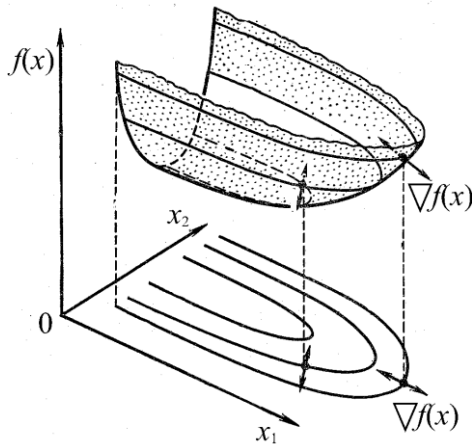


Vasile Moraru

**M E T O D E Ș I M O D E L E
D E C A L C U L**

Modulul I Metode numerice



Chișinău
2020

*Facultatea Calculatoare, Informatică și
Microelectronică
Departamentul Informatică și Ingineria Sistemelor*

Conf. univ., dr. Vasile MORARU

**M E T O D E Ș I M O D E L E
D E C A L C U L**

Modulul I Metode numerice

Chișinău
U.T.M.
2020

În lucrare sunt prezentate principalele metode de calcul numeric pentru rezolvarea unor probleme ce pot fi întâlnite frecvent în practică. Ea se adresează în primul rând studenților Universității Tehnice a Moldovei și va constitui un sprijin efectiv la predarea cursurilor de *Metode numerice*, *Cercetări operaționale*, *Matematică de calcul* ș.a. Cartea însă poate fi folosită și de toți cei care sunt preocupați de utilizarea metodelor numerice și a mijloacelor electronice de calcul la soluționarea problemelor practice.

Autor: conf. univ., dr. Vasile Moraru

© U.T.M., 2020

PREFATĂ

Metodele de calcul numeric au pătruns eficient în toate domeniile științei, tehnicii și economiei. Cursurile corespunzătoare se predau studenților de la instituțiile de învățământ superior cu profiluri tehnic și economic, matematicienilor și fizicienilor. Lor și le este adresată această lucrare, precum și celor care doresc să se inițieze în aplicarea metodelor numerice și mijloacelor electronice de calcul la rezolvarea problemelor ce se întâlnesc des în practică.

Lucrarea este structurată astfel: o introducere, cinci capitole, o anexă și își propune să prezinte principalele metode numerice de rezolvare a ecuațiilor algebrice și transcendente,

Un loc important îl ocupă rezolvarea numerică a ecuațiilor și a sistemelor de ecuații neliniare, a sistemelor de ecuații liniare și neliniare, a problemelor de optimizare necondiționată și condiționată. Se propun spre soluționare exerciții, rezolvate sau însoțite de indicațiile corespunzătoare.

Pentru înțelegerea materialului de bază sunt suficiente cunoștințele pe care studenții le-au obținut la cursurile de matematici și programare.

În *Noțiuni introductive* sunt expuse primele concepții despre metodele numerice și despre algoritmi de calcul, se fac estimări privind viteza de convergență a acestora.

Capitolul întâi este dedicat aproximării numerelor reale prin reprezentări computaționale finite cu ajutorul virgulei mobile și erorilor care le implică. Prin exemple simple se arată că erorile de rotunjire, propagându-se de la o operație aritmetică la alta, ne pot conduce la rezultate eronate.

În capitolul al doilea se prezintă metode numerice de calcul al rădăcinilor ecuațiilor algebrice și transcendente.

Este vorba de metoda înjumătățirii intervalului, metoda aproximațiilor succesive, metoda tangentei (Newton), metoda secantei și alte metode numerice. Un loc aparte îl ocupă ecuațiile algebrice, istoria cărora începe cu Evul Mediu și cu Renașterea. Se expun schema lui Horner și metoda Newton de determinare a tuturor rădăcinilor reale ale ecuațiilor algebrice.

Capitolul al treilea, destinat metodelor algebrei liniare, pune la îndemâna cititorului atât elemente de analiză matriceală, cât și algoritmi cei mai reprezentativi ce intervin în problemele de rezolvare a sistemelor de ecuații liniare și de calcul ai valorilor și vectorilor proprii. Sunt expuse metodele directe și iterative de rezolvare a sistemelor de ecuații liniare (metoda Gauss și metoda Cholesky cu factorizările sale triunghiulare, metoda Jacobi, metoda Gauss-Seidel, metode de ortogonalizare ș.a.), făcându-se totodată aprecieri asupra eficacității și stabilității numerice a acestora. Se subliniază faptul că metodele bazate pe transformări de asemănare ortogonală sunt mai eficiente decât metodele clasice de determinare a valorilor și vectorilor proprii.

În *capitolul al patrulea* se tratează metoda simplex în vederea utilizării acesteia în determinarea soluțiilor optime a problemelor de programare liniară, de programare în numere întregi, de programare liniar-fracționară. Se prezintă problema de transport și se abordează dualitatea în programarea liniară. Tot aici cititorul este familiarizat cu noțiunea de funcție convexă folosită pe larg în următorul capitol al lucrării.

În *capitolul al cincilea* se prezintă principalele metode numerice de rezolvare a sistemelor de ecuații neliniare și a problemelor de optimizare necondiționată. Problema de optimizare se poate reduce la rezolvarea unui sistem de ecuații și invers. De aceea s-a considerat necesar a expune în cadrul aceluiași capitol metodele menționate. Se

tratează metoda iterației, metoda Gauss-Seidel neliniară, metoda Newton care sunt o extindere a metodelor studiate în capitolele doi și trei. Se aduc condițiile necesare și suficiente de extrem în optimizarea necondiționată. Se face o trecere în revistă a metodei gradientului, metodei Newton-Raphson, metodelor cvasi-Newton și a metodelor de direcții conjugate, punându-se în evidență drept cea mai eficace metoda Fletcher-Powell în versiunea Polak-Ribiere.

Bibliografia include referințele în care cititorul poate afla detalii suplimentare asupra metodelor numerice de calcul prezentate în lucrare. Unele date bibliografice au fost incluse chiar în text. Menționăm că din mulțimea cărților scrise în limba română se fac trimiteri doar la acele surse pe care autorul le-a avut la îndemână în timpul scrierii lucrării.

LISTA DE NOTAȚII

R	mulțimea numerelor reale
R^n	spațiu liniar n -dimensional (3.1.1)
$\{x^{(k)}\}$	șir de vectori; $x^{(k)} \in R^n$
x^*	soluția problemei; $x^* \in R^n$
$\ x\ $	norma vectorului $x \in R^n$
$\ A\ $	norma matricei A (3.1.1)
A^T	transpusa matricei A
A^{-1}	inversa matricei A
I	matricea unitate
(x,y)	produsul scalar al vectorilor $x,y \in R^n$
$\{x P\}$	mulțimea elementelor x cu proprietatea P
$\nabla f(x)$	gradientul funcției $f(x)$.
$\nabla^2 f(x)$	matricea Hesse a funcției $f(x)$
$o(x)$	$o(x)/x \rightarrow 0$ dacă $x \rightarrow 0$.
$\mathfrak{J}(x)$	matricea Jacobi (5.2)

NOTIUNI INTRODUCTIVE

Metodele de calcul numeric au devenit în zilele noastre deosebit de importante. Ele se aplică aproape peste tot: în inginerie și în economie, în matematică și fizică, în medicină, în astronomie, în chimie, în geologie etc. Acest lucru se datorește atât progreselor obținute în domeniul calculatoarelor electronice, cât și experimentelor din ce în ce mai complicate în modelarea matematică.

Prin *metode numerice* se subînțeleg metode de rezolvare a problemelor cu ajutorul operațiilor cu caracter aritmetic – logic asupra numerelor reale, adică cu ajutorul acelor operații care pot fi executate în mod automat de un calculator electronic.

Rezolvarea unei probleme impuse de practică începe cu construcția *modelului matematic*. Modelul matematic reprezintă efectiv formularea matematică a problemei enunțate, adică constituie exprimarea matematică a relațiilor și restricțiilor dintre parametrii problemei.

După formularea matematică a problemei se realizează *alegerea metodei numerice* și se *elaborează algoritmul de calcul*. Aceste etape sunt cele mai importante în procesul soluționării problemelor. La alegerea metodei numerice se ia în considerație viteza de convergență, precizia, stabilitatea, timpul de execuție și necesarul de memorie.

Algoritmul metodei numerice constă dintr-un număr finit de operații aritmetice și logice, care trebuie să le efectueze calculatorul electronic pentru rezolvarea problemei date. Regulile de calcul alcătuiesc *pașii algoritmului*.

Subliniem faptul că noțiunea de algoritm în forma sa generală se situează printre noțiunile fundamentale ale matematicii și sta la baza programării calculatoarelor electronice. Orice algoritm este caracterizat prin următoarele proprietăți:

- *Generalitate.* Prin aceasta se înțelege că algoritmul nu trebuie să rezolve numai o problemă, ci toate problemele din clasa respectivă de probleme.
- *Finitudine.* Numărul de transformări intermediare, aplicate datelor de intrare pentru a obține datele de ieșire, este finit.
- *Unicitate.* Toate transformările intermediare trebuie să fie determinate univoc de regulile algoritmului.

După elaborarea algoritmului metodei numerice de calcul se trece la *scrierea programului* de rezolvare a problemei într-un limbaj de programare. În continuare se trece la *testarea și verificarea programului*. După testarea programului din punct de vedere sintactic, este necesar ca programul să fie verificat prin exemple de probleme concrete (probleme - test) ale căror soluții sunt cunoscute.

Prin urmare, rezolvarea unei probleme la calculator necesită parcurgerea următoarelor etape:

- I. enunțarea problemei și formularea matematică, evidențiind ce se dă și ce se cere;
- II. alegerea unei metode numerice pentru obținerea soluției;
- III. elaborarea algoritmului de calcul;
- IV. testarea și verificarea programului;
- V. analiza rezultatelor obținute.

Majoritatea metodelor numerice de calcul reprezintă în esență proceduri iterative. Aceasta înseamnă că, plecând de la un x_0 dat, se construiește un șir $x_0, x_1, \dots, x_k, \dots, x_K$ (notat de obicei prin $\{x_k\}$) care în anumite condiții converge către soluția exactă x_* a problemei considerate. Elementele x_k , $k = 0, 1, 2, \dots, K$ pot fi atât numere reale cât și vectori sau matrice.

În metodele numerice drept *criteriu de oprire (stopare)* al iterațiilor, de obicei, se folosește următorul: șirul $\{x_n\}$ se

trunchiază la un indice m determinat pe parcursul calculului în funcție de precizia dată astfel încât termenul curent x_m constituie o aproximație satisfăcătoare a soluției căutate x_* . De aceea un fapt esențial în compararea metodelor de calcul numeric îl constituie aprecierea vitezei de convergență a metodelor. În metodele de convergență rapidă sunt necesare un număr mai mic de iterații pentru a atinge precizia prescrisă decât în metodele cu convergență lentă. Dintre metodele care necesită același volum de calcul la fiecare iterație în practică se alege metoda cu convergența mai rapidă.

De obicei viteza de convergență a șirului $x_k \rightarrow x_*$ se stabilește cu ajutorul unei funcții $e(x)$, numită *funcție de estimare a erorii*, care satisface condițiilor:

$$e(x) \geq 0, \text{ pentru } \forall x \text{ și } e(x_*) = 0.$$

În practică se utilizează pe larg în calitate de funcție de estimare a erorii cantitatea

$$e(x) = |x - x_*|,$$

în cazul $x, x_* \in R$, sau

$$e(x) = \|x - x_*\|,$$

dacă metoda generează un șir de vectori sau matrice. Aici $\| \cdot \|$ este o normă vectorială sau matriceală (vezi 3.1.2).

Fie șirul $\{x_k\}$ convergent către x_* , adică

$$\lim_{k \rightarrow \infty} x_k = x_* \text{ și } \lim_{k \rightarrow \infty} e(x_k) = 0.$$

Se spune că șirul $\{x_k\}$ *converge liniar* către x_* dacă există o constantă q , $0 < q < 1$, astfel încât

$$\lim_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} = q.$$

Altfel spus, există un număr întreg $N \geq 0$ astfel că pentru orice $k \geq N$ avem

$$|x_{k+1} - x_*| \leq q \|x_k - x_*\|,$$

sau, în cazul vectorilor și matricelor,

$$\|x_{k+1} - x_*\| \leq q \|x_k - x_*\|.$$

Dacă printr-o metodă numerică de calcul se obține un șir care converge liniar atunci se spune că metoda este de *convergență liniară* sau de *ordinul întâi de convergență*. Se mai spune în acest caz că metoda *converge tot atât de repede ca și progresia geometrică de rație q* sau că *converge liniar cu rația q*.

Se spune că șirul $\{x_k\}$ *converge supraliniar* către x_* dacă

$$\lim_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} = 0.$$

Cu alte cuvinte, șirul este convergent supraliniar dacă

$$|x_{k+1} - x_*| \leq q_k |x_k - x_*|,$$

sau, în cazul vectorilor și matricelor,

$$\|x_{k+1} - x_*\| \leq q_k \|x_k - x_*\|,$$

unde $q_k \rightarrow 0$ pentru $k \rightarrow \infty$.

Convergența supraliniară înseamnă că $e(x_k) \rightarrow 0$ mai repede decât orice progresie geometrică de rație q , adică converge mai repede către zero decât orice șir de forma cq^k , unde $c > 0$ și $0 < q < 1$.

Dacă există o constantă $C \neq 0$ astfel încât

$$\lim_{k \rightarrow \infty} \frac{e(x_{k+1})}{[e(x_k)]^\beta} = C,$$

atunci se spune că șirul $\{x_k\}$ are *ordinul β de convergență*. În cazul când $\beta = 2$ sau $\beta = 3$ se spune că șirul $\{x_k\}$ este de *ordinul doi*, corespunzător de *ordinul trei de convergență*. Se mai spune că metoda numerică cu ajutorul căreia se construiește acest șir este cu *convergența pătratică*, respectiv cu *convergența cubică* sau că este o *metodă de ordinul al doilea*, respectiv *al treilea*.

Prin urmare, șirul $\{x_k\}$ are ordinul β de convergență dacă există un număr $N \geq 0$ astfel încât pentru orice $k \geq N$ are loc

$$|x_{k+1} - x_*| \leq C|x_k - x_*|^\beta,$$

sau, în cazul vectorilor și matricelor,

$$\|x_{k+1} - x_*\| \leq C\|x_k - x_*\|^\beta.$$

Evident, cu cât este mai mare β cu atât șirul $\{x_k\}$ converge mai repede către x_* .

În cele ce urmează vom ilustra prin exemple concrete definițiile date.

Exemple de convergență liniară.

1) Șirurile

$$x_0 = 2, x_1 = \frac{3}{2}, x_2 = \frac{5}{4}, x_3 = \frac{9}{8}, \mathbf{K} \quad x_k = 1 + 2^{-k}, \mathbf{K},$$

$$x_0 = 3, x_1 = \sqrt{x_0}, x_2 = \sqrt{x_1}, x_3 = \sqrt{x_2}, \mathbf{K}, \quad x_k = \sqrt{x_{k-1}} = 3^{2^{-k}}, \mathbf{K}$$

converg către $x_* = 1$ liniar cu rația $q = \frac{1}{2}$.

Într-adevăr,

$$\lim_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} = \lim_{k \rightarrow \infty} \frac{1 + 2^{-(k+1)} - 1}{1 + 2^{-k} - 1} = \frac{1}{2},$$

și

$$\lim_{k \rightarrow \infty} \frac{3^{2^{-(k+1)}} - 1}{3^{2^{-k}} - 1} = \lim_{k \rightarrow \infty} \frac{1}{3^{2^{-(k+1)}} + 1} = \frac{1}{2}.$$

2) Șirurile

a) $x_0 = c, x_1 = cq, x_2 = cq^2, x_3 = cq^3, \mathbf{K}, x_k = cq^k, \mathbf{K}$

b) $x_0 = c, x_1 = c(q+1), x_2 = c\left(q + \frac{1}{2}\right)^2, \mathbf{K}, x_k = c\left(q + \frac{1}{k}\right)^k, \mathbf{K}$

c) $x_0 = c, x_1 = c(q-1), x_2 = c\left(q - \frac{1}{2}\right)^2, \mathbf{K}, x_k = c\left(q - \frac{1}{k}\right)^k, \mathbf{K}$

d) $x_0 = c, x_1 = cq^2, x_2 = cq^{\frac{5}{2}}, x_3 = cq^{\frac{10}{3}}, \mathbf{K}, x_k = cq^{\frac{k+1}{k}}, \mathbf{K}$

unde $c > 0$ și $0 < q < 1$, converg către $x_* = 0$ liniar cu rația q .
 Demonstrația rezultă imediat din definiția convergenței liniare:

$$\lim_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} = \lim_{k \rightarrow \infty} \frac{x_{k+1} - 0}{x_k - 0} = q.$$

Exerciții.

1. Fie $0 < q_1 < q_2 < 1$. Definem șirul $\{x_k\}$ în modul următor:

$$x_0 = c + 1, x_1 = c + q_1, x_2 = c + q_1 q_2, x_3 = c + q_1^2 q_2,$$

$$x_4 = c + q_1^2 q_2^2, x_5 = c + q_1^3 q_2^2, \mathbb{K}, x_{2k} = c + q_1^k q_2^k, x_{2k+1} = c + q_1^{k+1} q_2^k, \mathbb{K}$$

unde $c > 0$. Să se arate că acest șir converge liniar către $x_* = c$ cu rația $q = \sqrt{q_1 q_2}$.

2. Să se demonstreze că șirul $\{x_k\}$ converge liniar cu rația $0 < q < 1$ dacă și numai dacă $\lim_{k \rightarrow \infty} \sqrt[k]{e(x_k)} = \lim_{k \rightarrow \infty} \sqrt[k]{|x_k - x_*|} = q$.

Exemple de convergență supraliniară. Șirurile

a) $x_0 = x_1 = 1, x_2 = \frac{1}{2}, x_3 = \frac{1}{27}, \mathbb{K}, x_k = \frac{1}{k^k}, \mathbb{K}$

b) $x_0 = x_1 = 1, x_2 = \frac{1}{2}, x_3 = \frac{1}{6}, \mathbb{K}, x_k = \frac{1}{k!}, \mathbb{K}$

c) $x_0 = x_1 = c, x_2 = \frac{c^2}{2}, x_3 = \frac{c^3}{6}, \mathbb{K}, x_k = \frac{c^k}{k!}, \mathbb{K}$

converg către $x_* = 0$ cu viteza supraliniară. Aceasta rezultă din faptul că

$$\lim_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} = \lim_{k \rightarrow \infty} \frac{x_{k+1} - 0}{x_k - 0} = 0.$$

Exercițiu. Să se demonstreze că orice șir de forma

$$x_0 = cq, x_1 = cq^m, x_2 = cq^{m^2}, \mathbb{K}, x_k = cq^{m^k}, \mathbb{K},$$

unde $c > 0$, $0 < q < 1$ și $m > 1$, converge către $x_* = 0$ cu viteza supraliniară.

Exemple de convergență pătratică.

1. Șirul $x_0 = cq, x_1 = cq^2, x_2 = cq^4, x_3 = cq^8, \mathbb{K}, x_k = cq^{2^k}, \mathbb{K}$,

unde constanța q satisface condiției $0 \leq q < 1$ și $c > 0$, este cu convergența pătratică.

Într-adevăr, avem $x_* = 0$ și

$$\lim_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} = \lim_{k \rightarrow \infty} \frac{x_{k+1} - 0}{(x_k - 0)^2} = \frac{1}{c}.$$

2. Șirul

$$x_0 = \frac{3}{2}, x_1 = \frac{5}{4}, x_2 = \frac{17}{16}, x_3 = \frac{257}{256}, \mathbb{K} \quad x_k = 1 + 2^{-2^k}, \mathbb{K},$$

care converge către $x_* = 1$, are viteza de convergență pătratică:

$$\lim_{k \rightarrow \infty} \frac{x_{k+1} - 1}{(x_k - 1)^2} = 1.$$

Exercițiu. Să se demonstreze că șirul definit prin formula

$$x_{k+1} = \frac{1}{2} \left(x_k + \frac{1}{x_k} \right), k = 0, 1, 2, \mathbb{K},$$

converge către $x_* = 1$ o oricare ar fi x_0 și viteza de convergență este pătratică.

Alte detalii și aspecte de diferite nuanțe legate de viteza de convergență a șirurilor pot fi studiate de cititor în lucrarea fundamentală **Ortega J.M., Rheinboldt W.C. Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, London and New York, 1970.** Lucrarea a apărut și în limba rusă în anul 1975 la editura Mir.

NUMERE APROXIMATIVE

1.1 Erori absolute și erori relative

Notăm cu x_* valoarea aproximativă pentru numărul exact x . Dacă $x_* < x$, atunci spunem că x_* aproximează numărul x prin lipsă, iar dacă $x_* > x$, atunci aproximarea lui x prin x_* este prin adaos.

De obicei, în procesul de calcul se înlocuiește valoarea exactă (care, în caz general, nu este cunoscută) prin valoarea sa aproximativă. În felul acesta comitem o eroare. Expresia

$$\Delta(x_*) = |x - x_*|$$

poartă numele de *eroare absolută*.

Eroarea absolută nu caracterizează suficient de bine precizia cu care se obțin rezultatele. Astfel, de exemplu, dacă $x = 1$ și $x_* = 2$, atunci eroarea absolută $\Delta(x_*) = 1$ indică o precizie slabă a măsurării. Dacă $x = 10^{10} + 1$ iar $x_* = 10^{10}$, aceeași eroare absolută $\Delta(x_*) = 1$ caracterizează o precizie remarcabilă. Aceasta ne conduce la noțiunea de *eroare relativă* $\delta(x_*)$ care reprezintă raportul dintre eroarea absolută și valoarea aproximativă, adică

$$\delta(x_*) = \frac{|x - x_*|}{|x_*|}$$

dacă $x_* \neq 0$.

În exemplele de mai sus erorile relative sunt egale cu 0.5 respectiv cu 10^{-10} , ceea ce confirmă buna precizie a măsurării în cazul al doilea.

Dacă se cunosc numerele x și x_* , atunci calculul erorii absolute și relative este imediat. Dar de obicei, în majoritatea cazurilor se cunoaște numai aproximarea x_* . De aceea se introduce noțiunea de margine (sau limită) a erorii absolute și relative. Numărul pozitiv ε este o *margine* (sau o *limită*) a erorii

absolute a numărului aproximativ x_* dacă

$$|x-x_*| \leq \varepsilon$$

iar numărul pozitiv r este o *limită a erorii relative* dacă

$$\frac{|x-x_*|}{|x_*|} \leq r$$

Notăția $x = x_* \pm \varepsilon$ semnifică întotdeauna faptul, că $|x-x_*| \leq \varepsilon$, adică

$$x_* - \varepsilon \leq x \leq x_* + \varepsilon.$$

Orice număr aproximativ x_* poate fi scris sub forma

$$x_* = c_1 10^m + c_2 10^{m-1} + \dots + c_n 10^{m-n+1},$$

unde c_1, c_2, \dots, c_n sunt cifrele zecimale ale numărului aproximativ x_* . Se știe că zerourile de la începutul numărului servesc numai pentru a fixa poziția virgulei zecimale. Cifrele cuprinse între prima și ultima cifră diferită de zero sau care indică ordinele păstrate în calcule se numesc *cifre semnificative*.

Exemplu. Numărul aproximativ

$$x_* = 3 \cdot 10^1 + 6 \cdot 10^0 + 0 \cdot 10^{-1} + 5 \cdot 10^{-2} + 8 \cdot 10^{-3}$$

are cinci cifre semnificative, iar numărul

$$y_* = -(2 \cdot 10^{-3} + 8 \cdot 10^{-4} + 0 \cdot 10^{-5}) = -0.00280$$

are trei cifre semnificative (primele trei zerouri sunt ne semnificative).

Dacă mărimea erorii x_* nu depășește $0.5 \cdot 10^{-t}$ se spune că numărul aproximativ x_* are t cifre zecimale corecte.

Dacă numărul aproximativ se scrie fără a indica limita erorii absolute, atunci în scrierea lui se consideră că toate cifrele sunt corecte. În acest caz zerourile de la sfârșitul numărului nu se aruncă.

De pildă numerele 0.0345 și 0.034500 sunt diferite; eroarea absolută a primului număr nu depășește 0.0001, iar eroarea absolută al celui de al doilea număr este mai mică ca 10^{-6} .

Exemple: 0.010224 ± 0.000004 are cinci zecimale corecte și patru cifre semnificative; 0.001234 ± 0.000006 are patru zecimale

corecte și două cifre semnificative (deoarece valoarea maximă a numărului poate fi 0.001240 iar minima 0.001228 și deci ultimele două zecimale sunt nesigure).

Numărul de zecimale corecte ne permite să ne facem o idee despre mărimea erorii absolute în timp ce numărul de cifre semnificative ne dă o idee sumară despre mărimea erorii relative.

1.2 Propagarea și sursele erorilor

În procesul de calcul aproximativ eroarea se propagă de la o operație la alta. Fie date valorile aproximative x_* și y_* ale valorilor exacte x și y , afectate de erorile ε_x și ε_y , adică fie

$$x = x_* \pm \varepsilon_x, y = y_* \pm \varepsilon_y$$

Atunci avem

$$x_* - \varepsilon_x + y_* - \varepsilon_y \leq x + y \leq x_* + \varepsilon_x + y_* + \varepsilon_y$$

$$x_* + y_* - (\varepsilon_x + \varepsilon_y) \leq x + y \leq x_* + y_* + (\varepsilon_x + \varepsilon_y),$$

sau

$$x + y = x_* + y_* \pm (\varepsilon_x + \varepsilon_y)$$

În mod analog ca în cazul adunării, se va obține

$$x - y = x_* - y_* \pm (\varepsilon_x + \varepsilon_y)$$

Deci, *la adunare sau scădere, marginea erorii absolute a rezultatului este dată de suma marginilor pentru erorile absolute ale termenilor.*

Se poate demonstra de asemenea că *la înmulțire și împărțire marginile erorilor relative ale factorilor se adună.*

Fie acum date două numere pozitive x_* și y_* aproximativ egale, afectate de erorile absolute $\Delta(x_*)$ și $\Delta(y_*)$. Atunci

$$\delta(x_* - y_*) = \frac{\Delta(x_* - y_*)}{|x_* - y_*|} \leq \frac{\Delta(x_*) + \Delta(y_*)}{|x_* - y_*|}$$

și deci, eroarea relativă a diferenței poate fi destul de mare, dacă diferența $|x_* - y_*|$ este foarte mică. Aceasta ne arată că exactitatea

relativă poate fi foarte slabă atunci când efectuăm diferența a două numere aproximativ egale.

Exemplu. Vom considera numerele aproximativ egale

$$x=0.1234\pm 0.5\cdot 10^{-4} \quad \text{și} \quad y=0.1233\pm 0.5\cdot 10^{-4}$$

atunci $x-y = 0.0001\pm 0.0001$ și marginea erorii este tot atât de mare ca și estimarea rezultatului.

Acest fenomen poartă denumirea de *anulare prin scădere* sau de *neutralizare a termenilor*. Cele mai serioase erori care apar în calculele efectuate cu ajutorul calculatorului electronic sunt datorate acestui fenomen.

De câte ori este posibil neutralizarea termenilor se evită prin rescrierea formulelor de calcul sau prin alte schimbări în algoritm. De exemplu, o expresie de forma $(\alpha + \gamma)^2 - \alpha^2$ poate fi scrisa sub forma $\gamma(\gamma + 2\alpha)$, sau expresia

$$\frac{\sqrt{\alpha + \gamma} - \sqrt{\alpha}}{b}$$

sub forma

$$\frac{\gamma}{b(\sqrt{\alpha + \gamma} + \sqrt{\alpha})}$$

Vom mai da un exemplu în care se arată cum poate fi evitată anularea prin scădere. Ecuația de gradul doi

$$x^2 + 1000.01 \cdot x - 2.5245315 = 0$$

are una din rădăcini egală cu 0.0025245. Dacă calculăm rădăcina cu ajutorul formulei

$$x_1 = \frac{-1000.01 + \sqrt{(1000.01)^2 + 4 \cdot 2.5245315}}{2},$$

efectuând calculele cu opt cifre semnificative, se va obține

$$x_1 = \frac{-1000.0100 + 1000.0150}{2} = 0.0025.$$

Rezultatul obținut are numai două cifre corecte cu toate că radicalul de gradul doi a fost calculat cu opt cifre semnificative.

Dacă vom face calculul aceleiași rădăcini utilizând formula

$$x_1 = \frac{2 \cdot 2.524315}{1000.01 + \sqrt{(1000.01)^2 + 4 \cdot 2.5245315}} = 0.0025245$$

se va obține rezultatul exact.

Precizia calculului numerice este criteriul cel mai eficient pentru alegerea metodelor de calcul. Analiza erorii dintr-un rezultat numeric este o chestiune esențială în orice calcul, fie că este executat manual, fie de un calculator. Cu toate performanțele calculatoarelor electronice, precizia rezultatelor este influențată de diferite erori. Se pot distinge trei surse de erori:

1. Erori provenite din simplificarea modelului fizic, pentru a fi descris într-un model matematic; erori din măsurările inițiale sau din soluții aproximative ale altor probleme etc. Aceste tipuri de erori se numesc *erori inerente*. Ele nu pot fi influențate de metoda de calcul.
2. *Erori de metodă sau de trunchiere*. Majoritatea metodelor numerice necesită un număr infinit de operații aritmetice pentru a ajunge la soluția exactă a problemei. De aceea suntem nevoiți să trunchiem metoda după un număr finit de operații. Ceea ce ometem constituie eroarea de trunchiere.
3. *Erori de rotunjire* în datele de intrare în calcule și în datele de ieșire. Multe numere nu pot fi reprezentate exact printr-un număr dat de cifre. Dacă în calcule numerice trebuie să folosim numărul π , îl putem scrie 3.14, 3.14159 sau 3.1415926 etc. Nici un număr irațional nu poate fi reprezentat printr-un număr finit de cifre. Chiar și unele numere raționale nu au o reprezentare exactă. Numărul $1/3$ poate fi scris ca 0.3333..., o succesiune a cifrei 3 la partea zecimală.

Datorită proprietăților constructive ale calculatoarelor electronice este necesară limitarea numărului cifrelor semnificative. Un exemplu instructiv este furnizat de numărul rațional $1/10$ care se folosește de multe ori ca dimensiune a «pasului» în multe algoritme. În sistemul binar (care folosește pentru reprezentarea numerelor cifrele 0 și 1) fracția $1/10$ are o

reprezentare infinită 0.00011001100... . În calcule trebuie să ne mărginim la un număr finit de cifre semnificative. Când se adună de zece ori numărul care reprezintă o aproximație binară a numărului 1/10, rezultatul nu va fi egal exact cu unitatea.

1.3 Numere cu virgulă mobilă

Este bine cunoscut că pe majoritatea calculatoarelor moderne numerele reale se reprezintă cu ajutorul virgulei mobile . Un număr scris în virgula mobilă este compus dintr-o fracție, numită *mantisă* și un întreg, numit *exponent*. Deci,

$$x = \pm m \cdot \beta^e,$$

unde β este baza sistemului de numerație (binar, octal sau hexazecimal), m este mantisa numărului și e este exponentul, afectat de semn. Frația m satisface

$$\frac{1}{\beta} \leq m < 1$$

și are forma

$$m = \frac{d_1}{\beta^1} + \frac{d_2}{\beta^2} + \dots + \frac{d_t}{\beta^t},$$

unde numerele întregi d_1, d_2, \dots, d_t , numite cifre, verifică inegalitățile

$$0 \leq d_i \leq \beta - 1, i = 1, 2, \dots, t$$

și $L \leq e \leq U$.

Dacă prima cifră din mantisă este diferită de zero, atunci numărul reprezentat în virgulă mobilă se numește *normalizat*.

Prin urmare, *sistemul de calcul cu numere cu virgulă mobilă* este o mulțime $F(\beta, t, L, U)$ caracterizată de patru parametri: baza β , precizia mașinii t și intervalul exponenților $[L, U]$.

F este o mulțime finită care conține

$$2 \cdot (\beta - 1) \cdot \beta^{t-1} \cdot (U - L + 1) + 1$$

numere. În fig.1.1 este reprezentată mulțimea F formată din 33 puncte pentru sistemul de calcul cu virgulă mobilă, ilustrativ cu următorii parametri: $\beta=2$, $t=3$, $L=-1$, $U=2$.

