

Приклад. Використовуючи метод найменших квадратів, підібрати найкращу апроксимацію поліномом (визначити його степінь) для функції, заданої таблицею:

x	-1,5	-1	-0,5	0	0,5	1	1,5	2	2,5
y	9,8	6,4	5,2	1,7	2,9	5,6	7,2	16,5	27,5

Розв'яжемо за допомогою Excel. Побудуємо точкову діаграму та добавимо лінію тренду, яка будується за методом найменших квадратів. Визначимо для тренду вид апроксимуючого поліному та коефіцієнт детермінації. Коефіцієнт детермінації показує ступінь відповідності апроксимуючого поліному таблично заданій функції. На рисунках 1,2,3 показано апроксимацію функції поліномами першого, другого та третього ступені. За значенням коефіцієнта детермінації можна зробити висновок, яка апроксимація яким поліномом є найкращою (таблиця).

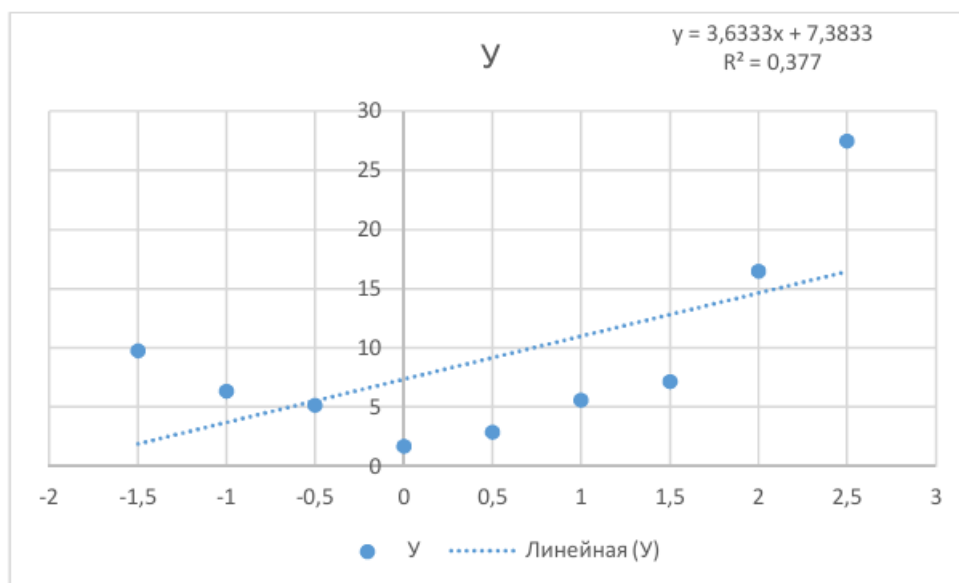


Рисунок 1- Апроксимація функції лінійним поліномом

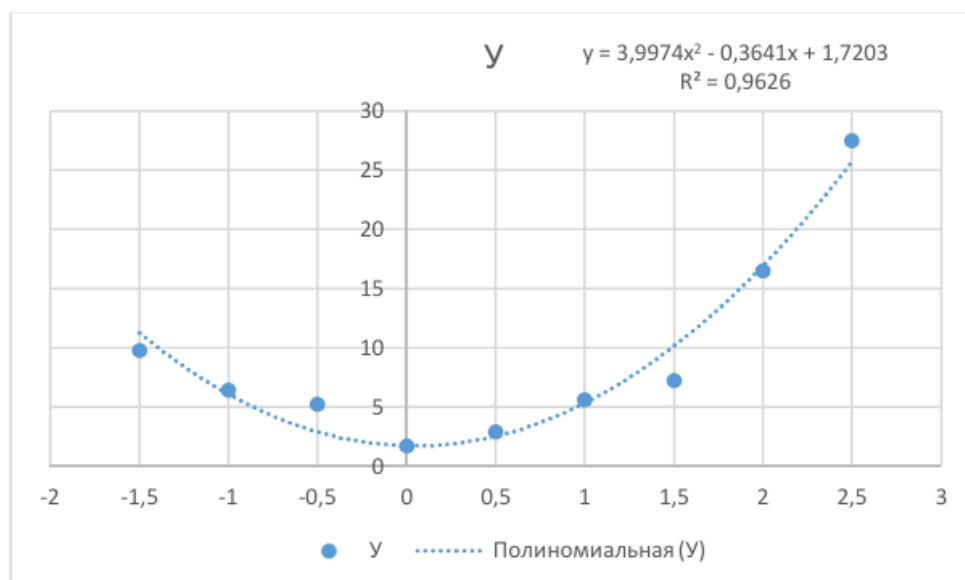


Рисунок 2- Апроксимація функції поліномом другого ступеня

№	Апроксимуючий поліном	Вид поліному	Коефіцієнт детермінації
1	Лінійний	$y = 3,6333 \cdot x + 7,3833$	$R^2 = 0,377$
2	Квадратичний	$y = 3,9974x^2 - 0,3641x + 1,7203$	$R^2 = 0,9626$
3	Кубічний	$y = 0,7811x^3 + 2,8257x^2 - 2,08$	$R^2 = 0,9884$

