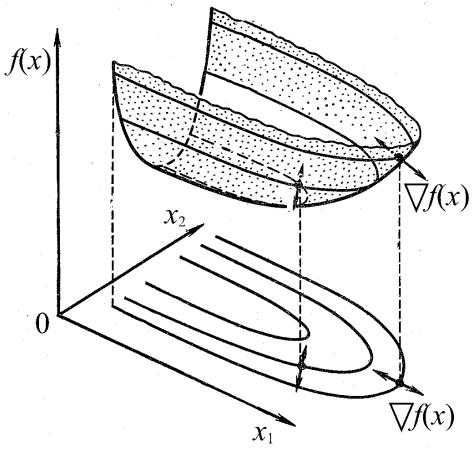
UNIVERSITATEA TEHNICĂ A MOLDOVEI

Vasile Moraru

M E T O D E ŞI M O D E L E DE C A L C U L

Modulul I Metode numerice



Chişinău
2021

UNIVERSITATEA TEHNICĂ A MOLDOVEI

Facultatea Calculatoare, Informatică şi Microelectronică
Departamentul Informatică şi Ingineria Sistemelor

Prof. univ., dr. Vasile MORARU

M E T O D E ŞI M O D E L E DE C A L C U L

Modulul I Metode numerice

Chişinău
U.T.M.
2021

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 studenţilor Universităţii Tehnice a Moldovei şi va constitui un sprijin efectiv la predarea cursurilor de *Metode numerice*, *Cercetări operaţionale*, *Matematică de calcul* ş.a. Cartea însă poate fi folosită şi de toţi cei care sunt preocupaţi de utilizarea metodelor numerice şi a mijloacelor electronice de calcul la soluţionarea problemelor practice.

Autor: prof. univ., dr. Vasile Moraru

PREFAŢĂ

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

Lucrarea este structurată astfel: o introducere, cinci capitole, o anexă şi îşi propune să prezinte principalele metode numerice de rezolvare a ecuaţiilor algebrice şi transcendente,

Un loc important îl ocupă rezolvarea numerică a ecuaţiilor şi a sistemelor de ecuaţii neliniare, a sistemelor de ecuaţii liniare şi neliniare, a problemelor de optimizare necondiţionată şi condiţionată. Se propun spre soluţionare exerciţii, rezolvate sau însoţite de indicaţiile corespunzătoare.

Pentru înţelegerea materialului de bază sunt suficiente cunoştinţele pe care studenţii le-au obţinut la cursurile de matematici şi programare.

În *Noţiuni introductive* sunt expuse primele concepţii despre metodele numerice şi despre algoritmii de calcul, se fac estimări privind viteza de convergenţă a acestora.

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

În capitolul al doilea se prezintă metode numerice de calcul al rădăcinilor ecuaţiilor algebrice şi transcendente.

Este vorba de metoda înjumătăţirii intervalului, metoda aproximaţiilor succesive, metoda tangentei (Newton), metoda secantei şi alte metode numerice. Un loc aparte îl ocupă ecuaţiile algebrice, istoria cărora începe cu Evul Mediu şi cu Renaşterea. Se expun schema lui Horner şi metoda Newton de determinare a tuturor rădăcinilor reale ale ecuaţiilor algebrice.

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

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

În *capitolul al cincilea* se prezintă principalele metode numerice de rezolvare a sistemelor de ecuaţii neliniare şi a problemelor de optimizare necondiţionată. Problema de optimizare se poate reduce la rezolvarea unui sistem de ecuaţii şi invers. De aceea s-a considerat necesar a expune în cadrul aceluiaşi capitol metodele menţionate. Se

tratează metoda iteraţiei, metoda Gauss-Seidel neliniară, metoda Newton care sunt o extindere a metodelor studiate în capitolele doi şi trei. Se aduc condiţiile necesare şi suficiente de extrem în optimizarea necondiţionată. Se face o trecere în revistă a metodei gradientului, metodei Newton-Raphson, metodelor cvasi-Newton şi a metodelor de direcţii conjugate, punându-se în evidenţă drept cea mai eficace metoda Fletcher-Powell în versiunea Polak-Ribiere.

Bibliografia include referinţele în care cititorul poate afla detalii suplimentare asupra metodelor numerice de calcul prezentate în lucrare. Unele date bibliografice au fost incluse chiar în text. Menţionăm că din mulţimea cărţilor 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 NOTAŢII

R	mulţimea numerelor reale
R^n	spaţiu liniar n-dimensional (3.1.1)
$\{x^{(k)}\}$	şir de vectori; $x^{(k)} \in R^n$
x^*	soluţia 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 $x,y \in R^n$
$\{x \mid P\}$	mulţimea elementelor x cu proprietatea P
$\nabla f(x)$	gradientul funcţiei $f(x)$.
$\nabla^2 f(x)$	matricea Hesse a funcţiei $f(x)$
$o(x)$	$o(x)/x \to 0$ dacă $x \to 0$.
$\Im(x)$	matricea Jacobi (5.2)

NOŢIUNI INTRODUCTIVE

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

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

Rezolvarea unei probleme impuse de practică începe cu construcţia *modelului matematic*. Modelul matematic reprezintă efectiv formularea matematică a problemei enunţate, adică constituie exprimarea matematică a relaţiilor şi restricţiilor dintre parametrii problemei.

După formularea matematică a problemei se realizează *alegerea metodei numerice* şi se *elaborează algoritmul de calcul*. Aceste etape sunt cele mai importante în procesul soluţionării problemelor. La alegerea metodei numerice se ia în consideraţie viteza de convergenţă, precizia, stabilitatea, timpul de execuţie şi necesarul de memorie.

Algoritmul metodei numerice constă dint-un număr finit de operaţii aritmetice şi logice, care trebuie să le efectueze calculatorul electronic pentru rezolvarea problemei date. Regulile de calcul alcătuiesc *paşii algoritmului.*

Subliniem faptul că noţiunea de algoritm în forma sa generală se situează printre noţiunile fundamentale ale matematicii şi sta la baza programării calculatoarelor electronice. Orice algoritm este caracterizat prin următoarele proprietăţi:

- *Generalitate.* Prin aceasta se înţelege 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 obţine datele de ieşire, 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 şi 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 soluţii sunt cunoscute.

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

I. enunţarea problemei şi formularea matematică, evidenţiind ce se dă şi ce se cere;
II. alegerea unei metode numerice pentru obţinerea soluţiei;
III. elaborarea algoritmului de calcul;
IV. testarea şi verificarea programului;
V. analiza rezultatelor obţinute.

Majoritatea metodelor numerice de calcul reprezintă în esenţă proceduri iterative. Aceasta înseamnă că, plecând de la un x_0 dat, se construieşte un şir $x_0, x_1, \ldots, x_k, \ldots$ (notat de obicei prin $\{x_k\}$) care în anumite condiţii converge către soluţia exactă x_* a problemei considerate. Elementele x_k, $k = 0,1,2,\ldots$ pot fi atât numere reale cât şi vectori sau matrice.

În metodele numerice drept *criteriu de oprire (stopare)* al iteraţiilor, de obicei, se foloseşte următorul: şirul $\{x_n\}$ se

trunchiază la un indice *m* determinat pe parcursul calculului în funcţie de precizia dată astfel încât termenul curent x_m constituie o aproximaţie satisfătătoare a soluţiei căutate x_*. De aceea un fapt esenţial în compararea metodelor de calcul numeric îl constituie aprecierea vitezei de convergenţă a metodelor. În metodele de convergenţă rapidă sunt necesare un număr mai mic de iteraţii pentru a atinge precizia prescrisă decât în metodele cu convergenţă lentă. Dintre metodele care necesită acelaşi volum de calcul la fiecare iteraţie în practică se alege metoda cu convergenţa mai rapidă.

De obicei viteza de convegeţă a şirului $x_k \to x_*$ se stabileşte cu ajutorul unei funcţii $e(x)$, numită *funcţie de estimare a erorii*, care satisface condiţiilor:

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

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

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

în cazul $x, x_* \in R$, sau

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

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

Fie şirul $\{x_k\}$ convergegent catre x_*, adică

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

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

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

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 şi matricelor,

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

Dacă printr-o metodă numerică de calcul se obţine un şir care converge liniar atunci se spune că metoda este de *convergenţă liniară* sau de *ordinul întâi de convergenţă*. Se mai spune în acest caz că metoda *converge tot atât de repede ca şi progresia geometrică de raţie q* sau că *converge liniar cu raţia q*.

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

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

Cu alte cuvinte, şirul este convergent supraliniar dacă

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

sau, în cazul vectorilor şi matricelor,

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

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

Convergenţa supraliniară înseamnă că $e(x_k) \to 0$ mai repede decât orice progresie geometrică de raţie q, adică converge mai repede către zero decît orice şir de forma cq^k, unde $c > 0$ şi $0 < q < 1$.

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

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

atunci se spune că şirul $\{x_k\}$ are *ordinul* β *de convergenţă*. În cazul când $\beta = 2$ sau $\beta = 3$ se spune că şirul $\{x_k\}$ este de *ordinul doi*, corespunzător de *ordinul trei de convergenţă*. Se mai spune că metoda nuimerică cu ajutorul căreia se construieşte acest şir este cu *convergenţa pătratică*, respectiv cu *convergenţa cubică* sau că este o *metodă de ordinul al doilea*, respectiv *al treilea*.

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

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

sau, în cazul vectorilor şi matricelor,

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

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

În cele ced urmează vom ilustra prin exemple concrete definiţiile date.

Exemle de convergenţă liniară.

1) Şirurile

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

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

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

Într-adevăr,

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

şi

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

2) Şirurile

a) $x_0=c, x_1=cq, x_2=cq^2, x_3=cq^3, \ldots, x_k=cq^k, \ldots$

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

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

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

unde $c>0$ şi $0<q<1$, converg către $x_*=0$ liniar cu raţia q. Demonstraţia rezultă imediat din definiţia convergenţei liniare:

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

Exerciţii.

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

$$x_0=c+1, x_1=c+q_1, x_2=c+q_1q_2, x_3=c+q_1^2q_2,$$

$$x_4=c+q_1^2q_2^2, x_5=c+q_1^3q_2^2,\ldots, x_{2k}=c+q_1^kq_2^k, x_{2k+1}=c+q_1^{k+1}q_2^k,\ldots$$

unde $c>0$. Să se arate că acest şir converge liniar către $x_*=c$ cu raţia $q=\sqrt{q_1q_2}$.

2. Să se demonstreze că şirul $\{x_k\}$ converge liniar cu raţia $0<q<1$ dacă şi numai dacă $\lim\limits_{k\to\infty}\sqrt[k]{e(x_k)}=\lim\limits_{k\to\infty}\sqrt[k]{|x_k-x_*|}=q$.

Exemple de convergenţă supraliniară. Şirurile

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

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

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

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

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

Exerciţiu. Să se demonstreze că orice şir de forma

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

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

Exemple de convergenţă pătratică.

1. Şirul $x_0=cq, x_1=cq^2, x_2=cq^4, x_3=cq^8\ldots, x_k=cq^{2^k},\ldots,$

unde constsnta q satisface condiţiei $0 \le q < 1$ şi $c > 0$, este cu convergenţa pătratică.

Într-adevăr, avem $x_* = 0$ şi

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

2. Şirul

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

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

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

Exerciţiu. Să se demonstreze că şirul definit prin formula

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{1}{x_k}\right), k = o,1,2,\ldots,$$

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

Alte detalii şi aspecte de diferite nuanţe legate de viteza de convergenţă a şirurilor 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 Zork, 1970.* Lucrarea a apărut şi în limba rusă în anul 1975 la editura Mir.

NUMERE APROXIMATIVE

1.1 Erori absolute şi 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 înlocuieşte 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 obţin rezultatele. Astfel, de exemplu, dacă $x=1$ x=1 si x_*=2, atunci eroarea absolută $\Delta(x_*)$=1 indica o precizie slabă a măsurării. Dacă $x=10^{10}+1$ iar $x_*=10^{10}$, aceeaşi eroare absolută $\Delta(x_*)$=1 caracterizează o precizie remarcabilă. Aceasta ne conduce la noţiunea de *eroare relativă* $\delta(x_*)$ care reprezintă raportul dintre eroarea absolută şi 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 şi x_*, atunci calculul erorii absolute şi relative este imediat. Dar de obicei, în majoritatea cazurilor se cunoaşte numai aproximarea x_*. De aceea se introduce noţiunea de margine (sau limită) a erorii absolute şi relative. Numărul pozitiv ε este o *margine* (sau o *limita*) *a erorii*

absolute a numărului aproximativ x_* dacă

$$|x\text{-}x_*| \le \varepsilon$$

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

$$\frac{|x\text{-}x_*|}{|x_*|} \le r$$

Notaţia $x=x_* \pm \varepsilon$ semnifică întotdeauna faptul, că $|x\text{-}x_*| \le \varepsilon$, adică

$$x_* \text{-} \varepsilon \le x_* \le x + \varepsilon .$$

Orice număr aproximativ x_* poate fi scris sub forma

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

unde $c_1, c_2, \ldots, c_n$ sunt cifrele zecimale ale numărului aproximativ x_*. Se ştie că zerourile de la începutul numărului servesc numai pentru a fixa poziţia virgulei zecimale. Cifrele cuprinse intre prima si ultima cifră diferită de zero sau care indică ordinele păstrate in 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 nesemnificative).

Dacă mărimea erorii x_* nu depăşeşte $0.5\cdot 10^{-t}$ 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 sfârşitul numărului nu se aruncă.

De pildă numerele 0.0345 şi 0.034500 sunt diferite; eroarea absolută a primului număr nu depăşeşte 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 şi patru cifre semnificative; 0.001234 ± 0.000006 are patru zecimale

corecte şi două cifre semnificative (deoarece valoarea maximă a numărului poate fi 0.001240 iar minima 0.001228 şi 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 şi sursele erorilor

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

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

Atunci avem

$$x_* - \varepsilon_x + y_x - \varepsilon_y \le x + y \le x_* + \varepsilon_x + y_x + \varepsilon_y$$

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

sau

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

În mod analog ca în cazul adunării, se va obţine

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

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 înmulţire şi împărţire marginile erorilor relative ale factorilor se adună.*

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

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

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

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

Exemplu. Vom considera numerele aproximativ egale

$$x=0.1234\pm0.5\cdot10^{-4} \quad \text{şi} \quad y=0.1233\pm0.5\cdot10^{-4}$$

atunci $x-y = 0.0001\pm0.0001$ şi marginea erorii este tot atât de mare ca şi 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. Ecuaţia 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

$$x_1=\frac{-1000.01+\sqrt{(1000.01)^2+4\cdot2.5245315}}{2},$$

efectuând calculele cu opt cifre semnificative, se va obţine

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

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

Dacă vom face calculul aceleiaşi rădăcini utilizând formula

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

se va obţine rezultatul exact.

Precizia calculelor numerice este criteriul cel mai eficient pentru alegerea metodelor de calcul. Analiza erorii dintr-un rezultat numeric este o chestiune esenţială în orice calcul, fie că este executat manual , fie de un calculator. Cu toate performanţele calculatoarelor electronice, precizia rezultatelor este influenţată 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 iniţiale sau din soluţii aproximative ale altor probleme etc. Aceste tipuri de erori se numesc *erori inerente*. Ele nu pot fi influenţate de metoda de calcul.
2. *Erori de metodă sau de trunchiere.* Majoritatea metodelor numerice necesită un număr infinit de operaţii aritmetice pentru a ajunge la soluţia exactă a problemei. De aceea suntem nevoiţi să trunchiem metoda după un număr finit de operaţii. Ceea ce omitem constituie eroarea de trunchiere.
3. *Erori de rotunjire* în datele de intrare în calcule şi în datele de ieşire. 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 iraţional nu poate fi reprezentat printr-un număr finit de cifre. Chiar şi unele numere raţionale nu au o reprezentare exactă. Numărul 1/3 poate fi scris ca 0.3333..., o succesiune a cifrei *3* la partea zecimală.

Datorită proprietăţilor constructive ale calculatoarelor electronice este necesară limitarea numărului cifrelor semnificative. Un exemplu instructiv este furnizat de numărul raţional 1/10 care se foloseşte de multe ori ca dimensiune a «pasului» în multe algoritme . În sistemul binar (care foloseşte pentru reprezentarea numerelor cifrele 0 şi 1) fracţia 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 aproximaţie 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 fracţie, numită *mantisă* şi un întreg, numit *exponent*. Deci,

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

unde β este baza sistemului de numeraţie (binar, octal sau hexazecimal), m este mantisa numărului şi e este exponentul, afectat de semn. Fracţia m satisface

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

şi are forma

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

unde numerele întregi $d_1, d_2, \ldots, d_t$, numite cifre, verifică inegalităţile

$$0 \le d_i \le \beta - 1, i = 1,2,\ldots,t$$

şi $L \le e \le U$.

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

Prin urmare, *sistemul de calcul cu numere cu virgulă mobilă* este o mulţime *F (β, t, L, U)* caracterizată de patru parametri: baza *β*, precizia maşinii *t* şi intervalul exponenţilor *[L, U]*.

F este o mulţime finită care conţine

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

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

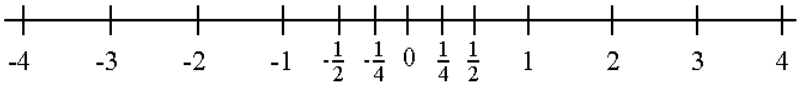


Fig.1.1. Sistemul de calcul *F (2,3,-1,-2).*

Mulţimea *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 depăşeşte 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

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

de unde rezultă, că dacă $d_1 \neq 0$, atunci maximul mărimii *m* este egal cu $1-\beta^{-t}$, ce corespunde cazului $m_i = \beta - 1$, *i*=1, 2, ... , *t*; valoarea minimală va fi β^{-1} şi se obţine pentru $d_1 = 1, d_2 = d_3 = \ldots = d_t = 0$.

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

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

Eroarea relativă în *fl(x)* este

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

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

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

apropiat element de *x*, care aparţine lui *F*. Dacă se foloseşte rotunjirea prin tăiere (se elimină compararea primei cifre neglijate), atunci $\varepsilon_M = \beta^{1-t}$ şi *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 răspândiţi încă doi parametri σ şi λ: cel mai mic element pozitiv şi elementul maximal al lui *F*. Pentru ilustrare vom analiza sistemul *F(*2, 3, -1, 1*)*. El conţine numărul zero şi toate numerele care au o exprimare binară de forma

$$x = \pm 0.1 d_2 d_3 \cdot 2^e,$$

unde $-1 \le e \le 1$ iar fiecare din cifrele d_2 şi d_3 este 0 sau 1 . prin urmare avem trei posibilităţi pentru valoarea exponentului *(*-1, 0 ori +1 *)* şi patru pentru reprezentarea părţilor fracţionare *(*0.100, 0.101, 0.110 şi 0.111*)*. Mulţimea *F* constă din $2 \cdot 1 \cdot 4 \cdot 3 + 1 = 25$ de numere cu virgulă mobilă. În sistemul zecimal de numeraţie fracţiile de mai sus pot fi scrise respectiv 1/2, 5/8, 3/4 şi 7/8 (de exemplu, *fracţia 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 mulţimii *F*.

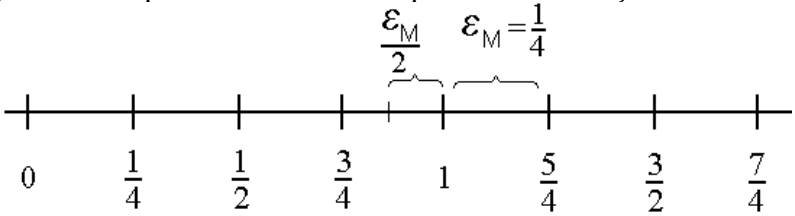


Fig.1.2. Numerele pozitive ale sistemului de calcul *F(2,3, -1, 1).*

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

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

Unitatea de rotunjire a maşinii ε_M se mai numeşte *epsilon al maşinii* şi 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 distanţa dintre numărul $x \in F$ şi numărul cel mai apropiat de el în sistemul dat nu e mai mică decât $\varepsilon_M|x|/\beta$ şi 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 distanţa dintre oricare două numere cu virgulă mobilă este mai mică decât σ. Prin urmare aceste numere nu pot fi obţinute unul din altul prin adunare. Ele pot fi căpătate decât în calculul expresiilor aritmetice.