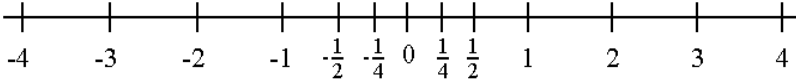


Fig.1.1. Sistemul de calcul $F(2,3,-1,-2)$.

Mulțimea F nu poate reproduce oricât de detaliat structura continuă a numerelor reale. Mai mult, în general nu putem reprezenta în calculator numerele al căror modul depășește cel mai mare element al lui F sau care sunt mai mici în modul decât cel mai mic număr din F .

Mantisa m poate fi scrisă

$$m = \beta^{-t} (d_1 \beta^{t-1} + d_2 \beta^{t-2} + \dots + d_t)$$

de unde rezultă, că dacă $d_1 \neq 0$, atunci maximum mării m este egal cu $1 - \beta^{-t}$, ce corespunde cazului $m_i = \beta - 1$, $i=1, 2, \dots, t$; valoarea minimală va fi β^{-1} și se obține pentru $d_1 = 1, d_2 = d_3 = \dots = d_t = 0$.

Fie x un număr real care nu depășește limitele mulțimii F și $x \neq 0$; în calculator acest număr este reprezentat de numărul cu virgulă mobilă notat $fl(x)$, a cărui mantisă m_* se obține din mantisa m a lui x rotunjind-o la t cifre (de aceea spunem că precizia mașinii este t). Dacă se efectuează rotunjirea corectă atunci

$$|m - m_*| \leq \frac{1}{2} \beta^{-t}.$$

Eroarea relativă în $fl(x)$ este

$$\frac{|fl(x) - x|}{|x|} \leq \frac{1}{2} \beta^{1-t},$$

deoarece $\frac{|x_* - x|}{|x|} = \frac{|m - m_*|}{|m|}$ și $m \geq \frac{1}{\beta}$

Numărul $\varepsilon_M = \frac{1}{2} \beta^{1-t}$ se numește *unitatea (de rotunjire a) mașinii*. Efectuând rotunjirea corectă, numărul $fl(x)$ este cel mai

apropiat element de x , care aparține lui F . Dacă se folosește rotunjirea prin tăiere (se elimină compararea primei cifre neglijate), atunci $\varepsilon_M = \beta^{1-t}$ și $fl(x)$ este cel mai apropiat element din F , inferior lui x (vezi 1.4).

În afară de parametrul ε_M în practică sunt larg răspândiți încă doi parametri σ și λ : cel mai mic element pozitiv și elementul maximal al lui F . Pentru ilustrare vom analiza sistemul $F(2, 3, -1, 1)$. El conține numărul zero și toate numerele care au o exprimare binară de forma

$$x = \pm 0.1d_2d_3 \cdot 2^e,$$

unde $-1 \leq e \leq 1$ iar fiecare din cifrele d_2 și d_3 este 0 sau 1. Prin urmare avem trei posibilități pentru valoarea exponentului ($-1, 0$ ori $+1$) și patru pentru reprezentarea părților fracționare (0.100, 0.101, 0.110 și 0.111). Mulțimea F constă din $2 \cdot 1 \cdot 4 \cdot 3 + 1 = 25$ de numere cu virgulă mobilă. În sistemul zecimal de numerație fracțiile de mai sus pot fi scrise respectiv $1/2, 5/8, 3/4$ și $7/8$ (de exemplu, fracția binară 0.101 devine $1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} = 1/2 + 1/8 = 5/8$). Prin urmare în sistemul de calcul $F(2,3,-1,1)$ cel mai mic element pozitiv este egal cu $\sigma = 1/2 \cdot 2^{-1} = 1/4$, iar cel mai mare element $\lambda = 7/8 \cdot 2^1 = 7/4$. În fig.1.2 sunt reprezentate numerele pozitive ale mulțimii F .

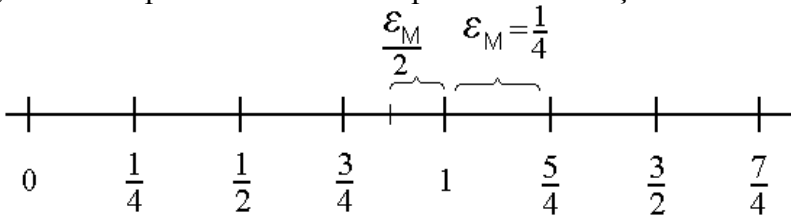


Fig.1.2. Numerele pozitive ale sistemului de calcul $F(2,3, -1, 1)$.

Parametrii σ și λ se reprezintă prin parametrii β, t, L și U după cum urmează

$$\sigma = \beta^{L-1}, \lambda = \beta^U (1 - \beta^{-t})$$

Unitatea de rotunjire a mașinii ε_M se mai numește *epsilon al mașinii* și este cel mai utilizat parametru ce caracterizează un sistem de calcul dat. Acest parametru ne dă măsura de «discretizare» a sistemului F care are loc pentru tot intervalul numerelor nenule în virgulă mobilă. Deci distanța dintre numărul $x \in F$ și numărul cel mai apropiat de el în sistemul dat nu e mai mică decât $\varepsilon_M/x//\beta$ și nu e mai mare decât $\varepsilon_M/x/$ (numai dacă numărul x nu este situat în vecinătatea lui zero).

Din fig.1.2 se vede că în vecinătatea lui σ distanța dintre oricare două numere cu virgulă mobilă este mai mică decât σ . Prin urmare aceste numere nu pot fi obținute unul din altul prin adunare. Ele pot fi căpătate decât în calculul expresiilor aritmetice.

Un alt exemplu. Fie $\beta=10$, $t=6$ și $L=-100$. Atunci $\varepsilon_M=10^{-5}$, $\sigma=10^{-101}$. Între zero și σ nu există nici un număr ce aparține sistemului dat, în timp ce între σ și 10σ sunt 899999 de numere cu virgulă mobilă.

1.4 Aritmetica virgulei mobile și erorile de rotunjire

Pe mulțimea F , care o considerăm un model al mulțimii numerelor reale, se definesc operațiile aritmetice în așa fel încât ele să reproducă modul de efectuare al acestor operații de către calculatorul electronic.

Fie x și $y \in F$. Atunci suma lor exactă nu aparține, însă, întotdeauna mulțimii F . În exemplul în care F este o mulțime formată din 33 de puncte și cu parametrii $\beta=2$, $t=3$, $L=-1$, $U=2$ (vezi fig.1.1 din 1.3) putem constata aceasta luând $x=5/4$ și $y=3/8$, sau $x=3/4$ și $y=7/8$. În particular, în cazul al doilea vom avea

$$0.110 \cdot 2^0 + 0.111 \cdot 2^0 = 0.1101 \cdot 2^1 \quad (3/4 + 7/8 = 13/8)$$

Suma calculată nu aparține sistemului de calcul F deoarece pentru reprezentarea părților fracționare a ei sunt necesare patru cifre binare. Prin urmare, operația de adunare trebuie modelată ea însăși

în calculatorul electronic cu ajutorul unei aproximații a ei numită *adunare cu virgulă mobilă* (pe care o vom nota cu \oplus).

Dacă x și y sunt numere cu virgulă mobilă iar numărul $x+y$ nu iese din limitele mulțimii F atunci ar fi ideal ca

$$x \oplus y = fl(x+y)$$

Acest ideal este atins sau aproape atins de majoritatea calculatoarelor electronice. În sistemul de calcul F din exemplul considerat ne putem aștepta că $5/4 \oplus 3/8$ este egal cu $3/2$ sau cu $7/4$ (deoarece ambele elemente se află la distanță egală de $5/4+3/8$).

Diferența dintre $x \oplus y$ și $x+y$ (pentru $x, y \in F$) reprezintă o eroare de rotunjire care se comite în cazul adunării cu virgula mobilă. Proprietăți asemănătoare sunt adevărate și pentru celelalte operații aritmetice cu virgulă mobilă.

Revenind la exemplul dat de fig. 1 din 1.3 constatăm că $5/4+3/8$ sau $3/4+7/8$ nu aparține lui F din cauza repartizării elementelor lui F . Pe de altă parte, suma $7/2+7/2$ nu aparține lui F deoarece 7 este mai mare decât elementul maximal al lui F . Încercarea de a forma o astfel de sumă atrage la majoritatea calculatoarelor apariția unui *semnal de depășire*, după care calculele se întrerup deoarece nu este posibil de dat o aproximație de sens pentru numerele care ies din limitele lui F .

Deși majoritatea sumelor $x+y$ cu $x \in F, y \in F$, aparțin ele însele lui F , foarte rar produsul (exact) obișnuit $x \cdot y$ aparține lui F , deoarece, de regulă, el are $2t$ sau $2t-1$ cifre semnificative. În afară de aceasta, depășirea este mai probabilă la înmulțire. În fine, în cazul înmulțirii cu virgulă mobilă, este posibilă apariția unui *zero-mașină*, atunci când pentru $x \neq 0$ și $y \neq 0$, produsul $x \cdot y$ este nenul dar este mai mic, în modul, decât cel mai mic element pozitiv al lui F (apariția unui zero-mașină este posibilă și în cazul scăderii deși aceasta se întâmplă foarte rar).

Operațiile de adunare și înmulțire cu virgulă mobilă sunt comutative dar nu sunt asociative; de asemenea nu este adevărată nici distributivitatea.

În continuare ne vom ocupa de erorile de rotunjire. Eroarea de rotunjire reprezintă diferența $|fl(x)-x|$, unde x este un număr real, exponentul cărui aparține intervalului $[L, U]$ iar $fl(x)$ este numărul cu virgulă mobilă ce semnifică numărul dat x în memoria calculatorului. Pentru $x \neq 0$ eroarea relativă în $fl(x)$ se definește astfel

$$\delta(x) = \frac{|fl(x) - x|}{|x|}$$

și se îndeplinește condiția

$$\delta(x) \leq \varepsilon_M = \begin{cases} \beta^{1-t} & , \text{daca are loc rotunjirea prin trunchiere} \\ \frac{1}{2}\beta^{1-t} & , \text{daca are loc rotunjirea corecta.} \end{cases}$$

Într-adevăr, fie numărul întreg e , $L < e < U$ și

$$\beta^{e-1} \leq x \leq \beta^e$$

În intervalul $[\beta^{e-1}, \beta^e]$ numerele în virgulă mobilă sunt repartizate uniform cu pasul β^{e-t} . În cazul rotunjirii prin trunchiere $fl(x)$ se află la depărtare de x ce nu depășește mărimea β^{e-t} iar în cazul rotunjirii corecte mărimea $\beta^{e-t/2}$, adică

$$|fl(x) - x| \leq \begin{cases} \beta^{e-t} & , \text{in cazul rotunjirii prin taiere} \\ \frac{1}{2}\beta^{e-t} & , \text{in cazul rotunjirii corecte.} \end{cases}$$

Deoarece $x \geq \beta^{e-1}$:

$$\delta(x) \leq \begin{cases} \frac{\beta^{e-t}}{\beta^{e-1}} = \beta^{1-t} & , \text{In cazul rotunjirii prin taiere} \\ \frac{\beta^{e-t}/2}{\beta^{e-1}} = \frac{1}{2}\beta^{1-t} & , \text{pentru rotunjirea corecta.} \end{cases}$$

Din cele de mai sus reiese că rotunjirea prin trunchiere a numerelor se efectuează mai repede decât rotunjirea corectă. Însă eroarea relativă este de două ori mai mare. În afară de aceasta eroarea rotunjirii prin trunchiere are permanent același semn (opus

semnului numărului dat) ceea ce în cazul calculelor masive poate să conducă la o acumulare rapidă de erori. De aceea ar fi mai bine să se utilizeze rotunjirea corectă.

Pentru exemplificare vom analiza sistemul de calcul $F(10,4,-50,50)$. Numărului $x=12.467$ prin trunchiere îi va corespunde numărul în virgula mobilă $fl(x)=0.1246 \cdot 10^2$ cu eroarea relativă de rotunjire

$$\delta(x)=0.007/12.467 \approx 0.00056 < \varepsilon_M=10^{-3}=0.001$$

În cazul rotunjirii corecte vom avea $fl(x)=0.1247 \cdot 10^2$ și

$$\delta(x)=0.003/12.467 \approx 0.00024 < \varepsilon_M=10^{-3}/2=0.0005$$

1.5 Determinarea parametrilor unui sistem de calcul

Pentru început vom analiza principiile generale de realizare a operațiilor aritmetice la un calculator. Fie doua numere nenule reprezentate de numerele cu virgulă mobilă

$$x=m_x \cdot \beta^p \quad \text{și} \quad y=m_y \cdot \beta^q,$$

La adunare și scădere numerele se aduc la exponentul cel mai mare și apoi se adună (scad) mantisele:

$$z = x \pm y = \begin{cases} (m_x \pm m_y \cdot \beta^{-(p-q)}) \cdot \beta^p, & \text{daca } p > q, \\ (m_x \cdot \beta^{-(q-p)} \pm m_y) \cdot \beta^q, & \text{daca } p \leq q. \end{cases}$$

Evident că mantisa rezultatului poate deveni mai mare ca unitatea dar întotdeauna este mai mică decât doi. Dacă în urma operației de adunare mantisa sumei depășește unitatea, atunci trebuie normalizată aceasta prin deplasarea ei cu o poziție spre dreapta, adăugând o unitate la exponent pentru a compensa deplasarea.

Operațiile de înmulțire și împărțire se definesc astfel:

$$z = x \cdot y = (m_x \cdot m_y) \cdot \beta^{p+q},$$

$$z = \frac{x}{y} = \frac{m_x}{m_y} \beta^{(p-q)}, y \neq 0,$$

adică se înmulțesc (împart) mantisele numerelor x și y , se normalizează rezultatul, oprindu-se t cifre la mantisă și la partea exponențială se adună (scad) exponenții.

După cum vedem, deoarece calculul se face cu un număr anumit de cifre semnificative t , modul de rotunjire are o mare importanță. De modul de rotunjire depinde parametrul ε_M care caracterizează exactitatea relativă a sistemului de calcul F . Acest parametru poate fi definit ca cel mai mic număr pozitiv care adăugat în sistemul de calcul F la unitate dă în rezultat un număr cu virgulă mobilă ce aparține din nou lui F și este strict mai mare ca 1 , adică

$$fl(1+\varepsilon_M) > 1.$$

De exemplu, în sistemul de calcul $F(10,4,-50,50)$ la rotunjirea prin trunchiere $\varepsilon_M = 10^{-3} = 0.001$, deoarece

$$fl(1+0.001) = fl(0.1001 \cdot 10^1) = 0.1001 \cdot 10^1 > 1$$

și nu există un alt număr ε_M mai mic cu această proprietate. La rotunjirea corectă vom avea $\varepsilon_M = 0.005$, fiindcă

$$fl(1+0.0005) = fl(0.10005 \cdot 10^1) = 0.1001 \cdot 10^1 > 1.$$

La majoritatea mașinilor de calcul se folosesc instrucțiuni speciale care determină modul de rotunjire. De aceea unul și același program poate da rezultate diferite în dependență de modul de rotunjire stabilit de translator. Un mare număr de compilatoare generează programul obiect ca să folosească tăierea. Acest tip de rotunjire introduce o eroare mai mare decât regula obișnuită de rotunjire, care, după cum s-a mai spus, folosește mult timp de calculator dacă este aplicată la fiecare operație aritmetică.

Pentru a afla care mod de rotunjire e folosit în programele compilatoare este suficient să calculăm unitatea de rotunjire a mașinii ε_M . Dacă $\varepsilon_M = \beta^{1-t}$, atunci mașina de calcul generează modul de rotunjire prin tăiere, iar dacă $\varepsilon_M = \frac{1}{2}\beta^{1-t}$ - generează modul de rotunjire corect.

De reținut că și ceilalți parametri (β , t , U , L , σ și λ) pot fi estimați direct la calculator utilizând tehnicile de programare (software) ai acesteia. Programele și argumentarea algoritmilor acestor programe poate fi găsită în [8,15,17,24,33,37], lucrări la care și-l trimitem pe cititor pentru a afla și alte detalii.

Trebuie de menționat importanța ordinii efectuării operațiilor aritmetice, deoarece într-o aritmetică a virgulei mobile nu au loc legile asociativă și distributivă. Pentru unele sisteme de calcul

$$(A + 1.0) - A \neq A + (1.0 - A)$$

Într-adevăr, fie de exemplu $t=40$ și $\beta=2$. Atunci (vezi [9], pag.30) în cazul rotunjirii corecte

$$(2^{40} + 1.0) - 2^{40} = 2$$

în cazul rotunjirii prin tăiere

$$(2^{40} + 1.0) - 2^{40} = 0,$$

iar dacă vom muta parantezele, pentru orice mod de rotunjire, vom avea

$$2^{40} + (1.0 - 2^{40}) = 1.$$

Deci putem sintetiza că operațiile aritmetice $+$, $-$, $*$, $/$ se realizează la calculatoarele numerice după cum urmează (prin x este notat rezultatul exact al operației aritmetice):

- 1) dacă $\sigma \leq |x| \leq \lambda$, atunci rezultatul operației se rotunjește;
- 2) dacă $|x| < \sigma$, atunci rezultatul se anulează, adică apare un zero-mașină;
- 3) dacă $|x| > \lambda$, atunci calculele se întrerup și apare un semnal de depășire.

1.6 Efectul erorilor de rotunjire

Erorile introduse de calculator în procesul de calcul sunt erorile de rotunjire. Aceste erori se propagă de la o operație

aritmetică la alta. Felul în care o eroare se propagă depinde și de algoritmul de calcul. Aducem câteva exemple ilustrative

1.6.1. Calculul mediei aritmetice

Media aritmetică dintre două numere a și b poate fi calculată prin formulele :

$$c = \frac{a+b}{2} \quad (1.1)$$

$$a = a + \frac{b-a}{2} \quad (1.2)$$

Formula (1.1) necesită cu o operație de adunare mai puțin decât formula (1.2) dar din punct de vedere al exactității nu este întotdeauna cea mai bună. Într-adevăr, fie că calculele se efectuează într-o aritmetică zecimală ($\beta=10$) a virgulei mobile cu trei cifre semnificative ($t=3$) și fie $a=0.596$ și $b=0.600$. Presupunem de asemenea că are loc rotunjirea corectă. Atunci

$$c = \frac{0.596+0.600}{2} = \frac{1.20}{2} = 0.600,$$

cu toate că valoarea corectă a lui c este egală cu 0.598. Efectuând calculele după formula (1.2), vom avea

$$c = 0.596 + \frac{0.600-0.596}{2} = 0.596 + \frac{0.004}{2} = 0.598.$$

Este necesar să menționăm că în exemplul dat în care înclinăm pentru formula (1.2) numerele a și b au același semn.

Să considerăm acum un alt exemplu în care $t=4$ și se aplică rotunjirea prin tăiere. Fie $a=-3.483$ iar $b=8.765$. Folosind formula (1.1) obținem

$$c = \frac{-3.483+8.765}{2} = \frac{5.282}{2} = 2.641$$

și acest rezultat este exact. Calculul după formula (2) ne dă:

$$\begin{aligned} c &= -3.483 + \frac{8.765+3.483}{2} = -3.483 + \frac{12.24}{2} = \\ &= -3.483 + 6.120 = 2.637. \end{aligned}$$

Chiar dacă s-ar efectua rotunjirea corectă rezultatul obținut prin formula (1.2) ar fi diferit de valoarea exactă, deoarece în acest caz am avea $c=2.642$. Prin urmare, în exemplul de mai sus, unde a și b sunt de semne diferite, vom prefera formula (1.1).

În concluzie, este necesar de-a ne folosi de una din formulele (1.1) sau (1.2), în dependență de semnele lui a și b : dacă $sign(a) \neq sign(b)$ atunci $c=(a+b)/2$; în caz contrar $c=a+(b-a)/2$.

1.6.2. Evaluarea recurentă a unei integrale

Ne propunem să calculăm următoarea integrală :

$$I_n = \int_0^1 x^n e^{x-1} dx, n = 1, 2, K$$

Integrând prin părți, obținem

$$\int_0^1 x^n e^{x-1} dx = x^n e^{x-1} \Big|_0^1 - \int_0^1 nx^{n-1} e^{x-1} dx,$$

sau

$$I_n = 1 - n \cdot I_{n-1}, n = 2, 3, K$$

Pentru $n=1$ avem

$$I_1 = \int_0^1 x e^{x-1} dx = \frac{1}{e}.$$

Fie aritmetica virgulei mobile cu $\beta=10$ și $t=6$. Utilizând formula de recurență $I_n = 1 - n \cdot I_{n-1}$, obținem [29]:

$$\begin{aligned} I_1 &\approx 0.367879, & I_6 &\approx 0.127120, \\ I_2 &\approx 0.264242, & I_7 &\approx 0.110160, \\ I_3 &\approx 0.207274, & I_8 &\approx 0.118720, \\ I_4 &\approx 0.170904, & I_9 &\approx -0.068480, \\ & & I_5 &\approx 0.145480, \end{aligned}$$

Deși $x^n e^x \geq 0, \forall x \in [0, 1]$ se observă că pentru $n=9$ am obținut $I_n < 0$, evident contradictoriu. Apariția rezultatului eronat se datorează algoritmului utilizat. Pentru explicarea acestui fapt notăm cu E_n eroarea din I_n . Atunci E_1 este eroarea din calculul $1/e=0.367879$,

deci $E_1 \approx 4.412 \cdot 10^{-7}$. Se vede din formula de recurență că $E_2 = -2E_1$, $E_3 = -3E_2 = (-2) \cdot (-3)E_1 = 3!E_1$ ș.a.m.d. Calculând E_n pentru $n=9$ vom găsi $E_9 \approx 9! \cdot 4.412 \cdot 10^{-7} \approx 0.1601$, care este în mod evident mai mare decât I_9 , întrucât valoarea exactă I_9 (cu trei cifre semnificative) este egală cu $-0.06848 + 0.1601 = 0.0916$. Se spune că algoritmul de calcul folosit este *instabil*.

Deci, se vede că algoritmul constituie aici sursa erorilor. Pentru înlăturarea acestui neajuns vom rescrie formula de recurență sub forma

$$I_{n-1} = \frac{1 - I_n}{n}, n = 2, 3, K$$

În acest caz eroarea din I_n la fiecare pas se înmulțesc la factorul $1/n$. Prin urmare, pornind cu I_N ($N \gg 1$) am obține $I_{N-1}, I_{N-2}, \dots, I_3, I_2$ și eroarea de rotunjire s-ar micșora la fiecare iterație. Despre astfel de algoritme se spune că au o *stabilitate numerică*. Pentru a calcula valoarea inițială I_N observăm că

$$I_n \leq \int_0^1 x^n dx = \left. \frac{x^{n+1}}{n+1} \right|_0^1 = \frac{1}{n+1},$$

și deci

$$\lim_{n \rightarrow \infty} I_n = 0.$$

Alegând $I_{20} = 0$, se admite o eroare ce nu depășește $1/21$. Atunci eroarea în I_{19} nu va întrece $(1/20) \cdot (1/21) \approx 0.0024$. Această eroare se va micșora până la $4 \cdot 10^{-8}$ în timpul calcului I_{15} și devine mai mică decât eroarea de rotunjire. Rezultatele obținute prin recurență inversă sunt următoarele [29]:

$I_{20} \approx 0.0,$	$I_{14} \approx 0.0627322,$
$I_{19} \approx 0.0500000,$	$I_{13} \approx 0.0669477,$
$I_{18} \approx 0.0500000,$	$I_{12} \approx 0.0717733,$
$I_{17} \approx 0.0527778,$	$I_{11} \approx 0.0773523,$
$I_{16} \approx 0.0557190,$	$I_{10} \approx 0.0838771,$
$I_{15} \approx 0.0590176,$	$I_9 \approx 0.0916123.$

Deci este importantă alegerea corectă a algoritmului de calcul. Un algoritm de calcul se numește *numeric stabil* dacă aplicat unei probleme cu date inițiale «ușor perturbate» produce o soluție care aproape coincide cu soluția exactă (soluția problemei cu datele inițiale neperturbate).

1.6.3. Exemple de sisteme rău condiționate

Fie sistemul liniar

$$\begin{cases} 5x - 331y = 5, \\ 6x - 397y = 7. \end{cases}$$

Rezolvarea se poate face folosind algoritmul dat de regula lui Cramer

$$\Delta = \begin{vmatrix} 5 & -331 \\ 6 & -397 \end{vmatrix} = 1, x = \begin{vmatrix} 5 & -331 \\ 7 & -397 \end{vmatrix} = 332, y = \begin{vmatrix} 5 & 5 \\ 6 & 7 \end{vmatrix} = 5.$$

Soluția exactă a sistemului $x=332, y=5$. Dacă se reia sistemul de mai sus, admițând însă o variație mică a coeficientului 5 de pe lângă x din prima ecuație, adică

$$\begin{cases} 5.01x - 331y = 5, \\ 6x - 397y = 7, \end{cases}$$

același procedeu de calcul ne conduce la soluția $x = -111.7845\dots, y = -1.7070\dots$. S-a produs o catastrofă! Despre un astfel de sistem se spune că este *rău* (sau *prost*) *condiționat*.

Un alt exemplu de sistem rău condiționat :

$$\begin{cases} x + 2y = 3, \\ 0.499x + 1.001y = 1.5. \end{cases}$$

Soluția exactă este $x=y=1.0$, ceea ce se poate verifica prin substituție. Dacă înlocuim ecuația a doua a sistemului dat cu ecuația $0.5x + 1.001y = 1.5$, atunci soluția devine $x=3, y=0$.

Sistemul de ecuații

$$\begin{cases} 14x + 13y - 66z = 1, \\ 12x + 11y - 13z = 1, \\ 11x + 10y + 4z = 1, \end{cases}$$

admite o soluție unică $x=1, y=-1, z=0$. Înlocuim elementele din partea dreaptă a sistemului $(1, 1, 1)$ cu $1.001, 0.999$ și 1.001 respectiv. Atunci, lucrând doar cu trei cifre semnificative, vom obține $x = -0.683, y = 0.843, z = 0.006$.

În practică soluția exactă nu se cunoaște, deci nu putem avea certitudinea că soluția aproximativă calculată este suficient de aproape de soluția exactă. Acest lucru depinde, după cum se vede din cele câteva exemple de mai sus, atât de algoritmul de calcul, cât și de însuși problema considerată. Prin urmare, rezolvarea numerică a unei probleme nu este o chestiune simplă, adică nu este suficient să facem niște calcule pentru a ajunge la soluția problemei. Mai este necesar de a studia și de a face aprecieri, privind condiționarea problemei și stabilitatea numerică a algoritmilor de calcul.

1.7 Exerciții

1. Să se calculeze eroarea absolută și relativă a numărului π dacă se ia drept număr aproximativ $x_* = 3.1416$

2. Fie aritmetica virgulei mobile cu $\beta = 10, t = 4$ și se aplică rotunjirea prin tăiere. Să se determine eroarea relativă a diferenței $x - y$, dacă

a) $x = 0.3258 \times 10^3$ și $y = 0.3257 \times 10^3$;

b) $x = \frac{2}{3}$ și $y = 0.6665$.

3. Presupunând că lucrăm într-o aritmetică a virgulei mobile dată, să se găsească eroarea relativă în calculul expresiilor u și v :

$$a) \quad u = \frac{1}{x} - \frac{1}{x+1} \quad \text{și} \quad v = \frac{1}{x(x+1)}, x \neq 0, x \neq -1;$$

$$b) \quad u = (1+x)^2 \quad \text{și} \quad v = x^2 + 2x + 1;$$

$$c) \quad u = \frac{x-1}{x^4-1} \quad \text{și} \quad v = \frac{1}{(x+1)(x^2+1)}, x \neq \pm 1.$$

Întotdeauna va avea loc $u=v$?

4. Să se calculeze $x = \frac{a^3 \sqrt{b}}{c^2}$, unde $x=7.45 \pm 0.001$,

$b=50.46 \pm 0.02$, $c=15.4 \pm 0.03$. Să se estimeze eroarea absolută și eroarea relativă.

5. Fie aritmetica virgulei mobile $\beta=10$, $t=3$ și rotunjirea prin tăiere. Să se calculeze integrala [1]:

$$I_n = \int_0^1 \frac{x^n}{x+5} dx$$

pentru $n=1, 2, 3, 4$, utilizând formula de recurență

$$I_n = \frac{1}{n} - 5I_{n-1}$$

6. Care va fi rezultatul în urma rulării următoarei secvențe de program [37] ?

```

X=0.0
H=0.1
DO 10 I=1.10
X=X+H
10 CONTINUE
Y=1.0-X
WRITE (6.20) X,Y
20 FORMAT (2E20.10)
STOP
END

```

7. Fie date trei sisteme de calcul cu virgulă mobilă în care baza sistemului de numerație $\beta=2$, $\beta=8$ și respectiv $\beta=16$. Cum se reprezintă fracțiile $1/2$, $2/3$ și $3/5$ în aceste sisteme? Explicați rezultatele obținute în urma rulării programului [37]:

$$H=1./2.$$

```
X=2./3.-H
Y=3./5.-H
E=(X+X+X)-H
F=(Y+Y+Y+Y+Y)-H
Q=F/E
WRITE(6,10) Q
10 FORMAT(1X,G20.10)
STOP
END
```

Pentru o inițiere mai aprofundată în materialul prezentat în capitolul de față, recomandăm lucrările [2,15,16,17,22,37].

**Rezolvarea numerică a ecuațiilor algebrice
și transcendente**

2.1 Introducere

În acest capitol este tratată problema rezolvării ecuațiilor neliniare de forma: $f(x)=0$, unde $f(x)$ este un polinom sau o funcție transcendentă.

Dacă $f(x)$ este un polinom sau în urma unor transformări poate fi adusă la forma polinomială, ecuația se numește algebrică.

Exemple:

$$4x^5 - 12x^4 + x^3 - 2x + 10 = 0;$$

$$\sqrt{x+1} = x^2 - 2.$$

Cea de-a doua ecuație este tot algebrică, deoarece prin ridicare la pătrat și ordonare devine:

$$x^4 - 4x^2 - x + 3 = 0.$$

O ecuație algebrică în forma generală se va scrie:

$$a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 = 0 \quad (2.1)$$

cu $n \geq 1$ și coeficienții reali $a_0, a_1, \dots, a_n; a_n \neq 0$.

Orice ecuație algebrică (2.1) are exact n rădăcini, fiecare rădăcină multiplă fiind socotită ca atâtea rădăcini confundate cât arată ordinul ei de multiplicitate. Această afirmație poartă denumirea de *teorema fundamentală a algebrei* și pentru prima dată se întâlnește în lucrările lui Girard și Descartes. Prima demonstrare completă a acestei teoreme a fost dată în anul 1799 de

către Carl Gauss, unul din cei mai mari matematicieni germani din secolul al XIX-lea.

Ecuatiile care nu sunt algebrice se numesc *ecuații transcendente*.

Exemple:

$$x^2 - \sin x - 1 = 0; \quad 2^x - \lg(x+1) = 0.$$

Rezolvarea ecuației $f(x)=0$ (algebrică sau transcendentă) implică parcurgerea a două etape importante:

- 1) separarea rădăcinilor, care constă în determinarea unui interval $[a, b]$, în care este situată o rădăcină reală a ecuației;
- 2) calculul aproximativ ai fiecărei rădăcini și evaluarea erorii care s-a comis, considerând că separarea deja s-a efectuat.

Să presupunem, de exemplu, că în intervalul $[a, b]$ a fost separată rădăcina reală r pentru ecuația respectivă. Atunci graficul funcției $f(x)$ intersectează axa absciselor în punctul $x=r$. În general, prin *rădăcină aproximativă* se înțelege o valoare x_* suficient de apropiată de rădăcina exactă r . Cu alte cuvinte, trebuie să avem:

$$|x_* - r| < \varepsilon$$

unde $\varepsilon > 0$ și suficient de mic.

