Conf. dr. ing. GHEORGHE DODESCU

METODE NUMERICE ÎN ALGEBRĂ

EDITURA TEHNICĂ
București — 1979

Lucrarea are un caracter pronunțat aplicativ, prezentînd în fiecare capitol tipuri de procese fizice care au modele matematice din algebră, precum și modul cum trebuie abordate aceste modele în scopul unei prelucrări electronice. Se face o ierarhizare a metodelor numerice prezentate după eficiența acestora, în ceea ce privește convergența, consistența, stabilitatea, eroarea, numărul de operații, necesarul de memorie etc. Metodele care se utilizează frecvent în practică sînt ilustrate cu exemple rulate pe calculator.

Cartea se adresează inginerilor, informaticienilor, automaticienilor, analiștilor de sisteme, studenților etc.

Control științific: Conf. dr. **Alexandru Șchiop**
Redactor: **Valentina Crețu**
Tehnoredactor: **Elena Geru**
Coperta: **Constantin Guluță**

Bun de tipar 29. X. 1979. Tiraj 7250 + 90 exemplare broșate Coli de tipar 21. C.Z. 512

c. 44 I. P. Informația, str. Brezoianu nr. 23—25, București

Apariția unei lucrări privind metodele de calcul numeric aplicabile pe calculator se înscrie pe linia generală de promovare a calculului automat. Lucrarea de față se încadrează în preocupările utilizării tehnicilor electronice de calcul în activitatea de cercetare, proiectare și învățămînt.

În cele șase capitole se prezintă principalele metode de calcul numeric pentru o mare varietate de modele matematice ce pot fi întîlnite frecvent în practică, folosind drept instrument de calcul calculatorul electronic. Metodele de calcul prezentate sînt analizate, comparate și ierarhizate după următoarele criterii: convergență, consistență, stabilitate, număr de operații, timp de execuție, necesarul de memorie, tehnicile de control asupra modului de propagare a erorilor. Avînd în vedere scopul aplicativ al lucrării, în primul capitol se prezintă o serie de elemente necesare pe parcursul lucrării cum ar fi: rolul, posibilitățile și limitele de utilizare ale unui calculator; aspectele matematice și de calcul ale unui algoritm; tipurile de erori introduse la executarea unui algoritm; instabilitatea numerică a algoritmilor și natura problemelor; tehnici de investigare privind precizia rezultatelor etc. Noțiunile teoretice sînt însoțite în majoritatea cazurilor de aplicații, diagrame logice, programe rezultate, precum și de un material grafic adecvat, în felul acesta fiind mult mai accesibile cititorilor. La începutul fiecărui capitol sînt prezentate o serie de modele fizice și modelele matematice corespunzătoare, încercînd în acest fel realizarea unei legături între domeniile furnizoare de probleme și matematică, analiză numerică și calculator, scop foarte important în etapa actuală cînd calculatorul trebuie folosit

mai mult în activitatea de cercetare, producție, învățămînt pentru obținerea unor rezultate de calitate.

Lucrarea se adresează economiștilor, inginerilor, fizicienilor, cadrelor didactice, studenților de la facultățile care au în planul de învățămînt o asemenea disciplină sau discipline înrudite, precum și tuturor celor care doresc să-și finalizeze lucrările de cercetare și proiectare cu ajutorul tehnicilor electronice de calcul.

AUTORUL

CUPRINS

INTRODUCERE

1.1. Rolul, posibilităţile şi limitele unui calculator în activitatea curentă

Matematica aplicată, din care face parte şi analiza numerică, s-a dezvoltat în ultimul timp datorită, mai ales, amplificării utilizării sistemelor electronice de calcul. Analiza numerică se ocupă cu elaborarea, analiza, evaluarea şi ierarhizarea algoritmilor numerici, care se pot executa pe un calculator electronic, pentru obţinerea soluţiei problemei considerate.

Metodele şi rezultatele analizei clasice, folosite adesea, oferă de obicei numai baza şi/sau punctul de start pentru analiza numerică. De exemplu, un matematician cu preocupări în domeniul matematicii pure va fi complet satisfăcut dacă poate demonstra că la o problemă dată soluţia există şi este unică, dar este foarte mult pentru un matematician sau inginer care lucrează în domeniul analizei numerice să realizeze o procedură pentru calculul soluţiei, cu o tehnică de calcul existentă, să se menţină în cadrul unei precizii impuse şi în cadrul unui timp de calcul rezonabil.

Pentru dezvoltarea unui algoritm de calcul folosit la rezolvarea unei probleme date, analistul trebuie să fie preocupat nu numai de numărul operaţiilor aritmetice şi de precizia teoretică, dar şi de erorile de rotunjire şi trunchiere, care se comit cînd algoritmul este implementat

pe un calculator electronic. Scopul acestei lucrări este de a studia algoritmii numerici orientați pe calculator pentru rezolvarea diverselor tipuri de probleme întîlnite în cercetare, proiectare etc.

Calculatoarele sînt sisteme fizice proiectate pentru a implementa modelele matematice și pentru manipularea lor în mod automat [30, 39]. Noile tehnologii, arhitecturile diferite, microprogramarea unor activități, memoriile virtuale etc. au avut o mare influență asupra calculatoarelor care se construiesc în prezent. Un calculator folosit în scopuri generale, construit în prezent, este mult mai rapid, mai mic, mai performant, mai ieftin decît predecesorul său, calități care permit ca acesta să fie din ce în ce mai mult utilizat în economie, producție și cercetare. Din acest punct de vedere aplicațiile pot fi :

— *aplicații care solicită capacitatea de memorare și manipulare a unui volum important de informații*;

— *aplicații care implică precizie și viteză în executarea unor calcule matematice.*

Ambele tipuri de aplicații pot fi executate pe un calculator de uz general.

În ultimii ani calculatoarele au început să fie utilizate intens în noi domenii ca : tehnica comunicațiilor, controlul și conducerea proceselor, stocarea și sortarea unor volume mari de date, roboți etc. În toate aceste domenii, calculatorul prelucrează cantități mari de date cu viteze foarte mari.

Un calculator poate fi programat să rezolve orice problemă care a fost corect definită. Prin definirea unei probleme se înțelege alcătuirea unui algoritm de calcul constituit din etape ce pot fi codificate cu ajutorul unor secvențe de instrucțiuni ale calculatorului.

Domeniile în care sînt utilizate calculatoarele sînt următoarele :

● *Domeniul transmiterii informației.* Calculatorul poate fi folosit în procesul de sincronizare a transmisiei, comutației, codificării și memorării informației pentru sistemele de comunicație și în special în comunicațiile dintre calculatoarele numerice și terminale la distanță.

● *Controlul și conducerea proceselor cu ajutorul calculatorului.* Calculatoarele pot fi foarte bune instrumente pentru urmărirea și conducerea automată a producției, dacă sînt programate să conducă procese tehnologice sau mașini-unelte, cu mai multă rapiditate și precizie decît este posibil să o facă omul. Calculatorul controlează procesul luînd decizii în timp real, ceea ce conduce la creșterea calitativă și cantitativă a producției.

● *Cercetarea științifică și experiențele de laborator.* Calculatorul se utilizează în activitatea de laborator pentru evaluarea și memorarea informației culese de la diverse și numeroase dispozitive electronice de măsură și control, folosite în experiența de analizat. Există experiențe unde parametrii (sau semnalele) trebuie percepuți și înregistrați cu viteză foarte mare, altfel informația respectivă se pierde, fapt care impune prelucrări rapide atît pentru regimurile dinamice cît și pentru cele staționare. Dispozitivele de calcul în timp real sau „on-line" servesc drept componente ale sistemului de măsură și control, realizînd următoarele funcțiuni:

— implementează relațiile matematice între variabilele fizice (generează funcții, predictează valori parametrilor, optimizează și reglează valoarea unor parametri etc.);

— inițializează și controlează din punct de vedere logic secvențe de operații și experiențe.

● *Simularea proceselor cu ajutorul calculatorului.* În general este scump, nepractic și periculos să încerci un avion, un tren, un vapor, în condiții normale. Calculatorul poate permite simularea în toate aceste condiții de încercare, răspunde la toate acțiunile modelului și furnizează rezultatele încercării, realizînd astfel o economie de timp și de instalație fără a se risca și fără a se folosi obiectul de încercat. Asemenea aplicații necesită prelucrarea de informații numerică și analogică. Informația analogică constă din mărimile fizice continue care pot fi generate și controlate, cum ar fi tensiuni, curenți, unghiuri de rotație etc. Informația numerică constă din valori numerice discrete, care reprezintă variabilele problemei. În majoritatea cazurilor valorile analogice sînt convertite (cu ajutorul convertoarelor analogic/numeric) în valori numerice pentru

rezolvarea numerică a problemei. În general, calculatoarele folosite în astfel de aplicații combină caracteristicile unui calculator numeric și cele ale unui calculator analogic în cadrul unui singur sistem de calcul, denumit *sistem de calcul hibrid*. Simularea este utilizată de asemenea în cazul unor experiențe care se desfășoară în mod normal într-un timp foarte lung sau imposibile datorită condițiilor reale. Uneori experiențele de simulare sau testele implică părți sau componente ale sistemului real. Simularea este folosită cu rezultate foarte bune în domenii ca proiectare, cercetare, învățămînt, planificare, jocuri strategice etc.

● *Rezolvarea problemelor numerice și prelucrarea datelor*. Calculatorul este un instrument indispensabil în probleme de proiectare. La proiectarea unui dispozitiv sau instalații care depinde de foarte mulți parametri, proiectantul descrie comportarea acestor parametri și interdependențele dintre aceștia cu ajutorul unor ecuații matematice adecvate. Se folosește în continuare un limbaj de programare pentru scrierea, codificarea algoritmului și scrierea programului de calcul. În final calculatorul este folosit pentru executarea acestui program [35, 44]. În cadrul acestui domeniu de aplicații sînt incluse calcule de proiectare (care implică rezolvarea sistemelor de ecuații liniare, rezolvarea ecuațiilor și sistemelor neliniare, rezolvarea numerică a ecuațiilor diferențiale ordinare și a ecuațiilor diferențiale cu derivate parțiale, calcule matriceale, problema valorilor și vectorilor proprii etc.), calcule statistice, studii genetice, calcule de gestiune etc. Rezultatul obținut în urma rulării programului asociat problemei poate fi un rezultat numeric, ori de descriere, dar întotdeauna servește pentru luarea unei decizii.

După cum s-a arătat, soluționarea unei probleme presupune definirea ei corectă, construirea unui algoritm de calcul și codificarea acestui algoritm cu ajutorul unei secvențe de instrucțiuni calculator. Cu ani în urmă se considera că un calculator poate fi programat să rezolve orice problemă care poate fi corect pusă (corect definită). Practica a demonstrat că anumite probleme, de exemplu translatarea limbajului natural, sînt foarte greu de definit. Totuși este destul de ușor să translatezi o listă de

cuvinte dintr-o limbă în alta [39, 44] dar este foarte dificil să translatezi propoziții pentru că există o serie de nuanțe și sensuri asociate cuvintelor individuale și combinațiilor de cuvinte. Acest element arată că nu este practic să comunici cu un calculator, folosind limbajul natural. Datorită acestui fapt au fost realizate limbaje specifice pentru dialogul om-calculator și limbaje de programare specifice unor domenii de preocupări ca economie, inginerie, matematică etc.

Astfel se poate considera următoarea clasificare a limbajelor :

• *Limbaje universale* (orientate pe proceduri) : FORTRAN, ALGOL, COBOL, PL/I, BASIC etc., limbaje cu structura și sintaxa lor proprie. Aceste limbaje sînt orientate pe tipuri de aplicații și conțin cuvinte și expresii familiare domeniului de aplicație. Însușirea acestor limbaje și a tehnicii de scriere a programelor se poate realiza într-un timp relativ scurt.

Multe firme producătoare de echipament de calcul au adoptat limbaje de programare standard și au implementat aceste limbaje pe calculatoarele lor. Un program scris într-un anumit limbaj universal poate fi rulat pe un număr mare de calculatoare fără schimbări esențiale, dacă calculatoarele considerate dispun de compilatorul asociat limbajului respectiv.

• *Limbaje de programare specializate* (orientate pe tipuri de aplicații) au fost proiectate pentru controlul programat al mașinilor-unelte, pentru calculatoare specializate pe activități cum ar fi culegerea datelor, culegerea textelor și tipărirea cărților, compunerea muzicii, probleme de instruire și alte aplicații.

Orice sistem de calcul poate fi considerat ca format din două componente : hardware și software, care cooperează la rezolvarea unei probleme ce poate fi foarte complexă sau laborioasă. Partea de hardware constă din calculatorul propriu-zis (pentru calcule și control) și diverse periferice (dispozitive de I/E) pentru introducerea datelor și tipărirea rezultatelor. Partea de software constă dintr-o colecție de programe utilizate pentru a extinde facilitățile componentei hardware. Fiecare program constă dintr-o

secvență de instrucțiuni în limbaj mașină necesare calculatorului pentru rezolvarea unei probleme. Selectarea și optimizarea componentelor hardware și software pentru a satisface o aplicație considerată are ca scop realizarea unor performanțe îmbunătățite la un preț scăzut. Dependența reciprocă dintre hardware, software (metode și tehnici de programare) și aplicații este dată în fig. 1.1.

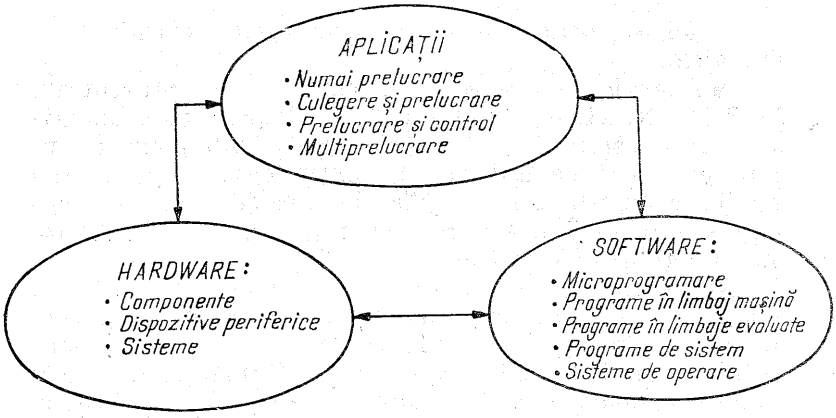


Fig. 1.1.

Performanțele hardware sînt determinate în cea mai mare parte de tipul și calitatea componentelor ce-l constituie. Datorită acestui fapt a avut loc o evoluție hardware de la utilizarea circuitelor logice și memoriilor cu parametri scăzuți (tuburi electronice, memorii pe tambur) la circuite integrate cu viteze mari de comutație și la memorii pe inele de ferită. Odată cu creșterea complexității componentei de hardware sînt necesare alte forme de software. Datorită acestui fapt au apărut sisteme de operare proiectate cu scopul de a gestiona toate resursele hardware și alte resurse informaționale existente în cadrul unui sistem de calcul, într-o manieră eficientă. Sistemele de operare evoluate oferă posibilitatea prelucrării în time-sharing (prelucrare prin divizarea timpului), culegerea datelor și controlul

gestiunii în timp real, o execuţie a mai multor programe în acelaşi timp, distribuirea şi utilizarea aparent simultană a tututror tipurilor de resurse etc. [40, 110].

În rezolvarea unei probleme cu ajutorul calculatorului pot fi evidenţiate următoarele etape [128, 123]:

● *Enunţarea problemei şi formularea matematică.* În această etapă se exprimă matematic relaţiile şi restricţiile dintre parametrii problemei, se pun în evidenţă condiţiile iniţiale precum şi restricţiile referitoare la soluţie.

● *Alegerea metodei numerice.* Rezolvarea problemei presupune existenţa unui algoritm de calcul. Avînd în vedere utilizarea calculatorului electronic, la alegerea metodei numerice trebuie ţinut seamă de următoarele elemente: precizia impusă, viteza de calcul, necesarul de memorie în funcţie de volumul datelor, simplitatea formulelor de calcul, controlul erorilor, consistenţa, stabilitatea şi convergenţa metodei, timpul de răspuns etc.

