

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Київський національний університет будівництва і архітектури

О.І. Баліна, І.С. Безклубенко, Ю.І. Буценко

ЧИСЕЛЬНІ МЕТОДИ

Конспект лекцій

для студентів, які навчаються за спеціальністю
151 «Автоматизація та комп'ютерно-інтегровані технології»

Київ 2019

УДК 517
Б39

Рецензент О.О. Терентьев, д-р техн. наук, профессор

Затверджено на засіданні вченої ради факультету автоматизації і інформаційних технологій, протокол № 8 від 28 січня 2019 року.

В авторській редакції.

Баліна О.І.

Б39 Чисельні методи: конспект лекцій / О.І. Баліна, І.С. Безклубенко, Ю.І. Буценко. – Київ: КНУБА, 2019. – 68 с.

Розглянуто методи наближеного розв'язання лінійних і нелінійних рівнянь і їх систем, методи інтерполяції і екстраполяції, методи наближеного диференціювання та інтегрування як методи сучасної обчислювальної математики для зручного користування ЕОМ; методи оцінювання похибок, збіжності розв'язку у наближених обчисленнях. Містить теоретичний і довідковий матеріал, приклади розв'язання типових задач, контрольні запитання, список літератури.

Призначено для студентів, які навчаються за спеціальністю 151 «Автоматизація та комп'ютерно-інтегровані технології»

УДК 517

О.І. Баліна, І.С. Безклубенко, 2019

КНУБА, 2019

ЗМІСТ

Вступ.....	4
<i>Лекція 1. Обчислювальні процеси та похибки.....</i>	<i>5</i>
<i>Лекція 2. Наближене розв’язання алгебраїчних та трансцендентних рівнянь (метод хорд, метод дотичних).....</i>	<i>12</i>
<i>Лекція 3. Наближене розв’язання алгебраїчних та трансцендентних рівнянь (метод ітерацій, метод Лобачевського).....</i>	<i>16</i>
<i>Лекція 4. Задача інтерполяції. Інтерполяційні многочлени Ньютона та Лагранжа</i>	<i>22</i>
<i>Лекція 5. Локальна апроксимація. Сплайни. Рівномірні наближення...</i>	<i>29</i>
<i>Лекція 6. Прямі методи розв’язання систем лінійних алгебраїчних рівнянь.....</i>	<i>36</i>
<i>Лекція 7. Ітераційні методи розв’язання СЛАР</i>	<i>45</i>
<i>Лекція 8. Методи розв’язання систем лінійних алгебраїчних рівнянь із матрицями спеціального вигляду.....</i>	<i>49</i>
<i>Лекція 9. Часткова проблема власних значень та власних функцій..</i>	<i>59</i>
<i>Лекція 10. Повна проблема власних значень та власних функцій....</i>	<i>64</i>
Список літератури.....	68

ВСТУП

Багато технічних, економічних, інформаційних, фізичних задач потребують дослідження складних математичних моделей. Ці моделі зазвичай описуються нелінійними, зокрема диференційними та інтегральними рівняннями та їх системами. На практиці такі математичні задачі мають аналітичний розв'язок тільки в деяких випадках. Основним інструментом розв'язання широкого кола математичних задач є чисельні методи та наближені розв'язки.

Курс «Числові методи і моделювання на ЕОМ» призначений для ознайомлення майбутніх спеціалістів із застосуванням чисельних методів до наближених обчислень інтегралів, диференційних рівнянь, значень функцій тощо.

Для успішного оволодіння матеріалом студент повинен **вміти**:

- математично моделювати технологічні, технічні та соціально-економічні процеси в межах тих технологічних, технічних та соціально-економічних знань, здобутих у процесі вивчення відповідних природничих і спеціальних дисциплін;
- застосувати чисельні методи до наближених обчислень, дослідити збіжність отриманого результату і виявити його похибку, а також отримати результат з наперед заданою похибкою.

У результаті вивчення дисципліни студенти повинні **знати**:

- основні математичні поняття сучасної математичної символіки, елементи теорії множин і математичної логіки як основних можливостей мінімально-збиткового представлення математично формалізованих процесів;
- теорію функцій однієї та багатьох змінних, яка дає змогу якісно аналізувати дискретні і неперервні функціональні зв'язки, даючи їм геометричну й аналітичну інтерпретацію;
- теорію і практику нескінченних рядів як конструктивного апарата апроксимації неперервних й дискретних функціональних залежностей;
- методи наближеного розв'язку лінійних і нелінійних рівнянь і їх систем, методи інтерполяції й екстраполяції, методи наближеного диференціювання та інтегрування як методи сучасної обчислювальної математики для зручного пристосування ЕОМ;
- методи оцінки похибок: збіжності розв'язку у наближених обчисленнях;
- метод координат як загальний метод геометрії для дослідження плоских кривих першого і другого порядку, поверхонь першого та другого порядку.

Лекція 1. ОБЧИСЛЮВАЛЬНІ ПРОЦЕСИ ТА ПОХИБКИ

У лекції запроваджується поняття абсолютної похибки, методи визначення незсувної похибки, вимоги до обчислювальних алгоритмів.

Зазвичай вихідні дані для інженерних розрахунків, одержані шляхом вимірювань, містять похибки внаслідок обмеженої точності вимірювальної апаратури. Крім того, в результаті обчислювального процесу виникають неусувні похибки внаслідок дій з наближеними числами. Слід, у зв'язку з цим, вказати на такі принципові моменти, які свого часу визначив видатний фахівець з прикладної математики академік О. М. Крилов:

- 1) у прикладних задачах немає потреби використовувати лише абсолютні точні формули;
- 2) немає потреби добиватися ідеальної точності в обчисленнях за цими формулами;
- 3) можна використовувати напевно неточні формули або методи, контролюючи лише похибку, яка має не перевищувати максимально допустимого значення;
- 4) користувача обчислювальних методів цікавить не процес обчислення, а результат, який має бути отриманий з достатньою точністю за найменших затрат часу та праці.

Наведемо далі з основні поняття, з якими доводиться мати справу в реалізації чисельних методів.

Нехай:

a_0 – істинне значення деякої величини, яке найчастіше є невідомим;

a – число, яким в обчисленнях можна замінювати a_0 , наближене до значення a_0 :

$$a \approx a_0; \quad (1.1)$$

абсолютною похибкою наближеного числа a називається величина $\Delta = |a - a_0|$; відносною похибкою наближеного числа a – відношення його абсолютної похибки до абсолютної величини цього числа:

$$\delta = \frac{\Delta}{|a|}. \quad (1.2)$$

У конкретних задачах звичайно вдається обмежити абсолютну похибку деякою величиною Δa :

$$\Delta \leq \Delta a, a_0 \in [a - \Delta a; a + \Delta a]. \quad (1.3)$$

Наприклад, під час вимірювання звичайною лінійкою $\Delta a = 0,5 \text{ мм}$. Точні значення як Δ , так і δ є невідомими. На практиці δ також обмежують деякою граничною величиною δ_a :

$$\delta = \frac{\Delta}{|a|} \leq \delta_a; \quad \delta_a = \frac{\Delta a}{|a|}. \quad (1.4)$$

Величина Δ має розмірність вимірюваної величини, δ - величина безрозмірна, її часто наводять у відсотках (наприклад, вказуючи точність вимірювальних приладів).

Записуючи число у деякій системі числення, в переважній більшості випадків маємо справу, як впливає із структури множини дійсних чисел, з нескінченною кількістю значущих цифр (наприклад, числа $\sqrt{2}, e, \pi$). В наближених обчисленнях використовують лише числа, записані за допомогою скінченної кількості розрядів. Якщо число a , що містить n значущих цифр, потрібно замінити числом a_1 з меншою кількістю значущих цифр (округлити), то:

- 1) для того щоб округлити число до n значущих цифр, відкидають всі цифри, що стоять правіше від n -ї значущої цифри, або, якщо це потрібно для збереження розрядності, замінюють їх нулями;
- 2) якщо перша з відкинутих цифр менша за «5», то залишені зберігаються без змін;
- 3) якщо перша з відкинутих цифр більше, а ніж «5», то до останньої із залишених додають «1»;
- 4) якщо перша з відкинутих цифр рівна «5» і серед останніх відкинутих є ненульові, то до останньої залишеної додається «1»;
- 5) якщо перша відкинута цифра дорівнює 5, а всі решта – нулі, то залишені не змінюються.

Називатимемо цифру наближеного числа правильною, якщо абсолютна похибка не перевищує (у записі в десятковій системі числення) п'яти одиниць наступного за нею розряду, в інших випадках цифра називатиметься сумнівною. У математичних таблицях та в записі вихідних даних наближені числа записують так, щоб усі значущі цифри були правильними.

Це дає змогу не вказувати спеціально абсолютну похибку, а знаходити її з наведених міркувань. В той же час у наближених числах, які мають використовуватися в обчисленнях зазвичай залишають декілька сумнівних цифр. Наприклад, якщо число $a = 3,1472$ належить до вихідних

даних, то $\Delta a = 0,00005$, якщо число $a = 6,472$ має похибку $0,003$, то його записують у вигляді $6,472 \pm 0,003$. Додаючи похибку округлення та власну похибку числа, результат завжди округлюють у більший бік: наприклад, якщо число $a = 6,4738$ має абсолютну похибку $a = 0,0026$, округлюємо до трьох значущих цифр після коми, отже, маємо $6,474 \pm 0,003$.

Розглядаючи великі числа, застосовують «степеневий» запис. Нехай у числі 12732000 цифра «7» – правильна, а «3» – сумнівна, тоді число записують (не відкидаючи сумнівної цифри) у вигляді $1,273 \cdot 10^7$. Для збереження сумнівних цифр іноді доводиться залишати певну кількість нулів: $5 \pm 0,003$ записують у вигляді $5,000$ (з однією сумнівною цифрою). Зауважимо, що може бути використане й означення правильної цифри в так званому широкому сенсі: цифру a_k у числі, що містить n значущих цифр, вважають правильною, якщо

$$\Delta a \leq \omega \cdot 10^{n-k-1}, \quad (1.5)$$

де $5 \leq \omega \leq 10$ – деякий наперед визначений параметр.

Такий підхід дає змогу диференційовано підходити до наближених чисел, які з'являються в результаті операцій з високоточними початковими даними (ω обирають близьким до 5) і початковими даними низької точності (ω обирають близьким до 10). Для знаходження відносної похибки наближеного значення, яке має m правильних значущих цифр із загальної кількості n , можна скористатися такою оцінкою:

$$\begin{aligned} \delta_u = \frac{\Delta a}{|a|} &\leq \frac{5 \cdot 10^{n-m}}{|a_0 \cdot 10^n + \dots + a_{m-1} \cdot 10^{n-1+1\dots}|} \leq \frac{5 \cdot 10^{n-m}}{a_0 \cdot 10^n} = \\ &= \frac{1}{2a \cdot 10^{m-1}}. \end{aligned} \quad (1.6)$$

Розглянемо тепер методи визначення неусувної похибки під час обчислення. Нехай

y^0, x_i^0 – точні значення величин;

y, x_i – їх наближені значення;

$|\Delta y| = |y - y^0|, |\Delta x| = |x_i - x_i^0|$ – абсолютні похибки;

$\Delta y, \Delta x_i, \delta_y, \delta_x$ – відповідні абсолютні та відносні похибки.

Нехай стоїть задача обчислити функції $y = f(x_1, \dots, x_n)$. Розглянемо спочатку найпростіші варіанти задати таку функцію:

а) Нехай $y = \sum_{i=1}^n x_i$. Очевидно, що цьому випадку абсолютна похибка суми не перевищує суми абсолютних похибок (незалежних!) доданків.

$$\text{Відповідно } \Delta y = \sum_{i=1}^n \Delta x_i.$$

Гранична відносна похибка суми доданків, що мають однакові знаки, не перевищує найбільшої з граничних відносних похибок доданків:

$$\delta_{\text{EMBED}} \leq \max \delta_{\text{EMBED}_x}. \quad (1.7)$$

Додаючи наближені числа з різною кількістю десяткових знаків, дотримуються таких правил:

- 1) залишають без змін числа з найменшою кількістю десяткових знаків;
- 2) останні числа округлюють за зразком залишених без зміни, залишаючи один запасний десятковий знак;
- 3) виконавши додавання, результат округлюють на один знак;
- 4) повну абсолютну похибку обчислюють за формулою $\Delta = \Delta_1 + \Delta_2 + \Delta_3$,

де: Δ_1 – сума абсолютних похибок вихідних даних;

Δ_2 – абсолютна величина суми похибок округлених доданків;

Δ_3 – абсолютна похибка округлення результату.

б) Нехай $y = x_1 - x_2$; $0 < x_1 < x$. Граничну відносну похибку визначають за формулою

$$\delta_y = \frac{x_1 \cdot \delta_{x_1} + x_2 \cdot \delta_{x_2}}{y}. \quad (1.8)$$

Зазначимо, що за малих значень y (близьких до x_1 і x_2) похибка δ_y може бути великою навіть за малих δ_{x_1} та δ_{x_2} (втрата точності), тому в обчисленнях бажано позбутись такого роду різниць, наприклад:

$$\sqrt{1,01} - 1 = \frac{1,01 - 1}{\sqrt{1,01} + 1} = \frac{0,01}{\sqrt{1,01} + 1}.$$

в) Нехай $y = x_1 \cdot x_2 \cdot \dots \cdot x_n$, $x_i > 0$, тоді, оскільки $\ln y = \sum_{i=1}^n \ln x_i$,

$$\Delta \ln y \approx \frac{\Delta y}{y} = \left| \sum_{i=1}^n \frac{\Delta x_i}{x_i} \right| \leq \sum_{i=1}^n \frac{\Delta x_i}{x_i}. \text{ Таким чином, відносна похибка добутку}$$

(додатних співмножників) не перевищує суми відносних похибок співмножників, гранична відносна похибка такого добутку дорівнює сумі граничних відносних похибок множників. Граничну абсолютну похибку обчислюють за формулою

$$\Delta y = y \cdot \delta_y. \quad (1.9)$$

Зауважимо, що у співмножниках з різною кількістю правильних значущих цифр не варто залишати їх зайву кількість, тому що це збільшує відносну похибку. Виконуючи множення, слід керуватися такими правилами:

- 1) округлити співмножники так, щоб кожен з них містив на одну значущу цифру більше, ніж співмножники з найменшою кількістю значущих цифр;
- 2) результат множення округлюється до кількості значущих цифр, яку мали співмножники, залишені без змін.

г) Нехай $y = \frac{x_1}{x_2}$, $x_1, x_2 > 0$. Тоді

$$\ln y = \ln x_1 - \ln x_2, \quad \frac{\Delta y}{y} = \frac{\Delta x_1}{x_1} - \frac{\Delta x_2}{x_2};$$

$$\left| \frac{\Delta y}{y} \right| \leq \left| \frac{\Delta x_1}{x_1} \right| + \left| \frac{\Delta x_2}{x_2} \right|, \quad \delta_y = \delta_{x_1} + \delta_{x_2}.$$

д) Нехай $y = f(x_1 \dots x_n)$, $f(\bullet)$ - диференційована функція,

$$|\Delta y| = |y - y^0| = |f(x_1, \dots, x_n) - f(x_1^0, \dots, x_n^0)| \approx \left| \sum_{i=1}^n \frac{df}{dx_i} \Delta x_i \right| \leq \sum_{i=1}^n \left| \frac{df}{dx_i} \right| \cdot |\Delta x_i|,$$

звідки

$$\Delta y = \sum_{i=1}^n \left| \frac{df}{dx_i} \right| \Delta x_i;$$

$$\delta_y = \sum_{i=1}^n \left| \frac{df}{dx_i} \cdot f(\bullet) \right| \Delta x_i,$$

або

$$\delta_y = \sum_{i=1}^n \left| \frac{d \ln f}{dx_i} \right| \Delta x_i = \sum_{i=1}^n \left| \frac{d \ln f}{dx_i} \right| \cdot |x_i| \delta_{x_i}. \quad (1.10)$$

Взагалі кажучи, обчислювальна математика пропонує користувачеві різноманітні алгоритми розв'язування математичних задач. В зв'язку з цим наведемо ряд базисних значень теорії алгоритмів:

абеткою – називається набір скінченної кількості символів (літер),

словом – скінченна послідовність літер цієї абетки,

алгоритмом – (у деякій абетці A) називається точний загальнозрозумілий пропис P , що визначає потенціально здійснений процес перетворення слів у A , причому цей припис має бути:

а) скінченним і таким, щоб його виконання не потребувало додаткової інформації, крім тієї, що міститься в слові y у P ;

б) однозначно здійсненим;

в) відтворюваним (застосування до одного і того самого слова має призводити до однакових результатів);

г) процес застосування алгоритму має бути скінченним;

д) алгоритм має бути масовим (мати широку область застосування).

До обчислювальних алгоритмів пред'являються такі вимоги:

стійкість, тобто незначне зростання обчислювальної похибки в процесі його реалізації;

збіжність – близькість одержаного результату до точного значення, точніше, можливість досягти належної точності за допомогою використання деякого різновиду алгоритму;

коректність – можливість розв'язувати всі задачі в цьому класі, які мають розв'язок і до того ж єдиний, що неперервним чином залежить від вихідних даних;

економічність (якомога менша кількість здійснюваних арифметичних дій на обсяг пам'яті комп'ютера).

Наведемо деякі приклади.

Приклад 1. Ще в Давній Греції виникла так звана делійська задача про знаходження довжини ребра куба, об'єм якого подвоєно порівняно з об'ємом заданого куба. Ясно, що вона еквівалентна розв'язуванню рівняння

$$x^3 - 2 = 0.$$

Зауважимо, що розв'язком цієї задачі є границя послідовності

$$x_{n+1} = \frac{2}{3} \left(x_n + \frac{1}{x_n^2} \right).$$

Дійсно, переходячи до границі, яку позначатимемо як x , у лівій та правій частині рівності матиме: $x = \frac{2}{3} \left(x + \frac{1}{x^2} \right)$, або $x^3 = \frac{2}{3} (x^3 + 1)$,

$$\frac{1}{3} x^3 = \frac{2}{3}, \quad x^3 = 2.$$

Поклавши $x_0 = 1$, обчислюватимемо послідовні наближення x_n аж до досягнення потрібної точності $|x_n^3 - 2| < \varepsilon$ (або $|x_{n+1} - x_n| < \varepsilon$).

Відзначимо, що такий алгоритм є стійким (початкове наближення x_0 може і відрізнятись від одиниці у певних межах), збіжним (дає змогу досягти будь-якої наперед заданої точності), коректним (дає відповідь, і до того ж однозначно), економічним (може бути реалізований навіть найпростішими обчислювальними пристроями).

Приклад 2. Нехай треба обчислити інтеграл $I_n = \int x^n e^{x-1} dx$. Інтегрування частинами дає просту рекурентну формулу:

$$I_n = 1 - n \cdot I_{n-1}.$$

Зауваживши, що $I_1 = e^{-1}$, маємо, що, наприклад, $I_x = -0,06848$, але $x^n e^{x-1}$ за $0 \leq x \leq 1$ – невід’ємна функція, тому такий результат є неприйнятним. Виникає він внаслідок накопичування обчислювальних похибок, тобто цей алгоритм є нестійким.