Таким чином, найкращім апроксимуючим поліномом є поліном третього ступеня, а саме: $y = 0,7811x^3 + 2,8257x^2 - 2,0826x + 2,7749$

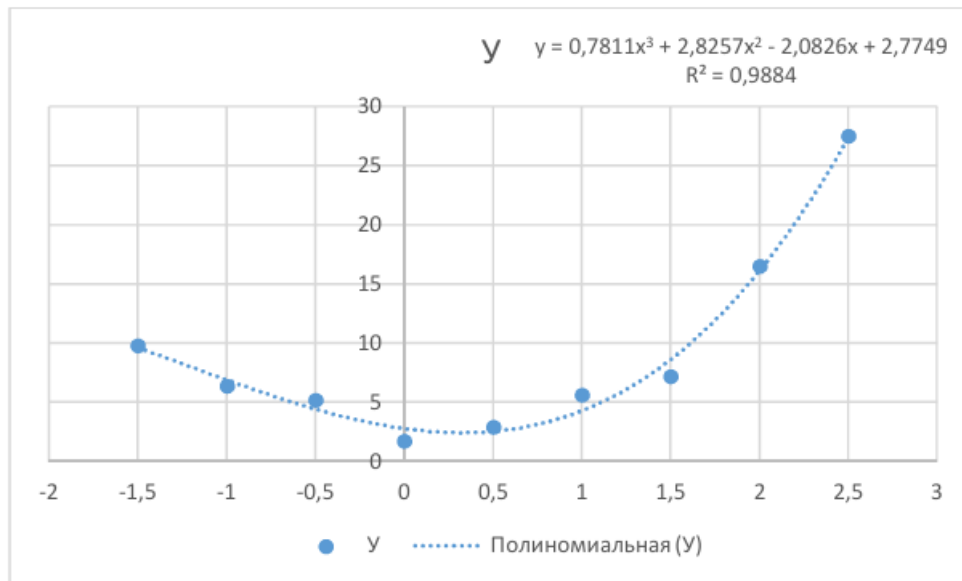


Рисунок 3- Апроксимація функції кубічним поліномом

У випадках, коли функція $g(x;a)$ є лінійною по параметрам a , точку мінімуму функції $\Phi(a)$ можна знайти аналітично з необхідної умови мінімуму 1-го порядку $\Phi'(a) = 0$, тобто:

$$\frac{\partial \Phi}{\partial a_i} = 0, \quad \forall j = \overline{1, k} \quad (3)$$

Коли функція $g(x;a)$ лінійна по параметрам, тобто має вигляд:

$$g(x, a) = \sum_{m=1}^k a_m \psi_m(x) \quad (4)$$

де $\psi_m(x)$, $m = \overline{1, k}$ – деякі відомі функції від x . Тоді з (2) - (4) отримуємо $\forall j = \overline{1, k}$

$$\frac{\partial \Phi}{\partial a_j} = -2 \sum_{i=1}^n (y_i - g(x_i, a)) \psi_j(x_i) = -2 \sum_{i=1}^n (y_i - \sum_{m=1}^k a_m \psi_m(x_i)) \psi_j(x_i) = 0$$

Звідки для $\forall j = \overline{1, k}$

$$\sum_{m=1}^k a_m \left(\sum_{i=1}^n \psi_m(x_i) \psi_j(x_i) \right) = \sum_{i=1}^n y_i \psi_j(x_i) \quad (5)$$

Таким чином, параметри a можуть бути знайдені як розв'язок системи лінійних рівнянь (5).