● *Descrierea algortimului metodei numerice.* Pentru aceasta se folosesc schemele logice. Schema logică trebuie să evidenţieze succesiunea logică a etapelor importante din algoritmul de calcul şi deciziile logice necesare obţinerii soluţiei, adică o reprezeare grafică a algoritmului de calcul.

● *Întocmirea programului de calcul.* După ce algoritmul metodei numerice alese a fost reprezentat grafic cu ajutorul diagramei logice, are loc codificarea lui cu ajutorul unui limbaj de programare, în vederea executării cu ajutorul calculatorului electronic. Programul de calcul se poate scrie folosind schema logică, care pune în evidenţă algoritmul, specificînd datele problemei, formulele de calcul, deciziile logice şi modul de descriere a rezultatelor. Programul de calcul se scrie într-un limbaj de programare ca FORTRAN, COBOL, ALGOL, PL/I, BASIC etc. cu ajutorul unor instrucţiuni.

● *Testarea rezultatelor.* Sînt necesare diferite procede de control care să permită verificarea unor rezultate parţiale, detectarea eventualelor erori apărute în calcule şi modul de propagare a erorilor. Aceste elemente oferă informaţii necesare opririi sau continuării calculelor.

Interpretarea rezultatelor, din punct de vedere al problemei practice propuse.

Schemele logice sînt reprezentări grafice ale fluxului de informaţii care stabilesc legătura între operaţiile implicate în rezolvarea problemei. Utilitatea schemelor logice apare la depanarea programelor şi la schimbul de informaţie între diverse grupuri de programatori. În cazul aplicaţiilor complexe elaborarea schemei logice este obligatorie [35, 98].

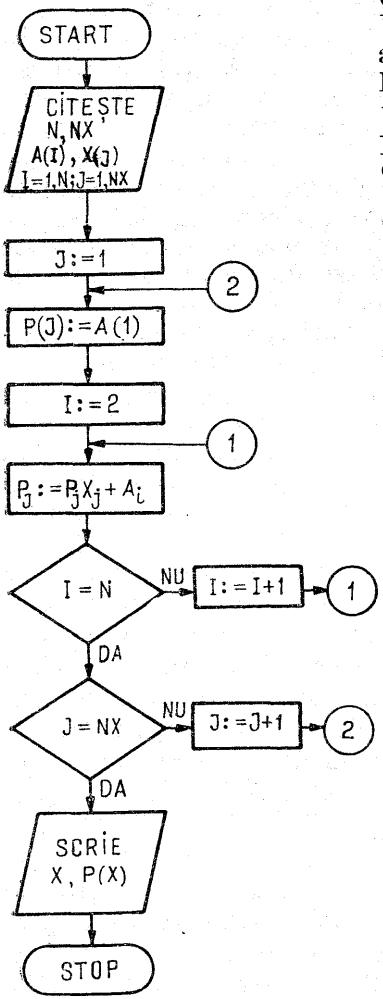


Fig. 1.2.

Fie schema logică pentru evaluarea polinomului $P(x) = a_1x^n + a_2x^{n-1} + \ldots + a_nx + a_{n+1}$, variabila şi coeficienţii polinomului fiind reali. Datele iniţiale ale problemei sînt :

n — gradul polinomului,

$a_1, a_2, \ldots, a_n, a_{n+1}$ — coeficienţii polinomului,

x — valoarea reală în care se cere evaluarea polinomului.

Formulele de calcul utile etapei de programare sînt :

$P := a_1$ — formula de start,

$P := Px + a_i$. $i = 2, 3, \ldots, N+1$, — formulă recursivă, pentru calculul lui $P(x)$ pentru valoarea x dată.

În fig. 1.2 este reprezentată schema logică pentru problema considerată, iar programul 1.1 în FORTRAN codifică algoritmul prezentat pentru polinomul

$$P_6(x) = 1x^6 + 2x^5 + 3x^4 + 4x^3 + 5x^2 + 6x + 7$$

```
      DIMENSION A(20),X(20),P(20)
      INTEGER A,X,P
      READ(105,100) N,NX
100   FORMAT(2(I3))
      READ(105,101) (A(I),I=1,N),(X(J),J=1,NX)
101   FORMAT(12(I5))
      DO 4 J=1,NX
      P(J)=A(1)
      DO 4 I=2,N
      P(J)=P(J)*X(J)+A(I)
      WRITE(108,102) (X(J),J=1,NX),(P(I),I=1,NX)
102   FORMAT(' ',9X,'X  I',5(I7,3X),/,9X,52('-'),/,8X,'P(X) I',
     *                    5(I7,3X))
      STOP
      END
          LINK
          RUN

   X  I        1        2        3        4        5
 ------------------------------------------------------
 P(X) I       28      247     1636     7279    24412

          EOJ
```

Programul 1.1.

în punctele $x = 1, 2, 3, 4, 5$. Rezultatele sînt date sub norma unui tabel:

x	1	2	3	4	5
$P(x)$	$P(1)$	$P(2)$	$P(3)$	$P(4)$	$P(5)$

1.2. Aspecte matematice şi de calcul ale unui algoritm

Analiza numerică se ocupă cu aplicarea matematicii la construcţia şi algoritmizarea metodelor care pot fi utilizate la obţinerea soluţiei numerice a problemelor cu ajutorul calculatorului electronic.

De foarte multe ori se vede că anumite rezultate ale analizei clasice nu sînt integral folositoare analizei numerice. Un exemplu în acest sens îl constituie teoremele de existenţă şi unicitate ale soluţiei pentru anumite clase de probleme, teoreme care se demonstrează presupunînd că soluţia nu există şi astfel se ajunge la o contradicţie. Astfel de demonstraţii nu oferă nici un fel de informaţie utilă despre modul de găsire a soluţiei pentru care s-a demonstrat că există şi este unică [128, 110]. De asemenea,

uneori, chiar dacă soluţia analitică a unei probleme date poate fi găsită, aceasta nu întotdeauna poate servi la obţinerea soluţiei numerice. De exemplu seria de forma

$$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots \qquad (1.1)$$

converge absolut către e^x pentru orice valoare a lui x. Dar dacă seria se utilizează pentru a calcula e^{-100}, va fi complet impracticabilă, deoarece volumul de calcul implicat va fi foarte mare, chiar dacă se utilizează un calculator electronic performant, timpul cerut pentru execuţie va fi excesiv de mare, iar precizia va fi foarte scăzută. Funcţia e^x poate fi evaluată uşor şi mult mai precis pentru $x = -100$ prin alte metode, utilizînd alt algoritm.

Un alt exemplu, care ilustrează faptul că o soluţie analitică a unei probleme date nu serveşte în mod practic la găsirea soluţiei numerice a problemei considerate este următorul sistem de ecuaţii algebrice:

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, 2, \ldots, n. \qquad (1.2)$$

Cînd acest sistem are o soluţie unică, se poate rezolva analitic cu ajutorul regulii lui Cramer. Dar metoda lui Cramer pentru rezolvarea sistemelor liniare algebrice cu ajutorul calculatorului este neindicată pentru $n \geqslant 3$, deoarece trebuie calculaţi $n + 1$ determinanţi de ordinul n, pentru aflarea soluţiilor sistemului, iar evaluarea fiecăruia dintre aceşti determinanţi implică în general $P_n n!$ operaţii de înmulţire, dacă dezvoltarea se face în funcţie de minori [128, 30], unde

$$P_n = \sum_{j=2}^{n} \frac{1}{(j-1)!} \text{ şi } \lim_{n\to\infty} P_n = e - 1. \qquad (1.3)$$

De asemenea este necesar aproximativ acelaşi număr de operaţii de adunare. În concluzie, rezolvarea sistemului (1.2) prin metoda lui Cramer implică $2P_n(n + 1)!$ operaţii elementare.

Dacă $n = 20$, numărul operațiilor aritmetice elementare va fi de aproximativ 16×10^{19}, adică un calculator modern, capabil să execute $2 \cdot 10^6$ operații pe s, va trebui să ruleze continuu la această problemă $2 \cdot 10^6$ ani [127]. Chiar și metode mai sofisticate de evaluare a determinanților (care reduc fantastic numărul de operații) nu sînt suficient de competitive cu metodele numerice ce se vor prezenta în lucrare.

Numărul mare al operațiilor din cadrul unor metode analitice nu deranjează numai din punctul de vedere al timpului de execuție consumat de calculator, dar și din punctul de vedere al preciziei de calcul, permițînd acumularea erorilor de rotunjire. Algoritmii orientați pe calculator sînt caracterizați de simplitate și manipulare ușoară. Aproape frecvent reducerea numărului operațiilor pierde în favoarea simplicității. Atenție sporită este acordată controlului acumulării erorilor de rotunjire. Se comite o mare greșeală cînd sînt neglijate aspectele matematice ale analizei numerice și se ține seama numai de aspectul calculator și invers. În acest sens în cadrul unei aplicații trebuie să se acorde o atenție egală atît aspectelor matematice cît și aspectelor legate de calculator, pentru selectarea unui algoritm adecvat rezolvării poblemei considerate.

Un exemplu care sugerează necesitatea îmbinării aspectelor matematice cu cele legate de calculator pentru un algoritm este problema evaluării funcției

$$g(x) = \operatorname{tg} x - \sin x$$

pentru valori mici ale argumentului x. Astfel pentru $x = 0{,}1250$ din tabele rezultă valorile

$$\operatorname{tg} 0{,}1250 \approx 0{,}1257, \quad \sin 0{,}1250 \approx 0{,}1247,$$

de unde rezultă că $g(0{,}1250) \approx 0{,}0010$. Se poate obține un rezultat mult mai precis dacă se formulează altfel problema, de exemplu prin dezvoltarea în serie Taylor a celor două funcții obținîndu-se

$$\operatorname{tg} x = x + \frac{1}{3} x^3 + \frac{2}{15} x^5 + \frac{17}{315} x^7 + \ldots$$

$$\sin x = x - \frac{1}{6} x^3 + \frac{1}{120} x^5 - \frac{1}{5040} x^7 + \ldots$$

astfel că $g(x)$ devine

$$g(x) = \frac{1}{2}x^3 + \frac{1}{8}x^5 + \frac{13}{240}x^7, \quad g(0{,}1250) \approx 0{,}009804.$$

Din acest exemplu se vede necesitatea examinării problemei și, dacă este necesar, reformularea ei matematică în sensul obținerii unui răspuns mai precis cu un timp de execuție rezonabil.

La aplicarea unui algoritm în practică există considerente matematice mai importante sau mai puțin importante de care trebuie să se țină seama. De asemenea operațiile aritmetice nu pot fi executate riguros, în general erorile de rotunjire pot afecta serios rezultatele. Nu se poate garanta că au loc egalitățile

$$a\left(\frac{b}{c}\right) = b\left(\frac{a}{c}\right) = \frac{ab}{c}$$

datorită erorilor de rotunjire ce pot fi introduse de calculator.

1.3. Tipuri de erori introduse la executarea unui algoritm

Precizia calculelor numerice este parametrul important în alegerea metodelor de calcul. Un algoritm de calcul este eficient cînd precizia calculelor este bună. Cu toate performanțele calculatoarelor electronice, precizia rezultatelor este influențată de mai mulți factori. Soluția depinde de datele inițiale, acestea fiind datele unor măsurări, observații sau soluții aproximative ale altor probleme, fapt care face ca la rezolvarea numerică a unei probleme să se introducă erori. Uneori erorile sînt introduse de modelul matematic cînd acesta nu corespunde în toată intimitatea fenomenului fizic modelat, datorită unor aproximații efectuate în fazele de modelare. Aceste tipuri de erori se numesc *erori inerente* (inițiale). La soluționarea numerică a unei probleme se folosește o anume metodă, care poate introduce o eroare; astfel de eroare se numește

eroare de metodă care poate fi micşorată prin alegerea metodei celei mai adecvate.

În procesul de calcul apar erori de trunchiere şi erori de rotunjire.

În concluzie, *eroarea totală* se compune din cele trei erori amintite : *eroarea inerentă*, *eroarea metodei* şi *eroarea de calcul*.

Fie x o valoare adevarată şi x^* o valoare aproximativă a lui x rezultată în urma unei măsurări, observaţii sau a unui calcul numeric. Dacă $x^* < x$, x^* aproximează pe x prin lipsă, dacă $x^* > x$, x^* aproximează pe x prin adaos (exces).

● Diferenţa $\varepsilon_x = x - x^*$ poartă denumirea de *eroare* iar $|\varepsilon_x| = |x - x^*|$ se numeşte *eroare absolută*.

● *Eroarea relativă* este raportul dintre eroarea absolută $|\varepsilon_x|$ şi valoarea absolută a lui x, adică

$$\varepsilon_r = \frac{|x - x^*|}{|x|} = \frac{|\varepsilon_x|}{|x|}. \tag{1.4}$$

Diferenţa dintre x şi x^* se măsoară în funcţie de eroarea absolută şi eroarea relativă din (1.4).

În cazul aproximării funcţiilor prin polinoame sau funcţii raţionale, analiza erorilor este făcută cu ajutorul funcţiei eroare. Dacă $R(x)$ este aproximaţia lui $F(x)$, atunci funcţia eroare absolută şi funcţia eroare relativă sînt

$$\varepsilon_a(x) = |R(x) - F(x)|, \quad \varepsilon_r(x) = \frac{|R(x) - F(x)|}{|F(x)|}. \tag{1.5}$$

Calitatea aproximării lui $F(x)$ prin $R(x)$ se măsoară cu ajutorul celor două funcţii eroare date în (1.5).

Din cele prezentate se observă că noţiunea de eroare (absolută şi relativă) se referă la metoda de măsură a erorii, iar termenul de eroare de trunchiere şi eroare de rotunjire se referă la sursa de eroare. Erorile de trunchiere şi erorile de rotunjire apar în procesul de calcul. Fie un program pentru evaluarea funcţiei sin x pentru $-1 < x < 1$,

cu ajutorul polinomului $P(x)$ ce aproximează pe sin x, unde $P(x)$ a fost obținut prin trunchierea seriei Maclaurin de dezvoltare a lui sin x. Eroarea implicată în aproximarea lui sin x prin polinomul $P(x)$ este *eroare de trunchiere.* Eroarea absolută și eroarea relativă de trunchiere sînt date de expresiile

$$\varepsilon_a(x) = |P(x) - \sin x|, \quad \varepsilon_r(x) = \frac{|P(x) - \sin x|}{|\sin x|}. \quad (1.6)$$

Eroarea implicată în aproximarea polinomului $P(x)$ prin valoarea calculată $P^*(x)$ se numește *eroare de rotunjire.* Eroarea absolută și eroarea relativă de rotunjire sînt date prin expresiile

$$\varepsilon_a(x) = |P^*(x) - P(x)|, \quad \varepsilon_r(x) = \frac{|P^*(x) - P(x)|}{|P(x)|}. \quad (1.7)$$

Dacă se face evaluarea unei funcții $f(x)$ pentru un anumit x din intervalul $[a, b]$, programul calculează un număr $f^*(x)$ care aproximează pe $f(x)$. Fie $\bar{f}(x)$ aproximarea lui $f(x)$ calculată teoretic cu ajutorul calculelor ordinare fără a se utiliza calculatorul electronic. Și de această dată se recunosc două surse de erori în $f^*(x)$, ce aproximează pe $f(x)$: eroarea implicată în aproximarea lui $f(x)$ prin $\bar{f}(x)$, numită eroare de trunchiere absolută și cea relativă, date prin formulele

$$\varepsilon_a(x) = |\bar{f}(x) - f(x)|, \quad \varepsilon_r(x) = \frac{|\bar{f}(x) - f(x)|}{|f(x)|}, \quad (1.8)$$

și eroare implicată în aproximarea lui $\bar{f}(x)$ prin $f^*(x)$, numită eroare de rotunjire absolută și cea relativă:

$$\varepsilon_a(x) = |f^*(x) - \bar{f}(x)|, \quad \varepsilon_r(x) = \frac{|f^*(x) - \bar{f}(x)|}{|\bar{f}(x)|}. \quad (1.9)$$

Eroarea absolută totală [42, 107] are expresia $|f^*(x) - f(x)|$ și poate fi exprimată sub forma

$$|f^*(x) - f(x)| = |[f^*(x) - \bar{f}(x)] + [\bar{f}(x) - f(x)]|, \quad (1.10)$$

adică este dată de combinația dintre eroarea absolută de trunchiere și eroarea absolută de rotunjire.