Se poate defini o rădăcină aproximativă și altfel: numărul

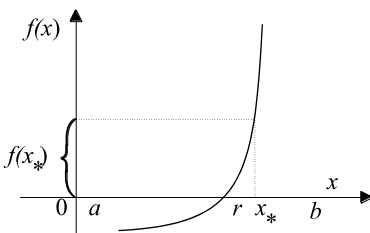


Fig. 2.1

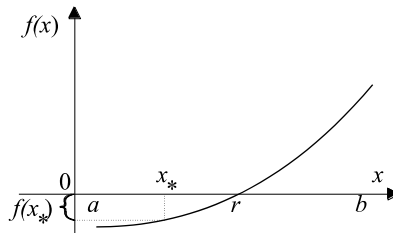


Fig. 2.2

x_* cu proprietatea că $f(x_*)$ este aproape de zero, adică

$$|f(x_*)| < \varepsilon_1, \quad \varepsilon_1 > 0.$$

Menționăm că aceste două moduri de definire a rădăcinii aproximative nu coincid, după cum se poate vedea din fig.2.1 și fig.2.2. Ecuația $f(x)=0$ este echivalentă ecuației $C \times f(x)=0$ oricare ar fi constanta $C \neq 0$. De aceea nu se recomandă de aplicat ultima relație pentru determinarea rădăcinii aproximative, deoarece ea depinde în mare măsură de scara funcției $f(x)$.

În fig.2.1 valoarea lui $|x_* - r|$ este foarte mică, dar $|f(x_*)|$ nu este apropiată de zero. În fig. 2.2 $|f(x_*)|$ este un număr foarte mic, în timp ce $|x_* - r|$ este un număr mare. În general ar fi bine să fie satisfăcute ambele condiții.

2.2 Separarea rădăcinilor

Separarea rădăcinilor poate fi efectuată prin diferite metode. În continuare vom prezenta metoda grafică și metoda analitică de separare a rădăcinilor unei ecuații cu o singură necunoscută.

Metoda grafică. Fie dată o ecuație $f(x)=0$. Construim graficul funcției $y=f(x)$. Abscisele punctelor în care graficul funcției intersectează axa Ox sunt rădăcinile ecuației. Aceste rădăcini pot fi citite într-o primă aproximare pe grafic și pentru

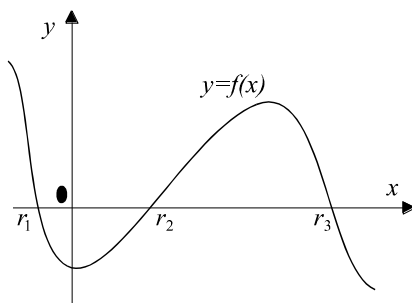


Fig. 2.3

fiecare rădăcină poate fi indicat un interval în care se află (fig.2.3).

Adeseori ecuația dată poate fi prezentată sub forma: $\varphi(x)=g(x)$. Rădăcinile ultimei ecuații sunt abscisele punctelor de intersecție ale curbelor $y=\varphi(x)$ și $y=g(x)$ (fig.2.4).

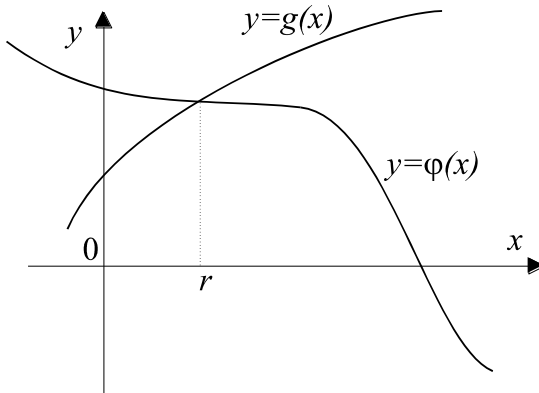


Fig. 2.4

Ecuația $f(x)=0$ este un caz particular al ecuației $\varphi(x)=g(x)$, și anume cazul în care $g(x)$ este funcția constantă zero. Graficul acestei funcții este axa Ox , care are acum rolul curbei $y=g(x)$.

Exemplu: Considerăm ecuația:

$$x^3 - 3x - 1 = 0$$

Graficul funcției $y=x^3 - 3x - 1$ se vede în fig.2.5. Deci ecuația dată are trei rădăcini reale: r_1 , r_2 , și r_3 :

$$r_1 \in (-2, -1), r_2 \in (-1, 0), r_3 \in (1, 2)$$

Ecuația $x^3 - 3x - 1 = 0$ poate fi pusă sub forma:

$$x^3 = 3x + 1.$$

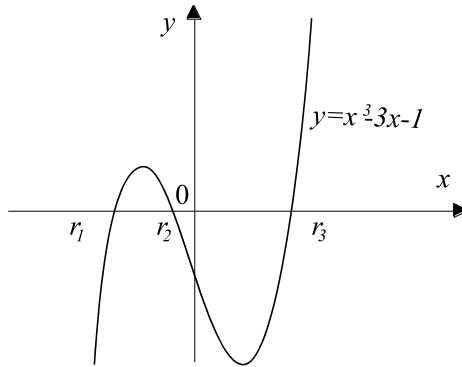


Fig. 2.5

Atunci rădăcinile ei sunt abscisele punctelor de intersecție ale curbei $y = x^3$ cu dreapta $y = 3x + 1$ (vezi fig.2.6).

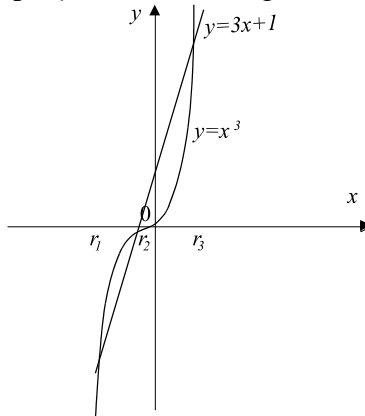


Fig. 2.6

Observație. În fig.2.7 se vede graficul funcției:

$$y=x^3-1.5x^2+0.5.$$

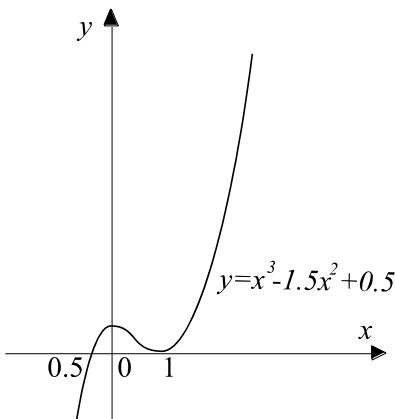


Fig. 2.7

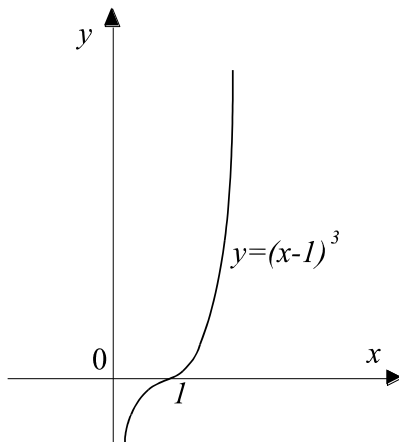


Fig. 2.8

Astfel ecuația $x^3-1.5x^2+0.5=0$ are trei rădăcini: $r_1=-0.5$, $r_2=r_3=1$. În punctul $r=1$ axa Ox este tangentă la graficul funcției date. Se spune că $r=1$ este o rădăcină dublă a ecuației considerate.

În general, dacă în descompunerea funcției în factori apare factorul $(x-r)^k$ și nu apare o putere mai mare a lui $(x-r)$, se spune că numărul r este o rădăcină multiplă de ordinul k . În exemplul de mai sus avem:

$$x^3-1.5x^2+0.5=(x-0.5)(x-1)^2.$$

Ecuația $x^3-3x^2+3x-1=0$ are o rădăcină triplă $r=1$ (fig.2.8), deoarece

$$x^3-3x^2+3x-1=(x-1)^3.$$

În cursul de analiză matematică se demonstrează că un număr r este o rădăcină multiplă de ordinul k a unei ecuații $f(x)=0$, dacă și numai dacă $f(r)=0$ și primele sale $k-1$ derivate

$$f'(r)=f''(r)=\dots=f^{(k-1)}(r)=0 \text{ și } f^{(k)}(r) \neq 0.$$

O rădăcină care nu este multiplă se numește rădăcină simplă.

Metoda analitică. Se știe că o funcție continuă nu trece de la o valoare la alta fără să treacă prin toate valorile intermediare. Fie f o funcție continuă pe un interval închis $[a, b]$ și fie valorile funcției $f(x)$ la capetele acestui interval $f(a)$ și $f(b)$ sunt de semne contrare (pentru a exprima că valorile $f(a)$ și $f(b)$ sunt de semne

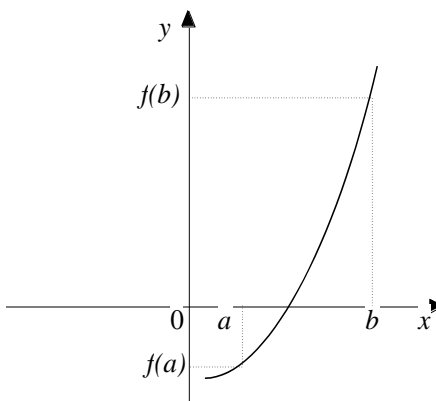


Fig. 2.9

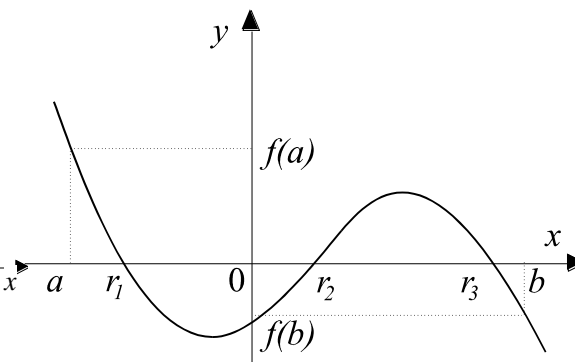


Fig. 2.10

contrare, se scrie $f(a)f(b) < 0$. Atunci există între a și b cel puțin un punct r , astfel încât avem $f(r) = 0$. În fig.2.9 și fig.2.10 se dă o justificare intuitivă a acestei afirmații.

Condiția $f(a)f(b) < 0$ arată că ecuația $f(x) = 0$ are un număr impar de rădăcini (cel puțin una) pe intervalul $[a, b]$.

Condiția de continuitate a funcției $f(x)$, după cum se vede din fig.2.11, este esențială în enunțul afirmației de mai sus.

Dacă $f(a)f(b) > 0$, adică $f(a)$ și $f(b)$ au același semn, aceasta

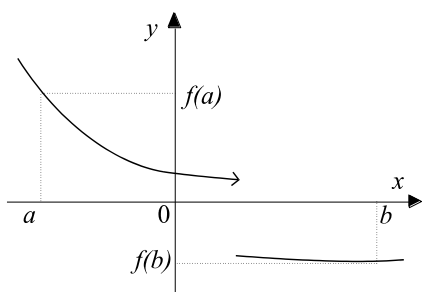


Fig. 2.11

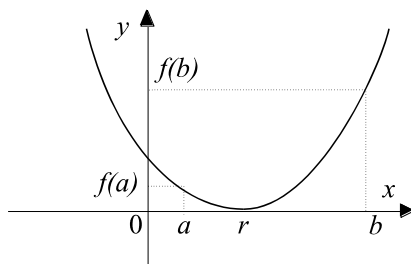


Fig. 2.12

încă nu înseamnă că ecuația $f(x)=0$ nu are pe intervalul $[a, b]$ o rădăcină reală (fig.2.12).

În cursul de analiză matematică se demonstrează, că între două rădăcini reale consecutive ale derivatei funcției $f(x)$ există cel mult o rădăcină reală a ecuației $f(x)=0$. Interpretarea geometrică a acestei afirmații este dată în fig.2.13.

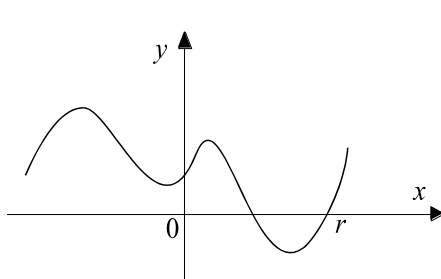


Fig. 2.13

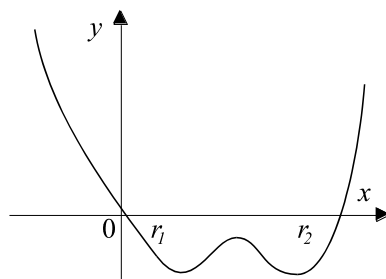


Fig. 2.14

De asemenea, între două rădăcini consecutive ale ecuației $f(x)=0$ există cel puțin o rădăcină a ecuației $f'(x)=0$ (fig.2.14).

Fie $a < x_1 < x_2 < \dots < x_k < b$ rădăcinile ecuației $f'(x)=0$, așezate în ordine crescătoare. Șirul

$$f(a), f(x_1), f(x_2), \dots, f(x_k), f(b)$$

se numește *șirul lui Rolle*.

Ecuația $f(x)=0$ are atâtea rădăcini reale, câte variații de semn prezintă șirul lui Rolle. Într-adevăr, în fiecare interval

$$(a, x_1), (x_1, x_2), \dots, (x_{k-1}, x_k), (x_k, b),$$

conform celor enunțate mai sus, se află cel mult o rădăcină reală a funcției, numai dacă la capetele intervalului funcția ia valori de semne contrare.

De aici rezultă următorul procedeu de separare a rădăcinilor unei ecuații $f(x)=0$. Se scriu în ordine crescătoare rădăcinile derivatei, x_1, x_2, \dots, x_k , precum capetele a și b ale intervalului dat, iar dedesubt valorile corespunzătoare ale funcției.

x	a	x_1	x_2	\dots	x_k	b
$f(x)$	$f(a)$	$f(x_1)$	$f(x_2)$	\dots	$f(x_k)$	$f(b)$

Dacă la capetele unuia dintre intervalele $(a, x_1), (x_1, x_2), \dots, (x_k, b)$ funcția $f(x)$ ia valori de semne contrare (prezintă o variație), ecuația are în acel interval o singură rădăcină; în caz contrar ecuația nu are în acel interval nici o rădăcină.

Exemplu. Fie ecuația algebrică:

$$f(x) = x^4 - x^3 - 2x^2 + 3x - 3 = 0.$$

Derivata $f'(x) = 4x^3 - 3x^2 - 4x + 3 = 4x(x^2 - 1) - 3(x^2 - 1) = (x^2 - 1)(4x - 3)$ se anulează pentru $x_1 = -1, x_2 = 3/4, x_3 = 1$. Șirul lui Rolle este următorul

x	-2	-1	3/4	1	2
$f(x)$	7	-6	-1.98	-2	3

La extremitățile intervalului $(-2, -1)$ funcția are valori de semne contrare, deci în acest interval ecuația are o singură rădăcină; în

intervalul $(-1, 2)$ situația este aceeași; la extremitățile intervalului $(-1, 3/4)$ și $(3/4, 1)$ funcția are același semn, deci în acest interval ecuația nu are nici o rădăcină.

Prin urmare, avem două variații de semn, deci ecuația propusă are două rădăcini reale, r_1 și r_2 , și anume:

$$r_1 \in (-2, -1), r_2 \in (1, 2).$$

Celelalte două rădăcini sunt complexe.

2.4 Metoda înjumătățirii intervalului

Fie ecuația

$$f(x)=0, \tag{2.2}$$

unde funcția $f(x)$ este continuă pe $[a, b]$, are o singură rădăcină reală în acest interval și $f(a)f(b)<0$.

Metoda constă în construirea recurentă a unui șir de subintervale $[a_k, b_k]$ și a unui șir de puncte $c_k=(a_k+b_k)/2$, astfel ca

$$f(a_k)f(b_k)<0 \tag{2.3}$$

și

$$b_k - a_k = \frac{1}{2^k}(b - a). \tag{2.4}$$

Fie $a_0=a$, $b_0=b$ și $c_0=(a_0+b_0)/2$ jumătatea intervalului $[a_0, b_0]$. Dacă $f(c_0)=0$, atunci c_0 este chiar rădăcina căutată. Dacă nu, atunci rădăcina reală se găsește într-unul din intervalele $[a_0, c_0]$, acolo unde funcția ia valori de semne contrare la capetele intervalului. Fie acesta notat cu $[a_1, b_1]$, unde

$$a_1 = \begin{cases} c_0, & \text{dacă } \text{sign } f(a_0) = \text{sign } f(c_0) \\ a_0, & \text{dacă } \text{sign } f(a_0) \neq \text{sign } f(c_0) \end{cases}$$

$$b_1 = \begin{cases} c_0, & \text{dacă } \text{sign } f(b_0) = \text{sign } f(c_0) \\ b_0, & \text{dacă } \text{sign } f(b_0) \neq \text{sign } f(c_0) \end{cases}$$

Evident că $\text{sign } f(a_1) = \text{sign } f(a_0)$ și $\text{sign } f(b_1) = \text{sign } f(b_0)$ și prin urmare $f(a_1)f(b_1) < 0$. Continuând, se obține succesiunea de subintervale

$$[a, b] = [a_0, b_0] \supset [a_1, b_1] \supset [a_2, b_2] \supset \dots \supset [a_k, b_k] \supset \dots$$

la extremitățile cărora funcția ia valori de semne contrare,

$$c_k = \frac{a_k + b_k}{2}$$

și

$$a_{k+1} = \begin{cases} c_k, & \text{dacă } \text{sign } f(a_k) = \text{sign } f(c_k) \\ a_k, & \text{dacă } \text{sign } f(a_k) \neq \text{sign } f(c_k) \end{cases}$$

$$b_{k+1} = \begin{cases} c_k, & \text{dacă } \text{sign } f(b_k) = \text{sign } f(c_k) \\ b_k, & \text{dacă } \text{sign } f(b_k) \neq \text{sign } f(c_k) \end{cases}$$

Se obțin astfel șirul $\{a_k\}$ nedescrescător, mărginit superior, și șirul $\{b_k\}$ nedescrescător mărginit inferior. Trecând la limită în egalitatea (2.4), obținem:

$$r = \lim_{k \rightarrow \infty} a_k = \lim_{k \rightarrow \infty} b_k \quad (2.5)$$

Din (2.3) și (2.5) rezultă că

$$\lim_{k \rightarrow \infty} f(a_k)f(b_k) = [f(r)]^2 \leq 0,$$

de unde concluzionăm că șirurile $\{a_k\}$, $\{b_k\}$ și $\{c_k\}$ sunt convergente către rădăcina ecuației (2.2) și avem:

$$|c_k - r| \leq \frac{b - a}{2^{k+1}}.$$

În fig. 2.15. este ilustrat procesul de înjumătățire a intervalului.

Fie $\varepsilon > 0$ marginea superioară a erorii absolute, care se admite. Dacă $|b_k - a_k| < 2\varepsilon$, atunci c_k aproximează rădăcina r cu eroarea dorită, deoarece $|c_k - r| < \varepsilon$.

Dacă vrem ca precizia să fie foarte mare (eroarea mică),

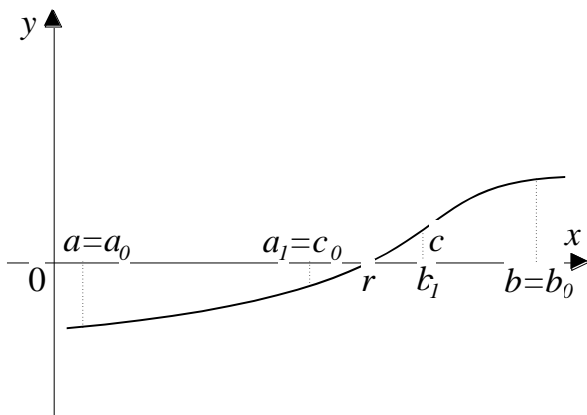


Fig. 2.15. Metoda înjumătățirii intervalului

atunci numărul k de înjumătățiri crește. În cazul când au fost separate mai multe rădăcini, algoritmul expus anterior poate fi utilizat pentru calculul aproximativ al fiecărei rădăcini în parte.

Observație. În programele de calculator operația de înjumătățire se recomandă de scris astfel: $c = a + (b - a) / 2$. În sistemele de calcul cu rotunjirea prin tăiere formula $c = (a + b) / 2$ ne poate scoate în afara intervalului $[a, b]$. De exemplu, fie o aritmetică a virgulei mobile $\beta = 10$ și $t = 3$. Atunci pentru $a = 0.982$ și $b = 0.987$ vom avea :

$$c = (a + b) / 2 = (0.982 + 0.987) / 2 = 1.969 / 2 = 0.9845 < a.$$

2.5 Metoda aproximațiilor succesive

Fie

$$f(x) = 0 \tag{2.6}$$

o ecuație algebrică sau transcendentă care admite o singură rădăcină reală în intervalul $[a, b]$.

Ecuția (2.6) o punem sub forma echivalentă:

$$x = \varphi(x). \quad (2.7)$$

Rescrierea ecuației (2.6) sub forma (2.7) se poate face utilizând deferite artificii de calcul.

Exemplu . Ecuția $x^3 - 2x - 9 = 0$ poate fi scrisă:

$$x = \sqrt[3]{2x + 9},$$

$$x = \frac{2x + 9}{x^2}, \quad x = \frac{9}{x^2 - 2}.$$

Plecând de la o valoare inițială arbitrară $x_0 \in [a, b]$, generăm șirul x_k conform regulii:

$$x_{k+1} = \varphi(x_k), \quad k=0, 1, 2, \dots, \quad (2.8)$$

adică $x_1 = \varphi(x_0)$, $x_2 = \varphi(x_1)$, \dots , $x_k = \varphi(x_{k-1})$, \dots

Șirul definit mai sus prin relația (2.8) se numește *șir de iterare* .

Să presupunem că există $\lim_{k \rightarrow \infty} x_k = r$. În acest caz se spune că șirul de iterare este *convergent* și *converge către r*.

Dacă $\varphi(x)$ este o funcție continuă, atunci avem:

$$r = \lim_{k \rightarrow \infty} x_{k+1} = \lim_{k \rightarrow \infty} \varphi(x_k) = \varphi(r),$$

deci limita r este chiar rădăcina ecuației (2.7).

Din punct de vedere geometric, rădăcina reală r este abscisa punctului de intersecție a curbei $y = \varphi(x)$ cu dreapta $y = x$. Modul cum șirul aproximațiilor succesive $x_0, x_1, \dots, x_k, \dots$ conduce spre soluția exactă este ilustrat în fig.2.16 și fig.2.17 (în funcție de forma curbei $y = \varphi(x)$).

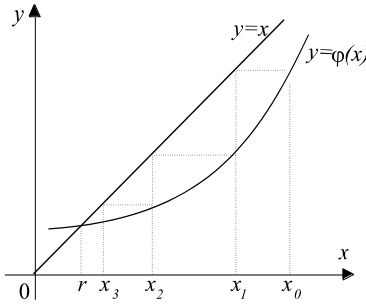


Fig. 2.16

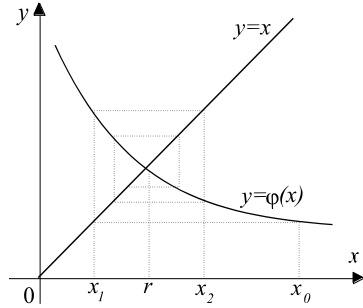


Fig. 2.17

Dacă graficul funcției $\varphi(x)$ are forma din fig.2.18 sau din fig.2.19, atunci șirul de iterație $\{x_k\}$ nu converge către rădăcina căutată, deci șirul este divergent și se spune că *metoda diverge*.

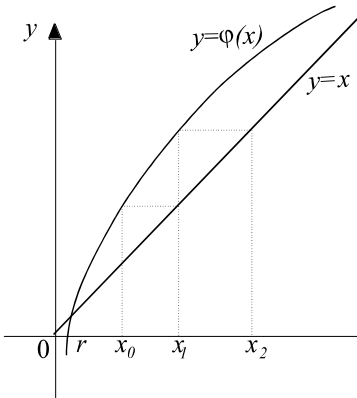
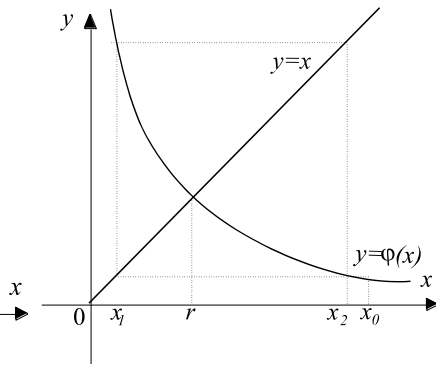


Fig. 2.18



Fif. 2.19

Deci șirul de iterație poate fi divergent chiar dacă x_0 se alege oricât de apropiat de rădăcină. O condiție suficientă de convergență este dată de următoarea teoremă.

Teoremă. Fie funcția $\varphi(x)$ definită pe intervalul $[a, b]$ și $\varphi(x) \in [a, b]$ pentru $\forall x \in [a, b]$. Dacă funcția φ este derivabilă și derivata sa φ' satisface inegalitatea $|\varphi'(x)| \leq \alpha < 1$, oricare ar fi $x \in [a, b]$, atunci ecuația (2) are în $[a, b]$ o singură rădăcină reală r , putem forma șirul de iterare $x_0, x_1, \dots, x_k, \dots$ după regula (2.8), astfel încât $x_k \in [a, b]$ pentru $k=0, 1, 2, \dots$ și acest șir converge către rădăcina r . În plus eroarea este evaluată prin

$$|x_k - r| \leq \frac{\alpha}{1 - \alpha} |x_k - x_{k-1}| \leq \frac{\alpha^k}{1 - \alpha} |x_1 - x_0|, \quad k \geq 1.$$

Demonstrație. Datorită faptului că $\varphi(x) \in [a, b]$, $\forall x \in [a, b]$, rezultă că $\varphi(a) \geq a$ și $\varphi(b) \leq b$. Punem $g(x) = \varphi(x) - x$. Atunci $g(a) \geq 0$, $g(b) \leq 0$ și deoarece $g(x)$ este continuă, rezultă că există cel puțin un punct $r \in [a, b]$, astfel încât $g(r) = 0$, adică r este rădăcina ecuației inițiale (2.7) în intervalul $[a, b]$. Să considerăm acum diferența:

$$x_{k+1} - r = \varphi(x_k) - \varphi(r).$$

Aplicând teorema lui Lagrange, se obține:

$$x_{k+1} - r = \varphi'(\xi_k)(x_k - r), \quad \xi_k \in (x_k, r).$$

Deoarece

$$|\varphi'(x)| \leq \alpha, \quad \forall x \in [a, b],$$

avem:

$$|x_{k+1} - r| \leq \alpha |x_k - r|.$$

Atunci, pentru orice $k \geq 0$, avem:

$$|x_1 - r| \leq \alpha |x_0 - r|,$$

$$|x_2 - r| \leq \alpha^2 |x_0 - r|,$$

.....

$$|x_{k+1} - r| \leq \alpha^{k+1} |x_0 - r|.$$

Din ultima relație rezultă (fiindcă $0 \leq \alpha < 1$):

$$\lim_{k \rightarrow \infty} x_k = r,$$

Dar $\lim_{k \rightarrow \infty} x_{k+1} = \varphi(\lim_{k \rightarrow \infty} x_k)$, întrucât funcția $\varphi(x)$ este continuă. Prin urmare, $r = \varphi(r)$.

Să arătăm că rădăcina astfel obținută este unică. Într-adevăr, fie s o altă rădăcină a ecuației (2.7) în intervalul $[a, b]$. Atunci $s = \varphi(s)$. Evaluăm diferența:

$$r - s = \varphi(r) - \varphi(s) = \varphi'(\xi)(r - s), \quad \xi \in (r, s) \subset [a, b].$$

Deci

$$(r - s)[1 - \varphi'(\xi)] = 0,$$

de unde rezultă că $r = s$.

Pentru a analiza cum se propagă eroarea în calcule vom scrie:

$$x_{k-1} - r = x_{k-1} - x_k + \varphi(x_{k-1}) - \varphi(r) = x_{k-1} - x_k + \varphi'(\xi)(x_{k-1} - r), \quad \xi \in (x_{k-1}, r).$$

Datorită inegalității din enunțul teoremei, se obține:

$$|x_{k-1} - r| \leq |x_{k-1} - x_k| + \alpha |x_{k-1} - r|,$$

Sau

$$|x_{k-1} - r| \leq \frac{1}{1 - \alpha} |x_k - x_{k-1}|.$$

Prin urmare:

$$|x_k - r| = |\varphi(x_{k-1}) - \varphi(r)| \leq \alpha |x_{k-1} - r| \leq \frac{\alpha}{1 - \alpha} |x_k - x_{k-1}|.$$

Teorema este demonstrată.

Deci, pentru rezolvarea ecuației (2.6) prin metoda aproximațiilor succesive va trebui să o aducem, în prealabil, la forma (2.7), alegându-l pe $\varphi(x)$ în mod special, ca să se satisfacă condiția de convergență.

Exemplu. Fie dată ecuația:

$$x^3 - 2x - 9 = 0.$$

Prin metoda grafică sau prin metoda analitică se stabilește că ecuația admite o singură rădăcină reală în intervalul (2, 3). Rescriem ecuația sub forma echivalentă:

$$x = \sqrt[3]{2x+9}$$

Pentru a verifica condiția de convergență, calculăm derivata:

$$\varphi'(x) = \frac{2}{3} \cdot \frac{1}{\sqrt[3]{(2x+9)^2}}$$

Condiția de convergență $|\varphi'(x)| < 1$ este îndeplinită pentru intervalul (2, 3) și deci șirul de iterare este dat de

$$x_{k+1} = \sqrt[3]{2x_k + 9}, \quad k=0, 1, 2, \dots$$

cu valoarea inițială (de start) $x_0 \in (2, 3)$.

De remarcat că dacă am fi rearanjat ecuația inițială

$$x^3 - 2x - 9 = 0$$

în felul următor: $x = x^3 - x - 9$ și deci am fi folosit șirul de iterare: $x_{k+1} = x_k^3 - x_k - 9$, atunci metoda aproximațiilor succesive diverge, condiția suficientă de convergență nefiind îndeplinită:

$$|\varphi'(x)| = |3x^2 - 1| > 1, \quad \forall x \in (2, 3).$$

Din teorema demonstrată mai sus rezultă că pentru determinarea rădăcinii aproximative x_* cu eroarea $\varepsilon > 0$ procesul de calcul îl vom opri când

$$\frac{\alpha}{1-\alpha} |x_{k+1} - x_k| < \varepsilon.$$

Acest criteriu pentru terminarea calculelor necesită aprecierea parametrului subunitar α , care nu se cunoaște, în mod general, apriori. În 2.5 ne vom ocupa de examinarea altor criterii de stopare în metodele iterative de rezolvare a ecuațiilor algebrice și transcendente.

În încheiere remarcăm că metoda aproximațiilor succesive se utilizează cu succes și pentru studiul aplicațiilor numite contracții în așa-numite spații metrice.

2.5 Criterii de oprire în metodele iterative

În metodele iterative oprirea procesului de calcul se face prin trunchierea șirului de iterare $\{x_k\}$ la un indice m , astfel încât termenul x_m să constituie aproximația satisfăcătoare a rădăcinii exacte. Definirea apropierei rădăcinii aproximative x_* de rădăcina exactă r este o chestiune delicată și e departe a fi perfectă.

Presupunem că rădăcina simplă r a ecuației $f(x)=0$ este izolată într-un interval $[a, b]$. Vom deduce o estimare a erorii, care să fie independentă de metoda de rezolvare a ecuației considerate. Fie funcția $f(x)$ continuă și derivabilă pe intervalul $[a, b]$ și fie

$$m = \min |f'(x)| > 0, \quad x \in (a, b).$$

Aplicând teorema lui Lagrange, se obține:

$$f(x_k) - f(r) = (x_k - r) f'(\xi), \quad \xi \in (x_k, r) \subset [a, b],$$

de unde rezultă

$$|x_k - r| \leq \frac{|f(x_k)|}{m}.$$

Această relație evidențiază, în primul rând, că dacă $|f'(r)|$ este mic, atunci $|f(x_k)|/m$ este mare și perturbații slabe în x_k pot produce perturbații mari în rădăcină; în acest caz se spune că problema determinării lui r este rău condiționată. Pentru ilustrare a se vedea fig.2.2 din paragraful 2.1. În al doilea rând, dacă dorim să determinăm rădăcina r cu eroarea prescrisă $\varepsilon > 0$ am putea opri iterațiile de îndată ce

$$|f(x_k)| < \varepsilon m,$$

ceea ce presupune cunoașterea majorantei m a derivatei $f'(x)$.

Deoarece derivatele, în caz general, sunt greu de estimat, înclinăm să facem câteva iterații în plus, decât să folosim formulele de mai sus sau alte formule complicate de evaluare a erorii.