Варто розглянути ще випадок, коли апроксимуюча функція $g(x; a)$ є поліномом заданого $(k-1)$ -го степеню. Тоді вона в загальному випадку може бути записана у вигляді:

$$g(x, a) = b_0 + b_1 x + b_2 x^2 + \dots + b_{k-1} x^{k-1} = \sum_{i=0}^{k-1} b_i x^i$$

де $a \in R^k$, $a_1 = b_0$, $a_2 = b_1$, ..., $a_k = b_{k-1}$ - коефіцієнти полінома.

При цьому функція $g(x; a)$ є лінійною по параметрам a і має вигляд (4), де $\psi_1(x) = 1$, $\psi_2(x) = x$, ..., $\psi_k(x) = x^{k-1}$, тобто $\psi_m(x) = x^{m-1}$, $m = \overline{1, k}$.

Тоді параметри a можна знайти, розв'язавши систему лінійних рівнянь (5).

Інтерполяція поліномами має такі переваги, як простота обчислень їхніх значень, диференціювання та інтегрування.

Запитання для самоконтролю.

1. Назвіть постановки задач які приводять до апроксимації функцій.
2. У чому полягає суть методу найменших квадратів?
3. Назвіть способи знаходження значення параметру a в залежності від вигляду функції $g(x; a)$.

ЛЕКЦІЯ 14. Інтерполяція функції. Види інтерполяцій функції.

У випадках, коли вид функції $g(x, a)$ наперед невідомий, а значення y_i , $i = \overline{1, n}$, містять малу погрішність, задачу наближення таблично заданої функції $f(x)$ часто розв'язують як задачу інтерполяції.

Кусочно-лінійна інтерполяція. Такий вид інтерполяції полягає в тому, що на кожному окремому інтервалі (x_i, x_{i+1}) функція $f(x)$ наближається лінійною функцією

$$g_i(x, a_i, b_i) = a_i x + b_i, \quad i = \overline{1, n-1}$$

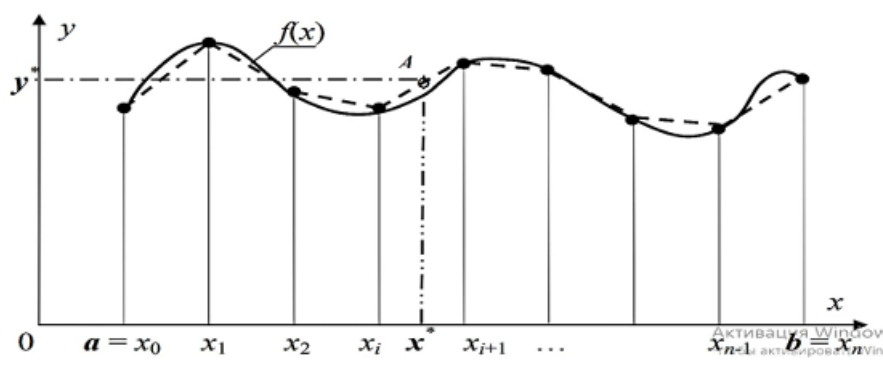


Рисунок 1 - Графічна інтерпретація кусочно-лінійної інтерполяції

Оскільки значення функцій $f(x)$ і $g_i(x, a_i, b_i)$ в точках x_i і x_{i+1} повинні співпадати, то:

$$\begin{aligned} g_i(x_i, a_i, b_i) &= a_i x_i + b_i = y_i \\ g_i(x_{i+1}, a_i, b_i) &= a_i x_{i+1} + b_i = y_{i+1} \end{aligned}$$

Тоді, розв'язуючи отриману систему рівнянь відносно a_i, b_i отримуємо:

$$\begin{aligned} a_i &= \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \\ b_i &= y_i - a_i x_i \end{aligned}$$

Звідси випливає, що параметри a_i, b_i функції $g_i(x)$ повністю визначаються значеннями координат точок (x_i, y_i) і (x_{i+1}, y_{i+1}) . У цьому випадку лінійна функція може бути записана в зручнішому вигляді:

$$g_i(x) = \frac{(x - x_{i+1})}{(x_i - x_{i+1})} y_i + \frac{(x - x_i)}{(x_{i+1} - x_i)} y_{i+1} \quad (6)$$