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

1.4 Aritmetica virgulei mobile şi erorile de rotunjire

Pe mulţimea F, care o considerăm un model al mulţimii numerelor reale, se definesc operaţiile aritmetice în aşa fel încât ele să reproducă modul de efectuare al acestor operaţii de către calculatorul electronic.

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

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

Suma calculată nu aparţine sistemului de calcul F deoarece pentru reprezentarea părţilor fracţionare a ei sunt necesare patru cifre binare. Prin urmare, operaţia de adunare trebuie modelată ea însăşi

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

Dacă x şi y sunt numere cu virgulă mobilă iar numărul $x+y$ nu iese din limitele mulţimii 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 aştepta că $5/4 \oplus 3/8$ este egal cu 3/2 sau cu 7/4 (deoarece ambele elemente se află la distanţă egală de 5/4+3/8).

Diferenţa dintre $x \oplus y$ şi $x+y$ (pentru $x, y \in F$) reprezintă o eroare de rotunjire care se comite în cazul adunării cu virgula mobilă. Proprietăţi asemănătoare sunt adevărate şi pentru celelalte operaţii 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 aparţine lui F din cauza repartizării elementelor lui F. Pe de altă parte, suma 7/2+7/2 nu aparţine 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 apariţia unui *semnal de depăşire*, după care calculele se întrerup deoarece nu este posibil de dat o aproximaţie de sens pentru numerele care ies din limitele lui F.

Deşi majoritatea sumelor $x+y$ cu $x \in F$, $y \in F$, aparţin ele însele lui F, foarte rar produsul (exact) obişnuit $x \cdot y$ aparţine lui F, deoarece, de regulă, el are $2t$ sau $2t$-1 cifre semnificative . În afară de aceasta, depăşirea este mai probabilă la înmulţire. În fine, în cazul înmulţirii cu virgulă mobilă, este posibilă apariţia unui *zero-maşină*, atunci când pentru $x \neq 0$ şi $y \neq 0$, produsul $x \cdot y$ este nenul dar este mai mic, în modul, decât cel mai mic element pozitiv al lui F (apariţia unui zero-maşină este posibilă şi în cazul scăderii deşi aceasta se întâmplă foarte rar).

Operaţiile de adunare şi înmulţire 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 diferenţa $|fl(x)-x|$, unde x este un număr real, exponentul căruia aparţine 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 defineşte astfel

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

şi se îndeplineşte condiţia

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

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

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

În intervalul $\left[\beta^{e-1},\beta^{e}\right]$ 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 depăşeşte mărimea β^{e-t} iar în cazul rotunjirii corecte mărimea $\beta^{e-t/2}$, adică

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

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

$$\delta(x)\le\begin{cases}\dfrac{\beta^{e-t}}{\beta^{e-1}}=\beta^{1-t}, & \text{]n cazul rotungirii prin taier e}\\ \dfrac{\beta^{e-t}/2}{\beta^{e-1}}=\dfrac{1}{2}\beta^{1-t}, & \text{pentru rotungirea corecta .}\end{cases}$$

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 acelaşi 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 $fl(x)=0.1247\cdot 10^2$ şi

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

1.5 Determinarea parametrilor unui sistem de calcul

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

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

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

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

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

Operaţiile de înmulţire şi împărţire 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 înmulţesc (împart) mantisele numerelor *x* şi *y*, se normalizează rezultatul, oprindu-se *t* cifre la mantisă şi la partea exponenţială se adună (scad) exponenţii.

După cum vedem, deoarece calculul se face cu un număr anumit de cifre semnificative *t*, modul de rotunjire are o mare importanţă. 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 aparţine din nou lui *F* şi 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

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

şi 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 maşinilor de calcul se folosesc instrucţiuni speciale care determină modul de rotunjire. De aceea unul şi acelaşi program poate da rezultate diferite în dependenţă 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 obişnuită de rotunjire, care, după cum s-a mai spus, foloseşte mult timp de calculator dacă este aplicată la fiecare operaţie aritmetică.

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

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

Trebuie de menţionat importanţa ordinii efectuării operaţiilor aritmetice, deoarece într-o aritmetică a virgulei mobile nu au loc legile asociativă şi distributivă. Pentru unele sisteme de calcul

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

Într-adevăr, fie de exemplu t=40 şi β=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 operaţiile aritmetice +, -, *, / se realizează la calculatoarele numerice după cum urmează (prin x este notat rezultatul exact al operaţiei aritmetice):

1) dacă $\sigma \le |x| \le \lambda$, atunci rezultatul operaţiei se rotunjeşte;

2) dacă $|x| < \sigma$, atunci rezultatul se anulează, adică apare un zero-maşină;

3) dacă $|x| > \lambda$, atunci calculele se întrerup şi apare un semnal de depăşire.

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 operaţie

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

1.6.1. Calculul mediei aritmetice

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

$$c = \frac{a+b}{2} \tag{1.1}$$

$$a = a + \frac{b-a}{2} \tag{1.2}$$

Formula (1.1) necesită cu o operaţie de adunare mai puţin decât formula (1.2) dar din punct de vedere al exactităţii nu este întotdeauna cea mai bună. Într-adevăr, fie că calculele se efectuează într-o aritmetică zecimală *(β*=10*)* a virgulei mobile cu trei cifre semnificative *(t*=3*)* şi fie *a*=0.596 şi *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 menţionăm că în exemplul dat în care înclinăm pentru formula (1.2) numerele *a* şi *b* au acelaşi semn.

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

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

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

$$c = -3.483 + \frac{8.765+3.483}{2} = -3.483 + \frac{12.24}{2} =$$

$$= -3.483 + 6.120 = 2.637.$$

Chiar dacă s-ar efectua rotunjirea corectă rezultatul obţinut 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 şi 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 dependenţă de semnele lui a şi 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 :

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

Integrând prin părţi, obţinem|

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

sau

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

Pentru n=1 avem

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

Fie aritmetica virgulei mobile cu β=10 şi t=6. Utilizând formula de recurenţa $I_n = 1 - n \cdot I_{n-1}$, obţinem [29]:

$$I_1 \approx 0.367879, \quad I_6 \approx 0.127120,$$
$$I_2 \approx 0.264242, \quad I_7 \approx 0.110160,$$
$$I_3 \approx 0.207274, \quad I_8 \approx 0.118720,$$
$$I_4 \approx 0.170904, \quad I_9 \approx -0.068480,$$
$$I_5 \approx 0.145480,$$

Deşi $x^n e^x \geq 0, \forall x \in [0,1]$ se observă că pentru n=9 am obţinut I_n<0, evident contradictoriu. Apariţia 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 recurenţă că $E_2 = -2E_1$, $E_3 = -3E_2 = (-2)\cdot(-3)E_1 = 3!E_1$ ş.a.m.d. Calculând E_n pentru $n=9$ vom găsi $E_9 \approx 9!*4.412*10^{-7} \approx 0.1601$, 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 recurenţă sub forma

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

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

$$I_n \le \int_0^1 x^n dx = \frac{x^{n+1}}{n+1}\Big|_0^1 = \frac{1}{n+1},$$

şi deci

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

Alegând $I_{20}=0$, se admite o eroare ce nu depăşeşte 1/21. Atunci eroarea în I_{19} nu va întrece $(1/20)*(1/21) \approx 0.0024$. Această eroare se va micşora până la $4 \cdot 10^{-8}$ în timpul calcului I_{15} şi devine mai mică decât eroarea de rotunjire. Rezultatele obţinute prin recurenţă 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 numeşte *numeric stabil* dacă aplicat unei probleme cu date iniţiale «uşor perturbate» produce o soluţie care aproape coincide cu soluţia exactă (soluţia problemei cu datele iniţiale neperturbate).

1.6.3. Exemple de sisteme rău condiţionate

Fie sistemul liniar

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

Rezolvarea se poate face folosind algoritmul dat de regula lui Cramer

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

Soluţia exactă a sistemului x=332, y=5. Dacă se reia sistemul de mai sus, admiţând însă o variaţie mică a coeficientului 5 de pe lângă x din prima ecuaţie, adică

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

acelaşi procedeu de calcul ne conduce la soluţia x= -111.7845..., y=-1.7070.... S-a produs o catastrofă! Despre un astfel de sistem se spune că este *rău* (sau *prost*) *condiţionat.*

Un alt exemplu de sistem rău condiţionat :

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

Soluţia exactă este x=y=1.0, ceea ce se poate verifica prin substituire. Dacă înlocuim ecuaţia a doua a sistemului dat cu ecuaţia $0.5x+1.001y=1.5$, atunci soluţia devine x=3, y=0.

Sistemul de ecuaţii

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

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

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

1.7 Exerciţii

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

2. Fie aritmetica virgulei mobile cu β=10, t=4 şi se aplică rotunjirea prin tăiere. Să se determine eroarea relativă a diferenţei x-y, dacă

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

b) $x = \frac{2}{3}$ şi $y = 0.6665$.

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

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

b) $u=(1+x)^2$ şi $v=x^2+2x+1;$

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

Întotdeauna va avea loc $u=v$?

4. Să se calculeze $x=\frac{a^3\sqrt{b}}{c^2}$, unde x=7.45±0.001, b=50.46±0.02, c=15.4±0.03. Să se estimeze eroarea absolută şi eroarea relativă.

5. Fie aritmetica virgulei mobile β=10, t=3 şi 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 recurenţă

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

6. Care va fi rezultatul în urma rulării următoarei secvenţe 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 numeraţie β=2, β=8 şi respectiv β=16. Cum se reprezintă fracţiile 1/2, 2/3 şi 3/5 în aceste sisteme? Explicaţi rezultatele obţinute 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 iniţiere mai aprofundată în materialul prezentat în capitolul de faţă, recomandăm lucrările [2,15,16,17,22,37].

CAPITOLUL II

Rezolvarea numerică a ecuaţiilor algebrice şi transcendente

2.1 Introducere

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

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

Exemple:

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

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

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

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

O ecuaţie algebrică în forma generală se va scrie:

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

cu $n\geq 1$ şi coeficienţii reali $a_0, a_1, ..., a_n$; $a_n\neq 0$.

Orice ecuaţie algebrică (2.1) are exact n rădăcini, fiecare rădăcină multiplă fiind socotită ca atâtea rădăcini confundate cât arată ordinul ei de multiplicitate. Această afirmaţie poartă denumirea de *teorema fundamentală a algebrei* şi pentru prima dată se întâlneşte în lucrările lui Girard şi 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.

Ecuaţiile care nu sunt algebrice se numesc ecuaţii *transcendente.*

Exemple:

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

Rezolvarea ecuaţiei $f(x)=0$ (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 ecuaţiei;
2) calculul aproximativ ai fiecărei rădăcini şi 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 ecuaţia respectivă. Atunci graficul funcţiei $f(x)$ intersectează axa absciselor în punctul $x=r$. În general, prin *rădăcină aproximativă* se înţelege o valoare x_* suficient de apropiată de rădăcina exactă r. Cu alte cuvinte, trebuie să avem:

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

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

Se poate defini o rădăcină aproximativă şi altfel: numărul

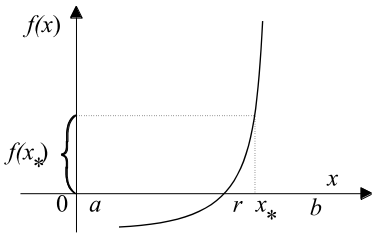


Fig. 2.1

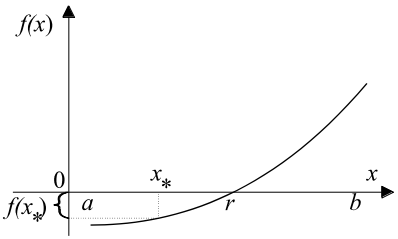


Fig. 2.2

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

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

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

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 condiţii.

2.2 Separarea rădăcinilor

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

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

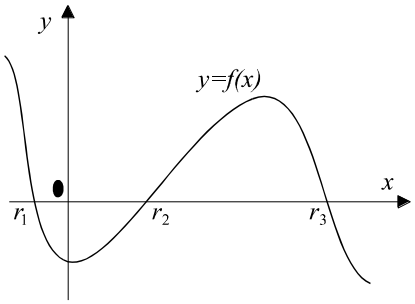


Fig. 2.3

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

Adeseori ecuaţia dată poate fi prezentată sub forma: $\varphi(x)=g(x)$. Rădăcinile ultimei ecuaţii sunt abscisele punctelor de intersecţie ale curbelor $y=\varphi(x)$ *şi* $y=g(x)$ (fig.2.4).

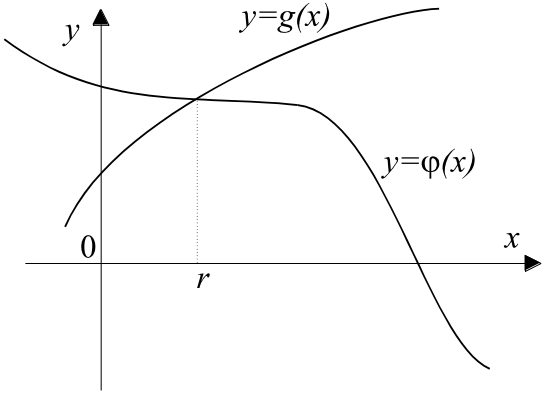


Fig. 2.4

Ecuaţia $f(x)=0$ este un caz particular al ecuaţiei $\varphi(x)=g(x)$, şi anume cazul în care $g(x)$ este funcţia constantă zero. Graficul acestei funcţii este axa *Ox*, care are acum rolul curbei $y=g(x)$.

Exemplu: Considerăm ecuaţia:

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

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

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

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

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

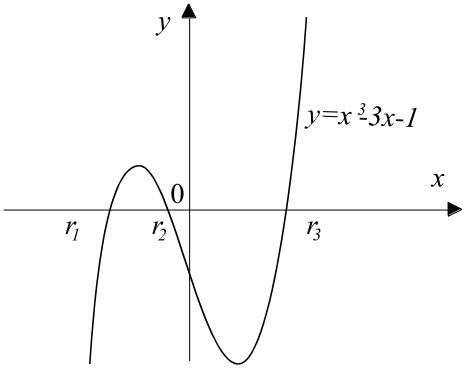


Fig. 2.5

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

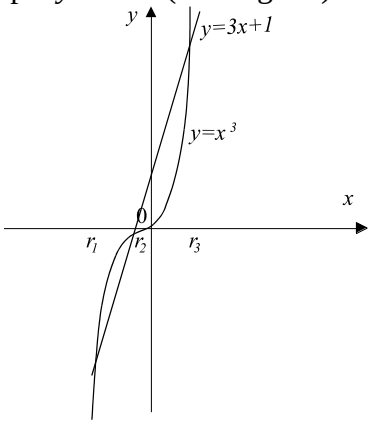


Fig. 2.6

Observaţie. În fig.2.7 se vede graficul funcţiei:

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

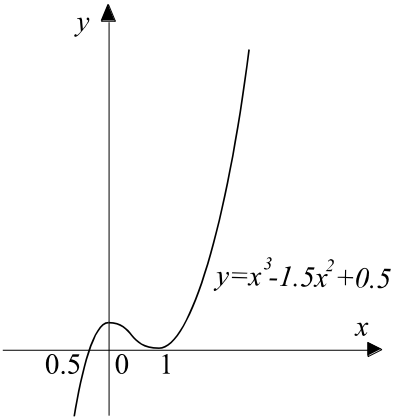


Fig. 2.7

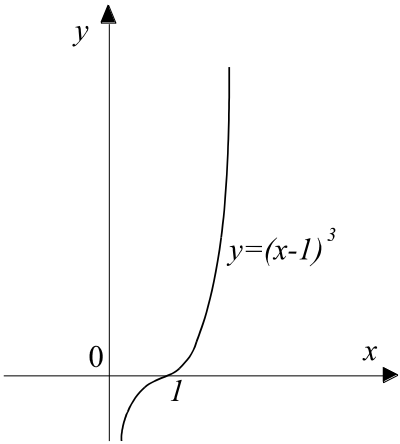


Fig. 2.8

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

În general, dacă în descompunerea funcţiei în factori apare factorul $(x-r)^k$ şi 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:

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

Ecuaţia $x^3-3x^2+3x-1=0$ 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 ecuaţii $f(x)=0$, dacă şi numai dacă $f(r)=0$ şi primele sale $k-1$ derivate

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

O rădăcină care nu este multiplă se numeşte rădăcină *simplă.*

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

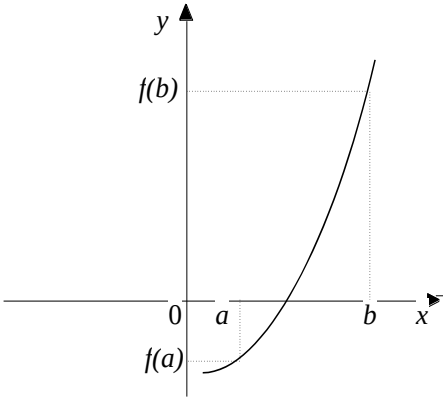


Fig. 2.9

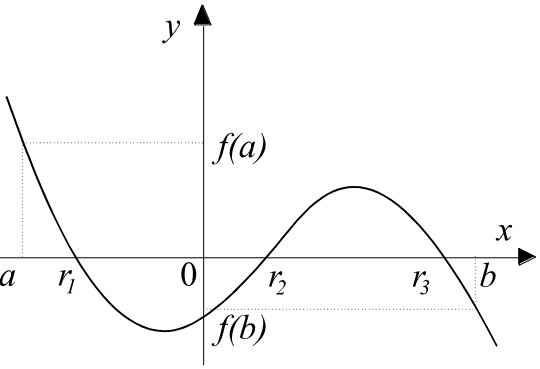


Fig. 2.10

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

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

Condiţia de continuitate a funcţiei $f(x)$, după cum se vede din fig.2.11, este esenţială în enunţul afirmaţiei de mai sus.

Dacă $f(a)f(b)>0$, adică $f(a)$ şi $f(b)$ au acelaşi semn, aceasta

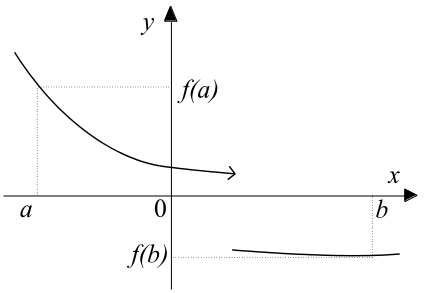


Fig. 2.11

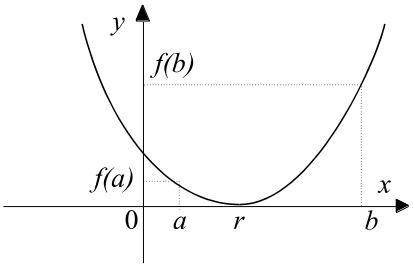


Fig. 2.12

încă nu înseamnă că ecuaţia $f(x)=0$ 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 funcţiei $f(x)$ există cel mult o rădăcină reală a ecuaţiei $f(x)=0$. Interpretarea geometrică a acestei afirmaţii este dată în fig.2.13.

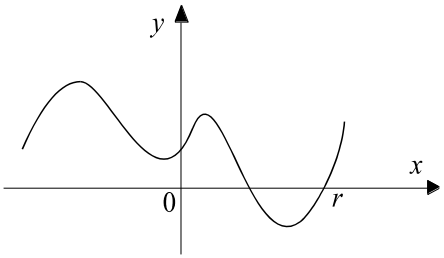


Fig. 2.13

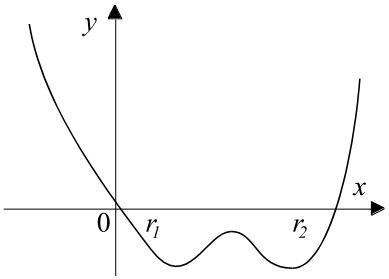


Fig. 2.14

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

Fie $a<x_1<x_2<\ldots<x_k<b$ rădăcinile ecuaţiei $f'(x)=0$, aşezate în ordine crescătoare. Şirul

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

se numeşte *şirul lui Rolle*.

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

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

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

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

x	a	x_1	x_2	…	x_k	b
$f(x)$	$f(a)$	$f(x_1)$	$f(x_2)$	…	$f(x_k)$	$f(b)$

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

Exemplu. Fie ecuaţia 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 $x_1=-1$, $x_2=3/4$, $x_3=1$. Şirul lui Rolle este următorul

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

La extremităţile intervalului (–2, –1) funcţia are valori de semne contrare, deci în acest interval ecuaţia are o singură rădăcină; în

intervalul (–1, 2) situaţia este aceeaşi; la extremităţile intervalului (–1, 3/4) şi (3/4, 1) funcţia are acelaşi semn, deci în acest interval ecuaţia nu are nici o rădăcină.