În practică, rezolvând problema la calculatorul electronic, putem folosi următorul criteriu de stopare. Fie

$$|f(x_{k+1})| < \varepsilon_1, \tag{2.9}$$

unde $\varepsilon_1 > 0$ și este suficient de mic; de exemplu, $\varepsilon_1 = \sqrt{\varepsilon_M}$, unde ε_M este unitatea de rotunjire a calculatorului. Atunci putem termina iterațiile și accepta x_{k+1} ca rădăcină aproximativă, dacă

$$|x_{k+1} - x_k| < \varepsilon_2. \quad (2.10)$$

Aici $\varepsilon_2 > 0$ și se alege, astfel încât $\varepsilon_2 \geq \varepsilon_1$. În cazul când se verifică inegalitățile (2.9) și (2.10) cantitatea $|x_{k+1} - x_k|$ este, de regulă, o bună estimare a lui $|x_k - r|$. Menționăm că în practică cel mai des este utilizat în calitate de criteriu de oprire a algoritmului cel care verifică doar inegalitatea (2.10) cu $\varepsilon_2 > 0$ și suficient de mic. Pentru aprofundarea problemelor acestui paragraf recomandăm referințele [15,17,20].

2.6 Metoda lui Newton (metoda tangentei)

Fie ecuația algebrică sau transcendentă $f(x)=0$ care admite o

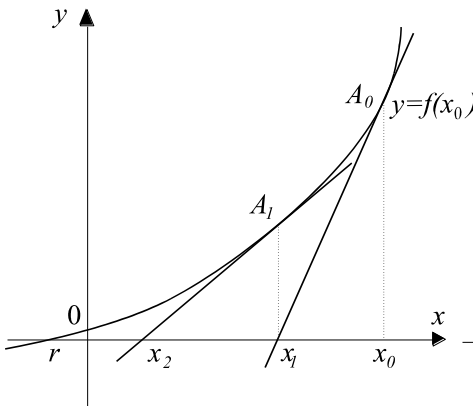


Fig. 2.20

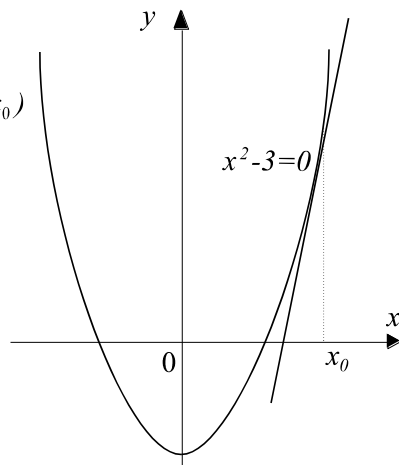


Fig. 2.21

singură rădăcină reală r în intervalul $[a, b]$. Să presupunem în plus că derivatele $f'(x)$ și $f''(x)$ păstrează un semn constant pe intervalul $[a, b]$. Ducem în punctul A_0 (fig.2.20) tangenta la curba $y=f(x)$.

Punctul x_1 în care tangenta întretaie axa Ox ne dă o valoare aproximativă a rădăcinii. Deoarece

$$x_1 = x_0 - p \quad \text{și} \quad f'(x_0) = \frac{f(x_0)}{p}.$$

vom avea:

$$f'(x_0)p = f(x_0).$$

Prin urmare, abscisa punctului de intersecție a acestei tangente cu axa Ox este

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

Considerăm punctul de coordonate $A_1(x_1, f(x_1))$ și construim tangenta la curbă în acest punct. În mod analog ca mai sus se arată că

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

Procedeul se va repeta în mod asemănător. Se obține metoda tangențelor definită de următoarea formulă de iterare:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k=0, 1, 2, \dots \quad (2.11)$$

Această metodă se mai numește și *metoda lui Newton*.

Exemplu. Să aplicăm metoda lui Newton la calculul $\sqrt{3}$. Pentru aceasta scriem ecuația $x^2 - 3 = 0$ (fig.2.21). Formula de iterare a lui Newton este în acest caz

$$x_{k+1} = x_k - \frac{x_k^2 - 3}{2x_k},$$

sau

$$x_{k+1} = \frac{1}{2} \left(x_k + \frac{3}{x_k} \right).$$

Alegând $x_0=2$, obținem:

$$x_1=1,75; x_2=1,732; x_3=1,7320508.$$

Se observă că avem o convergență rapidă a șirului x_1, x_2, \dots către $\sqrt{3}$. Acest lucru nu este întâmplător. Analizăm cazul general. Fie dat $a>0$ și să se găsească rădăcina ecuației:

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

Atunci

$$x_{k+1} = x_k - \frac{x_k^2 - a}{2x_k} = \frac{x_k}{2} + \frac{a}{2x_k}$$

și

$$x_{k+1} - \sqrt{a} = \frac{x_k}{2} + \frac{a}{2x_k} - \sqrt{a} = \frac{(x_k - \sqrt{a})^2}{2x_k},$$

ori, utilizând eroarea relativă:

$$\frac{x_{k+1} - \sqrt{a}}{\sqrt{a}} = \left(\frac{\sqrt{a}}{2x_k} \right) \left(\frac{x_k - \sqrt{a}}{\sqrt{a}} \right)^2.$$

Astfel, în cazul rezolvării ecuației $x^2 - a = 0$, putem spune că fiecare iterație în metoda lui Newton, în mod aproximativ, ridică eroarea la pătrat. Deci numărul cifrelor zecimale (sau binare) corecte aproape se dublează, la fiecare iterație. Acest rezultat este adevărat și în cazul rezolvării ecuațiilor scrise în forma generală. Se mai spune că metoda lui Newton este cu convergența pătratică sau că este o metodă de ordinul al doilea.

Teoremă. Fie funcția $f(x)$ definită și de două ori derivabilă pe $[a, b]$. Presupunem că există $m>0, M<\infty$, astfel încât

$$|f'(x)| \geq m > 0, \quad |f''(x)| \leq M < \infty, \quad \forall x \in [a, b]$$

și $r \in [a, b]$ este rădăcina ecuației $f(x)=0$. Atunci șirul de iterare determinat de relația (2.11) converge către r dacă aproximația inițială x_0 este aleasă într-o vecinătate a rădăcinii r . Eroarea este

estimată de relația:

$$|x_{k+1}-r| \leq C |x_k-r|^2.$$

Demonstrație. Se observă că metoda lui Newton este un caz particular al metodei aproximațiilor succesive cu funcția:

$$\varphi(x) = x - \frac{f(x)}{f'(x)}.$$

Se verifică imediat că $r = \varphi(r)$ și $\varphi'(r) = 0$. Deci putem afirma că într-o vecinătate a rădăcinii r se îndeplinește condiția $|\varphi'(x)| \leq \alpha < 1$. Rezultă că șirul (2.11) converge (către rădăcina r), dacă aproximația inițială x_0 este aleasă suficient de aproape de rădăcină.

Pentru a analiza viteza de convergență se evaluează diferența:

$$|x_{k+1}-r| = \left| x_k - r - \frac{f(x_k)}{f'(x_k)} \right| = \frac{1}{|f'(x_k)|} |f'(x_k)(x_k-r) - (f(x_k) - f(r))|.$$

Prin aplicarea repetată a teoremei de medie a lui Lagrange, ținând seama de condițiile din enunț, rezultă:

$$|x_{k+1}-r| \leq \frac{1}{m} |(f'(x_k) - f'(\xi_1))(x_k-r)| = \frac{1}{m} |f''(\xi_2)| |x_k - \xi_1| |x_k - r|,$$

unde

$$\xi_1 = r + \theta_1(x_k - r), \quad \xi_2 = \xi_1 + \theta_2(x_k - \xi_1), \quad 0 \leq \theta_1, \quad \theta_2 \leq 1.$$

Avem

$$|x_{k+1}-r| \leq C |x_k-r|^2, \quad C = \frac{(1-\theta_1)M}{m}$$

și teorema este demonstrată.

O dificultate în aplicarea metodei lui Newton (și în general a unei metode iterative) o reprezintă alegerea aproximației inițiale x_0 . Să presupunem, de exemplu, că se rezolvă ecuația $\arctg x = 0$ (fig.2.22).

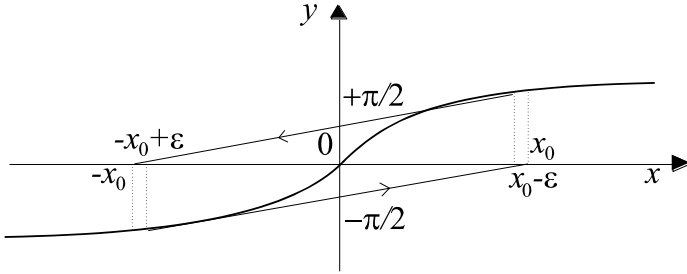


Fig. 2.22

Alegem ca aproximație inițială punctul $x_0 \in [1,39; 1,40]$. Atunci tangenta dusă în punctul $(x_0, f(x_0))$ va trece prin punctul de abscisă $-x_0$ și apoi tangenta dusă în $(-x_0, f(-x_0))$ trece prin x_0 . În acest caz, șirul x_k începe să "cicleze". Se numește *zonă de convergență* a rădăcinii r mulțimea tuturor aproximațiilor inițiale x_0 pentru care șirul iterativ $\{x_k\}$ tinde către r . În fig.2.22 zona de convergență a rădăcinii $r=0$ este intervalul $(-x_0+\varepsilon, x_0-\varepsilon)$. Alegerea aproximației inițiale în afara zonei de convergență a rădăcinii dorite nu permite să găsim această rădăcină (putem însă nimeri, eventual, în zona de convergență a altei rădăcini).

În cazul ecuației $x^3 - x = 0$ zona de convergență a rădăcinii $r=0$ este intervalul deschis $(-\frac{\sqrt{5}}{5}, \frac{\sqrt{5}}{5})$ (fig.2.23). Pentru

$$x_0 = \pm \frac{\sqrt{5}}{5}$$

vom avea $x_1 = -x_0$, $x_2 = -x_1 = x_0$, $x_3 = -x_2 = -x_0$, ... , un șir care "ciclează". Dacă $x_0 = \pm \sqrt{3}/3$, atunci $f'(x_0) = 0$ și deci tangenta la curba dată în aceste puncte este paralelă cu axa Ox . Alegând

aproximația inițială $x_0 < -\sqrt{3}/3$, vom avea că șirul iterativ $\{x_k\}$ tinde către rădăcina $r=-1$; alegând $x_0 > \sqrt{3}/3$, se asigură convergența către rădăcina $r=1$. În practică se recomandă de a alege valoarea aproximativă inițială x_0 astfel ca

$$f(x_0)f'(x_0) > 0.$$

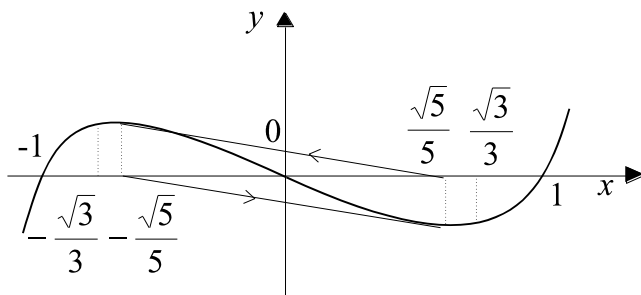


Fig. 2.23

Această alegere a punctului de start asigură aplicarea metodei lui Newton, după cum se observă din fig.2.20-fig.2.23.

Drept criteriu de oprire al iterațiilor poate servi următorul: iterațiile se întrerup atunci când $|x_{k+1}-x_k|$ și (sau) $|f(x_{k+1})|$ devine mai mic decât $\varepsilon > 0$, o eroare maximă pe care o fixăm pentru determinarea rădăcinii.

Am văzut că în cazul rădăcinilor simple ($f'(r) \neq 0$) metoda lui Newton are gradul doi de convergență. Dacă r este o rădăcină multiplă, atunci convergența șirului $\{x_k\}$ este liniară.

Exemplu. Fie dată ecuația $x^2=0$ cu rădăcina dublă $r=0$. Potrivit metodei lui Newton putem scrie:

$$x_{k+1} = x_k - \frac{x_k^2}{2x_k} = \frac{1}{2}x_k.$$

Prin urmare

$$|x_{k+1} - r| = \frac{1}{2} |x_k - r|.$$

Dacă se cunoaște gradul de multiplicitate a rădăcinii, atunci putem accelera convergența șirului construit prin metoda lui Newton. Dacă r este o rădăcină multiplă de gradul p , adică

$$f'(r) = f''(r) = \dots = f^{(p-1)}(r) = 0, \quad f^{(p)}(r) \neq 0,$$

se recomandă de a efectua calculele conform formulei de iterare:

$$x_{k+1} = x_k - p \frac{f(x_k)}{f'(x_k)}, \quad k=0, 1, \dots$$

2.7 Metoda secantei

În metoda lui Newton fiecare "pas" necesită calculul valorilor funcției f și ale derivatei f' în punctele x_k . Există funcții pentru care calculul valorilor derivatei este dificil sau aproape imposibil, de exemplu, când nu se cunoaște expresia analitică a lui f , ci este definită cu ajutorul unui tabel de valori. Pentru astfel de funcții pentru care derivatele se evaluează greu, o alegere mai bună este metoda secantei.

Metoda secantei se deduce din metoda lui Newton, înlocuind derivata:

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$

Obținem:

$$x_{k+1} = x_k - f(x_k) \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}, \quad k=1, 2, \dots \quad (2.12)$$

Pentru startul iterațiilor în metoda secantei sunt necesare de două aproximații inițiale x_0 și x_1 . Valoarea x_{k+1} este abscisa punctului de intersecție dintre secanta

$$\frac{x - x_k}{x_{k-1} - x_k} = \frac{y - f(x_k)}{f(x_{k-1}) - f(x_k)}$$

care trece prin punctele $(x_{k-1}, f(x_{k-1}))$ și $(x_k, f(x_k))$ și Ox (fig.2.24); de aici și denumirea metodei.

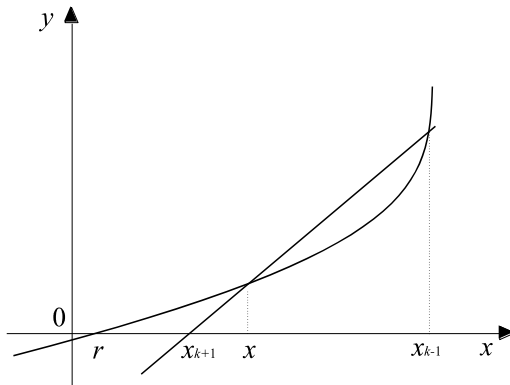


Fig. 2.24. Metoda secantei

Un criteriu de oprire a algoritmului, după cum am stabilit în 2.5, este ca să fie verificate inegalitățile:

$$|f(x_{k+1})| < \varepsilon_1, \quad |x_{k+1} - x_k| < \varepsilon_2, \quad \varepsilon_1 \geq \varepsilon_2 > 0.$$

Dacă se cunoaște

$$m = \min_{x \in [a, b]} |f'(x)| > 0$$

și dorim să determinăm rădăcina cu eroarea $\varepsilon > 0$, vom întrerupe calculele când $|f(x_k)|/m < \varepsilon$. În cazul când se știe și constanta

$$M = \max_{x \in [a, b]} |f''(x)| < \infty$$

calculele le vom opri, dacă

$$\frac{M}{2m} |x_{k+1} - x_k| |x_k - x_{k-1}| < \varepsilon,$$

ceea ce garantează inegalitatea $|x_{k+1} - r| < \varepsilon$.

Să studiem în continuare viteza de convergență a metodei secantei, presupunând că se îndeplinesc condițiile: $f(r) = 0$, $f'(r) \neq 0$,

iar $f'(x)$ sunt funcții continue și păstrează un semn constant pe intervalul $[a, b]$. Pentru aceasta se dezvoltă funcția $f(x)$ în serie Taylor în vecinătatea punctului $x=r$:

$$f(x) = f(r) + (x-r)f'(r) + \frac{(x-r)^2}{2} f''(r) + \dots$$

Notăm prin $\varepsilon_{k-1} = x_{k-1} - r$, $\varepsilon_k = x_k - r$, $\varepsilon_{k+1} = x_{k+1} - r$. În dezvoltarea Taylor vom pune succesiv $x = x_{k-1}$ și $x = x_k$, apoi o trunchiem după termenul al doilea (ținând seama că $f(r) = 0$):

$$f(x_{k-1}) \approx \varepsilon_{k-1} f'(r) + \frac{\varepsilon_{k-1}^2}{2} f''(r),$$

$$f(x_k) \approx \varepsilon_k f'(r) + \frac{\varepsilon_k^2}{2} f''(r).$$

Înlocuind în formula de iterare (2.13), obținem:

$$\varepsilon_{k+1} \approx \varepsilon_k - \frac{[\varepsilon_k f'(r) + \frac{\varepsilon_k^2}{2} f''(r)](\varepsilon_k - \varepsilon_{k-1})}{(\varepsilon_k - \varepsilon_{k-1})f'(r) + \frac{\varepsilon_k^2 - \varepsilon_{k-1}^2}{2} f''(r)},$$

sau

$$\varepsilon_{k+1} \approx \frac{1}{2} \frac{f''(r)}{f'(r)} \varepsilon_k \varepsilon_{k-1}.$$

Deci putem scrie:

$$x_{k+1} - r \approx a(x_k - r)(x_{k-1} - r), \quad a = \frac{1}{2} \frac{f''(r)}{f'(r)}. \quad (2.13)$$

Relația de recurență (2.13) o vom pune sub forma:

$$x_{k+1} - r = a^\alpha (x_k - r)^\beta,$$

unde α și β urmează a fi determinate. Substituind această formă în (2.13), vom obține (pentru demonstrație vezi [18,27]):

$$\alpha\beta = 1, \quad \beta^2 - \beta - 1 = 0. \quad (2.14)$$

Se va lua numai rădăcina pozitivă a ecuației pătrate (2.14), deoarece numai ea garantează convergența șirului $\{x_k\}$. Prin urmare, pentru metoda secantei vom avea:

$$x_{k+1} - r \approx a^{1/\beta} (x_k - r)^\beta$$

unde $\beta = \frac{1}{2}(\sqrt{5} + 1) \approx 1.62$, $1/\beta \approx 0.62$.

În metoda lui Newton (vezi paragraful 2.6) $\beta=2$ și deci metoda lui Newton converge mai repede decât metoda secantei. Pe de altă parte, metoda lui Newton reclamă necesitatea evaluării funcției și a derivatei sale, iar metoda secantei necesită numai calculul funcției. De aceea la aceeași cantitate de operații în metoda secantei se poate face de două ori mai mulți "pași" și deci se poate obține rădăcina cu o precizie mai înaltă.

Să observăm că la fiecare pas nou în metoda secantei se calculează o singură valoare nouă pentru funcția f . Formula (2.12) se mai poate pune sub forma:

$$x_{k+1} = \frac{x_{k-1}f(x_k) - x_k f(x_{k-1})}{f(x_k) - f(x_{k-1})},$$

care nu se recomandă la programare, deoarece dacă $f(x_k) \cdot f(x_{k-1}) > 0$ și $x_k \approx x_{k-1}$, atunci poate avea loc o neutralizare a termenilor.

2.8 Rezolvarea ecuațiilor algebrice

Considerăm ecuația algebrică

$$P_n(x) \equiv a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 = 0, \quad (2.15)$$

unde coeficienții $a_0, a_1, a_2, \dots, a_n$ sunt reali, $a_n > 0$.

2.8.1 Proprietățile ecuațiilor algebrice

Fără a putea rezolva ecuația algebrică (2.15) se poate stabili câte rădăcini are ecuația, unele relații între coeficienții polinomului $P_n(x)$ și rădăcinile ecuației (2.15) etc. În cele ce urmează vom aduce unele dintre aceste proprietăți:

1. O ecuație algebrică (2.15) de gradul n are exact n rădăcini (printre care pot fi atât reale, cât și complexe), fiecare rădăcină multiplă fiind considerată ca atâtea

rădăcini confundate cât indică ordinul ei de multiplicitate.

2. Între coeficienții ecuației algebrice (2.15) și rădăcinile ei r_1, r_2, \dots, r_n există relațiile (formulele lui Viète):

$$r_1 + r_2 + \dots + r_n = -\frac{a_{n-1}}{a_n},$$

$$r_1 r_2 + r_1 r_3 + \dots + r_{n-1} r_n = \frac{a_{n-2}}{a_n},$$

$$r_1 r_2 r_3 + r_1 r_3 r_4 + \dots + r_{n-2} r_{n-1} r_n = -\frac{a_{n-3}}{a_n},$$

.....

$$r_1 r_2 \dots r_n = (-1)^n \frac{a_0}{a_n}.$$

3. Rădăcinile complexe ale ecuației algebrice (2.15) cu coeficienți reali sunt conjugate două câte două. (Numărul $a-ib$, $i = \sqrt{-1}$, se numește conjugatul lui $a+ib$).
4. Dacă o ecuație algebrică (2.15) cu coeficienți raționali admite ca rădăcină un irațional pătratic $m+n\sqrt{p}$ ($m, n \in \mathbb{Q}$, $n \neq 0$, $p \in \mathbb{N}$), ea admite și conjugatul său, $m-n\sqrt{p}$, ca rădăcină.
5. Ecuația algebrică (2.15) de ordin n par și cu $a_0 < 0$ are cel puțin două rădăcini reale de semne diferite.
6. Rădăcinile reale și complexe ale ecuației algebrice (2.15) sunt situate în inelul circular $R_1 < |x| < R_2$, unde

$$R_1 = \frac{1}{1 + \frac{1}{|a_0|} \max_{1 \leq k \leq n} |a_k|}; \quad R_2 = 1 + \frac{1}{|a_n|} \max_{0 \leq k \leq n-1} |a_k|.$$

Aceste și alte proprietăți ale ecuațiilor algebrice sunt tratate în manualele de algebră superioară.

2.8.2 Schema lui Horner

Schema lui Horner constituie un procedeu efectiv de calcul al valorii unui polinom și al derivatelor lui.

Considerăm polinomul de gradul n :

$$P_n(x) \equiv a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0.$$

Pentru stabilirea schemei lui Horner se transcrie polinomul astfel:

$$P_n(x) = a_0 + x(a_1 + x(a_2 + \dots + x((a_{n-1} + x a_n) \dots))).$$

Deci putem afla valoarea acestui polinom în punctul $x = \xi$, calculând succesiv mărimile:

$$b_n = a_n,$$

$$b_{n-1} = a_{n-1} + \xi b_n = a_{n-1} + \xi a_n,$$

$$b_{n-2} = a_{n-2} + \xi b_{n-1} = a_{n-2} + \xi(a_{n-1} + \xi a_n),$$

.....

$$b_0 = a_0 + \xi b_1 = a_0 + \xi(a_1 + \dots + \xi(a_{n-1} + \xi a_n) \dots) = P_n(\xi).$$

Această schemă de obținere a șirului finit $\{b_i\}$ se numește *schema lui Horner*. Ea necesită cel mult $2n$ operații aritmetice. Se demonstrează, că în cazul general, când toți coeficienții polinomului dat $P_n(x)$ sunt diferiți de zero, nu există o schemă mai eficientă de calcul, decât cea a lui Horner.

Câtul împărțirii lui $P_n(x)$ la $x - \xi$ este dat de polinomul

$$P_{n-1}(x) \text{ de gradul } n-1 \text{ cu coeficienții } b_1, b_2, \dots, b_n:$$

$$P_{n-1}(x) = b_n x^{n-1} + b_{n-1} x^{n-2} + \dots + b_2 x + b_1,$$

adică

$$P_n(x) = (x - \xi)P_{n-1}(x) + P_n(\xi).$$

Ultima relație ne dă:

$$P_{n-1}(x) = \frac{P_n(x) - P_n(\xi)}{x - \xi}.$$

Trecând la limită în această identitate, obținem:

$$P'_n(\xi) = P_{n-1}(\xi).$$

Prin urmare, cu ajutorul lui $P_{n-1}(x)$ putem calcula valoarea derivatei polinomului $P_n(x)$ în punctul $x = \xi$. Valoarea derivatei de asemenea se obține cu ajutorul schemei lui Horner:

$$\begin{aligned} c_n &= b_n = a_n, \\ c_{n-1} &= b_{n-1} + \xi c_n = a_{n-1} + 2a_n \xi, \\ c_{n-2} &= b_{n-2} + \xi c_{n-1} = a_{n-2} + 2a_{n-1} \xi + 3a_n \xi^2, \\ &\dots\dots\dots \\ c_1 &= b_1 + \xi c_2 = a_1 + 2a_2 \xi + \dots + na_{n-1} \xi^{n-1} = P'_n(\xi). \end{aligned}$$

Procedeul poate fi continuat obținându-se valorile oricărei derivate a lui $P_n(x)$ într-un punct fixat $x = \xi$.

Pentru ușurința calculelor coeficienții polinoamelor

$$P'_n, P''_n, K, P_n^{(n)}$$

se pot determina cu ajutorul schemei concise a lui Horner.

Schema concisă a lui Horner

a_n	a_{n-1}	a_{n-2}	...	a_2	a_1	a_0
↓	↓	↓		↓	↓	↓
(o)	ξb_n	ξb_{n-1}	...	ξb_3	ξb_2	ξb_1
↓	↓ ↗	↓ ↗		↓	↓	↓ ↗
b_n	b_{n-1}	b_{n-2}	...	b_2	b_1	$b_0 = P_n(\xi)$
↓	↓	↓		↓	↓	
(o)	ξc_n	ξc_{n-1}	...	ξc_3	ξc_2	
↓	↓ ↗	↓ ↗		↓	↓	↓ ↗
c_n	c_{n-1}	c_{n-2}	...	c_2	$c_1 = 1/1! P'_n(\xi)$	
·	·	·	...	·	$n(\xi)$	
·	·	·	...	·		
·	·	·	...	·		
s_n	s_{n-1}	$s_{n-2} = \frac{1}{(n-2)!} \times$				
↓	↓					
(0)	ξq_n	$\times P_n^{(n-2)}(\xi)$				
↓	↓ ↗	↓				
q_n	$q_{n-1} = \frac{1}{(n-1)!} P_n^{(n-1)}(\xi)$					
↓						
(0)	↓ ↗					
↓						
$r_n = \frac{1}{n!} P_n^{(n)}(\xi)$						

2.8.3 Metoda lui Newton

Fie dată ecuația algebrică

$$P_n(x) \equiv a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 = 0,$$

cu coeficienți reali.

Vom folosi metoda lui Newton pentru a determina o rădăcină reală a ecuației date. Șirul de iterare Newton (vezi paragraful 2.6) devine:

$$x_{k+1} = x_k - \frac{P_n(x_k)}{P'_n(x_k)}, \quad k=1, 2, 3, \dots$$

Valorile lui $P_n(x)$ și $P'_n(x)$ în punctele fixate x_k , $k=0, 1, 2, \dots$, se calculează cu ajutorul schemei lui Horner (vezi 2.8.2):

$$b_n = a_n,$$

$$b_j = a_j + x_k b_{j+1}, \quad j=n-1, \dots, 0,$$

$$P_n(x_k) = b_0,$$

$$c_n = b_n,$$

$$c_j = b_j + x_k c_{j+1}, \quad j=n-1, \dots, 1,$$

$$P'_n(x_k) = c_1.$$

Prin urmare, expresia șirului de iterare va fi

$$x_{k+1} = x_k - \frac{b_0}{c_1}, \quad k=0, 1, 2, \dots \quad (2.16)$$

Cu ajutorul acestei metode pot fi calculate toate rădăcinile reale ale ecuației algebrice date. Într-adevăr, fie r_1 rădăcina (simplă) obținută prin metoda lui Newton (2.16). Atunci

$$P_n(x) = (x - r_1)P_{n-1}(x).$$

Deci pentru a găsi o altă rădăcină reală, avem ecuația algebrică de gradul $n-1$:

$$P_{n-1}(x) = b_n x^{n-1} + b_{n-1} x^{n-2} + \dots + b_2 x + b_1 = 0.$$

Ultima ecuație o rezolvăm reluând metoda lui Newton și schema lui Horner după cum s-a arătat anterior. Astfel se calculează toate rădăcinile reale.

Această metodă pentru determinarea rădăcinilor reale ale ecuațiilor algebrice se mai numește metoda iterativă Birge-Viète (vezi [16]).

Exemplu [30]. Să se calculeze rădăcina reală a ecuației:

$$P_3(x) = x^3 - x - 1 = 0$$

utilizând $x_0 = 1.3$.

Folosind schema concisă a lui Horner, se obține:

+	1	0	-1	-1
	0	1.3	1.69	0.897
=	1	1.3	0.69	-0.103 = $b_0 = P_3(1.3)$
+	0	1.3	3.38	
=	1	2.6	4.07 = $c_1 = P'_3(1.3)$	

$$x_1 = x_0 - \frac{b_0}{c_1} = 1.3 - \frac{-0.103}{4.07} = 1.325$$

+	1	0	-1	-1
	0	1.325	1.755625	1.001203
=	1	1.325	0.755625	0.001203 = $b_0 = P_3(1.325)$
+	0	1.325	3.511375	
=	1	2.65	4.267 = $c_1 = P'_3(1.325)$	

$$x_2 = x_1 - \frac{b_0}{c_1} = 1.325 - \frac{0.001203}{4.267} = 1.3247181$$

+	1	0	-1	-1
	0	1.32471	1.154878	1.0000004
		8		
=	1	1.32471	0.154878	0.0000004 = $b_0 = P_3(1.324718)$
		8		
+	0	1.32471	4.109756	
		8		
=	1	2.64943	4.264634 = $c_1 = P'_3(1.324718)$	
		6		

$$x_3 = x_2 - \frac{b_0}{c_1} = 1.324718 - \frac{0.0000004}{4.264634} = 1.3247179$$

Deci, una din rădăcinile reale ale polinomului $P_3(x)$ este $r_1=1.324718$ cu șapte cifre semnificative corecte.

2.9 Alte metode numerice

Considerăm o ecuație algebrică sau transcendentă

$$f(x)=0.$$

Presupunem că am stabilit printr-un mijloc oarecare că ea are o singură rădăcină reală r în intervalul $[a, b]$ și $f(a)f(b)<0$.

Metoda coardei este o variantă a metodei secantei, în care

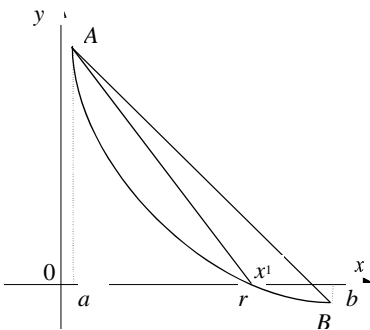


Fig. 2.25

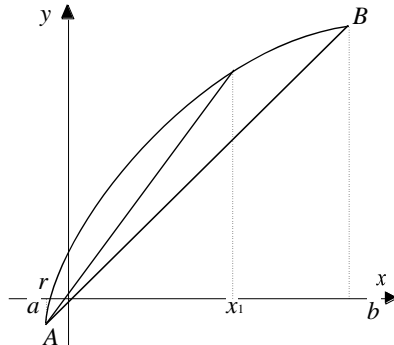


Fig. 2.26

se alege coarda care trece prin $(a, f(a))$ și $(x_k, f(x_k))$ sau $(b, f(b))$ și $(x_k, f(x_k))$, unde se alege capătul a sau b , dacă $f(a)f(x_k)<0$ sau $f(b)f(x_k)<0$. Figurile 2.25-2.28 arată toate cazurile posibile.

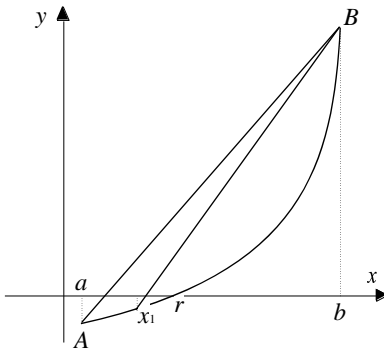


Fig. 2.27

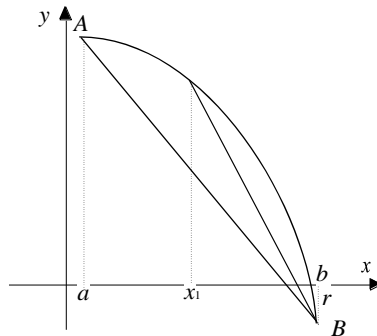


Fig. 2.28

Ecuatia dreptei care trece prin punctele A și B este:

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

Pentru a calcula abscisa punctului de intersecție a coardei AB cu axa Ox punem $y=0$ și deci

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

sau

$$x_1 = \frac{af(b) - bf(a)}{f(b) - f(a)}.$$

Punctul x_1 împarte (a, b) în două intervale (a, x_1) și (x_1, b) .