Таким чином, у ході використання кусочно-лінійної інтерполяції для таблично заданої функції, для обчислення наближеного значення функції $f(x)$ у деякій точці $\bar{x} \in (x_1, x_n)$ діють таким чином. Спочатку знаходять інтервал (x_i, x_{i+1}) якому точка $\bar{x}$ належить, а потім обчислюють наближене значення функції $f(x)$ у точці $\bar{x}$ як $g_i(x)$, де функція $g_i(x)$ визначена в (6).

При цьому функція $g(x)$ по суті є наближеним розв'язком задачі наближення функції $f(x)$, тобто істинного розв'язку. Тому виникає питання

про точність розв'язку цієї задачі, тобто про погрішність отриманого розв'язку.

Оцінка погрішності. Нехай функція $f(x)$ двічі неперервно-диференційована на відрізку (x_1, x_n) , тоді для всіх $x \in (x_1, x_n)$:

$$|f(x) - g(x)| \leq Mh^2$$

$$\text{де } M = |f''(\xi)|, \quad h = |x_{i+1} - x_i|$$

Таким чином, погрішність кусочно-лінійної інтерполяції має порядок $O(h)$. Кусочно-лінійна інтерполяція застосовується тоді, коли не вимагається великої точності наближення. Наприклад, при побудові графіків по точках тощо.

Кусочно-квадратична інтерполяція. Подібно до кусочно-лінійної інтерполяції можна використовувати для наближення на кожному окремому інтервалі (x_i, x_{i+1}) і поліноми вищого порядку, наприклад квадратичну

функцію $g(x, a, b, c) = ax^2 + bx + c$. У цьому випадку для знаходження коефіцієнтів a, b, c слід використовувати додаткову точку x_{i+2} (три невідомих, тому необхідно і три рівняння для їх визначення) (рис. 2).

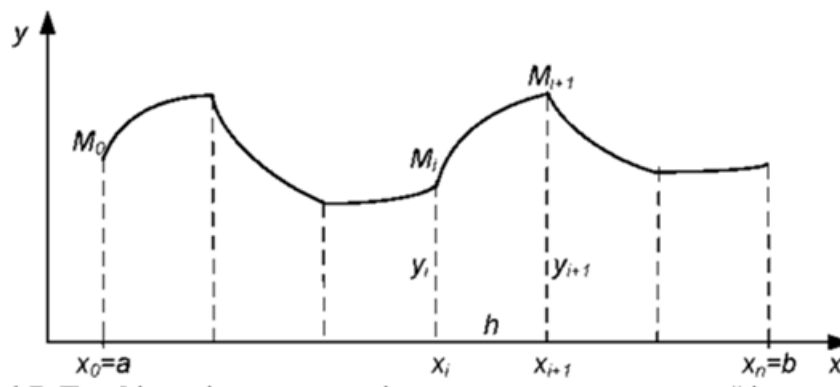


Рисунок 2- Графічна інтерпретація кусочно-квадратичної інтерполяції

Квадратична функція, що проходить через 3 точки (x_i, y_i) , (x_{i+1}, y_{i+1}) , (x_{i+2}, y_{i+2}) , може бути записана у вигляді:

$$g_i(x) = \frac{(x-x_{i+1})(x-x_{i+2})}{(x_i-x_{i+1})(x_i-x_{i+2})}y_i + \frac{(x-x_i)(x-x_{i+2})}{(x_{i+1}-x_i)(x_{i+1}-x_{i+2})}y_{i+1} + \\ + \frac{(x-x_i)(x-x_{i+1})}{(x_{i+2}-x_i)(x_{i+2}-x_{i+1})}y_{i+2}$$

Оцінка погрішності. Нехай функція $f(x)$ тричі неперервно-диференційована на відрізку (x_1, x_n) , поді для всіх $x \in (x_1, x_n)$:

$$|f(x) - g(x)| \leq Mh^3$$

$$\text{де } M = |f^{(3)}(\xi)|, \quad h = |x_{i+1} - x_i|$$

Таким чином, погрішність кусочно-квадратичної інтерполяції має порядок $O(h^3)$.