Nivelul erorilor de rotunjire în diversele rutine de evaluare depinde de o serie de factori:

— precizia cu care se execută calculele;

— ordinea în care se execută operațiile aritmetice;

— dacă rezultatele intermediare sînt rotunjite sau numai trunchiate;

— aspectele aritmetice ale calculatorului utilizat.

În sistemul de reprezentare în virgulă fixă are loc inegalitatea

$$\left| \frac{f^*(x) - \bar{f}(x)}{f^*(x)} \right| \leqslant A(x), \quad (1.11)$$

unde $A(x)$ satisface relația

$$\frac{1}{2} B^{-t} < A(x) \leqslant \frac{1}{2} B^{-t+1}, \quad (1.12)$$

B fiind baza de reprezentare și t numărul caracterelor din baza B conținut în mantisa numărului reprezentat în virgulă mobilă.

Pentru $B = 16$ și $t = 6$,

$$A(x) \approx B^{-t+1} = 16^{-6+1} = 16^{-5} = 2^{-20}.$$

Propagarea erorilor de rotunjire poate fi controlată prin extinderea tehnicilor de programare. Protecția contra erorilor de rotunjire poate fi obținută prin folosirea preciziei duble în anumite etape ale algoritmului de calcul. De asemenea, în cadrul unei rutine de evaluare scrise pentru virgula mobilă, se pot folosi calcule în virgulă fixă pentru etape intermediare, deoarece precizia în virgulă fixă pentru unele calculatoare este mai bună decît în virgulă mobilă de simplă precizie.

1.4. Instabilitatea numerică a algoritmilor şi natura problemelor

Dacă se face calculul numeric al valorilor rădăcinilor unei ecuaţii de gradul al doilea cu o formulă obişnuită, se observă că, pentru anumite valori ale coeficienţilor A, B şi C, una din rădăcini nu este calculată cu aceeaşi precizie cu care este calculată cealaltă.

La obţinerea soluţiei numerice [1.04, 1.01], un algoritm este numeric instabil dacă pentru coeficienţii A, B, C ai ecuaţiei de gradul al doilea există o pierdere a cifrelor semnificative. Dacă nu are loc pierderea cifrelor semnificative, algoritmul se numeşte *stabil numeric.*

Fie sistemul de două ecuaţii algebrice

$$\begin{aligned} 4{,}0000\ x + 0{,}8889\ y &= 4{,}0000, \\ 1{,}0000\ x + 0{,}2222\ y &= 1{,}0000. \end{aligned} \tag{1.13}$$

Dacă a doua ecuaţie se înmulţeşte cu 4 şi se scade din prima, rezultă $0{,}0001y = 0{,}0000$; se obţine o soluţie unică $x = 1{,}0000$, $y = 0{,}0000$.

Dacă se consideră sistemul (1.13) doar cu trei cifre după virgulă, rezultă sistemul

$$\begin{aligned} 4{,}000\ x + 0{,}888\ y &= 4{,}000, \\ 1{,}000\ x + 0{,}222\ y &= 1{,}000, \end{aligned} \tag{1.14}$$

care are o infinitate de soluţii:

$$x = 1{,}000 - 0{,}222\ k,\ y = k. \tag{1.15}$$

Din prezentarea acestui sistem se vede că *o mică perturbaţie* într-un singur coeficient schimbă problema dintr-o problemă cu o soluţie unică într-o problemă cu o infinitate de soluţii. Acest lucru reprezintă o proprietate matematică a problemei care este complet independentă de orice algoritm utilizat.

Dacă o problemă are proprietatea ca o mică perturbaţie în una din datele (sau în toate) conduce la mici perturbaţii în soluţia matematică, atunci problema se numeşte *bine condiţionată.*

Dacă mici perturbaţii chiar numai într-o parte din datele problemei conduce la mari perturbaţii în soluţia

matematică, atunci problema se numește *slab condiționată*.

Fie D datele exacte care caracterizează o problemă și G funcția matematică care conduce la obținerea soluției exacte $G(D)$. Aceasta poate fi scrisă sub forma unei aplicații

$$D \xrightarrow{G} G(D). \tag{1.16}$$

În cazul în care datele sînt afectate de erori sau perturbate, se lucrează cu date perturbate D^*. Soluția exactă a problemei cu date perturbate este prezentată sub forma

$$D^* \xrightarrow{G} G(D^*). \tag{1.17}$$

Aplicația caracterizată de datele D este bine condiționată dacă D^* este apropiată de D și $G(D^*)$ este apropiată de $G(D)$ într-un anumit sens, altfel aplicația este slab condiționată.

Pentru a măsura distanța dintre D^* și D, precum și distanța dintre $G(D^*)$ și $G(D)$, sînt necesare o serie de cunoștințe despre forma datelor și forma soluției. În cazul în care D și $G(D) \in C$ (sînt numere complexe), atunci se va examina $|D - D^*|$ și $|G(D) - G(D^*)|$. În cazul în care D și $G(D)$ sînt vectori sau matrice, se va examina distanța dintre D și D^*, respectiv $G(D)$ și $G(D^*)$ cu ajutorul normei: $\|D - D^*\|$ și $\|G(D) - G(D^*)\|$.

Fie G^* un algoritm de calcul pentru rezolvarea unei probleme $\mathcal{P}$ caracterizate de datele D. Un algoritm G^* este stabil dacă există D^* apropiată de D astfel că $G(D^*)$ este apropiată de $G^*(D)$ într-un anumit sens, altfel algoritmul este instabil.

Elementul ce caracterizează un algoritm de calcul stabil este faptul că soluția obținută cu ajutorul algoritmului este apropiată într-un anumit sens de soluția exactă a problemei ușor perturbate.

Desigur nu se așteaptă ca un algoritm stabil să rezolve o problemă slab condiționată cu mare precizie; acest lucru depinde de precizia datelor. De asemenea un algoritm

instabil aplicat unei probleme bine condiționate poate conduce la rezultate imprecise și desigur se impune evitarea aplicării unui algoritm de calcul instabil la soluționarea numerică a unei probleme slab condiționate.

Dacă se consideră o aplicație oarecare, în care D reprezintă datele exacte, $G(D)$ soluția exactă a aplicației considerate și $A^*(D)$ soluția obținută de calculator (cu toate erorile posibile incluse), atunci se presupune că există un set de date perturbate D^* pentru care $G(D^*) = A^*(D)$. Astfel relația

$$\| G(D^*) - G(D) \| = f(\| D^* - D \|) \tag{1.18}$$

evidențiază perturbațiile din soluție exprimate în funcție de perturbațiile din date, precum și modul de analiză a erorilor, care constă din găsirea unei metode prin care toate sursele posibile de erori luate împreună pot fi evidențiate printr-o perturbație în problema originală. Dacă A^* este puțin mai general decît G^* și D trebuie să fie perturbate în sensul introducerii în calculator, atunci G^* operează pe date perturbate și produce $G^*(D)$.

Pentru o reprezentare grafică a elementelor introduse se consideră spațiul datelor și spațiul soluțiilor, obținîndu-se aplicațiile din fig. 1.3.

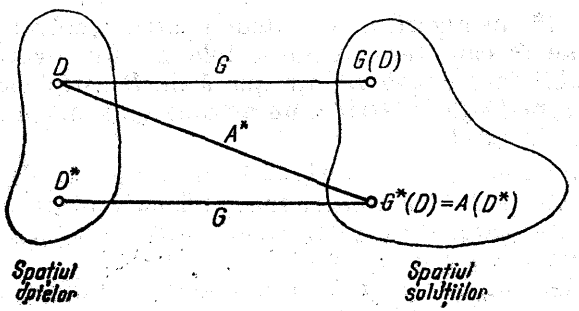


Fig. 1.3

1.5. Metode de investigare privind precizia rezultatelor

Erorile introduse în rezultatele obținute cu ajutorul unor rutine pot fi investigate sau prin metode manuale sau prin metode experimentale, metoda manuală fiind cea mai dificilă. Obiectivul metodei manuale este de a găsi în mod riguros cîte o margine pentru eroare absolută și eroarea relativă :

$$|f^*(x) - f(x)|, \quad \left| \frac{f^*(x) - f(x)}{f(x)} \right|.$$

Pentru găsirea unor astfel de margini este necesară o analiză pas cu pas a procedurii de calcul utilizate pentru evaluarea lui $f^*(x)$, ținîndu-se seama de caracteristicile calculatorului folosit. Acest tip de analiză evidențiază eroarea în $f^*(x)$ la o perturbație a argumentului x. Scopul este de a găsi o margine pentru $|x^* - x|$ sau pentru $\left| \frac{x^* - x}{x} \right|$, unde x^* satisface relația $f^*(x) = f(x^*)$. Evident o astfel de analiză este dificilă și anevoioasă.

Este mult mai des întîlnită investigarea experimentală cu ajutorul unei rutine de evaluare a funcțiilor. Cel mai simplu tip de test experimental constă din calculul lui $f^*(x)$ pentru valori selectate ale argumentului x, verificîndu-le cu valorile cunoscute ale lui $f(x)$. Astfel de testări care implică analiza rezultatelor nu pot fi foarte complete.

Procesul testării rezultatelor în mod experimental poate fi făcut automat prin scrierea unor programe de test, utilizînd un generator de numere aleatoare care calculează un șir de n argumente de test $x_1, x_2, \ldots, x_n$. Pentru fiecare x_k se calculează : $f^*(x_k)$ — aproximația lui $f(x_k)$ obținută prin rutina de evaluare a funcției ce se testează; $f^{**}(x_k)$ — o altă aproximare a lui $f(x_k)$ care este suficient de precisă pentru a fi utilizată ca o valoare de verificare pentru $f^*(x_k)$, [$f^{**}(x_k)$ trebuie calculată într-o precizie superioară precizei în care a fost calculată $f^*(x_k)$, respectiv în dublă precizie].

Calculele statistice vor oferi informații imediate privind mărimea erorii absolute sau relative conținute în

$f^*(x_k)$. De obicei, testele statistice utilizate dau maximul erorii relative, respectiv rădăcina mediei pătratice a erorii relative, adică

$$\max_{1\leqslant k\leqslant n}\left|\frac{f^*(x_k)-f^{**}(x_k)}{f^{**}(x_k)}\right| \text{ și } \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left[\frac{f^*(x_k)-f^{**}(x_k)}{f^{**}(x_k)}\right]^2}. \tag{1.19}$$

Valoarea lui n, adică a numărului argumentelor de test poate fi foarte mare, la unele programe $n = 100\,000$ [42, 56].

1.6. Elemente necesare la proiectarea rutinelor de calcul

Adesea în aplicațiile executate cu ajutorul unui calculator numeric se folosesc programe standard pentru evaluarea funcțiilor ca : e^x, $\sqrt{x}$, $\sin x$, $\sqrt[3]{x}$ etc. Rolul acestui paragraf este prezentarea metodelor matematice utilizate în scrierea programelor de calcul pentru unele funcții.

Orice sistem de calcul dispune de o bibliotecă de rutine standard pentru evaluarea unor funcții ca tg x, sh x, arccos x, $\ln x$, e^x etc. Utilizatorii calculatoarelor electronice îmbogățesc aceste biblioteci cu noi rutine pentru calculul unor funcții care apar cu o anumită frecvență în programele lor. Aceste programe standard pentru evaluarea funcțiilor sînt scrise în limbaje ca ALGOL, FORTRAN, PL/I etc., în funcție de compilatoarele de care dispune sistemul de calcul considerat. În anumite biblioteci rutinele pentru evaluarea aceleiași funcții pot fi în : aritmetică simplă precizie, dublă precizie sau în aritmetică complexă simplă precizie sau aritmetică complexă dublă precizie. Există programe pentru evaluarea unor funcții din matematica aplicată cum ar fi funcțiile Bessel, integralele Fresnel, integralele eliptice etc. Trebuie menționat că tipul de aritmetică (virgulă fixă sau virgulă mobilă) utilizat în rutinele pentru evaluarea funcțiilor, afectează selectarea procedurilor de calcul folosite.

În mod curent calculele de precizie sînt efectuate utilizîndu-se aritmetica în virgulă mobilă. La proiectarea rutinelor pentru evaluarea funcțiilor trebuie avute în vedere o serie de elemente. Se scrie un program de evaluare a funcției $f(x)$ pentru anumite valori ale argumentului x. Prin $f^*(x)$ se notează aproximarea lui $f(x)$ cu ajutorul calculatorului. Elemente ce trebuie avute în atenție la proiectarea rutinelor sînt :

● *Precizie.* Programele de interes general trebuie să aibă precizia de un cuvînt, adică pentru fiecare argument x, $f^*(x)$ va reprezenta valoarea lui $f(x)$ rotunjită la precizia de un cuvînt al calculatorului considerat.

● *Viteză de calcul și lungimea rutinei.* În general se cere ca viteza de execuție a unei rutine să fie mare și dimensiunea rutinei să fie redusă. Timpul de execuție al multor rutine poate fi redus dacă crește posibilitatea ei de memorare. Aceste elemente fac dificilă operația găsirii unui echilibru între viteză și lungime deoarece o rutină perfectă pentru un tip de calculator poate să fie imperfectă pentru alt calculator.

● *Argumente speciale.* Adesea există un argument a pentru care $f^*(a) = f(a)$. De obicei acest lucru are loc numai dacă $f(a)$ se întîmplă să fie un număr întreg astfel că eroarea în $f^*(x)$ poate fi ușor recunoscută. De exemplu dacă $f(x) = \sin x$, se va găsi că $f(0) = \sin 0 = 0$. La fel dacă $f(x) = \sqrt{x}$, atunci $f^*(0) = 0$ și $f^*(1) = 1$.

● *Argumente invalide.* Apar situații în cazul rutinei de evaluare a funcției, cînd argumentul este invalid. Datorită acestui fapt trebuie executate o serie de teste asupra rutinei pentru a determina invaliditatea argumentului. Dacă apare un argument invalid, poate apărea un mesaj sau nu, calculul poate sau nu să fie terminat. Dacă calculul nu este terminat, trebuie folosite anumite proceduri standard care să permită continuarea calculului.

De exemplu dacă un argument $x < 0$ a fost detectat în evaluarea unei funcții ca $f(x) = \sqrt{x}$, calculul poate fi continuat prin calcularea lui $f^*(|x|)$. De asemenea, în cazul unei rutine pentru evaluarea lui $f(x) = e^x$, dacă un argument este invalid pentru că $e^x > L$ (unde L este cel mai mare număr care poate fi reprezentat în calculatorul considerat), calculul poate fi continuat prin scrierea lui $f(x) = L$ și afișarea unei depășiri ce poate fi testată în afara rutinei.

În cadrul unei colecții de rutine pentru evaluarea funcțiilor este necesară introducerea unei asigurări privind mesajele de eroare și procedurile pentru terminarea calculelor.

● *Limitele lui* $f^*(x)$. Este posibil a se cere ca $f^*(x)$ să satisfacă anumite inegalități. De exemplu, dacă $f(x) = \sin x$, este necesar ca $-1 \leqslant f^*(x) \leqslant 1$.

● *Monotonie.* Dacă $f(x)$ este monoton crescătoare sau descrescătoare, trebuie ca și $f^*(x)$ să satisfacă aceeași proprietate. Adică, dacă $f(p) < f(q)$ pentru $p < q$, atunci se impune și $f^*(p) < f^*(q)$ pentru $p < q$. Anumite rutine pentru evaluarea unor funcții sînt scrise să satisfacă condiția de monotonie.