Приклад 3. Многочлен $P(x) = (x-1)(x-2)\dots(x-20) = x^{20} + 210x^{19} + \dots$ має корені: $x_1 = 1, x_2 = 2, \dots, x_{20} = 20$. Збільшивши коефіцієнт за x^{19} на $2^{-23} \approx 10^{-7}$ (відносна похибка становить при цьому $\approx 5 \cdot 10^{-10}$!) і проводячи обчислення з точністю до 11 значущих цифр, ми одержимо істотно відмінні значення коренів (серед них навіть з’являться комплексні). В такому разі (так звана задача Вілкінсона) маємо справу із некоректністю поставленої задачі – відсутністю неперервної залежності розв’язку від вихідних даних.

Відмітимо, що для розв’язання некоректних задач часто можна запропонувати так звані процедури регуляризацій, приклади яких будуть розглянуті у наступних розділах.

Запитання для самоконтролю

1. Що таке абсолютна похибка?
2. Які вимоги ставлять до обчислювальних алгоритмів?
3. Що таке гранична відносна похибка?

Лекція 2. НАБЛИЖЕНЕ РОЗВ'ЯЗАННЯ АЛГЕБРАЇЧНИХ ТА ТРАНСЦЕНДЕНТНИХ РІВНЯНЬ (МЕТОД ХОРД, МЕТОД ДОТИЧНИХ)

У лекції запроваджується поняття про метод хорд, метод дотичних (Ньютона) розв'язання алгебраїчних рівнянь.

У шкільному курсі елементарної математики розглядаються методи розв'язання різноманітних алгебраїчних, ірраціональних, тригонометричних, показникових, логарифмічних рівнянь, які дають змогу знайти їх корені шляхом арифметичних операцій, добування коренів, знаходження значень показникової логарифмічної, тригонометричних та обернених тригонометричних функцій з використанням величин, що входять до рівняння. В той же час існують об'єктивні труднощі у розв'язанні навіть алгебраїчних рівнянь, пов'язані із складністю формул Кардано для розв'язання рівнянь третього та четвертого степенів і неможливістю одержання загальних формул для розв'язання рівнянь п'ятого та вищих степенів. Зважаючи на те, що коефіцієнти рівнянь, які розглядають в інженерних дослідженнях, знаходять з неминучими похибками вимірювання, доводиться розглядати наближені методи розв'язання різноманітних класів рівнянь.

Нехай розглядається рівняння $f(x)=0$, $f(x)$ – визначена та неперервна на деякому відрізку $[A,B]$, для якого потрібно визначити всі корені (нулі) на цьому відрізку, тобто всі числа є такими, що $\xi \in [A,B]$, такими, що $f(\xi)=0$.

Зазвичай, розв'язуючи таку задачу, доводиться спочатку відокремити корені (тобто поділити проміжок $[A,B]$ на такі проміжки $[a_i,b_i]$, що кожен з них містить лише один корінь рівняння, після чого знайти (уточнити) значення кожного з цих коренів з наперед заданою точністю. На початковому етапі розв'язування задачі зручно користуватися графіком функції.

Приклад.

Розглянемо рівняння

$$x^3-3x+2=0.$$

Зауважимо, що функція $f(x)=x^3-3x+2$, визначена для всіх дійсних значень x , прямує до мінус нескінченності за необмеженого спадання x і до плюс нескінченності за його необмеженого зростання. Її похідна $y^1=3x^2-3=3(x+1)(x-1)$,

що вказує на зростання функції від мінус нескінченності до чотирьох за зміни x від мінус нескінченності до -1 , а це означає наявність кореня рівняння на цьому проміжку. Відповідно спадання функції від 4 до 0 на проміжку від -1 до 1 дає змогу точно знайти ще один її корінь $x=1$, а подальше зростання від 0 до плюс нескінченності вказує на те, що інших коренів немає. У такому методі локалізації коренів використовують одну з властивостей неперервних функцій: якщо на кінцях деякого інтервалу неперервна функція набуває значень різних знаків, то на цьому інтервалі знайдеться точка, в якій функція дорівнює нулеві. Уточнення коренів можна виконувати методами випробувань: ділячи інтервали, на яких локалізовані корені, наприклад, навпіл і перевіряючи знаки функції $f(x)$ у знову одержаних точках, можемо будь-як сильно звузити інтервал, на якому знаходиться корінь. У прикладі, розглянутому раніше, корінь (або корені) рівняння належить нескінченному інтервалу, тому проміжну точку можна обирати довільно, наприклад:

$$f(-10)=-968, f(-5)=-107, f(-3)=-11, f(-2)=0.$$

Недоліком такого методу уточнення кореня є великий обсяг обчислень, потрібних для досягнення заданої точності. Розглянемо більш економічні методи локалізації коренів.

А. МЕТОД ХОРДА

Потребуватимемо від функції $f(x)$ наявності неперервної похідної $f'(x)$ на інтервалі $[a;b]$ (вважатимемо, що завдяки попередньо виконаній локалізації встановлено знаходження на цьому інтервалі одного кореня рівняння, при цьому похідні $f'(x)$ та $f''(x)$ повинні мати сталі знаки на цьому інтервалі). Нехай $f(a)<0$; $f(b)>0$; $f'(x)>0$; $f''(x)>0$ (інші випадки розглядають аналогічно). Удосконалення попереднього методу полягає в тому, що перевіряється значення функції $f(x)$ не в середній точці відрізка, на якому локалізоване значення кореня, а в точці, геометрично відповідній абсцисі точки перетину прямої, що з'єднує точки $M_1(a; f(a))$ та $M_2(b; f(b))$ з горизонтальною віссю. З вкладки зрозуміло стає назва методу. У розглядуваному випадку (рис.1):

$$x_1 = a - \frac{b-a}{f(b)-f(a)} f(a);$$

$$x_2 = x_1 - \frac{b - x_1}{f(b) - f(x_1)} f(x_1) (f(x_1) < 0);$$

...

$$x_n = x_{n-1} - \frac{b - x_{n-1}}{f(b) - f(x_{n-1})} f(x_{n-1});$$

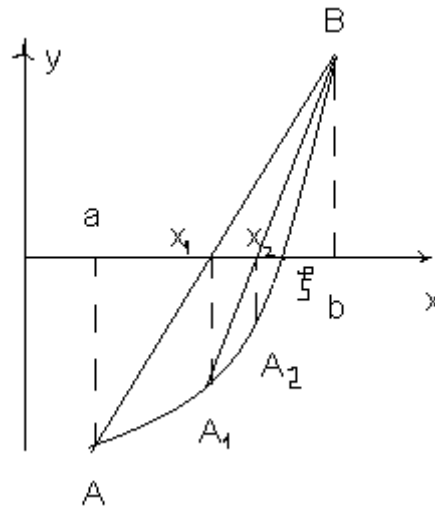


Рис.1

Збіжність такої послідовності до шуканого значення кореня буде доведена далі. Як вже було сказано, такий метод дає економію часу в обчисленнях. Застосовують і його видозміни, які дають змогу:

а) спростити обчислення (модифікований метод хорд):

$$x_n = x_{n-1} - \frac{b - x_1}{f(b) - f(x_1)} f(x_{n-1}); \quad (2.1)$$

б) додатково пришвидшити збіжність (удосконалений метод хорд):

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n). \quad (2.2)$$

Б. МЕТОД ДОТИЧНИХ (МЕТОД НЬЮТОНА)

Нехай x_0 – деяке початкове (знайдене довільним чином) наближення кореня. Тоді, скориставшись формулою Тейлора в околі точки $x=x_0$, маємо:

$$f(x) = f(x_0) + \frac{f'(x_0)}{1}(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 + \dots = 0; \quad (2.3)$$

наближений варіант цієї рівності має такий вигляд:

$$f(x_0) + f'(x_0)(x - x_0) \approx 0, \quad (2.4)$$

звідки

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \quad (2.5)$$

Геометрично це означає побудову дотичної до графіка функції $y=f(x)$ у точці $M_0(x_0;f(x_0))$ та знаходження абсциси x_1 точки її перетину з горизонтальною віссю. Розглядаючи x_1 як наступне наближення і повторюючи такі дії, маємо:

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}. \quad (2.6)$$

що являє собою рекурентну формулу згаданого методу. Має місце теорема, яка доводить, що такий процес збігається до значення кореня за $f(a)f(b)<0$; $f(x)\neq 0$; $f'(x)\neq 0$; $(x\in[a;b])$; $f(x_0)f'(x_0)>0$. Зазначимо, що найкращим (найшвидшим) чином метод Ньютона «спрацьовує» за великих значень $f(x)$ на проміжку $[a;b]$. Та якщо ці значення малі (функція «повільно змінюється»), то більш доцільним є застосування попереднього методу.

У разі застосування методу Ньютона можливо (рис.2):

а) спростити обчислення, поклавши $x_n = x_{n-1} - \frac{f(x_n - 1)}{f'(x_0)}$;

б) одночасно користуватись методом хорд, локалізуючи корінь між наближеннями, одержаними різними методами (одне – із надлишком, друге – з недостатчею).

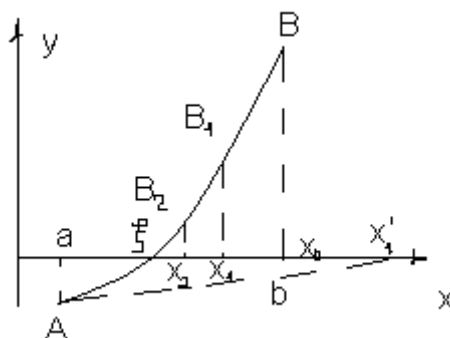


Рис.2

Запитання для самоконтролю

1. У чому полягає метод хорд?
2. Наведіть основні критерії застосування методу дотичних.

Лекція 3. НАБЛИЖЕНЕ РОЗВ'ЯЗАННЯ АЛГЕБРАЇЧНИХ ТА ТРАНСЦЕНДЕНТНИХ РІВНЯНЬ (МЕТОД ІТЕРАЦІЙ, МЕТОД ЛОБАЧЕВСЬКОГО)

У лекції запроваджується поняття про метод ітерацій та метод Лобачевського розв'язання алгебраїчних рівнянь.

А. МЕТОД ІТЕРАЦІЙ (ПОСЛІДОВНИХ НАБЛИЖЕНЬ)

Перепишемо рівняння у вигляді $x = \varphi(x)$. Нехай знову x_0 – початкове наближення.

Для знаходження наступних наближень використовуватимемо формулу $x_n = \varphi(x_{n-1})$ (рис. 3; 4).

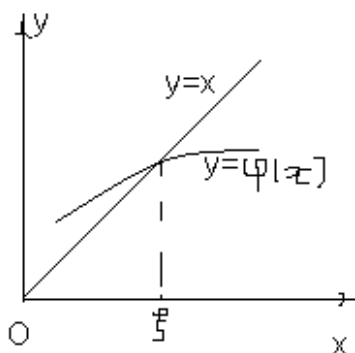


Рис. 3

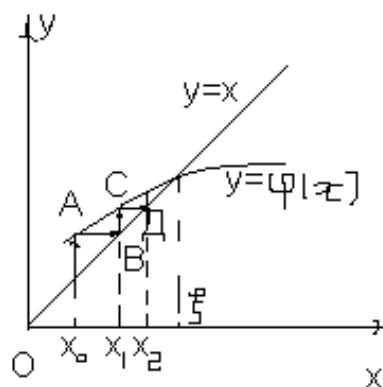


Рис. 4

Збіжність такої послідовності до значення кореня гарантує

Теорема. Нехай $\varphi(x)$ - неперервно диференційовна на $[a,b]$, причому $\varphi(x) \in [a,b]$ за $x \in [a,b]$; $|\varphi'(x)| < q < 1$; $x \in [a,b]$, тоді:

- 1) ітераційний процес збігається для довільного $x_0 \in [a,b]$;
- 2) граничне значення послідовності ітерацій є єдиним коренем рівняння $x = \varphi(x)$ на проміжку $[a,b]$.

Доведення. Відповідно до ітераційної формули $x_n = \varphi(x_{n-1})$; $x_{n+1} = \varphi(x_n)$. Звідси, користуючись формулою Лагранжа для скінченних приростів, маємо:

$$\begin{aligned} x_{n+1} - x_n &= \varphi(x_n) - \varphi(x_{n-1}) = \varphi'(\xi)(x_n - x_{n-1}); \\ |x_{n+1} - x_n| &\leq q |x_n - x_{n-1}|; \\ x_n &= x_0 + (x_1 - x_0) + (x_2 - x_1) + \dots + (x_n - x_{n-1}). \end{aligned} \tag{3.1}$$

$$\text{Ряд} \quad \lim_{n \rightarrow \infty} x_n = x_0 + \sum_{n=1}^{\infty} (x_n - x_{n-1}) \quad (3.2)$$

є абсолютно збіжним з огляду на можливість мажорувати його члени за абсолютною величиною членами збіжного ряду – суми нескінченно спадної геометричної прогресії.

$$\text{Дійсно,} \quad |x_n - x_{n-1}| \leq q |x_{n-1} - x_{n-2}| \leq q^2 |x_{n-2} - x_{n-3}| \leq \dots \leq q^{n-1} |x_1 - x_0| \quad (3.3)$$

Таким чином доведено перше твердження теореми. Друге твердження доведемо від супротивного. Нехай $\xi = \lim_{n \rightarrow \infty} \xi_n$, $\xi_1 \neq \xi$, ξ_1 – інший корінь рівняння, тобто $\xi_1 = \varphi(\xi_1)$, тоді $\xi_1 - \xi = \varphi(\xi_1) - \varphi(\xi) = \varphi'(c)(\xi_1 - \xi)$, звідки $(1 - \varphi'(c))(\xi_1 - \xi) = 0$, але $\varphi'(c) \neq 1$, тобто $\xi_1 = \xi$.

Зазначимо, що таким чином доведено також збіжність методів хорд та Ньютона.

Зауваження:

1) Метод ітерацій є самовиправним, тобто поява похибки на деякому кроці лише уповільнює збіжність, але не порушує її.

2) Геометричний зміст методу ітерації полягає у тому, що для точки x_n , розміщеної на горизонтальній осі, точку x_{n+1} знаходять як значення ординати відповідної точки на графіку функції.

Б. ОЦІНКА ПОХИБКИ МЕТОДУ ІТЕРАЦІЇ

Оцінимо абсолютну похибку методу:

$$\begin{aligned} |x_{n+k} - x_n| &= |x_{n+k} - x_{n+k-1} + x_{n+k-1} - x_{n+k-2} + \dots + x_{n+1} - x_n| \leq |x_{n+k} - x_{n+k-1}| + \dots + |x_{n+1} - x_n| \leq \\ &\leq q^n (1 + q + \dots + q^{k-1}) |x_1 - x_0| = q^n \frac{1 - q^k}{1 - q} |x_1 - x_0| \leq \frac{q^n}{1 - q} |x_1 - x_0| \end{aligned} \quad (3.4)$$

$$\text{Переходячи до границі за } k \rightarrow +\infty, \text{ маємо } |x_n - \xi| \leq \frac{q_n}{1 - q} |x_1 - x_0|.$$

Зазначимо, що похибка тим менша, чим менше значення q , тобто найкраще цей метод придатний для функцій, що «повільно змінюються», на противагу методу Ньютона.

Повернемося до початкового вигляду рівняння $f(x) = 0$, звідки:

$$f(x) = x - \varphi(x);$$

$$f(x_n) = x_n - \varphi(x_n) = x_n - x_{n+1} = f(x_n) - f(\xi) = f'(c)(x_n - \xi);$$

$$|x_n - \xi| = \left| \frac{x_n - x_{n+1}}{f'(c)} \right| \leq \frac{|x_{n+1} - x_n|}{1 - q},$$

якщо

$$\varphi'(c) < q < 1,$$

або

$$|x_n - \xi| \leq \frac{q}{1 - q} |x_n - x_{n-1}| \leq \dots$$

Якщо

$$q \leq \frac{1}{2}, \text{ то } |x_n - \xi| \leq |x_n - x_{n-1}|.$$

Зауваження:

- 1) методи хорд та Ньютона є частинними випадками методу ітерацій.
- 2) можна запропонувати такі критерії для вибору методу (з погляду його найшвидшої збіжності):

а) якщо в околі кореня $(x \in [a, b]) \left| \left(x - \frac{f(x)(b-x)}{f(b) - f(x)} \right)' x \right| \leq \frac{1}{2}$, то є сенс

застосувати метод хорд;

б) якщо в аналогічній області $\left| \left(x - \frac{f(x)}{f'(x)} \right)' x \right| \leq \frac{1}{2}$, то найшвидшої

збіжності досягають методом дотичних.

В. РОЗВ'ЯЗАННЯ АЛГЕБРАЇЧНИХ РІВНЯНЬ ЗА МЕТОДОМ ЛОБАЧЕВСЬКОГО

Зауважимо, що локалізація коренів рівняння суттєво ускладнюється за наявності великої кількості коренів, а також якщо серед них є комплексні корені. У розв'язанні численних задач з електро- та радіотехніки, оптимального управління виникають саме такі ситуації, причому окрім дійсних коренів відповідних алгебраїчних рівнянь, необхідно знаходити і комплексні. Описаний нижче метод вперше був запропонований М. В. Лобачевським.

Нехай
$$f(x) = x^n + a_1 x^{n-1} + \dots + a_n = 0, \quad (3.5)$$

всі корені цього рівняння дійсні, причому $|x_1| > |x_2| > \dots > |x_n|$.

Рівняння $F(x) = x^n + A_1 x^{n-1} + \dots + A_n = 0$ (надалі позначатиметься(*))
має корені $x_1^m, -x_2^m, \dots, -x_n^m$,
при цьому у відповідності з теоремою Вієта

$$A_1 = x_1^m + \dots + x_n^m = x_1^m \left(1 + \left(\frac{x_2}{x_1} \right)^m + \dots + \left(\frac{x_n}{x_1} \right)^m \right);$$

$$A_2 = (x_1 x_2)^m \left(1 + \left(\frac{x_1 x_3}{x_1 x_2} \right)^m + \dots + \left(\frac{x_{n-1} x_n}{x_1 x_2} \right)^m \right);$$

$$A_n = (x_1 \cdot x_2 \cdot \dots \cdot x_n)^m.$$

За достатньо великих значень m $A_1 \approx x_1^m$;

$$A_2 \approx (x_1 x_2)^m;$$

.
.
.

$$A_n \approx (x_1 x_2 \dots x_n)^m;$$

за парних m $|x_1| \approx A_1^{\frac{1}{m}};$

$$|x_2| \approx \left(\frac{A_1}{A_2} \right)^{\frac{1}{m}};$$

.
.
.

$$|x_n| \approx \left(\frac{A_n}{A_{n-1}} \right)^{\frac{1}{m}}.$$

Знаки коренів встановлюють завдяки простій перевірці: підстановці у вихідне рівняння $\pm |x_i|$. Таким чином, наближене знаходження коренів рівняння потребує:

- а) складення за даним рівнянням рівняння (•);
- б) визначення у відповідності з заданою точністю значення m .