Din (a, x_1) și (x_1, b) este ales acel interval la extremitățile căruia funcția $f(x)$ are semne contrare și procedura se repetă în mod analog. Concis, formulele de iterare a metodei coardei pentru generarea șirului de aproximații ale lui r , plecând de la aproximația inițială x_0 , se scriu astfel:

$$x_{k+1} = x_k - f(x_k) \cdot \frac{b - x_k}{f(b) - f(x_k)}, \quad x_0 = a, \quad k = 0, 1, 2, \dots \quad (2.17)$$

sau

$$x_{k+1} = x_k - f(x_k) \cdot \frac{x_k - a}{f(x_k) - f(a)}, \quad x_0 = b, \quad k = 0, 1, 2, \dots \quad (2.18)$$

În dependență de forma funcției pe intervalul $[a, b]$ se va alege una sau alta din formulele de iterare a metodei coardei. Fie intervalul $[a, b]$, astfel încât derivata a doua $f''(x)$ să păstreze același semn când x variază de la a la b , adică curba este tot timpul convexă sau tot timpul concavă. Dacă

$$f'(x)f''(x) < 0, \quad \forall x \in [a, b],$$

(vezi fig.2.25 și fig.2.26), atunci se aplică formula (2.18). În caz contrar (vezi fig.2.27 și fig.2.28) se va aplica formula (2.17).

Această metodă se mai numește și *regula falsei poziții*.

Avantajul metodei coardei este că ea produce un șir care, pentru funcțiile continue, este întotdeauna convergent (spre deosebire de metoda secantei sau metoda lui Newton). După cum se vede din fig.2.25 - fig.2.28 convergența șirului construit prin metoda coardei către rădăcina ecuației nu este rapidă. Deci metoda coardei este o metoda bună de start, dar nu trebuie utilizată în vecinătatea rădăcinii.

În metoda lui Newton gradul de convergență este pătratic, adică eroarea la fiecare iterație este proporțională cu pătratul erorii de la iterația anterioară. Pentru metoda secantei gradul de convergență este aproximativ egal cu 1.618. Metoda coardei în general, este de ordinul întâi. Se demonstrează, în ipoteze foarte slabe, că nici o metodă iterativă care folosește doar o evaluare a funcției la fiecare pas nu poate avea ordinul doi de convergență.

O metodă de ordinul doi care la fiecare pas folosește două evaluări pentru funcție, dar nici una pentru derivate, este *metoda lui Steffensen*:

$$x_{k+1} = x_k - f(x_k) \cdot \frac{f(x_k)}{f(x_k + f(x_k)) - f(x_k)}, \quad k=0, 1, 2, \dots$$

Există și alte metode de ordin superior (vezi [36]), adică metode în care șirul de iterare să tindă mai repede către rădăcina r decât șirul Newton sau Steffensen.

Metode de ordinul trei de convergență:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k) - \frac{f(x_k)f''(x_k)}{2f'(x_k)}};$$

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \left[1 + \frac{f(x_k)f''(x_k)}{2f'^2(x_k)} \right].$$

Metode de ordinul patru de convergență:

$$x_{k+1} = x_k - \frac{f(x_k)f'^2(x_k) - \frac{1}{2}f^2(x_k)f''(x_k)}{f'^3(x_k) - f(x_k)f'(x_k)f''(x_k) + \frac{1}{6}f^2(x_k)f''(x_k)};$$

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f^2(x_k)f''(x_k)}{2f'^3(x_k)} + f^3(x_k) \left(\frac{f''(x_k)}{6f'^4(x_k)} - \frac{f''^2(x_k)}{2f'^5(x_k)} \right).$$

2.10 Exerciții

1. Să se separe rădăcinile reale ale ecuațiilor:

$$5^x - 6x - 3 = 0;$$

$$x - \cos x = 0;$$

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

2. Presupunem că există un interval $[a, b]$, astfel încât $f(a)f(b) < 0$ și în intervalul $[a, b]$ funcția este continuă, iar ecuația $f(x) = 0$ admite soluție unică. Care este numărul

maximal de înjumătățiri al intervalului dat pentru a obține soluția cu eroarea dată $\varepsilon > 0$?

3. Considerăm ecuația $f(x)=0$, unde

$$f(x) = \begin{cases} (x+1)^2 - 1, & \text{daca } -1 \leq x \leq 0, \\ -(x-1)^2 + 1, & \text{daca } 0 \leq x \leq 1, \end{cases}$$

Indicați zona de convergență a metodei lui Newton. Ce se întâmplă dacă se alege $x_0 = -2/3$?

4. Să se calculeze $\sqrt[3]{3}$ cu patru zecimale, aplicând metoda lui Newton ecuației $x^3 - 3 = 0$. Să se scrie formula generală pentru calculul $\sqrt[m]{a}$.
5. Să se calculeze o rădăcină reală a ecuațiilor:

$$x^3 - 2x^2 - 4x + 7 = 0, \quad x \in (-1, 0),$$

$$\lg(2x + 3) + 2x - 1 = 0, \quad x \in (0, 0.05),$$

folosind metoda înjumătățirii intervalului, metoda aproximațiilor succesive și metoda lui Newton. Să se compare rezultatele.

6. Să se determine condiția necesară și suficientă pe care trebuie să o satisfacă numerele p și q pentru ca ecuația $x^3 + px + q = 0$ să aibă o rădăcină dublă.
7. Să se arate că șirul Newton pentru determinarea rădăcinii ecuației $1/x - a = 0$ este convergent, dacă se alege x_0 , astfel încât $|1 - ax_0| < 1$. Să se calculeze $1/12$, alegând $x_0 = 0.1$. Ce se întâmplă dacă se alege $x_0 = 1$?
8. Să se rezolve ecuația algebrică $x^3 - 2x^2 - 4x + 7 = 0$, utilizând metoda Birge-Viïte.

METODE NUMERERICE ÎN ALGEBRA LINIARĂ

3.1 Elemente de analiză matriceală

3.1.1. Vectori și matrice

Un tablou dreptunghiular de $m \times n$ numere reale așezate pe m linii și n coloane:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix}$$

se numește *matrice*. numerele a_{ij} se numesc *elementele matricei*.

Matricea se mai poate reprezenta simbolic astfel: $A = (a_{ij})$, $i = 1, 2, \dots, m$; $j = 1, 2, \dots, n$, sau $A = (a_{ij})_{mn}$. Vom spune că matricea A este de dimensiune $m \times n$. În cazul când $m = n$, matricea se numește *pătrată de ordinul n* și se notează $A = (a_{ij})_n$. Dacă $m \neq n$, matricea se numește *rectangulară (dreptunghiulară)*. O matrice $1 \times n$ se numește *vector linie*, iar o matrice $n \times 1$ este un *vector coloană*.

Un sistem ordonat de n numere reale se numește *vector n -dimensional*. Un vector se reprezintă printr-o matrice cu o singură linie sau o singură coloană. În lucrarea de față prin vector vom înțelege întotdeauna vector-coloană.

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Vom nota prin A^T matricea *transpusă* (matricea obținută din cea dată, transformând liniile în coloane și coloanele în linii):

$$A^T = \begin{pmatrix} a_{11} & a_{21} & a_{31} & \dots & a_{m1} \\ a_{12} & a_{22} & a_{32} & \dots & a_{m2} \\ a_{13} & a_{23} & a_{33} & \dots & a_{m3} \\ \dots & \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & a_{3n} & \dots & a_{mn} \end{pmatrix}$$

În particular, transpusa unui vector coloană x este un vector linie:

$$x^T = (x_1, x_2, \dots, x_n).$$

Matricea A se numește *matrice simetrică* dacă $A=A^T$, adică $a_{ij} = a_{ji}$.

Mulțimea tuturor vectorilor n -dimensionali (n natural, fixat) se numește *spațiul liniar n -dimensional* și se notează cu R^n .

Suma matricelor A și B , ambele de dimensiuni $m \times n$, este o matrice C de dimensiune $m \times n$ cu elementele $c_{ij} = a_{ij} + b_{ij}$.

Produsul a două matrice se definește numai în cazul când numărul coloanelor primului factor este egal cu numărul liniilor celui de-al doilea factor. Astfel, dacă $A=(a_{ij})_{mn}$ și $B=(b_{ij})_{np}$, atunci $C=AB$, unde $C=(c_{ij})_{mp}$ și

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj}, i=1,2,\dots,m; j=1,2,\dots,p.$$

Considerăm doi vectori $x, y \in R^n$. Ca un caz particular, vom obține:

$$x^T y = (x_1, x_2, \dots, x_n) \cdot \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = x_1 y_1 + x_2 y_2 + \dots + x_n y_n.$$

$$xy^T = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \cdot (y_1, y_2, \dots, y_n) = \begin{pmatrix} x_1 y_1 & x_2 y_2 & \dots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_n \\ x_3 y_1 & x_3 y_2 & \dots & x_3 y_n \\ \dots & \dots & \dots & \dots \\ x_n y_1 & x_n y_2 & \dots & x_n y_n \end{pmatrix}.$$

$x^T y$ se numește *produsul scalar* al vectorilor $x, y \in \mathbb{R}^n$ și se mai notează (x, y) sau $\langle x, y \rangle$.

xy^T se numește *produsul diadic* al vectorilor $x, y \in \mathbb{R}^n$; este o matrice pătrată de ordinul n și se mai notează astfel: $\triangleright x, y \triangleleft$.

Pentru orice vectori din \mathbb{R}^n au loc proprietățile:

1. $(x, y) = (y, x)$;
2. $(x + y, z) = (x, z) + (y, z)$;
3. $(ax, y) = a(x, y)$; ($a \in \mathbb{R}$);
4. $(x, x) \geq 0$; $(x, x) = 0$ atunci și numai atunci când $x = 0$.

Produsul scalar este comutativ. În caz general pentru matrice $A \cdot B \neq B \cdot A$.

Orice matrice pătrată se poate înmulți cu ea însăși:

$$A \cdot A = A^2; A^2 \cdot A = A^3; \dots; A^{n-1} \cdot A = A^n; \dots$$

Au loc inegalitățile:

1. $(A \cdot B)C = A(B \cdot C)$, legea asociativității,
2. $A(B + C) = AB + AC$, legea distributivității la stânga,
3. $(B + C)A = BA + CA$, legea distributivității la dreapta,
4. $\alpha(AB) = (\alpha A)B = A(\alpha B)$, $\alpha \in \mathbb{R}$.

3.1.2. Norme de vectori și matrice

Norma unui vector $x \in R^n$ este un număr real, notat $\|x\|$, cu proprietățile:

1. $\|x\| \geq 0$ pentru orice $x \in R^n$
2. $\|x\| = 0$ dacă și numai dacă $x = 0$
3. $\|\alpha x\| = |\alpha| \|x\|$ pentru orice $x \in R^n$ și $\alpha \in R$.
4. $\|x + y\| \leq \|x\| + \|y\|$ pentru orice x și $y \in R^n$.

Pentru orice vector $x \in R^n$ se definesc normele:

$$\|x\|_1 = \sum_{i=1}^n |x_i|$$

$$\|x\|_2 = \left(\sum_{i=1}^n x_i^2 \right)^{1/2}$$

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

Ele satisfac proprietățile din definiția de mai sus a normei. Norma $\|x\|_2$ se numește și *normă euclidiană*. Ea provine din produsul scalar:

$$\|x\|_2 = \sqrt{(x, x)}$$

și generalizează noțiunea de lungime a vectorului.

Au loc următoarele inegalități:

$$\|x\|_\infty \leq \|x\|_1 \leq n \|x\|_\infty,$$

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \|x\|_\infty,$$

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1.$$

Exemplu. Fie $x = (1, -2, -3)^T$. Atunci

$$\|x\|_1 = 6, \|x\|_2 = \sqrt{14} \text{ și } \|x\|_\infty = 3.$$

Oricare ar fi $x, y \in R^n$ avem:

$$|(x, y)| \leq \|x\|_2 \|y\|_2.$$

Această relație se numește *inegalitatea lui Schwarz Cauchy Buniacovski*.

Unghiul dintre doi vectori x, y din R^n se definește prin formula:

$$\cos \theta = \frac{(x, y)}{\|x\|_2 \|y\|_2}.$$

Doi vectori x, y din R^n se zic *ortogonali* dacă $(x, y) = 0$.

Pe mulțimea matricelor pătrate se poate introduce o normă $\|A\|$ în sensul definit mai sus pentru vectori. Mai importante și mai utilizate sunt însă normele matriceale definite astfel:

$$\|A\| = \max_{|x|=1} \|Ax\|.$$

Această normă satisface următoarea condiție:

$$\|A \cdot B\| \leq \|A\| \cdot \|B\|$$

oricare ar fi matricele A și B .

Dacă în plus, oricare ar fi vectorul $x \in R^n$ avem:

$$\|Ax\| \leq \|A\| \|x\|$$

se zice că norma matriceală este *compatibilă* cu norma vectorială sau *subordonată* normei vectoriale.

Normele:

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|;$$

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|;$$

sunt norme matriceale compatibile cu normele vectoriale $\|x\|_\infty$ și $\|x\|_1$.

Numărul λ se numește *valoare proprie* a lui A dacă există un vector nenul $x \in R^n$, astfel încât $Ax = \lambda x$.

Mulțimea valorilor proprii ale matricei A formează *spectrul* lui A și se notează cu $\sigma(A)$.

Raza spectrală a lui A se definește prin relația:

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|.$$

Norma matriceală subordonată normei euclidiene $\|x\|_2$ este:

$$\|A\|_2 = \sqrt{\rho(A^T A)}.$$

În aplicații se utilizează des următoarea normă:

$$\|A\|_F = \left(\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2 \right)^{1/2}.$$

numită *norma lui Frobenius*, care însă nu este subordonată unei norme vectoriale.

Prin I vom nota matricea unitate:

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Oricare ar fi matricea A avem $IA=AI=A$. Matricea A se numește inversabilă dacă există o matrice, notată cu A^{-1} , astfel încât $A^{-1}A = AA^{-1} = I$.

Dacă $\|A\| < 1$ atunci matricea $I-A$ este inversabilă și

$$(I-A)^{-1} = I + A + A^2 + A^3 + \dots,$$

$$\|(I-A)^{-1}\| \leq \frac{1}{1-\|A\|}.$$

Pentru ca matricea $I-A$ să fie inversabilă este suficient ca:

$$\lim_{n \rightarrow \infty} A^n = 0.$$

unde 0 este matricea nulă (o matrice cu elementele egale cu zero).

3.1.3. Matrice speciale

O matrice pătrată de forma:

$$D = \begin{pmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ 0 & 0 & d_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & d_n \end{pmatrix}$$

se numește *matrice diagonală*. Astfel de matrice se mai notează:

$$D = \text{Diag}(d_1, d_2, \dots, d_n).$$

Matricea A se numește *inferior triunghiulară* (*superior triunghiulară*) dacă elementele sale satisfac relațiile:

$$a_{ij} = 0 \text{ pentru } i < j \text{ (} i > j \text{)}, i, j = 1, 2, \dots, n.$$

Matricele inferior triunghiulare se notează de obicei prin L , iar cele superior triunghiulare prin U ; de exemplu:

$$L = \begin{pmatrix} -1 & 0 & 0 \\ 2 & 3 & 0 \\ 7 & -4 & 5 \end{pmatrix} \quad U = \begin{pmatrix} 4 & 2 & 1 \\ 0 & 3 & -7 \\ 0 & 0 & 6 \end{pmatrix}$$

Matricea A se numește *tridiagonală* dacă elementele sale satisfac relațiile:

$$a_{ij} = 0 \text{ pentru } |i - j| > 1, \quad i, j = 1, 2, \dots, n.$$

Astfel:

$$A = \begin{pmatrix} a_{11} & a_{12} & 0 & 0 & \text{K} & 0 \\ a_{21} & a_{22} & a_{23} & 0 & \text{K} & 0 \\ 0 & a_{32} & a_{33} & a_{34} & \text{K} & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & \text{K} & a_{n-1,n} \\ 0 & 0 & \text{K} & a_{n,n-1} & a_{nn} & \end{pmatrix}$$

O matrice Q se numește *ortogonală* dacă:

$$Q^T Q = Q Q^T = I \text{ sau } Q^T = Q^{-1}.$$

Imediat se verifică ca

$$(Qx)^T Qy = x^T y, \quad \text{pentru } \forall x \in R^n,$$

$$\|Qx\|_2 = \|x\|_2 \quad \text{pentru } \forall x \in R^n,$$

$$\|Q\|_2 = 1, \quad \|QA\|_2 = \|AQ\|_2 = \|A\|_2 \quad \text{pentru } \forall A.$$

Prin urmare, matricele ortogonale păstrează produsul scalar, lungimea vectorilor și norma matricelor.

Exemplu.

$$Q = \begin{pmatrix} \cos \theta & -\sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}, \quad Q^T = Q^{-1} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}$$

Matricea Q rotește orice vector cu unghiul θ , iar matricea Q^T îl rotește în direcție inversă cu unghiul $-\theta$.

Matricea obținută din matricea unitate prin reordonarea coloanelor ei se numește matrice de *permutare* și se notează P . De exemplu:

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

O matrice de forma uv^T , unde $u, v \in R^n$, se numește matrice de *rangul întâi*. Matricele $I + \alpha uv^T$, unde α este un scalar, se numesc matrice *elementare*. Matricea

$$H = I - \frac{2uu^T}{\|u\|_2^2}, \quad u \neq 0, u \in R^n.$$

se numește *reflector* sau matricea lui *Householder*.

Se verifică ușor că matricele P și H sunt ortogonale.

Identitatea Sherman-Morrison-Woodbury. Fie $x, y \in R^n$ și presupunem că matricea A este inversabilă. Matricea $A + xy^T$ va fi inversabilă atunci și numai atunci când $1 + y^T A^{-1} x \neq 0$. În plus:

$$(A + xy^T)^{-1} = A^{-1} - \frac{A^{-1}xy^T A^{-1}}{1 + y^T A^{-1}x}$$

O matrice A simetrică se numește *pozitiv definită* dacă

1. Sistemul de ecuații are o soluție unică. În acest caz se spune că sistemul (3.1) este *compatibil determinat*. De exemplu, sistemul:

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

este compatibil determinat cu soluția $x_1^* = x_2^* = 1$. În fig. 3.1 este dată interpretarea geometrică a sistemului considerat, din care se vede că dreptele $x_2 = 2 - x_1$ și $x_2 = -1 + 2x_1$ se intersectează numai într-un singur punct.

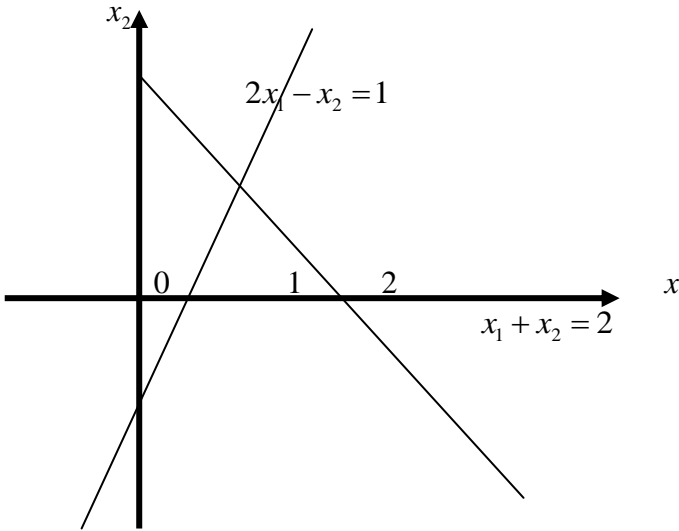


Fig. 3.1 Interpretarea geometrică a sistemului compatibil determinat

2. Sistemul de ecuații are o infinitate de soluții. Despre astfel de sisteme se spune că sunt *compatibil nedeterminate*. În fig.3.2 avem interpretarea geometrică a sistemului:

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

care are o infinitate de soluții; aceste două soluții descriu una și aceeași dreaptă $x_2 = 1 - x_1$.

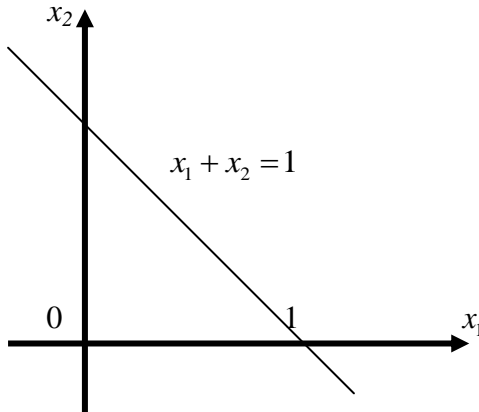


Fig. 3.2 Interpretarea geometrică a sistemului compatibil nedeterminat

3. Sistemul de ecuații nu are soluții, adică este *incompatibil*. De exemplu, sistemul:

$$\begin{cases} x_1 + 2x_2 = 2 \\ 2x_1 + 4x_2 = 7 \end{cases}$$

nu este compatibil. Dreptele $x_2 = 1 - \frac{1}{2}x_1$ și $x_2 = \frac{7}{4} - \frac{1}{2}x_1$ (vezi fig. 3.3) sunt paralele.

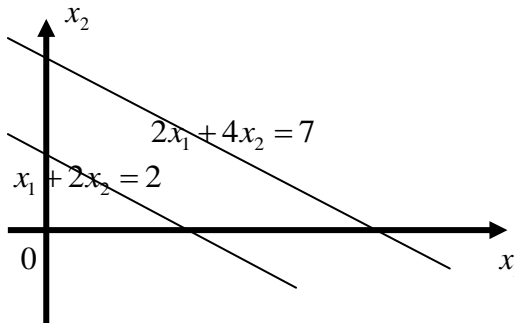


Fig.3.3 Interpretarea geometrică a sistemului incompatibil

Dacă matricea A este nesingulară ($\det A \neq 0$), atunci oricare ar fi vectorul $b \in R^n$ sistemul $Ax=b$ este compatibil determinat. Soluția sistemului poate fi scrisă sub forma:

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

unde A^{-1} este inversa lui A .

Inversarea matricelor este o operație costisitoare (vezi de ex., [2]) care trebuie evitată în practică. În calitate de exemplu ilustrativ considerăm “sistemul” dintr-o ecuație cu o singură necunoscută:

$$7x = 21.$$

Mijlocul cel mai bun de rezolvare a acestei probleme este împărțirea:

$$x^* = \frac{21}{7} = 3.$$

Aplicarea matricei inverse ne-ar duce la:

$$x^* = 7^{-1} \times 21 = 0.142857 \times 21 = 2.99997.$$

Al doilea procedeu necesită cu o operație aritmetică mai mult și dă un rezultat mai puțin precis. Același lucru, dar într-un mod mai pronunțat, este adevărat și în cazul rezolvării sistemelor cu multe ecuații. De aceea relația (3.2) trebuie interpretată doar în sensul de exprimare a faptului că x^* este soluția unică a sistemului $Ax = b$, dar nu și ca o cale de obținere a acestei soluții.

După cum se știe din matematica elementară, sistemele de ecuații liniare pot fi rezolvate prin formulele lui Cramer:

$$x_i^* = \frac{\Delta_i}{\Delta}, \quad \Delta = \det(A), \quad \Delta_i = \sum_{j=1}^n A_{ij} b_j, \quad i = 1, 2, \dots, n,$$

A_{ij} fiind complementul algebric al lui a_{ij} .

Metoda de rezolvare a sistemelor prin formulele lui Cramer din punct de vedere practic rămâne inutilizabilă, deoarece cere un număr mare de operații aritmetice, și anume este necesar să se calculeze $n+1$ determinanți ($\Delta, \Delta_1, \dots, \Delta_n$) și să se efectueze n împărțiri. Pentru calculul unui determinant sunt necesare $(n-1) \times n$

înmulțiri și $(n-1)$ adunări. De exemplu, rezolvarea unui sistem cu 20 de ecuații prin formulele lui Cramer presupune efectuarea a $19 \times 20 \times 21$ înmulțiri. S-a apreciat că dacă am executa aceste înmulțiri la un calculator electronic de viteză 10^5 operații pe secundă ne-ar trebui aproximativ 3×10^6 ani!

Metodele numerice de rezolvare a sistemelor de ecuații liniare sunt de două tipuri: metode directe și metode iterative.

Metodele directe constau în transformarea sistemului $Ax=b$ într-un sistem echivalent pentru care rezolvarea este cu mult mai simplă. În metodele directe soluția exactă se obține după un număr finit de operații aritmetice elementare (adunare, scădere, înmulțire, împărțire și rădăcină pătrată) și acest număr de operații este de ordinul n^3 . Subliniem că soluția exactă se obține în cazurile (ideale) în care erorile de rotunjire sunt absente. La fiecare operație elementară efectuată de calculator avem o eroare de rotunjire și prin urmare metodele directe în caz general furnizează doar o soluție aproximativă. Metodele directe se utilizează pentru rezolvarea sistemelor nu prea “mari”, de dimensiune $n \leq 200$.

Rezolvarea sistemelor de ecuații liniare printr-o *metodă iterativă* înseamnă construirea unui șir de vectori $x^{(k)}$, $k=0,1,\dots$ (pornind de la un vector $x^{(0)}$ ales arbitrar) convergent către soluția sistemului considerat. În metodele iterative, de obicei, o iterație necesită efectuarea unui număr de ordinul n^2 operații aritmetice. De aceea metodele iterative se utilizează pentru rezolvarea sistemelor “mari”, de dimensiune $n \geq 10^2$ (în cazul asigurării unei viteze sporite de convergență pentru o alegere a aproximării inițiale adecvate). Trunchierea șirului $\{x^{(k)}\}$ are loc la un indice m astfel încât $x^{(m)}$ constituie o aproximație satisfăcătoare a soluției căutate x^* (de exemplu, $\|x^{(m)} - x^*\| < \varepsilon$, unde $\varepsilon > 0$ este eroarea admisă).

3.3 Metoda eliminării a lui Gauss

Metoda eliminării a lui Gauss constă în a aduce sistemul inițial la un sistem echivalent având matricea coeficienților superior triunghiulară. Transformarea sistemului dat într-un sistem de formă triunghiulară fără ca să se modifice soluția sistemului se realizează cu ajutorul următoarelor trei operații de bază:

1. rearanjarea ecuațiilor (schimbarea a două ecuații între ele);
2. înmulțirea unei ecuații cu o constantă (diferită de zero);
3. scăderea unei ecuații din alta și înlocuirea celei de a doua cu rezultatul scăderii.

Exemplificăm această metodă pentru următorul sistem de ecuații liniare:

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

Putem elimina necunoscuta x_1 din ultimele două ecuații, înmulțind prima ecuație respectiv cu factorii:

$$\mu_{21} = \frac{a_{21}}{a_{11}} = \frac{4}{2} = 2, \quad \mu_{31} = \frac{a_{31}}{a_{11}} = \frac{-2}{2} = -1$$

și scăzând-o din ecuația a doua și apoi din ecuația a treia.

Obținem astfel sistemul echivalent

$$\begin{cases} 2x_1 + x_2 + x_3 = 1, \\ -x_2 - 2x_3 = -4, \\ 3x_2 + 2x_3 = 8. \end{cases}$$

Coeficientul $a_{11} = 2$ din prima ecuație se numește *elementul pivot* al primului pas de eliminare, iar linia corespunzătoare se numește *linie pivot*.

În mod analog putem elimina necunoscuta x_2 din ultima ecuație. La pasul al doilea elementul pivot este $a'_{22} = -1$. ecuația a doua o înmulțim cu $\mu_{32} = \frac{a'_{32}}{a'_{22}} = \frac{3}{-1} = -3$ și o scădem din ecuația a treia. Deci se obține sistemul de formă triunghiulară:

$$\begin{cases} 2x_1 + x_2 + x_3 = 1, \\ -x_2 - 2x_3 = -4, \\ -4x_3 = -4. \end{cases}$$

În continuare se determină necunoscutele începând cu ecuația a treia: $x_3^* = 1$; înlocuind rezultatul obținut în ecuația a doua vom obține pe $x_2^* = 2$; în sfârșit din prima ecuație avem $x_1^* = -1$.

Să generalizăm această metodă. Fie dat sistemul de ecuații liniare:

$$Ax = b. \quad (3.3)$$

unde $A = (a_{ij})_n$, $x, b \in R^n$, $\det A \neq 0$.

Să presupunem că $a_{11} \neq 0$; dacă $a_{11} = 0$ se aduce elementul nenul din prima coloană pe locul (1,1), permutând ecuațiile respective ale sistemului. Primul pas constă în eliminarea necunoscutei x_1 din ecuațiile sistemului începând cu a doua, multiplicând ecuația întâia cu raportul:

$$\mu_{i1} = \frac{a_{i1}}{a_{11}}, i = 2, 3, \dots, n$$

și scăzând rezultatul obținut din ecuația i pentru $\forall i \geq 2$.

Obținem în acest caz sistemul echivalent:

$$A^{(2)}x = b^{(2)}. \quad (3.4)$$

cu coeficienții:

$$\begin{aligned}
a_{1j}^{(2)} &= a_{1j}^{(1)}, \quad j = 1, 2, \dots, n; \\
a_{i1}^{(2)} &= 0, \quad i = 2, 3, \dots, n; \\
a_{ij}^{(2)} &= a_{ij}^{(1)} - \mu_{i1} a_{1j}^{(1)}, \quad i, j = 2, 3, \dots, n; \\
b_1^{(2)} &= b_1^{(1)}, \quad b_i^{(2)} = b_i^{(1)} - \mu_{i1} b_1^{(1)}, \quad j = 2, 3, \dots, n.
\end{aligned}$$

Mai sus s-a notat $a_{ij}^{(1)} = a_{ij}$; $i, j = 1, 2, \dots, n$ și $b_i^{(1)} = b_i$; $i = 1, 2, \dots, n$.

Prima ecuație a sistemului (3.4) coincide cu prima ecuație a sistemului (3.3). În continuare se repetă procedeul de mai sus pentru eliminarea necunoscutei x_2 din sistemul (34) ș.a.m.d. La pasul k se obține sistemul:

$$A^{(k)}x = b^{(k)}$$

unde

$$A^{(k)} = \begin{pmatrix}
a_{11}^{(1)} & a_{12}^{(1)} & \text{K} & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & \text{K} & a_{1n}^{(1)} \\
0 & a_{22}^{(2)} & \text{K} & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & \text{K} & a_{2n}^{(2)} \\
\text{K} & \text{K} & \text{K} & \text{K} & \text{K} & \text{K} & \text{K} \\
0 & 0 & \text{K} & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \text{K} & a_{k-1,n}^{(k-1)} \\
0 & 0 & \text{K} & 0 & a_{kk}^{(k)} & \text{K} & a_{kn}^{(k)} \\
\text{K} & \text{K} & \text{K} & \text{K} & \text{K} & \text{K} & \text{K} \\
0 & 0 & \text{K} & 0 & a_{nk}^{(k)} & \text{K} & a_{nn}^{(n)}
\end{pmatrix}; b^{(k)} = \begin{pmatrix}
b_1^{(1)} \\
b_2^{(2)} \\
\text{M} \\
b_{k-1}^{(k-1)} \\
b_k^{(k)} \\
\text{M} \\
b_n^{(k)}
\end{pmatrix}$$

Elementele $a_{ij}^{(k)}$ ale lui $A^{(k)}$ și $b_i^{(k)}$ ale lui $b^{(k)}$ se calculează recursiv prin formulele:

$$a_{ij}^{(k)} = \begin{cases}
a_{ij}^{(k-1)} & , \text{pentru } i \leq k-1, \\
0 & , \text{pentru } i \geq k, j \leq k-1, \\
a_{ij}^{(k-1)} - \mu_{i,k-1} \cdot a_{k-1,j}^{(k-1)} & \text{pentru } i \geq k, j \geq k,
\end{cases}$$

unde

$$\mu_{i,k-1} = \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}},$$

iar

$$b_i^{(k)} = \begin{cases} b_i^{(k-1)} & , \text{pentru } i \leq k-1, \\ b_i^{(k-1)} - \mu_{i,k-1} \cdot b_{k-1}^{(k-1)} & , \text{pentru } i \geq k. \end{cases}$$

După n pași necunoscuta x_{n-1} va fi eliminată din ultima ecuație, obținându-se un sistem cum matricea superior triunghiulară:

$$\left\{ \begin{array}{l} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \text{K} + a_{1k}^{(1)}x_k + \text{K} + a_{1n}^{(1)}x_n = b_1^{(1)}, \\ a_{22}^{(2)}x_2 + \text{K} + a_{2k}^{(2)}x_k + \text{K} + a_{2n}^{(2)}x_n = b_2^{(2)}, \\ \text{K K K K K K K K K K K K K K K K K K K K} \\ a_{kk}^{(k)}x_k + \text{K} + a_{kn}^{(k)}x_n = b_k^{(k)}, \\ \text{K K K K K K K K K K K K K K K K K K K K} \\ a_{nn}^{(n)}x_n = b_n^{(n)}. \end{array} \right.$$

Acest sistem se rezolvă începând cu ultima ecuație cu ajutorul *procesului de eliminare inversă* care se poate descrie astfel:

$$x_n = \frac{b_n^{(n)}}{a_{nn}^{(n)}}, \quad x_{n-1} = \frac{b_{n-1}^{(n-1)} - a_{n-1,n}^{(n-1)}x_n}{a_{n-1,n-1}^{(n-1)}},$$

$$x_k = \frac{b_k^{(k)} - \sum_{j=k+1}^n a_{kj}^{(k)}x_j}{a_{kk}^{(k)}}, \quad k = n-2, n-3, \text{K}, 2, 1.$$

Metoda eliminării a lui Gauss prezentată mai sus presupune că elementele pivot trebuie să fie diferite de zero. Dacă la

efectuarea pasului k elementul pivot $a_{kk}^{(k)} = 0$, atunci cel puțin unul din celelalte elemente din coloana k și din liniile $k+1, k+2, \dots, n$ este nenul; în caz contrar matricea A ar fi singulară ($\det(A) = 0$). Permutând ecuațiile sistemului putem aduce pe locul (k, k) elementul nenul și, deci, este posibil să reluăm eliminarea.