Prin urmare, avem două variaţii de semn, deci ecuaţia propusă are două rădăcini reale, r_1 şi r_2, şi anume:

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

Celelalte două rădăcini sunt complexe.

2.4 Metoda înjumătăţirii intervalului

Fie ecuaţia

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

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

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

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

şi

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

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

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

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

Evident că $sign\, f(a_1) = sign\, f(a_0)$ şi $sign\, f(b_1) = sign\, f(b_0)$ şi prin urmare $f(a_1)f(b_1)<0$. Continuând, se obţine succesiunea de subintervale

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

la extremităţile cărora funcţia ia valori de semne contrare,

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

şi

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

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

Se obţin astfel şirul $\{a_k\}$ nedescrescător, mărginit superior, şi şirul $\{b_k\}$ nedescrescător mărginit inferior. Trecând la limită în egalitatea (2.4), obţinem:

$$r = \lim_{k\to\infty} a_k = \lim_{k\to\infty} b_k \qquad (2.5)$$

Din (2.3) şi (2.5) rezultă că

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

de unde concluzionăm că şirurile $\{a_k\}$, $\{b_k\}$ şi $\{c_k\}$ sunt convergente către rădăcina ecuaţiei (2.2) şi avem:

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

În fig. 2.15. este ilustrat procesul de înjumătăţire 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),

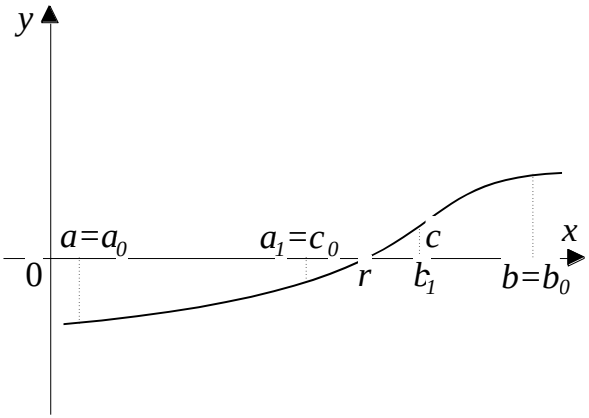


Fig. 2.15. Metoda înjumătăţirii intervalului

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

Observaţie. În programele de calculator operaţia de înjumătăţire 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 $\beta=10$ şi $t=3$. Atunci pentru $a=0.982$ şi $b=0.987$ vom avea :

$$c=(a+b)/2=(0.982+0.987)/2=1.96/2=0.980<a.$$

2.5 Metoda aproximaţiilor succesive

Fie

$$f(x)=0 \qquad (2.6)$$

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

Ecuaţia (2.6) o punem sub forma echivalentă:

$$x=\varphi(x). \tag{2.7}$$

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

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

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

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

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

$$x_{k+1}=\varphi(x_k), \qquad k=0, 1, 2, \ldots, \tag{2.8}$$

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

Şirul definit mai sus prin relaţia (2.8) se numeşte *şir de iterare* .

Să presupunem că există $\lim\limits_{k\to\infty} x_k = r$. În acest caz se spune că şirul de iterare este *convergent* şi *converge către r.*

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

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

deci limita r este chiar rădăcina ecuaţiei (2.7).

Din punct de vedere geometric, rădăcina reală r este abscisa punctului de intersecţie a curbei $y=\varphi(x)$ cu dreapta $y=x$. Modul cum şirul aproximaţiilor succesive $x_0, x_1, \ldots, x_k, \ldots$ conduce spre soluţia exactă este ilustrat în fig.2.16 şi fig.2.17 (în funcţie de forma curbei $y=\varphi(x)$).

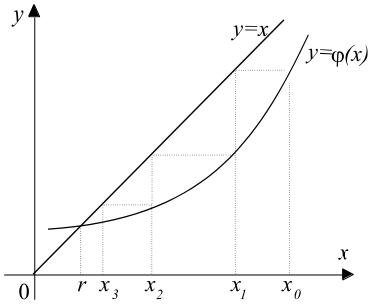


Fig. 2.16

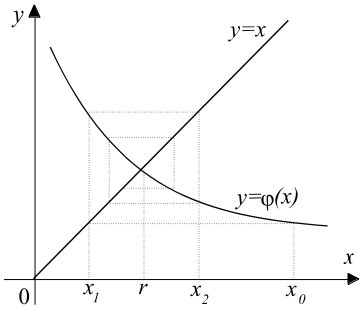


Fig. 2.17

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

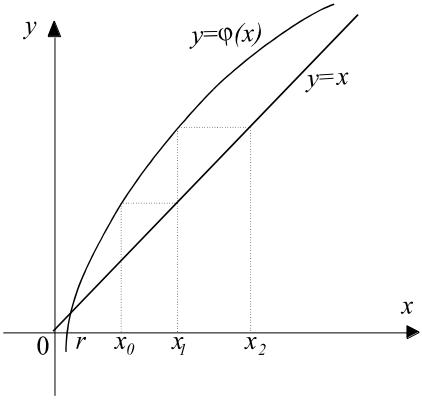
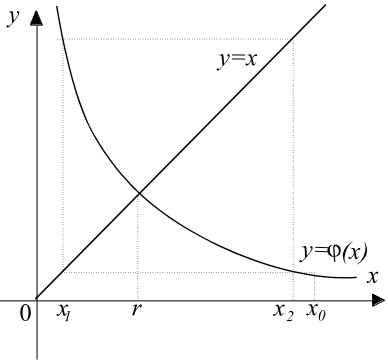


Fig. 2.18



Fif. 2.19

Deci şirul de iterare poate fi divergent chiar dacă x_0 se alege oricât de apropiat de rădăcină. O condiţie suficientă de convergenţă este dată de următoarea teoremă.

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

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

Demonstraţie. Datorită faptului că $\varphi(x)\in[a, b]$, $\forall x\in[a, b]$, rezultă că $\varphi(a)\ge a$ şi $\varphi(b)\le b$. Punem $g(x)=\varphi(x)-x$. Atunci $g(a)\ge 0$, $g(b)\le 0$ şi deoarece $g(x)$ este continuă, rezultă că există cel puţin un punct $r\in[a, b]$, astfel încât $g(r)=0$, adică r este rădăcina ecuaţiei iniţiale (2.7) în intervalul $[a, b]$. Să considerăm acum diferenţa:

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

Aplicând teorema lui Lagrange, se obţine:

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

Deoarece

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

avem:

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

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

$$|x_1-r| \le \alpha |x_0-r|,$$
$$|x_2-r| \le \alpha^2 |x_0-r|,$$
$$\ldots\ldots\ldots\ldots$$
$$|x_{k+1}-r| \le \alpha^{k+1} |x_0-r|.$$

Din ultima relaţie rezultă (fiindcă $0\le\alpha<1$):

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

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

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

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

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 inegalităţii din enunţul teoremei, se obţine:

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

Sau

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

Prin urmare:

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

Teorema este demonstrată.

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

Exemplu. Fie dată ecuaţia:

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

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

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

Pentru a verifica condiţia de convergenţă, calculăm derivata:

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

Condiţia de convergenţă $|\varphi'(x)|<1$ este îndeplinită pentru intervalul (2, 3) şi deci şirul de iterare este dat de

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

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

De remarcat că dacă am fi rearanjat ecuaţia iniţială

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

în felul următor: $x=x^3-x-9$ şi deci am fi folosit şirul de iterare: $x_{k+1}=x_k^3-x_k-9$, atunci metoda aproximaţiilor succesive diverge, condiţia suficientă de convergenţă nefiind îndeplinită:

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

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 cunoaşte, în mod general, apriori. În 2.5 ne vom ocupa de examinarea altor criterii de stopare în metodele iterative de rezolvare a ecuaţiilor algebrice şi transcendente.

În încheiere remarcăm că metoda aproximaţiilor succesive se utilizează cu succes şi pentru studiul aplicaţiilor numite contracţii în aşa-numite spaţii metrice.

2.5 Criterii de oprire în metodele iterative

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

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

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

Aplicând teorema lui Lagrange, se obţine:

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

de unde rezultă

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

Această relaţie evidenţiază, în primul rând, că dacă $|f'(r)|$ este mic, atunci $|f(x_k)|/m$ este mare şi perturbaţii slabe în x_k pot produce perturbaţii mari în rădăcină; în acest caz se spune că problema determinării lui r este rău condiţionată. 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 iteraţiile de îndată ce

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

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

Deoarece derivatele, în caz general, sunt greu de estimat, înclinăm să facem câteva iteraţii 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

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

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

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

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

2.6 Metoda lui Newton (metoda tangentei)

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

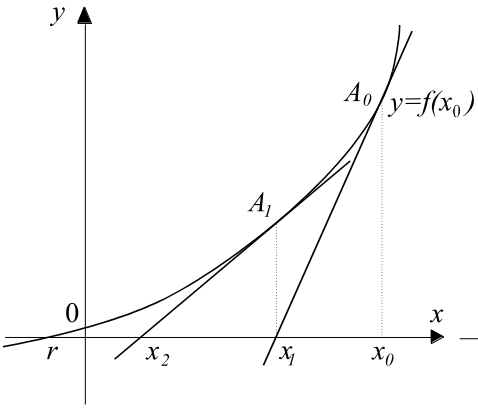


Fig. 2.20

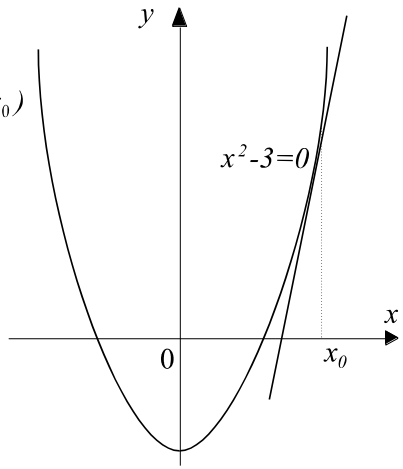


Fig. 2.21

singură rădăcină reală r în intervalul $[a, b]$. Să presupunem în plus că derivatele $f'(x)$ şi $f''(x)$ 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

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

vom avea:

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

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

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

Considerăm punctul de coordonate $A_1(x_1, f(x_1))$ şi 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 obţine metoda tangentelor definită de următoarea formulă de iterare:

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

Această metodă se mai numeşte şi *metoda lui Newton.*

Exemplu. Să aplicăm metoda lui Newton la calculul $\sqrt{3}$. Pentru aceasta scriem ecuaţia $x^2-3=0$ (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

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

Alegând x_0=2, obţinem:

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

Se observă că avem o convergenţă rapidă a şirului x_1, x_2, ... către $\sqrt{3}$. Acest lucru nu este întâmplător. Analizăm cazul general. Fie dat $a>0$ şi să se găsească rădăcina ecuaţiei:

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

Atunci

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

şi

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

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

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

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

şi $r\in[a, b]$ este rădăcina ecuaţiei $f(x)=0$. Atunci şirul de iterare determinat de relaţia (2.11) converge către r dacă aproximaţia iniţială x_0 este aleasă într-o vecinătate a rădăcinii r. Eroarea este

estimată de relaţia:

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

Demonstraţie. Se observă că metoda lui Newton este un caz particular al metodei aproximaţiilor succesive cu funcţia:

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

Se verifică imediat că $r=\varphi(r)$ şi $\varphi'(r)=0$. Deci putem afirma că într-o vecinătate a rădăcinii r se îndeplineşte condiţia $|\varphi'(x)| \le \alpha < 1$. Rezultă că şirul (2.11) converge (către rădăcina r), dacă aproximaţia iniţială x_0 este aleasă suficient de aproape de rădăcină.

Pentru a analiza viteza de convergenţă se evaluează diferenţa:

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

Prin aplicarea repetată a teoremei de medie a lui Lagrange, ţinând seama de condiţiile din enunţ, rezultă:

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

unde

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

Avem

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

şi teorema este demonstrată.

O dificultate în aplicarea metodei lui Newton (şi în general a unei metode iterative) o reprezintă alegerea aproximaţiei iniţiale x_0. Să presupunem, de exemplu, că se rezolvă ecuaţia arctg x=0 (fig.2.22).

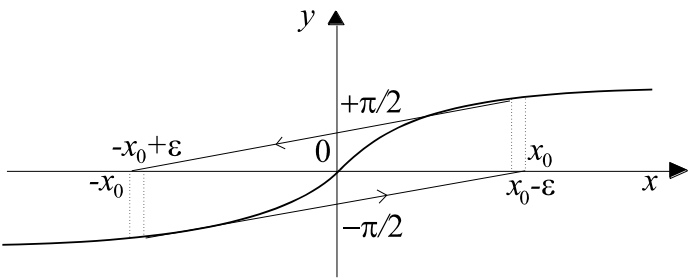


Fig. 2.22

Alegem ca aproximaţie iniţială punctul $x_0 \in [1{,}39;\ 1{,}40]$. Atunci tangenta dusă în punctul $(x_0, f(x_0))$ va trece prin punctul de abscisă $-x_0$ şi apoi tangenta dusă în $(-x_0, f(-x_0))$ trece prin x_0. În acest caz, şirul x_k începe să "cicleze". Se numeşte *zonă de convergenţă* a rădăcinii r mulţimea tuturor aproximaţiilor iniţiale x_0 pentru care şirul iterativ $\{x_k\}$ tinde către r. În fig.2.22 zona de convergenţă a rădăcinii $r=0$ este intervalul $(-x_0+\varepsilon,\ x_0-\varepsilon)$. Alegerea aproximaţiei iniţiale în afara zonei de convergenţă a rădăcinii dorite nu permite să găsim această rădăcină (putem însă nimeri, eventual, în zona de convergenţă a altei rădăcini).

În cazul ecuaţiei $x^3-x=0$ zona de convergenţă 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 şir care "ciclează". Dacă $x_0=\pm\sqrt{3}/3$, atunci $f'(x_0)=0$ şi deci tangenta la curba dată în aceste puncte este paralelă cu axa Ox. Alegând aproximaţia iniţială $x_0<-\sqrt{3}/3$, vom avea că şirul iterativ $\{x_k\}$

tinde către rădăcina $r=-1$; alegând $x_0 > \sqrt{3}/3$, se asigură convergenţa către rădăcina $r=1$. În practică se recomandă de a alege valoarea aproximativă iniţială x_0 astfel ca

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

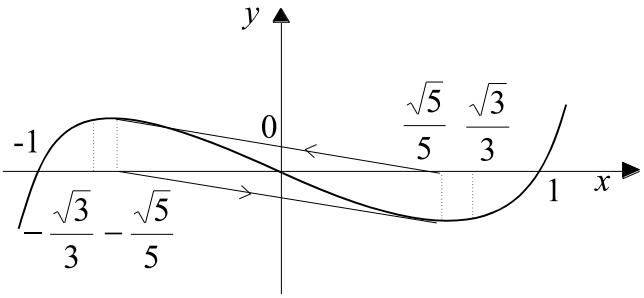


Fig. 2.23

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 iteraţiilor poate servi următorul: iterţiile se întrerup atunci când $|x_{k+1}-x_k|$ şi (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 convergenţă. Dacă r este o rădăcină multiplă, atunci convergenţa şirului $\{x_k\}$ este liniară.

Exemplu. Fie dată ecuaţia $x^2=0$ 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 cunoaşte gradul de multiplicitate a rădăcinii, atunci putem accelera convergenţa şirului construit prin metoda lui Newton. Dacă r este o rădăcină multiplă de gradul p, adică

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

se recomandă de a efectua calculele conform formulei de iterare:

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

2.7 Metoda secantei

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

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

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

Obţinem:

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

Pentru startul iteraţiilor în metoda secantei sunt necesare de două aproximaţii iniţiale x_0 şi x_1. Valoarea x_{k+1} este abscisa punctului de intersecţie 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 $(x_{k-1}, f(x_{k-1}))$ şi $(x_k, f(x_k))$ şi Ox (fig.2.24); de aici şi denumirea metodei.

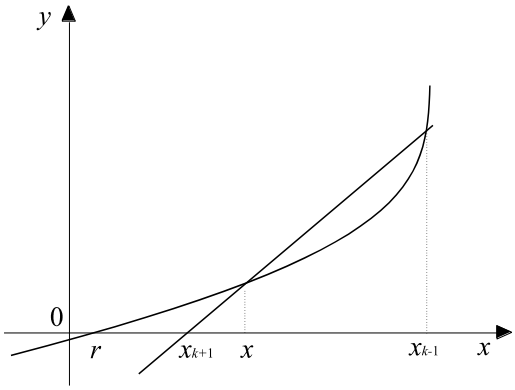


Fig. 2.24. Metoda secantei

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

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

Dacă se cunoaşte

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

şi dorim să determinăm rădăcina cu eroarea $\varepsilon>0$, vom întrerupe calculele când $|f(x_k)|/m<\varepsilon$. În cazul când se ştie şi 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 convergenţă a metodei secantei, presupunând că se îndeplinesc condiţiile: $f(r)=0$, $f'(r)\neq 0$, iar $f''(x)$ sunt funcţii continue şi păstrează un semn constant pe intervalul $[a, b]$. Pentru aceasta se dezvoltă funcţia $f(x)$ în serie Taylor în vecinătatea punctului $x=r$:

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

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 $x=x_{k-1}$ şi $x=x_k$, apoi o trunchiem după termenul al doilea (ţinând 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), obţinem:

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

sau

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

Deci putem scrie:

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

Relaţia de recurenţă (2.13) o vom pune sub forma:

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

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

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

Se va lua numai rădăcina pozitivă a ecuaţiei pătrate (2.14), deoarece numai ea garantează convergenţa şirului $\{x_k\}$. 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) $\beta=2$ şi deci metoda lui Newton converge mai repede decât metoda secantei. Pe de altă parte, metoda lui Newton reclamă necesitatea evaluării funcţiei şi a derivatei sale, iar metoda secantei necesită numai calculul funcţiei. De aceea la aceeaşi cantitate de operaţii în metoda secantei se poate face de două ori mai mulţi "paşi" şi deci se poate obţine rădăcina cu o precizie mai înaltă.

Să observăm că la fiecare pas nou în metoda secantei se calculează o singură valoare nouă pentru funcţia 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 $f(x_k)\, f(x_{k-1})>0$ şi $x_k \approx x_{k-1}$, atunci poate avea loc o neutralizare a termenilor.

2.8 Rezolvarea ecuaţiilor algebrice

Considerăm ecuaţia algebrică

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

unde coeficienţii $a_0, a_1, a_2, ..., a_n$ sunt reali, $a_n>0$.

2.8.1 Proprietăţile ecuaţiilor algebrice

Fără a putea rezolva ecuaţia algebrică (2.15) se poate stabili câte rădăcini are ecuaţia, unele relaţii între coeficienţii polinomului $P_n(x)$ şi rădăcinile ecuaţiei (2.15) etc. În cele ce urmează vom aduce unele dintre aceste proprietăţi:

1. O ecuaţie algebrică (2.15) de gradul n are exact n rădăcini (printre care pot fi atât reale, cât şi complexe), fiecare rădăcină multiplă fiind considerată ca atâtea rădăcini confundate cât indică ordinul ei de multiplicitate.

2. Între coeficienţii ecuaţiei algebrice (2.15) şi rădăcinile ei r_1, r_2, ... , r_n există relaţiile (formulele lui Viète):

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

$$r_1r_2+r_1r_3+...+r_{n-1}r_n=\frac{a_{n-2}}{a_n},$$

$$r_1r_2r_3+r_1r_3r_4+...+r_{n-2}r_{n-1}r_n=-\frac{a_{n-3}}{a_n},$$

.....

$$r_1r_2 \dots r_n=(-1)^n\frac{a_0}{a_n}.$$

3. Rădăcinile complexe ale ecuaţiei algebrice (2.15) cu coeficienţi reali sunt conjugate două câte două.(Numărul $a-ib$, $i=\sqrt{-1}$, se numeşte conjugatul lui $a+ib$).
4. Dacă o ecuaţie algebrică (2.15) cu coeficienţi raţionali admite ca rădăcină un iraţional pătratic $m+n\sqrt{p}$ (m, $n\in Q$, $n\neq 0$, $p\in N$), ea admite şi conjugatul său, $m-n\sqrt{p}$, ca rădăcină.
5. Ecuaţia algebrică (2.15) de ordin n par şi cu $a_0<0$ are cel puţin două rădăcini reale de semne diferite.
6. Rădăcinile reale şi complexe ale ecuaţiei algebrice (2.15) sunt situate în inelul circular $R_1<|x|<R_2$, unde

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

Aceste şi alte proprietăţi ale ecuaţiilor 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 şi al derivatelor lui.