Щодо пункту «а» зазначимо, що в разі квадратування (тобто побудови рівняння, коренями якого є $-x_1^2, -x_2^2, \dots, -x_n^2$, де $x_1, \dots, x_n$ – корені вихідного рівняння), коефіцієнти визначають за формулами

$$b_i = a_i^2 + 2 \sum_{s=1} (-1)^s a_{i-s} a_{i+s} \quad (3.6)$$

Одержане таким чином рівняння $x^n + bx^{n-1} + \dots + b_n = 0$ знову піддається квадратуванню і так далі, звідки $m = 2^k$.

Стосовно пункту «б» зауважимо, що позначивши коефіцієнти многочлена, одержаного після k -го квадратування, як $b_i^{(k)}$, матимемо

$$b_i^{(m+1)} \approx (x_1 \cdot x_2 \cdot \dots \cdot x_1)^{2^{m+1}} = \left((x_1 \cdot x_2 \cdot \dots \cdot x_i)^{2^m} \right)^2 = \left(b_i^{(m)} \right)^2. \quad (3.7).$$

Таким чином, якщо після чергового квадратування коефіцієнти одержаного рівняння в межах припустимої похибки дорівнюють квадратам коефіцієнтів попереднього рівняння, то потрібної точності досягнуто.

Розглянемо тепер випадок, коли серед коренів рівняння є комплексні

$$x_{k_1} \cdot k + 1 = re^{\pm i\phi} = r(\cos \phi + \sin \phi).$$

Нехай корені рівняння знову впорядковані за абсолютною величиною (модулем):

$$|x_1| > |x_2| > \dots > |x_k| = |x_{k+1}| > \dots > |x_n|.$$

Оскільки

$$(x_{k,k+1})^m = r^m e^{\pm im\phi},$$

то

$$(x_k \cdot x_{k+1})^m = r^{2m};$$

$$\begin{aligned} (x_1 \cdot x_2 \cdot \dots \cdot x_k)^m + (x_1 \cdot x_2 \cdot \dots \cdot x_{k-1} \cdot x_{k+1})^m &= (x_1 \cdot x_2 \cdot \dots \cdot x_{k-1})^m (x_k^m + x_{k+1}^m) = \\ &= (x_1 \cdot x_2 \cdot \dots \cdot x_{k-1})^m \cdot 2r^m \cos m\phi. \end{aligned}$$

Звідси система набуває вигляду

$$\begin{aligned} A_1 &= x_1^m (1 + \alpha_1) \\ &\vdots \\ A_{k-1} &= (x_1 x_2 \cdot \dots \cdot x_{k-1})^m (1 + \alpha_{k-1}) \\ A_k &= (x_1 x_2 \cdot \dots \cdot x_{k-1})^m (\cos m\phi + \alpha_k) \cdot 2r^m \\ &\vdots \\ A_n &= (x_1 x_2 \cdot \dots \cdot x_{k-1})^m r^{2m} (x_{k+2} \cdot \dots \cdot x_n)^m. \end{aligned}$$

Знехтувавши всіма значеннями α_i , як й у випадку дійсних коренів, але за винятком α_k , отримуємо:

$$\begin{aligned}
& {}^m A_1 \approx x_1^m \\
& \vdots \\
& A_{k-1} \approx (x_1 \cdot \dots \cdot x_{k-1})^m \\
& A_k \approx (x_1 \cdot \dots \cdot x_{k-1})^m (\cos m\varphi + \alpha_k) \cdot 2r^m \\
& A_{k+1} \approx (x_1 \cdot \dots \cdot x_{k-1})^m \cdot r^{2m} \\
& \vdots \\
& A_n = (x_1 \cdot \dots \cdot x_{k-1} \cdot x_{k+2} \cdot \dots \cdot x_n)^m \cdot r^{2m}.
\end{aligned}$$

Як було показано раніше, після достатньо великої кількості квадратур стане помітною тенденція, що виявляється в близькості коефіцієнтів, одержаних після чергового квадратування, до попередніх коефіцієнтів. Ця закономірність не буде справедливою щодо коефіцієнтів $b^{(n)}$, що дає змогу відразу відділяти комплексні корені від дійсних, які знаходять аналогічно попередньому. Далі

$$r = \left(\frac{A_{k+1}}{A_k} \right)^{\frac{1}{2m}}; a_1 = - \sum_{i=1}^n (x_1 + \dots + x_{k-1} + 2rc \cos \phi + x_{k+2} + \dots + x_m),$$

що дає змогу знайти $\cos \phi$, отже, й аргумент комплексного числа або його дійсну частину (при цьому якщо)

$$\begin{aligned}
Z &= u + iv; \\
v &= \sqrt{|Z|^2 - u^2} = \sqrt{r^2 - u^2}.
\end{aligned}$$

Запитання для самоконтролю

1. У чому полягає метод ітерацій?
2. Назвіть основні критерії застосування методу Лобачевського.

Лекція 4. ЗАДАЧА ІНТЕРПОЛЯЦІЇ: ІНТЕРПОЛЯЦІЙНІ МНОГОЧЛЕНИ НЬЮТОНА ТА ЛАГРАНЖА

У лекції запроваджено поняття точкової та інтегральної апроксимації, застосування степеневих рядів. Розглядається використання інтерполяційних многочленів Лагранжа та Ньютона і мінімізація залишкового члена інтерполяційної формули.

Часто для встановлення функціональної залежності $y = f(x)$ доводиться починати з того, що відомі лише значення $y_i = f(x_i), i = 1, 2, \dots, n$. Трапляється також, що точний вигляд функціональної залежності є відомим, але настільки громіздким, що практичні обчислення потребують переходу до менш трудомісткої формули. Обчислювальні аспекти таких задач є предметом теорії наближення (апроксимації) функцій.

А. ТОЧКОВА ТА ІНТЕГРАЛЬНА АПРОКСИМАЦІЯ. ВИКОРИСТАННЯ СТЕПЕНЕВИХ РЯДІВ.

Апроксимація функції називається

точковою, якщо шукають її значення (точні або наближені) на дискретній (скінченній або зліченній) множині точок;

інтегральною, якщо шукають її значення на неперервній множині точок.

Із курсу математичного аналізу відомо, що $(n+1)$ раз диференційована на інтервалі $[a, b]$ функція $y = f(x)$ може бути представлена за формулою Тейлора $f(x) = P_n(x) + R_n(x)$, де $P_n(x)$ – многочлен Тейлора n -го степеня;

$R_n(x)$ – залишковий член формули Тейлора n -го порядку;

$$P_n(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k; \quad R_n(x) = \frac{f^{(n+1)}(x_0 + \theta(x - x_0))}{(n+1)!} (x - x_0)^{n+1};$$
$$x \in (a, b), \theta \in [0; 1] \quad (4.1)$$

Це дає змогу звести обчислення значень, наприклад трансцендентних функцій $y = \sin x, y = \ln x, y = e^x$, до обчислення значень поліномів, тобто виконання чотирьох арифметичних операцій. Залишковий член $R_n(x)$ досить швидко спадає (за абсолютною величиною) за збільшення n та (або) прямування x до x_0 . В той же часу міру віддалення x від x_0 похибка обчислень за формулою $f(x) \approx P_n(x)$ за фіксованих значень n має

тенденцію до швидкого зростання. Наприклад, обчислюючи значення функції $y = \sin x$ за допомогою многочлена п'ятого степеня, тобто $\sin x \approx x - \frac{x^3}{6} + \frac{x^5}{120}$, можна легко побачити значну похибку навіть за

$x = \frac{\pi}{3}$. Щоправда, якоюсь мірою така ситуація може бути виправлена

завдяки належному вибору значення x_0 , але це потребує обчислення наново коефіцієнтів розкладу, що часто є неможливим і практично завжди обтяжливим. Для усунення такого явища, що може бути назване нерівномірним розподілом похибки на інтервалі $[a; b]$, застосовують рівномірні наближення функцій, які можуть реалізовуватись, наприклад, за допомогою многочленів Чебишова.

Означення. Многочлен n -го степеня, що визначається на проміжку $[-1; 1]$ співвідношенням

$$T_n(x) = \cos(n \arccos x) = \frac{1}{2} \left[(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n \right], \quad (4.2)$$

називається многочленом (поліномом) Чебишова (відповідного степеня).

Відзначимо таке:

- 1) $T_0(x) = 1$;
- 2) $T_1(x) = x$;
- 3) $T_2(x) = 2x^2 - 1$;
- 4) за парних n $T_n(x)$ містить тільки парні степені x , за непарного – тільки непарні;

5) старший коефіцієнт $T_n(x)$ ($n \geq 1$) дорівнює 2^{n-1} ;

6) $T_n(x)$ має n дійсних коренів на інтервалі $(-1; 1)$;

$$x = \cos \frac{(2i+1)\pi}{2n}, i = \overline{0, n-1}, \max |T_n(x)| = 1, T_n(x_m) = (-1)^m,$$

$$\text{де } x_m = \cos \frac{m\pi}{n}, m = \overline{0, n};$$

7) $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), n = 1, 2, \dots$;

8) многочлен $\overline{T}_n(x) = 2^{1-n} T_n(x), n \geq 1$ серед усіх многочленів n -го степеня зі старшим коефіцієнтом 1 має на $[-1; 1]$ найменше значення максимуму модуля, $\max_{[-1; 1]} |\overline{T}_n(x)| = 2^{1-n}$.

З огляду на це, $T_n(x)$ називаються многочленами, які найменше відхиляються від нуля. Якщо тепер замість ряду Тейлора $f(x) = a_n + a_1x + a_2x + \dots$ використовувати ряд $f(x) = c_n + c_1T_1(x) + c_2T_2(x) + \dots$, то похибка апроксимації буде розподілена рівномірно на проміжку наближення. Зазначимо, що проміжок $[a; b]$ може бути переведений у проміжок $[-1; 1]$ за допомогою підстановки $t = \frac{a+b}{2} + \frac{b-a}{2}x$. При цьому оцінку похибки виконують, виходячи з таких співвідношень: оскільки $T_n(x) = b_n + b_1x + \dots + 2^{n-1}x^n$, то $x^n = 2^{1-n}(b_n + b_1x + \dots + b_{n-1}x^{n-1}) + 2^{1-n}T_n(x)$, отже, нехтуючи останнім доданком, припускаємося похибки $|\Delta| \leq 2^{1-n}(|T_n(x)| \leq 1)$, а підставляючи вираз для x^n у скінченну суму ряду, одержуємо многочлен $(n-1)$ степеня. Такий процес триває аж до перевищення допустимої величини похибки.

Б. ЗАДАЧА ІНТЕРПОЛЯЦІЇ.

ІНТЕРПОЛЯЦІЙНИЙ МНОГОЧЛЕН ЛАГРАНЖА

Нехай на проміжку $[a; b]$ задано сукупність точок (вузлів інтерполяції) $x_0, x_1, x_2, \dots, x_n$ (попарно різних), у яких відомо значення функції $f(x)$.

Розглянемо одночасно деяку систему функцій $\varphi_i(x)$ (наприклад, $\varphi_i(x) = x^i$ – степенева система, або $\{\varphi_i(x)\} = \{1; \sin x; \cos x; \sin 2x; \cos 2x, \dots\}$ – тригонометрична система).

Називатимемо лінійну комбінацію таких функцій $\varphi(x) = \sum_{i=0}^n a_i \varphi_i(x)$ узагальненим інтерполяційним многочленом для $f(x)$, якщо $\varphi(x_j) = f(x_j), j = \overline{0, n}$.

За такої умови коефіцієнти a_i знаходять із системи $(n+1)$ рівняння $\sum_{i=0}^n a_i \varphi_i(x_j) = f(x_j), j = \overline{0, n}$, яка має єдиний розв'язок, якщо система функцій $\{\varphi_i(x)\}$ є системою Чебишова (довільна лінійна комбінація $\sum_{i=0}^n c_i \varphi_i(x)$ не обертається в нуль частіше, ніж у n різних точках проміжку $[a; b]$).

В обчисленнях зручно користуватись так званими фундаментальними узагальненими многочленами $\Phi_j(x)$:

$$1) \quad \Phi_j(x) = \sum_{i=0}^j b_i \varphi_i(x)$$

$$2) \quad \Phi_j(x_i) = \delta_{ij}, (\delta_{ij} - \text{символ Кронекера}),$$

У випадку степеневї системи

$$\Phi_j(x) = l_j(x) = \frac{(x-x_0)\dots(x-x_{j-1})(x-x_{j+1})\dots(x-x_n)}{(x_j-x_0)\dots(x_j-x_{j-1})(x_j-x_{j+1})\dots(x_j-x_n)} \quad (4.3)$$

маємо так звані фундаментальні інтерполяційні поліноми Лагранжа. При цьому $l_n(x) = \sum_{j=0}^n f(x_j) l_j(x)$ називається інтерполяційним поліномом Лагранжа.

Узагальненням цього полігона є інтерполяційний поліном Ерміта:

$$H_m(x) = \sum_{i=0}^m c_i \varphi_i(x), H_m^{(j)}(x_i) = f_i^{(j)}, j = 0, 1, \dots, a_i - 1, j = \overline{0, n}, m = \sum_{i=0}^n a_i - 1, \quad (4.4)$$

в цьому випадку

$$H_m(x) = \sum_{i=0}^n \sum_{j=0}^{a_i-1} f_i^{(j)} H_{ij}(x), H_{ij}^{(p)}(x_k) = \delta_{ik} \cdot \delta_{jp}, p = 0, 1, \dots, a_k - 1, k = \overline{0, n}.$$

Конкретні вирази для $H_{ij}(x)$ можуть бути одержані, але тут з огляду на їх громіздкість не наводимо.

Легко зрозуміти, що для того, щоби система $\{\varphi_i(x)\}, i = \overline{0, n}$ була системою Чебишова, достатньо диференційованості $n+1$ раз цих функцій, та відмінності від нуля їх вронскіана. Ці вимоги справедливі для степеневї системи (а тому інтерполяційний многочлен Лагранжа n -го степеня існує, і при тому єдиний), тригонометричної (те саме), системи поліномів Ерміта (те саме).

В. РОЗДІЛЕНІ РІЗНИЦІ. ІНТЕРПОЛЯЦІЙНИЙ МНОГОЧЛЕН НЬЮТОНА

Суттєвим недоліком інтерполяційного многочлена Лагранжа є те, що поява додаткового вузла інтерполяції веде до повного перерахунку всіх його доданків. Для того щоби позбутися цього недоліку, запроваджується інтерполяційний многочлен Ньютона.

Розглянемо спочатку деякі найпростіші поняття так званого числення різниць. Нехай для функцій $y = f(x)$ відомі її значення $y_i = f(x_i)$ у вузлах $x_i = x_0 + ih, i = 0, 1, 2, \dots$.

Називатимемо *скінченною різницею першого порядку* вираз $\Delta y_i = y_{i+1} - y_i$ *скінченною різницею $n+1$ порядку* вираз $\Delta^{n+1} y_i = \Delta^n y_{i+1} - \Delta y_i, i = 0, 1, \dots$ (вводяться індуктивно).

Розділеними різницями 1-го порядку називаються числа

$$f(x_{i-1}, x_i) = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \quad (4.5)$$

Розділена різниця n -го порядку визначається за рекурентною формулою $f(x_0, x_1, \dots, x_n) = \frac{f(x_1, x_2, \dots, x_n) - f(x_0, x_1, \dots, x_{n-1})}{x_n - x_0}$. У

розглядуваному випадку рівновіддалених вузлів справедливою є така формула зв'язку між розділеними та скінченними різницями:

$$f(x_0, x_1, \dots, x_n) = \frac{\Delta^n f_0}{n! h^n}.$$

Не важко показати, що многочлен

$$P_n(x) = f(x_0) + (x - x_0)f(x_0, x_1) + (x - x_0)(x - x_1)f(x_0, x_1, x_2) + \dots + (x - x_0)\dots(x - x_{n-1})f(x_0, x_1, \dots, x_n) \quad (4.6)$$

є інтерполяційним, тобто $P_n(x_i) = f(x_i), i = \overline{0, n}$. Такий многочлен називається інтерполяційним многочленом Ньютона (іноді додатково вказують – для рівних або нерівних проміжків). Оскільки інтерполяційний многочлен n -го степеня для цієї системи вузлів має властивість єдності, то $P_n(x) = L_n(x)$.

Зауважимо, що в разі рівновіддалених вузлів

$$P_n = y_0 + \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{2! h^2}(x - x_0)(x - x_1) + \dots + \frac{\Delta^n y_0}{n! h^n}(x - x_0)\dots(x - x_{n-1}),$$

або, скориставшись заміною $t = \frac{x - x_0}{h}$, можемо записати:

$$P_n(x) = y_0 + t\Delta y_0 + \frac{t(t-1)}{2!}\Delta^2 y_0 + \dots + \frac{t(t-1)\dots(t-n+1)}{n!}\Delta^n y_0. \quad (4.7)$$

У практичному застосуванні інтерполяційного многочлена Ньютона цю формулу використовують для $x \in [x_0, x_1] (t \in [0; 1])$. У випадку $x = [x_i; x_{i+1}]$ використовують запис

$$P_n(x_i + th) = y_i + t\Delta y_i + \frac{t(t-1)}{2!}\Delta^2 y_i + \dots + \frac{t(t-1)\dots(t-n+1)}{n!}\Delta^n y_i \quad (4.8)$$

Такий многочлен називають інтерполяційним многочленом Ньютона для інтерполювання вперед (входять значення $y_k, k \geq i$), використовують його для лівої частини проміжку інтерполяції та для екстраполяції (знаходження значень функції поза відрізком ліворуч від x_0). Для інтерполяції на правій частині відрізка та екстраполяції у точках, які лежать праворуч від x_n використовують інтерполяційний многочлен Ньютона для інтерполювання вперед:

$$P_n(x_n + th) = y_n + t\Delta y_{n-1} + \frac{t(t+1)}{2!}\Delta^2 y_{n-2} + \dots + \frac{t(t+1)\dots(t+n-1)}{n!}\Delta^n y_0. \quad (4.9)$$

Нагадаємо ще раз, що визначальною властивістю інтерполяційного многочлена Ньютона є те, що в разі появи додаткового вузла (ліворуч від x_0 , праворуч від x_n) у відповідному многочлені (для інтерполювання відповідно вперед чи назад) з'являється додатковий доданок. Існують також інтерполяційні многочлени Гаусса, Бесселя, Стірлінга.

Г. МІНІМІЗАЦІЯ ЗАЛИШКОВОГО ЧЛЕНА ІНТЕРПОЛЯЦІЙНОЇ ФОРМУЛИ

Ясно, що в разі заміни функції $f(x)$ інтерполяційним многочленом $N_n(x)$ виникає похибка інтерполяції $R_n(x) = f(x) - N_n(x)$, яку інакше називають залишковим членом інтерполяційної формули. Можна довести, що за $f(x) \in C^{n+1}[a, b]$ справедливим є таке представлення залишкового

члена:
$$R_n(x) = \omega(x) f(x, x_0, \dots, x_n) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega(x),$$

де
$$\xi \in [a; b], \omega(x) = (x - x_0) \dots (x - x_n),$$

звідки
$$|R_n(x)| \leq \frac{M_{n+1}}{(n+1)!} |\omega(x)|,$$

де
$$M_{n+1} = \max_{x \in [a, b]} |f^{(n+1)}(x)|.$$

Величина $\omega(x)$ може бути мінімізована шляхом оптимального вибору вузлів інтерполяції (за практичних вимірювань, це є одним з етапів планування експерименту).