Considerăm un exemplu simplu:

$$\begin{pmatrix} 0 & 3 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 4 \end{pmatrix}$$

Evident, nici o multiplicare a primei ecuații nu poate fi utilizată pentru a elimina pe x_1 din ecuația a doua. Schimbând ecuațiile între ele cu locul, obținem:

$$\begin{cases} 2x_1 + x_2 = 4. \\ 3x_2 = 0. \end{cases}$$

un sistem sub formă triunghiulară, care se rezolvă imediat prin eliminarea inversă: $x_2^* = 0$, $x_1^* = 2$.

Analizăm un alt exemplu:

$$\begin{cases} 0.000100x_1 + x_2 = 1, \\ x_1 + x_2 = 2 \end{cases}$$

cu soluția exactă $x_1^* = 1.00010$, $x_2^* = 0.99990$.. Vom utiliza o aritmetică a virgulei mobile cu $\beta = 10$ și $t = 3$: se păstrează în calcule numai trei cifre zecimale semnificative și presupunem că rezultatul se rotunjește corect. Aplicând metoda eliminării a lui Gauss obținem sistemul:

$$\begin{cases} 0.000100x_1 + x_2 = 1, \\ -10000x_2 = -10000. \end{cases}$$

Din ultima ecuație avem $x_2^* = 1.000$ care înlocuită în prima ecuație ne dă $x_1^* = 0.000$, evident un rezultat eronat. ***S-a produs o catastrofă de calcul!*** Permutând ecuațiile între ele, avem sistemul:

$$\begin{cases} x_1 + x_2 = 2. \\ 0.000100x_1 + x_2 = 1. \end{cases}$$

și metoda eliminării lui Gauss îl transformă în:

$$\begin{cases} x_1 + x_2 = 2. \\ x_2 = 1. \end{cases}$$

cu soluția $x_1^* = x_2^* = 1.00$.

Prin urmare, dacă un element pivot este exact egal cu zero sau chiar aproape egal cu zero, din motive de stabilitate numerică, trebuie să efectuăm rearanjarea ecuațiilor.

Există două strategii de alegere al elementului pivot pentru a preveni ca influența erorilor de rotunjire să devină catastrofală. Prima strategie se numește *pivotare parțială* și constă în următoarele: la pasul k pivotul se ia egal cu primul element maxim în modul din coloana k subdiagonală a lui $A^{(k)}$:

$$|a_{rk}^{(k)}| = \max_{k \leq i \leq n} |a_{ik}^{(k)}|$$

și se permută liniile k și r .

O altă strategie de permutare constă în *pivotarea completă (totală)*; se schimbă liniile k și r ($r \geq k$) și coloanele k și s , ($s \geq k$) astfel încât pivotul $a_{kk}^{(k)}$ obținut după permutare să coincidă cu primul element maxim în modul din submatricea delimitată de ultimile $n-k$ linii și coloane ale lui $A^{(k)}$:

$$|a_{rs}^{(k)}| = \max_{k \leq i, j \leq n} |a_{ij}^{(k)}|.$$

Matricea A se numește *diagonal dominantă* dacă

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|, \quad i = 1, 2, \dots, n.$$

Fie A o matrice simetrică și diagonal dominantă. După primul pas de eliminare gaussiană matricea $A^{(2)}$ devine:

$$\begin{pmatrix} a & z \\ 0 & \bar{A}_1 \end{pmatrix}$$

unde submatricea \overline{A}_1 este de asemenea diagonal dominantă. Se poate demonstra că procesul eliminării în cazul matricelor diagonal dominante nu depinde de alegerea elementului pivot. Nu este necesară pivotarea și în cazul când matricea A este pozitiv definită.

Putem estima numărul de operații aritmetice în metoda eliminării lui Gauss. Procedura de eliminare a necunoscutei x_1 cere $n(n-1) = n^2 - n$ operații aritmetice. Eliminarea necunoscutei x_k necesită $k(k-1) = k^2 - k$ operații. Prin urmare procedura directă cere următorul număr de operații aritmetice:

$$\begin{aligned} N_1 &= (n^2 - n) + K + (k^2 - k) + K + (1^2 - 1) = \\ &= \sum_{k=1}^n k^2 - \sum_{k=1}^n k = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)}{2} = \frac{n^2 - n}{3} \end{aligned}$$

Procesul de eliminare inversă se efectuează cu mult mai repede. Necunoscuta x_n se află cu ajutorul unei singure operații (împărțirea la elementul pivot); calculul x_{n-1} cere două operații (împărțire - scădere și apoi împărțire) ș.a.m.d.; pasul k necesită numai k operații. Prin urmare eliminarea inversă necesită

$$N_2 = \sum_{k=1}^n k = \frac{n(n+1)}{2}$$

de operații aritmetice.

Mulți ani s-a crezut că metoda eliminării lui Gauss este optimă în sensul că orice altă metodă directă de rezolvare a sistemelor din n ecuații liniare necesită cel puțin $\frac{n^3}{3}$ operații aritmetice. În momentul de față se cunosc metode în care numărul de operații este redus la Cn^α ($2 < \alpha < 3$). Aceste metode se bazează pe un rezultat remarcabil obținut în 1971 de către A. Schonhage și V. Strassen care au arătat că, teoretic, înmulțirea poate avea o complexitate numai cu puțin superioară adunării. Nu ne vom opri aici asupra acestor metode. Pentru a arăta că metoda

lui Gauss nu este optimă să examinăm algoritmul lui Strassen de multiplicare a două matrice.

Fie de exemplu,

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}.$$

Algoritmul lui Strassen (vezi, de exemplu, [18], pag.47) se bazează pe identitatea matriceală:

$$A \times B = \begin{pmatrix} C & D \\ E & F \end{pmatrix},$$

unde:

$$C = (a_{11} + a_{22})(b_{11} + b_{22}) + a_{22}(-b_{11} + b_{21}) - (a_{11} + a_{12})b_{22} + (a_{12} - a_{22})(b_{21} + b_{22}),$$

$$F = (a_{11} + a_{22})(b_{11} + b_{22}) + a_{11}(b_{12} - b_{22}) - (a_{21} + a_{22})b_{11} + (-a_{11} + a_{21})(b_{11} + b_{12}),$$

$$D = a_{11}(b_{12} - b_{22}) + (a_{11} + a_{12})b_{22},$$

$$E = (a_{21} + a_{22})b_{11} + a_{22}(-b_{11} + b_{21}).$$

Astfel, pentru a obține produsul a două matrice este suficient de a efectua șapte înmulțiri și 18 adunări. Dacă am multiplica matricele în mod tradițional, ne-ar trebui opt înmulțiri. În exemplul de mai sus nu s-au concretizat elementele matricelor A și B și nu s-a utilizat proprietatea de comutativitate a produsului. Prin urmare elementele a_{ij} , b_{ij} pot fi considerate matrice și avem o procedură de multiplicare a matricelor de orice dimensiune n . Fie $n = 2^m$, m – natural; în caz contrar adăugăm atâtea linii și coloane nule în matricele A și B încât n ar deveni o putere a lui doi. Dacă $N(2^m)$ este numărul de operații efectuate la înmulțirea a două matrice de dimensiune 2^m , atunci în baza identității de mai sus avem:

$$N(2^m) = 7 \cdot N(2^{m-1}) + 18 \cdot 2^{2m-2}.$$

Ținând seama că $N(2) = 7 + 18$, ultima relație implică

$$N(2^m) = 7^m + 6(7^m + 4^m)$$

sau

$$N(n) = n^{\log_2 7} + 6(n^{\log_2 7} + n^{\log_2 4}).$$

Deoarece $\alpha = \log_2 7 \approx 2.81 < 3$, algoritmul lui Strassen pentru n suficient de mare este mai avantajos decât procedeul obișnuit de înmulțire a matricelor.

3.4 Factorizarea LU

Să reluăm exemplul numeric din paragraful 3.3 de rezolvare a sistemului $Ax = b$, cu

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 4 & 1 & 0 \\ -2 & 2 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ -2 \\ 7 \end{pmatrix}.$$

În urma procesului de eliminare a necunoscutei x_1 se obține sistemul $A^{(2)}x = b^{(2)}$, unde

$$A^{(2)} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 3 & 2 \end{pmatrix}, \quad b^{(2)} = \begin{pmatrix} 1 \\ -4 \\ 8 \end{pmatrix}.$$

Notăm prin M_1 matricea interior triunghiulară:

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ -\mu_{21} & 1 & 0 \\ -\mu_{31} & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

care se obține din matricea unitate prin înlocuirea elementelor subdiagonale din coloana întâia cu multiplicatorii $-\mu_{21}, -\mu_{31}$. Se verifică ușor că

$$M_1 A = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 \\ 4 & 1 & 0 \\ -2 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 3 & 2 \end{pmatrix} = A^{(2)},$$

$$M_1 b = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 7 \end{pmatrix} = \begin{pmatrix} 1 \\ -4 \\ 8 \end{pmatrix} = b^{(2)}.$$

În etapa finală de transformare necunoscuta x_2 va fi eliminată din ultima ecuație, obținându-se un sistem sub formă triunghiulară $Ux = c$, unde

$$U = \begin{pmatrix} 2 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & -4 \end{pmatrix}, \quad c = \begin{pmatrix} 1 \\ -4 \\ -4 \end{pmatrix}.$$

Se observă că $U = M_2 A^{(2)}$ și $c = M_2 b^{(2)}$, unde

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\mu_{32} & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix}.$$

Prin urmare procesul de transformare a sistemului $Ax = b$, într-un sistem echivalent de formă triunghiulară $Ux = c$ poate fi reprezentat ca înmulțirea sistemului inițial succesiv la matricele M_1, M_2 :

$$M_2 M_1 A x = M_2 M_1 b.$$

Relația $M_2 M_1 A = U$ permite de a da o altă interpretare metodei lui Gauss. Multiplicând această relație la stânga cu matricea inferior triunghiulară $L = M_1^{-1} M_2^{-1}$ obținem:

$$A = LU.$$

Deci, cu ajutorul metodei eliminării a lui Gauss matricea A se descompune în produsul de doi factori L și U , unde L este o matrice inferior triunghiulară, iar U este o matrice superior

triunghiulară. Această descompunere se numește *factorizarea LU* a matricei A .

Vom arăta că pentru orice matrice nesingulară există o “*factorizare LU*” care este echivalentă metodei eliminării lui Gauss. Pentru început presupunem că matricea A este astfel, încât eliminarea să se poată face fără permutări de linii sau de coloane.

Fie

$$m_k = \begin{pmatrix} 0 \\ 0 \\ M \\ 0 \\ \mu_{k+1,k} \\ M \\ \mu_{nk} \end{pmatrix}, \quad e_k = \begin{pmatrix} 0 \\ 0 \\ M \\ 1 \\ 0 \\ M \\ 0 \end{pmatrix} \leftarrow \text{componenta } k$$

unde multiplicatorii $\mu_{ik}, i = k + 1, k + 2, \dots, n$ sunt cei utilizați la pasul $k+1$ pentru eliminarea necunoscutei x_{k+1} (vezi paragraful 3.3).

Definim o matrice M_k astfel

$$M_k = I - m_k e_k^T.$$

Această matrice diferă de matricea unitate I numai prin elementele subdiagonale nenule din coloana k :

$$M_k = \begin{pmatrix} 1 & 0 & K & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ 0 & 0 & \dots & -\mu_{k+1,k} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -\mu_{nk} & \dots & 1 \end{pmatrix}.$$

Metoda eliminării a lui Gauss constă (vezi paragraful 3.3) în determinarea șirului de matrice $A = A^{(1)}, A^{(2)}, K, A^{(n)}$. Se constată ușor că

$$A^{(k+1)} = M_k \cdot M_{k-1} \cdot K \cdot M_1 A, k = 1, 2, K, n-1.$$

Scriind această relație pentru $k = n-1$ și notând $U = A^{(n)}$, obținem:

$$M_{n-1} \cdot M_{n-2} \cdot K \cdot M_2 \cdot M_1 \cdot A = U,$$

sau

$$A = M_1^{-1} \cdot M_2^{-1} \cdot K \cdot M_{n-1}^{-1} \cdot U.$$

Se verifică imediat că

$$M_k^{-1} = I + m_k e_k^T.$$

De mai sus deducem:

$$A = L \cdot U,$$

unde

$$L = M_1^{-1} \cdot M_2^{-1} \cdot K \cdot M_{n-1}^{-1} = I + n-1 \sum_{k=1} m_k e_k^T.$$

Prin urmare, metoda eliminării lui Gauss calculează o factorizare LU a matricei A , unde

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 & \dots & 0 \\ \mu_{21} & 1 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \mu_{k1} & \mu_{k2} & \dots & 1 & \dots & 0 \\ \mu_{k+1,1} & \mu_{k+1,2} & \dots & \mu_{k+1,k} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \mu_{n1} & \mu_{n2} & \dots & \mu_{nk} & \dots & 1 \end{pmatrix},$$

$$U = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \dots & \dots & \dots & \dots \\ M & M & \dots & M \\ 0 & 0 & \dots & a_{nn}^{(n)} \end{pmatrix}.$$

După cum s-a arătat în paragraful 3.3 din motive de stabilitate este indicat să utilizăm o strategie de pivotare parțială. Se demonstrează că orice matrice admite o factorizare LU , eventual efectuând asupra liniilor permutările care apar prin pivotare parțială. Cu alte cuvinte, există o matrice de permutare P astfel încât

$$PA=LU.$$

Metoda eliminării lui Gauss și factorizarea LU sunt echivalente. O factorizare LU a lui A poate fi calculată și direct, printr-o procedură compactă numită *factorizarea lui Crout*. Această factorizare impune U cu diagonala unitate și este mai avantajoasă pe calculatoarele ce permit calcularea rapidă a produselor scalare. Pentru o inițiere mai aprofundată în factorizarea lui Crout recomandăm lucrările [2,34,38].

Presupunem că se cunoaște o factorizare LU a matricei A . atunci sistemul de ecuații liniare $Ax=b$ este echivalent cu $LUx=b$, care se desface în două sisteme triunghiulare:

$$Ly=b, Ux=y.$$

Se rezolvă mai întâi sistemul inferior triunghiular $Ly = b$ printr-o procedură tipică de *substituție "înainte"* începând cu prima ecuație:

$$y_1 = \frac{b_1}{l_{11}}, \quad y_i = \frac{b_i - \sum_{k=1}^{i-1} l_{ik} y_k}{l_{ii}}, \quad i = 2, 3, \dots, n.$$

Aici l_{ik} sunt elementele matricei L . Apoi se rezolvă sistemul superior triunghiular $Ux = y$ prin procedura de *substituție "înapoi"*, începând cu ultima ecuație:

$$x_n = \frac{y_{nn}}{u_{nn}}, \quad x_i = \frac{y_i - \sum_{k=i+1}^n u_{ki} x_k}{u_{ii}}, \quad i = n-1, n-2, \dots, 1,$$

unde u_{ik} sunt elementele matricei U .

Utilizând o factorizare LU a lui A putem rezolva simultan mai multe sisteme de ecuații, având aceeași matrice A , fără a relua calculele de la început.

Metoda eliminării lui Gauss ne permite să calculăm și determinantul matricei A . Într-adevăr, să observăm că

$$\det(A) = \det(L) \cdot \det(U).$$

Deoarece $\det(L) = 1$, avem:

$$\det(A) = a_{11}^{(1)} \cdot a_{22}^{(2)} \cdot \dots \cdot a_{nn}^{(n)},$$

adică determinantul este produsul elementelor pivoți. Dacă se aplică una din strategiile de pivotare atunci:

$$\det(A) = (-1)^m a_{11}^{(1)} \cdot a_{22}^{(2)} \cdot \dots \cdot a_{nn}^{(n)}$$

unde m este numărul total de permutări efectuate.

3.5 Factorizarea Cholesky

Fie $A = (a_{ij})_{n \times n}$ o matrice simetrică și pozitiv definită, adică:

$$(Ax, x) > 0, \quad \forall x \in \mathbb{R}^n, x \neq 0.$$

Vom arăta că în factorizarea LU a lui A se poate alege $U = L^T$. Descompunerea

$$A = L \cdot L^T$$

se numește *factorizarea Cholesky*.

Teoremă. Dacă matricea A este simetrică și pozitiv definită, atunci există o matrice inferior triunghiulară L , cu elementele diagonale pozitive, unică, astfel încât $A = L \cdot L^T$.

Demonstrație. Vom demonstra teorema prin inducție asupra lui n . Pentru $n = 2$ avem:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad L = \begin{pmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{pmatrix}.$$

Dacă formăm produsul $L \cdot L^T$ și-l identificăm cu A , obținem:

$$\begin{aligned} a_{11} &= l_{11}^2, & a_{12} &= l_{11}l_{21}, \\ a_{21} &= l_{21}l_{11}, & a_{22} &= l_{21}^2 + l_{22}^2. \end{aligned}$$

Deoarece matricea A este pozitiv definită, elementele de pe diagonală sunt pozitive și deci putem extrage rădăcină pătrată: $l_{11} = \sqrt{a_{11}}$. Al doilea element l_{21} se determină din ecuația ce conține a_{21} ($a_{21} = a_{12}$ pentru că A este o matrice simetrică):

$$l_{21} = \frac{a_{21}}{\sqrt{a_{11}}}.$$

În fine elementul l_{22} se determină astfel :

$$l_{22} = \sqrt{a_{22} - l_{21}^2} = \sqrt{a_{22} - \frac{a_{12}^2}{a_{11}}}.$$

Să arătăm că extragerea rădăcinii pătrate de mai sus este posibilă. Într-adevăr, fie vectorul x cu componentele a_{12} și $-a_{11}$:

$$x = \begin{pmatrix} a_{12} \\ -a_{11} \end{pmatrix}.$$

Atunci $(Ax, x) = x^T Ax = a_{11}^2 a_{22} - a_{12}^2 a_{11} > 0$ fiindcă matricea A este pozitiv definită. Împărțind ultima inegalitate la $a_{11}^2 > 0$ obținem

$$a_{22} - \frac{a_{12}^2}{a_{11}} > 0.$$

Prin urmare pentru $n = 2$ descompunerea $A = L \cdot L^T$ există și este unică. Să presupunem teorema adevărată pentru $n = k - 1$: $A = L \cdot L^T$ unde A și L sunt matrice de dimensiune $(k - 1) \times (k - 1)$. Fie A' și L' astfel:

$$A' = \begin{pmatrix} A & y \\ y^T & a_{kk} \end{pmatrix}, \quad L' = \begin{pmatrix} L & 0 \\ w^T & l_{kk} \end{pmatrix},$$

unde y și w sunt vectorii coloană, având $k - 1$ componente.

Formăm produsul $L' \cdot (L')^T$ și-l identificăm cu A' . Atunci obținem:

$$\begin{aligned} A &= LL^T, & y &= Lw, \\ y^T &= w^T L^T, & a_{kk} &= l_{kk}^2 + w^T w. \end{aligned}$$

Prin presupunerea de inducție matematică, matricea L este determinată în mod unic astfel ca $A = LL^T$. De aici rezultă că vectorul w este și el unic și se calculează ca soluție al sistemului $y = Lw$. Mai departe, elementul l_{kk} se definește din formula:

$$l_{kk} = \sqrt{a_{kk} - w^T w}.$$

Arătăm ca și în cazul matricelor de dimensiune 2×2 că expresia de sub rădăcină este pozitivă. În calitate de vectorul x vom lua :

$$x = \begin{pmatrix} A^{-1}y \\ -1 \end{pmatrix}.$$

Notăm $z = A^{-1}y$; atunci

$$\begin{aligned} (Ax, x) &= x^T Ax = z^T Az - 2z^T y + a_{kk} = \\ &= -z^T y + a_{kk} = a_{kk} - y^T A^{-1}y = a_{kk} - y^T (LL^T)^{-1} y = \\ &= a_{kk} - (L^{-1}y)^T (L^{-1}y) = a_{kk} - w^T w > 0. \end{aligned}$$

Deci matricea L' cu proprietatea $A' = L' \cdot (L')^T$ este unic determinată și **teorema este demonstrată**.

Se poate arăta că factorizarea Cholesky a unei matrice $A = (a_{ij})_{n \times n}$ simetrice pozitiv definite necesită aproximativ $n^3/6$ operații (înmulțiri și adunări) și n extrageri de radical (necesare pentru a calcula elementele diagonale l_{kk}).

Metoda lui Cholesky de rezolvare a sistemelor de ecuații liniare se mai numește *metoda rădăcinii pătrate* și constă în descompunerea sistemului $Ax = b$ în sistemele triunghiulare:

$$L^T y = b, \quad Lx = y.$$

Elementele l_{ij} ale matricei inferior triunghiulare L pot fi calculate în modul următor: se determină prima coloană a matricei L

$$l_{11} = \sqrt{a_{11}}, \quad l_{i1} = \frac{a_{i1}}{l_{11}}, \quad i = 2, 3, \dots, n;$$

după ce s-au obținut primele $(k-1)$ coloane ale matricei L se calculează coloana k

$$l_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2},$$

$$l_{ik} = \frac{1}{l_{kk}} \left(a_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj} \right), \quad i = k+1, \dots, n.$$

O caracteristică remarcabilă al algoritmului Cholesky constă în stabilitatea lui numerică. Acest lucru rezultă din faptul că elementul maxim în modul al unei matrice simetrice și pozitiv definite este situat pe diagonală principală. În plus, elementele

diagonale ale matricei A și elementele l_{ij} ale matricei L satisfac relației:

$$l_{1k}^2 + l_{2k}^2 + \dots + l_{kk}^2 = a_{kk}, \quad k = 1, 2, \dots, n.$$

Astfel avem o limitare a creșterii elementelor matricei L : orice element l_{ij} nu depășește elementul maxim în modul $|a_{kk}|$.

3.6 Perturbații. Numărul de condiționare

Considerăm sistemul liniar $Ax = b$. Soluția exactă a sistemului considerat poate fi scrisă sub forma: $x^* = A^{-1}b$. Să presupunem că matricea nesingulară A și vectorul b suferă perturbațiile δA și δb .

Mai întâi să analizăm cazul când perturbăm numai termenul liber. Soluția perturbată \tilde{x} a sistemului cu partea dreaptă $b + \delta b$ satisface egalitatea:

$$A\tilde{x} = b + \delta b.$$

Obținem:

$$\tilde{x} - x^* = A^{-1}\delta b,$$

de unde rezultă că

$$\|\tilde{x} - x^*\| \leq \|A^{-1}\| \|\delta b\| \quad (3.5)$$

oricare ar fi norma matricială subordonată unei norme vectoriale. Pentru orice A și b există o perturbație δb astfel încât să avem egalitate în (3.5). Prin urmare $\|A^{-1}\|$ evaluează cu cât poate să crească eroarea furnizată de δb .

Pentru determinarea efectului relativ al aceleiași perturbații δb să observăm că:

$$\|b\| = \|Ax^*\| \leq \|A\| \|x^*\|,$$

sau

$$\frac{1}{\|x^*\|} \leq \frac{\|A\|}{\|b\|}.$$

Ținând seama de aceasta, inegalitatea (3.5) implică

$$\frac{\|\tilde{x} - x^*\|}{\|x^*\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|} \quad (3.6)$$

Presupunem acum că perturbăm elementele lui A ; atunci soluția perturbată \tilde{x} va verifica egalitatea:

$$(A + \delta A)\tilde{x} = b.$$

Rezultă că:

$$\tilde{x} - x^* = A^{-1} \delta A \tilde{x}$$

și obținem estimarea:

$$\|\tilde{x} - x^*\| \leq \|A^{-1}\| \|\delta A\| \|\tilde{x}\|. \quad (3.7)$$

Această inegalitate o putem pune sub forma:

$$\frac{\|\tilde{x} - x^*\|}{\|\tilde{x}\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta A\|}{\|A\|} \quad (3.8)$$

Se observă că atât în (3.6) cât și (3.8) numărul $\|A\| \|A^{-1}\|$ estimează eroarea relativă în soluție furnizată de δb sau δA . Acest număr se numește *număr de condiționare* al matricei A în raport cu normă matriceală considerată și se notează:

$$\text{cond}(A) = \|A\| \|A^{-1}\|.$$

Deoarece pentru orice normă matriceală subordonată unei norme vectoriale se îndeplinește egalitatea $\|I\| = 1$, avem:

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\| \|A^{-1}\|,$$

și deci, $\text{cond}(A) \geq 1$ oricare ar fi matricea A .

Numărul de condiționare caracterizează efectul maximal al perturbărilor δb și δA la rezolvarea sistemului $Ax = b$. Dacă numărul de condiționare $\text{cond}(A)$ este mare, atunci perturbații mici ale lui A și b vor produce perturbații relativ mari ale lui x^* ; în acest caz se spune că matricea A este *rău (sau prost) condiționată*. Matricele cu numărul de condiționare $\text{cond}(A)$ "mic" se numesc *bine condiționate*.

Subliniem că dimensiunea matricei nu are o influență directă asupra numărului ei de condiționare: dacă $A=I$ sau chiar $A = \frac{1}{10}I$ avem $cond(A)=1$. Pentru comparație, determinantul matricei nu este un indice adecvat al condiționării, deoarece valoarea determinantului depinde și de dimensiunea n a matricei. Dacă A este o matrice aproape singulară încă nu înseamnă că este prost condiționată. În exemplul $A = \frac{1}{10}I$ avem $\det(A)=10^{-n}$; această matrice "aproape singulară" este maxim de bine condiționată.

Exemplu. Considerăm sistemul de ecuații $Ax = b$, unde:

$$A = \begin{pmatrix} 1 & -1 & -1 & K & -1 & -1 \\ 0 & 1 & -1 & K & -1 & -1 \\ 0 & 0 & 1 & K & -1 & -1 \\ \dots\dots\dots & & & K & & \\ 0 & 0 & 0 & K & 1 & -1 \\ 0 & 0 & 0 & K & 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ -1 \\ -1 \\ M \\ -1 \\ 1 \end{pmatrix}.$$

Menționăm că $\det(A)=1 \neq 0$. Sistemul considerat în formă desfășurată este:

$$\begin{cases} x_1 - x_2 - x_3 - K - x_n = -1, \\ x_2 - x_3 - K - x_n = -1, \\ \dots\dots\dots \\ x_{n-1} - x_n = -1, \\ x_n = 1. \end{cases} \quad (3.9)$$

Sistemul de ecuații (3.9) admite o soluție exactă unică $x_1^* = x_2^* = K = x_{n-1}^* = 0, x_n^* = 1$. Presupunem că perturbăm vectorul liber b cu $\mathcal{B} = (0,0,K,0,\varepsilon)^T$. Atunci soluția exactă al sistemului

perturbat $A\tilde{x} = b + \delta b$ devine $\tilde{x} = x^* + r$ și eroarea $r = (r_1, r_2, K, r_n)^T$ satisface ecuațiilor:

$$\left\{ \begin{array}{l} r_1 - r_2 - r_3 - K - r_n = 0, \\ r_2 - r_3 - K - r_n = 0, \\ \dots\dots\dots \\ r_{n-1} - r_n = 0, \\ r_n = \varepsilon. \end{array} \right.$$

De unde rezultă, că:

$$\begin{aligned} r_n &= \varepsilon, \\ r_{n-1} &= r_n = \varepsilon, \\ r_{n-2} &= r_n + r_{n-1} = 2\varepsilon, \\ r_{n-3} &= r_n + r_{n-1} + r_{n-2} = 4\varepsilon = 2^2\varepsilon, \\ &\dots\dots\dots \\ r_{n-k} &= r_n + r_{n-1} + K + r_{n-(k+1)} = 2^{k-1}\varepsilon, \\ &\dots\dots\dots \\ r_1 &= r_{n-(n-1)} = 2^{(n-1)-1}\varepsilon = 2^{n-2}\varepsilon. \end{aligned}$$

Astfel $\tilde{x}_i = 2^{n-i-1}\varepsilon$, $i = 1, 2, K, n-1$; $\tilde{x}_n = 1 + \varepsilon$.

Să estimăm $cond(A)$, luând în calitate de normă $\|\cdot\|_\infty$. Vom avea:

$$\|\tilde{x} - x^*\|_\infty = \|r\|_\infty = 2^{n-2}\varepsilon, \quad \|x^*\|_\infty = 1, \quad \|\delta b\|_\infty = \varepsilon, \quad \|b\|_\infty = 1.$$

Prin urmare

$$cond(A) = \|A\|_\infty \cdot \|A^{-1}\|_\infty \geq \frac{\|r\|_\infty}{\|x^*\|_\infty} \cdot \frac{\|b\|_\infty}{\|\delta b\|_\infty} = 2^{n-2}.$$

Deoarece $\|A\|_\infty = n$, rezultă că norma matricei inverse este destul de mare, cu toate că $\det(A^{-1}) = \frac{1}{\det(A)} = 1$. De exemplu pentru $n = 102$ avem:

$$\|A\|_\infty = 102, \quad \text{cond}(A) \geq 2^{100} > 10^{30}, \quad \text{iar } \|A^{-1}\|_\infty > 10^{27}.$$

În particular, dacă $\varepsilon = 10^{-15}$ (o eroare suficient de mică), atunci $\|\tilde{x} - x^*\|_\infty = \|r\|_\infty > 10^{15}$; o perturbație destul de mică a termenului liber a produs o perturbație atât de mare în soluție!