● *Limbajul de programare.* Pentru considerente de eficiență rutinele de evaluare a funcțiilor sînt de obicei programate în limbaj ASEMBLER. De exemplu, la evaluarea rădăcinii pătrate în virgulă mobilă, trebuie separată mantisa și exponentul, operație destul de dificilă în alte limbaje evoluate și destul de simplă în ASEMBLER.

● *Compatibilitatea.* Dacă două rutine sînt scrise pentru valuarea aceleiași funcții, li se impune a fi compatibile în anumit sens. Aceeași rutină poate fi scrisă pentru același calculator dar în limbaje diferite: ALGOL sau FORTRAN. Ele pot fi diferite dar compatibile pe același calculator, dacă pentru același argument x, $f^*(x)$ dintr-un program să coincidă cu $f^*(x)$ din alt program.

Din cele prezentate se evidențiază faptul că la construirea unui algoritm trebuie să se țină seamă de considerațiile matematice și de calculator (cum ar fi erorile de rotunjire, acumularea și propagarea erorilor, necesarul de memorie, timpul de execuție etc.).

1.7. Noțiuni generale privind metodele iterative

Utilizarea calculatoarelor electronice oferă posibilitatea folosirii metodelor iterative la rezolvarea anumitor tipuri de probleme. În cadrul unei metode iterative se alege o aproximație inițială (număr sau funcție) și succesiv se

îmbunătățește această aproximație a soluției (prin iterare) în așa fel ca șirul de soluții îmbunătățite să conveargă către soluția problemei considerate. Astfel de metode sînt destul de simple și foarte atractive pentru a fi utilizate pe un calculator electronic, cu toate că implică destul de multe operații aritmetice, lucru care nu devine un impediment avînd în vedere viteza de calcul pentru calculatoarele din actuală etapă. O trăsătură importantă a metodelor iterative este că acestea sînt „auto-corectoare".

În timp ce metodele directe, cel puțin teoretic, converg într-un număr finit de etape, metodele iterative necesită un număr infinit de etape pentru convergență.

La realizarea metodelor iterative trebuie avute în vedere următoarele elemente: metoda să fie convergență, să se poată determina viteza de convergență și să stabilească criterii pentru stoparea procesului iterativ în momentul obținerii unei aproximări acceptabile pentru soluția problemei considerate.

Pentru $n \geqslant 1$, fie $g_n(t_0, t_1, t_2, \ldots, t_{n-1})$ o funcție de n variabile. Dîndu-se astfel de funcție și $k+1$ valori de start $x_0, x_1, \ldots, x_k$, se poate defini șirul $x_{k+1}, x_{k+2}, \ldots$ cu ajutorul relației

$$x_{n+1} = g_{n+1}(x_0, x_1, \ldots, x_n). \tag{1.20}$$

O astfel de metodă este numită metodă iterativă *nestaționară.*

Dacă pentru fiecare n, funcția g_{n+1} depinde de cel mult una din variabilele $x_{n-s+1}, x_{n-s+2}, \ldots, x_{n-1}, x_n$, metoda se numește metodă iterativă *într-un pas.*

Dacă funcția g_{n+1} nu depinde de n, metoda se numește *staționară.*

Metoda bisecției și metoda poziției false sînt exemple de metode *nestaționare.* Metoda secantei este un exemplu de metodă staționară în doi pași, în acest caz $k+1=2$ și pentru orice n, $g(x_{n-1}, x_n)$ este definită prin formula

$$g(x_{n-1}, x_n) = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})} = x_n - \frac{(x_n - x_{n-1})f(x_n)}{f(x_n) - f(x_{n-1})}. \tag{1.21}$$

O clasă importantă de metode este clasa metodelor staționare într-un pas. În acest caz se alege o singură valoare de start x_0 și o funcție $g(x)$, iar șirul $x_1, x_2, x_3, \ldots$ este calculat cu ajutorul relației:

$$x_{k+1} = g(x_k), \qquad k = 0, 1, 2, \ldots \tag{1.22}$$

Funcția $g(x)$ se numește uneori funcția iterativă [99].

Considerîndu-se o funcție iterativă $g(x)$, există o serie de întrebări care trebuie puse în legătură cu metoda iterativă respectivă [64, 124]:

- Șirul construit prin (1.22) converge către o limită unică?
- Dacă α este soluția exactă și valoarea de start x_0 este suficient de aproape de α, șirul construit prin (1.22) converge și dacă converge, limita lui este α?
- Dacă șirul construit prin (1.22) converge, atunci converge el către soluția problemei considerate?

Metoda iterativă definită prin (1.22) este consistentă dacă orice soluție a problemei considerate este o limită a șirului construit prin (1.22).

Fie x o soluție a problemei considerate și (1.22) o metodă iterativă consistentă; atunci dacă x_0 și $g(x)$ satisfac oricare din următoarele două condiții:

1) $g(x)$ este continuă și diferențiabilă în intervalul I:

$$\bar{x} - |x_0 - \bar{x}| \leqslant x \leqslant \bar{x} + |x_0 - \bar{x}| \quad \text{și} \quad |g'(x) < M < 1, \quad (\forall)\, x \in I;$$

2) $g(x)$ este continuă și diferențiabilă în intervalul închis $[\bar{x}, x_0]$ și $0 \leqslant g'(x) \leqslant M < 1$, în același interval, atunci șirul $x_0, x_1, x_2, x_3 \ldots$, definit prin (1.22), converge către $\bar{x}$.

Teorema contracției permite o serie de analize asupra metodelor iterativ, precum și consecințele ei [85, 42].

METODE DE CALCUL PENTRU REZOLVAREA ECUAȚIILOR ALGEBRICE NELINIARE, TRANSCENDENTE ȘI A SISTEMELOR NELINIARE

2.1. Introducere

Un număr considerabil de modele matematice asociate fenomenelor fizice se reduc în final la o ecuație de forma

$$\mathbf{f}(\mathbf{x}) = 0, \tag{2.1}$$

unde $\mathbf{f}$ și $\mathbf{x}$ sînt vectori de aceeași dimensiune k. Pentru $k = 1$ rezultă o ecuație cu o singură necunoscută, pentru $k = n$ un sistem de n ecuații cu n necunoscute.

În acest capitol se vor prezenta o serie de metode de calcul pentru determinarea rădăcinilor ecuației (2.1). În unele cazuri există posibilitatea rezolvării acestor ecuații în mod analitic.

Exemple

$$\left.\begin{array}{ll} x^2 - 5x + 6 = 0, & x_1 = 2,\ x_2 = 3; \\ \ln x - 3 = 0, & x = e^3; \\ 10^x - 142 = 0, & x = \lg 142; \\ \sin x - 1 = 0,\ x \in [0, 2\pi], & x = \dfrac{\pi}{2}. \end{array}\right\} \tag{2.2}$$

Astfel de metode de rezolvare a ecuației $f(x) = 0$ devin imposibile cînd expresia analitică a lui f este complicată.

Exemple

$$\left.\begin{array}{ll} x^3 \lg x - 3{,}9 = 0 & e^{-x} - \sin(\pi x/2) = 0 \\ & e^x + \lg x - 3 = 0 \\ \sin x - 3{,}2 \ln x = 0 & x^4 - x - \lg(x^2 - x + 1) = 0 \end{array}\right\} \tag{2.3}$$

Ecuațiile prezentate în (2.3) sînt ecuații transcendente și neliniare, iar pentru rezolvarea lor se folosesc metode grafice sau metode aproximative.

Ecuația (2.1) poate avea rădăcini reale sau complexe. Zerourile funcției $f(x)$ sînt egale cu rădăcinile ecuației (2.1). Rădăcinile reale ale ecuației (2.1) pot fi puse în evidență cu ajutorul metodelor grafice dar rădăcinile complexe nu. Dacă se trasează graficul funcției $f(x)$, rădăcinile reale ale ecuației sînt reprezentate prin punctele unde graficul intersectează axa Ox.

Se vor prezenta în continuare trei modele fizice care conduc la probleme de acest gen.

Exemple. 1. Fie circuitul din fig. 2.1 format dintr-o sursă $V(t)$, o rezistență R și o bobină L legate în serie, unde

$$V(t) = 100\sqrt{2}\sin 5t,\ R = 20\,\Omega,\ L = 4H,\ i|_{t=0} = 0.$$

Expresia curentului în circuitul din fig. 2.1 este dată de relația

$$i(t) = 5e^{-5t}\sin\frac{\pi}{4} + 5\sin\left(5t - \frac{\pi}{4}\right). \tag{2.4}$$

Se cere timpul $t \in (0{,}5\ ;\ 1{,}4)$ pentru care curentul este nul ($i = 0$), iar $e = 2{,}71828$. În aceste condiții relația (2.4) se reduce la o ecuație transcendentă în variabila t, avînd expresia

$$5 \cdot (2{,}71828)^{-5t}\sin\frac{\pi}{4} + 5\sin\left(5t - \frac{\pi}{4}\right) = 0. \tag{2.5}$$

2. În foarte multe situații privind transportul energiei electrice se utilizează izolatori tubulari pentru liniile de înaltă tensiune (fig. 2.2). Care trebuie să fie raportul x al diametrului exterior $D = 2R$ la diametrul interior $d = 2r$, în scopul obținerii unei secțiuni transversale S minime [107, 15]?

Expresia secțiunii transversale S este

$$S(x) = \pi q^2(x^2 - 1)\ /(\ln^2 x), \tag{2.6}$$

unde q este raportul dintre tensiunea liniei și tensiunea maximă admisibilă (de această dată q fiind o constantă în problemă), iar $x = D/d$.

Pentru a determina secțiunea minimă S se derivează $S(x)$ din (2.6) în raport cu x;

$$S'(x) = \pi q^2\,\frac{2x\ln^2 x - \dfrac{2}{x}(x^2 - 1)\ln x}{\ln^4 x}. \tag{2.7}$$

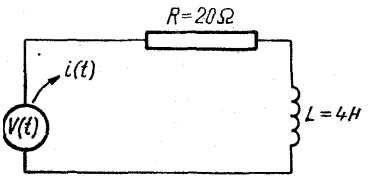


Fig. 2.1

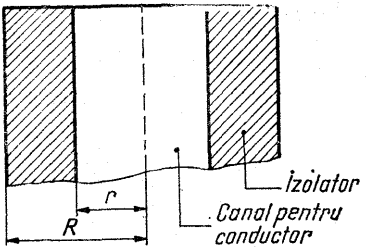


Fig.2.2

făcîndu-se derivata egală cu zero, se obține

$$2x^2 \ln^2 x - 2(x^2 - 1) \ln x = 0,$$

de unde

$$x^2 \ln x - (x^2 - 1) = 0, \tag{2.8}$$

care va conduce la determinarea lui x dorit.

3. Acest exemplu este luat din domeniul hidraulicii, privind curgerea apei într-un canal deschis (fig. 2.3), canalul avînd un unghi de înclinare [88, 89].

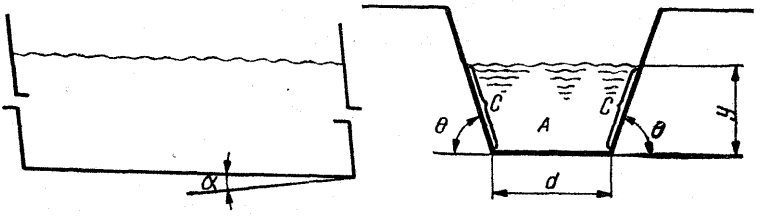


Fig. 2.3

O relație empirică pentru debitul de curgere Q are expresia

$$Q = \frac{1{,}49\, A^{5/3} S^{1/2}}{n P^{2/3}} \text{ sau } Q = \frac{1{,}49}{n} A R^{2/3} S^{1/2}, \tag{2.9}$$

unde $P = 2C + d$ este permimetrul canalului, n este coficientul de rugozitate determinat experimental, care este cuprins între 0,25 și 0,35 pentru multe din canalele rîurilor, A aria secțiunii transversale a canalului, R raza hidraulică, fiind definită de raportul ariei A la perimetrul P al secțiunii canalului, $S = \operatorname{tg} \alpha$ panta canalului (α unghiul de înclinare).

Se consideră un canal cu o secțiune dreptunghiulară și se presupune că sînt date mărimile Q, n, S și d. Să se calculeze adîncimea y a apei în canal.

În relația (2.9) se înlocuiesc elementele canalului considerat și atunci (2.9) devine

$$Q = \frac{1{,}49}{n}\, dy \left(\frac{dy}{d+2y}\right)^{2/3} S^{1/2}$$

iar după ridicarea la cub în ambele părți și ordonare se obține

$$\left(\frac{1{,}49}{n}\right)^3 d^5 S^{3/2} y^5 - 4Q^2 y^2 - 4Q^3 dy - Q^3 d^2 = 0. \qquad (2.10)$$

Se observă că problema a fost redusă la rezolvarea unei ecuații de gradul cinci, care are o singură rădăcină reală pozitivă.

2.1.1. *Metode iterative*

Cele mai simple metode iterative utilizate pentru rezolvarea ecuației (2.1) pot fi scrise sub forma

$$x_{n+1} = g(x_n), \qquad n = 0, 1, 2, \ldots \qquad (2.11)$$

O schemă iterativă se numește convergentă dacă aplicația g îndeplinește pe un domeniu $D \subset C^n$ într-o normă oarecare condiția

$$\|g(x) - g(y)\| \leqslant M \|x - y\|; \quad M < 1 \text{ și } x, y \in D. \qquad (2.12)$$

Se pune problema ce condiții trebuie să îndeplinească funcția g și aproximația inițială x^0 pentru asigurarea convergenței acestui proces iterativ, dat de (2.11).

Fie $f : D_1 \to D_2$, $D_1 \in R^n$ și $D_2 \in R^n$ și $f(x) = 0$ (ecuație sau sistem) care admite în vecinătatea $V_\alpha \in D_1$ o rădăcină unică $x = \alpha$. Prin x^0 se notează aproximația inițială a rădăcinii, $x^0 \in V_\alpha$.

Numim *formulă de iterare* (șir de iterare) pentru determinarea rădăcinii α, un șir de forma

$$x^{(k)} = g_k(x^{(0)}, x^{(1)}, \ldots, x^{(k-1)}) \qquad (2.13)$$

care satisface condițiile $x^{(k)} \to \infty$, cînd $k \to \infty$. Funcția g_k depinde de f și de rangul termenilor din șir. În cazul în

care g nu depinde de rangul termenilor din şir, adică

$$x^{(k)} = g(x^{(0)},\ x^{(1)},\ldots,x^{(k-1)}), \tag{2.14}$$

atunci formula de iterare este de *tip staţionar*.

Marea majoritate a metodelor pe care le vom întîlni sînt de tip staţionar şi de ordinul întîi, şirul iterativ în acest caz are forma

$$x^{(k)} = g(x^{(k-1)}) \tag{2.15}$$

şi prezintă o serie de avantaje din punctul de vedere al spaţiului de memorie faţă de şirul iterativ (2.14), care necesită spaţiu de memorie pentru stocarea a $k+1$ vectori $\mathbf{x}^{(k)}$, iar schema (2.15) necesită spaţiu doar pentru memorarea a doi vectori.

În cadrul metodelor iterative, ecuaţia (sistemul) $f(x)=0$ se pune sub formă echivalentă $x = g(x)$ cel puţin în vecinătatea $V_\alpha \in D_1$, în sensul:

$$f(x) = 0 \Leftrightarrow \alpha = g(\alpha). \tag{2.16}$$

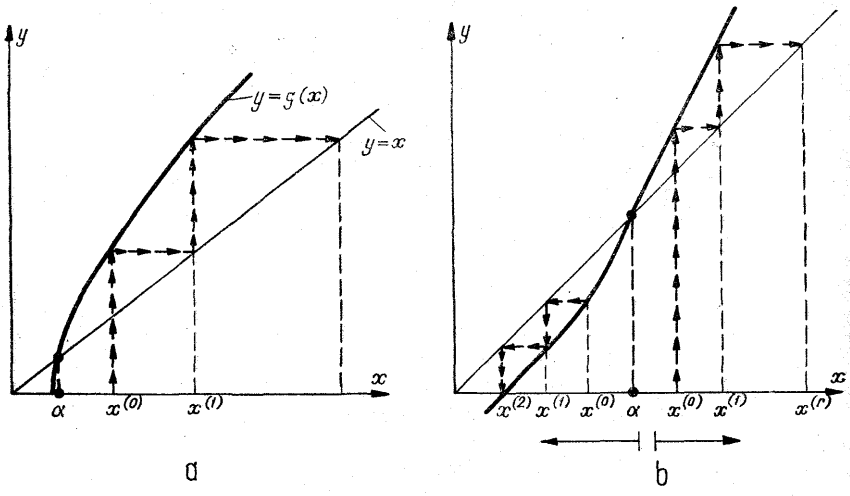
Convergenţa şirului de iterare, existenţa şi unicitatea soluţiei sînt garantate în anumite condiţii de teorema de punct fix pentru contracţii [42].