Considerăm polinomul de gradul n:

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

Pentru stabilirea schemei lui Horner se transcrie polinomul astfel:

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

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

$$b_n = a_n ,$$

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

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

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

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

Această schemă de obţinere a şirului finit $\{b_i\}$ se numeşte *schema lui Horner*. Ea necesită cel mult $2n$ operaţii aritmetice. Se demonstrează, că în cazul general, când toţi coeficienţii polinomului dat $P_n(x)$ sunt diferiţi de zero, nu există o schemă mai eficientă de calcul, decât cea a lui Horner.

Câtul împărţirii lui $P_n(x)$ la $x-\xi$ este dat de polinomul $P_{n-1}(x)$ de gradul $n-1$ cu coeficienţii $b_1, b_2, \ldots, b_n$:

$$P_{n-1}(x)=b_nx^{n-1}+b_{n-1}x^{n-2}+...+b_2x+b_1 ,$$

adică

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

Ultima relaţie ne dă:

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

Trecând la limită în această identitate, obţinem:

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

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 obţine cu ajutorul schemei lui Horner:

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

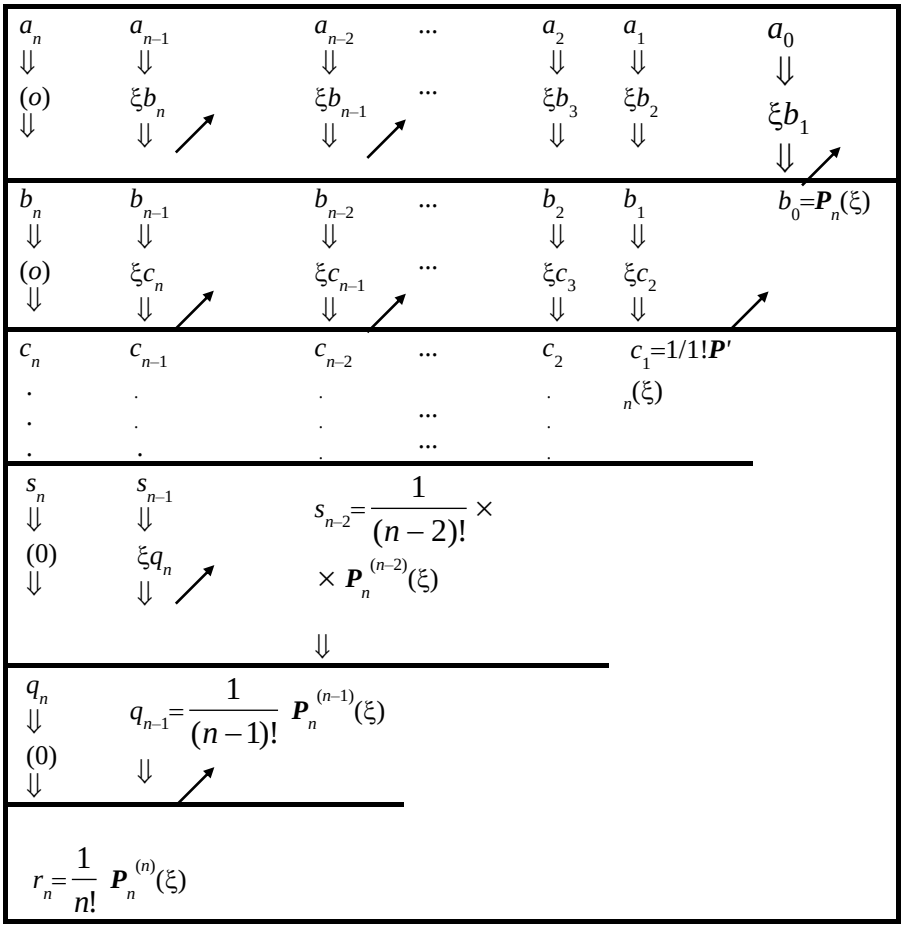
Procedeul poate fi continuat obţinându-se valorile oricărei derivate a lui $P_n(x)$ într-un punct fixat $x=\xi$.

Pentru uşurinţa calculelor coeficienţii polinoamelor

$$P_n^{'},P_n^{''},\dots,P_n^{(n)}$$

se pot determina cu ajutorul schemei concise a lui Horner.

Schema concisă a lui Horner



2.8.3 Metoda lui Newton

Fie dată ecuaţia algebrică

$$P_n(x)\equiv a_n x^n+a_{n-1}x^{n-1}+...+a_1x+a_0=0,$$

cu coeficienţi reali.

Vom folosi metoda lui Newton pentru a determina o rădăcină reală a ecuaţiei date. Şirul de iterare Newton (vezi paagraful 2.6) devine:

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

Valorile lui $P_n(x)$ şi $P_n'(x)$ în punctele fixate x_k, k=0, 1, 2, ... , se calculează cu ajutorul schemei lui Horner (vezi 2.8.2):

$$b_n=a_n,$$

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

$$P_n(x_k)=b_0,$$

$$c_n=b_n,$$

$$c_j=b_j+x_k c_{j+1}, \; j=n-1,\ldots,1,$$

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

Prin urmare, expresia şirului de iterare va fi

$$x_{k+1}=x_k-\frac{b_0}{c_1}, \qquad k=0, 1, 2, \ldots \tag{2.16}$$

Cu ajutorul acestei metode pot fi calculate toate rădăcinile reale ale ecuaţiei algebrice date. Într-adevăr, fie r_1 rădăcina (simplă) obţinută 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 ecuaţia algebrică de gradul n–1:

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

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

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

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

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

utilizând x_0=1.3.

Folosind schema concisă a lui Horner, se obţine:

+	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	1.324718	1.154878	1.0000004
=	1	1.324718	0.154878	$0.0000004 = b_0 = P_3(1.324718)$
+	0	1.324718	4.109756	
=	1	2.649436		$4.264634 = c_1 = P'_3(1.324718)$

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

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

2.9 Alte metode numerice

Considerăm o ecuaţie algebrică sau transcendentă

$$f(x)=0.$$

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

Metoda coardei este o variantă a metodei secantei, în care

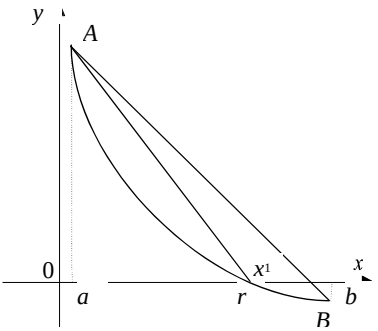


Fig. 2.25

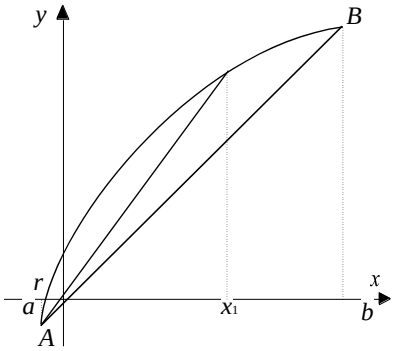


Fig. 2.26

se alege coarda care trece prin $(a, f(a))$ şi $(x_k(f(x_k))$ sau $(b, f(b))$ şi $(x_k(f(x_k))$, 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.

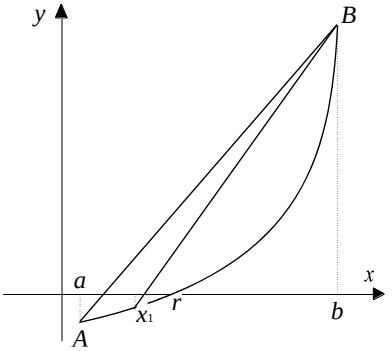


Fig. 2.27

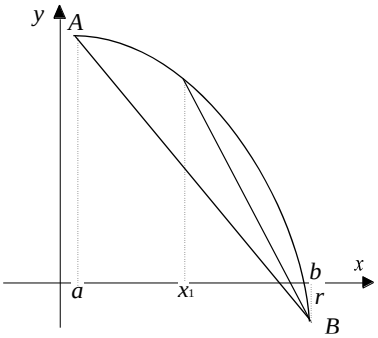


Fig. 2.28

Ecuaţia dreptei care trece prin punctele *A* şi *B* este:

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

Pentru a calcula abscisa punctului de intersecţie a coardei *AB* cu axa *Ox* punem *y*=0 şi 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) şi (x_1,b). Din (a,x_1) şi (x_1,b) este ales acel interval la extremităţile căruia funcţia *f*(*x*) are semne contrare şi procedura se repetă în mod analog. Concis, formulele de iterare a metodei coardei pentru generarea şirului de aproximaţii ale lui *r*, plecând de la aproximaţia iniţială x_0, se scriu astfel:

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

sau

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

În dependenţă de forma funcţiei 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 acelaşi 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 şi fig.2.26), atunci se aplică formula (2.18). În caz contrar (vezi fig.2.27 şi fig.2.28) se va aplica formula (2.17).

Această metodă se mai numeşte şi *regula falsei poziţii*.

Avantajul metodei coardei este că ea produce un şir care, pentru funcţiile continue, este întotdeauna convergent (spre deosebire de metoda secantei sau metoda lui Newton). După cum se vede din fig.2.25 - fig.2.28 convergenţa şirului construit prin metoda coardei către rădăcina ecuaţiei 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 convergenţă este pătratic, adică eroarea la fiecare iteraţie este proporţională cu pătratul erorii de la iteraţia anterioară. Pentru metoda secantei gradul de convergenţă 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 foloseşte doar o evaluare a funcţiei la fiecare pas nu poate avea ordinul doi de convergenţă.

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

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

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

Metode de ordinul trei de convergenţă:

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

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

Metode de ordinul patru de convergenţă:

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

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

2.10 Exerciţii

1. Să se separe rădăcinile reale ale ecuaţiilor:

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

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

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

2. Presupunem că există un interval $[a, b]$, astfel încât $f(a) f(b) < 0$ şi în intervalul [a,b] funcţia este continuă, iar ecuaţia $f(x)=0$ admite soluţie unică. Care este numărul

maximal de înjumătăţiri al intervalului dat pentru a obţine soluţia cu eroarea dată $\varepsilon>0$?

3. Considerăm ecuaţia $f(x)=0$, unde

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

Indicaţi zona de convergenţă 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 ecuaţiei $x^3-3=0$. Să se scrie formula generală pentru calculul $\sqrt[m]{a}$.

5. Să se calculeze o rădăcină reală a ecuaţiilor:

$$x^3-2x^2-4x+7=0, \quad x\in(-1, 0),$$
$$\lg(2x+3)+2x-1=0, \quad x\in(0, 0.05),$$

folosind metoda înjumătăţirii intervalului, metoda aproximaţiilor succesive şi metoda lui Newton. Să se compare rezultatele.

6. Să se determine condiţia necesară şi suficientă pe care trebuie să o satisfacă numerele p şi q pentru ca ecuaţia $x^3+px+q=0$ să aibă o rădăcină dublă.

7. Să se arate că şirul Newton pentru determinarea rădăcinii ecuaţiei $1/x-a=0$ este convergent, dacă se alege x_0, astfel încât $|1-ax_0|<1$. Să se caluleze 1/12, alegând $x_0=0.1$. Ce se întîmplă dacă se alege $x_0=1$?

8. Să se rezolve ecuaţia algebrică $x^3-2x^2-4x+7=0$, utilizând metoda Birge-Viĭte.

CAPITOLUL III

METODE NUMERERICE ÎN ALGEBRA LINIARĂ

3.1 Elemente de analiză matriceală

3.1.1. Vectori şi matrice

Un tablou dreptunghiular de $m \times n$ numere reale aşezate pe *m* linii şi *n* coloane:

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

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

Matricea se mai poate reprezenta simbolic astfel: $A = (a_{ij}),\ i = 1,2,...,m;\ j = 1,2,...,n,$ sau $A = (a_{ij})_{mn}$. Vom spune că matricea A este de dimensiune $m \times n$. În cazul când $m = n$, matricea se numeşte *pătrată de ordinul n* şi se notează $A = (a_{ij})_n$. Dacă $m \neq n$, matricea se numeşte *rectangulară (dreptunghiulară).* O matrice *1xn* se numeşte *vector linie*, iar o matrice *nx1* este un *vector coloană*.

Un sistem ordonat de *n* numere reale se numeşte *vector n-dimensional.* Un vector se reprezintă printr-o matrice cu o singură linie sau o singură coloană. În lucrarea de faţă prin vector vom înţelege întotdeauna vector-coloană:

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

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

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

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

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

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

Mulţimea tuturor vectorilor *n*-dimensionali (*n* natural, fixat) se numeşte *spaţiul liniar n-dimensional* şi se notează cu R^n.

Suma matricelor A şi *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 defineşte numai în cazul când numărul coloanelor primului factor este egal cu numărul liniilor celui de-al doilea factor. Astfel, dacă $A=(a_{ij})_{mn}$ şi $B=(b_{ij})_{np}$, atunci $C=A \cdot B$, unde $C=(c_{ij})_{mp}$ şi

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

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

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

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

$x^T y$ se numeşte *produsul scalar* al vectorilor $x,y \in R^n$ şi se mai notează (x,y) sau $\langle x,y \rangle$.

xy^T se numeşte *produsul diadic* al vectorilor $x,y \in R^n$; este o matrice pătrată de ordinul n şi se mai notează astfel: $>x,y<$.

Pentru orice vectori din R^n au loc proprietăţile:

1. $(x,y)=(y,x)$;
2. $(x+y,z)=(x,z)+(y,z)$;
3. $(ax,y)=a(x,y)$; $(a \in R)$;
4. $(x,x) \geq 0$; $(x,x)=0$ atunci şi 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 înmulţi cu ea însăşi:

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

Au loc inegalităţile:

1. $(A \cdot B)C = A(B \cdot C)$, legea asociativităţii,
2. $A(B+C) = AB + AC$, legea distributivităţii la stânga,
3. $(B+C)A = BA + CA$, legea distributivităţii la dreapta,
4. $\alpha(AB) = (\alpha A)B = A(\alpha B), \alpha \in R$.

3.1.2. Norme de vectori şi matrice

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

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

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 proprietăţile din definiţia de mai sus a normei. Norma $\|x\|_2$ se numeşte şi *normă euclidiană.* Ea provine din produsul scalar:

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

şi generalizează noţiunea de lungime a vectorului.

Au loc următoarele inegalităţi:

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

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

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

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

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

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

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

Această relaţie se numeşte *inegalitatea lui Schwarz Cauchy Buniacovski.*

Unghiul dintre doi vectori *x,y* din R^n se defineşte 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 mulţimea matricelor pătrate se poate introduce o normă $\|A\|$ în sensul definit mai sus pentru vectori. Mai importante şi mai utilizate sunt însă normele matriceale definite astfel:

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

Această normă satisface următoarea condiţie:

$$\|A \cdot B\| \le \|A\| \cdot \|B\|$$

oricare ar fi matricele *A* şi *B*.

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

$$\|Ax\| \le \|A\| \|x\|$$

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

Normele:

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

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

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

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

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

Raza spectrală a lui *A* se defineşte prin relaţia:

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

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

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

În aplicaţii 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:

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

Oricare ar fi matricea *A* avem *IA=AI=A.* Matricea *A* se numeşte 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 şi

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

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

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

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

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

3.1.3. Matrice speciale

O matrice pătrată de forma:

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

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

$$D=Diag(d_1,d_2,...,d_n).$$

Matricea A se numeşte *inferior triunghiulară (superior triunghiulară)* dacă elementele sale satisfac relaţiile:

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

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

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

Matricea A se numeşte *tridiagonală* dacă elementele sale satisfac relaţiile:

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

Astfel:

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

O matrice Q se numeşte *ortogonală* dacă:

$$Q^TQ=QQ^T=I \quad sau \quad Q^T=Q^{-1}.$$

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 şi 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 roteşte orice vector cu unghiul θ, iar matricea Q^T îl roteşte în direcţie inversă cu unghiul $-\theta$.

Matricea obţinută din matricea unitate prin reordonarea coloanelor ei se numeşte matrice de *permutare* şi se notează *P*. De exemplu:

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

O matrice de forma uv^T, unde $u, v \in R^n$, se numeşte 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 numeşte *reflector* sau matricea lui *Householder*.

Se verifică uşor că matricele *P* şi *H* sunt ortogonale.

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

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

O matrice *A* simetrică se numeşte *pozitiv definită* dacă

$$(Ax, x) > 0 \quad pentru \forall\, x \neq 0, x \in R^n.$$

Elementele de pe diagonala principală a matricei pozitiv definite sunt pozitive. Pe diagonala principală se află şi elementul maximal (în modul) al unei matrice pozitiv definite.

Exemplu. Matricea

$$A = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$$

este pozitiv definită, deoarece

$$(Ax, x) = x^T Ax = (x_1 - x_2)\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1^2 - 2x_1x_2 + 2x_2^2 =$$

$$= (x_1 - x_2)^2 + x_2^2 > 0 \quad pentru \forall x \neq 0.$$

3.2 Sisteme de ecuaţii algebrice liniare

Considerăm sistemul de ecuaţii algebrice liniare

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2 \\ ..... \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = b_n \end{cases} \tag{3.1}$$

Acest sistem poate fi scris sub formă matriceală:

$$Ax = b.$$

unde

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ ..... & ..... & ..... & ..... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix} \qquad b = \begin{pmatrix} b_1 \\ b_2 \\ ..... \\ b_n \end{pmatrix} \qquad x = \begin{pmatrix} x_1 \\ x_2 \\ ..... \\ x_n \end{pmatrix}.$$

A rezolva sistemul de ecuaţii dat înseamnă a determina un vector $x^* \in R^n$ care satisface egalităţile (3.1).

Oricare ar fi sistemul de ecuaţii liniare există trei şi numai trei posibilităţi.

1. Sistemul de ecuaţii are o soluţie 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 soluţia $x_1^* = x_2^* = 1$. În fig. 3.1 este dată interpretarea geometrică a sistemului considerat, din care se vede că dreptele $x_2 = 2 - x_1$ şi $x_2 = -1 + 2x_1$ se intersectează numai într-un singur punct.

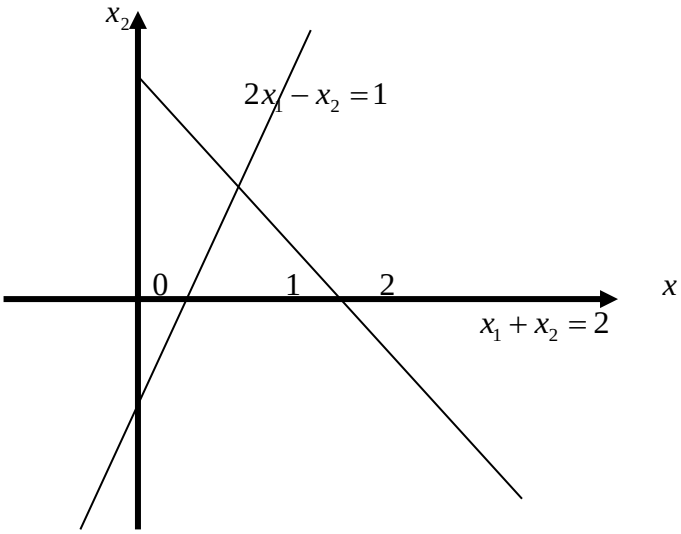


Fig. 3.1 Interpretarea geometrică a sistemului compatibil determinat

2. Sistemul de ecuaţii are o infinitate de soluţii. Despre astfel de sisteme se spune că sunt *compatibil nedeterminate*. In 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 soluţii; aceste două soluţii descriu una şi aceeaşi dreaptă $x_2 = 1 - x_1$.