Valoarea numărului de condiționare al unei matrice depinde de norma matriceală întrebuințată. Presupunem că A este o matrice simetrică și pozitiv definită cu valorile proprii pozitive ordonate astfel : $0 < \lambda_1 \leq \lambda_2 \leq K \leq \lambda_n$. În mod analog se arată că pentru astfel de matrice numărul de condiționare este:

$$\text{cond}(A) = \frac{\lambda_n}{\lambda_1} = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Exemplu: Valorile proprii ale matricei:

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1.0001 \end{pmatrix}$$

aproximativ sunt egale cu $\lambda_1 = \frac{1}{2}10^{-4}$ și $\lambda_2 = 2$. De aceea numărul

de condiționare $\text{cond}(A)$ este aproape egal cu $4 \cdot 10^4$. Ne putem aștepta că perturbații mici ale datelor inițiale vor produce mari schimbări în soluție. Într-adevăr, fie sistemele de ecuații $Ax = b$ și $Ax = b + \delta b$ unde

$$b = \begin{pmatrix} 2 \\ 2.0001 \end{pmatrix}, \quad \delta b = \begin{pmatrix} 0 \\ 0.0001 \end{pmatrix}.$$

Soluția se schimbă de la $x^* = (1 \ 1)^T$ până la $\tilde{x} = (0 \ 2)^T$:

$$\frac{\|x^* - \tilde{x}\|_2}{\|x^*\|_2} = \frac{\left\| \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right\|_2}{\left\| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\|_2} = 1, \quad \frac{\|\delta b\|_2}{\|b\|_2} = \frac{\left\| \begin{pmatrix} 0 \\ 0.0001 \end{pmatrix} \right\|_2}{\left\| \begin{pmatrix} 2 \\ 2.0001 \end{pmatrix} \right\|_2} = \frac{10^{-4}}{2\sqrt{2}}.$$

Fie dat un sistem de ecuații liniare. Reprezentarea cu virgulă mobilă a elementelor A și B în calculatorul electronic nu este exactă. Prin urmare, efectiv în memoria mașinii electronice de calcul vom avea sistemul $\tilde{A}x = \tilde{b}$ unde \tilde{A} și \tilde{b} sunt rotunjirile corespunzătoare. Există matricele de perturbare P și D (D este o matrice diagonală) astfel încât

$$\tilde{A} = A(I + P) \quad \tilde{b} = (I + D)b.$$

Dacă notăm cu ε_M unitatea de rotunjire a mașinii (vezi paragraful 1.3), atunci $\|P\| \leq \varepsilon_M$ și $\|D\| \leq \varepsilon_M$. Astfel obținem:

$$\|\delta A\| = \|\tilde{A} - A\| \leq \varepsilon_M \|A\|, \quad \|\delta b\| = \|\tilde{b} - b\| \leq \varepsilon_M \|b\|.$$

Din inegalitățile de mai sus și din (3.7) rezultă că erorile de rotunjire produc o perturbație estimată prin:

$$\|\tilde{x} - x^*\| \leq 2\varepsilon_M \|\tilde{x}\| \text{cond}(A).$$

Subliniem că determinarea numărului de condiționare $\text{cond}(A)$ este o problemă dificilă, deoarece conține calculul lui $\|A^{-1}\|$. Calculul matricei inverse și normei sale necesită aproximativ $n^3 + 2n^2$ operații suplimentare și aproape de patru ori majorează cheltuielile necesare pentru rezolvarea sistemului $Ax = b$. Un procedeu practic de calcul aproximativ al lui $\|A^{-1}\|$ constă în următoarele. Se observă că dacă $w = A^{-1}y$ atunci $\|w\| \leq \|A^{-1}\| \|y\|$ și prin urmare

$$\|A^{-1}\| \geq \frac{\|w\|}{\|y\|}.$$

Deci, se poate alege k vectori y_i , apoi se rezolvă sistemul de ecuații $Aw_i = y_i$, $i = 1, 2, \dots, k$, și se pune

$$\|A^{-1}\| \approx \max_{1 \leq i \leq k} \frac{\|w_i\|}{\|y_i\|}.$$

3.7 Calculul valorilor și vectorilor proprii

3.7.1. Formularea problemei. Proprietăți fundamentale

Fie A o matrice de dimensiune $n \times n$. Numărul λ (real sau complex) se numește *valoare proprie* a matricei A dacă există un vector nenul $x \in R^n$, astfel încât

$$Ax = \lambda x \quad (3.10)$$

Vectorul $x \neq 0$ se numește *vector propriu* al lui A asociat valorii proprii λ .

Ecuația (3.10) poate fi rescrisă sub forma $Ax - \lambda x = 0$ sau

$$(A - \lambda I)x = 0, \quad (3.11)$$

unde

$$\lambda I = \begin{pmatrix} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \lambda \end{pmatrix}, \quad 0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Ecuația (3.11) poate fi scrisă dezvoltat astfel:

$$\begin{pmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}.$$

Relația de mai sus este un sistem omogen de ecuații liniare. Condiția necesară și suficientă ca acest sistem omogen să admită soluție nenulă este ca:

$$\det(A - \lambda I) = 0. \quad (3.12)$$

Acest determinant este un polinom de grad n cu coeficienți reali și se notează de obicei:

$$P_n(\lambda) = (-1)^n \lambda^n + p_1 \lambda^{n-1} + \dots + p_{n-1} \lambda + p_n.$$

Polinomul $P_n(\lambda)$ se numește *polinom caracteristic* asociat matricii A , iar ecuația (3.12) se numește *ecuația caracteristică* a matricii A .

Ecuația caracteristică este o ecuație algebrică de grad n cu coeficienți reali care în virtutea teoremei fundamentale a algebrei are exact n rădăcini $\lambda_1, \lambda_2, \dots, \lambda_n$, în general complexe și nu neapărat distincte.

Mulțimea valorilor proprii ale matricii A se numește *spectrul* lui A și se notează cu $\sigma(A)$:

$$\sigma(A) = (\lambda_1, \lambda_2, \dots, \lambda_n).$$

Raza spectrală a lui A se notează $\rho(A)$ și se definește prin relația:

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|.$$

De aceea norma matriceală $\|A\|_2 = \sqrt{\rho(A^T A)}$ subordonată normei euclidiene $\|x\|_2 = \sqrt{\sum x_i^2}$ se mai numește *normă spectrală*.

Pentru orice normă matriceală subordonată unei norme vectoriale avem:

$$\rho(A) \leq \|A\|.$$

Mai mult ca atât:

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n} \|A\|_2,$$

unde $\|A\|_F$ este norma lui Frobenius (definiția vezi în paragraful 3.1.2.).

Dacă A este o matrice simetrică, adică $A = A^T$, atunci

$$\|A\|_2 = \max_{\lambda \in \sigma(A)} |\lambda| = \rho(A)$$

și

$$\|A\|_F = \sqrt{\sum_{i=1}^n \lambda_i^2}.$$

Dacă se cunoaște o valoare proprie atunci un vector propriu asociat acestei valori proprii este soluția nenulă a sistemului omogen (2). Pe de altă parte dacă se știe un vector propriu v atunci $Av = \lambda v$, de unde $\lambda(v, v) = (Av, v)$: deci valoarea proprie corespunzătoare se obține imediat:

$$\lambda = \frac{(Av, v)}{(v, v)}$$

Exemplu: Să se calculeze valorile proprii și vectorii proprii pentru matricea:

$$A = \begin{pmatrix} 4 & -5 \\ 2 & -3 \end{pmatrix}.$$

Polinomul caracteristic este:

$$\begin{aligned} P_2(\lambda) &= \det(A - \lambda I) = \begin{vmatrix} 4 - \lambda & -5 \\ 2 & -3 - \lambda \end{vmatrix} = \\ &= (4 - \lambda)(-3 - \lambda) + 10 = \lambda^2 - \lambda - 2 \end{aligned}$$

Se obține: $\lambda_1 = -1$, $\lambda_2 = 2$. Prin urmare matricea A are două valori proprii distincte. Înlocuind fiecare valoare proprie în sistemul omogen (3.11) obținem vectorul propriu asociat valorii proprii. Pentru $\lambda_1 = -1$ avem:

$$(A - \lambda_1 I)x = \begin{pmatrix} 5 & -5 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

de unde vectorul propriu :

$$v^1 = c \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad c = \text{const} \neq 0.$$

În mod analog pentru $\lambda_2 = 2$ vom avea:

$$(A - \lambda_2 I)x = \begin{pmatrix} 2 & -5 \\ 2 & -5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad v^2 = c \begin{pmatrix} 5 \\ 2 \end{pmatrix}, \quad c \neq 0.$$

În exemplul considerat spectrul $\sigma(A) = (-1, -2)$, iar raza spectrală $\rho(A) = 2$.

Subliniem că vectorii proprii al unei matrice se determină neunivoc, deoarece dacă x este un vector propriu atunci orice vector cx , unde c este un scalar, va fi tot un vector propriu.

Dăm câteva rezultate remarcabile referitor la valorile proprii ale unor matrice speciale:

1. Valorile proprii ale matricelor simetrice sunt reale.
2. Valorile proprii ale matricelor pozitiv definite sunt pozitive.
3. Valorile proprii ale matricelor diagonale, inferior triunghiulare și superior triunghiulare coincid cu elementele de pe diagonala principală.

Proprietățile 1-3 se demonstrează imediat, reieșind din definiția valorilor și vectorilor proprii.

4. Matricele asemenea au aceleași valori proprii.

Demonstrație: Fie matricele A și B sunt asemenea. Aceasta înseamnă că există o matrice nesingulară M astfel încât

$$B = M^{-1}AM.$$

Dacă $Ax = \lambda x$ atunci $M^{-1}Ax = \lambda M^{-1}x$. Vom nota $x = My$.

Rezultă că

$$M^{-1}AM = \lambda y \text{ și } By = \lambda y.$$

Prin urmare la matricele asemenea valorile proprii coincid. În plus, vectorii proprii sunt legați prin relația $x = My$.

5. Suma valorilor proprii ale unei matrice A este egală cu suma elementelor de pe diagonala principală:

$$\lambda_1 + \lambda_2 + \dots + \lambda_n = a_{11} + a_{22} + \dots + a_{nn}.$$

Această sumă se numește *urma matricei* A și se notează $Tr(A)$ sau $Sp(A)$.

6. Produsul valorilor proprii coincide cu determinantul matricei:

$$\lambda_1 \cdot \lambda_2 \cdot \dots \cdot \lambda_n = \det(A).$$

7. Dacă $\lambda_i, 1 \leq i \leq n$ sunt valorile proprii ale matricei A , atunci $\lambda_i^k, 1 \leq i \leq n$, sunt valorile proprii ale matricei

$$A^k = A \cdot A \cdot \dots \cdot A.$$

Notăm

$$Q_n(\lambda) = (-1)^n P_n(\lambda) = \lambda^n + q_1 \lambda^{n-1} + \dots + q_{n-1} \lambda + q_n.$$

Polinomul $Q_n(\lambda)$ se numește *polinom propriu*.

8. Formulele lui Newton:

$$\mu_k + \sum_{i=1}^{k-1} q_i \mu_{k-i} = -k q_k, \quad k = 1, 2, \dots, n,$$

unde

$$\mu_k = \sum_{i=1}^n \lambda_i^k, \quad k = 1, 2, \dots, n,$$

iar $\lambda_1, \lambda_2, \dots, \lambda_n$ sunt rădăcinile ecuației caracteristice.

Se observă că $\mu_k = \text{Tr}(A^k)$. Suma $\sum_{i=1}^n \lambda_i^k$ se mai numește

moment de ordinul k al valorilor proprii.

9. **Identitatea lui Cayley - Hamilton.** Orice matrice pătrată A este o rădăcină a polinomului său caracteristic:

$$(-1)^n A^n + p_1 A^{n-1} + \dots + p_{n-1} A + p_n I = 0.$$

10. **Teorema despre cercurile lui Gershgorin.** Orice valoare proprie λ a matricei $A = (a_{ij})_{n \times n}$ se află, în planul complex, în reuniunea cercurilor :

$$\bigcup_{i=1}^n \left\{ z : |z - a_{ii}| \leq r_i; \quad r_i = \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| \right\}.$$

11. **Teorema (Rayleigh-Ritz).** Fie A o matrice simetrică și valorile proprii ale lui A sunt așezate în ordine crescătoare

$$\lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_{n-1} \leq \lambda_n = \lambda_{\max}.$$

Atunci

$$\lambda_1 \|x\|_2^2 \leq (Ax, x) \leq \lambda_n \|x\|_2^2, \quad \forall x \in R^n,$$

$$\lambda_1 = \lambda_{\min} = \min_{x \neq 0} \frac{(Ax, x)}{(x, x)} = \min_{\|x\|_2=1} (Ax, x),$$

$$\lambda_n = \lambda_{\max} = \max_{x \neq 0} \frac{(Ax, x)}{(x, x)} = \max_{\|x\|_2=1} (Ax, x).$$

Expresia $\frac{(Ax, x)}{(x, x)}$ se numește *câtul lui Rayleigh*.

Metodele de calcul ale valorilor proprii se împart în două grupe:

- 1) metode care determină întâi coeficienții polinomului caracteristic și apoi se rezolvă ecuația caracteristică;
- 2) metode care determină valorile proprii și vectorii proprii prin procedee iterative fără a calcula coeficienții polinomului caracteristic .

Se cunosc metode de determinare a coeficienților polinomului caracteristic: metoda lui Krîlov ,metoda lui Leverrier ,metoda lui Fadeev , metoda lui Lanczos (vezi, de exemplu, [12,16,18,27]). Subliniem că aceste metode sunt recomandabile doar în cazurile când matricea A este de ordin mic și rădăcinile ecuației caracteristice sunt bine separate. Motivul este că numărul de operații aritmetice pentru determinarea coeficienților polinomului caracteristic este foarte mare (de exemplu, metoda lui Fadeev necesită n^4 operații). Pe de altă parte, coeficienții se obțin cu erori de rotunjire inerente care pot conduce la variații mari ale rădăcinilor, deoarece problema rezolvării ecuațiilor algebrice este rău condiționată .

Fie ,de exemplu ,matricea A de forma [2,28] :

$$A(\varepsilon) = \begin{pmatrix} \alpha & 1 & 0 & K & 0 & 0 \\ 0 & \alpha & 1 & K & 0 & 0 \\ 0 & 0 & \alpha & K & 0 & 0 \\ M & M & M & M & M & M \\ \varepsilon & 0 & 0 & K & 0 & \alpha \end{pmatrix},$$

unde ε este o perturbație mică iar α valoarea proprie de multiplicitate n a matricei neperturbate $A(0)$. Polinomul caracteristic al lui $A(\varepsilon)$ este

$$P_n(\varepsilon, \lambda) = (\lambda - \alpha)^n - (-1)^n \varepsilon.$$

Se observă că, de exemplu, dacă $n=10$, $\alpha=0$, și $\varepsilon=10^{-10}$, atunci practic matricele $A(0)$ și $A(10^{-10})$ reprezintă una și aceeași matrice în memoria mașinii electronice de calcul, în timp ce polinomul caracteristic $P_{10}(10^{-10}, \lambda)$ are rădăcina multiplă 0,1. Prin urmare, valoarea proprie $\alpha=0$ suferă o deplasare de 10^9 ori mai mare decât o perturbația ε care a produs-o.

3.7.2. Metode bazate pe transformări de asemănare ortogonală

Metodele practice de determinare a valorilor și vectorilor proprii constau în proceduri iterative de aducere a matricei considerate la forma canonică Schur prin transformări de asemănare ortogonală.

Două matrice A și B se numesc *ortogonal asemenea* dacă există o matrice ortogonală Q astfel încât

$$B = Q^T A Q$$

Transformarea $Q^T A Q$ a lui A se numește *transformare de asemănare ortogonală*.

Matricele A și $Q^T A Q$ au aceleași valori proprii, deoarece $Q^T = Q^{-1}$ (vezi proprietatea 4 din 3.7.1.). Dacă x este un vector propriu pentru matricea A , atunci $Q^T x$ este vector propriu pentru matricea $Q^T A Q$.

Teoremă (Schur). Oricare ar fi matricea $A = (a_{ij})_{n \times n}$ există o matrice ortogonală Q de dimensiune $n \times n$ astfel încât :

$$Q^T A Q = S = \begin{pmatrix} s_{11} & s_{12} & \mathbf{K} & s_{1k} \\ 0 & s_{22} & \mathbf{K} & s_{2k} \\ \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} \\ 0 & 0 & 0 & s_{11} \end{pmatrix},$$

unde $s_{ii}, i=1,2,\mathbf{K},k$, sunt blocuri diagonale de dimensiune 1×1 sau 2×2 . Blocurile s_{ii} de dimensiune 1×1 reprezintă valorile proprii reale ale matricei A iar fiecare bloc de dimensiune 2×2 reprezintă valorile proprii complex conjugate.

Demonstrația acestei teoreme poate fi găsită în [2,40].

Matricea S se numește *formă canonică Schur reală* a lui A .

Dacă matricea A are numai valori proprii reale atunci matricea S este superior triunghiulară . Dacă A are și valori proprii complexe atunci S este *cvasi-triunghiulară* .

Stabilirea unei forme canonice Shur este precedată de o pregătire inițială a matricei A , aducând-o la forma compactă numită *forma Hessenberg*:

$$A = \begin{pmatrix} h_{11} & h_{12} & h_{13} & \mathbf{K} & h_{1,n-1} & h_{1n} \\ h_{21} & h_{22} & h_{23} & \mathbf{K} & h_{2,n-1} & h_{2n} \\ 0 & h_{32} & h_{33} & \mathbf{K} & h_{3,n-1} & h_{3n} \\ \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} \\ 0 & 0 & 0 & \mathbf{K} & h_{n-1,n-1} & h_{n-1,n} \\ 0 & 0 & 0 & \mathbf{K} & h_{n,n-1} & h_{nn} \end{pmatrix}$$

Pentru matricele simetrice forma Hessenberg devine o formă simetrică tridiagonală.

Metoda lui Householder

Reducerea oricărei matrice A la forma Hessenberg se poate face după un număr finit de rotații plane printr-un procedeu numit *metoda lui Givens*. O altă metodă mult mai eficientă este *metoda reflexiilor a lui Householder*.

Reamintim că matricea

$$H = I - 2 \frac{vv^T}{\|v\|_2^2}$$

se numește *reflector* sau *transformarea lui Householder*.

Fie $z = (1, 0, K, 0)^T \in \mathbb{R}^n$, $\sigma = \|x\|_2$ și $v = x + \sigma z$. Atunci

$$Hx = -\sigma z = (-\sigma, 0, K, 0)^T.$$

Într-adevăr,

$$\begin{aligned} Hx &= x - 2 \frac{vv^T x}{\|v\|_2^2} = x - (x + \sigma z) \frac{2(x + \sigma z)^T x}{(x + \sigma z)^T (x + \sigma z)} = \\ &= x - (x + \sigma z) = -\sigma z, \end{aligned}$$

deoarece $x^T x = \sigma^2$.

Procedura directă de adunare a matricei A la forma lui Householder constă din $n-2$ etape și se bazează pe egalitatea mai sus demonstrată. În prima etapă se determină matricea U_1 astfel încât ultimele $n-2$ elemente din prima coloană a matricei $U_1 A$ să fie nule. Pentru aceasta notăm

$$x = \begin{pmatrix} a_{21} \\ a_{31} \\ \mathbf{M} \\ a_{n1} \end{pmatrix}, \quad z = \begin{pmatrix} 1 \\ 0 \\ \mathbf{M} \\ 0 \end{pmatrix}, \quad Hx = \begin{pmatrix} -\sigma \\ 0 \\ \mathbf{M} \\ 0 \end{pmatrix}.$$

Aici H este matricea lui Householder de dimensiune $(n-1) \times (n-1)$. Calculul lui U_1 se face conform relației :

$$U_1 = \begin{pmatrix} 1 & 0 & K & 0 \\ 0 & & & \\ M & & H & \\ 0 & & & \end{pmatrix} = U_1^T = U_1^{-1}.$$

Deoarece în colțul “stânga-sus” este situată unitatea, multiplicarea matricei U_1 cu matricea A nu modifică prima linie a lui A dar modifică ultimele $n-1$ elemente din coloanele următoare $j=2,3,K,n$. După ce matricea U_1A a fost obținută, se calculează matricea $(U_1A)U_1$. Multiplicarea matricei U_1A cu U_1 nu modifică prima coloană a lui U_1A , astfel încât matricea

$$U_1^{-1}AU_1 = \begin{pmatrix} \alpha_{11} & * & K & * \\ -\sigma & * & K & * \\ 0 & * & K & * \\ M & M & M & M \\ 0 & * & K & * \end{pmatrix}$$

este de formă Hessenberg în prima coloană. Prima etapă este complet terminată.

Etapă a doua este asemănătoare primei: se ia x egal cu vectorul din $n-2$ elemente ale coloanei a doua a matricei $U_1^{-1}AU_1$, z un vector unitate de dimensiune corespunzătoare, iar H_2 va fi o matrice de dimensiune $(n-2) \times (n-2)$. În mod analog obținem :

$$U_2 = U_2^{-1} = \begin{pmatrix} 1 & 0 & 0 & K & 0 \\ 0 & 1 & 0 & K & 0 \\ 0 & 0 & & & \\ M & M & & H_2 & \\ 0 & 0 & & & \end{pmatrix}, U_2(U_1AU_1)U_2 = \begin{pmatrix} * & * & * & K & * \\ * & * & * & K & * \\ 0 & * & * & K & * \\ 0 & 0 & * & K & * \\ M & M & M & K & M \\ 0 & 0 & * & K & * \end{pmatrix}$$

Astfel matricea:

$$U_{n-2}U_{n-3}K U_1AU_1U_2K U_{n-3}U_{n-2}$$

Obținută în final, după parcurgerea celor $n-2$ etape ale procedurii, este de formă Hessenberg. Numărul de operații aritmetice necesar este egal cu $5n^3/3$.

Dacă aplicăm transformările descrise unei matrice simetrice, atunci toate matricele $U_kU_{k-1}K U_1AU_1U_2K U_{k-1}U_k$ vor fi de asemenea simetrice și în final obținem o matrice Hessenberg simetrică tridiagonală. În acest caz este nevoie doar de $2n^3/3$ operații aritmetice .

Exemplu: Considerăm matricea [35]:

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Vom avea:

$$x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad z = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad v = x + \sigma z = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$H = I - 2 \frac{vv^T}{\|v\|_2^2} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \quad Hx = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Deci

$$U = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}, \quad U^{-1}AU = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Algoritmul QR

Algoritmul *QR* construiește iterativ un șir de matrice ortogonal asemenea $A_0 = A, A_1, A_2, K$ convergent către forma canonică Schur a matricei A . Prin urmare acest algoritm rezolvă problema completă a valorilor proprii .

Fie A o matrice de dimensiune $n \times n$. Notăm $A_0 = A$ și considerăm o factorizare QR a matrice A_0 (vezi compartimentul 3.9.3):

$$A_0 = R_0 Q_0,$$

unde Q_0 este o matrice ortogonală, iar R_0 este o matrice superior triunghiulară. Matricea următoare A_1 se obține multiplicând factorii Q_0 și R_0 în ordine inversă :

$$A_1 = R_0 Q_0$$

Matricele A_0 și A_1 sunt ortogonal asemenea:

$$Q_0^{-1} A_0 Q_0 = Q_0^{-1} (Q_0 R_0) Q_0 = R_0 Q_0 = A_1.$$

În general se calculează șirul matriceal $A_0, A_1, A_2, \dots, A_K$ definit recurent prin formulele:

$$A_0 = Q_0 R_0, \quad A_1 = R_0 Q_0,$$

$$A_1 = Q_1 R_1, \quad A_2 = R_1 Q_1,$$

.....

$$A_k = Q_k R_k, \quad A_{k+1} = R_k Q_k.$$

Se poate demonstra (vezi, de exemplu, [2,22,26,32]) că în condiții destul de generale , șirul $\{A_k\}_{k=0}^{\infty}$ converge și matricea limită $A_{\infty} = S$ coincide cu forma canonică Schur a lui A . În plus, matricea limită A_{∞} are pe diagonală valorile proprii ale lui A .

Dacă unele rapoarte $\frac{\lambda_{m+1}}{\lambda_m}$ sunt aproape de valoarea 1 convergența este încetă. În acest caz algoritmul QR se modifică astfel:

$$A_k - \alpha_k I = Q_k R_k, \quad A_{k+1} = R_k Q_k + \alpha_k I, \quad k = 0, 1, \dots, K.$$

unde α_k este un parametru de accelerare a convergenței numit *deplasare*. Deoarece

$$Q_k^{-1} A_k Q_k = Q_k^{-1} (Q_k R_k + \alpha_k I) Q_k = R_k Q_k + \alpha_k I = A_{k+1}$$

fiecare pas în algoritmul QR cu deplasări este o transformare cu asemănare ortogonală.

Dacă deplasările α_k se aleg suficient de apropiate de o valoare proprie λ a lui A vom avea o convergență rapidă a șirului $\{A_k\}$ către forma canonică Schur.

O iterație al algoritmului QR necesită $\frac{4n^3}{3}$ operații, ceea ce este exagerat. În practică se recomandă inițial de-a aduce matricea considerată la forma Hessenberg prin metoda lui Householder; apoi asupra formei Hessenberg se aplică algoritmul QR . Astfel numărul de operații la o iterație se reduce la $4n^2$, în cazul matricelor nesimetrice, și la aproximativ $12n$ în cazul matricelor simetrice.

3.7.3. Metoda puterii

Metoda puterii este cea mai simplă metodă de determinare a celei mai mari valori proprii (în modul) a unei matrice reale A de dimensiune $n \times n$ și a vectorului propriu corespunzător.

Fie A o matrice simplă. Matricea A de dimensiune $n \times n$ se numește simplă dacă are exact n vectori proprii liniar independenți e_1, e_2, \dots, e_n . Acești vectori formează o bază a spațiului n -dimensional și pot fi aleși astfel încât:

$$\|e_i\|_2 = \sqrt{(e_i, e_i)} = 1, \quad i = 1, 2, \dots, n.$$

De exemplu, orice matrice simetrică este o matrice simplă. Vom presupune că matricea A admite o valoare proprie reală dominantă λ_1 adică

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|.$$

Alegem un vector inițial arbitrar $x^{(0)}$ nenul. Deoarece vectorii proprii e_1, e_2, \dots, e_n formează o bază, există scalarii c_1, c_2, \dots, c_n nu toți nuli, astfel încât:

$$x^{(0)} = c_1 e_1 + c_2 e_2 + \dots + c_n e_n.$$

Vom presupune că $c_1 \neq 0$; în caz contrar se poate alege alt vector inițial $x^{(0)}$ astfel încât coeficientul corespunzător $c_1 \neq 0$.

Recurent formăm șirul de vectori:

$$x^{(k)} = Ax^{(k-1)} = A^k x^{(0)}, \quad k = 1, 2, K$$

Deoarece $Ae_i = \lambda_i e_i$ putem scrie:

$$\begin{aligned} x^{(1)} &= Ax^{(0)} = A(c_1 e_1 + c_2 e_2 + \dots + c_n e_n) = \\ &= c_1 A e_1 + c_2 A e_2 + \dots + c_n A e_n = \\ &= c_1 \lambda_1 e_1 + c_2 \lambda_2 e_2 + \dots + c_n \lambda_n e_n. \end{aligned}$$

În general șirul $\{x^{(k)}\}$ are reprezentarea:

$$x^{(k)} = c_1 \lambda_1^k e_1 + c_2 \lambda_2^k e_2 + \dots + c_n \lambda_n^k e_n = \lambda_1^k (c_1 e_1 + \eta^{(k)}).$$

unde

$$\eta^{(k)} = c_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k e_2 + \dots + c_n \left(\frac{\lambda_n}{\lambda_1} \right)^k e_n.$$

Prin ipoteză $\left| \frac{\lambda_i}{\lambda_1} \right| < 1$ pentru $i \geq 2$; deci $\|\eta^{(k)}\|_2 = 0 \left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right)$ și

$$\lim_{k \rightarrow \infty} \|\eta^{(k)}\|_2 = 0.$$

Calculăm produsul scalar:

$$\begin{aligned} (Ax^{(k-1)}, x^{(k-1)}) &= (x^{(k)}, x^{(k-1)}) = \lambda_1^{2k-1} (c_1 e_1 + \eta^{(k)}, c_1 e_1 + \eta^{(k-1)}) = \\ &= \lambda_1^{2k-1} [c_1^2 (e_1, e_1) + c_1 (e_1, \eta^{(k-1)}) + c_1 (\eta^{(k)}, e_1) + (\eta^{(k)}, \eta^{(k-1)})] \end{aligned}$$

Deoarece $(e_1, e_1) = \|e_1\|_2^2 = 1$ și

$$\begin{aligned} |(e_1, \eta^{(k-1)})| &\leq \|e_1\|_2 \|\eta^{(k-1)}\|_2 = \|\eta^{(k-1)}\|_2, \\ |(\eta^{(k)}, e_1)| &\leq \|\eta^{(k)}\|_2 \|e_1\|_2 = \|\eta^{(k)}\|_2, \\ |(\eta^{(k)}, \eta^{(k-1)})| &\leq \|\eta^{(k)}\|_2 \|\eta^{(k-1)}\|_2, \end{aligned}$$

deducem că

$$(Ax^{(k-1)}, x^{(k-1)}) = \lambda_1^{2k-1} \left[c_1^2 + o\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{k-1}\right) \right].$$

În mod analog obținem:

$$(x^{(k-1)}, x^{(k-1)}) = \lambda_1^{2k-2} \left[c_1^2 + o\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{k-1}\right) \right].$$

Prin urmare cotelul lui Rayleigh al lui $x^{(k-1)}$

$$\frac{(Ax^{(k-1)}, x^{(k-1)})}{(x^{(k-1)}, x^{(k-1)})} = \lambda_1 + o\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{k-1}\right)$$

va tinde către valoarea proprie (maximă în modul) λ_1 .

Se observă că

$$\|x^{(k)}\|_2 = \sqrt{(x^{(k)}, x^{(k)})} = |\lambda_1|^k \left[c_1 + o\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \right].$$

De aici rezultă că pentru $|\lambda_1| > 1$ avem $\|x^{(k)}\|_2 \rightarrow \infty$ iar dacă $|\lambda_1| < 1$ atunci $\|x^{(k)}\|_2 \rightarrow 0$. Rezolvând problema la mașina electronică de calcul în primul caz putem avea apariția unui semnal de depășire, după aceea, calculele se întrerup. În cazul al doilea $\|x^{(k)}\|_2$ poate deveni un zero-mașină. De aceea în practică este necesar să schimbăm factorul de scară al șirului $\{x^{(k)}\} = \{Ax^{(k-1)}\}$; în locul șirului $\{x^{(k)}\}$ se formează un alt șir de vector $\{z^{(k)}\}$ cu $\|z^{(k)}\|_2 = 1$ și construit astfel:

$$z^{(k)} = \frac{Az^{(k-1)}}{\|Az^{(k-1)}\|}, \quad k = 1, 2, K,$$

de unde

$$\frac{(Az^{(k-1)}, z^{(k-1)})}{(z^{(k-1)}, z^{(k-1)})} = (z^{(k)}, z^{(k-1)}) \rightarrow \lambda_1.$$

Exemplu: Să se calculeze valoarea proprie maximă λ_1 pentru matricea:

$$A = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 2 & 0 \\ -1 & 0 & 2 \end{pmatrix}$$

Folosim ca vector de start pe $x^{(0)} = (0, 0, 1)^T$. Calculele le aranjăm într-o schemă ușor de urmărit:

A	$x^{(0)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$...	$x^{(\infty)}$
1 -1 -1	0	-1	-3	-9	-25	-75	...	$-\infty$
-1 2 0	0	0	1	3	15	45	...	$+\infty$
-1 0 2	1	2	5	13	35	95	...	$+\infty$
	0	$-1/\sqrt{5}$	$3/\sqrt{35}$	$-9/\sqrt{259}$...	$-\frac{1}{\sqrt{5}}$
	0	0	$1/\sqrt{35}$	$3/\sqrt{259}$...	$\frac{1}{\sqrt{35}}$
	1	$2/\sqrt{5}$	$5/\sqrt{35}$	$13/\sqrt{259}$...	$\frac{1}{\sqrt{5}}$
		—	3	3	2.8	3	...	3
		—	—	3	3	3	...	3
		2	2.5	2.6	2.69	2.7	...	3

În schema de mai sus, în penultim tabel avem vectorii $z^{(k)} = \frac{x^{(k)}}{\|x^{(k)}\|_2}$ iar în ultimul – câtul $\frac{(x^{(k)})_i}{(x^{(k-1)})_i}$, unde prin $(x^{(k)})_i$ s-a notat componenta i a vectorului $x^{(k)}$. Valorile proprii ale matricei considerate sunt: $\lambda_1 = 3, \lambda_2 = 2$ și $\lambda_3 = 0$. Câtul $\frac{(x^{(k)})_i}{(x^{(k-1)})_i}$ tinde către valoarea maximă λ_1 , iar șirul de vectori $\{z^{(k)}\}$ converge către

vectorul propriu $\left(\frac{-1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)^T$ corespunzător acestei valori proprii.

Metoda puterii poate fi folosită și pentru calculul valorii proprii minime (în modul) λ_n cu condiția ca să avem $|\lambda_n| < |\lambda_{n-1}|$. Într-adevăr, dacă matricea A este nesingulară, atunci $A^{-1}x = \lambda^{-1}x$; deci valoarea proprie maximă a lui A^{-1} va fi valoarea proprie minimă a lui A .