Інтерполяційний поліном Лагранжа. Якщо відомі значення y_i функції $f(x)$ в n точках $x_i, i = \overline{1, n}$, то поліномом мінімального степеня, що інтерполює функцію $f(x)$, матиме степінь $n-1$. Якщо він записаний у вигляді:

$$L_{n-1}(x) = \sum_{i=1}^n \left[\prod_{j=1, j \neq i}^n \frac{x-x_j}{x_i-x_j} \right] y_i \quad (7)$$

то його називають інтерполяційним багаточленом Лагранжа.

Погрішність наближення інтерпольованої функції багаточленом Лагранжа сильно залежить від розкиду точок x_i , тобто чим менше інтервал (x_1, x_n) , тим точніше $L_{n-1}(x)$ наближає значення функції $f(x)$ в точках $x \in (x_1, x_n)$.

Оцінка погрішності. Нехай функція $f(x)$ n разів неперервно-диференційована на відрізку (x_1, x_n) . Тоді для всіх $x \in (x_1, x_n)$:

$$|f(x) - L_{n-1}(x)| \leq \frac{M_n h^n}{n!}$$

$$\text{де } M_n = |f^{(n)}(\xi)|, \quad h = |x_{i+1} - x_i|$$

Таким чином, погрішність інтерполяції за допомогою багаточлена Лагранжа має порядок $O(h^n)$.

Однак між вузлами інтерполяції x_i поліном Лагранжа, як правило, нестійкий до погрішностей обчислень, причому нестійкість зростає зі збільшенням n . Крім того, навіть невеликі погрішності значень y_i можуть

сильно змінити поведінку полінома між вузлами. Зазначені ефекти нестійкості виявляються вже при $n > 15$, а в практичних задачах, коли n близьке до 100, нестійкість полінома Лагранжа дуже велика.

Для обчислення полінома (многочлена) Лагранжа не треба попереднього визначення коефіцієнтів полінома шляхом розв'язування системи рівнянь, проте для кожного значення x^* поліном доводиться перераховувати знову. Тобто практичне застосування полінома Лагранжа виправдано лише у випадку, якщо інтерполяційна функція обчислюється у порівняно невеликій кількості значень x .

Інтерполяційний поліном Ньютона. На відміну від (7), поліном мінімального степеня, що інтерполіює функцію $f(x)$ в n точках x_i , може бути записаний в іншому вигляді. Вводяться такі позначення:

розділені різниці 1-го порядку:

$$f(x_1, x_2) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}, \dots, f(x_{n-1}, x_n) = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

розділені різниці 2-го порядку:

$$f(x_1, x_2, x_3) = \frac{f(x_2, x_3) - f(x_1, x_2)}{x_3 - x_1}, \dots$$

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

розділена різниця (n-1)-го порядку:

$$f(x_1, x_2, \dots, x_n) = \frac{f(x_1, x_2, \dots, x_n) - f(x_1, x_2, \dots, x_{n-1})}{x_n - x_1}$$

Тоді розглядається поліном $(n - 1)$ -го степеня виду

$$H_{n-1}(x) = f(x_1) + (x - x_1)f(x_1, x_2) + (x - x_1)(x - x_2)f(x_1, x_2, x_3) + \dots +$$

$$+ (x - x_1)(x - x_2) \dots (x - x_{n-1})f(x_1, x_2, \dots, x_n) \quad (8)$$

Поліном $H_{n-1}(x)$ є інтерполяційним для функції $f(x)$. Його називають інтерполяційним поліномом Ньютона. Інтерполяційний поліном Ньютона виявляється дуже корисним при розв'язанні інших задач чисельного аналізу.

Сплайн-інтерполяція. Інтерполяція за допомогою поліномів Лагранжа або Ньютона з використанням великої кількості вузлів часто призводить до поганого наближення, що пояснюється накопиченням великої погрішності в процесі обчислень. З іншого боку, кусочно-лінійна і кусочно-квадратична

інтерполяція не дозволяють добитися гарного наближення зважаючи на їх теоретичну неточність. Одним із способів добитися гарного наближення в практичних задачах є інтерполяція за допомогою сплайн- функцій.