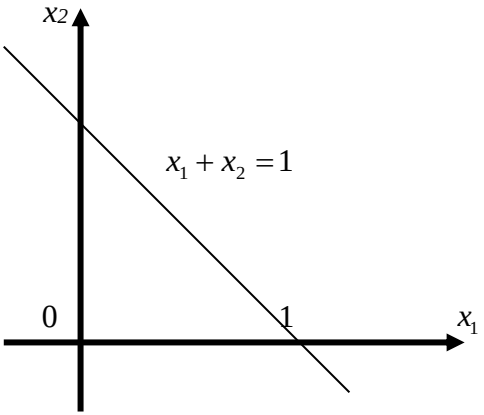


Fig. 3.2 Interpretarea geometrică a sistemului compatibil nedeterminat

3. Sistemul de ecuaţii nu are soluţii, adică este *incompatibil*. De exemplu, sistemul:

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

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

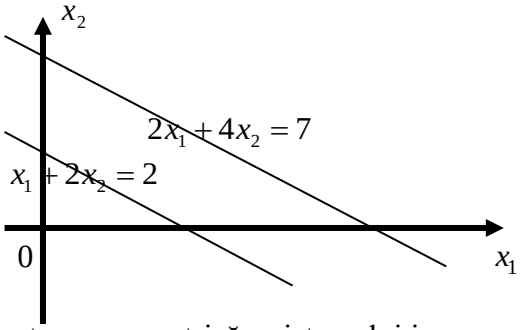


Fig.3.3 Interpretarea geometrică a sistemului incompatibil

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

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

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

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

$$7x = 21.$$

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

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

Aplicarea matricei inverse ne-ar duce la:

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

Al doilea procedeu necesită cu o operaţie aritmetică mai mult şi dă un rezultat mai puţin precis. Acelaşi lucru, dar într-un mod mai pronunţat, este adevărat şi în cazul rezolvării sistemelor cu multe ecuaţii. De aceea relaţia (3.2) trebuie interpretată doar în sensul de exprimare a faptului că x^* este soluţia unică a sistemului $Ax = b$, dar nu şi ca o cale de obţinere a acestei soluţii.

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

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

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 operaţii aritmetice, şi anume este necesar să se calculeze $n+1$ determinanţi $(\Delta, \Delta_1, ..., \Delta_n)$ şi să se efectueze n împărţiri. Pentru calculul unui determinant sunt necesare $(n-1) \times n$

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

Metodele numerice de rezolvare a sistemelor de ecuaţii liniare sunt de două tipuri: metode directe şi 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 soluţia exactă se obţine după un număr finit de operaţii aritmetice elementare (adunare, scădere, înmulţire, împărţire şi rădăcină pătrată) şi acest număr de operaţii este de ordinul n^3. Subliniem că soluţia exactă se obţine în cazurile (ideale) în care erorile de rotunjire sunt absente. La fiecare operaţie elementară efectuată de calculator avem o eroare de rotunjire şi prin urmare metodele directe în caz general furnizează doar o soluţie aproximativă. Metodele directe se utilizează pentru rezolvarea sistemelor nu prea "mari", de dimensiune $n \le 200$.

Rezolvarea sistemelor de ecuaţii liniare printr-o *metodă iterativă* înseamnă construirea unui şir de vectori $x^{(k)}$, $k = 0,1,...$ (pornind de la un vector $x^{(0)}$ ales arbitrar) convergent către soluţia sistemului considerat. În metodele iterative, de obicei, o iteraţie necesită efectuarea unui număr de ordinul n^2 operaţii aritmetice. De aceea metodele iterative se utilizează pentru rezolvarea sistemelor "mari", de dimensiune $n \ge 10^2$ (în cazul asigurării unei viteze sporite de convergenţă pentru o alegere a aproximării iniţiale adecvate). Trunchierea şirului $\{x^{(k)}\}$ are loc la un indice m astfel încât $x^{(m)}$ constituie o aproximaţie satisfăcătoare a soluţiei 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 iniţial la un sistem echivalent având matricea coeficienţilor superior triunghiulară. Transformarea sistemului dat într-un sistem de formă triunghiulară fără ca să se modifice soluţia sistemului se realizează cu ajutorul următoarelor trei operaţii de bază:

1. rearanjarea ecuaţiilor (schimbarea a două ecuaţii între ele);
2. înmulţirea unei ecuaţii cu o constantă (diferită de zero);
3. scăderea unei ecuaţii din alta şi înlocuirea celei de a doua cu rezultatul scăderii.

Exemplificăm această metodă pentru următorul sistem de ecuaţii liniare:

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

Putem elimina necunoscuta x_1 din ultimele două ecuaţii, înmulţind prima ecuaţie respectiv cu factorii:

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

şi scăzând-o din ecuaţia a doua şi apoi din ecuaţia a treia.

Obţinem astfel sistemul echivalent

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

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

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

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

În continuare se determină necunoscutele începând cu ecuaţia a treia: $x_3^* = 1$; înlocuind rezultatul obţinut în ecuaţia a doua vom obţine pe $x_2^* = 2$; în sfârşit din prima ecuaţie avem $x_1^* = -1.$

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

$$Ax = b. \tag{3.3}$$

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

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

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

şi scăzând rezultatul obţinut din ecuaţia *i* pentru $\forall i \geq 2.$

Obţinem în acest caz sistemul echivalent:

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

cu coeficienţii:

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

Mai sus s-a notat $a_{ij}^{(1)} = a_{ij}; \quad i, j = 1,2,...,n$ şi $b_i^{(1)} = b_i$;. i=1,2,…,n. Prima ecuaţie a sistemului (3.4) coincide cu prima ecuaţie a sistemului (3.3). În continuare se repetă procedeul de mai sus pentru eliminarea necunoscutei x_2 din sistemul (34) ş.a.m.d. La pasul *k* se obţine sistemul:

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

unde

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

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

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

unde

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

iar

$$b_i^{(k)} = \begin{cases} b_i^{(k-1)} & , pentru \quad i \le k-1, \\ b_i^{(k-1)} - \mu_{i,k-1} \cdot b_{k-1}^{(k-1)} & , pentru \quad i \ge k. \end{cases}$$

După n paşi necunoscuta x_{n-1} va fi eliminată din ultima ecuaţie, obţinându-se un sistem cum matricea superior triunghiulară:

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \quad \ldots \quad + a_{1k}^{(1)}x_k + \quad \ldots \quad + a_{1n}^{(1)}x_n = b_1^{(1)}, \\ \qquad a_{22}^{(2)}x_2 + \quad \ldots \quad + a_{2k}^{(2)}x_k + \quad \ldots \quad + a_{2n}^{(2)}x_n = b_2^{(2)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad\qquad a_{kk}^{(k)}x_k + \quad \ldots \quad + a_{kn}^{(k)}x_n = b_k^{(k)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad\qquad\qquad a_{nn}^{(n)}x_n = b_n^{(n)}. \end{cases}$$

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

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

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

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 puţin unul din celelalte elemente din coloana *k* şi din liniile *k+1,k+2,...,n* este nenul; în caz contrar matricea *A* ar fi singulară ($\det(A) = 0$). Permutând ecuaţiile sistemului putem aduce pe locul *(k,k)* elementul nenul şi, 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 ecuaţii nu poate fi utilizată pentru a elimina pe x_1 din ecuaţia a doua. Schimbând ecuaţiile între ele cu locul, obţinem:

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

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

Analizăm un alt exemplu:

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

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

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

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

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

şi metoda eliminării lui Gauss îl transformă în:

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

cu soluţia $x_1^* = x_2^* = 1.00$.

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 ecuaţiilor.

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

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

şi se permută liniile *k* şi *r*.

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

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

Matricea *A* se numeşte *diagonal dominantă* dacă

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

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

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

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 şi în cazul când matricea *A* este pozitiv definită.

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

$$N_1=(n^2-n)+\ldots+(k^2-k)+\ldots+(1^2-1)=$$

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

Procesul de eliminare inversă se efectuează cu mult mai repede. Necunoscuta x_n se află cu ajutorul unei singure operaţii (împărţirea la elementul pivot); calculul x_{n-1} cere două operaţii (împărţire – scădere şi apoi împărţire) ş.a.m.d.; pasul *k* necesită numai *k* operaţii. Prin urmare eliminarea inversă necesită

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

de operaţii aritmetice.

Mulţi ani s-a crezut că metoda eliminării lui Gauss este optimă în sensul că orice altă metodă directă de rezolvare a sistemelor din *n* ecuaţii liniare necesită cel puţin $\frac{n^3}{3}$ operaţii aritmetice. În momentul de faţă se cunosc metode în care numărul de operaţii este redus la Cn^{α} $(2<\alpha<3)$. Aceste metode se bazează pe un rezultat remarcabil obţinut în 1971 de către A. Schonhage şi V. Strassen care au arătat că , teoretic, înmulţirea poate avea o complexitate numai cu puţin 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,

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

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

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

Ţinând seama că $N(2)=7+18$, ultima relaţie 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 obişnuit de înmulţire 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 obţine 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 obţine din matricea unitate prin înlocuirea elementelor subdiagonale din coloana întâia cu multiplicatorii $-\mu_{21}, -\mu_{31}$. Se verifică uşor că

$$M_1A=\begin{pmatrix}1&0&0\\-2&1&0\\1&0&1\end{pmatrix}\begin{pmatrix}2&1&1\\4&1&0\\-2&2&1\end{pmatrix}=\begin{pmatrix}2&1&1\\0&-1&-2\\0&3&2\end{pmatrix}=A^{(2)},$$

$$M_1b=\begin{pmatrix}1&0&0\\-2&1&0\\1&0&1\end{pmatrix}\begin{pmatrix}1\\-1\\7\end{pmatrix}=\begin{pmatrix}1\\-4\\8\end{pmatrix}=b^{(2)}.$$

În etapa finală de transformare necunoscuta x_2 va fi eliminată din ultima ecuaţie, obţinându-se un sistem sub formă triunghiulară $Ux=c$, unde

$$U=\begin{pmatrix}2&1&1\\0&-1&-2\\0&0&-4\end{pmatrix},\ c=\begin{pmatrix}1\\-4\\-4\end{pmatrix}.$$

Se observă că $U=M_2A^{(2)}$ şi $c=M_2b^{(2)}$, 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 înmulţirea sistemului iniţial succesiv la matricele M_1, M_2:

$$M_2M_1Ax=M_2M_1b\,.$$

Relaţia $M_2M_1A=U$ permite de a da o altă interpretare metodei lui Gauss. Multiplicând această relaţie la stânga cu matricea inferior triunghiulară $L=M_1^{-1}M_2^{-1}$ obţinem:

$$A=LU\,.$$

Deci, cu ajutorul metodei eliminării a lui Gauss matricea *A* se descompune în produsul de doi factori *L* şi *U*, unde *L* este o matrice inferior triunghiulară, iar *U* este o matrice superior

triunghiulară. Această descompunere se numeşte *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

$$m_k = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \mu_{k+1,k} \\ \vdots \\ \mu_{nk} \end{pmatrix} \begin{matrix} \\ \\ \} k \\ \\ \\ \\ \\ \end{matrix}, \quad e_k = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow \text{componenta} k$$

unde multiplicatorii $\mu_{ik}, i = k+1, k+2, \ldots n$ sunt cei utilizaţi 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*:

$$M_k = \begin{pmatrix} 1 & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ 0 & 0 & \dots & -\mu_{k+1,k} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & -\mu_{nk} & \dots & 1 \end{pmatrix}.$$

Metoda eliminării a lui Gauss constă (vezi paragraful 3.3) în determinarea şirului de matrice $A = A^{(1)}, A^{(2)}, \dots, A^{(n)}$. Se constată uşor că

$$A^{(k+1)} = M_k \cdot M_{k-1} \cdot \dots \cdot M_1 A, k = 1,2,\dots n-1.$$

Scriind această relaţie pentru $k = n-1$ şi notând $U = A^{(n)}$, obţinem:

$$M_{n-1} \cdot M_{n-2} \cdot \dots \cdot M_2 \cdot M_1 \cdot A = U,$$

sau

$$A = M_1^{-1} \cdot M_2^{-1} \cdot \dots \cdot M_{n-1}^{-1} \cdot U.$$

Se verifică imediat că

$$M_k^{-1} = I + m_k e_k^T.$$

De mai sus deducem:

$$A = L \cdot U,$$

unde

$$L = M_1^{-1} \cdot M_2^{-1} \cdot \dots \cdot M_{n-1}^{-1} = I + n-1\sum_{k=1} m_k e_k^T.$$

Prin urmare, metoda eliminării lui Gauss calculează o factorizare *LU* a matricei *A*, unde

$$L=\begin{pmatrix} & & & & & \\ 1 & 0 & ... & 0 & ... & 0 \\ \mu_{21} & 1 & ... & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ \mu_{k1} & \mu_{k2} & .... & 1 & ... & 0 \\ \mu_{k+1,1} & \mu_{k+1,2} & ... & \mu_{k+1,k} & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ \mu_{n1} & \mu_{n2} & ... & \mu_{nk} & ... & 1 \end{pmatrix},$$

$$U=\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & ... & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & ... & a_{2n}^{(2)} \\ .... & ..... & ... & ... \\ \vdots & \vdots & .... & \vdots \\ 0 & 0 & ... & a_{nn}^{(n)} \end{pmatrix}.$$

După cum s-a arătat în paragraful 3.3 din motive de stabilitate este indicat să utilizăm o strategie de pivotare parţială. Se demonstrează că orice matrice admite o factorizare *LU*, eventual efectuând asupra liniilor permutările care apar prin pivotare parţială. Cu alte cuvinte, există o matrice de permutare P astfel încât

$$PA=LU.$$

Metoda eliminării lui Gauss şi factorizarea *LU* sunt echivalente. O factorizare *LU* a lui *A* poate fi calculată şi direct, printr-o procedură compactă numită f*actorizarea lui Crout*. Această factorizare impune *U* cu diagonala unitate şi este mai avantajoasă pe calculatoarele ce permit calcularea rapidă a produselor scalare. Pentru o iniţiere mai aprofundată în factorizarea lui Crout recomandăm lucrările [2,34,38].

Presupunem că se cunoaşte o factorizare *LU* a matricei *A*. atunci sistemul de ecuaţii 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 *substituţie "înainte"* începând cu prima ecuaţie:

$$y_1 = \frac{b_1}{l_{11}}, \quad y_i = \frac{b_i - \sum_{k-1}^{i-1} l_{ik} y_k}{l_{ii}}, i = 3,3,\ldots,n.$$

Aici l_{ik} sunt elementele matricei *L*. Apoi se rezolvă sistemul superior triunghiular $Ux = y$ prin procedura de *substituţie "înapoi"*, începând cu ultima ecuaţie:

$$x_n = \frac{y_{nn}}{u_{nn}}, \quad x_i = \frac{y_i - \sum_{k-i+1}^{n} u_{ki} x_k}{u_{ii}}, i = n-1, n-2,\ldots,1,$$

unde u_{ik} sunt elementele matricei *U*.

Utilizând o factorizare *LU* a lui *A* putem rezolva simultan mai mult sisteme de ecuaţii, având aceeaşi matrice *A*, fără a relua calculele de la început.

Metoda eliminării lui Gauss ne permite să calculăm şi determinantul matricei *A*. Într-adevăr, să observăm că

$$\det(A) = \det(L) \times \det(U).$$

Deoarece $\det(L) = 1$, avem:

$$\det(A) = a_{11}^{(1)} \cdot a_{22}^{(2)} \cdot \ldots \cdot a_{nn}^{(n)},$$

adică determinantul este produsul elementelor pivoţi. Dacă se aplică una din strategiile de pivotare atunci:

$$\det(A) = (-1)^m a_{11}^{(1)} \cdot a_{22}^{(2)} \cdot \ldots \cdot a_{nm}^{(n)}$$

unde *m* este numărul total de permutări efectuate.

3.5 Factorizarea Cholesky

Fie $A=(a_{ij})_{n\times n}$ o matrice simetrică şi pozitiv definită, adică:

$$(Ax,x)>0,\quad \forall x\in R^n, x\neq 0.$$

Vom arăta că în factorizarea LU a lui A se poate alege $U=L^T$. Descompunerea

$$A=L\cdot L^T$$

se numeşte *factorizarea Cholesky*.

Teoremă. Dacă matricea A este simetică şi pozitiv definită, atunci există o matrice inferior triunghiulară L, cu elementele diagonale pozitive, unică, astfel încât $A=L\cdot L^T$.

Demonstraţie. Vom demonstra teorema prin inducţie asupra lui n. Prntru $n=2$ 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 $L\cdot L^T$ şi-l identificăm cu A, obţinem:

$$a_{11}=l_{11}^2,\qquad a_{12}=l_{11}l_{21},$$
$$a_{21}=l_{21}l_{11},\quad a_{22}=l_{21}^2+l_{22}^2.$$

Deoarece matricea A este pozitiv definită, elementele de pe diagonală sunt pozitive şi deci putem extrage rădăcină pătrată: $l_{11}=\sqrt{a_{11}}$ .Al doilea element l_{21} se determină din ecuaţia ce conţine 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. Intr-adevăr, fie vectorul x cu componentele a_{12} şi $-a_{11}$:

$$x = \begin{pmatrix} a_{12} \\ -a_{11} \end{pmatrix}.$$

Atunci $(Ax, x) = x^T Ax = a_{11}^2 a_{22} - a_{12}^2 a_{11} > 0$ fiindcă matricea A este pozitiv definită. Împărţind ultima inegalitate la $a_{11}^2 > 0$ obţinem

$$a_{22} - \frac{a_{12}^2}{a_{11}} > 0 .$$

Prin urmare pentru $n = 2$ descompunerea $A = L \cdot L^T$ există şi este unică. Să presupunem teorema adevărată pentru $n = k-1$: $A = L \cdot L^T$ unde A şi L sunt matrice de dimensiune $(k-1)\times(k-1)$. Fie A' şi 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* şi *w* sunt vectorii coloană ,având $k-1$ componente .

Formăm produsul $L' \cdot (L')^T$ şi-l identificăm cu A'. Atunci obţinem:

$$A = LL^T, \qquad y = Lw,$$
$$y^T = w^T L^T, \quad a_{kk} = l_{kk}^2 + w^T w.$$

Prin presupunerea de inducţie matematică, matricea L este determinată în mod unic astfel ca $A = LL^T$. De aici rezultă că vectorul w este şi el unic şi se calculează ca soluţie al sistemului $y = Lw$. Mai departe, elementul l_{kk} se defineşte din formula:

$$l_{kk} = \sqrt{a_{kk} - w^T w} .$$

Arătăm ca şi 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

$$(Ax, x) = x^T Ax = z^T Az - 2z^T y + a_{kk} =$$

$$= -z^T y + a_{kk} = a_{kk} - y^T A^{-1} y = a_{kk} - y^T \left(LL^T\right)^{-1} y =$$

$$= a_{kk} - \left(L^{-1}y\right)^T \left(L^{-1}y\right) = a_{kk} - w^T w > 0.$$

Deci matricea L' cu proprietatea $A' = L' \cdot (L')^T$ este unic determinată şi ***teorema este demonstrată***.

Se poate arăta că factorizarea Cholesky a unei matrice $A = \left(a_{ij}\right)_{n\times n}$ simetrice pozitiv definite necesită aproximativ $n^3/6$ operaţii (înmulţiri şi adunări) şi n extrageri de radical (necesare pentru a calcula elementele diagonale l_{kk}).

Metoda lui Cholesky de rezolvare a sistemelor de ecuaţii liniare se mai numeşte *metoda rădăcinii pătrate* şi 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

$$l_{11} = \sqrt{a_{11}}, \quad l_{i1} = \frac{a_{i1}}{l_{11}}, \quad i = 2,3,\ldots,n;$$

după ce s-au obţinut primele $(k-1)$coloane ale natricei L se calculează coloana k

$$l_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2},$$

$$l_{ik} = \frac{1}{l_{kk}} \left(a_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj} \right), \quad i = k+1,\ldots,n.$$

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 şi pozitiv definite este situat pe diagonală principală. În plus, elementele

diagonale ale matricei A şi elementele l_{ij} ale matricei L satisfac relaţiei:

$$l_{1k}^2 + l_{2k}^2 + \ldots + l_{kk}^2 = a_{kk}, \quad k = 1,2,\ldots,n.$$

Astfel avem o limitare a creşterii elementelor matricei L: orice element l_{ij} nu depăşeşte elementul maxim în modul $|a_{kk}|$.

3.6 Perturbaţii. Numărul de condiţionare

Considerăm sistemul liniar $Ax = b$. Soluţia exactă a sistemului considerat poate fi scrisă sub forma : $x^* = A^{-1}b$. Să presupunem că matricea nesingulară A şi vectorul b suferă perturbaţiile δA şi δb.