Fie ecuaţia $x = g(x)$. Dacă $g(x)$ satisface condiţia $|g'(x)| \leqslant \lambda < 1$, în vecinătatea lui α (soluţia unică a ecuaţiei), atunci procesul este convergent. Dacă

$$|g'(x)| \geqslant \lambda > 1, \tag{2.17}$$

atunci condiţia de convergenţă a şirului de iterare nu este îndeplinită şi s-ar putea ca procedeul să fie divergent. O interpretare geometrică a celor afirmate se poate vedea în fig. 2.4. În asemenea cazuri se poate considera ecuaţia echivalentă $x = \varphi(x)$, unde φ este funcţia inversă lui g. Pentru φ este realizată condiţia

$$|\varphi'(x)| = \left|\frac{1}{g'(\varphi(x))}\right| \leqslant \frac{1}{\lambda} = k < 1. \tag{2.18}$$



a Fig. 2.4 *b*

Exemple. Fie următoarea ecuație:

$$f(x) = x^3 - x - 2$$

care are o singură rădăcină reală în intervalul [1, 2]. Se consideră ecuația echivalentă

$$x = x^3 - 2 \equiv g(x)$$

cu $g'(x) = 3x^2$; în intervalul $[1,2]$ avem $3 \leqslant g'(x) \leqslant 6$, de unde se vede că procesul iterativ este divergent. De aceea se va folosi funcția inversă a lui g, adică φ (obținînd o ecuație echivalentă cu $f(x) = 0$):

$$x = \sqrt[3]{x+2} = \varphi(x) \equiv g^{-1}(x).$$

În acest caz

$$\varphi'(x) = \frac{1}{3\sqrt[3]{(x+2)^2}} \quad \text{și} \quad |\varphi'(x)| \leqslant \frac{1}{3\sqrt[3]{9}} < 1, \qquad x \in [1, 2],$$

de unde se vede că în acest caz procesul este convergent dacă se folosește funcția φ, inversa lui g.

O b s e r v a ţ i e. Se vor prezenta în continuare două exemple care sînt interesante prin modul în care se alege valoarea de start $x^{(0)}$.

Fie ecuaţia (fig. 2.5,a)

$$x^3 - 23x^2 + 62x - 40 = 0 \tag{2.19}$$

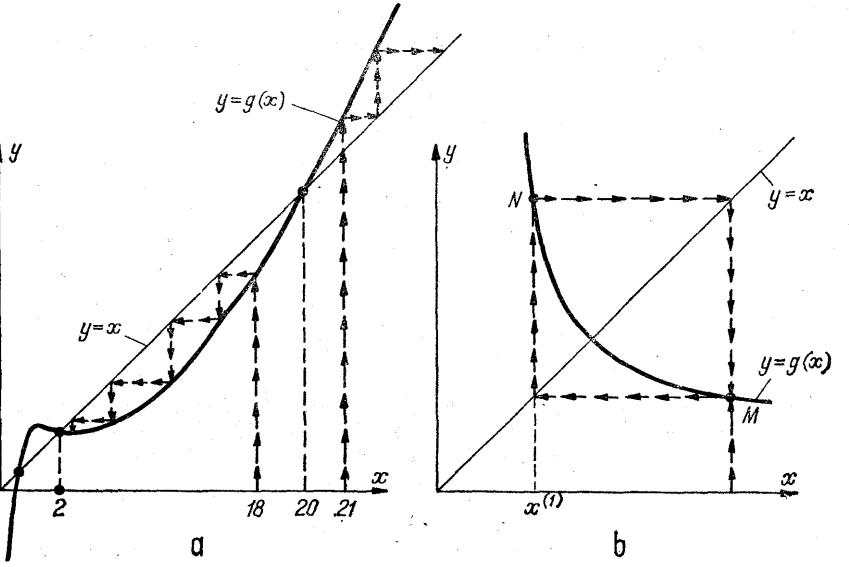


Fig. 2.5

care se scrie sub forma

$$x = g(x) \equiv 23 - \frac{62}{x} + \frac{40}{x^2}. \tag{2.20}$$

Dacă se foloseşte ca valoare de start $x^{(0)} = 21$, rezultă

$$x^{(0)} = 21; \quad x^{(1)} = 23 - \frac{62}{21} + \frac{40}{21^2} \approx 20{,}1;$$

$$x^{(1)} = 20{,}1; \quad x^{(2)} = 23 - \frac{62}{20{,}1} + \frac{40}{20{,}1^2} \approx 20;$$

$$x^{(2)} = 20; \quad x^{(3)} = 23 - \frac{62}{20} + \frac{40}{20^2} \approx 20,$$

deci rădăcina este $x = 20$.

Dar ecuația (2.19) se mai pune și sub forma

$$x = g(x) \equiv \frac{x^2}{23} + \frac{62}{23} - \frac{40}{23x}. \tag{2.21}$$

Dacă se alege valoarea de stat $x^{(0)}=21$, se obține următorul șir iterativ:

$$x^{(0)} = 21; \quad x^{(1)} = g(21) = \frac{21^2}{23} + \frac{62}{23} - \frac{40}{23 \times 21} = 21{,}8;$$

$$x^{(1)} = 21{,}8; \quad x^{(2)} = g(21{,}8) = \frac{21{,}8^2}{23} + \frac{62}{23} - \frac{40}{23 \times 21{,}8} \approx 23{,}2;$$

$$x^{(2)} = 23{,}2; \quad x^{(3)} = g(23{,}2) = \frac{23{,}2^2}{23} + \frac{62}{23} - \frac{40}{23 \times 23{,}2} \approx 26{,}1$$

ș.a.m.d. Se observă că în acest caz șirul iterativ nu converge, totuși dacă se alege $x^{(0)} = 18$, procesul converge și șirul de iterații construit cu (2.21) converge către soluția $x = 2$, care este o altă soluție a ecuației (2.19).

O alegere a valorii de start cît mai aproape de rădăcină de obicei conduce la un șir iterativ convergent, altfel la unul divergent sau care ciclează. De exemplu $|g'(x)| < 1$ în punctul M și în imediata vecinătate a rădăcinii, dar în punctul N avem $|g'(x)| > 1$ (fig. 2.5, b).

Fie un sistem de ecuații neliniare $\mathbf{f}(\mathbf{x}) = 0$ și

$$\mathbf{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \cdot \\ \cdot \\ x_n^{(0)} \end{bmatrix} \tag{2.22}$$

un punct fix din R^n, $r > 0$ un număr real și

$$D = \left\{ \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \; |x_i - x_i^{(0)}| \leqslant r, \; i = 1, 2, \ldots, n \right\} \tag{2.23}$$

Fie, de asemenea, $g_1, g_2, \ldots, g_n$, n funcţii definite pe D cu valori reale, avînd derivate de ordinul întîi continue pe D, iar

$$\varphi_i = \sum_{j=1}^{n} \left| \frac{\partial g_i}{\partial x_j} \right|, \quad K_i = \sup \ \varphi_i(\zeta), \ \lambda = \max_{1 \leqslant i \leqslant n} K_i. \tag{2.24}$$

Dacă

$$| g_i(x_1^{(0)}, \ x_2^{(0)}, \ldots, x_n^{(0)}) - x_i^{(0)} | \leqslant r(1 - \lambda), \tag{2.25}$$

$0 < \lambda < 1$ pentru orice $i = 1, 2, \ldots, n$, atunci sistemul de ecuaţii

$$\left.\begin{aligned} x_1 &= g_1(x_1, x_2, \ldots, x_n) \\ x_2 &= g_2(x_1, x_2, \ldots, x_n) \\ &\cdots\cdots\cdots\cdots \\ x_n &= g_n(x_1, x_2, \ldots, x_n) \end{aligned}\right\}, \tag{2.26}$$

sau scris sub formă vectorială $\mathbf{x} = \mathbf{g}(\mathbf{x})$, va avea în domeniul D o soluţie unică:

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}. \tag{2.27}$$

Dacă $\mathbf{x}^{(0)}$ este un vector oarecare din D şi se construieşte şirul $\{x^{(1)}, x^{(2)}, \ldots, x^{(k)}, \ldots\}$ după formula de recurenţă

$$\mathbf{x}^{(k+1)} = \begin{bmatrix} g_1(x^{(k)}) \\ g_2(x^{(k)}) \\ \vdots \\ g_n(x^{(k)}) \end{bmatrix}, \tag{2.28}$$

atunci au loc relaţiile

$$| x_i^{(k)} - \alpha_i | \leqslant \frac{\max\limits_{1 \leqslant j \leqslant n} | x_j^{(1)} - x_j^{(2)} |}{1 - \lambda} \lambda^{k-1}, \ i = 1, 2, \ldots, n. \tag{2.29}$$

În particular

$$\lim_{k\to\infty} x_i^{(k)} = \alpha_i. \qquad (2.30)$$

2.1.2. *Propagarea erorilor*

Avînd în vedere că toate aceste calcule se execută pe un calculator electronic, nu este posibil ca funcţia $\mathbf{g}(\mathbf{x})$ să fie evaluată exact.

Pentru orice x se poate reprezenta aproximarea lui $\mathbf{g}(\mathbf{x})$ prin $\mathbf{G}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \varepsilon(\mathbf{x})$, unde $\varepsilon(\mathbf{x})$ este eroarea comisă în evaluarea funcţiei

De obicei se poate cunoaşte o margine pentru $\varepsilon(\mathbf{x})$, adică $|\varepsilon(\mathbf{x})| < \varepsilon$. În acest caz schema iterativă care se utilizează poate fi reprezentată astfel:

$$\mathbf{x}^{(k+1)} = \mathbf{g}(\mathbf{x}^{(k)}) + \varepsilon^{(k)}, \quad k = 0,1,2,\ldots, \qquad (2.31$$

unde $\mathbf{x}^{(k)}$ sînt valorile obţinute din calcul şi $\varepsilon^{(k)}$ satisface relaţia

$$|\varepsilon^{(k)}| \leqslant \varepsilon, \qquad k = 0,1,2,\ldots.$$

Este foarte greu de afirmat că şirul iterativ obţinut prin (2.31) este convergent, totuşi în anumite condiţii va fi posibilă găsirea unei soluţii aproximative la o precizie determinată în principal de precizia de calcul ε. Din fig. 2.6 se vede că pentru un caz particular cînd $g(x) = \alpha + \lambda(x - \alpha)$, eroarea în rădăcina α este mărginită prin $\pm\, \varepsilon/(1 - \lambda)$.

Se observă că dacă $\lambda \approx 1$, problema este slab condiţionată.

În cazul în care schema de iteraţie este convergentă, prezenţa erorii în calculul funcţiei $\mathbf{g}(\mathbf{x})$, de mărime mărginită prin ε, face ca schema iterativă să estimeze rădăcina α cu o imprecizie mărginită prin $\pm\, \varepsilon/(1 - \lambda)$.

Fie $\mathbf{x}^{(0)}$ orice valoare astfel ca $|\alpha - \mathbf{x}^{(0)}| < \rho_0$, unde $0 < \rho_0 \leqslant \rho - \dfrac{\varepsilon}{1 - \lambda}$. În acest caz iteraţia $\mathbf{x}^{(k)}$, calcu-

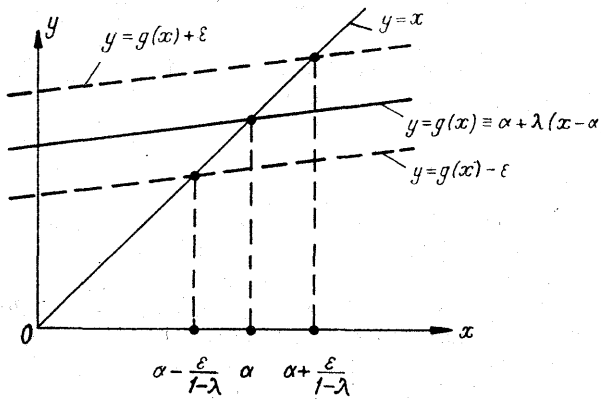


Fig. 2.6

lată prin (2.31) cu o eroare mărginită prin ε, se află [79, 127] în intervalul $|\alpha - \mathbf{x}^{(k)}| \leqslant \rho$ și

$$|\alpha - \mathbf{x}^{(k)}| \leqslant \frac{\varepsilon}{1-\lambda} + \lambda^k \left(\rho_0 - \frac{\varepsilon}{1-\lambda} \right), \quad (2.32)$$

unde $\lambda^k \to 0$, cînd $k \to \infty$.

Din afirmația precedentă se vede că eroarea de calcul care apare în evaluarea lui $\mathbf{g}(\mathbf{x})$ este de cel mult $\varepsilon/(1-\lambda)$. Numărul de iterații necesar [106, 104] este

$$n \approx \lg \left[\frac{\varepsilon}{(1-\lambda)\rho_0} \right] / \lg \lambda \approx \lg \left[\frac{(1-\lambda)\rho_0}{\varepsilon} \right] / \lg \frac{1}{\lambda}. \quad (2.33)$$

Desigur dacă eroarea acceptabilă este mai mare de $\varepsilon/(1-\lambda)$ numărul iterațiilor dat prin (2.33) este estimat adecvat.

2.2. Metode pentru rezolvarea ecuațiilor transcendente și neliniare

Uneori este convenabil să se folosească metode grafice pentru determinarea valorii aproximative a rădăcinilor reale ale ecuației $f(x) = 0$. În continuare vor fi prezentate

două metode grafice care se utilizează mai frecvent în practică.

În cazul în care se utilizează o metodă grafică pentru găsirea soluției unei ecuații de forma

$$f(x) = 0, \tag{2.34}$$

se calculează $y = f(x)$ pentru un număr destul de mare de valori ale argumentului x, iar după aceea cu ajutorul unui plotăr se materializează punctele (x, y), rezultînd o curbă care va trece prin aceste puncte. Punctele unde curba intersectează axa Ox reprezintă o valoare aproximativă a unei rădăcini reale. În cazul în care graficul prezintă o serie de dubii în anumite zone, pentru elucidare se calculează o serie de puncte adiționale. Se vede că în cadrul acestei metode grafice este foarte convenabil a avea diferite scări de reprezentare pentru x și y.

A doua metodă grafică constă în trasarea a două grafice y_1 și y_2, adică se scrie funcția $f(x)$ sub forma diferenței a două funcții $f_1(x)$ și $f_2(x)$ astfel încît

$$f(x) = f_1(x) - f_2(x). \tag{2.35}$$

Evident $f(x) = 0$ dacă și numai dacă $y_1 = y_2$. Metoda constă în trasarea celor două grafice y_1 și y_2, după care se determină punctul (x, y), unde cele două curbe se intersectează. Abscisa x a punctului de intersecție va fi rădăcina reală a ecuației (2.34).

În fig. 2.7 se prezintă metoda grafică pentru ecuația $f(x) \equiv \cos x + 2x - 2 = 0$, $y_1 = \cos x$, $y_2 = 2 - 2x$. Avantajul acestei metode față de metoda precedentă este evidentă, deoarece în foarte multe cazuri este mult mai simplu să trasezi graficele pentru $f_1(x)$ și $f_2(x)$ decît pentru $f(x) = f_1(x) - f_2(x)$.

Adeseori rădăcina aproximativă obținută prin metodele grafice poate fi utilizată ca valoare de start pentru o metodă iterativă, care după un număr de iterație permite obținerea unei soluții mult îmbunătățite.