Визначення. Сплайном (з англ. spline - рейка, лінійка) називається кусочно-поліноміальна функція, визначена на відрізку (x_1, x_n) , і яка має на цьому відрізку не менше двох неперервних похідних. Сплайн буде позначено через $S(x)$. Інтерполяція за допомогою сплайнів називається *сплайн-інтерполяцією*.

Розглянемо кубічну сплайн-інтерполяцію, яка найчастіше застосовується на практиці, тобто на кожному інтервалі (x_i, x_{i+1}) , функція $S(x)$ задається поліномом третьої степені:

$$S_i(x) = y_i + c_{1i}(x - x_i) + c_{2i}(x - x_i)^2 + c_{3i}(x - x_i)^3 \quad (9)$$

Слід зазначити, що у випадку застосування кубічного сплайна він повинен мати дві неперервні похідні на відрізку (x_1, x_n) .

Для того щоб побудувати кубічний сплайн $S(x)$, необхідно знайти всі коефіцієнти c_{1i}, c_{2i}, c_{3i} ($i = \overline{1, n-1}$) поліномів $S_i(x)$. Кількість цих невідомих коефіцієнтів $3(n-1)$.

Коефіцієнти c_{1i}, c_{2i}, c_{3i} ($i = \overline{1, n-1}$) підбираються так, щоб на межах інтервалів (x_i, x_{i+1}) забезпечити неперервність, як функції $S(x)$, так і її першої $S'(x)$ і другої $S''(x)$ похідних. Згідно з (9) для всіх $i = \overline{1, n-1}$

$$S'_i(x) = c_{1i} + 2c_{2i}(x - x_i) + 3c_{3i}(x - x_i)^2$$

$$S''_i(x) = 2c_{2i} + 6c_{3i}(x - x_i)$$

Таким чином, повинні виконуватися умови:

$$\begin{aligned} S_i(x_{i+1}) &= S_{i+1}(x_{i+1}) = y_{i+1}, \quad i = \overline{1, n-2}; \quad S_{n-1}(x_n) = y_n \\ S'_i(x_{i+1}) &= S'_{i+1}(x_{i+1}) = c_{1,i+1}, \quad i = \overline{1, n-2} \\ S''_i(x_{i+1}) &= S''_{i+1}(x_{i+1}) = c_{2,i+1}, \quad i = \overline{1, n-2} \end{aligned} \quad (10)$$

Можна переписати (10) в більш наочному вигляді:

$$\{y_i + c_{1i}(x_{i+1} - x_1) + c_{2i}(x_{i+1} - x_i)^2 + c_{3i}(x_{i+1} - x_i)^3 = y_{i+1}, \quad i = \overline{1, n-1}\} \quad (11)$$

Ця система (11) визначає $3(n-1) - 2$ лінійних рівняння для пошуку $3(n-1)$ невідомих c_{1i}, c_{2i}, c_{3i} ($i = \overline{1, n-1}$). Тому необхідно довизначити ще два рівняння. Для цього задають так звані граничні умови - значення першої або другої похідної функції $f(x)$ на межах інтервалу (x_1, x_n) . Якщо значення однієї з похідних на межах відомі, то задавши їх, можна отримати дуже точну інтерполяційну схему. Якщо ж ці значення невідомі, то можна задати другу похідну на межах рівною нулю і отримати також достатньо гарні результати. Необхідно розглянути простий випадок, коли:

$$f''(x_1) = 0, \quad f''(x_n) = 0$$

тобто

$$\{S_1''(x_1) = 2c_{21} S_{n-1}''(x_n) = 2c_{2,n-1} + 6c_{3,n-1}(x_n - x_{n-1}) = 0 \quad (12)$$

Таким чином, для обчислення коефіцієнтів c_{1i}, c_{2i}, c_{3i} ($i = \overline{1, n-1}$), необхідно розв'язати систему лінійних алгебраїчних рівнянь (11), (12).

Оцінка погрішності. Нехай функція $f(x)$ 4 рази неперервно-диференційована на відрізьку (x_1, x_n) . Тоді для всіх $x \in (x_1, x_n)$

$$|f(x) - S(x)| \leq \frac{M_4 h^4}{8}$$

$$\text{де } M_4 = |f^{(4)}(\xi)|, \quad h = |x_{i+1} - x_i|$$

Таким чином, погрішність інтерполяції за допомогою кубічного сплайна має порядок $O(h^4)$. Перевагою кубічних сплайнів перед іншими наведеними способами інтерполяції є: по-перше, їх стійкість до процесу обчислень, і, по-друге, достатньо висока точність. При використуванні сплайн-інтерполяції для таблично заданої функції діють таким чином.