Нехай $[a, b]$ збігається з $[-1; 1]$, а вузли інтерполяції сумістимо з коренями полінома Чебишова $T_{n+1}(x) : x_i = \cos \frac{(2i+1)\pi}{2n+2}, i = \overline{0, n}$.

Беручи до уваги властивості 5), 6), 8) многочленів Чебишова, можемо стверджувати, що $\omega_n(x) = 2^{-n} T_{n+1}(x), |R_n(x)| \leq \frac{M_{n+1}}{(n+1)! 2^n}$, причому таку оцінку поліпшити неможливо.

У разі інтерполювання на довільному відрізку $[a, b]$ задача зводиться до раніше розглянутої задачі шляхом заміни $x = \frac{1}{2}[(b-a)t + b + a]$, тоді за вузли інтерполяції обирають точки

$$x = \frac{1}{2} \left[(b-a) \cos \frac{(2i+1)\pi}{2n+2} + b + a \right], \omega_n(x) = \frac{(b-a)^{n+1}}{2^{n+1}} \overline{T_{n+1}(t)}, |R_n(x)| \leq \frac{M_{n+1}(b-a)^{n+1}}{(n+1)! 2^{2n+1}}.$$

Зауважимо, що отримана при цьому похибка в 2^n разів менша за похибку тейлорівського наближення (і до того ж, вона більш рівномірно розподілена на проміжку!)

Здається абсолютно природним, що $R_n(x)$ має прямувати до 0 за нескінченного збільшення кількості вузлів інтерполяції. Але це не так! Ще на початку століття С. Н. Бернштейн продемонстрував, що для функції $f(x) = |x|$ на проміжку $[-1; 1]$ послідовність інтерполяційних многочленів, побудованих на рівновіддалених вузлах, не збігається до $f(x)$ у жодній точці, окрім точок -1; 0; 1. В той же час Марцинкевичем доведено, що в разі побудови системи вузлів спеціальним чином для кожного n та кожної неперервної функції $f(x)$ інтерполяційний процес буде рівномірно збіжним. Попри це такі методи побудови системи вузлів та інтерполяційних многочленів високого степеня використовують дуже рідко через їх громіздкість.

Запитання для самоконтролю

1. Що таке точкова та інтегральна апроксимація?
2. Охарактеризуйте особливості застосування інтерполяційного многочлена Лагранжа.
3. Охарактеризуйте застосування інтерполяційного многочлена Ньютона.

Лекція 5. ЛОКАЛЬНА АПРОКСИМАЦІЯ. СПЛАЙНИ. РІВНОМІРНІ НАБЛИЖЕННЯ

У лекції запроваджується поняття локальної апроксимації, поняття сплайнів, розглядається задача найкращого наближення функції та рівномірного наближення.

Розглянуті раніше інтерполяційні процедури називають глобальними (на всьому проміжку функція наближається за допомогою одного многочлена).

Як уже вказувалося, збільшення кількості вузлів не обов'язково веде до підвищення точності, при цьому також відбувається накопичення похибок. Тому часто застосовують кусково-поліноміальну або локальну апроксимацію: відрізок $[a; b]$ поділяють на частини $a = x_0 < x_1 < \dots < x_n < b$ (поділити Δ) і на кожному відрізку $[x_i, x_{i+1}]$, $i = \overline{0, n-1}$ будуть окремий інтерполяційний многочлен так, щоби їх спряження було гладким. Таким чином виникає поняття кусково-многочленних функцій однорідної структури – сплайнів.

Сплайном $S_{\Delta}^m(f, x)$ порядку m називають функцію, яка є многочленом m -го степеня

$$S_{\Delta}^m(f, x) = P_{k,m}(x) = a_{k0} + a_{k1} + \dots + a_{km}x^m, x \in [x_{k-1}, x_k], k = \overline{1, n} \quad (5.1)$$

і задовольняє такі умови гладкості в точках

$$x_i, i = \overline{1, n-1}: P_{im}^{(k)}(x_i) = P_{i+1}^{(k)}(x_i), k = \overline{0, m-1}.$$

Таким чином одержуємо для знаходження $(m+1)n$ коефіцієнтів $a_{ij}, i = \overline{1, n}, j = \overline{0, m}, (n-1)m$ рівнянь. Для знаходження останніх рівнянь використовують додаткові умови.

Наприклад, у побудові сплайна першого степеня ($m=1$) маємо систему співвідношень:

$$P_{i1}(x_{i-1}) = a_{i0} + a_{i1}x_{i-1} = f(x_{i-1}), \quad P_{i1}(x_i) = a_{i0} + a_{i1}x_i = f(x_i), \quad i = \overline{1, n}, \quad \text{звідки}$$

$$a_{i1} = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}, \quad a_{i0} = f(x_{i-1}) - a_{i1}x_{i-1}, \quad \text{тобто на кожному з відрізків між}$$

вузлами сплайн співпадає з інтерполяційним многочленом першого степеня для $f(x)$ з вузлами x_{i-1}, x_i . Аналогічно можна побудувати сплайни другого степеня, але найчастіше використовують сплайни третього степеня $S_{\Delta}^3(x)$. Зауважимо також, що можна розглядати сплайни, для яких в умовах

гладкості зумовлюється рівність похідних не до $m-1$ порядку, при цьому різниця між степенем сплайна і найвищим порядком похідної в умовах гладкості називається його дефектом.

Побудуємо кубічний сплайн з одиничним дефектом. На кожному з відрізків $[x_{i-1}, x_i], i = \overline{1, n}$:

$$S_{\Delta, i}^3(x) = S_i(x) = a_i + b_i(x - x_i) + \frac{c_i}{2}(x - x_i)^2 + \frac{d_i}{6}(x - x_i)^3;$$

$$a_i = S_i(x_i), b_i = S'_i(x_i) \text{ — нахил сплайна у точці } x_i;$$

$$c_i = S''(x_i), d_i = S'''(x_i).$$

Знайдемо тепер ці коефіцієнти.

1) Виходячи із задачі інтерполяції, $a_i = f(x_i), i = \overline{1, n}$. Додатково покладемо $a_0 = f(x_0)$.

2) Виходячи з неперервності сплайна $(S_i(x_i) = S_{i+1}(x_1), i = \overline{1, n})$,

$$a_i = a_{i+1} + b_{i+1}(x_i - x_{i+1}) - \frac{c_{i+1}}{2}(x_{i+1})^2 + \frac{d_{i+1}}{6}(x_i - x_{i+1})^3, i = \overline{0, n-1},$$
 звідки,
 поклавши $h_i = x_i - x_{i-1}, i = \overline{1, n}; h_i b_i \frac{h_i^2}{2} c_i + \frac{h_i^3}{6} d_i = f_i - f_{i-1}$.

Аналогічно маємо, спираючись на неперервність першої похідної,

$$c_i h_i - \frac{d_i}{2} h_i^2 = b_i - b_{i-1}, i = \overline{2, 4}, \text{ другої — } d_i h_i = c_i - c_{i-1}, i = \overline{2, n}.$$

3) Таким чином отримано $4n-2$ рівнянь з $4n$ необхідних. Останні два рівняння можуть полягати у деяких граничних умовах для $S(x)$, наприклад: $S''(a) = S''(b) = 0$, або $c_i - d_i h_i = 0, c_n = 0$.

Одержана при цьому система лінійних алгебраїчних рівнянь має єдиний розв'язок. Її матриця має певні спеціальні властивості, які дають змогу застосувати для розв'язування таких систем відповідні спеціалізовані алгоритми.

Розглянутий вище спосіб побудови сплайна називається глобальним, відмінність від нього певного локального способу може полягати, наприклад, у тому, що відомими є значення $f'_i = f'(x_i)$, тоді покладають $b_i = S'_i(x_i)$.

За спрощеної побудови сплайна покладають $b_i = \frac{f_{i+1} - f_{i-1}}{2h}, i = \overline{1, n-1}$,

$$b_0 = \frac{4f_1 - f_2 - 3f_0}{2h}, b_n = \frac{3f_n - f_{n-2} + 4f_{n-1}}{2h}.$$

За необмеженого збільшення кількості вузлів інтерполяції n послідовність сплайн-функцій збігається до функцій $f(x)$. Такі сплайни називаються інтерполяційними, оскільки дотримується умова рівності їх значень значенням інтерпольованої функції у вузлах.

Зазначимо, що серед функцій, які мають інтегровану з квадратом другу похідну, умовам

$$1) u(x_i) = f_i, i = \overline{0, n};$$

$$2) \int_a^b [u''(x)]^2 dx - \min$$

задовольняє лише кусково-кубічна сплайн-функція. Проте не завжди доцільно користуватись для наближення саме многочленами (навіть локально). Наприклад, функцію, яка є періодичною з періодом 1 природно наближати за допомогою тригонометричної системи (відрізка ряду Фур'є).

$$\varphi_k(x) = a_k \cos \frac{\pi k x}{l} + b_k \sin \frac{\pi k x}{l}, k = \overline{0, n} \text{ часто також використовують систему}$$

$$\text{дробово-раціональних функцій} \quad \varphi_{k1}(x) = \frac{a_k x^k + a_{k-1} x^{k-1} + \dots + a_0}{x^k + b_{k-1} x^{k-1} + \dots + b_0}, \varphi_{k1}(x_i) = f(x_i),$$

$$j = \overline{0, n}, n = k + e, \text{ зокрема дробово-лінійних } \varphi(x) = \frac{a_k + a_0}{x + b_0}.$$

У випадку функції двох змінних можна використовувати лінійну інтерполяцію для апроксимації її частинних похідних: нехай $u = u(x, y), u(x_i, y_i) = u_i, i = 1, 2, 3$, тоді, поклавши $u(x, y) = a(x - x_i) + b(y - y_i) + e$, де $a = \frac{du}{dx}, b = \frac{du}{dy}$, маємо можливість знайти a та b у точці (x_1, y_1) .

Зауважимо, що лінійна інтерполяція функції двох змінних по трьох точках має просту геометричну інтерполяцію: вона являє собою побудову площини за трьома точками з координатами $M_1(x_1, y_1, u_1), M_2(x_2, y_2, u_2), M_3(x_3, y_3, u_3)$. Як відомо, така площина визначається

$$\text{рівнянням: } \begin{vmatrix} u - u_1 & x - x_1 & y - y_1 \\ u_2 - u_1 & x_2 - x_1 & y_2 - y_1 \\ u_3 - u_1 & x_3 - x_1 & y_3 - y_1 \end{vmatrix} = 0.$$

Загальна задача найкращого наближення функцій

Нехай розглядається деяка множина функцій H , яка має назву лінійного нормованого простору. На такій множині кожній функції ставиться у відповідність число $\|f\|$, яке називається її нормою. При цьому у H існує єдиний елемент 0 , норма якого дорівнює нулеві. Отже,

близькість до нуля $\|f - \varphi\|$ означає близькість між собою елементів f та φ . Загальна задача наближення формується таким чином: нехай $f, \varphi_0, \varphi_1, \dots, \varphi_n \in H, \varphi = c_0 \varphi_0 + \dots + c_n \varphi_n$ – узагальнений многочлен, треба знайти такі коефіцієнти $c_0, c_1, \dots, c_n$, щоби $\|f - \varphi\|$ була мінімальною.

Такий елемент φ називається елементом найкращого наближення (ЕНН). Найчастіше розглядають множину функцій, визначених на інтервалі $[a, b]$ та інтегрованих з квадратом: $\int_a^b f^2(x) dx < +\infty$.

Для таких функцій можна визначити скалярний добуток – $(f, g) = \int_a^b f(x)g(x)dx$ та норму $\|f\| = \sqrt{\int_a^b f^2(x)dx}$.

Якщо функції $\varphi_0, \varphi_1, \dots, \varphi_n$ – лінійно незалежні, то можна довести, що задача пошуку ЕНН зводиться до розв'язання системи лінійних алгебраїчних рівнянь $((f - \varphi, \varphi_k) = 0, k = \overline{0, n})$, тобто $\sum_{i=0}^n (\varphi_i, \varphi_k) = (f, \varphi_k)$, визначник якої є визначником Грама системи лінійно незалежних функцій $\{\varphi_k\}_{k=0}^n$, відповідно відмінний від 0.

Зауважимо, що у випадку, коли $\{\varphi_k\}_{k=0}^{+\infty}$ – повна система, тобто справедливо є рівність Парсеваля $\sum_{k=0}^n |c_k|^2 = \|f\|^2$, то послідовність ЕНН збігається до функції f . Такий спосіб наближення функції називається середньоквадратичним наближенням, а сам спосіб його пошуку – методом найменших квадратів.

Надалі розглянемо простір $H = L_{2,a}[a, b]$ ($[a, b]$ – нескінченний інтервал) $(u, v)_{2,a} = \int_a^b u(x)v(x)d\alpha(x), \|u\| = \sqrt{(u, u)}$. При цьому (теорема Вейерштрасса) для довільного $\xi > 0$ існує многочлен $P_n(x)$ такий, що $\|f - P_n\| < \xi$, тобто система многочленів $P_n(x)$, є повною (відзначимо, що цю систему можна вважати ортонормованою, тобто $(\varphi_i, \varphi_j) = \delta_{ij}$ де δ_{ij} – символ Кронекера).

На практиці, беручи $\varphi_i = x_i$, одержуємо систему лінійних алгебраїчних рівнянь, яка є погано збумовленою (тобто незручною для розв'язання стандартними методами через значні неусувні похибки). Тому найчастіше використовують многочлени ортогональні з вагою $\rho(x)$ на інтервалі $[a, b]$:

$$P_{in}(x) = A_n \begin{pmatrix} a_0 & a_1 \dots a_n \\ a_1 & a_2 \dots a_{n+1} \\ \vdots & \vdots \\ \vdots & \vdots \end{pmatrix},$$

де

$$\begin{matrix} a_{n-1} & a_n \dots a_{2n-1} \\ 1x & x^n \end{matrix}$$

A_n – константа нормування;

$a_n = \int_a^b x^n \rho(x) dx$ – момент вагової функції.

До таких многочленів належать:

а) описані вище многочлени Чебишова $\left([a;b] = [-1;1], \rho(x) = \frac{1}{\sqrt{1-x^2}}\right)$;

б) многочлени Лежандра $P_0(x) = p_i(x) = 1, ([a;b] = [1;1], \rho = 1)$

$$P_{i+1}(x) = \frac{[(2i+1)xP_i(x) - iP_{i-1}(x)]}{(i+1)}, i=1,2,\dots;$$

в) многочлени Лагерра $L_0^{(0)}(x) = 1, L_i^{(0)}(x) = 1-x$;

$$L_{i+1}^{(0)} = \frac{[(2i+1+x)L_i^{(0)}(x) - iL_{i-1}^{(0)}(x)]}{(i+1)}, ([a,b] = [a_1 + \infty), \rho = e^{-x}), i=1,2,\dots$$

г) многочлени Ерміта, $H_0(x) = 1, H_1(x) = 2x, H_{i+1}(x) = 2xH_i(x) - 2iH_{i-1}(x), i=1,2,\dots, ([a;b] = (-\infty; +\infty), \rho = e^{-x^2})$.

Згадана теорема Вєєрштрасса є справедливою для обмеженого інтервалу. Можна довести, що ряд Фур'є, побудований за нормованою системою класичних многочленів збігається (у просторі функцій $L_{2,a}[a,b]$ до $f(x)$).

В експериментальних дослідженнях часто доводиться мати справу із ступінчастими і, взагалі кажучи, негладкими функціями. Відповідні методи апроксимації, які враховують необхідність згладжування (нівелювання похибок вимірювання) викладено далі.

Рівномірні наближення. Чебишовський альтернанс

У попередньому випадку показником близькості функції $f(x)$ та ЕНН було значення норми їх різниці у просторі L_2 або близькому до нього просторі $L_{2,a}$. Значно складнішою є задача з пошуку *многочлена*

найкращого рівномірного наближення (МНРН), тобто пошук серед множини π_n многочленів виду $P_n(x) = \sum_{i=0}^n a_i x^i$ такого $Q_n(x)$, щоби

$$\Delta_n(f) = \Delta_n(f, Q_n) = \min_{P_n \in \pi_n} \max_{a \leq x \leq b} |f(x) - P_n(x)|.$$

Відзначимо таке:

1) немає методу знаходження за скінченної кількості кроків МНРН;

2) доведено, що існує єдиний такий многочлен степеня, не вищого за n (для довільного n та неперервної функції $f(x)$).

3) якщо $f(x) = 0$, то $Q_n(x)$ є многочленом, що найменше відхиляється від 0, а серед усіх многочленів степеня n з одиничним старшим коефіцієнтом таким є многочлен Чебишова, так що можна довести, що для функції x^{n+1} МНРН є $\Phi_0(x) = x^{n+1} - \frac{1}{2^n} T_{n+1}(x)$, де T_{n+1} – многочлен Чебишова, при цьому $\Delta_n(x^{n+1}) = \frac{1}{2^n}$.

Чебишовим свого часу показано, що для того, щоби многочлен $Q_n(x)$ був МНРН для $f(x) \in C[a; b]$, необхідно і достатньо наявності на $[a; b]$ хоча б $n + 2$ точок $a = x_0 < x_1 < \dots < x_{n+1} = b$, у яких $f(x_i) - Q_n(x_i) = \sigma \cdot (-1)^i \cdot \|f - Q_n\|_{C[a, b]}$, де $\sigma = 1$, або $\sigma = -1$, одночасно для всіх точок $\|f\|_{C[a, b]} = \max_{x \in [a, b]} |f(x)|$. Точки x_i , які задовольняють такі вимоги, називаються точками Чебишовського альтернансу.

Опишемо, наприклад, метод побудови таких точок і відповідно МНРН для випадку, коли $f(x)$ – двічі неперервно-диференційовна, опукла функція $n = 1$. Нехай МНРН $Q_1(x) = a_0 + a_1 x$, тоді функція $\varphi(x) = f(x) - Q_1(x)$ має на відрізку $[a; b]$ одну внутрішню точку екстремуму $x = c$. Точки a і b – точки Чебишовського альтернансу. Знайдемо третю таку точку: $f(a) - Q_1(a) = R$, $f(b) - Q_1(b) = R$, $f(c) - Q_1(c) = -R$, де $R = \sigma \|f - Q\|$.

У точці C маємо $f(c) - a_1 c = 0$, тобто невідомі R, c, a_1, a_0 визначаються

з рівнянь;

$$\begin{cases} R + a_0 + a_1 a = f(a) \\ R + a_0 + a_1 b = f(b) \\ R - a_0 - a_1 c = f(c) \\ a_1 = f'(c) \end{cases}.$$

Геометрично це означає, що пряма $y = a_0 + a_1 x$ є паралельною хорді, що з'єднує точки $(a; f(a)), (b; f(b))$, дотичній проведеній паралельно цій хорді у точці C та лежить посередині між ними.

Запитання для самоконтролю

1. Що таке локальна апроксимація ?
2. У чому полягає задача найкращого наближення?
3. У чому суть рівномірного наближення?

Лекція 6. ПРЯМІ МЕТОДИ РОЗВ'ЯЗАННЯ СИСТЕМ ЛІНІЙНИХ АЛГЕБРАЇЧНИХ РІВНЯНЬ

У лекції розглядаються основні поняття та точні методи розв'язування систем лінійних алгебраїчних рівнянь, прикладами яких слугують методи Крамера, оберненої матриці, Гаусса. Розглянуто також метод відображень, метод обертань і метод обведень. Визначено умови застосування цих методів.