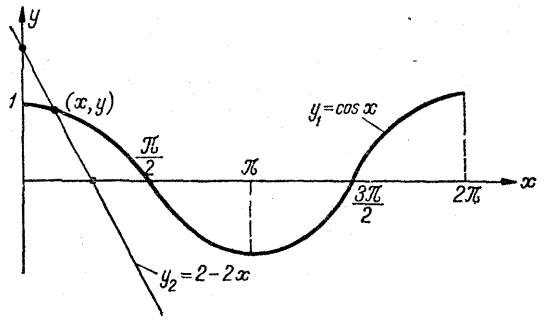


Fig. 2.7

2.2.1. *Metoda bisecției*

Această metodă constă în determinarea unui zero al funcției $f(x)$, cuprins între a și b. Dacă $\alpha \in (a, b)$ și $f(\alpha) = 0$, iar $f(a) \cdot f(b) < 0$, atunci intervalul $[a, b]$ se împarte în două părți egale, după care se testează în care jumătate de interval se află rădăcina ș.a.m.d.

Algoritmul de calcul constă din următoarele etape. Se notează $a_0=a$, $b_0=b$, după care se calculează o estimație inițială

$$c_0 = \frac{a_0 + b_0}{2} = \frac{a + b}{2}.$$

Dacă $f(c_0) = 0$, procesul este terminat, altfel se fac atribuirile

$$a_1 = a_0 \text{ și } b_1 = c_0, \text{ dacă } f(c_0)\, f(a_0) < 0.$$

și

$$a_1 = c_0,\ b_1 = b_0, \text{ dacă } f(c_0)\, f(a_0) > 0.$$

După n iterații se obțin a_n și b_n astfel că $f(a_n)\, f(b_n) < 0$, după care se calculează

$$c_n = \frac{a_n + b_n}{2}.$$

Dacă $f(c_n) = 0$, procesul de calcul este terminat și c_n este rădăcina căutată, altfel se fac atribuirile

$$\left.\begin{aligned} a_{n+1} &= a_n,\ b_{n+1} = c_n \text{ dacă } f(c_n)\ f(a_n) < 0, \\ a_{n+1} &= c_n,\ b_{n+1} = b_n \text{ dacă } f(c_n)\ f(a_n) > 0. \end{aligned}\right\} \quad (2.36)$$

Pentru a arăta convergența acestui proces se poate arăta că șirul $a_0, a_1, a_2, \ldots$ este un șir crescător mărginit superior, iar șirul $b_0, b_1, b_2, \ldots$ este un șir descrescător mărginit inferior. În concluzie cele două șiruri $\{a_n\}_{n\in N}$ și $\{b_n\}_{n\in N}$ converg. Fie α și β limitele celor două șiruri $\{a_n\}$ și $\{b_n\}$ respectiv. Deoarece $a_n \leqslant c_n \leqslant b_n$, atunci din

$$\lim_{n\to\infty} |a_n - b_n| = 0$$

rezultă că $\alpha = \beta$ și

$$\alpha = \lim_{n\to\infty} b_n = \lim_{n\to\infty} c_n = \lim_{n\to\infty} a_n.$$

Datorită faptului că $f(a_n)\ f(b_n) < 0$ pentru orice n, rezultă

$$0 \geqslant \lim_{n\to\infty} [f(a_n)\ f(b_n)] = [\lim_{n\to\infty} f(a_n)]\ [\lim_{n\to\infty} f(b_n)] = [f(\alpha)]^2,$$

dar $[f(\alpha)]^2 \geqslant 0$, de unde rezultă că $f(\alpha) = 0$ și $\lim_{n\to\infty} f(a_n) = 0$ datorită continuității lui $f(x)$.

În practică se lucrează cu valoarea lui $f(x)$ notată prin $\bar{f}(x)$. Pentru determinarea rădăcinii α a funcției $f(x)$ pe intervalul $[a, b]$ se introduc două constante pozitive ε_1 și ε_2, astfel că α este acceptată ca rădăcină dacă $|\bar{f}(\alpha)| \leqslant \varepsilon_1$ sau dacă α se găsește în intervalul $[\beta, \gamma]$ astfel încît

$$\bar{f}(\beta)\bar{f}(\gamma) \leqslant 0, \text{ cu } \gamma - \beta \leqslant \varepsilon_2.$$

Cu aceste două constante pozitive, ε_1 (pentru funcția) și ε_2 (pentru interval) algoritmul de calcul se desfășoară după următoarele etape.

Pentru a și b date se testează dacă

$$|\bar{f}(a)| \leqslant \varepsilon_1 \text{ sau } |\bar{f}(b)| \leqslant \varepsilon_1. \quad (2.37)$$

Dacă una din inegalităţi are sens, atunci a sau b este zeroul funcţiei şi procesul de calcul s-a terminat. De asemenea se testează dacă

$$\bar{f}(a)\,\bar{f}(b) < 0. \tag{2.38}$$

Dacă inegalitatea (2.38) nu este satisfăcută, atunci metoda de înjumătăţire pentru acest caz nu converge şi procesul este oprit. Dacă inegalitatea (2.38) are sens, atunci se testează dacă

$$b - a \leqslant \varepsilon_2. \tag{2.39}$$

Dacă (2.39) are sens, atunci se acceptă ca rădăcină

$$c_0 = \frac{a+b}{2}. \tag{2.40}$$

Dacă (2.39) nu este satisfăcută, atunci se notează $a_0 = a$, $b_0 = b$ şi se determină a_1 şi b_1 cu ajutorul relaţiilor (2.36).

În fig. 2.8 este prezentată o schemă logică însoţită de programul 2.1 în FORTRAN care codifică algoritmul introdus, în scopul găsirii unui zero al funcţiei $f(x)=x^3-11$ cu $a = 2$ şi $b = 3$.

Constantele ε_1 şi ε_2 trebuie alese cu multă atenţie pentru că ele influenţează numărul de iteraţii şi convergenţă. De asemenea la alegerea lui ε_1 şi ε_2 trebuie să ţină seama şi de tipul calculatorului pe care se rulează algoritmul codificat. În acest sens dacă se alege ε_1 mai mic decît valoarea erorii de rotunjire în cazul evaluării lui $f(x)$ în $[a, b]$ şi dacă se alege ε_2 mai mic decît minimul distanţei dintre două numere consecutive reprezentate în calculator [din intervalul $[a, b] \in R$), atunci procesul de calcul devine infinit. Pentru a preveni astfel de necazuri este bine să se testeze la fiecare pas că $c_n \neq a_n$ şi $c_n \neq b_n$ sau altfel se va specifica numărul maxim de iteraţii care poate fi executat, după care procesul este stopat.

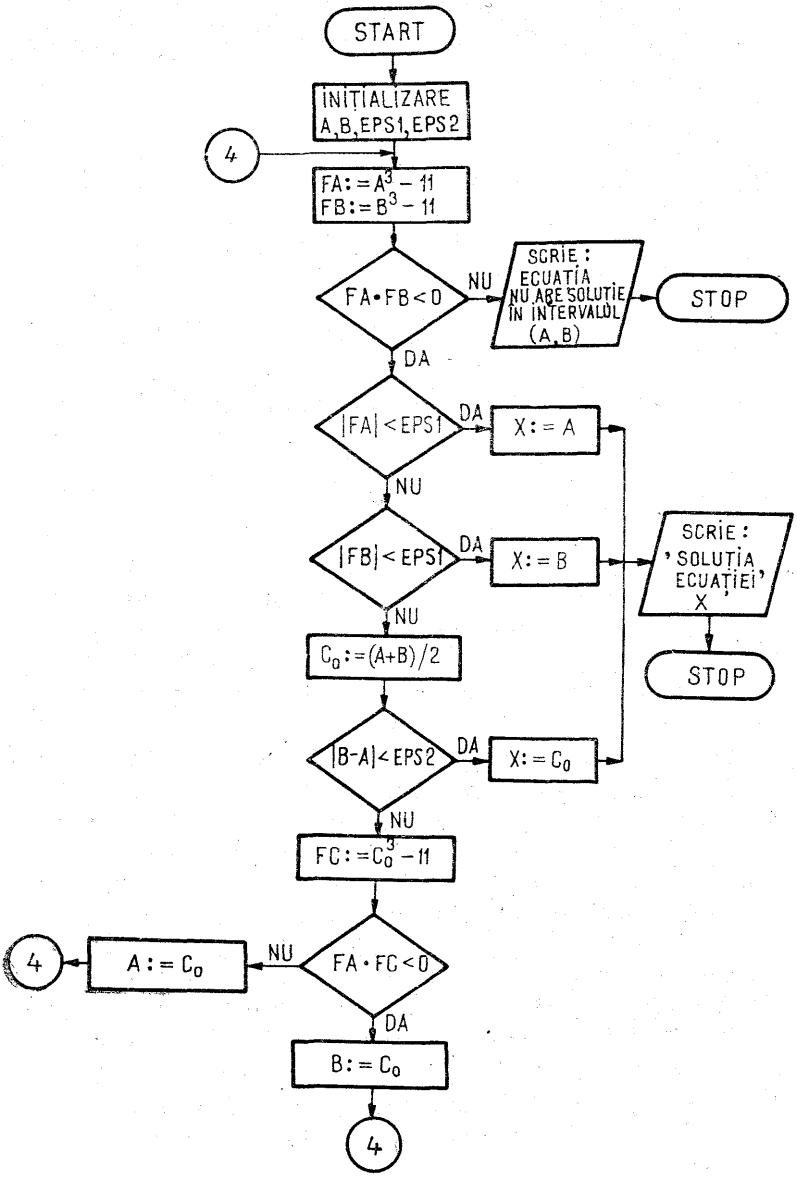

START
INIȚIALIZARE A,B,EPS1,EPS2
4
FA:=A^3-11
FB:=B^3-11
FA•FB<0
NU
SCRIE: ECUAȚIA NU ARE SOLUȚIE ÎN INTERVALUL (A,B)
STOP
DA
|FA|<EPS1
DA
X:=A
NU
|FB|<EPS1
DA
X:=B
SCRIE: 'SOLUȚIA ECUAȚIEI' X
STOP
NU
C0:=(A+B)/2
|B-A|<EPS2
DA
X:=C0
NU
FC:=C0^3-11
FA•FC<0
NU
A:=C0
4
DA
B:=C0
4

Fig. 2.8

```
   METODA BISECTIEI
     DATA A/2./,B/3./,EPS1/1.E-6/,EPS2/1.E-5/
4    FA=A**3-11.
     FB=B**3-11.
     IF(ABS(FA).LT.EPS1) GO TO 6
     IF(ABS(FB).LT.EPS1) GO TO 7
     IF((FA*FB).LT.0.) GO TO 5
     WRITE(108,101)
101  FORMAT(' ','ECUATIA NU ARE SOLUTIE IN INTERVALUL (A,B)')
     GO TO 10
5    CO=(A+B)/2
     IF(ABS(B-A).LT.EPS2) GO TO 9
     FC=CO**3-11.
     IF((FA*FC).LT.0) GO TO 11
     A=CO
     GO TO 4
11   B=CO
     GO TO 4
6    X=A
     GO TO 8
7    X=B
     GO TO 8
9    X=CO
8    WRITE(108,102) X
102  FORMAT(' ','SOLUTIA ECUATIEI ESTE X=',F6.4)
10   STOP
     END
         LINK
         RUN

       SOLUTIA ECUATIEI ESTE X = 2.2240
         EOJ
```

Programul 2.1

2.2.2. *Metoda poziției false (metoda secantei)*

Fie $f(x)$ o funcție continuă definită pe R cu $[a, b] \in R$ astfel că $f(a)\ f(b) < 0$ (deci există cel puțin o rădăcină reală între a și b). Metoda constă în determinarea unei funcții liniare

$$G(x) = Ax + B, \tag{2.41}$$

astfel încît

$$G(a) = f(b),\ G(b) = f(b) \tag{2.42}$$

și se cere α, astfel ca $G(\alpha) = 0$.

Folosind relațiile (2.41), se pot determina constantele A și B cu ajurorul sistemului:

$$\left.\begin{aligned} Aa + B &= f(a) \\ Ab + B &= f(b) \end{aligned}\right\}, \tag{2.43}$$

de unde rezultă

$$A = \frac{f(b) - f(a)}{b - a}, \quad B = \frac{bf(a) - af(b)}{b - a}. \tag{2.44}$$

Astfel expresia funcției liniare $G(x)$, cu A și B determinați, este următoarea :

$$G(x) = \frac{f(b) - f(a)}{b - a}\, x + \frac{bf(a) - af(b)}{b - a}. \tag{2.45}$$

Datorită faptului că $G(\alpha) = 0$, rezultă

$$\alpha = -\frac{B}{A} = \frac{af(b) - bf(a)}{f(b) - f(a).} \tag{2.46}$$

Dacă $a_n = a$, $b_n = b$, $x_n = \alpha$, (2.46) devine

$$x_n = \frac{a_n f(b_n) - b_n(a_n)}{f(b_n) - f(a_n)}. \tag{2.47}$$

În fig. 2.9 se reprezintă grafic modul în care se desfășoară etapele în metoda punctului fals. Graficul funcției $G(x)$ este o dreaptă care trece prin punctele de coordonate $(a, f(a))$ și $(b, f(b))$. Ecuația dreptei care trece prin cele

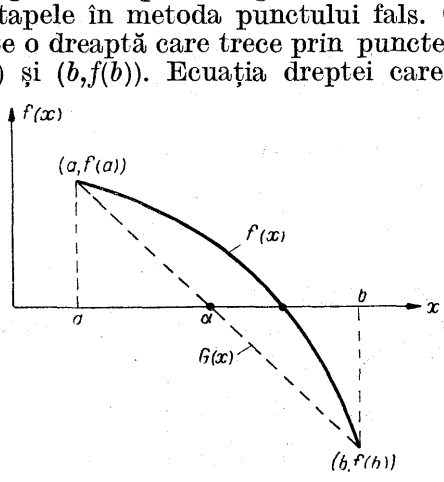


Fig. 2.9

două puncte este

$$\frac{y - f(a)}{x - a} = \frac{f(b) - f(a)}{b - a}. \tag{2.48}$$

Se poate constata ușor [128] că procesul de convergență în metoda punctului fals este mult mai rapid decît în metoda bisecției.