Mai întâi să analizăm cazul când perturbăm numai termenul liber. Soluţia perturbată $\tilde{x}$ a sistemului cu partea dreaptă $b + \delta b$ satisface egalitatea:

$$A\tilde{x} = b + \delta b.$$

Obţinem:

$$\tilde{x} - x^* = A^{-1}\delta b,$$

de unde rezultă că

$$\|\tilde{x} - x^*\| \le \|A^{-1}\|\|\delta b\| \tag{3.5}$$

oricare ar fi norma matricială subordonată unei norme vectoriale. Pentru orice A şi b există o perturbaţie 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 aceleiaşi perturbaţii δb să observăm că:

$$\|b\| = \|Ax^*\| \le \|A\|\|x^*\|,$$

sau

$$\frac{1}{\|x^*\|} \le \frac{\|A\|}{\|b\|}.$$

Ţinând seama de aceasta, inegalitatea (3.5) implică

$$\frac{\left\|\tilde{x}-x^*\right\|}{\left\|x^*\right\|} \le \left\|A\right\|\left\|A^{-1}\right\|\frac{\left\|\delta b\right\|}{\left\|b\right\|} \tag{3.6}$$

Presupunem acum că perturbăm elementele lui *A*; atunci soluţia perturbată $\tilde{x}$ va verifica egalitatea:

$$(A+\delta A)\tilde{x} = b.$$

Rezultă că:

$$\tilde{x}-x^* = A^{-1}\delta A\tilde{x}$$

şi obţinem estimarea:

$$\left\|\tilde{x}-x^*\right\| \le \left\|A^{-1}\right\|\left\|\delta A\right\|\left\|\tilde{x}\right\|. \tag{3.7}$$

Această inegalitate o putem pune sub forma:

$$\frac{\left\|\tilde{x}-x^*\right\|}{\left\|\tilde{x}\right\|} \le \left\|A\right\|\left\|A^{-1}\right\|\frac{\left\|\delta A\right\|}{\left\|A\right\|} \tag{3.8}$$

Se observă că atât în (3.6) cât şi (3.8) numărul $\left\|A\right\|\left\|A^{-1}\right\|$ estimează eroarea relativă în soluţie furnizată de δb sau δA. Acest număr se numeşte *număr de condiţionare* al matricei *A* în raport cu normă matriceală considerată şi se notează:

$$cond(A) = \left\|A\right\|\left\|A^{-1}\right\|.$$

Deoarece pentru orice normă matriceală subordonată unei norme vectoriale se îndeplineşte egalitatea $\left\|I\right\| = 1$,avem:

$$1 = \left\|I\right\| = \left\|AA^{-1}\right\| \le \left\|A\right\|\left\|A^{-1}\right\|,$$

şi deci , $cond(A) \ge 1$ oricare ar fi matricea *A*.

Numărul de condiţionare caracterizează efectul maximal al perturbărilor δb şi δA la rezolvarea sistemului $Ax = b$. Dacă numărul de condiţionare *cond (A)* este mare, atunci perturbaţii mici ale lui *A* şi *b* vor produce perturbaţii relativ mari ale lui x^*; în acest caz se spune că matricea *A* este *rău (sau prost) condiţionată.* Matricele cu numărul de condiţionare *cond (A) "*mic" se numesc *bine condiţionate.*

Subliniem că dimensiunea matricei nu are o influenţă directă asupra numărului ei de condiţionare: dacă $A=I$ sau chiar $A=\frac{1}{10}I$ avem $cond(A)=1$. Pentru comparaţie, determinantul matricei nu este un indice adecvat al condiţionării, deoarece valoarea determinantului depinde şi de dimensiunea n a matricei. Dacă A este o matrice aproape singulară încă nu înseamnă că este prost condiţionată. În exemplul $A=\frac{1}{10}I$ avem $\det(A)=10^{-n}$; această matrice ″aproape singulară″ este maxim de bine condiţionată .

Exemplu. Considerăm sistemul de ecuaţii $Ax=b$, unde:

$$A=\begin{pmatrix} 1 & -1 & -1 & \dots & -1 & -1 \\ 0 & 1 & -1 & \dots & -1 & -1 \\ 0 & 0 & 1 & \dots & -1 & -1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}, \quad b=\begin{pmatrix} -1 \\ -1 \\ -1 \\ \vdots \\ -1 \\ 1 \end{pmatrix}.$$

Menţionăm că $\det(A)=1\neq 0$. Sistemul considerat în formă desfăşurată este:

$$\begin{cases} x_1-x_2-x_3-\dots-x_n=-1, \\ \qquad x_2-x_3-\dots-x_n=-1, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \qquad\qquad\qquad x_{n-1}-x_n=-1, \\ \qquad\qquad\qquad\qquad x_n=1. \end{cases} \tag{3.9}$$

Sistemul de ecuaţii (3.9) admite o soluţie exactă unică $x_1^*=x_2^*=\dots=x_{n-1}^*=0, x_n^*=1$. Presupunem că perturbăm vectorul liber b cu $\delta b=(0,0,\dots,0,\varepsilon)^T$. Atunci soluţia exactă al sistemului

perturbat $A\tilde{x} = b + \delta b$ devine $\tilde{x} = x^* + r$ şi eroarea $r = (r_1, r_2, \ldots, r_n)^T$ satisface ecuaţiilor:

$$\begin{cases} r_1 - r_2 - r_3 - \ldots - r_n = 0, \\ \qquad r_2 - r_3 - \ldots - r_n = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad\qquad\qquad r_{n-1} - r_n = 0, \\ \qquad\qquad\qquad\qquad r_n = \varepsilon. \end{cases}$$

De unde rezultă, că:

$$r_n = \varepsilon,$$
$$r_{n-1} = r_n = \varepsilon,$$
$$r_{n-2} = r_n + r_{n-1} = 2\varepsilon,$$
$$r_{n-3} = r_n + r_{n-1} + r_{n-2} = 4\varepsilon = 2^2\varepsilon,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$r_{n-k} = r_n + r_{n-1} + \ldots + r_{n-(k+1)} = 2^{k-1}\varepsilon,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$r_1 = r_{n-(n-1)} = 2^{(n-1)-1}\varepsilon = 2^{n-2}\varepsilon.$$

Astfel $\tilde{x}_i = 2^{n-i-1}\varepsilon, \quad i = 1,2,\ldots,n-1; \quad \tilde{x}_n = 1 + \varepsilon$.

Să estimăm *cond(A)*, luând în calitate de normă $\|.\|_\infty$. Vom avea:

$$\left\|\tilde{x} - x^*\right\|_\infty = \|r\|_\infty = 2^{n-2}\varepsilon, \quad \left\|x^*\right\|_\infty = 1, \quad \|\delta b\|_\infty = \varepsilon, \quad \|b\|_\infty = 1.$$

Prin urmare

$$cond(A) = \|A\|_\infty \cdot \left\|A^{-1}\right\|_\infty \geq \frac{\|r\|_\infty}{\left\|x^*\right\|_\infty} \cdot \frac{\|b\|_\infty}{\|\delta b\|_\infty} = 2^{n-2}.$$

Deoarece $\|A\|_\infty = n$, rezultă că norma matricei inverse este destul de mare, cu toate că $\det(A^{-1}) = \dfrac{1}{\det(A)} = 1$. De exemplu pentru $n = 102$ avem:

$$\|A\|_\infty = 102, \quad cond(A) \geq 2^{100} > 10^{30}, \quad \text{iar } \|A^{-1}\|_\infty > 10^{27}.$$

În particular, dacă $\varepsilon = 10^{-15}$ (o eroare suficient de mică), atunci $\|\tilde{x} - x^*\|_\infty = \|r\|_\infty > 10^{15}$; o perturbaţie destul de mică a termenului liber a produs o perturbaţie atât de mare în soluţie!

Valoarea numărului de condiţionare al unei matrice depinde de norma matriceală întrebuinţată. Presupunem că *A* este o matrice simetrică şi pozitiv definită cu valorile proprii pozitive ordonate astfel : $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. În mod analog se arată că pentru astfel de matrice numărul de condiţionare este:

$$cond(A) = \frac{\lambda_n}{\lambda_1} = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Exemplu:Valorile proprii ale matricei:

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1.0001 \end{pmatrix}$$

aproximativ sunt egale cu $\lambda_1 = \dfrac{1}{2}10^{-4}$ şi $\lambda_2 = 2$ .De aceea numărul de condiţionare $cond(A)$ este aproape egal cu $4 \cdot 10^4$. Ne putem aştepta că perturbaţii mici ale datelor iniţiale vor produce mari schimbări în soluţie. Într-adevăr, fie sistemele de ecuaţii $Ax = b$ şi $Ax = b + \delta b$ unde

$$b = \begin{pmatrix} 2 \\ 2.0001 \end{pmatrix}, \qquad \delta b = \begin{pmatrix} 0 \\ 0.0001 \end{pmatrix}.$$

Soluţia se schimbă de la $x^* = (1 \quad 1)^T$ până la $\tilde{x} = (0 \quad 2)^T$:

$$\frac{\left\|x^* - \tilde{x}\right\|_2}{\left\|x^*\right\|_2} = \frac{\left\|\begin{pmatrix} -1 \\ 1 \end{pmatrix}\right\|_2}{\left\|\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right\|_2} = 1, \quad \frac{\left\|\delta b\right\|_2}{\left\|b\right\|_2} = \frac{\left\|\begin{pmatrix} 0 \\ 0.0001 \end{pmatrix}\right\|_2}{\left\|\begin{pmatrix} 2 \\ 2.0001 \end{pmatrix}\right\|_2} = \frac{10^{-4}}{2\sqrt{2}}.$$

Fie dat un sistem de ecuaţii liniare. Reprezentarea cu virgulă mobilă a elementelor A şi B în calculatorul electronic nu este exactă. Prin urmare, efectiv în memoria maşinii electronice de calcul vom avea sistemul $\tilde{A}x = \tilde{b}$ unde $\tilde{A}$ şi $\tilde{b}$ sunt rotunjirile corespunzătoare. Există matricele de perturbare P şi 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 maşinii (vezi paragraful 1.3), atunci $\|P\| \le \varepsilon_M$ şi $\|D\| \le \varepsilon_M$. Astfel obţinem:

$$\|\delta A\| = \left\|\tilde{A} - A\right\| \le \varepsilon_M \|A\|, \quad \|\delta b\| = \left\|\tilde{b} - b\right\| \le \varepsilon_M \|b\|.$$

Din inegalităţile de mai sus şi din (3.7) rezultă că erorile de rotunjire produc o perturbaţie estimată prin:

$$\left\|\tilde{x} - x^*\right\| \le 2\varepsilon_M \|\tilde{x}\| cond(A).$$

Subliniem că determinarea numărului de condiţionare $cond(A)$ este o problemă dificilă, deoarece conţine calculul lui $\left\|A^{-1}\right\|$. Calculul matricei inverse şi normei sale necesită aproximativ $n^3 + 2n^2$ operaţii suplimentare şi aproape de patru ori majorează cheltuielile necesare pentru rezolvarea sistemului $Ax = b$. Un procedeu practic de calcul aproximativ al lui $\left\|A^{-1}\right\|$ constă în următoarele. Se observă că dacă $w = A^{-1}y$ atunci $\|w\| \le \left\|A^{-1}\right\| \|y\|$ şi prin urmare

$$\left\|A^{-1}\right\| \geq \frac{\|w\|}{\|y\|}.$$

Deci, se poate alege k vectori y_i, apoi se rezolvă sistemul de ecuaţii $Aw_i = y_i$, $i = 1,2,\ldots,k$, şi se pune

$$\left\|A^{-1}\right\| \approx \max_{1\leq i\leq k} \frac{\|w_i\|}{\|y_i\|}.$$

3.7 Calculul valorilor şi vectorilor proprii

3.7.1. Formularea problemei. Proprietăţi fundamentale

Fie A o matrice de dimensiune $n \times n$. Numărul λ (real sau complex) se numeşte *valoare proprie* a matricei A dacă există un vector nenul $x \in R^n$, astfel încât

$$Ax = \lambda x \tag{3.10}$$

Vectorul $x \neq 0$ se numeşte *vector propriu* al lui A asociat valorii proprii λ.

Ecuaţia (3.10) poate fi rescrisă sub forma $Ax - \lambda x = 0$ sau

$$(A - \lambda I)x = 0, \tag{3.11}$$

unde

$$\lambda I = \begin{pmatrix} \lambda & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ .. & ... & .. & .. \\ .0 & 0 & 0 & \lambda \end{pmatrix}, \quad 0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Ecuaţia (3.11) poate fi scrisă dezvoltat astfel:

$$\begin{pmatrix} a_{11} - \lambda & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} - \lambda & ... & a_{21} \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a - \lambda_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Relaţia de mai sus este un sistem omogen de ecuaţii liniare. Condiţia necesară şi suficientă ca acest sistem omogen să admită soluţie nenulă este ca:

$$\det(A-\lambda I)=0\,. \tag{3.12}$$

Acest determinant este un polinom de grad n cu coeficienţi reali şi se notează de obicei:

$$P_n(\lambda)=(-1)^n\lambda^n+p_1\lambda^{n-1}+\ldots+p_{n-1}\lambda+p_n\,.$$

Polinomul $P_n(\lambda)$ se numeşte *polinom caracteristic* asociat matricei *A*, iar ecuaţia (3.12) se numeşte *ecuaţia caracteristică* a matricei *A*.

Ecuaţia caracteristică este o ecuaţie algebrică de grad *n* cu coeficienţi reali care în virtutea teoremei fundamentale a algebrei are exact *n* rădăcini $\lambda_1,\lambda_2,\ldots,\lambda_n$, în general complexe şi nu neapărat distincte.

Mulţimea valorilor proprii ale matricei *A* se numeşte *spectrul* lui *A* şi se notează cu $\sigma(A)$:

$$\sigma(A)=(\lambda_1,\lambda_2,\ldots,\lambda_n).$$

Raza spectrală a lui *A* se notează $\rho(A)$ şi se defineşte prin relaţia:

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

De aceea norma matriceală $\|A\|_2=\sqrt{\rho(A^TA)}$ subordonată normei euclidiene $\|x\|_2=\sqrt{\sum x_t^{\,2}}$ se mai numeşte *normă spectrală.*

Pentru orice normă matriceală subordonată unei norme vectoriale avem:

$$\rho(A)\le\|A\|\,.$$

Mai mult ca atât:

$$\|A\|_2\le\|A\|_F\le\sqrt{n}\|A\|_2\,,$$

unde $\|A\|_F$ este norma lui Frobenius (definiţia vezi în paragraful 3.1.2.) .

Dacă *A* este o matrice simetrică, adică $A=A^T$, atunci

$$\|A\|_2 = \max_{\lambda \in \sigma(A)} |\lambda| = \rho(A)$$

şi

$$\|A\|_F = \sqrt{\sum_{i=1}^{n} \lambda_i^2}\ .$$

Dacă se cunoaşte o valoare proprie atunci un vector propriu asociat acestei valori proprii este soluţia nenulă a sistemului omogen (2). Pe de altă parte dacă se ştie un vector propriu *v* atunci $Av = \lambda v$, de unde $\lambda(v,v) = (Av,v)$: deci valoarea proprie corespunzătoare se obţine imediat:

$$\lambda = \frac{(Av,v)}{(v,v)}$$

Exemplu: Să se calculeze valorile proprii şi vectorii proprii pentru matricea:

$$A = \begin{pmatrix} 4 & -5 \\ 2 & -3 \end{pmatrix}.$$

Polinomul caracteristic este:

$$P_2(\lambda) = \det(A - \lambda I) = \begin{vmatrix} 4-\lambda & -5 \\ 2 & -3-\lambda \end{vmatrix} = .$$

$$= (4-\lambda)(-3-\lambda) + 10 = \lambda^2 - \lambda - 2$$

Se obţine: $\lambda_1 = -1,\ \lambda_2 = 2.$ Prin urmare matricea *A* are două valori proprii distincte. Înlocuind fiecare valoare proprie în sistemul omogen (3.11) obţinem 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 :

$$v^1 = c \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad c = const \neq 0.$$

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 şi superior triunghiulare coincid cu elementele de pe diagonala principală.

Proprietăţile 1-3 se demonstrează imediat, reieşind din definiţia valorilor şi vectorilor proprii.

4. Matricele asemenea au aceleaşi valori proprii .

Demonstraţie: Fie matricele A şi 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

$$M^{-1}AM=\lambda y \text{ şi } By=\lambda y.$$

Prin urmare la matricele asemenea valorile proprii coincid. În plus, vectorii proprii sunt legaţi prin relaţia $x=My$.

5. Suma valorilor proprii ale unei matrice A este egală cu suma elementelor de pe diagonala principală:

$$\lambda_1+\lambda_2+....+\lambda_n=a_{11}+a_{22}+...+a_{nn}.$$

Această sumă se numeşte *urma matricei* A şi se notează $Tr(A)$ sau $Sp(A)$.

6. Produsul valorilor proprii coincide cu determinantul matricei:

$$\lambda_1 \cdot \lambda_2 \cdot \ldots \cdot \lambda_n = \det(A).$$

7. Dacă $\lambda_i, 1 \le i \le n$ sunt valorile proprii ale matricei A, atunci $\lambda_i^k, 1 \le i \le n$, sunt valorile proprii ale matricei

$$A^k = A \cdot A \cdot \ldots \cdot A.$$

Notăm

$$Q_n(\lambda) = (-1)^n P_n(\lambda) = \lambda^n + q_1\lambda^{n-1} + \ldots + q_{n-1}\lambda + q_n.$$

Polinomul $Q_n(\lambda)$ se numeşte *polinom propriu*.

8. Formulele lui Newton:

$$\mu_k + \sum_{i=1}^{k-1} q_i \mu_{k-1} = -kq_k, \quad k = 1,2,\ldots,n,$$

unde

$$\mu_k = \sum_{i=1}^{n} \lambda_i^k, \quad k = 1,2,\ldots,n,$$

iar $\lambda_1, \lambda_2, \ldots, \lambda_n$ sunt rădăcinile ecuaţiei caracteristice .

Se observă că $\mu_k = Tr(A^k)$. Suma $\sum_{i=1}^{n} \lambda_i^k$ se mai numeşte *moment de ordinul k* al valorilor proprii.

9. ***Identitatea lui Cayley - Hamilton***. Orice matrice pătrată *A* este o rădăcină a polinomului său caracteristic:

$$(-1)^n A^n + p_1 A^{n-1} + \ldots + p_{n-1} A + p_n I = 0.$$

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 :

$$\bigcup_{i=1}^{n} \left\{ z : |z - a_{ii}| \le r_i; \quad r_i = \sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{ii}| \right\}.$$

11. ***Teorema (Rayleigh-Ritz)***. Fie *A* o matrice simetrică şi valorile proprii ale lui *A* sunt aşezate în ordine crescătoare

$$\lambda_{\min} = \lambda_1 \le \lambda_2 \le \ldots \le \lambda_{n-1} \le \lambda_n = \lambda_{\max}.$$

Atunci

$$\lambda_1 \|x\|_2^2 \le (Ax, x) \le \lambda_n \|x\|_2^2, \quad \forall x \in R^n,$$

$$\lambda_1 = \lambda_{\min} = \min_{x \neq 0} \frac{(Ax, x)}{(x, x)} = \min_{\|x\|_2 = 1} (Ax, x),$$

$$\lambda_n = \lambda_{\max} = \max_{x \neq 0} \frac{(Ax, x)}{(x, x)} = \max_{\|x\|_2 = 1} (Ax, x).$$

Expresia $\frac{(Ax, x)}{(x, x)}$ se numeşte *câtul lui Rayleigh.*

Metodele de calcul ale valorilor proprii se împart în două grupe:

1) metode care determină întâi coeficienţii polinomului caracteristic şi apoi se rezolvă ecuaţia caracteristică;
2) metode care determină valorile proprii şi vectorii proprii prin procedee iterative fără a calcula coeficienţii polinomul caracteristic .

Se cunosc metode de determinare a coeficienţilor 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 şi rădăcinile ecuaţiei caracteristice sunt bine separate. Motivul este că numărul de operaţii aritmetice pentru determinarea coeficienţilor polinomului caracteristic este foarte mare (de exemplu, metoda lui Fadeev necesită n^4 operaţii). Pe de altă parte, coeficienţii se obţin cu erori de rotunjire inerente care pot conduce la variaţii mari ale rădăcinilor, deoarece problema rezolvării ecuaţiilor algebrice este rău condiţionată .