Спочатку один раз визначають всі коефіцієнти c_{1i}, c_{2i}, c_{3i} ($i = \overline{1, n-1}$), а потім вже для обчислення наближеного значення функції $f(x)$ у деякій точці

$\bar{x} \in (x_1, x_n)$ спочатку знаходять інтервал (x_i, x_{i+1}) , якому точка $\bar{x}$ належить, а потім обчислюють наближене значення функції $f(x)$ в точці $\bar{x}$ як $Si(x)$, де функція визначена в (9).

Запитання для самоконтролю.

1. У чому полягає лінійна інтерполяція функцій? Коли її слід застосовувати?
2. У чому полягає квадратична інтерполяція функцій? Коли її слід застосовувати?
3. Що називається інтерполяційним поліномом Лагранжа? Коли його слід застосовувати?
4. Що називають сплайном?
5. У чому полягає лінійна сплайн-інтерполяція функцій? Коли її слід застосовувати?

ЛЕКЦІЯ 15. Лінійні емпіричні формули. Лінеаризація емпіричної формули.

Лінійні емпіричні формули. Припустимо, що емпірична (теоретична) формула має вигляд $y = \varphi(x) = ax + b$. Тоді формула сумарного відхилення представляється так

$$S(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2 \rightarrow \min$$

Дорівнюючи до нуля частинні похідні функції S по a і b , одержимо систему рівнянь

$$\left\{ \frac{\partial S}{\partial a} = \sum_{i=1}^n 2(y_i - ax_i - b)(-x_i) = 0 \quad \frac{\partial S}{\partial b} = \sum_{i=1}^n 2(y_i - ax_i - b)(-1) = 0 \right.$$

або

$$\left\{ a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \quad a \sum_{i=1}^n x_i + bn = \sum_{i=1}^n y_i \right.$$

В остаточному вигляді невідомі параметри знаходять з виразів

$$\left\{ a = \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i - n \sum_{i=1}^n x_i y_i}{\left[\sum_{i=1}^n x_i \right]^2 - n \sum_{i=1}^n x_i^2} \right. \quad b = \frac{1}{n \left(\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i \right)}$$

Приклад апроксимації лінійної емпіричної формули. Нехай для $n=5$ дослідів отримана наступна таблиця

x	2	4	6	8	10
y	5,5	6,3	7,2	8	8,6

Припустимо, що емпірична формула має лінійний вигляд, і скористаємося наведеними вище формулами, тоді

$$\sum_{i=1}^n x_i = 30; \quad \sum_{i=1}^n y_i = 35,6; \quad \sum_{i=1}^n x_i y_i = 229,4; \quad \sum_{i=1}^n x_i^2 = 220$$

Знаходимо невідомі параметри:

$$a = \frac{(30 \cdot 35,6 - 5 \cdot 229,4)}{30^2 - 5 \cdot 220} = 0,395$$

$$b = \frac{1}{5 \cdot (35,6 - 0,395 \cdot 30)} = 4,75$$

Емпірична формула має вигляд $\varphi(x) = 0,395x + 4,75$

Прорахувавши значення функції у вихідних точках, одержимо:

x	2	4	6	8	10
y	5,5	6,3	7,2	8	8,6
$\varphi(x)$	5,54	6,33	7,12	7,91	8,7

Лінеаризація емпіричної формули. Подібно наведеному прикладу, можна знайти формули для обчислення параметрів будь-якої емпіричної формули, однак система може виявитися важкою для розв'язання. Один із способів відходу від виникаючих труднощів складається в «лінеаризації» емпіричної формули, тобто приведення її до лінійного вигляду. Розрахунки по лінійному випадку прості і розповсюджені і в деяких алгоритмічних мовах лінійна апроксимація представлена у вигляді оператора.