Metoda secantei este convergentă. Șirul $a_0, a_1, a_2, \ldots$ este crescător dar mărginit de b, deci el converge către o limită c. De asemenea șirul $b_0, b_1, b_2, \ldots$ este descrescător și converge către o limită $c_1 \geqslant c$. Presupunînd că c și c_1 nu sînt zerouri ale funcției f, deci $f(c) \neq 0$ și $f(c_1) \neq 0$, va rezulta din continuitatea lui $f(x)$ că pentru $\varepsilon > 0$ și pentru un anumit N

$$|f(a_n)| \geqslant \varepsilon, \qquad |f(b_n)| \geqslant \varepsilon \tag{2.49}$$

pentru orice $n > N$, dar din (2.47) se vede că

$$x_n - a_n = \frac{(a_n - b_n) f(a_n)}{f(b_n) - f(a_n)}, \quad x_n - b_n = \frac{(a_n - b_n) f(b_n)}{f(b_n) - f(a_n)}. \tag{2.50}$$

Funcția f este mărginită pe intevalul $[a, b]$, adică $|f(x)| \leqslant M$, deci pentru orice n suficient de mare

$$|a_{n+1} - b_{n+1}| \leqslant \max(|x_n - a_n|, |x_n - b_n|) \leqslant \frac{|a_n - b_n| M}{M + \varepsilon}, \tag{2.51}$$

prin urmare

$$|a_{N+k} - b_{N+k}| \leqslant \left(\frac{M}{M + \varepsilon}\right)^k |a_N - b_N|. \tag{2.52}$$

În acest caz

$$\lim_{n \to \infty} a_n = \lim_{n \to \infty} b_n = c. \tag{2.53}$$

Datorită faptului că

$$f(a_n) f(b_n) < 0 \tag{2.54}$$

pentru orice n, rezultă ca și în metoda bisecției că sau c sau c_1 este un zerou al funcției $f(x)$ [42, 41]. Metoda secantei este exemplificată prin programul 2.2, dat pentru determinarea rădăcinii ecuației

$$f(x) \equiv x + \lg x - 3 = 0, \quad x_0 = 2, \quad x_1 = 6.$$

```
C    METODA SECANTEI
     DIMENSION X(100)
     DATA XO/2./,X1/6./,PAS/0.05/
     AN=XO
     BN=X1
     I=1
     X(I)=0.
 4   FAN=AN-ALOG10(AN)-3
     FBN=BN-ALOG10(BN)-3
     X(I+1)=(AN*FBN-BN*FAN)/(FBN-FAN)
     IF(ABS(X(I+1)-X(I)).LT.1.E-3) GO TO 6
     AN=AN+PAS
     BN=BN-PAS
     I=I+1
     GO TO 4
 6   WRITE(108,100) X(I+1),I
100  FORMAT(' ',5X,'SOLUTIA ECUATIEI ESTE X =',F6.2,/6X,'NR. ITERATII =
    * ',I3)
     STOP
     END
         LINK
         RUN

       SOLUTIA ECUATIEI ESTE X =  3.55
       NR. ITERATII = 30
         EOJ
```

Programul 2.2

2.2.3. *Metode iterative (metoda lui Newton)*

Fie o ecuație $f(x) = 0$. Se urmărește alegerea unei funcții de iterare $g(x)$, astfel încît convergența corespunzătoare metodei iterative să fie cît mai rapidă. Aici va fi prezentată o schemă pentru alegerea lui $g(x)$. Funcția $f(x)$ se va scrie ca o diferență de două funcții $t(x)$ și $h(x)$, adică

$$f(x) = t(x) - h(x). \tag{2.55}$$

În acest fel se poate ușor rezolva ecuația $t(x) = y$ sau $h(x) = y$ pentru orice y. Astfel putem considera metoda iterativă definită prin

$$t(x_{n+1}) = h(x_n) \tag{2.56}$$

și/sau metoda definită prin

$$h(x_{n+1}) = t(x_n). \tag{2.57}$$

În mod simbolic (2.56) și (2.57) se pot scrie sub forma

$$x_{n+1} = t^{-1}[h(x_n)] \text{ sau } x_{n+1} = h^{-1}[(t(x_n)]. \tag{2.58}$$

Exemplu. Se consideră ecuația

$$f(x) \equiv \cos x - 3x + 2 = 0.$$

Fie $t(x) = \cos\ x$ și $h(x) = 3x - 2$. Atunci

$$f(x) = t(x) - h(x) = \cos\ x - (3x - 2).$$

În acest caz se poate utiliza metoda iterativă dată prin formula

$$x_{n+1} = \cos^{-1}(3x_n - 2) = t_1(x_n)$$

sau

$$x_{n+1} = \frac{2}{3} + \frac{1}{3}\cos\ x_n = t_2(x_n).$$

Se poate verifica fără dificultate că ambele metode sînt convergente.

Dacă se consideră

$$f(x) = t(x) - h(x) \text{ cu } t(x) = x,\ h(x) = x + kf(x), \tag{2.59}$$

unde k este o constantă, atunci din (2.56) rezultă

$$t(x_{n+1}) = x_n + kf(x_n), \text{ deci } x_{n+1} = x_n + kf(x_n). \tag{2.60}$$

Se urmărește minimizarea expresiei $\max\limits_{x} |\ h'(x)\ |$ pentru $\alpha - |\ \alpha - x_n| \leqslant x \leqslant \alpha + |\ \alpha - x_n|$, unde α este o rădăcină a ecuației $t(x) - h(x) = 0$. Datorită faptului că funcția iterativă $g(x) \equiv h(x)$ este cunoscută numai pentru anumite valori ale lui x, se va selecta o valoare $x = a$ astfel încît $g'(a) = 0$, adică

$$0 = g'(a) = 1 + kf'(a), \quad k = -\frac{1}{f'(a)}.$$

Formula iterativă devine

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(a)}. \tag{2.61}$$

Metoda se poate extinde, căutîndu-se o funcţie $t(x)$ astfel ca

$$g(x) = x + t(x)f(x). \tag{2.62}$$

După derivare (2.62) devine

$$g'(x) = 1 + t'(x)f(x) + t(x)f'(x). \tag{2.63}$$

Dacă $g'(\alpha) = 0$ şi $f'(\alpha) \neq 0$, atunci

$$t(\alpha) = -\frac{1}{f'(\alpha)}. \tag{2.64}$$

Datorită faptului că α nu se cunoaşte, o bună alegere este

$$t(x) = -\frac{1}{f'(x)} \text{ şi } g(x) = x - \frac{f(x)}{f'(x)}. \tag{2.65}$$

În acest fel rezultă metoda iterativă

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{2.66}$$

care se numeşte *metoda lui Newton*. Aceasta este o metodă staţionară pas cu pas care are interpretarea geometrică din fig. 2.10. Din punctul $(x_n, f(x_n))$ se duce tangenta

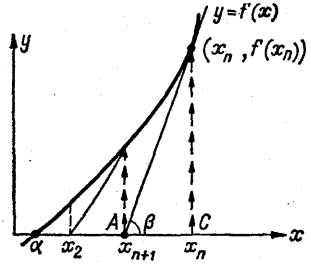


Fig. 2.10

la graficul funcției, tangentă care intersectează axa Ox în punctul $(x_{n+1}, 0)$ și face cu aceasta un unghi β. Astfel se poate scrie

$$\operatorname{tg}\ \beta = \frac{BC}{AC} = \frac{f(x_n)}{x_n - x_{n+1}} = f'(x_n), \tag{2.67}$$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{2.68}$$

Metoda lui Newton este consistentă dacă este definită astfel:

$$x_{n+1} = \begin{cases} x_n - \dfrac{f(x_n)}{f'(x_n)}, & \text{dacă } f'(x_n) \neq 0, \\ x_n, & \text{dacă } f(x_n) = f'(x_n) = 0. \end{cases} \tag{2.69}$$

Se observă că x_{n+1} este nedefinit dacă $f'(x_n) = 0$ și $f(x_n) \neq 0$. Pentru analiza convergenței acestei metode se presupune că $f(x)$ este continuă și de două ori derivabilă în jurul rădăcinii α. Dacă se derivează funcția iterativă $g(x) = x - \dfrac{f(x)}{f'(x)}$ se obține

$$g'(x) = 1 - \frac{[f'(x)]^2 - f(x)\, f''(x)}{[f'(x)]^2} = \frac{f(x) f''(x)}{[f'(x)]^2}. \tag{2.70}$$

Dacă α este o rădăcină simplă a lui $f(x) = 0$, atunci $f'(\alpha) \neq 0$; din continuitatea lui $f(x)$, $|f'(x)| \geqslant \varepsilon$ pentru orice $\varepsilon > 0$, într-o vecinătate V_α a lui α. În interiorul vecinătății V_α se alege un subinterval V'_α astfel ca $|f(x)f''(x)| < \varepsilon^2$ pentru $x \in V'_\alpha \subset V$, lucru posibil deoarece $f(\alpha) = 0$ și deoarece $f(x)$ și $f''(x)$ sînt continue; deci în acest subinterval $g'(x) < 1$, iar metoda este convergentă.

Dacă α este o rădăcină a ecuației $f(x) = 0$, de ordinul de multiplicitate r, atunci un proces de convergență rapidă se poate obține cu schema iterativă dată prin relația

$$x_{n+1} = x_n - r\frac{f(x_n)}{f'(x_n)}. \tag{2.71}$$

Propagarea erorii de la o iterare la alta se pune în evidenţă cu ajutorul relaţiei [37, 21]

$$|x_k - \alpha| = \frac{M}{2} |x_{k-1} - \alpha|^2, \text{ unde } M = \sup_{x \in [\alpha, \beta]} |g''(x)|. \tag{2.72}$$

În inegalitatea (2.72) partea stîngă conţine eroarea absolută în iteraţia k, iar în dreapta se află pătratul erorii absolute în iteraţia $k-1$, fapt care se reflectă în convergenţa pătratică a metodei lui Newton.

Figura 2.11 prezintă diagrama logică, iar programul 2.3 este scris în FORTRAN pentru metoda lui Newton-la găsirea unei rădăcini a ecuaţiei

$$f(x) \equiv x^5 - 6x^4 + 1{,}5\, x^3 - 20\, x^2 + 14\, x - 4 = 0,$$

folosind ca valoare de start $x_0 = 0{,}8$ şi algoritmul (2.66). De asemenea se aplică din nou metoda lui Newton pentr. aceeaşi ecuaţie şi în fig. 2.12, cu programul 2.4, de această dată folosind algoritmul (2.71) pentru $r = 2$. Se vor interpreta rezultatele obţinute cu cele două programe.

2.2.4. *Metoda lui Müller*

Această metodă este o extindere a metodei secantei. Dîndu-se trei puncte distincte x_1, x_2, x_3, se construieşte un polinom $T(x)$ de gradul doi astfel că

$$T(x_1) = f(x_1),\ T(x_2) = f(x_2),\ T(x_3) = f(x_3). \tag{2.73}$$

Se găsesc zerourile acestui polinom de gradul doi iar unul din ele se alege a fi o nouă aproximaţie x_4 a rădăcinii. Procesul se repetă cu x_2, x_3 şi x_4, construind din nou expresia lui $T(x)$, după rezolvare se găsesc două rădăcini, una din ele se alege ca fiind o nouă aproximaţie a rădăcinii ş.a.m.d.

Pentru construcţia lui $T(x)$, se caută coeficienţii a_0, a_1, a_2 astfel ca

$$T(x) = a_2 x^2 + a_1 x + a_0, \tag{2.74}$$

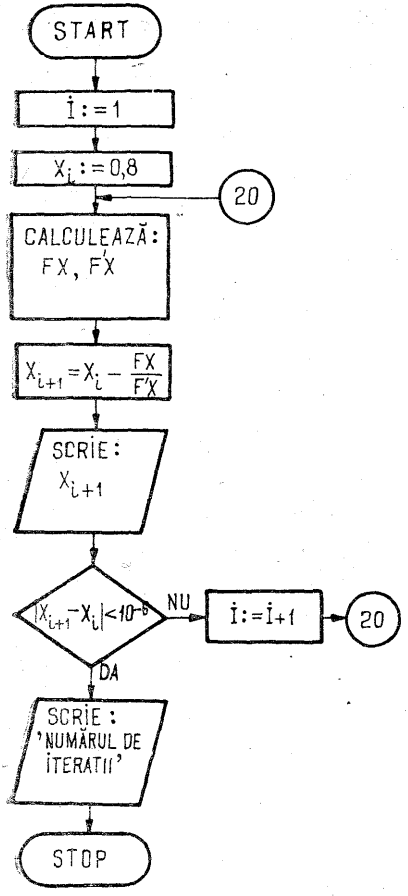


Fig. 2.11

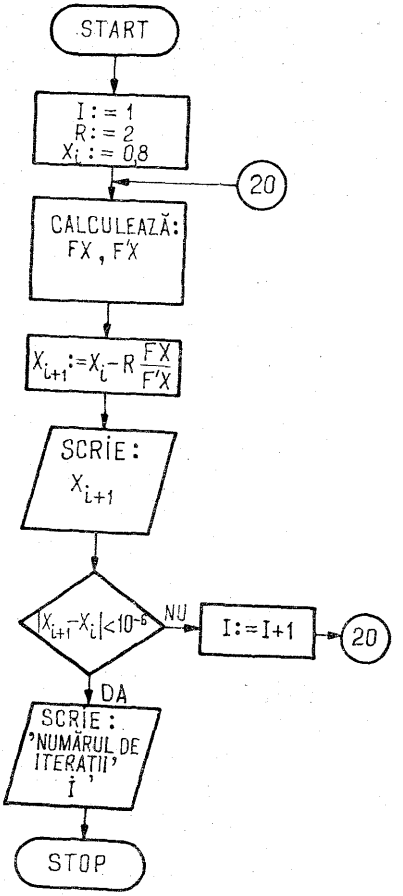


Fig. 2.12

coeficienții ce se pot determina din condițiile (2.73), de unde rezultă sistemul

$$\left.\begin{aligned} a_0 + a_1x_1 + a_2x_1^2 &= f(x_1) \\ a_0 + a_1x_2 + a_2x_2^2 &= f(x_2) \\ a_0 + a_1x_3 + a_2x_3^2 &= f(x_3) \end{aligned}\right\} \qquad (2.75)$$

```
      DIMENSION X(500)
      I=1
      X(I)=0.8
20    FX=X(I)**5-6.*X(I)**4+15.*X(I)**3-20.*X(I)**2+14.*X(I)-4
      FDX=5.*X(I)**4-24.*X(I)**3+45.*X(I)**2-40.*X(I)+14
      X(I+1)=X(I)-FX/FDX
      WRITE(108,100) X(I+1)
100   FORMAT(' ',5X,'X =',F6.3)
      IF(ABS(X(I+1)-X(I)).LT.1.E-6) GO TO 10
      I=I+1
      GO TO 20
10    WRITE(108,101) I
101   FORMAT(' ',5X,'NUMARUL DE ITERATII =',I5)
      STOP
      END
          LINK
          RUN


        X =  .889
        X =  .941
        X =  .970
        X =  .985
        X =  .992
        X =  .996
        X =  .999
        X = 1.000
        X = 1.071
        X = 1.035
        X = 1.071
        X = 1.009
        X = 1.004
        X = 1.004
        X = 1.004
        NUMARUL DE ITERATII = 15

          EOJ
```

Programul 2.3

```
      DIMENSION X(500)
      R=2.
      I=1
      X(I)=0.8
20    FX=X(I)**5-6.*X(I)**4+15.*X(I)**3-20.*X(I)**2+14.*X(I)-4
      FDX=5.*X(I)**4-24.*X(I)**3+45.*X(I)**2-40.*X(I)+14
      X(I+1)=X(I)-R*(FX/FDX)
      WRITE(108,100) X(I+1)
100   FORMAT(' ',5X,'X=',F6.3)
      IF(ABS(X(I+1)-X(I)).LT.1.E-6) GO TO 10
      I=I+1
      GO TO 20
10    WRITE(108,101) I
101   FORMAT(' ',5X,'NUMARUL DE ITERATII =',I5)
      STOP
      END


        X =  .978
        X =  .999
        X = 1.036
        X =  .999
        X =  .988
        X = 1.000
        X = 1.000
        NUMARUL DE ITERATII = 7
```

Programul 2.4

Dacă în (2.74) se face substituţia $x = x_1 + (x - x_1)$, atunci

$$T(x) = a_2[x_3 + (x - x_3)]^2 + a_1[x_3 + (x - x_3)] + a_0 =$$
$$= a_2(x - x_3)^2 + (2a_2x_3 + a_1)(x - x_3) + a_2x_3^2 + a_1x_3 + a_0 =$$
$$= a_2'(x - x_3)^2 + a_1'(x - x_3) + a_0',$$

unde

$$a_2' = a_2,\ a_1' = a_1 + 2a_2x_3,\ a_0' = a_2x_3^2 + a_1x_3 + a_0.$$

Lucrul cu polinomul

$$T(x) = a_2'(x - x_3) + a_1'(x - x_3) + a_0' \qquad (2.76)$$

este mult mai avantajos

Reprezentarea grafică a acestei metode este dată în fig. 2.13, pentru două iteraţii consecutive. Din figură se vede că x_1, x_2, x_3 au fost folosite la determinarea polinomului $T^1(x)$ (curba punctată). Acesta intersectează Ox în x_4, care se alege ca o valoare îmbunătăţită a rădăcinii căutate. În continuare punctele x_2, x_3, x_4 se vor folosi la construirea polinomului $T^2(x)$ ş.a.m.d.