Fie ,de exemplu ,matricea *A* de forma [2,28] :

$$A(\varepsilon)=\begin{pmatrix} \alpha & 1 & 0 & \dots & 0 & 0 \\ 0 & \alpha & 1 & \dots & 0 & 0 \\ 0 & 0 & \alpha & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \varepsilon & 0 & 0 & \dots & 0 & \alpha \end{pmatrix},$$

unde ε este o perturbaţie mică iar α valoarea proprie de multiplicitate n a matricei neperturbate $A(0)$. Polinomul caracteristic al lui $A(\varepsilon)$este

$$P_n(\varepsilon,\lambda)=(\lambda-a)^n-(-1)^n\varepsilon\,.$$

Se observă că, de exemplu, dacă $n=10,\alpha=0,$şi $\varepsilon'10^{-10}$, atunci practic matricele $A(0)$şi $A(10^{-10})$ reprezintă una şi aceeaşi matrice în memoria maşinii 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 perturbaţia ε care a produs-o.

3.7.2. Metode bazate pe transformări de asemănare ortogonală

Metodele practice de determinare a valorilor şi vectorilor proprii constau în proceduri iterative de aducere a matricei considerate la forma canonică Schur prin transformări de asemănare ortogonală.

Două matrice *A* şi *B* se numesc *ortogonal asemenea* dacă există o matrice ortogonală *Q* astfel încât

$$B=Q^TAQ$$

Transformarea Q^TAQ a lui *A* se numeşte *transformare de asemănare ortogonală*.

Matricele A şi $Q^T AQ$ au aceleaşi 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 $Q^T AQ$.

Teoremă (Schur). Oricare ar fi matricea $A = (a_{ij})_{n\times n}$ există o matrice ortogonală Q de dimensiune $n\times n$ astfel încât :

$$Q^T AQ = S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1k} \\ 0 & s_{22} & \dots & s_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & s_{11} \end{pmatrix},$$

unde $s_{ii}, i = 1,2,\dots,k$, 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.

***Demonstraţi**a* acestei teoreme poate fi găsită în [2,40].

Matricea S se numeşte *formă canonică Schur reală* a lui A.

Dacă matricea A are numai valori proprii reale atunci matricea S este superior triunghiulară . Dacă A are şi valori proprii complexe atunci S este *cvasi-triunghiulară* .

Stabilirea unei forme canonice Shur este precedată de o pregătire iniţială a matricei A, aducând-o la forma compactă numită *forma Hessenberg:*

$$A = \begin{pmatrix} h_{11} & h_{12} & h_{13} & \dots & h_{1,n-1} & h_{1n} \\ h_{21} & h_{22} & h_{23} & \dots & h_{2,n-1} & h_{2n} \\ 0 & h_{32} & h_{33} & \dots & h_{3,n-1} & h_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & h_{n-1,n-1} & h_{n-1,n} \\ 0 & 0 & 0 & \dots & h_{n,n-1} & h_{nn} \end{pmatrix}$$

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 rotaţii 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 numeşte *reflector* sau *transformarea lui Householder*.

Fie $z = (1,0,\ldots,0)^T \in R^n$, $\sigma = \|x\|_2$ şi $v = x + \sigma z$. Atunci

$$Hx = -\sigma z = (-\sigma,0,\ldots,0)^T .$$

Într-adevăr,

$$H = x - 2\frac{vv^T x}{\|v\|_2^2} = x - (x+\sigma z)\frac{2(x+\sigma z)^T x}{(x+\sigma z)^T (x+\sigma z)} = ,$$

$$= x - (x+\sigma z) = -\sigma z,$$

deoarece $x^T x = \sigma^2$.

Procedura directă de adunare a matricei *A* la forma lui Householder constă din *n-2* etape şi 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

$$x = \begin{pmatrix} a_{21} \\ a_{31} \\ \vdots \\ a_{n1} \end{pmatrix}, \quad z = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad Hx = \begin{pmatrix} -\sigma \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Aici *H* este matricea lui Householder de dimensiune $(n-1)\times(n-1)$ Calculul lui U_1 se face conform relaţiei :

$$U_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & H & \\ 0 & & & \end{pmatrix} = U_1^T = U_1^{-1}.$$

Deoarece în colţul "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 $j = 2,3,\dots,n$. După ce matricea U_1A a fost obţinută, 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

$$U_1^{-1}AU_1 = \begin{pmatrix} \alpha_{11} & * & \dots & * \\ -\sigma & * & \dots & * \\ 0 & * & \dots & * \\ \vdots & \vdots & \vdots & \vdots \\ 0 & * & \dots & * \end{pmatrix}$$

este de formă Hessenberg în prima coloană. Prima etapă este complet terminată.

Etapa 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 obţinem :

$$U_2 = U_2^{-1} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & & & \\ \vdots & \vdots & & H_2 & \\ 0 & 0 & & & \end{pmatrix}, U_2(U_1AU_1)U_2 = \begin{pmatrix} * & * & * & \dots & * \\ * & * & * & \dots & * \\ 0 & * & * & \dots & * \\ 0 & 0 & * & \dots & * \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & * & \dots & * \end{pmatrix}$$

Astfel matricea:

$$U_{n-2}U_{n-3}\ldots U_1AU_1U_2\ldots U_{n-3}U_{n-2}$$

Obţinută în final, după parcurgerea celor $n-2$ etape ale procedurii, este de formă Hessenberg. Numărul de operaţii aritmetice necesar este egal cu $5n^3/3$.

Dacă aplicăm transformările descrise unei matrice simetrice, atunci toate matricele $U_kU_{k-1}\ldots U_1AU_1U_2\ldots U_{k-1}U_k$ vor fi de asemenea simetrice şi în final obţinem o matrice Hessenberg simetrică tridiagonală. În acest caz este nevoie doar de $2n^3/3$ operaţii 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* construieşte iterativ un şir de matrice ortogonal asemenea $A_0=A, A_1, A_2,\ldots$ 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 $A_0 = A$ şi 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 obţine multiplicând factorii Q_0 şi R_0 în ordine inversă :

$$A_1 = R_0 Q_0$$

Matricele A_0 şi 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 şirul matriceal $A_0, A_1, A_2, \ldots$ definit recurent prin formulele:

$$\begin{aligned} &A_0 = Q_0 R_0, \quad A_1 = R_0 Q_0, \\ &A_1 = Q_1 R_1, \quad A_2 = R_1 Q_1, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &A_k = Q_k R_k, \quad A_{k+1} = R_k Q_k. \end{aligned}$$

Se poate demonstra (vezi, de exemplu, [2,22,26,32]) că în condiţii destul de generale , şirul $\{A_k\}_{k=0}^{\infty}$ converge şi 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 convergenţa este înceată. În acest caz algoritmul QR se modifică astfel:

$$A_k - \alpha_k I = Q_k R_k, \quad A_{k+1} = R_k Q_k + \alpha_k I, \quad k = 0,1,\ldots.$$

unde α_k este un parametru de accelerare a convergenţei 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 convergenţă rapidă a şirului $\{A_k\}$ către forma canonică Schur.

O iteraţie al algoritmului *QR* necesită $\frac{4n^3}{3}$ operaţii,ceea ce este exagerat. În practică se recomandă iniţial de-a aduce matricea considerată la forma Hessenberg prin metoda lui Householder; apoi asupra formei Hessenberg se aplică algoritmul *QR*. Astfel numărul de operaţii la o iteraţie se reduce la $4n^2$, în cazul matricelor nesimetrice, şi 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 $n \times n$ şi a vectorului propriu corespunzător .

Fie *A* o matrice *simplă*. Matricea *A* de dimensiune $n \times n$ se numeşte *simplă* dacă are exact *n* vectori proprii liniar independenţi $e_1, e_2, \ldots, e_n$. Aceşti vectori formează o bază a spaţiului *n*-dimensional şi pot fi aleşi astfel încât:

$$\|e_i\|_2 = \sqrt{(e_i, e_i)} = 1, \quad i = 1,2,\ldots,n.$$

De exemplu, orice matrice simetrică este o matrice *simplă*. Vom presupune că matricea *A* admite o valoare proprie reală dominantă λ_1 adică

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n| .$$

Alegem un vector iniţial arbitrar $x^{(0)}$ nenul. Deoarece vectorii proprii $e_1, e_2, \ldots, e_n$ formează o bază, există scalarii $c_1, c_2, \ldots, c_n$ nu toţi nuli , astfel încât:

$$x^{(0)} = c_1 e_1 + c_2 e_2 + \ldots c_n e_n .$$

Vom presupune că $c_1 \neq 0$; în caz contrar se poate alege alt vector iniţial $x^{(0)}$ astfel încât coeficientul corespunzător $c_1 \neq 0$.

Recurent formăm şirul de vectori:

$$x^{(k)} = Ax^{(k-1)} = A^k x^{(0)}, \quad k = 1,2,\ldots$$

Deoarece $Ae_i = \lambda_i e_i$ putem scrie:

$$\begin{aligned} x^{(1)} &= Ax^{(0)} = A\left(c_1 e_1 + c_2 e_2 + \ldots c_n e_n\right) = \\ &= c_1 Ae_1 + c_2 Ae_2 + \ldots + c_n Ae_n = \\ &= c_1\lambda_1 e_1 + c_2\lambda_2 e_2 + \ldots + c_n\lambda_n e_n. \end{aligned}$$

În general şirul $\left\{x^{(k)}\right\}$ are reprezentarea:

$$x^{(k)} = c_1\lambda_1^k e_1 + c_2\lambda_2^k e_2 + \ldots + c_n\lambda_n^k e_n = \lambda_1^k\left(c_1 e_1 + \eta^{(k)}\right).$$

unde

$$\eta^{(k)} = c_2\left(\frac{\lambda_2}{\lambda_1}\right)^k e_2 + \ldots + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^k e_n.$$

Prin ipoteză $\left|\frac{\lambda_i}{\lambda_1}\right| < 1$ pentru $i \geq 2$; deci $\left\|\eta^{(k)}\right\|_2 = 0\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$ şi

$$\lim_{k\to\infty}\left\|\eta^{(k)}\right\|_2 = 0.$$

Calculăm produsul scalar:

$$\begin{aligned} &\left(Ax^{(k-1)}, x^{(k-1)}\right) = \left(x^{(k)}, x^{(k-1)}\right) = \lambda_1^{2k-1}\left(c_1 e_1 + \eta^{(k)}, c_1 e_1 + \eta^{(k-1)}\right) = \\ &= \lambda_1^{2k-1}\left[c_1^2\left(e_1, e_1\right) + c_1\left(e_1, \eta^{(k-1)}\right) + c_1\left(\eta^{(k)}, e_1\right) + \left(\eta^{(k)}, \eta^{(k-1)}\right)\right] \end{aligned}$$

Deoarece $\left(e_1, e_1\right) = \left\|e_1\right\|_2^2 = 1$ şi

$$\begin{aligned} \left|\left(e_1, \eta^{(k-1)}\right)\right| &\leq \left\|e_1\right\|_2\left\|\eta^{(k-1)}\right\|_2 = \left\|\eta^{(k-1)}\right\|_2, \\ \left|\left(\eta^{(k)}, e_1\right)\right| &\leq \left\|\eta^{(k)}\right\|_2\left\|e_1\right\|_2 = \left\|\eta^{(k)}\right\|_2, \\ \left|\left(\eta^{(k)}, \eta^{(k-1)}\right)\right| &\leq \left\|\eta^{(k)}\right\|_2\left\|\eta^{(k-1)}\right\|_2, \end{aligned}$$

deducem că

$$\left(Ax^{(k-1)}, x^{(k-1)}\right) = \lambda_1^{2k-1}\left[c_1^2 + o\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{k-1}\right)\right].$$

În mod analog obţinem:

$$\left(x^{(k-1)}, x^{(k-1)}\right) = \lambda_1^{2k-2}\left[c_1^2 + o\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{k-1}\right)\right].$$

Prin urmare câtul lui Rayleigh al lui $x^{(k-1)}$

$$\frac{\left(Ax^{(k-1)}, x^{(k-1)}\right)}{\left(x^{(k-1)}, x^{(k-1)}\right)} = \lambda_1 + o\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{k-1}\right)$$

va tinde către valoarea proprie (maximă în modul) λ_1.

Se observă că

$$\left\|x^{(k)}\right\|_2 = \sqrt{\left(x^{(k)}, x^{(k)}\right)} = \left|\lambda_1\right|^k\left[\left|c_1\right| + o\left(\left|\frac{\lambda_2}{\lambda_1}\right|\right)^k\right].$$

De aici rezultă că pentru $\left|\lambda_1\right| > 1$ avem $\left\|x^{(k)}\right\|_2 \to \infty$ iar dacă $\left|\lambda_1\right| < 1$ atunci $\left\|x^{(k)}\right\|_2 \to 0$. Rezolvând problema la maşina electronică de calcul în primul caz putem avea apariţia unui semnal de depăşire, după aceea , calculele se întrerup. În cazul al doilea $\left\|x^{(k)}\right\|_2$ poate deveni un zero-maşină. De aceea în practică este necesar să schimbăm factorul de scară al şirului $\left\{x^{(k)}\right\} = \left\{Ax^{(k-1)}\right\}$; în locul şirului $\left\{x^{(k)}\right\}$ se formează un alt şir de vector $\left\{z^{(k)}\right\}$ cu $\left\|z^{(k)}\right\|_2 = 1$ şi construit astfel:

$$z^{(k)} = \frac{Az^{(k-1)}}{\left\|Az^{(k-1)}\right\|}, \quad k = 1,2,\ldots,$$

de unde

$$\frac{\left(Az^{(k-1)}, z^{(k-1)}\right)}{\left(z^{(k-1)}, z^{(k-1)}\right)} = \left(z^{(k)}, z^{(k-1)}\right) \to \lambda_1.$$

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 uşor de urmărit:

A	$x^{(o)}$	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	…	$x^{(\infty)}$
$1\ -1\ -1$	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	$-1/\sqrt{5}$	$3/\sqrt{35}$	$-9/\sqrt{259}$			…	$\frac{-1}{\sqrt{3}}$
	0	0	$1/\sqrt{35}$	$3/\sqrt{259}$			…	$\frac{1}{\sqrt{3}}$
	1	$2/\sqrt{5}$	$5/\sqrt{35}$	$13\sqrt{259}$			…	$\frac{1}{\sqrt{3}}$
		—	3	3	2.8	3	…	3
		—	—	3	3	3	…	3
		2	2.5	2.6	2.69	2.71	…	3

În schema de mai sus, în penultim tabel avem vectorii $z^{(k)}=\dfrac{x^{(k)}}{\left\|x^{(k)}\right\|_2}$ iar în ultimul – câtul $\dfrac{\left(x^{(k)}\right)_i}{\left(x^{(k-1)}\right)_i}$, unde prin $\left(x^{(k)}\right)_i$ s-a notat componenta *i* a vectorului $x^{(k)}$. Valorile proprii ale matricei considerate sunt: $\lambda_1=3, \lambda_2=2$ şi $\lambda_3=0$. Câtul $\dfrac{\left(x^{(k)}\right)_i}{\left(x^{(k-1)}\right)_i}$ tinde către valoarea maximă λ_1, iar şirul de vectori $\left\{z^{(k)}\right\}$ 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 şi pentru calculul valorii proprii minime (în modul) λ_n cu condiţia 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 şirul de vectori $\{z^{(k)}\}$ se formează în felul următor:

$$z^{(k)} = \frac{A^{-1}z^{(k-1)}}{\left\|Az^{(k-1)}\right\|}, \quad k = 1,2,\ldots,$$

Această metodă se numeşte *metoda puterii inverse*. Subliniem că în practică în metoda puterii inverse vectorul $z^{(k)}$ se determină în urma rezolvării sistemului de ecuaţii liniare:

$$Az^{(k)} = \frac{z^{(k-1)}}{\left\|Az^{(k-1)}\right\|}.$$

O altă metodă mult mai eficientă este *metoda puterii inverse cu deplasări*. Observăm că matricea $A - \alpha I$ are valorile proprii $\lambda_i - \alpha$ şi aceeaşi vectori proprii ca şi *A*. De aceea, dacă se aleg deplasările α_m suficient de apropiate de valoarea proprie λ a lui *A*, atunci şirul $\{z^{(k)}\}$ 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 $\left\|z^{(k)}\right\|_2 = 1$. Dacă *A* este inversabilă şi luăm $\alpha_m = 0$ obţinem $Az^{(k)} = \beta_k z^{(k-1)}$ - metoda puterii inverse .

În practică în scopul accelerării convergenţei şirului $\left\{z^{(k)}\right\}$ către vectorul propriu cea mai bună alegere a deplasării α_m pentru matrice simetrice este câtul Rayleigh:

$$\alpha_m = \frac{\left(Az^{(m)}, z^{(m)}\right)}{\left(z^{(m)}, z^{(m)}\right)}.$$

Se poate demonstra că $\left\{\alpha_m\right\}$ converge către λ_1 extrem de rapid (cu viteza cubică de convergenţă).

Aici încheiem discutarea metodelor de calcul al valorilor proprii ăi vectorilor proprii. Subliniem încă o dată că algoritmul *QR* este unul dintre cele mai remarcabile metode ale matematicii aplicate. Cititorului dornic de a-şi aprofunda cunoştinţele în acest domeniu I se recomandă lucrările [2,22,26,28,35,40].

3.8 Metode iterative de rezolvare a sistemelor de ecuaţii liniare

Să considerăm sistemul de ecuaţii liniare:

$$Ax = b \qquad (3.13)$$

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 puţin $n^3/3$ operaţii aritmetice şi atât timp cât acest număr de operaţii este acceptabil putem utiliza această metodă. Pe de altă parte, când $n^3/3$ este mare, cu ajutorul metodelor iterative se poate obţine cu aproximaţie satisfăcătoare a soluţiei după efectuarea unui număr mult mai mic de operaţii aritmetice.

Metodele iterative se construesc utilizând desfacerea matriciei *A* difinită prin:

$$A = S - T.$$

Atunci sistemul (1) este echivalent cu sistemul :

$$Sx = Tx + b, \qquad (3.14)$$

sau

$$x = Qx + d, \qquad (3.15)$$

unde: $Q = S^{-1}T, d = S^{-1}b$. Prin urmare putem construi şirul $\{x^{(k)}\}$ utilizând relaţia recurentă:

$$Sx^{(k+1)} = Tx^{(k)} + b, \quad k = 0,1,2,\ldots, \qquad (3.16)$$

sau

$$x^{(k+1)} = Qx^{(k)} + d, \quad k = 0,1,2,\ldots, \qquad (3.17)$$

unde $x^{(0)} \in R^n$ este o aproximaţie iniţială a soluţiei $x*$.

Pentru a reduce sistemul (3.13) la forma (3.14) sau (3.16), potrivită pentru iteraţie, desfacerea matricei *A* trebuie să satisfacă condiţiile:

a) Sistemul (3.16) are o soluţie unică $x^{(k+1)}$ şi se rezolvă uşor. De aceea matricea *S* se alege de o formă simplă şi este inversabilă. Ea poate fi diagonală sau triunghiulară.

b) Şirul $\{x^{(k)}\}_{k=0}^{\infty}$ converge către soluţia exactă $x*$ oricare ar fi $x^{(0)} \in R^n$.

Deoarece $x* = Qx^* + d$ avem:

$$x^{(k+1)} - x* = Q\left(x^{(k)} - x*\right), \quad k = 0,1,2,\ldots,$$

de unde rezultă că:

$$x^{(k+1)} - x* = Q^k\left(x^{(0)} - x^*\right).$$

Evident că $\lim_{k\to\infty} x^{(k)} = x^*$ dacă şi numai dacă

$$\lim_{k\to\infty} Q^k = O.$$

În cursul algebrei se demonstrează că $Q^k \underset{k\to\infty}{\to} O$, dacă şi numai dacă raza spectrală a lui *Q* este mai mică ca unitatea: $\rho(Q) = \max_{\lambda\in\sigma(Q)} |\lambda| < 1$. Viteza de convergenţă a şirului matriceal $Q, Q^2, Q^3, \ldots, Q^k, \ldots$ către matricea nulă *O*, şi deci a şirului

$\{x^{(k)}\}_{k=0}^{\infty}$ către x^*, este cu atât mai mare cu cât raza spectrală $\rho(Q)$ este mai mică. Rezultă că este adevărată următoarea teoremă.

Teorema 1. (Condiţia necesară şi suficientă de convergenţă). Şirul $\{x^{(k)}\}$ definit prin (5) converge către soluţia unică x^* a sistemului (3) pentru orice aproximaţie iniţială $x^{(0)} \in R^n$ dacă şi numai dacă

$$\rho(Q) = \max_{\lambda \in \sigma(Q)} |\lambda| < 1. \tag{3.18}$$

În practică nu cunoaştem valorile proprii ale lui *Q*, de aceia teorema 1 este dificil de folosit. În locul teoremei 1 se utilizează următoarea teoremă.