Загальні поняття

До розв'язання системи лінійних алгебраїчних рівнянь (СЛАР) приводять численні практичні задачі (наприклад, дослідження лінійних електричних ланцюгів), а також переважна більшість алгоритмів обчислювальної математики. При цьому основною задачею є розв'язання системи з квадратною матрицею $A\bar{X} = \bar{b}$:

$\bar{X} = \{x^{(1)}, \dots, x^{(n)}\}$ – вектор невідомих;

$\bar{b} = \{b^{(1)}, \dots, b^{(n)}\}$ – вектор правих частин;

$A = \{a_{ij}\}_{i,j=1}^n$ – квадратна невироджена матриця ($\det A \neq 0$).

Як відомо з теорії, така система має єдиний розв'язок, для знаходження якого можуть бути застосовані методи Крамера, оберненої матриці, Гаусса. Визначимо, що для знаходження визначника матриці розмірності $n \times n$ за визначенням

$$D = \det A = \sum (-1)^k a_{1a_1} \cdot a_{2a_2} \cdot \dots \cdot a_{na_i} \quad (6.1)$$

треба виконати $n \cdot n! - 1$ арифметичних операцій, тобто визначник десятого порядку потребує для свого обчислення $3,6 \cdot 10^7$ операцій! Це вказує на неефективність практичного застосування методу Крамера. Аналогічно лише для обчислення оберненої матриці:

$$A^{-1} = \begin{pmatrix} \frac{A_{11}}{D} & \dots & \frac{A_{n1}}{D} \\ \vdots & \ddots & \vdots \\ \frac{A_{1n}}{D} & \dots & \frac{A_{nn}}{D} \end{pmatrix}$$

потрібно $n \cdot n! - 1$ операцій, що дає можливість дійти аналогічного висновку відносно методу оберненої матриці.

Взагалі кажучи, можна вибирати для розв'язання СЛАР між точними (прямими) та наближеними (ітераційними) методами.

Точні методи, у яких використовують скінченні алгоритми, дають змогу знайти розв'язок системи після певної кількості арифметичних операцій. Прикладами їх слугують вже згадані методи Крамера, оберненої матриці, Гаусса.

Треба наголосити таке:

а) прямі методи можуть називатися точними лише умовно, оскільки в процесі їх реалізацій відбувається неконтрольоване накопичення похибок;

б) вони потребують зберігання в оперативній пам'яті комп'ютера всієї матриці системи.

Зважаючи на особливості точних методів, їх використовують на практиці або для розв'язання досить невеликих ($n \sim 10^2$) систем зі щільно заповненою матрицею та істотно відмінним від нуля визначником, або для розв'язання більших $n \sim 10^3$ систем з розрідженими (тобто такими, які містять велику кількість нульових елементів) матрицями (відповідні методи – технологія розріджених матриць – будуть розглянуті далі). Порівняльна оцінка прямих методів потребує визначення їх **якості** – кількості потрібних для реалізації арифметичних дій.

Ітераційні методи (методи послідовних наближень) використовують для систем великого порядку ($n \sim 10^3 \dots 10^6$) і полягають у реалізації деякої кількості послідовних наближень (ітерацій) $\bar{X}$, причому $\lim_{k \rightarrow \infty} \bar{X}_k = \bar{X}$ (k – номер ітерацій).

Обчислення зазвичай виконують аж до досягнення потрібної точності ε ($\|\bar{X}_k - x\| \leq \varepsilon$), при цьому якість ітераційного методу визначається кількістю ітерацій $k(\varepsilon)$, потрібних для визначення наближення, що задовольняє цю вимогу.

Ітераційні методи не потребують зберігання в оперативній пам'яті всієї матриці (запам'ятовуються лише декілька n -компонентних векторів), похибки округлень не накопичуються, проте величина $k(\varepsilon)$ може виявитись великою (повільна збіжність ітерацій).

Для прискорення збіжності ітераційного процесу застосовують, наприклад, комбіновані (гібридні) методи, серед яких є такі, які полягають в ітераційному уточненні розв'язку, одержаного за допомогою прямих методів. Це потрібно у разі так званих **погано зумовлених систем**.

Обумовленість матриці СЛАР – це міра її чутливості до збурення правої частини. Нехай R^n - множина n -вимірних векторів $\bar{x} = \{x_1, \dots, x_n\}$.

Називатимемо *нормою вектора* величину

$$\|\bar{x}\| = \|\bar{x}\|_1 = \max_{1 \leq j \leq n} |x_j|,$$

Або
$$\|\bar{x}\| = \|\bar{x}\|_2 = \left(\sum_{j=1}^n x_j^2\right)^{\frac{1}{2}}.$$

Нормою матриці
$$\|A\|_i = \sup_{x \neq 0} \frac{\|Ax\|_i}{\|x\|_i}, i = 1, 2$$

можна довести, що
$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|, \|A\|_2 \leq \left(\sum_{i,j=1}^n a_{ij}\right)^{\frac{1}{2}}.$$

Якщо A – симетрична матриця, то

$$\|A\|_2 = \max_{1 \leq i \leq n} |\lambda_i(A)|;$$

$$\|A^{-1}\|_2 = \frac{1}{\min_{1 \leq i \leq n} |\lambda_i(A)|}.$$

де $\lambda_i(A), i = 1, n$ – власні числа A .

Якщо розглядати разом системи $A\bar{x} = \bar{b}$ та $A(\bar{x} + \bar{r}) = \bar{b} + \bar{\eta}$, де $\bar{\eta}$ – похибка правої частини, $\bar{r}$ – похибка розв'язку, то, зважаючи на те, що $A\bar{r} = \bar{\eta}$, $\bar{r} = A^{-1}\bar{\eta}$ ($\det A \neq 0$), маємо:

$$\frac{\|\bar{r}\|/\|\bar{x}\|}{\|\bar{\eta}\|/\|\bar{b}\|} = \frac{\|A\bar{x}\| \cdot \|A^{-1}\bar{\eta}\|}{\|\bar{x}\| \cdot \|\bar{\eta}\|} \leq \|A\| \cdot \|A^{-1}\|.$$

Останню величину називають мірою (числом) обумовленої матриці

$$V(A) = \|A\| \cdot \|A^{-1}\| \quad (6.2)$$

Вона характеризує ступінь залежності відносної похибки розв'язку $(\|\bar{r}\| \cdot \|\bar{x}\|)$ від відносної похибки правої частини $(\|\bar{\eta}\|/\|\bar{b}\|)$.

Відзначимо таке:

1) $V(A) \geq 1$;

2) $V(A) \geq \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}$ (рівність досягається, якщо обрано $\|\cdot\|_2$ та матриця

A є симетричною);

3) $V(A \cdot B) \leq V(A) \cdot V(B)$;

4) якщо $V(A)$ – велике, то A називається погано обумовленою (можливе сильне накопичення похибок), інакше A – добре обумовлена;

5) для погано обумовлених систем «якість розв'язку» не можна визначити шляхом обчислення величини $\bar{b} - A\bar{x}$ (за малих значень норми такої величини $\bar{x}$ може значно відхилятися від точного розв'язку $\bar{x}$).

Прямі методи розв'язання СЛАР. Найпоширенішим серед алгоритмів розв'язання СЛАР є метод виключення Гаусса, який ґрунтується на можливості складати лінійні комбінації рівнянь системи.

Нехай маємо систему
$$\begin{cases} a_{11}x_1 + \dots + a_{1n}x_n = b_1; \\ \vdots \\ a_{n1}x_1 + \dots + a_{nn}x_n = b_n. \end{cases}$$

Прямий хід – перша частина алгоритму Гаусса – полягає у зведенні системи до вигляду:

$$\begin{cases} x_1 + c_{12}x_2 + \dots + c_{1n}x_n = y_1 \\ x_2 + c_{23}x_3 + \dots + c_{2n}x_n = y_2 \\ \dots \\ \dots \\ \dots x_n = y_n. \end{cases}$$

Після цього реалізується обернений хід методу Гаусса, який полягає у знаходженні невідомих x_k , $k = \overline{1, n}$ за формулами:

$$x_k = y_k - \sum_{j=i+1}^n c_{ij} x_j, i = n-1, n-2, \dots, 1; x_n = y_n.$$

Відзначимо, що основний час реалізації методу Гаусса займає його прямий хід (приблизно $\frac{n^3}{3}$ дій, обернений хід – приблизно $\frac{n^2}{2}$ дій, обчислення значень однієї з невідомих потребує за великих n – $n - \frac{n^2}{2}$ дій).

Застосовують різні алгоритми реалізації прямого методу Гаусса – зведення основної матриці системи A до верхньої трикутної матриці C :

$$C = \begin{pmatrix} 1 & C_{12} & \dots & \dots & C_{1n} \\ 0 & 1 & C_{23} & \dots & C_{2n} \\ \vdots & \ddots & & & \\ \vdots & & \ddots & & \\ 0 & & & 0 & 1 \end{pmatrix}$$

Розглянемо алгоритми **одиничного ділення**. Нехай систему зведено

до вигляду:

$$\begin{cases} x_1 + C_{12}x_2 + \dots + C_{1n}x_n = y_1 \\ x_{k-1} + C_{k-1,k}x_k + \dots + C_{k-1,n}x_n = y_{k-1} \\ a_{kk}^{(k-1)}x_k + \dots + C_{kn}^{(k-1)}x_n = b_k^{(k-1)} \\ a_{nk}^{(k-1)}x_k + \dots + C_{nn}^{(k-1)}x_n = b_n^{(k-1)} \end{cases}$$

Розділимо обидві частини k -го рівняння системи на $a_{kk}^{(k-1)}$ ($a_{kk}^{(k-1)} \neq 0$), одержимо $x_k + c_{k,k+1}x_{k+1} + \dots + c_{kn}x_n = y_k$.

Множачи це рівняння на $a_{ik}^{(k-1)}$, $i = \overline{k+1, n}$ і віднімаючи одержане з i -го рівняння системи, завершимо виконання k -го кроку прямого ходу методу Гаусса:

$$\begin{cases} x_1 + C_{12}x_2 + \dots + C_{1n}x_n = y_1 \\ \vdots \\ x_k + C_{k,k+1}x_{k+1} + \dots + C_{kn}x_n = y_k \\ b_{k+1,k+1}^{(k)}x_{k+1} + \dots + C_{k+1,n}x_n = b_{k+1}^{(k)} \\ \vdots \\ b_{n,k+1}^{(k)}x_{k+1} + \dots + b_{n,n}^{(k)}x_n = b_n^{(k)} \end{cases}$$

При цьому в оперативній пам'яті зберігаються (для здійснення оберненого ходу) значення $c_{ij}, y_i (i = \overline{1, n}, j = \overline{i+1, n})$.

Зауважимо, що, застосовуючи метод Гаусса, ми припускаємо, що $a_{kk}^{(k-1)} \neq 0, k = \overline{1, n}$. Для того щоби позбутися цього припущення, вдаються до модифікованого методу Гаусса (із переставлянням рівнянь).

Фактично прямий хід методу Гаусса являє собою реалізацію представлення $A = B \cdot C$, де матриця B – нижня трикутна, а C – описана раніше. У теорії матриць використовують такі позначення: L – нижня трикутна матриця; U – верхня трикутна матриця з одиничною діагоналлю, і доводять теорему про те, що матрицю A з ненульованими кутовими мінорами $\Delta_j, j = \overline{1, n}$ можна представити у вигляді $A = L \cdot U$ і до того ж єдиним способом.

Конструктивно таке представлення реалізується, наприклад, за допомогою алгоритмів Краута або Дулітла.

Корисність реалізації цього представлення впливає, наприклад, з того, що часто доводиться розв'язувати послідовно декілька систем, які мають однакові основні матриці, але відрізняються стовпцями вільних членів. Крім того LU - розклад не змінюється в разі транспонування основної матриці. Метод Гаусса, прямий хід якого реалізується безпосередньо за допомогою алгоритмів розкладу, називається **компактною схемою** методу Гаусса. Вилучення невідомих у методі Гаусса можна представити такою формулою:

$$L_n L_{n-1} \dots L_1 A \bar{x} = L_n L_{n-1} \dots L_1 \bar{f},$$

де

$$L_k = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & \frac{1}{a_{kk}^{(k-1)}} & \dots & \dots & 0 \\ \vdots & \frac{a_{k-1,k}^{(k-1)}}{a_{kk}^{(k-1)}} & \dots & \ddots & 1 \\ 0 & \frac{a_{nk}^{(k-1)}}{a_{kk}^{(k-1)}} & & & 1 \end{pmatrix}.$$

Якщо існує таке k , що $\Delta_k = 0$, описані вище схеми не є справедливим.

У таких випадках застосовують *схему вибору головного елемента*, яка полягає у тому, що на черговому кроці вилучається така невідома, коефіцієнт при якій (у рядку, стовпці або матриці) є максимальним (відповідно до принципу його вибору розрізняють частковий або повний вибір головного елемента). Ясно, що частковий вибір головного елемента рівносильний перенумерації невідомих (якщо обирають максимальний елемент у рядку) або рівнянь (якщо обирають максимальний коефіцієнт при цій невідомій у відповідних рівняннях). Відзначимо, що такі алгоритми значно знижують вразливість розв'язку до похибок округлення в обчисленнях.

Обчислюючи обернену до матриці A , можна скористатися таким підходом: оскільки $A \cdot A^{-1} = E$, то $\sum_{k=1}^n a_{ik} x_{kj} = \delta_{ij}$, $j = \overline{1, n}$,

x_{kj} - елементи матриці A^{-1} ,

δ_{ij} - символ Кронекера.

Це – система рівнянь відносно n^2 невідомих, але вона розпадається на n незалежних систем з однаковими основними матрицями (рівними A). При цьому їх праві частини являють собою n - вимірні вектори, у яких один з елементів одиничний, а решта – нульові. Зрозуміло, що в розв'язанні такої сукупності систем методом Гаусса його прямий хід реалізовується лише один раз (як зазначено раніше). Кількість потрібних при цьому операцій становить приблизно n^3 .

Можливим є й інший спосіб знаходження оберненої матриці – *метод Жордана*.

Він полягає в тому, що, знайшовши спочатку у першому рядку найбільший за абсолютною величиною елемент a_{1e_1} , переставивши стовпці так, щоб він опинився на місці a_1 , розділивши перший рядок на a_{1e_1} та анулювавши всі останні елементи першого стовпця, ми представляємо одержану матрицю A_1 у вигляді $A_1 = L_1 A P_1$,

$$\text{де } L_1 = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ \vdots & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ l_{n1} & 0 & \cdots & 1 \end{pmatrix};$$

P_1 – матриця перестановки стовпців.

Після виконання такої процедури n разів, одержуємо одиничну матрицю $L_n L_{n-1} \dots L_1 A P_1 P_2 \dots P_{n-1} = E$, або $CA = E$,

тобто $C = A^{-1}$, $C = P B_n$, $P = P_1 P_2 \dots P_{n-1}$, $B_n = L_n \dots L_1$.

Прямий метод розв’язання СЛАР, у якому використано таку ідею, називається методом **оптимального вилучення** (метод Жордана – Гаусса). Застосовуючи такий метод, фактично об’єднують прямий та обернений ходи – основна матриця системи відразу зводиться до одиничної.

Застосовують також інші прямі методи розв’язання СЛАР.

а) **Метод відображень**

Нехай U – матриця відображень від $(n-1)$ - вимірної площини Q . Виявляється, що добираючи матриці M_k так, щоб $A_k = M_k A$ мала у k -му стовпці всі елементи, окрім k -го, рівними нулеві, можемо їх обирати матрицями відображень. Таким чином реалізується процес, аналогічний прямому ходу методу Гаусса. Цей метод потребує $\frac{2n^3}{3}$ арифметичних операцій.

б) **Метод обертань**

$$\text{Матриця вигляду } \{t_{ij}\}_{i,j=1}^n = T_{ij} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & C & S & \vdots \\ \vdots & S & C_1 & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}, C^2 + S^2 = 1$$

$C_{ii} = C_{jj} = C$, $C_{ij} = -C_{ji} = \pm S$ називається елементарною матрицею обертань. Множачи матрицю A зліва на елементарну матрицю обертань, можна змінювати її i -й та j -й стовпці. Так реалізується знову ж таки прямий хід методу Гаусса.

Розглянуті методи потребують більше арифметичних операцій (метод відображень – удвічі, метод обертань – в чотири рази), ніж метод одиничного ділення, але мають свої позитивні властивості (уніфікація алгоритмів для будь-яких матриць – у методу відображень, стійкість – у методу обертань).

в) **Метод обведень**

Цей метод ґрунтується на тому, що СЛАР порядку $i+1$ може бути представлена у такому вигляді:

$$A_{i+1} \bar{x}_q^{(i+1)} + f_q^{(i+1)} = 0,$$

де

$$A_{i+1} = \begin{pmatrix} A_i & \vdots & a_{1,i+1} \\ \dots & \dots & \vdots \\ a_{i+1,1} & \dots & a_{i+1,i+1} \end{pmatrix} = \begin{pmatrix} A_i & \vdots & f_{i+1}^{(i)} \\ \dots & \vdots & \dots \\ V^{(i+1)} & \vdots & a_{i+1,i+1} \end{pmatrix};$$

$$\bar{f}_q^{(i)} = (a_{iq}, \dots, a_{iq})^T.$$

Можна показати, що

$$\bar{X}_q^{(i+1)} = \begin{pmatrix} \bar{x}_q^{(i)} \\ 0 \end{pmatrix} - \frac{(\bar{V}^{(i+1)}, x_q^{(i)}) + a_{i+1,q}}{(\bar{V}^{(i+1)}, x_{i+1}^{(i)}) + a_{i+1,i+1}} \begin{pmatrix} x_{i+1}^{(i)} \\ 1 \end{pmatrix};$$

$$\bar{V}^{(i)} = (a_{i1}, \dots, a_{i-1}).$$

Розв'язок системи $\bar{x} = \bar{x}_{n+1}^{(n)}$ може бути одержаний обчисленням послідовності i -вимірних векторів $\bar{x}_q^{(i)}$ $i < q \leq n+1, i = \overline{1, n}$. Така схема потребує $n(2n^2 + 9n + 1)/6$ множень та ділень і фактично еквівалентна методу оптимального вилучення. На закінчення огляду прямих методів нагадаємо, що за загальною теорією СЛАР такі системи мають розв'язок (є сумісними) тільки тоді, коли ранги основної та розширеної матриці збігаються, причому якщо ці ранги менші за кількість невідомих у системі, то вона має нескінченну кількість розв'язків (розглядаються системи, матриці яких у загальному випадку є прямокутниками – кількість рівнянь не обов'язково збігається з кількістю невідомих). Навіть, коли вказані ранги є рівними і збігаються з кількістю невідомих, як зазначалось раніше, при великих числах обумовленості матриці A система є близькою до виродженої, що потребує спеціальних методів.

З огляду на викладене висловимо такі зауваження:

1. Ранг матриці може бути визначений під час реалізації прямого ходу методу Гаусса – він дорівнює r , якщо неможливо обрати ненульовий головний елемент на $(r+1)$ кроці.