Acum șirul de vectori $\{z^{(k)}\}$ se formează în felul următor:

$$z^{(k)} = \frac{A^{-1}z^{(k-1)}}{\|Az^{(k-1)}\|}, \quad k = 1, 2, \dots, K,$$

Această metodă se numește *metoda puterii inverse*. Subliniem că în practică în metoda puterii inverse vectorul $z^{(k)}$ se determină în urma rezolvării sistemului de ecuații liniare:

$$Az^{(k)} = \frac{z^{(k-1)}}{\|Az^{(k-1)}\|}.$$

O altă metodă mult mai eficientă este *metoda puterii inverse cu deplasări*. Observăm că matricea $A - \alpha I$ are valorile proprii $\lambda_i - \alpha$ și aceeași vectori proprii ca și A . De aceea, dacă se aleg deplasările α_m suficient de apropiate de valoarea proprie λ a lui A , atunci șirul $\{z^{(k)}\}$ definit prin

$$(A - \alpha_m I)z^{(k)} = \beta_k z^{(k-1)}$$

converge către valoarea proprie e_1 cu o viteză destul de rapidă. Parametrul β_k este un factor de normare, ales astfel încât $\|z^{(k)}\|_2 = 1$. Dacă A este inversabilă și luăm $\alpha_m = 0$ obținem $Az^{(k)} = \beta_k z^{(k-1)}$ - metoda puterii inverse.

În practică în scopul accelerării convergenței șirului $\{z^{(k)}\}$ către vectorul propriu cea mai bună alegere a deplasării α_m pentru matrice simetrice este câțul Rayleigh:

$$\alpha_m = \frac{(Az^{(m)}, z^{(m)})}{(z^{(m)}, z^{(m)})}.$$

Se poate demonstra că $\{\alpha_m\}$ converge către λ_1 extrem de rapid (cu viteza cubică de convergență).

Aici încheiem discutarea metodelor de calcul al valorilor proprii și vectorilor proprii. Subliniem încă o dată că algoritmul *QR* este unul dintre cele mai remarcabile metode ale matematicii aplicate. Cititorului dornic de a-și aprofunda cunoștințele în acest domeniu I se recomandă lucrările [2,22,26,28,35,40].

3.8 Metode iterative de rezolvare a sistemelor de ecuații liniare

Să considerăm sistemul de ecuații liniare:

$$Ax = b \tag{3.13}$$

unde $b \in R^n$ iar A este o matrice nesingulară ($\det(A) \neq 0$) de dimensiune $n \times n$. Metoda eliminării a lui Gauss de rezolvare a sistemului liniar (3.13) necesită cel puțin $n^3/3$ operații aritmetice și atât timp cât acest număr de operații este acceptabil putem utiliza această metodă. Pe de altă parte, când $n^3/3$ este mare, cu ajutorul metodelor iterative se poate obține cu aproximație satisfăcătoare a soluției după efectuarea unui număr mult mai mic de operații aritmetice.

Metodele iterative se construiesc utilizând desfacerea matriciei A difinită prin:

$$A = S - T.$$

Atunci sistemul (1) este echivalent cu sistemul :

$$Sx = Tx + b, \tag{3.14}$$

sau

$$x = Qx + d, \quad (3.15)$$

unde: $Q = S^{-1}T$, $d = S^{-1}b$. Prin urmare putem construi șirul $\{x^{(k)}\}$ utilizând relația recurentă:

$$Sx^{(k+1)} = Tx^{(k)} + b, \quad k = 0, 1, 2, \dots, K, \quad (3.16)$$

sau

$$x^{(k+1)} = Qx^{(k)} + d, \quad k = 0, 1, 2, \dots, K, \quad (3.17)$$

unde $x^{(0)} \in R^n$ este o aproximație inițială a soluției x^* .

Pentru a reduce sistemul (3.13) la forma (3.14) sau (3.16), potrivită pentru iterație, desfacerea matricei A trebuie să satisfacă condițiile:

- a) Sistemul (3.16) are o soluție unică $x^{(k+1)}$ și se rezolvă ușor. De aceea matricea S se alege de o formă simplă și este inversabilă. Ea poate fi diagonală sau triunghiulară.
- b) Șirul $\{x^{(k)}\}_{k=0}^{\infty}$ converge către soluția exactă x^* oricare ar fi $x^{(0)} \in R^n$.

Deoarece $x^* = Qx^* + d$ avem:

$$x^{(k+1)} - x^* = Q(x^{(k)} - x^*), \quad k = 0, 1, 2, \dots, K,$$

de unde rezultă că:

$$x^{(k+1)} - x^* = Q^k(x^{(0)} - x^*).$$

Evident că $\lim_{k \rightarrow \infty} x^{(k)} = x^*$ dacă și numai dacă

$$\lim_{k \rightarrow \infty} Q^k = O.$$

În cursul algebrei se demonstrează că $Q^k \rightarrow O$, dacă și numai dacă raza spectrală a lui Q este mai mică ca unitatea: $\rho(Q) = \max_{\lambda \in \sigma(Q)} |\lambda| < 1$. Viteza de convergență a șirului matriceal $Q, Q^2, Q^3, \dots, Q^k, \dots$ către matricea nulă O , și deci a șirului

$\{x^{(k)}\}_{k=0}^{\infty}$ către x^* , este cu atât mai mare cu cât raza spectrală $\rho(Q)$ este mai mică. Rezultă că este adevărată următoarea teoremă.

Teorema 1. (Condiția necesară și suficientă de convergență). Șirul $\{x^{(k)}\}$ definit prin (5) converge către soluția unică x^* a sistemului (3) pentru orice aproximație inițială $x^{(0)} \in R^n$ dacă și numai dacă

$$\rho(Q) = \max_{\lambda \in \sigma(Q)} |\lambda| < 1. \quad (3.18)$$

În practică nu cunoaștem valorile proprii ale lui Q , de aceea teorema 1 este dificil de folosit. În locul teoremei 1 se utilizează următoarea teoremă.

Teorema 2. (Condiția suficientă de convergență). Dacă există o normă matriceală subordonată unei norme vectoriale astfel încât $\|Q\| \leq q < 1$, atunci sistemul (3.15) are o soluție unică x^* , șirul $\{x^{(k)}\}$ definit prin (3.17) converge către x^* oricare ar fi aproximația inițială $x^{(0)} \in R^n$ și eroarea se evaluează prin:

$$\|x^{(k)} - x^*\| \leq \frac{q}{1-q} \|x^{(k)} - x^{(k-1)}\| \leq \frac{q^k}{1-q} \|x^{(1)} - x^{(0)}\|.$$

Demonstrație. Condiția $\|Q\| < 1$ implică condiția (3.18) (vezi paragraful 3.2). Pentru a obține o estimare a erorii folosim relația:

$$x^* - x^{(k-1)} = x^{(k)} - x^{(k-1)} + Q(x^* - x^{(k-1)}).$$

Deci

$$\|x^* - x^{(k-1)}\| \leq \|x^{(k)} - x^{(k-1)}\| + \|Q\| \|x^* - x^{(k-1)}\|,$$

sau

$$(1 - \|Q\|) \|x^* - x^{(k-1)}\| \leq \|x^{(k)} - x^{(k-1)}\|.$$

Deoarece prin ipoteză $1 - \|Q\| \geq 1 - q > 0$ avem:

$$\|x^* - x^{(k-1)}\| \leq \frac{1}{1-q} \|x^{(k)} - x^{(k-1)}\|.$$

Pe de altă parte

$$\|x^* - x^{(k)}\| = \|Q(x^* - x^{(k-1)})\| \leq \|Q\| \|x^* - x^{(k-1)}\|.$$

Prin urmare

$$\|x^* - x^{(k)}\| \leq \frac{q}{1-q} \|x^{(k)} - x^{(k-1)}\|.$$

Teorema este demonstrată.

Presupunem că elementele diagonale $a_{ii} \neq 0, i=1,2,\dots,n$. Atunci în calitate de matrice S se poate lua matricea diagonală atașată matricei A :

$$S = \text{Diag}(a_{11}, a_{22}, \dots, a_{nn}).$$

Avem

$$S^{-1} = \text{Diag}\left(\frac{1}{a_{11}}, \frac{1}{a_{22}}, \dots, \frac{1}{a_{nn}}\right)$$

În acest caz sistemul (3.15) devine:

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} x_j \right), \quad i = 1, 2, \dots, n.$$

Procesul iterativ (3.17) este definit prin:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} x_j^{(k)} \right), \quad i = 1, 2, \dots, n. \quad (3.19)$$

Astfel obținem o metodă de rezolvare a sistemului liniar (3.13) numită *metoda lui Jacobi*.

Exemplu. Fie dat sistemul de ecuații liniare:

$$\left. \begin{aligned} 2x_1 - x_2 &= 1, \\ -x_1 + 2x_2 &= 1 \end{aligned} \right\} \quad (3.20)$$

Vom avea:

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad S = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

$$Q = S^{-1}T = \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & 0 \end{pmatrix}, \quad d = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Valorile proprii ale matricei Q sunt: $\lambda_1 = -\frac{1}{2}$, $\lambda_2 = \frac{1}{2}$. Deci metoda lui Jacobi:

$$\left. \begin{aligned} x_1^{(k+1)} &= \frac{1}{2}x_2^{(k)} + \frac{1}{2}, \\ x_2^{(k+1)} &= \frac{1}{2}x_1^{(k)} + \frac{1}{2} \end{aligned} \right\}$$

converge către soluția exactă $x^* = (1, 1)^T$ oricare ar fi $x^{(0)} \in R^2$. În particular, pentru $x^{(0)} = (0, 0)^T$ obținem șirul:

$$x^{(1)} = \left(\frac{1}{2}, \frac{1}{2}\right)^T, \quad x^{(2)} = \left(\frac{3}{4}, \frac{3}{4}\right)^T, \quad x^{(3)} = \left(\frac{7}{8}, \frac{7}{8}\right)^T, \quad x^{(4)} = \left(\frac{15}{16}, \frac{15}{16}\right)^T, \dots$$

Observăm că pentru metoda lui Jacobi matricea Q are elementele

$$q_{ij} = \begin{cases} 0, & \text{dacă } i = j \\ -\frac{a_{ij}}{a_{ii}}, & \text{dacă } i \neq j \end{cases}$$

Utilizând teorema 2 cu norma $\|\bullet\|_\infty$ obținem:

$$\|Q\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |q_{ij}| = \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^n \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

Rezultă că dacă

$$|a_{ij}| > \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|, \quad i = 1, 2, \dots, n,$$

adică dacă matricea A este *diagonal dominantă*, atunci metoda lui Jacobi este convergentă.

În metoda lui Jacobi este necesar de a păstra în memoria calculatorului toate componentele vectorului $x^{(k)}$ atâta timp cât se calculează vectorul $x^{(k+1)}$. Putem modifica metoda lui Jacobi astfel încât la pasul $(k+1)$ să folosim în calculul componentei $x_i^{(k+1)}$, valorile deja calculate la același pas: $x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_{i-1}^{(k+1)}$. Această modificare a metodei lui Jacob se numește *metoda lui Gauss-Seidel* iar șirul iterativ (3.19) devine:

$$x_i^{(h+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(h+1)} - \sum_{j=i}^n a_{ij} x_j^{(h)} \right), \quad i = 1, 2, \dots, n.$$

Se observă ușor că metodei lui Gauss-Seidel corespunde desfacerea $A=S-T$ unde S este matricea subdiagonală atașată lui A , iar T este matricea strict supradiagonală cu elementele $-a_{ij}$ atașată la aceeași matrice A :

$$S = \begin{pmatrix} a_{11} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & 0 & \dots & 0 \\ a_{31} & a_{32} & a_{33} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix}, \quad T = \begin{pmatrix} 0 & -a_{12} & -a_{13} & \dots & -a_{1n} \\ 0 & 0 & -a_{23} & \dots & -a_{2n} \\ 0 & 0 & 0 & \dots & -a_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}.$$

Să reluăm exemplul de mai sus cu sistemul de ecuații (3.20). Pentru metoda Gauss-Seidel avem:

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad S = \begin{pmatrix} 2 & 0 \\ -1 & 2 \end{pmatrix}, \quad T = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

$$Q = S^{-1}T = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & \frac{1}{4} \end{pmatrix}, \quad d = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Șirul Gauss-Seidel arată astfel:

$$\left. \begin{aligned} x_1^{(k+1)} &= \frac{1}{2}x_2^{(k)} + \frac{1}{2}, \\ x_2^{(k+1)} &= \frac{1}{2}x_1^{(k+1)} + \frac{1}{2} \end{aligned} \right\}$$

ori:

$$\begin{pmatrix} 2 & 0 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Pentru aproximația inițială $x^{(0)} = (0, 0)^T$ obținem:

$$x^{(1)} = \left(\frac{1}{2}, \frac{1}{2}\right)^T, \quad x^{(2)} = \left(\frac{3}{4}, \frac{7}{8}\right)^T, \quad x^{(3)} = \left(\frac{15}{16}, \frac{31}{32}\right)^T, \text{K.}$$

Se observă că o iterație Gauss – Seidel aici este echivalentă cu două iterații Jacobi, deoarece valorile proprii ale matricei Q sunt 0 și $\frac{1}{4}$, deci raza spectrală este $\rho(S^{-1}T) = \frac{1}{4}$. Aceasta înseamnă că eroarea la fiecare iterație se împarte la 4; în metoda lui Jacobi $\rho(S^{-1}T) = \frac{1}{2}$ și eroarea se împarte la 2.

Subliniem că metoda lui Gauss – Seidel, necesitând același număr de operații aritmetice, este în caz general mai bună ca metoda lui Jacobi. Se poate arăta că dacă A este o matrice pozitiv definită atunci metoda Gauss – Seidel converge de două ori mai repede către soluție decât metoda lui Jacobi.

Se demonstrează și aici (vezi, de exemplu, [15], pag.181) că dacă matricea A este diagonal dominantă atunci metoda Gauss – Seidel converge. Se cunosc modele în care metoda lui Gauss – Seidel converge iar metoda lui Jacobi nu converge și invers. Pentru ilustrare anunțăm următoarea teoremă.

Teorema (Reich). Dacă matricea A este simetrică și are elementele diagonale $a_{ii} > 0$ pentru orice i atunci metoda Gauss – Seidel converge dacă și numai dacă A este o matrice pozitiv definită.

Este cunoscut că metoda lui Jacobi nu întotdeauna converge pentru matricea A pozitiv definită. De exemplu dacă A este o matrice pozitiv definită și nu este diagonal dominantă atunci este posibil ca raza spectrală $\rho(S^{-1}T) > 1$ pentru metoda lui Jacobi.

După cum sa văzut mai sus, dacă matricea A este diagonal dominantă, atunci metoda lui Jacobi și metoda lui Gauss – Seidel vor genera un șir de aproximații succesive care converge către soluția exactă oricare ar fi aproximația inițială $x^{(0)}$. Subliniem că această condiție este numai suficientă și nu necesară. De exemplu pentru matricea

$$A = \begin{pmatrix} 8 & 2 & 1 \\ 10 & 4 & 1 \\ 50 & 25 & 2 \end{pmatrix}$$

care nu este diagonal dominantă metoda lui Gauss – Seidel converge foarte rapid.

Metoda lui Gauss – Seidel poate fi modificată pentru îmbunătățirea vitezei de convergență a șirului aproximațiilor succesive. Fie $\tilde{x}^{(k)}$ vectorul obținut la pasul $k+1$ prin metoda Gauss – Seidel. Metoda iterativă definită prin:

$$x^{(k+1)} = x^{(k)} + \omega(\tilde{x}^{(k)} - x^{(k)})$$

este cunoscută cu numele de *metoda suprarelaxărilor succesive*. Parametrul de relaxare ω se alege astfel încât să crească viteza de convergență. Pentru $\omega = 1$ metoda se reduce la metoda lui Gauss – Seidel.

S-a găsit că pentru o alegere potrivită a parametrului ω convergența metodei suprarelaxării succesive este net superioară metodelor Jacobi și Gauss – Seidel. De exemplu, în cazul sistemului de ecuații liniare (3.20) o iterație prin metoda suprarelaxării succesive este echivalentă (vezi [2,35]) cu 30 de iterații prin metoda lui Jacobi.

Se poate arăta că $\omega \in (0, 2)$, de regulă în practică $\omega \approx 1.8 \div 1.9$.

Se demonstrează [39] că metoda suprarelaxărilor succesive converge pentru toate matricile A simetrice pozitiv definite. Pentru o înțelegere mai profundă a metodelor iterative recomandăm referințele [35,39]. În lucrarea [39] este adus și un subprogram – Fortran a metodei suprarelaxării succesive.

3.9 Sisteme liniare supradeterminate și metoda celor mai mici pătrate

3.9.1. Formularea problemei

Fie un sistem de m ecuații cu n necunoscute

$$Ax = b \quad (3.21)$$

unde A este o matrice de dimensiune $m \times n$ iar $b \in R^m$ este un vector cu m componente. Dacă $m > n$ sistemul (3.21) se numește sistem supradeterminat. Deoarece sistemul (3.21) conține mai multe ecuații decât necunoscute, nu putem găsi, în caz general, o soluție care să verifice exact toate ecuațiile sistemului. De exemplu sistemul:

$$2x_1 = b_1,$$

$$3x_1 = b_2,$$

$$4x_1 = b_3,$$

va avea soluție numai în cazul când termenii liberi b_1, b_2 și b_3 se află în raportul 2:3:4.

Deși sistemele supradeterminate în majoritatea lor nu sunt compatibile, ele se întâlnesc des în practică, de exemplu în probleme de statistică. Una din căile de rezolvare sistemelor supradeterminate constă în a determina pseudosoluția x^* care minimizează eroarea medie pentru toate cele m ecuații ale sistemului.

O pseudosoluție în sensul celor mai mici pătrate (CMMP) pentru sistemul supradeterminat (3.21) este un vector $x^* \in R^n$ cu proprietatea:

$$\|Ax^* - b\|_2^2 = \min_{x \in R^n} \|Ax - b\|_2^2 \quad (3.22)$$

Vectorul x^* se mai numește *soluție generalizată* în *sensul CMMP*. Acest vector x^* minimizează norma euclidiană a vectorului rezidual $r = Ax - b$, adică minimizează abaterea pătratică a lui Ax față de b :

$$\|r\|_2^2 = (r, r) = \sum_{i=1}^n r_i^2$$

De aici și denumirea de CMMP. În exemplul de mai sus

$$\|r\|_2^2 = (2x_1 - b_1)^2 + (3x_1 - b_2)^2 + (4x_1 - b_3)^2.$$

Se pune problema determinării unui vector $x^* \in R^n$ care să realizeze minimul expresiei:

$$E(x) = \|Ax - b\|_2^2 = \sum_{i=1}^m \left(\sum_{j=1}^n a_{ij} x_j - b_i \right)^2 \quad (3.23)$$

Această problemă revine la determinarea lui x^* cu proprietatea (3.22).

3.9.2. Metode bazate pe sisteme normale

Valoarea minimă a sumei (3.23) se obține anulând derivatele parțiale în raport cu x_1, x_2, \dots, x_n , adică anulând gradientul funcției $E(x)$:

$$\nabla E(x) = 2A^T(Ax - b) = 0. \quad (3.24)$$

Din ecuația (3.24) rezultă că orice pseudosoluție x^* în sensul CMMP a sistemului (3.21) satisface relația:

$$A^T(Ax^* - b) = 0,$$

sau

$$A^T Ax^* = A^T b \quad (3.25)$$

Sistemul de ecuații (3.25) se numește *sistem normal* asociat problemei (3.21). În acest sistem, $C = A^T A$ este o matrice de dimensiune $n \times n$, simetrică cu elementele:

$$c_{ij} = \bar{a}_i^T \bar{a}_j = (\bar{a}_i, \bar{a}_j),$$

unde $\bar{a}_i = (a_{1i}, a_{2i}, \dots, a_{mi})^T$ sunt vectorii coloană ai matricei A , $i=1,2,\dots,K, n$. Evident, matricea $C = A^T A$ este pozitiv semidefinită deoarece

$$(Cx, x) = x^T Cx = x^T A^T Ax = (Ax)^T Ax = \|Ax\|_2^2 \geq 0.$$

Dacă coloanele $\bar{a}_1, \bar{a}_2, \dots, \bar{a}_n$ ale matricei A sunt liniar independente atunci din $x \neq 0$ rezultă că și $Ax \neq 0$, deci matricea $C = A^T A$ este pozitiv definită. Prin urmare este adevărată următoare teoremă:

Teoreme de existență și unicitate. Dacă matricea A de dimensiune $m \times n$ are coloanele liniar independente atunci oricare ar fi vectorul $b \in R^m$ sistemul (3.21) are o pseudosoluție în sensul CMMP unică $x^* \in R^n$ și

$$x^* = (A^T A)^{-1} A^T b \quad (3.2)$$

Exemplu. Considerăm sistemul supradeterminat:

$$\left. \begin{aligned} 2x_1 - x_2 &= 9, \\ x_1 + 4x_2 &= 0, \\ 3x_1 + x_2 &= -3. \end{aligned} \right\}$$

Avem

$$A = \begin{pmatrix} 2 & -1 \\ 1 & 4 \\ 3 & 1 \end{pmatrix}, \quad A^T = \begin{pmatrix} 2 & 1 & 3 \\ -1 & 4 & 1 \end{pmatrix},$$

$$A^T A = \begin{pmatrix} 14 & 5 \\ 5 & 18 \end{pmatrix}, \quad A^T b = \begin{pmatrix} 9 \\ -12 \end{pmatrix}.$$

Sistemul normal asociat problemei propuse devine:

$$\left. \begin{aligned} 14x_1 + 5x_2 &= 9, \\ 5x_1 + 18x_2 &= -12. \end{aligned} \right\}$$

Sistemul normal permite determinarea pseudosoluției prin metodele prezentate în paragrafele 3.3 - 3.6, 3.8. Deoarece

matricea $A^T A$ este simetrică și pozitiv definită putem folosi factorizarea Cholesky. Pentru calculul lui $A^T A$ și $A^T b$ sunt necesare $m \cdot n(n+3)/2$ operații aritmetice, iar metoda lui Cholesky de rezolvare a sistemelor cere aproximativ $n^3/3$ operații. Astfel cea mai mare parte a efortului este cerut de formarea sistemului normal asociat problemei (3.21).

Numărul de condiționare a matricei $A^T A$ este egal cu pătratul numărului de condiționare a matricei A :

$$\text{cond}(A^T A) = [\text{cond}(A)]^2.$$

Rezultă că în general matricea $A^T A$ este prost condiționată și calculul său, deci, este afectat de erori de rotungere cu efect deseori catastrofal. De aceea în practică se evită formarea sistemelor normale și rezolvarea lor. Există metode mult mai bune de rezolvare în sensul CMMP n sistemelor supradeterminate. Ele se bazează pe factorizarea ortogonală a matricei A .

3.9.3. Metode de ortogonalizare

În cazul în care coloanele $\bar{a}_i^T, i=1,2,K,n$, ale matricei A sunt ortogonale putem ușor determina pseudosoluția sistemului supradeterminat (3.21). Într-adevăr, dacă $\bar{a}_i^T \bar{a}_j = 0, i \neq j$, matricea $A^T A$ devine o matrice diagonală cu elementele de pe diagonală egale cu $\bar{a}_i^T \bar{a}_i \neq 0$ și imediat se obține pseudosoluția:

$$x_i^* = \frac{b^T \bar{a}_i}{\bar{a}_i^T \bar{a}_i}, i=1,2,\dots,n.$$

Prin urmare, în locul formării sistemului normal putem ortogonaliza coloanele matricei A . Un procedeu clasic de ortogonalizare este *metoda lui Gram – Schmidt*. Șirul de vectori liniari independenți a_1, a_2, K, a_n se ortogonalizează după formulele:

$$v_1 = a_1, \quad v_i = a_i - \sum_{j=1}^{i-1} \frac{(a_i, v_j)v_j}{(v_j, v_j)}, \quad i = 2, 3, \dots, n. \quad (3.27)$$

Se constată ușor că vectorii v_1, v_2, \dots, v_n sunt ortogonali. Împărțind fiecare vector la lungimea lui, obținem un șir de vectori ortonormați: $q_1 = \frac{v_1}{\|v_1\|_2}, \quad q_2 = \frac{v_2}{\|v_2\|_2}, \dots, q_n = \frac{v_n}{\|v_n\|_2}$.

Exemplu. Fie vectorii:

$$a_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad a_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}.$$

Atunci $v_1 = a_1$, iar v_2 se calculează conform (3.27):

$$v_2 = a_2 - \frac{a_2^T v_1}{v_1^T v_1} v_1 = a_2 - \frac{1}{2} v_1 = \begin{pmatrix} 1/2 \\ -1/2 \\ 1 \end{pmatrix}$$

Vectorii ortonormați sunt:

$$q_1 = \frac{v_1}{\|v_1\|_2} = \sqrt{\frac{1}{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad q_2 = \frac{v_2}{\|v_2\|_2} = \sqrt{\frac{2}{3}} \begin{pmatrix} 1/2 \\ -1/2 \\ 1 \end{pmatrix}.$$

Pentru metoda eliminării lui Gauss o formă comodă de scriere al rezultatului constă în factorizarea matricei $A = LU$. Procesul de ortonormare Gram – Schimidt dă o altă factorizare pentru matricea A , numită factorizarea QR . În cazul exemplului de mai sus putem scrie:

$$a_1 = v_1, \quad \text{sau} \quad a_1 = \sqrt{2}q_1, \\ a_2 = -\frac{1}{2}v_1 + v_2, \quad \text{sau} \quad a_2 = \sqrt{\frac{1}{2}}q_1 + \sqrt{\frac{3}{2}}q_2.$$

Reprezentarea matriceală a acestor două ecuații este:

$$(a_1 \quad a_2) = (q_1 \quad q_2) \begin{pmatrix} \sqrt{2} & \sqrt{1/2} \\ 0 & \sqrt{3/2} \end{pmatrix},$$

adică $A=QR$ unde Q este cu coloane ortogonale iar R este superior triunghiulară.

Se poate arăta că orice matrice de dimensiune $m \times n, m \geq n$, cu coloanele liniar independente admite o factorizare QR , unde Q este o matrice (de dimensiune $m \times n$) cu coloanele ortonormate iar R este o matrice (pătrată de dimensiune $n \times n$) superior triunghiulară.

Dacă se cunoaște factorizarea QR a matricei A atunci CMMP se rezolvă ușor. Din (3.26) obținem:

$$x^* = (A^T A)^{-1} A^T = (R^T Q^T Q R)^{-1} R^T Q^T b,$$

de unde, ținând seama că $Q^T Q = I$, rezultă că:

$$x^* = (R^T R)^{-1} R^T Q^T b = R^{-1} Q^T b.$$

Prin urmare pseudosoluția în sensul CMMP se poate obține ușor rezolvând sistemul triunghiular:

$$Rx = Q^T b. \tag{3.28}$$

Pentru calculul lui R și $Q^T b$ sunt necesare aproximativ $n^2 m$ operații, iar pentru rezolvarea sistemului triunghiular (3.28) numai $n(n+1)/2$ operații. Deci numărul total de operații este aproximativ de două ori mai mare decât cel pentru formarea sistemului de ecuații normale.

Există o variantă nouă a metodei Gram – Schmidt numită *algoritmul lui Gram – Schmidt modificat*:

Pentru $k = 1, 2, K, n$

$$a_k = \frac{a_k}{\|a_k\|_2}$$

Pentru $j = k + 1, k + 2, K, n$

$$a_j = a_j - (a_j^T a_k) a_k$$

Algoritmul lui Gram – Schmidt modificat este numeric stabil datorită rearanjării ordinii de efectuare a calculelor. În plus, el necesită mai puțină memorie operativă decât metoda clasică de ortogonalizare. Vectorii q_k se calculează și se plasează în același loc de memorie care îl ocupă vectorii inițiali a_k .

Pentru completarea cunoștințelor cu alte metode de rezolvare a problemei CMMP se recomandă [2,22,29,34,35,37]. Menționăm lucrarea [29] care conține un program bun FORTAN bazat pe factorizarea valorilor singulare, numită DVS. Cititorul care stăpânește bine noțiunile de bază din algebra liniară recomandăm lucrarea fundamentală [29].

3.10 Exerciții

1. Fie matricea

$$A = \begin{pmatrix} 0 & 0 & 6 \\ 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \end{pmatrix}.$$

Să se calculeze A^3 .

2. Să se determine matricele pătrate de ordinul doi, satisfăcând relațiile:

a) $A^2 = 0$ deși $A \neq 0$

b) $B^2 = -I$

c) $CD = -DC$, dar $CD \neq 0$

d) $EF = 0$ cu toate că elementele matricelor E și F sunt nenule.

3. Să se arate că

$$(AB)^T = B^T A^T; (AB)^{-1} = B^{-1} A^{-1}; (A^{-1})^T = (A^T)^{-1}.$$

4. Matricea P se numește *idempotentă* dacă $P^2 = P$. Să se determine toate matricele idempotente de ordinul doi.

5. Să se calculeze $\|x\|_1, \|x\|_2, \|x\|_\infty$ pentru $x = (0, -1, -2, 0, 1)^T$.

6. Să se arate că oricare ar fi norma vectorială au loc inegalitățile:

$$\text{a) } \left| \|x\| - \|y\| \right| \leq \|x - y\|$$

$$\text{b) } \|x \pm y\| \leq \|x\| + \|y\|.$$

7. Să se determine unghiul dintre vectorii $x = (2, -2, 1)^T$ și $y = (1, 2, 2)^T$.

8. Să se arate că matricele A și B sunt egale între ele dacă și numai dacă $Ax = Bx$ pentru $\forall x$.

9. Fie $A = I - 2xx^T$ unde $x \in \mathbb{R}^n, (x, x) = 1$. Să se arate că matricea A este ortogonală și $A^2 = I$.

10. Fie A o matrice ortogonală și $Ax = \lambda x, x \neq 0$. Să se arate că $|\lambda| = 1$.

11. Fie matricea de rangul întâi

$$A = \begin{pmatrix} 1 & 3 \\ 3 & 9 \end{pmatrix}.$$

Să se pună matricea A de forma uv^T .

12. Fie matricea

$$A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}.$$

Să se arate că matricea A este ortogonală.

13. Să se arate că dacă A este o matrice pozitiv definită atunci matricele A^2 și A^{-1} sunt de asemenea pozitiv definite.

14. Să se calculeze factorizarea LU a matricei

$$A = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}.$$

15. Să se calculeze factorizarea Cholesky LL^T a matricei

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

16. Să se arate că $\text{cond}(AB) \leq \text{cond}(A) \times \text{cond}(B)$.

17. Să se arate că sistemul de ecuații

$$\begin{cases} x_1 + 100x_2 = 100 \\ x_2 = 0 \end{cases}$$

Este rău condiționat. Să se calculeze numărul de condiționare.

18. Fie $x = (3 \ 4)^T, z = (1 \ 1)^T$. Să se calculeze $\sigma = \|x\|_2$, $v = z + \sigma x$ și matricea corespunzătoare Householder.

19. Să se compare metodele Jacobi, Gauss – Seidel și a suprarelaxării succesive în cazul matricei A din exercițiul 15, iar

$$b = (1,0,1)^T \text{ și } x^{(0)} = (0,0,0)^T.$$

20. Fie

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 0.01 \\ 0.1 \end{pmatrix}, \quad \delta b = \begin{pmatrix} 0.0001 \\ 0 \end{pmatrix}.$$

Să se calculeze pseudosoluția în sensul celor mai mici pătrate a sistemelor supradeterminate $Ax = b, Ax = b + \delta b$. Să se compare rezultatul.

21. Să se calculeze valorile proprii și vectorii proprii ai matricei

$$A = \begin{pmatrix} 5 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix}.$$

22. Să se calculeze valorile proprii ai matricei din exercițiul nr. 15.

23. Fie sistemul supradeterminat $Ax = b, A = (a_{ij})_{m \times n}$, $b \in R^m$, cu pseudosoluția $x^* = (A^T A)^{-1} A^T b$. Să se arate că vectorul rezidual $r = Ax^* - b$ este ortogonal pe subspațiul

$$\text{Im } A = \{y | y = Ax, x \in R^n\} \subset R^m.$$

24. Matricea $A^+ = (A^T A)^{-1} A^T$ de dimensiune $n \times m$ se numește *pseudoinversa* lui A sau *generalizarea matricei inverse a lui More – Penrose*. Să se arate că

$$AA^+A = A; \quad A^+AA^+ = A^+;$$

$$(AA^+)^T = AA^+; \quad (A^+A)^T = A^+A.$$

25. $P_A = AA^+$ este *proiectorul ortogonal* al lui A pe spațiul $\text{Im } A$. Să se arate că $P_A^2 = P_A$ și $P_A^T = P_A$, adică matricea P_A este idempotentă și simetrică.