Приклад лінеаризації. Нехай $y = bx^a$. Прологарифмуємо цей вираз

$$\ln y = a \ln x + \ln b$$

Позначимо $V = \ln y$, $U = \ln x$, $\beta = \ln b$ тоді попередня формула матиме вигляд

$$V = aU + \beta$$

Тут V лінійно залежить від U . Заміняючи вихідні дані y_i і x_i на $V_i = \ln y_i$ і $U_i = \ln x_i$ і застосовуючи отримані вище формули для лінійного випадку знайдемо невідомі параметри вихідної системи (a і $b = e^\beta$).

Для найбільш розповсюджених формул, що застосовуються для апроксимації, розроблені спеціальні таблиці лінеаризації, що дозволяють відмовитися від громіздких обчислювань.

У загальному випадку нехай лінійна емпірична формула має вигляд $y' = a' \cdot x' + b'$. Тоді наведемо деякі перетворення, що зводять нелінійні формули до лінійних.

Формула	x'	y'	a	b
$a + bx$	x	y	a'	b'
$\frac{1}{(a+bx)}$	x	$\frac{1}{y}$	a'	b'
$a + \frac{b}{x}$	$\frac{1}{x}$	y	a'	b'
$\frac{x}{(a+bx)}$	x	$\frac{x}{y}$	a'	b'
$a \cdot b^x$	x	$\lg y$	$10a'$	$10b'$
$a \cdot e^{\frac{b}{x}}$	x	$\ln y$	$e^{a'}$	b'
$a + b \lg x$	$\lg x$	y	a'	b'
$a + bx^n$	x^n	y	a'	b'
$\frac{a}{(x+b)}$	x	$\frac{1}{y}$	$\frac{1}{b'}$	$\frac{a'}{b'}$

Користатися цією таблицею можна в такий спосіб: вибирається визначена емпірична формула, таблиця вихідних даних перераховується відповідно до 2 та 3 колонок, потім знаходяться параметри a' і b' лінійної формули і здійснюється зворотний перехід до 4 та 5 колонок.

Приклад апроксимації вихідних даних різними емпіричними формулами.

Нехай у результаті деякого дослідження отримані наступні значення експериментальних даних $n=10$

x_i	1	2	3	4	5	6	7	8	9	10
y_i	2,25	5,38	6,97	8,05	8,72	9,28	9,63	9,95	10,15	10,32

Застосуємо квадратичну емпіричну формулу $y = ax^2 + bx + c$. Після обчислення a, b, c одержимо $y = -0,118x^2 + 2,030x + 1,536$.

Для оцінки точності отриманої апроксимації можна застосувати середньоквадратичну похибку (чим менше ця величина тим точніше проведено апроксимацію)

$$\delta = \sqrt{\frac{\sum_{i=1}^n (y_i^0 - y_i)^2}{n}}$$

Для обраної квадратичної формули $\delta = 0,294$. Якщо для апроксимації в даному прикладі обрана формула $y = ae^{bx}$ то при лінеаризації даної формули одержимо $y = 4,207e^{0,11x}$ і похибка $\delta = 1,315$. Вибравши формулу типу $y = ax^b$ після лінеаризації та обчислень отримаємо $y = 3,531x^{0,516}$ при цьому $\delta = 0,681$. Звідси випливає що найкраще вихідні дані апроксимуються квадратичною емпіричною формулою.

Запитання для самоконтролю.

1. Наведіть приклад апроксимації лінійної емпіричної формули
2. Як обчислюється точність одержаної апроксимації?
3. Що називається відхиленням експериментальної точки?

Список літератури

1. Петренко В. Обчислювальні методи: базовий курс. Київ: КНУ, 2020. — 190 с.
2. Савчук А. Методи чисельного аналізу. Львів: ЛНУ, 2021. — 175 с.
3. Taylor, M. Introduction to Numerical Computation. Springer, 2021. — 250 с.
4. Nguyen, L. Numerical Techniques for Applied Sciences. Wiley, 2023. — 320 с.
5. Li, Q. Modern Numerical Methods and Algorithms. CRC Press, 2022. — 340 с.
6. Holbrook, J. Scientific Computing Foundations. Routledge, 2024. — 210 с.
7. Kaur, M. Numerical Modeling Toolkit. Elsevier, 2023. — 265 с.
8. Patel, D. College-Level Numerical Methods. Oxford University Press, 2025. — 230 с.