2. У випадку, коли система має нескінчену кількість розв'язків, їх списують, переносячи у праву частину ті невідомі, стовпці яких не належать базисному мінору, та розв'язуючи одержану систему відносно базисних невідомих.

3. Якщо кількість рівнянь перевищує кількість невідомих, може бути застосований, наприклад, метод **сингулярного розкладу** основної матриці A : $A=USV^T$, де U - ортогональна (квадратна) матриця, кількість рядків та стовпців якої збігається з кількістю m - рівнянь системи; V – ортогональна матриця, кількість рядків та стовпців якої збігається з кількістю n – невідомих системи; S – діагональна матриця $m \times n$ (її діагональні елементи називаються сингулярними числами A). Позначивши $\bar{Z} = V^T \bar{U}$, $\bar{d} = V^T \bar{f}$, можна записати систему $A\bar{U} = \bar{f}$ у вигляді: $S\bar{Z} = \bar{d}$.

Вектор $\bar{z}$, який мінімізує вираз $\|\bar{\delta}\| = \|S\bar{z} - \bar{d}\|$, визначається так:

$$z_i = \begin{cases} d_i / \sigma_i & \text{за } \sigma_i > \tau \\ \text{довільне} & \text{за } \sigma_i \leq \tau \end{cases}.$$

де $\bar{d} = U^T \bar{f}$.

4. Можливі підходи для розв'язання несумісних систем (несумісність виникає, наприклад, в результаті неточних значень вихідних даних), які ґрунтуються на методах **регуляризації**.

Запитання для самоконтролю

1. У чому полягає ідея метода Гаусса?
2. Формули Крамера?
3. Які методи застосовують для розв'язку систем великої розмірності?

Лекція 7. ІТЕРАЦІЙНІ МЕТОДИ РОЗВ'ЯЗАННЯ СИСТЕМ ЛІНІЙНИХ АЛГЕБРАЇЧНИХ РІВНЯНЬ

У лекції розглядаються найпростіші однокрокові (двошарові) методи, а саме метод простої ітерації і метод Зейделя. Визначено умови збіжності та похибки цих методів.

Розглянемо систему $A\bar{U} = \bar{f}$, у якій кількість рівнянь збігається з кількістю невідомих. Це означає, що вектори $\bar{U}$ (вектор невідомих) та $\bar{f}$ є елементами деякого простору R^n , матриця A визначає лінійний оператор, який діє з R^n у R^n .

Для реалізації будь-якого ітераційного методу потрібно вибрати початкове наближення $\bar{U}_0 \in R^n$. До найпростіших ітераційних методів належать **однокрокові (двошарові)** методи, під час реалізації яких для знаходження ітерації $\bar{U}_{k+1}$ використовують лише попередню ітерацію $\bar{U}_k$:

$$B \frac{\bar{U}_{k+1} - \bar{U}_k}{\tau_{k+1}} + A\bar{U}_k = \bar{f}, k = 0, 1, 2, \dots \quad (7.1)$$

При цьому $\tau_1, \tau_2, \dots, \tau_{k+1}$ – додатні числа – ітераційні параметри, B – невиворочена матриця. Якщо $B = E$, метод називається явним, інакше – неявним. Якщо $\tau = \tau_k, \forall k$, метод називається стаціонарним. Розглянемо найбільш вживані методи вказаного типу.

А. Метод простої ітерації

Нехай $\tau = \tau_k$. Схема є явною, тобто

$$\frac{\bar{U}_{k+1} - \bar{U}_k}{\tau_{k+1}} + A\bar{U}_k = \bar{f},$$

або в координатній формі:

$$U_{k+1}^{(i)} = U_k^{(i)} - \tau \left(\sum_{j=1}^n a_{ij} y_k^{(j)} - f^{(i)} \right), i = \overline{1, n}.$$

Якщо позначити $B = E - A$ і припустити, що $\|B\| < 1$, то метод простої ітерації збігається до **єдиного** розв'язку системи зі швидкістю геометричної прогресії:

$$\|\bar{U} - \bar{U}_{k+1}\| \leq \frac{\|B\|}{1 - \|B\|} \|\bar{U} - \bar{U}_{k-1}\|.$$

Необхідною і достатньою умовою збіжності (за додаткової умови, що система має єдиний розв'язок) є обмеженість згори одиницею власних чисел матриці B .

Б. Метод Зейделя

Обчислювальний алгоритм методу Зейделя може бути записаний формулою

$$\sum_{j=1}^{i-1} a_{ij} y_{k+1}^{(j)} + \sum_{j=i+1}^n a_{ij} y_k^{(j)} = f^{(i)}, i = \overline{1, n} \quad (7.2)$$

(за умови, що $a_{ii} \neq 0$, інакше $\sum_{j=1}^{i-1} a_{ij} y_k^{(j)} + \sum_{j=i}^n a_{ij} y_{k+1}^{(j)} = f^{(i)}, i = \overline{1, n}$).

Таку формулу можна переписати у вигляді

$$y_i^{(k)} = \sum_{j=1}^{i-1} \left(-\frac{a_{ij}}{a_{ii}} \right) y_j^{(k)} + \sum_{j=i+1}^n \left(-\frac{a_{ij}}{a_{ii}} \right) y_j^{k+1} + \frac{b_i}{a_{ii}} \quad (7.3).$$

Остання формула відображає той факт, що під час знаходження k -ї ітерації методом Зейделя її i -та компонента виражається через попередню знайдену першу, другу і так далі аж до $(i-1)$ -ї компоненти тієї самої ітерації (останні компоненти беруть з попередньої ітерації), а це прискорює збіжність методу.

Для того щоби метод Зейделя був збіжним, достатньо того, щоби

$$\sum_{j \neq i} |a_{ij}| < |a_{ii}|, i = \overline{1, n},$$

або щоби матриця A була симетричною, достатньо визначеною.

В. Метод верхньої релаксації

Цей метод визначається співвідношенням

$$(D + \omega A^-) \frac{\bar{U}_{k+1} - \bar{U}_k}{\omega} + A \bar{U}_k = \bar{f}, k = 0, 1, \dots \quad (7.4)$$

$$\text{де } D = \begin{pmatrix} a_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_{nn} \end{pmatrix}, A^- = \begin{pmatrix} 0 & \dots & \dots & 0 \\ a_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & \dots & a_{n,n-1} & 0 \end{pmatrix}$$

Розрахункова формула для компонент $(k+1)$ -ї ітерації має такий вигляд:

$$U_{k+1}^{(i)} = U_k^{(i)} + \frac{\omega}{a_{ii}} \left[f^{(i)} - \sum_{j=1}^{i-1} a_{ij} U_{k+1}^{(j)} - \sum_{j=i}^n a_{ij} U_k^{(j)} \right], i = \overline{1, n} \quad (7.5)$$

За $\omega=1$ цей метод являє собою метод Зейделя, умовою його збіжності (додатковою) є те, що ω має задовольнити нерівності $0 < \omega < 2$.

Так само, як метод Зейделя забезпечує прискорення збіжності порівняно з методом прискореної ітерації, метод верхньої релаксації збігається швидше, ніж метод Зейделя.

Серед інших ітераційних методів можна відзначити метод Якобі:

$$D(\bar{U}_{k+1} - \bar{U}_k) + A\bar{U}_k = \bar{f} \quad (7.6).$$

Повертаючись до загальної схеми однокрокових (двошарових) методів, зазначимо, що їх збіжність зумовлює дотримання такої умови:

A – симетрична, додатно визначена матриця, додатно визначеною є і матриця $B - \frac{\tau}{2}A$ при деякому $\tau > 0$. Якщо, наприклад, A – симетрична, додатно визначена, з діагональним переважанням:

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|,$$

то метод Якобі є збіжним.

Необхідною і достатньою умовою збіжності будь-якого стаціонарного ітераційного процесу за довільного початкового наближення $\bar{U}_0$ є така: всі власні числа матриці $S = E - \tau B^{-1}A$ мають бути менші за одиницю за абсолютною величиною.

Розглянемо тепер питання про похибку наближеного розв'язку системи (який може бути одержаний як ітераційним шляхом, так і будь-яким іншим).

Виявляється, що в більшості випадків можна перейти до розгляду **збуреної системи**:

$$A^* \bar{U}^* = \bar{f}^*, \quad (7.7)$$

де $A^* = A + \delta A$, $\bar{f}^* = \bar{f} + \delta \bar{f}$, відповідно $\bar{U}^* = \bar{U} + \delta \bar{U}$.

Якщо існує A^{-1} , $\|\delta A\| \leq \|A^{-1}\|^{-1}$, то відносна похибка може бути оцінена так:

$$\frac{\|\delta \bar{U}\|}{\|\bar{U}\|} \leq \frac{V(A)}{1 - V(A)} \frac{\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta \bar{f}\|}{\|\bar{f}\|} \right)}{\|A\|},$$

де $V(A)$ – число обумовленості матриці A .

Якщо, наприклад, права частина $\bar{f}$ не є збуреною ($\delta \bar{f} = 0$), а розв'язок виявляється наближеним внаслідок похибок округлень, то

$$\frac{\|A^* - A\|}{\|A\|} : (n2^{-k}) \xrightarrow{k \rightarrow +\infty} 0,$$

де k – кількість розрядів мантиси у двійковому представленні числа (комп'ютер з рухомою комою). Часто наближений розв'язок $\bar{U}_0$ (одержаний, скажімо, за методом Гаусса) обирають за початкове наближення у якомусь з ітераційних методів або коригують його за допомогою такої процедури.

Система має вигляд $A\bar{U} = \bar{f}$,

$\bar{U}_0$ – наближений розв’язок

$$A\bar{U}_0 = \bar{f}_0, A(\bar{U} - \bar{U}_0) = \bar{f} - \bar{f}_0 = \bar{r}.$$

Якщо позначити $\bar{\xi} = \bar{U} - \bar{U}_0$, то $A\bar{\xi} = \bar{r}$.

Розв’язавши останню систему, покладемо $\bar{U} = \bar{U}_0 + \bar{\xi}$ і не припинятимемо цей процес аж до досягнення заданої точності. Зазначимо насамкінець: якщо умови збіжності ітераційного процесу перевірити важко, відповідний обчислювальний алгоритм має обов’язково містити лічильник числа ітерації, що дає змогу уникнути непродуктивних витрат часу.

Запитання до самоконтролю

1. У чому полягає ідея методу простої ітерації?
2. У чому полягає ідея методу Зейделя?
3. Що таке збурена система?
4. У чому полягає похибка методу Зейделя
5. У чому полягає похибка методу простої ітерації?

Лекція 8. МЕТОДИ РОЗВ'ЯЗАННЯ СИСТЕМ ЛІНІЙНИХ АЛГЕБРАЇЧНИХ РІВНЯНЬ З МАТРИЦЯМИ СПЕЦІАЛЬНОГО ВИГЛЯДУ

У лекції розглянуто системи з симетричними матрицями, системи із стрічковими і з розрідженими матрицями, наведено алгоритми упорядкування симетричних матриць, а саме методи Даффа і блочної діагоналізації.

Системи із симетричними та стрічковими матрицями

Відзначимо, що використання для матриці A системи деяких додаткових умов (наприклад, симетрії відносно основної діагоналі, додатної визначеності тощо) дає змогу використовуючи спеціальні алгоритми їх розв'язання, суттєво заощаджувати час та пам'ять комп'ютера.

А. Системи з симетричними матрицями

Для зберігання в пам'яті комп'ютера елементів такої матриці потрібно удвічі менше місця ($a_{ij} = a_{ji}$), але у схемах з вибором головного елемента доводиться переставляти рядки та стовпці, що призводить до втрати симетрії.

В той же час якщо для симетричної матриці припустити додатково дотримання умови додатної визначеності, це дасть можливість застосувати, як було вказано раніше, спеціальні методи розв'язання, що сприяє згаданому ощадливому використанню пам'яті. Існують три основні необхідні умови додатної визначеності матриці:

- а) $(A\bar{x}, \bar{x}) > 0, \forall \bar{x} \neq 0$;
- б) всі власні значення A є додатними;
- в) всі основні мінори матриці A є додатними.

Однак через складність перевірки цих умов використовують, як правило, таку умову: для додатної визначеності симетричної матриці достатньо дотримання властивості діагонального переважання та додатності її діагональних елементів. Для вказаних матриць LU – має вигляд

$$A = LL^T,$$

де L – нижня трикутна матриця (розклад Холецкого). Ясно, що після знаходження матриці L розв'язок системи стає тривіальним. Методи

знаходження такої матриці (розкладу) називають методами Холецького, або методами квадратних коренів.

Якщо обчислити елементи матриці L за стовпцями, то одержимо такі співвідношення:

$$l_{11}^2 = a_{11};$$

$$l_{i1} l_{11} = a_{i1} \quad i = \overline{1, n};$$

$$l_{k1}^2 + \dots + l_{k, k-1}^2 + l_{kk}^2 = a_{kk}; \quad i = \overline{k+1, n};$$

$$l_{i1} l_{k1} + \dots + l_{i, k-1} l_{k, k-1} + l_{ik} l_{kk} = a_{ik};$$

$$\text{звідки } l_{11}^2 = \sqrt{a_{11}}, \quad l_{i1} = \frac{a_{i1}}{l_{11}}, \quad i = \overline{2, n}; \quad l_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2}, \quad l_{ik} = \frac{a_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj}}{l_{kk}}, \quad i = \overline{k+1, n}.$$

Обчислюючи ж елементи L за рядками, одержуємо

$$l_{11} = \sqrt{a_{11}}, \quad l_{k1} = \frac{a_{k1}}{l_{11}},$$

$$l_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2}, \quad l_{k, k-1} = \frac{a_{k, k-1} - \sum_{j=1}^{k-2} l_{kj} l_{k-1, j}}{l_{k-1, k-1}}, \quad i = \overline{k+1, n}.$$

Цей метод часто називають методом облямовування, тому він особливо зручний при обчисленні розкладу для матриці $L^{(n+1)}$, одержаної облямовуванням (додаванням стовпця та рядка) матриці $L^{(n)}$, для якої розклад відомий (потрібно лише $\frac{n^2}{2}$ додаткових операцій). Наведені методи є абсолютно стійкими і дають змогу досягати 50 відсоткової економії операцій порівняно із загальними методами.

Застосовують модифікацію метода Холецького, яка називається LDL^T -методом і полягає у представленні матриці A у вигляді $A = LDL^T$, де L – нижня трикутна матриця з одиничними діагональними елементами; D – діагональна (у випадку знаконевизначеності), A – блоково-діагональна матриця). Цей метод може бути використаний для знаконевизначених матриць A , у яких всі кутові мінори до $(n-1)$ -го порядку включно не дорівнюють нулеві (дотримання такої умови можна досягти, наприклад, виконанням умови діагонального переважання). Реалізовувати цей метод можна аналогічно до попереднього «за рядками» або «по стовпцях». Методами **квадратного кореня** називають також методи, які ґрунтують на представленні симетричної (ермітової – у випадку комплексних елементів) матриці A у вигляді $A = S^* Q S$,

де S – верхня трикутна матриця з додатними елементами головної діагоналі.

S^* – транспонована (комплексно спряжена – у разі комплексних елементів) до матриці S , Q – діагональна матриця, на діагоналі якої стоять числа, рівні ± 1 .

Б. Системи із стрічковими матрицями.

Називатимемо матрицю **стрічковою**, якщо в неї відмінні від нуля лише елементи головної та кількох сусідніх діагоналей, точніше: існують числа β_1, β_2 – натуральні та такі, що містяться між одиницею та $(n-1)$; $a_{ij} = 0$; $i, j: i - j > \beta_1, j - i > \beta_2$ (такі матриці називаються стрічковими зі стрічкою верхньої півширини β_2).

Якщо $\beta_1 = \beta_2 = 1$, матриця називається тридіагональною, $\beta_1 = \beta_2 = n-1$ будь-яка матриця підходить під таке означення, але найцікавішими, звичайно, випадки, коли β_1 та β_2 «близькі» до одиниці. Можна відзначити, що під час розв’язування таких систем досягається значна економія пам’яті (замість n^2 машинних слів потрібно лише $n(\beta_1 + \beta_2 + 1)$ для запису матриці). У разі використання прямого ходу методу Гаусса без вибору головного елемента одержані матриці залишатимуться стрічковими, у разі використання LV – розкладу L та V – також стрічкові.

Розглянемо «тридіагональну» систему:

$$\begin{aligned} A_j Z_{j-1} - C_j Z_j + B_j Z_{j+1} &= F_j; & j = \overline{1, N-1}; \\ Z_0 &= K_0 Z_1 + V_0, \\ Z_N &= K_N Z_{N-1} + V_N \end{aligned}$$

де $Z_0, Z_1, \dots, Z_N$ – невідомі; $A_{(\cdot)}, B_{(\cdot)}, C_{(\cdot)}, K_{(\cdot)}, F_{(\cdot)}, V_{(\cdot)}$ – коефіцієнти.

Система може бути записана у вигляді

$$\begin{pmatrix} 1 & -K_0 & 0 & 0 & 0 \\ A_1 & -C_1 & B_1 & 0 & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & A_{N-1} & -C_{N-1} & B_{N-1} \\ 0 & 0 & 0 & -K_N & 1 \end{pmatrix} \begin{pmatrix} Z_0 \\ \vdots \\ \vdots \\ Z_N \end{pmatrix} = \begin{pmatrix} V_0 \\ F_1 \\ \vdots \\ F_{N-1} \\ V_N \end{pmatrix}.$$

Варіант методу Гаусса, який застосовується у такому випадку, має назву методу **прогонки**. Щодо єдності розв’язку та стійкості обчислювального алгоритму, то їх можна досягти, наприклад, за допомогою діагонального переважання:

$$\begin{cases} |C_j| \geq |A_j| + |B_j| \\ |A_j| > 0 \\ |K_0| < 1, \quad |K_N| < 1 \end{cases} \quad j = \overline{1, N-1}.$$

Покажемо, як практично реалізувати метод прогонки, підставляючи $Z_0 = K_0 Z_1 + V_0$ у друге рівняння. Маємо.

$A_1(K_0 Z_1 + V_0) - C_1 Z_1 + B_1 Z_2 = F_1$, тобто

$$Z_1 = K_1 Z_2 + V_1; \quad K_1 = \frac{B_1}{C_1 - A_1 K_0}; \quad V_1 = \frac{A_1 V_0 - F_1}{C_1 - A_1 K_0}.$$

Продовжуючи цей процес, матимемо:

$$Z_K = K_K Z_{K+1} + V_K,$$

$$\text{де } K_K = \frac{B_K}{C_K - A_K K_{K-1}}; \quad V_K = \frac{A_K V_{K-1} - F_K}{C_K - A_K K_{K-1}}.$$

Останнє рівняння набуде такого вигляду: $Z_N = K_N(K_{N-1} Z_N + V_{N-1}) + V_N$,

звідки $Z_N = \frac{V_N + K_N V_{N-1}}{1 - K_N K_{N-1}}$, після чого завершується прямий хід методу

прогонки и починається обернений – обчислення значень невідомих $Z_{N-1}, Z_{N-2}, \dots$. При цьому витрачається $8(N+1) - 9$ арифметичних дій (порівняно з $\frac{n^3}{3}$ у класичній схемі методу Гаусса!).