Teorema 2. (Condiţia suficientă de convergenţă). Dacă există o normă matriceală subordonată unei norme vectoriale astfel încât $\|Q\| \le q < 1$, atunci sistemul (3.15) are o soluţie unică x^*, şirul $\{x^{(k)}\}$ definit prin (3.17) converge către x^* oricare ar fi aproximaţia iniţială $x^{(0)} \in R^n$ şi eroarea se evaluează prin:

$$\|x^{(k)} - x*\| \le \frac{q}{1-q}\|x^{(k)} - x^{(k-1)}\| \le \frac{q^k}{1-q}\|x^{(1)} - x^{(0)}\|.$$

Demonstraţie. Condiţia $\|Q\| < 1$ implică condiţia (3.18) (vezi paragraful 3.2). Pentru a obţine a estimare a erorii folosim relaţia:

$$x* - x^{(k-1)} = x^{(k)} - x^{(k-1)} + Q(x^* - x^{(k-1)}).$$

Deci

$$\|x* - x^{(k-1)}\| \le \|x^{(k)} - x^{(k-1)}\| + \|Q\|\|x^* - x^{(k-1)}\|,$$

sau

$$(1 - \|Q\|)\|x* - x^{(k-1)}\| \le \|x^{(k)} - x^{(k-1)}\|.$$

Deoarece prin ipoteză $1 - \|Q\| \ge 1 - q > 0$ avem:

$$\|x* - x^{(k-1)}\| \le \frac{1}{1-q}\|x^{(k)} - x^{(k-1)}\|.$$

Pe de altă parte

$$\|x* - x^{(k)}\| = \|Q(x* - x^{(k-1)})\| \le \|Q\|\|x* - x^{(k-1)}\|.$$

Prin urmare

$$\left\|x^* - x^{(k)}\right\| \le \frac{q}{1-q}\left\|x^{(k)} - x^{(k-1)}\right\|.$$

Teorema este demonstrată.

Presupunem că elementele diagonale $a_{ii} \neq 0, i - 1,2,\ldots,n$. Atunci în calitate de matrice *S* se poate lua matricea diagonală ataşată matricei *A*:

$$S = Diag(a_{11}, a_{22}, \ldots, a_{nn}).$$

Avem

$$S^{-1} = Diag(\frac{1}{a_{11}}, \frac{1}{a_{22}}, \quad ..... \quad , \frac{1}{a_{nn}})$$

În acest caz sistemul (3.15) devine:

$$x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j\right), \quad i = 1,2,...,n.$$

Procesul iterativ (3.17) este definit prin:

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j^{(k)}\right), \quad i = 1,2,...,n. \tag{3.19}$$

Astfel obţinem o metodă de rezolvare a sistemului liniar (3.13) numită *metoda lui Jacobi.*

Exemplu. Fie dat sistemul de ecuaţii liniare:

$$\left.\begin{aligned} 2x_1 - x_2 &= 1, \\ -x_1 + 2x_2 &= 1 \end{aligned}\right\} \tag{3.20}$$

Vom avea:

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad S = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

$$Q = S^{-1}T = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}, \quad d = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Valorile proprii ale matricei Q sunt: $\lambda_1 = -\frac{1}{2}, \quad \lambda_2 = \frac{1}{2}$. 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 soluţia exactă $x^* = (1,1)^T$ oricare ar fi $x^{(0)} \in R^2$. În particular, pentru $x^{(0)} = (0,0)^T$ obţinem şirul:

$$x^{(1)} = \left(\frac{1}{2}, \frac{1}{2}\right)^T, x^{(2)} = \left(\frac{3}{4}, \frac{3}{4}\right)^T, x^{(3)} = \left(\frac{7}{8}, \frac{7}{8}\right)^T, \; x^{(4)} = \left(\frac{15}{16}, \frac{15}{16}\right)^T, \ldots.$$

Observăm că pentru metoda lui Jacobi matricea Q are elementele

$$q_{ij} = \begin{cases} 0, & \text{dacă} \quad i = j \\ -\dfrac{a_{ij}}{a_{ii}}, & \text{dacă} = ..i \neq j \end{cases}$$

Utilizând teorema 2 cu norma $\|\bullet\|_\infty$ obţinem:

$$\|Q\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |q_{ij}| = \max_{1 \le i \le n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

Rezultă că dacă

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

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 acelaşi pas: $x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$. Această modificare a metodii lui Jacob se numeşte *metoda lui Gauss-Seidel* iar şirul iterativ (3.19) devine:

$$x_i^{(h+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(h+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(h)} \right), \quad i = 1,2,\ldots,n.$$

Se observă uşor că metodei lui Gauss-Sedel corespunde desfacerea $A=S-T$ unde S este matricea subdiagonală ataşată lui A, iar T este matricea strict supradiagonală cu elementele $-a_{ij}$ ataşată la aceeaşi matrice A:

$$S = \begin{pmatrix} a_{11} & 0 & 0 & \ldots & 0 \\ a_{21} & a_{22} & 0 & \ldots & 0 \\ a_{31} & a_{32} & a_{33} & \ldots & 0 \\ \ldots\ldots\ldots\ldots\ldots & & & & \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{nn} \end{pmatrix}, \quad T = \begin{pmatrix} 0 & -a_{12} & -a_{13} & \ldots & -a_{1n} \\ 0 & 0 & -a_{23} & \ldots & -a_{2n} \\ 0 & 0 & 0 & \ldots & -a_{3n} \\ \ldots\ldots\ldots\ldots\ldots & & & & \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix}.$$

Să reluăm exemplul de mai sus cu sistemul de ecuaţii (3.20). Pentru metoda Gauss-Seidel avem:

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad S = \begin{pmatrix} 2 & 0 \\ -1 & 2 \end{pmatrix}, \quad T = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

$$Q = S^{-1}T = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & \frac{1}{4} \end{pmatrix}, \quad d = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

Şirul 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 aproximaţia iniţială $x^{(0)} = (0,0)^T$ obţinem:

$$x^{(1)} = \left(\frac{1}{2}, \frac{1}{2}\right)^T,\ x^{(2)} = \left(\frac{3}{4}, \frac{7}{8}\right)^T, x^{(3)} = \left(\frac{15}{16}, \frac{31}{32}\right)^T, \ldots.$$

Se observă că o iteraţie Gauss – Seidel aici este echivalentă cu două iteraţii Jacobi, deoarece valorile proprii ale matricei *Q* sunt 0 şi $\frac{1}{4}$, deci raza spectrală este $\rho\left(S^{-1}T\right) = \frac{1}{4}$. Aceasta înseamnă că eroarea la fiecare iteraţie se împarte la 4; în metoda lui Jacobi $\rho\left(S^{-1}T\right) = \frac{1}{2}$ şi eroarea se împarte la 2.

Subliniem că metoda lui Gauss – Seidel, necesitând acelaşi număr de operaţii aritmetice, este în caz general mai bună ca metoda lui Jacobi. Se poate arăta că dacă *A* este o matrice pozitiv difinită atunci metoda Gauss – Seidel converge de două ori mai repede către soluţie decît metoda lui Jacobi.

Se demonstrează şi 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 şi invers. Pentru ilustrare anunţăm următoarea teoremă.

Teorema (Reich). Dacă matrcea *A* este simetrică şi are elementele diagonale $a_{ii} > 0$ pentru orice *i* atunci metoda Gauss – Seidel converge dacă şi 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 şi nu este diagonal dominantă atunci este posibil ca raza spectrală $\rho\left(S^{-1}T\right)>1$ pentru metoda lui Jacobi.

După cum sa văzut mai sus, dacă matricea A este diagonal dominantă, atunci metoda lui Jacobi şi metoda lui Gauss – Seidel vor genera un şir de aproximaţii succesive care converge către soluţia exactă oricare ar fi aproximaţia iniţială $x^{(0)}$. Subliniem că această condiţie este numai suficientă şi 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 îmbunătăţirea vitezei de convergenţă a şirului aproximaţiilor succesive. Fie $\tilde{x}^{(k)}$ vectorul obţinut la pasul $k+1$ prin metoda Gauss – Seidel. Metoda iterativă definită prin:

$$x^{(k+1)}=x^{(k)}+\omega\left(\tilde{x}^{(k)}-x^{(k)}\right)$$

este cunoscută cu numele de *metoda suprarelaxărilor succesive.* Parametrul de relaxare ω se alege astfel încât să crească viteza de convergenţă. Pentru $\omega=1$ metoda se reduce la metoda lui Gause – Seidel.

S-a găsit că pentru o alegere potrivită a parametrului ω convergenţa metodei suprarelaxării succesive este net superioară metodelor Jacobi şi Gauss – Sedei. De exemplu, în cazul sistemului de ecuaţii liniare (3.20) o iteraţie prin metoda suprarelaxării succesive este echivalentă (vezi [2,35]) cu 30 de iteraţii 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 înţelegere mai profundă a metodelor iterative recomandăm referinţele [35,39]. În lucrarea [39] este adus şi un subprogram – Fortran a metodei suprarelaxării succesive.

3.9 Sisteme liniare supradeterminate şi metoda celor mai mici pătrate

3.9.1. Formularea problemei

Fie un sistem de *m* ecuaţii cu *n* necunoscute

$$Ax = b \tag{3.21}$$

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 numeşte sistem supradeterminat. Deoarece sistemul (3.21) conţine mai multe ecuaţii decât necunoscute, nu putem găsi, în caz general, o soluţie care să verifice exact toate ecuaţiile sistemului. De exemplu sistemul:

$$2x_1 = b_1,$$
$$3x_1 = b_2,$$
$$4x_1 = b_3,$$

va avea soluţie numai în cazul când terminii liberi b_1, b_2 şi b_3 se află în raportul $2:3:4$.

Deşi 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 pseudosoluţia x^* care minimizează eroarea medie pentru toate cele *m* ecuaţii ale sistemului.

O pseudosoluţie în sensul celor mai mici pătrate (CMMP) pentru sistemul supradeterminat (3.21) este un vector $x^* \in R^n$ cu proprietatea:

$$\|Ax^*-b\|_2^2 = \min_{x\in R^n} \|Ax-b\|_2^2 \qquad (3.22)$$

Vectorul x^* se mai numeşte s*oluţie generalizată* în *sensul CMMP*. Acest vector x^* minimizează norma euclidiană a vectorului rezidual $r = Ax-b$, adică minimizează abaterea pătratică a lui Ax faţă de b:

$$\|r\|_2^2 = (r,r) = \sum_{i-1}^{n} r_i^2$$

De aici şi denumirea de CMMP. În exemplu 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:

$$E(x) = \|Ax-b\|_2^2 = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} a_{ij}x_j - b_i\right)^2 \qquad (3.23)$$

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 obţine anulând derivatele parţiale în raport cu $x_1, x_2, \ldots, x_n$, adică anulând gradientul funcţiei $E(x)$:

$$\nabla E(x) = 2A^T(Ax-b) = 0 . \qquad (3.24)$$

Din ecuaţia (3.24) rezultă că orice pseudosoluţie x^* în sensul CMMP a sistemului (3.21) satisface relaţia:

$$A^T(Ax^*-b) = 0 ,$$

sau

$$A^T Ax^* = A^T b \qquad (3.25)$$

Sistemul de ecuaţii (3.25) se numeşte *sistem normal* asociat problemei (3.21). În acest sistem, $C = A^T A$ este o matrice de dimensiune $n\times n$, simetrică cu elimentile:

$$c_{ij} = \bar{a}_i^T \bar{a}_j = (\bar{a}_i, \bar{a}_j),$$

unde $\bar{a}_i = (a_{1i}, a_{2i}, \ldots, a_{mi})^T$ sunt vectorii coloană ai matricei A, $i = 1,2,\ldots,n$. 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 $\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_n$ ale matricei A sunt liniar independente atunci din $x \neq 0$ rezultă că şi $Ax \neq 0$, deci matricea $C = A^T A$ este pozitiv definită. Prin urmare este adevărată următoare teoremă:

Teoreme de existenţă şi 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 pseudosoluţie în sensul CMMP unică $x^* \in R^n$ şi

$$x^* = \left(A^T A\right)^{-1} A^T b \qquad (3.2)$$

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 pseudosoluţiei prin metodele prezentate în paragrafele 3.3 - 3.6, 3.8. Deoarece

matricea $A^T A$ este simetrică şi pozitiv definită putem folosi faotorizarea Cholesky. Pentru calculul lui $A^T A$ şi $A^T b$ sunt necesare $m \cdot n(n+3)/2$ operaţii aritmetice, iar metoda lui Cholesky de rezolvare a sistemelor cere aproximativ $n^3/3$ operaţii. Astfel cea mai mare parte a efortului este cerut de formarea sistemului normal asociat problemei (3.21).

Numărul de condiţionare a matricei A^TA este egal cu pătratul numărului de condiţionare a matricei A:

$$cond\left(A^T A\right) = \left[cond\left(A\right)\right]^2.$$

Rezultă că în general matricea $A^T A$ este prost condiţionată şi calculul său, deci, este afectat de erori de rotungire cu efect deseori catastrofal. De aceea în practică se evită formarea sistemelor normale şi 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 $\overline{a}_i^{\,T}, i = 1,2,\ldots,n$, ale matricei A sunt ortogonale putem uşor determina pseudosoluţia sistemului supradeterminat (3.21). Într-adevăr, dacă $\overline{a}_i^{\,T}\overline{a}_j = 0, i \neq j$, matricea $A^T A$ devine o matrice diagonală cu elementele de pe diagonală egale cu $\overline{a}_i^{\,T}\overline{a}_i \neq 0$ şi imediat se obţine pseudosoluţia:

$$x_i^* = \frac{b^T \overline{a}_i}{\overline{a}_i^{\,T}\overline{a}_i}, . \, i = 1,2,\ldots\ldots, n.$$

Prin urmare, în locul formării sistemului normal putem ortogonaliza coloanele matricei A. Un procedeu clasic de ortogonalizare este *metoda lui Gram – Schmidt*. Şirul de vectori liniari independenţi $a_1, a_2, \ldots, a_n$ se ortogonalizează după formulele:

$$v_1 = a_1, \quad v_i = a_i - \sum_{j=1}^{i-1} \frac{(a_i, v_j)v_j}{(v_j, v_j)}, \quad i = 2,3,....., n. \qquad (3.27)$$

Se constată uşor că vectorii $v_1, v_2, \ldots, v_n$ sunt ortogonali. Împărţind fiecare vector la lungimea lui, obţinem un şir de vectori ortonormaţi: $q_1 = \frac{v_1}{\|v_1\|_2}, \quad q_2 = \frac{v_2}{\|v_2\|_2}, \ldots, q_n = \frac{v_n}{\|v_n\|_2}$.

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):

$$v_2 = a_2 - \frac{a_2^T v_1}{v_1^T v_1} = a_2 - \frac{1}{2} v_1 = \begin{pmatrix} 1/2 \\ -1/2 \\ 1 \end{pmatrix}$$

Vectorii ortonormaţi 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 – Schimdt 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 ecuaţii este:

$$\begin{pmatrix} a_1 & a_2 \end{pmatrix} = \begin{pmatrix} q_1 & q_2 \end{pmatrix} \begin{pmatrix} \sqrt{2} & \sqrt{1/2} \\ 0 & \sqrt{3/2} \end{pmatrix},$$

adică $A = QR$ unde Q este cu coloane ortogonate 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 cunoaşte factorizarea *QR* a matricei *A* atunci CMMP se rezolvă uşor. Din (3.26) obţinem:

$$x^* = \left(A^T A\right)^{-1} A^T = \left(R^T Q^T QR\right)^{-1} R^T Q^T b,$$

de unde, ţinând seama că $Q^T Q = I$, rezultă că:

$$x^* = \left(R^T R\right)^{-1} R^T Q^T b = R^{-1} Q^T b.$$

Prin urmare pseudosoluţia în sensul CMMP se poate obţine uşor rezovând sistemul triunghiular:

$$Rx = Q^T b. \tag{3.28}$$

Pentru calculul lui R şi $Q^T b$ sunt necesare aproximativ $n^2 m$ operaţii, iar pentru rezolvarea sistemului triunghiular (3.28) numai $n(n+1)/2$ operaţii. Deci numărul total de operaţii este aproximativ de două ori mai mare decât cel pentru formarea sistemului de ecuaţii normale.

Există o variantă nouă a metodei Gram – Schmidt numită *algoritmul lui Gram – Schmidt modificat:*

Pentru $k = 1,2,\ldots,n$

$$a_k = \frac{a_k}{\|a_k\|_2}$$

Pentru $j = k+1, k+2, \ldots, n$

$$a_j = a_j - \left(a_j^T a_k\right) a_k$$

Algoritmul lui Gram – Schmidt modificat este numeric stabil datorită rearanjării ordinii de efectuare a calculelor. În plus, el necesită mai puţină memorie operativă decât metoda clasică de ortogonalizare. Vectorii q_k se calculează şi se plasează în acelaş loc de memorie care îl ocupă vectorii iniţiali a_k.

Pentru completarea cunoştinţelor cu alte metode de rezolvare a problemei CMMP se recomandă [2,22,29,34,35,37]. Menţionăm lucrarea [29] care conţine un program bun FORTAN bazat pe factorizarea valorilor singulare, numită DVS. Cititorul care stăpâneşte bine noţiunile de bază din algebra liniară recomandăm lucrarea fundamentală [29].

3.10 Exerciţii

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 determenie matricele pătrate de ordinul doi, satisfăcând relaţiile:

a) $A^2 = 0$ deşi $A \neq 0$

b) $B^2 = -I$

c) $CD = -DC$, dar $CD \neq 0$

d) $EF = 0$ cu toate că elementele matricelor *E* şi *F* sunt nenule.

3. Să se arate că

$(AB)^T = B^T A^T; (AB)^{-1} = B^{-1}A^{-1}; \left(A^{-1}\right)^T = \left(A^T\right)^{-1}$.

4. Matricea *P* se numeşte *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 inegalităţile:

a) $\big|\|x\|-\|y\|\big| \le \|x-y\|$

b) $\|x \pm y\| \le \|x\| + \|y\|$.

7. Să se deetermine unghiul dintre vectorii $x=(2,-2,1)^T$ şi $y=(1,2,2)^T$.

8. Să se arate că matricele *A* şi *B* sunt egale între ele dacă şi numai dacă $Ax = Bx$ pentru $\forall x$.

9. Fie $A = I - 2xx^T$ unde $x \in R^n, (x,x)=1$. Să se arate că matricea *A* este ortogonală şi $A^2 = I$.

10. Fie *A* o matrice ortogonală şi $Ax = \lambda x, x \neq 0$. 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 şi 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 $cond(AB) \le cond(A) \times cond(B)$.

17. Să se arate că sistemul de ecuaţii

$$\begin{cases} x_1 + 100x_2 = 100 \\ x_2 = 0 \end{cases}$$

Este rău condiţionat. Să se calculeze numărul de condiţionare.

18. Fie $x = (3 \quad 4)^T, z = (1 \quad 1)^T$. Să se calculeze $\sigma = \|x\|_2$, $v = z + \sigma x$ şi matricea corespunzătoare Householder.

19. Să se compare metodele Jacobi, Gauss – Seidel şi a suprarelaxării succesive în cazul matricei *A* din exerciţiul 15, iar

$$b = (1,0,1)^T \text{ şi } x^{(0)} = (0,0,0)^T.$$

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 preudosoluţia 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 şi 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 exerciţiul nr. 15.

23. Fie sistemul supradeterminat $Ax = b, A = (a_{ij})_{m \times n}$, $b \in R^m$, cu pseudosoluţia $x^* = (A^T A)^{-1} Ab$. Să se arate că vectorul rezidual $r = Ax^* - b$ este ortogonal pe subspaţiul

$$\text{Im } A = \{y | y = Ax, x \in R^n\} \subset R^m.$$

24. Matricea $A^+ = \left(A^T A\right)^{-1} A^T$ de dimensiune $n \times m$ se numeşte *pseudoinversa* lui A sau *generalizarea matricei inverse a lui More – Penrose*. Să se arate că

$$AA^+A = A; \quad A^+AA^+ = A^+;$$

$$\left(AA^+\right)^T = AA^+; \quad \left(A^+A\right)^T = A^+A.$$

25. $P_A = AA^+$ este *proiectorul ortogonal* al lui A pe spaţiul Im A. Să se arate că $P_A^2 = P_A$ şi $P_A^T = P_A$, adică matricea P_A este idempotentă şi simetrică.