СИСТЕМИ З РОЗРІДЖЕНИМИ МАТРИЦЯМИ

Для розріджених (тобто таких, що містять велику кількість нульових елементів) матриць виникає можливість істотно заощадити об'єм пам'яті, який використовується в процесі розв'язання. Для розроблення конкретної схеми зберігання матриці користуються її **портретом** (формою розміщення ненульових елементів). Для розглянутих раніше стрічкових матриць, користуються наприклад, **діагональною схемою зберігання**. Це означає, наприклад, тридіагональну матрицю розмірності $n \times 3$ і т. д. Ще простіше зберігати симетричну стрічкову матрицю: достатньо зберігати її півстрічку. Для розрідженої матриці, портрет якої не виявляє певної закономірності, використовують більш складні схеми зберігання, що узагальнено називаються **розрідженим форматом** – вибір схеми зберігання визначається не портретом матриці, а наперед зумовлений її виглядом після упорядкування.

Принципово важливим для розріджених матриць є поняття **заповнення**, тобто переходу нульових елементів у ненульові в процесі виконання прямого ходу методу Гаусса.

Приклад.

Матриця A розмірності $n \times n$ має такий вигляд:

$$A = \begin{pmatrix} X & X & & \dots & X \\ X & X & 0 & 0 & 0 \\ \vdots & 0 & X & 0 & 0 \\ & 0 & 0 & \ddots & 0 \\ X & 0 & 0 & 0 & X \end{pmatrix}.$$

Кількість ненульових елементів цієї матриці дорівнює $n + (n-1) + (n-1) = 3n - 2$. Але вже перший крок методу Гаусса веде до заповнення другого рядка і т. д. У той же час якщо до початку вилучення поміняти місцями перший та останній рядки, а також перший та останній стовпці, то матриця набуде вигляду.

$$B = \begin{pmatrix} X & 0 & 0 & X \\ 0 & \ddots & 0 & X \\ 0 & 0 & & \vdots \\ X & X & \dots & X \end{pmatrix}$$

і заповнення не відбудеться.

Основна задача технології розріджених матриць полягає у знаходженні такого порядку рядків та стовпців, який, в міру можливості, дає можливість уникнути (або хоча б зменшити) заповнення в процесі реалізації методу Гаусса. Відзначимо, що зменшення заповнення сприяє зниженню вимог до обсягу пам'яті та використанню «швидких» алгоритмів. Задача мінімізації заповнення або не має розв'язку (для асиметричних матриць) або потребує складних методів його побудови (що фактично зводить нанівець вигоду від такої «мінімізації»).

Для подальшого вивчення розріджених матриць знадобляться деякі поняття теорії графів.

Граф $G = (V, E)$ складається з множини вершин V та множини ребер E , кожне з яких можна зіставити із парою різних вершин з V . Вершини графа нумерують (від 1 до n), з кожною з них пов'язується відповідний рядок матриці. Вершини зв'язані ребром (i, j) тоді і тільки тоді, коли $a_{ij} \neq 0$.

Для несиметричних матриць кожному з ребер надається напрям (ребро (i, j) відрізняється від (j, i) , може статися так, що одне з них існує, а другого немає), у такому разі граф є орієнтованим – ортографом. Симетричним матрицям є відповідними неорієнтовані графи. Важлива властивість графа полягає в тому, що коли матриця A шляхом переставляння рядків стовпців зведена до матриці B , то відповідний їй граф змінює лише нумерацію вершин. У пам'яті комп'ютера зручно зберігати граф за допомогою опису вершин, суміжних до кожної з його вершин. *Списки суміжності* кожної вершини зберігаються у цілочисельному масиві LIST. Компактне зберігання списків суміжності є ідентичним **рядковій** схемі зберігання портрета розрідженої матриці, асоційованої з цим графом (розріджений рядковий формат).

Існують також інші схеми (наприклад, Кнута, Ларкума), описані у тій самій роботі. Для зберігання розріджених трикутних матриць використовують схему Шермена.

Взагалі кажучи, сучасні програми, у яких використовуються розріджені матриці, містять у собі такі сегменти:

- 1) *сервісні підпрограми* – перетворення вхідної інформації (інтерфейсні), видачі попереджувальних повідомлень, статистичні (оцінка необхідних об'ємів пам'яті та числа операцій);
- 2) *підпрограми символного етапу* – реалізація перестановок елементів матриці;
- 3) *підпрограми трикутного розкладу* – дія їх впливає з назви;
- 4) *підпрограми чисельного розв'язку* – остаточний етап програми.

Алгоритми упорядкування симетричних матриць

Надалі називатимемо стрічковими такі алгоритми, які здійснюють «збирання» ненульових елементів ближче до головної діагоналі, що дає можливість розглядати матрицю як стрічкову. Наприклад, алгоритм Катхілл–Маккі полягає у тому, що вершини графа нумерують у відповідності до зростання їх степенів. При цьому для знаходження стартової вершини застосовують алгоритм Гіббса: у графі, відповідному матриці A , обирають довільну вершину K і будують рівневу структуру з коренем у цій вершині, після цього в останньому обирають вершину X з мінімальним степенем і знову будують рівневу структуру (тепер з коренем у X). Якщо кількість одержаних рівнів є більшою за попередню, то повторюємо цей цикл, якщо не більшою – вершина K є шуканою.

Головними перевагами стрічкових методів є простота символічного стану та компактність схеми зберігання. В той же час вони можуть виявитися неефективними внаслідок того, що ширина стрічки визначається найвіддаленішим від діагоналі ненульовим елементом. Така «штучна сталість» ширини стрічки зумовлює необхідність зберігання у пам'яті великої кількості нульових елементів. Оскільки алгоритм Гаусса дає змогу зберігати елементи поза стрічкою нульовими, то для уникнення цього недоліку запроваджують **профільні методи** (стрічка матриці має змінну ширину). Для симетричних матриць це означає, що оболонка трикутної матриці (множника Холецкого матриці A) збігається з оболонкою самої матриці A . Згадані методи мають більш складну (порівняно із стрічковими) схему зберігання - схему Дженкінса: всі елементи оболонки, упорядковані за рядками, зберігаються в одновимірному масиві, довжина якого дорівнює сумі профіля та порядку матриці (іноді діагональні елементи зберігаються у кінці рядка, інакше – формують окремий масив).

Приклад.

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 8 & 9 & 0 & 0 & 0 \\ 0 & 8 & 3 & 0 & 0 & 0 & 0 \\ 0 & 9 & 0 & 4 & 10 & 0 & 0 \\ 0 & 0 & 0 & 10 & 5 & 11 & 12 \\ 0 & 0 & 0 & 0 & 11 & 6 & 0 \\ 0 & 0 & 0 & 0 & 12 & 0 & 7 \end{pmatrix}.$$

Профільна схема зберігання

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	2	8	3	9	0	4	10	5	11	6	12	0	7
1	1	2	4	7	9	11	14							

Більш докладно ці питання висвітлено у роботі [11, §4.6], де розглянуто також алгоритми використання блокового вигляду матриці (алгоритм Джорджа – стор. 149 – ґрунтується на деревоподібному поділі графа матриці, схеми «блок у блочному стовпці», гіперматричну та надрозріджену – стор. 38-41). Існують алгоритми вкладених та паралельних перерізів, які спираються на розподільники графа, що дає змогу розкласти граф на зв'язні компоненти.

У впорядкуванні додатно визначених матриць головна увага приділяється локальній мінімізації заповнення та кількості операцій на кожному кроці вилучення за рахунок вибору головного елемента в тому рядку та стовпці, які забезпечують внесення найменшої кількості ненульових елементів у трикутні множники.

На символічному етапі імітується хід методу Гаусса з вибором головного елемента, який підпорядкований вимозі зменшення заповнення на кожному кроці, і виконується за правилом: головним обирають рядок активної підматриці, який містить мінімальну кількість ненульових елементів. Зазначимо, що такий алгоритм, який називається алгоритмом мінімального степеня не забезпечує глобального мінімуму заповнення (теоретично існують і кращі), але на практиці дає високі результати. Пакет підпрограм SPARSPAK, призначений для розв'язання симетричних, додатно визначених, розріджених систем, описаний у роботі [1].

Зазначимо, що додатно визначеність матриці гарантує стійкість методу вилучення у виборі головних елементів на головній діагоналі. Та якщо матриця симетрична, але не є знаковизначеною, то стає потрібним застосування методів вибору головного елемента (при цьому коли немає знаковизначеності - нехай навіть і симетричної матриці, використання загальних методів має перевагу перед використанням методів, спрямованих на дотримання симетрії). Якщо ухвалюється рішення про симетрію, то для цього можуть бути використані інші методи.

А. Метод Даффа – представлення матриці у вигляді різниці симетричної додатновизначеної M та симетричної N і розв'язання системи на кожному кроці за допомогою ітераційного процесу $M_{\bar{x}}^{(K-1)} = N_{\bar{x}}^K + \bar{b}$,

де $\bar{b}$ стовпець правої частини системи.

Такий метод застосовують, коли матриця близька до додатно визначеної.

В. Метод блочної діагоналізації обирають, коли не дотримано попередньої умови. До найбільш поширених стратегій вибору головного елемента (з погляду стійкості обчислювального процесу!) належать такі:

А. Діагональний вибір головного елемента – застосовують у разі додатно визначених матриць.

В. Стратегія часткового вибору головного елемента – вибір здійснюють серед елементів активної підматриці (за принципом найбільшого модуля).

С. Стратегія Ріда – порогів вибір – полягає у виборі деякого числа (найчастіше покладають $U = 0,25$), після чого обмежують множину можливих головних елементів тими елементами матриці, які не менші за $U \cdot M$, де M - максимум модуля її елемента. При цьому вибір $U = 1$ є відповідним стратегії часткового вибору головного елемента.

Д. Стратегія Златєва – використовують як введений у попередньому методі параметр допустимості U , так й обмеження на кількість рядків активної матриці, що проглядаються на кожному кроці вилучення.

Несиметричні системи

Для таких систем на символічному етапі доводиться розв'язувати одночасно дві задачі: зберігати розрідженість і зберігати достатню величину головних елементів (зазначимо, що у випадку діагонального переважання використовують методи, розроблені для симетричних, додатно визначених матриць). До блочно-трикутного вигляду матриця може бути зведена за допомогою алгоритмів Сарджента–Уестерберга (Тар'яна) [11, с.218-223], Холла, Хопкрофта–Карпа [11, с.210].

Відзначимо, що на відміну від симетричного (особливо додатно визначеного) випадку несиметричні системи не дають можливості для визначення величини заповнення попередньо (або навіть у процесі чисельного розкладу), тому місце раніше розглянутих статичних структур даних займають динамічні, які дають змогу просто здійснювати запис ненульових елементів, які виникають у процесі обчислень, та, відповідно, видалення інформації, що стала непотрібною. Серед таких структур даних назовемо пов'язані списки (наприклад, коли для кожного елемента матриці a_{ij} одночасно вказуються адреси запису наступних елементів i -го рядка та j -го стовбця), розріджені формати (розглянуті раніше), схеми Кнута та КРМ.

Стратегія вибору головного елемента для несиметричних матриць фактично є варіантами стратегії Марковіца – зменшення локального заповнення, що досягається вибором на головні тих елементів, які мають мінімальні ціни Марковіца $M_{ij} = (r_i - 1)(c_j - 1)$, де r_i, c_j - кількість ненульових елементів відповідного рядка та стовбця.

Раніше розглядалися ітераційні методи уточнення знайденого якимсь чином наближеного розв'язку системи. У роботі [14] був запропонований так званий «датський» метод уточнення, спеціально адаптований до розріджених матриць.

Нехай маємо розклад

$$A = LV$$

(точніше, внаслідок похибок округлення $LV = A + F$), ітераційний процес уточнення наближеного розв'язку x_1 системи $AX = b$ здійснюється за формулами:

$$\begin{aligned} \bar{r}_m &= \bar{b} - A\bar{x}_m \\ L(Vl_m) &= r_m \\ X_{m+1} &= X_m + l_m \end{aligned} \quad (8.1)$$

Ідея підходу полягає у встановленні «бар'єрного значення» $K > 0$, такого, що елементи матриці L та V , менші за нього (за модулем!), покладаються рівними нулеві. При цьому зростає матриця F , так що доводиться контролювати умову збіжності ітераційного процесу:

$$\|A^{-1}\| \|F\| \leq \sigma < \frac{1}{2}.$$

Як завжди, ітераційний процес закінчується або в результаті виконання запланованої кількості ітерацій або якщо послідовні ітерації стабілізуються навколо деякого вектора.

До недоліків «датського» підходу належать такі:

- 1) неможливість його застосування у разі погано обумовленої матриці A ;
- 2) неможливість його застосування у разі великої розмірності та обмеженого обсягу пам'яті;
- 3) відсутність реального виграшу за малозаповнених A .

Запитання для самоконтролю

1. Дайте означення симетричної матриці.
2. Наведіть означення стрічкової матриці.
3. У чому полягає ідея методу прогонки
4. У чому полягає основна задача технології розріджених матриць?
5. У чому полягає ідея методу Даффа

Лекція 9. ПРОБЛЕМА ВЛАСНИХ ЗНАЧЕНЬ ТА ВЛАСНИХ ФУНКЦІЙ

У лекції розглядається часткова проблема власних значень (степеневий метод, метод скалярних добутків, обернено-степеневий метод та метод лінеаризації).

Потреба розв'язувати задачі із знаходження власних чисел та власних векторів (функцій) матриць (операторів) пов'язана не тільки з їх використанням в алгоритмах обчислювальної математики, (наприклад, число зумовленості симетричної матриці визначається за допомогою її власних чисел), а й безпосередньо з багатьма науково-технічними задачами. Виділяють **повну проблему власних значень** – задачу знаходження всіх власних чисел (часто також і всіх власних векторів) матриці та **часткову проблему власних значень** – відповідно задачу знаходження одного або кількох власних чисел.

Часткова проблема власних значень

Зазначимо, що найчастіше доводиться стикатися із задачами знаходження максимального, близького до заданого числа, мінімального (наприклад у разі застосування скінченно-різницевого методів для звичайних диференціальних рівнянь) власного числа. Основним методом, який застосовують у таких випадках, є ітераційний. Помічено, що швидкість його збіжності визначається переважно вдалим (чи невдалим) вибором нульового наближення. Розглянемо тепер конкретні приклади таких методів.

А. Степеневий метод

Застосовують цей метод для знаходження максимального за модулем власного числа квадратної матриці A . Вважатимемо, що власні числа матриці упорядковані у порядку спадання їх модулів:

$$|\lambda_1| > |\lambda_2| \geq \dots |\lambda_n|.$$

Нехай початкове наближення $\bar{y}^{(0)} = (y_1^{(0)}, \dots, y_n^{(0)})$, ітераційний процес реалізується за формулою:

$$\bar{y}^{(k)} = A \bar{y}^{(k-1)} = A^k \bar{y}^{(0)}. \quad (9.1)$$

При цьому, якщо система власних векторів є повною, то, поклавши $\bar{y}^{(0)} = \sum_{j=1}^n \alpha_j \bar{x}_j$ (розклавши початкове наближення по базису власних векторів), маємо:

$$\bar{y}^{(k)} = A^k \bar{y}^{(0)} = \sum_{j=1}^n \alpha_j \bar{x}_j \lambda_j^k = \lambda_1^k \left[\alpha_1 \bar{x}_1 + \sum_{j=2}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \bar{x}_j \right], \quad (9.2)$$

$$\text{звідки} \quad \frac{y_s^{(k+1)}}{y_s^k} = \lambda_1 + 0 \left(\left(\frac{\lambda_2}{\lambda_1} \right)^k \right), \quad s = \overline{1, N} \quad (9.3)$$

Зазначимо таке:

1) після обчислення λ_1 на кожній ітерації вектор $\bar{y}^{(k+1)}$ потребує нормування (інакше можливим є переповнення або зниження порядку);

2) збіжність процесу є лінійною зі швидкістю $q = \left| \frac{\lambda_2}{\lambda_1} \right|$;

3) процес збігається, навіть якщо $\alpha_1 = 0$, але це потребує великої кількості ітерацій;

4) процес збігається також і тоді, коли найближче власне значення є кратним, а елементарний дільник матриці – лінійним. В той же час якщо $\lambda_1 \neq \lambda_2, |\lambda_1| = |\lambda_2|$ або елементарний дільник є нелінійним (клітина Жордана розмірністю, більшою за одиницю, наприклад), процес є, як правило, розбіжним;

5) найраціональніше використовувати цей процес для розріджених матриць (істотно зменшується кількість операцій).

Б. Метод скалярних добутків

Нехай

$$\bar{y}^{(0)} = \alpha_1 \bar{x}_1 + \dots + \alpha_n \bar{x}_n;$$

$$\bar{y}^{(k)} = A \bar{y}^{(k-1)} = \lambda_1^k (a_1 \bar{x}_1 + \bar{\eta}_k),$$

тоді

$$(\bar{y}^{(k)}, \bar{y}^{(k-1)}) = \lambda_1^{2k-1} (\alpha_1 \bar{x}_1 + \bar{\eta}_k, \alpha_1 \bar{x}_1 + \bar{\eta}_{k-1}) = \lambda_1^{2k-1} [\alpha_1^2 (\bar{x}_1, \bar{x}_1) + \alpha_1 (\bar{x}_1, \bar{\eta}_{k-1}) + \alpha_1 (\bar{\eta}_k, \bar{x}_1) + (\bar{\eta}_k, \bar{\eta}_{k-1})].$$

Оскільки з огляду на нерівність Коші-Буняковського

$$|(\bar{x}_1, \bar{\eta}_i)| \leq \|\bar{x}_1\| \|\bar{\eta}_i\| = \|\bar{\eta}_i\|;$$

$$|(\bar{\eta}_k, \bar{\eta}_{k-1})| \leq \|\bar{\eta}_k\| \|\bar{\eta}_{k-1}\|,$$

то

$$(\bar{y}^{(k)}, \bar{y}^{(k-1)}) = \lambda_1^{2k-1} \left[\alpha_1^2 + 0 \left(\left| \frac{\lambda_2}{\lambda_1} \right| \right)^{k-1} \right];$$

$$(\bar{y}^{(m)}, \bar{y}^{(m)}) = \lambda_1^{2m} \left[\alpha_1^2 + 0 \left(\frac{\lambda_2}{\lambda_1} \right)^m \right];$$

$$\lambda_1^{(k)} = \frac{(\bar{y}^{(k)}, \bar{y}^{(k-1)})}{(\bar{y}^{(k-1)}, \bar{y}^{(k-1)})} = \lambda_1 + 0 \left[\left(\frac{\lambda_2}{\lambda_1} \right)^{k-1} \right],$$

при цьому

$$\bar{x}_1^{-(k)} = \frac{\bar{y}^{-(k)}}{\|\bar{y}^{-(k)}\|} = [\operatorname{sgn} \lambda_1] (\operatorname{sgn} \alpha_1) \bar{x}_1 + 0(\mu_2 |^k); \quad (9.4)$$

$$\|\bar{y}^{-(k)}\| = (\bar{y}^{-(k)}, \bar{y}^{-(k)}) = |\lambda_1|^k \left[|\alpha_1| + 0(\mu_2 |^k) \right].$$

Наведені формули дають змогу визначати максимальне за модулем власне число λ_1 та відповідний йому власний вектор $\bar{x}_1$ з будь якою точністю.

Зазначимо такі:

1) за $|\lambda_1| < 1, \|\bar{y}^{-(k)}\| \rightarrow 0$ за $k \rightarrow \infty$, що може призвести до втрати інформації, відповідно за $|\lambda_1| > 1, \|\bar{y}^{-(k)}\| \rightarrow \infty$, що може призвести до переповнення порядку, тому доцільно використовувати таку модифікацію методу:

$$\bar{x}_1^{-0} = \frac{\bar{y}^{-0}}{\|\bar{y}^{-0}\|}, \bar{y}^{-(k)} = A \bar{x}_1^{-(k-1)}; \quad (9.5)$$

$$\lambda_1^{(k)} = (\bar{y}^{-(k)}, \bar{x}_1^{-(k-1)}); \quad (9.6)$$

$$\bar{x}_1^{-(k)} = \frac{\bar{y}^{-(k)}}{\|\bar{y}^{-(k)}\|}. \quad (9.7)$$

2) збіжність до одного і того самого (з точністю до знака) власного вектора за зміни $\bar{y}^{-(0)}$ означає, що, по-перше, λ_1 не є кратним, і, по-друге, що - λ_1 не є власним числом;

3) очевидно, за допомогою наведеного методу можна знайти і друге максимальне за модулем власне число λ_2 , і відповідний йому вектор $\bar{x}_2$; дійсно, якщо відомі $\lambda_1, \bar{x}_1 (\|\bar{x}_1\| = 1) \lambda_1 > |\lambda_2| > \dots > |\lambda_n|$, то, маючи ортонормований базис з власних векторів $\bar{x}_1, \dots, \bar{x}_n$, можна запровадити такий ітераційний процес $\bar{Z}^{(0)} = \bar{y}^{(0)} - (\bar{y}^{(0)}, \bar{x}_1) \bar{x}_1$ (початкове наближення має задовольняти умові $(\bar{y}^{(0)}, \bar{x}_2) \neq 0$):

$$\bar{Z}^{(k)} = \bar{y}^{(k)} - (\bar{y}^{(k)}, \bar{x}_1) \bar{x}_1;$$

$$\bar{x}_2^{-(k)} = \frac{\bar{Z}^{(k)}}{\|\bar{Z}^{(k)}\|};$$

звідки

$$\lambda_2^{(k)} = \lambda_2 + 0 \left(\left| \frac{\lambda_3}{\lambda_2} \right|^{2k-1} \right) \quad \bar{x}_2^{(k)} = (\text{sign} \lambda_2)^k \bar{x}_2' + \bar{\gamma}_k,$$

де $\bar{x}_2' \in$ або $\bar{x}_2$, або $-\bar{x}_2$

$$\|\bar{\gamma}\| = 0 \left(\left| \frac{\lambda_3}{\lambda_2} \right|^{2k} \right).$$

Задачу знаходження найменшого та найбільшого (не за абсолютною величиною!) власних чисел матриці A ефективно розв'язується у випадку, коли з власних векторів можна сформувати ортонормований базис (нагадаємо, що таке можливо, наприклад, якщо A є симетричною).

Відомо, що, у випадку, коли $B = a_0 E + a_1 A + a_2 A^2$ (тут E – одинична матриця),

$$\lambda_i(B) = P_2(\lambda_i(A)),$$

де $P_2(t) = a_0 + a_1 t + a_2 t^2$.

Отже, для розв'язання задачі може бути застосований такий алгоритм:

1) знайти (наприклад, за методом скалярних добутків) максимальне за модулем власне значення $\bar{\lambda}(A)$;

2) аналогічно знайти найбільше за модулем власне число матриці

$$B = A - \bar{\lambda}(A)E,$$

яке позначити $\bar{\lambda}(B)$;

3) якщо $\bar{\lambda}(A) > 0$, то $\max \lambda_i(A) = \bar{\lambda}(A)$, $\min \lambda_i(A) = \bar{\lambda}(A) + \bar{\lambda}(B)$;

$$\lambda_i(B) = \lambda_i(A) - \bar{\lambda}(A) \leq 0, \forall i;$$

$$\bar{\lambda}(B) = \min [\lambda_i(A) - \bar{\lambda}(A)],$$

та якщо $\bar{\lambda}(A) < 0$, $\min \lambda_i(A) = \bar{\lambda}(A)$, $\max \lambda_i(A) = \bar{\lambda}(A) + \bar{\lambda}(B)$.

Оскільки на практиці зазвичай невідомо, чи матриця має максимальні за модулем та протилежні за знаками власні значення, чи кратне максимальне за модулем власне значення, потрібно, як вже зазначено, обчислювати значення $\lambda_1^{(k)}$ за різних початкових векторів $\bar{y}^{(0)}$. Якщо при цьому проявляється збіжність до різних значень, то це свідчить про наявність максимальних за модулем власних значень з протилежними знаками. Після цього «зміщують спектр», тобто розглядають матрицю $A' = A + \delta E$, $\delta > 0$ і повторюють обчислення. Якщо при цьому маємо збіжність $\lambda_1^{(k)}$ до певного числа та в той же час послідовність $\bar{x}_1^{(k)}$ збігається до неколінеарних векторів $\bar{x}_1' = (\text{sign} \alpha_1) \bar{x}_1$, то це свідчить про кратність максимального за модулем власного значення.

В. Метод лінеаризації

Якщо $A\bar{x} = \lambda\bar{x}$, то $F_i(\bar{x}, \lambda) = \sum_{k=1}^n a_{ik}x_k - \lambda x_i = 0$

$i = \overline{1, n}$ – система нелінійних рівнянь. Надавши невідомим x_i, λ приростів $\Delta x_i, \Delta \lambda$ відповідно і лінеаризувавши попереднє співвідношення, маємо

$$\sum_{k=1}^n a_{ik} \Delta x_k - \lambda^{(s)} \Delta x_i - x_i^{(s)} \Delta \lambda = -F_i(\bar{x}^{(s)}, \lambda^{(s)}), \quad i = \overline{1, n} \quad (9.8)$$

(тут S - номер ітерації). Для замкненості доповнимо попередню систему співвідношенням $\Delta x_i = 0$.

Матриця одержаної лінійної системи є близькою до матриці A (додається один ненульовий стовпець). Зазначимо таке:

- за A – загального вигляду кожна ітерація потребує $2/3n^2$ арифметичних дій;

- за A – близькій до трикутної, відповідно - $2n^2$;

- за $A - \beta$ – стрічковій - $\frac{\beta^2 n}{2}$

(β - ширина стрічки). Як правило, за вдало обраному початковому наближенні процес збігається досить швидко (3-5 ітерацій дають прийнятну точність), якщо кількість ітерацій за відсутньої тенденції збіжності досягає 10, то змінюють початкове наближення.

Г. Обернено-степеневий метод

Цей метод полягає у застосуванні ітераційної формули

$$\bar{x}^{-(s+1)} = A\bar{x}^{-(s)} \quad (9.9)$$

Він збігається досить повільно (як і степеневий), проте існує можливість його покращення за допомогою **операції зсуву**:

Якщо відоме (наближено) одне з власних чисел λ_i , то розглядають ітерації $(A - \tilde{\lambda}_i E)\bar{x}^{-(s+1)} = \bar{x}^{-(s)}$, які збігаються досить швидко, визначаючи власне число $\lambda_i - \tilde{\lambda}_i$ (воно значно менше за модулем, ніж останні). При цьому, як і раніше, одержані ітерації слід нормувати. Під час уточнення зсуву (заміни в ітераційному співвідношенні $\tilde{\lambda}_i$ на $\tilde{\lambda}_i^{(s)}$) досягається квадратична збіжність.

Обернені ітерації найбільш ефективні за A - майже трикутній. В той же час у разі їх застосування слід уникати надто великої близькості $\tilde{\lambda}_i$ з шуканим власним значенням, оскільки це погіршує зумовленість системи.

Запитання для самоконтролю

1. У чому полягає часткова проблема власних значень?
2. У чому полягає ідея методу скалярних добутків?
3. Розкрийте ідею методу лінеаризації.

Лекція 10. ПОВНА ПРОБЛЕМА ВЛАСНИХ ЗНАЧЕНЬ ТА ВЛАСНИХ ФУНКЦІЙ

У лекції розглядається повна проблема власних значень, яка полягає у визначенні коефіцієнтів характеристичного многочлена (метод Данілевського, метод інтерполяції, метод оберненої ітерації).

Зазначимо, що розв'язання такої задачі зводиться фактично до знаходження коренів характеристичного многочлена, коефіцієнти якого відзначаються складністю обчислення, можливою є також наявність кратних його коренів. Через це більшість методів розв'язання «повної проблеми» полягає у визначенні коефіцієнтів характеристичного многочлена найраціональнішим способом.

А. Методи побудови характеристичного многочлена

Відомо, що для матриці Фробеніуса

$$\Phi = \begin{pmatrix} P_1 & P_2 & \dots & \dots & P_n \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \dots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix}$$

характеристичний многочлен має вигляд

$$P^n(\lambda) = (-1)^n (\lambda^n - p_1 \lambda^{n-1} - \dots - p_{n-1} \lambda - p_n) \quad (10.1)$$

Матриця A може бути зведена до «фробеніусової» за допомогою перетворення подібності $\Phi = C^{-1}AC$.

Метод Данілевського побудови C полягає у послідовному знаходженні $(n-1)$ перетворень подібності, які послідовно зводять до «фробеніусової» форми рядки A , (починаючи з останнього). Перше з цих перетворень реалізується матрицею

$$M_{n-1} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ -\frac{a_{n1}}{a_{n,n-1}} & \dots & -\frac{a_{nn}}{a_{n,n-1}} \\ 0 & \dots & 1 \end{pmatrix},$$

друге – матрицею

$$M_{n-2} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ (1) & \dots & \dots & (1) \\ -\frac{a_{n-1,1}}{a_{n-1,n-2}} & \dots & -\frac{a_{n-1,n}}{a_{n-1,n-2}} \\ a_{n-1,n-2} & \dots & a_{n-1,n-2} \\ 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

де $a_{ij}^{(1)}$ – коефіцієнти матриці, на яку одержано внаслідок застосування попереднього перетворення подібності і так далі. Зазначимо, що одержання на якомусь кроці реалізації цього методу елемента $a_{k,k-1}^{(j)}$ рівного нулеві, не веде до зупинки реалізації методу. В такому разі можна або переставляти стовпці (якщо існує елемент $a_{k,i}^{(j)} \neq 0 \quad i < k-1$) або, привести до канонічного вигляду верхню ліву матрицю, визначену цим рядком. При цьому характеристичний многочлен являтиме собою добуток характеристичних многочленів лівої (вже згаданої) та нижньої правої (одержаної в результаті перетворень подібності) матриць. Слід вказати, що власні вектори «фробеніусової» матриці мають вигляд $\bar{y}_k = (\lambda_k^{n-1}, \lambda_k^{n-2}, \dots, \lambda, 1)$, а власні вектори A знаходяться за формулою $\bar{x}_k = C \bar{y}_k$.

Метод Данілевського є достатньо економічним, проте вразливим до похибок обчислень (реально застосовується за $n < 10$).

Метод інтерполяції (метод Мікеладзе) полягає у тому, що характеристичний многочлен $|A - \lambda E|$ інтерполюється шляхом обчислення його значень за $\lambda = \lambda_k$ (якщо відомі $\lambda_{\min}$ та $\lambda_{\max}$, значення λ_k бажано рівномірно розміщувати між ними). Цей метод має аналогічні переваги та недоліки, застосовується також за $n < 10$.

Щодо знаходження власного вектора $\bar{x}_i$, відповідного власному значенню λ_i , слід зазначити: безпосереднє знаходження його із рівняння

$$(A - \tilde{\lambda}_i E) \bar{x}_i = 0 \quad (10.2)$$

є безперспективним! Це пов'язано з тим, що визначник матриці $A - \tilde{\lambda}_i E$ близький до нуля, але внаслідок наближеності $\tilde{\lambda}_i$ не дорівнює нулеві, так що система має тривіальний розв'язок.

Для знаходження власних векторів на практиці часто застосовують вже згадуваний **метод оберненої ітерації**.

Нехай $\bar{x}_1, \dots, \bar{x}_n$ – лінійно незалежні власні вектори матриці A , вектор $\bar{v}$ має своїми координатами багаторозрядні псевдовипадкові числа. Розглядається система:

$$(A - \tilde{\lambda}_i E) \bar{x} = \bar{v},$$

де

$$\bar{x} = \sum_{j=1}^n \xi_j \bar{x}_j, \bar{v} = \sum_{j=1}^n \beta_j \bar{x}_j.$$

Перетворюючи систему (зважаючи на те, що $A\bar{x}_j = \lambda_j \bar{x}_j$), маємо

$$\sum_{j=1}^n (\xi_j (\lambda_j - \bar{\lambda}_j) - \beta_j) \bar{x}_j = 0,$$

звідки

$$\xi_j = \frac{\beta_j}{\lambda_j - \bar{\lambda}_j}$$

Якщо $\bar{\lambda}_i$ є достатньо близьким до λ_i , то ξ_i відрізняється своєю величиною (за модулем), тобто вектор $\bar{x}$ близький до $\bar{x}_i$. Якщо власне значення знайдене неточно або невдало дібраний вектор $\bar{e}$, застосовують ітераційне уточнення $(A - \bar{\lambda}_i E) \bar{x}^{(s)} = \bar{x}^{(s-1)}$
 $\bar{x}^{(0)} = \bar{e}$.

Наведене є відповідним випадку простого власного числа, та якщо воно виявиться кратним, то відзначатимуться своєю величиною відразу декілька коефіцієнтів ξ_j , а власний вектор являтиме собою лінійну комбінацію векторів $\bar{x}_j$. Для знаходження всіх власних векторів, відповідних власному кратному значенню, беруть відповідну кількість лінійно незалежних векторів.

Наведений метод застосовують, знову ж таки за $n < 10$.

Аналогічно, обернену ітерацію можна реалізувати із зсувом δ (тобто підставляючи матрицю $A - \delta E$). В таких випадках звичайно будують послідовність Релея, яка дає вельми швидко збіжність до власного вектора.

Розв'язуючи задачу пошуку власних значень матриці A , менших від заданого числа λ , користуються також міркуваннями:

$$\text{нехай } A - \delta M = L_\delta \Delta_\delta L_\delta^*.$$

Δ_δ - діагональна, M - додатньо визначена, тоді

$$V(\Lambda - \delta E) = V(A - \delta M) = V(\delta) \quad (10.3)$$

де

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}$$

λ_i - власні значення, $V(\Lambda - \delta E)$ - кількість власних значень, менших за δ .
 Особливості застосування такого методу висвітлено у роботі [10].

Б. Симетрична проблема власних значень

Економічні та стійкі методи знаходження всіх власних значень матриць порядку, вищого від десятого, ґрунтуються на використанні

ортогональних матриць ($F^* = F^{-1}$). Очевидно, наприклад, що для тридіагональної матриці T , до якої симетрична матриця A зводиться шляхом скінченної кількості ортогональних перетворень, власні числа та власні вектори обчислюють із значною економією часу. Зазначимо, що для даної A матриця T визначається однозначно за допомогою, наприклад, *алгоритму Ланцоша*. У разі використання зсуву (див. вище) використовують відношення Релея, *зсув за Вілкінсом*, *зсув за Ньютоном*, *зсув за Саадом* [10].

В. Розріджені задачі на власні значення

Традиційно методи знаходження власних чисел та власних векторів матриць великої розмірності орієнтуються на розріджені матриці. Часто вони ґрунтуються на знаходженні підпростору φ^m інваріантного (або близького до інваріантного) відносно матриці A . Річ у тому, що базиси інваріантних підпросторів складаються з власних векторів. Основним алгоритмом при цьому є згаданий алгоритм Ланцоша (що являє собою реалізацію процедури Релея–Рітца) та його різновиди (алгоритм Ланцоша з *переортогоаналізацією*, *блоковий стрічковий*). Всі вони розглянуті у роботі [10] так само, як методи ітерування підпростору, бісекції. У книзі [11] розглянуто методи мінімізації сліду, який для невеликих за розмірністю матриць конкурує з методом Ланцоша.

Г. Загальні зауваження до проблеми власних значень

Щодо комплексної матриці, то у роботі [11] розглянуто відповідні модифікації методів знаходження власних чисел та власних векторів. У роботі [6] запропоновано методи «симетризації» задач на власні числа та власні вектори (тобто зведення розв’язання такої задачі для несиметричної матриці для використання методів, призначених до симетричних матриць). У роботі [3] можна ознайомитись із методами розв’язання задач пошуку власних коренів або чисел пари матриць (A, M) , яка називається пучком:

$$A\bar{z} = \lambda M\bar{z}.$$

Запитання для самоконтролю

1. У чому полягає метод Данілевського?
2. У чому полягає метод інтерполяції?
3. У чому полягає метод оберненої ітерації?

СПИСОК ЛІТЕРАТУРИ

1. Баліна О.І. Конспект лекцій з дисципліни «Чисельні методи» (системи лінійних алгебраїчних рівнянь; системи алгебраїчних рівнянь із матрицями спеціального вигляду; власні значення та власні функції) / О. І. Баліна, Ю. П. Буценко, І. С. Безклубенко. – Київ: КНУБА, 1999. 38 с.
2. Чисельні методи: методичні вказівки до практичних занять / О. І. Баліна, І. С. Безклубенко, О. В. Горда, Н. І. Полтораченко. – Київ: КНУБА, 2011. 44 с.
3. Бахвалов Н.С. Численные методы / Н. С. Бахвалов, Н. П. Жидков, Г. Н. Кобельников. – М.: Бином. Лаб. знаний, 2003. – 632 с.
4. Чисельні методи. Чисельне диференціювання на інтегрування. Системи рівнянь: методичні вказівки до практичних занять / О. В. Горда, Н. І. Полтораченко. – Київ: КНУБА, 2012. – 36 с.
5. Волков Е.А. Численные методы. СПб.: Лань, 2004. 248 с.
6. Чисельні методи. Наближені числа та дії над ними, методи обробки експериментальних даних, розв'язання рівнянь з однією змінною: методичні вказівки та контрольні завдання / Л. І. Турчанінова, О. І. Баліна, І. С. Безклубенко. – Київ: КНУБА, 2002. – 40 с.

Навчальне видання

БАЛІНА Олена Іванівна,
БЕЗКЛУБЕНКО Ірина Сергіївна,
БУЦЕНКО Юрій Павлович

ЧИСЕЛЬНІ МЕТОДИ

Конспект лекцій

для студентів, які навчаються за спеціальністю
151 «Автоматизація та комп'ютерно-інтегровані технології»

Комп'ютерне верстання *А.П. Селівестрової*

Підписано до друку 16.12.2019 Формат 60 × 84 1/16

Ум. друк. арк. 3,95 Обл.-вид. арк. 4,25

Електронний документ. Вид. № 13/I-19

Видавець і виготовлювач

Київський національний університет будівництва і архітектури

Повітрофлотський проспект, 31, Київ, Україна, 03680

Свідоцтво про внесення до Державного реєстру суб'єктів
видавничої справи ДК № 808 від 13.02.2002 р.