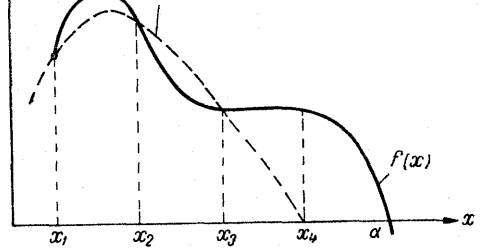


Fig. 2.13

Metoda lui Müller este o metodă iterativă staţionară în trei paşi care implică existenţa a trei valori iniţiale. Astfel se poate nota

$$x_{k+1} = g(x_{k-2},\ x_{k-1},\ x_k),\ k = 3,\ 4, \ldots, \qquad (2.77)$$

unde pentru cazul inițial se ia

$$g(x_3,\ x_2,\ x_1) = x_3 + (x_4 - x_3) = x_4, \tag{2.78}$$

x_4 fiind o valoare aproximativă îmbunătățită pentru α. O analiză completă asupra metodei lui Müller poate fi găsită în [127, 67].

Metoda lui Müller are o serie de avantaje asupra metodei lui Newton și metodei poziției false:

- nu cere existența derivatelor funcției $f(x)$;
- poate fi executată utilizînd numai valorile funcției $f(x)$;
- poate fi aplicată foarte bine la funcții care nu sînt date sub formă explicită.

Un dezavantaj al metodei lui Müller față de metoda lui Newton este problema convergenței în anumite cazuri [106, 107].

2.3. Metode de rezolvare a sistemelor de ecuații neliniare

În cadrul acestui paragraf se vor prezenta o serie de aspecte și metode privind rezolvarea sistemelor de ecuații neliniare.

Fie sistemul de ecuații neliniare de forma

$$f_i(x_1\ x_2,\ldots,x_n) = 0,\ i = 1,\ 2,\ldots,\ n, \tag{2.79}$$

care se poate scrie vectorial sub forma

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}, \tag{2.80}$$

unde $\mathbf{F}: R^n \to R$ este o aplicație vectorială ale cărei componente sînt f_i, adică $\mathbf{F}^T(\mathbf{x}) = (f_1(\mathbf{x}),\ f_2(\mathbf{x}),\ldots,f_n(\mathbf{x})]$.

Relația (2.79) se mai poate scrie dezvoltat și sub forma

$$\left.\begin{aligned} f_1(x_1,\ x_2,\ldots,\ x_n) &= 0\\ f_2(x_1,\ x_2,\ldots,\ x_n) &= 0\\ \cdots\cdots\cdots\cdots\cdots\\ f_n(x_1,\ x_2,\ldots,\ x_n) &= 0 \end{aligned}\right\} \tag{2.81}$$

sau

$$\mathbf{x} = \mathbf{g}(\mathbf{x}). \tag{2.82}$$

Soluția (sau rădăcina) unui astfel de sistem este un vector α, $\alpha^T = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, care este un punct din spațiul R^n. Folosind un vector de start $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$ și funcția iterativă

$$\mathbf{x}^{(k+1)} = \mathbf{g}(\mathbf{x}^{(k)}), \quad k = 0, 1, 2, \ldots, \tag{2.83}$$

se găsește vectorul α prin trecere la limită.

● Fie $\mathbf{g}(\mathbf{x})$ o funcție vectorială care satisface condiția

$$\| \mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y}) \| \leqslant \lambda \| \mathbf{x} - \mathbf{y} \| \tag{2.84}$$

pentru orice vectori $\mathbf{x}, \mathbf{y} \in R^n$, astfel ca $\| \mathbf{x} - \mathbf{x}^{(0)} \| \leqslant d$, $\| \mathbf{y} - \mathbf{x}^{(0)} \| \leqslant d$, cu constanta lui Lipschitz

$$0 \leqslant \lambda \leqslant 1,$$

vectorul inițial de iterație $\mathbf{x}^{(0)}$ satisface condiția

$$\| \mathbf{g}(\mathbf{x}^{(0)}) - \mathbf{x}^{(0)} \| \leqslant (1 - \lambda)\, d. \tag{2.85}$$

Atunci toți vectorii obținuți prin (2.83) satisfac inegalitatea

$$\| \mathbf{x}^{(k)} - \mathbf{x}^{(0)} \| \leqslant d; \tag{2.86}$$

iar șirul de vectori obținuți prin procesul iterativ (2.83) converge către vectorul α, adică

$$\lim_{k \to \infty} x^{(k)} = \alpha, \tag{2.87}$$

unde α este rădăcina sistemului (2.80).

Teoremă. *Fie* $\mathbf{x} = \mathbf{g}(\mathbf{x})$ *ecuație vectorială care are soluția* $\mathbf{x} = \alpha$, *iar componentele* $g_i(\mathbf{x})$ *ale vectorului* $\mathbf{g}$ *au derivate de ordinul întîi care satisfac relația*

$$\left| \frac{\partial g_i(\mathbf{x})}{\partial x_j} \right| \leqslant \frac{\lambda}{n}, \quad \lambda < 1, \tag{2.88}$$

pentru toți vectorii x care *satisfac relația*

$$\| \mathbf{x} - \alpha \|_\infty \leqslant d. \tag{2.89}$$

Atunci :

a) *toți vectorii* $\mathbf{x}^{(k)}$ *obținuți din procesul iterativ* (2.83), *pentru orice vector* $x^{(0)}$ *care verifică* (2.89), *satisfac de asemenea* (2.89).;

b) *vectorii obținuți prin procesul iterativ* (2.83), *pentru orice* $\mathbf{x}^{(0)}$ *care verifică* (2.89), *converg către rădăcina* α *a sistemului* (2.82) *care este unică.*

Demonstrație. Pentru doi vectori $\mathbf{x}, \mathbf{y}$ ce satisfac (2.89), se scrie formula Taylor

$$g_i(\mathbf{x}) - g_i(\mathbf{y}) = \sum_{j=1}^{n} \frac{\partial g_i(\zeta^{(i)})}{\partial x_j}(x_j - y_j),\ i=1, 2, \ldots, n, \tag{2.90}$$

unde $\zeta^{(i)}$ este un vector (un punct pe segmentul deschis ce unește punctele corespunzătoare vectorilor $\mathbf{x}$ și $\mathbf{y}$). Utilizînd o normă și (2.88), se obține

$$| g_i(\mathbf{x}) - g_i(\mathbf{y}) | \leqslant \sum_{j=1}^{n} \left| \frac{\partial g_i(\zeta^{(i)})}{\partial x_j} \right| | x_j - y_j | \leqslant$$

$$\leqslant \| \mathbf{x} - \mathbf{y} \|_\infty \sum_{j=1}^{n} \left| \frac{\partial g_i(\zeta^{(i)})}{\partial x_j} \right| \leqslant \lambda \| \mathbf{x} - \mathbf{y} \|_\infty. \tag{2.91}$$

Deoarece inegalitatea are loc pentru orice i, rezultă

$$\| g(\mathbf{x}) - g(\mathbf{y})_\infty \| \leqslant \lambda \| \mathbf{x} - \mathbf{y} \|_\infty. \tag{2.92}$$

Astfel (2.92) evidențiază faptul că $\mathbf{g}(\mathbf{x})$ este continuă Lipschitz în domeniul indicat prin (2.89), relativ la norma folosită.

Se observă că pentru orice $\mathbf{x}^{(0)}$ din domeniul (2.89) avem

$$\| \mathbf{x}^{(1)} - \alpha \|_\infty = \| \mathbf{g}(\mathbf{x}^{(0)}) - \mathbf{g}(\alpha) \|_\infty \leqslant \lambda \| \mathbf{x}^{(0)} - \alpha \|_\infty \leqslant \lambda d$$

și astfel vectorul $\mathbf{x}^{(1)}$ se găsește în domeniul (2.89). În mod evident

$$\|\mathbf{x}^{(k)} - \alpha\|_\infty = \|\mathbf{g}(\mathbf{x}^{(k-1)}) - \mathbf{g}(\alpha)\|_\infty \leqslant \lambda\|\mathbf{x}^{(k-1)} - \alpha\|_\infty \leqslant$$
$$\leqslant \lambda\|\mathbf{g}(\mathbf{x}^{(k-2)}) - \mathbf{g}(\alpha)\| \leqslant \lambda^2\|\mathbf{x}^{(k-2)} - \alpha\|_\infty \leqslant \ldots \leqslant \lambda^k d,$$

deci toți vectorii $\mathbf{x}^{(k)}$ se găsesc în domeniul (2.89).

Convergența șirului de vectori obținuți cu ajutorul relației iterative (2.83) rezultă din condiția $\lambda < 1$, deci

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \alpha.$$

2.3.1. *Scheme iterative explicite pentru rezolvarea sistemelor neliniare*

Sistemul de ecuații (2.82) se poate scrie în diverse moduri. Se vor examina variantele pentru $\mathbf{g}(\mathbf{x})$. De exemplu, fie

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} - \mathbf{A}(\mathbf{x})\mathbf{F}(\mathbf{x}), \tag{2.93}$$

unde $\mathbf{A}(\mathbf{x}) = (a_{ij}(\mathbf{x}))$ este o matrice pătrată de ordinul n. Dacă $\mathbf{A}(\mathbf{x})\mathbf{F}(\mathbf{x}) = \mathbf{0}$, atunci $\mathbf{F}(\mathbf{x}) = \mathbf{0}$. Dacă se introduce matricea

$$\mathbf{J}(\mathbf{x}) = \left(\frac{\partial f_i(\mathbf{x})}{\partial x_i}\right) \text{ și } \mathbf{A}(\mathbf{x}) \equiv \mathbf{A},$$

al cărei determinant este jacobianul funcțiilor $f_i(\mathbf{x})$ care sînt componentele vectorului $\mathbf{F}$, atunci din relațiile precedente rezultă

$$\mathbf{G}(\mathbf{x}) \equiv \left(\frac{\partial g_i(\mathbf{x})}{\partial x_j}\right) = \mathbf{I} - \mathbf{A}\mathbf{J}(\mathbf{x}). \tag{2.94}$$

În acest caz, folosind teorema dată, șirul vectorilor determinați prin schema iterativă

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{A}\mathbf{F}(\mathbf{x}^{(k)}), \quad k = 0, 1, 2, \ldots, \tag{2.95}$$

va converge pentru $\mathbf{x}^{(0)}$ suficient de apropiat de α, dacă elementele matricei $\mathbf{A}$ sînt suficient de mici, adică în cazul în care $\mathbf{J}(\alpha)$ este nesingulară şi $\mathbf{A}$ este aproximativ inversa lui $\mathbf{J}(\alpha)$.

În metoda lui Newton $\mathbf{A}(\mathbf{x}) = \mathbf{A}$ este aleasă a fi

$$\mathbf{A}(\mathbf{x}) = \mathbf{J}^{-1}(\mathbf{x}) \tag{2.96}$$

cu presupunerea că $\det |\mathbf{J}(\mathbf{x})| \neq 0$, pentru vectorul $\mathbf{x}$ aparţinînd domeniului $\| \mathbf{x} - \alpha \| < d$.

2.3.2. *Metoda lui Newton pentru rezolvarea sistemelor neliniare*

Utilizînd relaţiile (2.93) şi (2.96), iteraţiile pentru metoda lui Newton sînt date de schema

$$\mathbf{x}^{(k+1)} = g(\mathbf{x}^{(k)}) = \mathbf{x}^{(k)} - \mathbf{J}^{-1}(x^{(k)})\mathbf{F}(\mathbf{x}^{(k)}), \tag{2.97}$$

de unde se obţine după ordonare

$$\mathbf{J}(\mathbf{x}^{(k)})(\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}) = \mathbf{F}(\mathbf{x}^{(k)}) \tag{2.98}$$

care este un sistem ce se poate rezolva în raport cu vectorul $(\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)})$. Se poate arăta că aceasta este o metodă de ordinul doi [86].

Pentru o interpretare geometrică a metodei lui Newton pentru sisteme se consideră un sistem neliniar de două ecuaţii cu două necunoscute, unde $\mathbf{x}$ şi $\mathbf{F}$ se aleg astfel:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} f(x, y) \\ g(x, y) \end{bmatrix}, \quad \mathbf{F}(\mathbf{x}) = \mathbf{0}. \tag{2.99}$$

Pentru acest exemplu (2.98) se poate scrie sub forma:

$$\begin{aligned} (x^{k+1} - x^k) f_x(x^k, y^k) + (y^{k+1} - y^k) f_y(x^k, y^k) + f(x^k, y^k) &= 0, \\ (x^{k+1} - x^k) g_x(x^k, y^k) + (y^{k+1} - y^k) g_y(x^k, y^k) + g(x^k, y^k) &= 0. \end{aligned} \tag{2.100}$$

În spațiul $Oxyz$ ecuațiile

$$z = (x - x^k) f_x(x^k, y^k) + (y - y^k) f_y(x^k, y^k) + f(x^k, y^k),$$
$$z = (x - x^k)\, g_x(x^k, y^k) + (y - y^k)\, g_y(x^k, y^k) + g(x^k, y^k) \tag{2.101}$$

reprezintă două plane. Primul plan este tangent la suprafața $z = f(x, y)$ în punctul de coordonate $(x^k, y^k, f(x^k, y^k))$, iar planul reprezentat de cea de-a doua ecuație este tangent la suprafața $z = g(x, y)$ în punctul de coordonate $(x^k, y^k, g(x^k, y^k))$. De aici se vede că punctul de coordonate (x^{k+1}, y^{k+1}) determinat din sistemul (2.100) este punctul rezultat din intersecția celor două plane definite în (2.101) și planul $z = 0$, adică planul xOy.

În concluzie, geometric, pentru metoda lui Newton, prin trecerea de la o dimensiune la două dimensiuni tangenta la curbă se înlocuiește cu planul tangent la o suprafață, iar în cazul n dimensional interpretarea se face prin utilizarea hiperplanelor tangente. Fiecare din ecuațiile

$$z = \sum_{i=1}^{n} (x_i - x_i^{(k)}) \frac{\partial f_j(x^{(k)})}{\partial x_k} + f_j(x^{(k)}),\ j = 1,\ 2,\ \ldots,\ n, \tag{2.102}$$

reprezintă un hiperplan în spațiul $(x_1, x_2, \ldots, x_n, z)$ cu $n + 1$ dimensiuni care este tangent în punctul $(x_1^k, x_2^k, \ldots, x_n^k)$ la hipersuprafața corespunzătoare

$$z = f_i(x_1, x_2, \ldots, x_n).$$

O serie de dificultăți care pot apărea la rezolvarea sistemelor neliniare utilizînd metoda lui Newton pot fi interpretate cu ajutorul acestor considerații geometrice.

O metodă care se poate utiliza la rezolvarea sistemului (2.99) este metoda lui Newton generalizată. Pentru simplicitate se va descrie metoda pentru sistemul

$$\left.\begin{aligned} f(x, y) &= 0 \\ g(x, y) &= 0 \end{aligned}\right\}, \tag{2.103}$$

adică pentru un x şi un y daţi să se determine x' şi y' astfel ca pentru termenii de ordinul întîi în $\Delta x = x' - x$ şi $\Delta y = y' - y$ să avem

$$f(x',\ y') = g(x',\ y') = 0. \tag{2.104}$$

Din formula lui Taylor se obţine

$$\left.\begin{aligned} f(x', y') &= f(x, y) + \Delta x\, f'(x, y) + \Delta y\ f'(x, y) + \dots \\ g(x', y') &= g(x, y) + \Delta x\, g'(x, y) + \Delta y\, g'(x, y) + \dots \end{aligned}\right\} \tag{2.105}$$

Folosind (2.104) şi neglijînd termenii de ordin superior, se obţine sistemul

$$\left.\begin{aligned} f(x, y) + \Delta x\, f'(x, y) + \Delta y\, f'(x, y) &= 0 \\ g(x, y) + \Delta x\, g'(x, y) + \Delta y\, g'(x, y) &= 0 \end{aligned}\right\}.$$

După rezolvare în raport cu Δx şi Δy se obţin

$$\begin{aligned} x' - x = \Delta x &= \frac{-f(x, y)\, g'_y(x, y) + g(x, y)\, f'_y(x, y)}{f'_x(x, y)\ g'_y(x, y) - g'_x(x, y)\ f'_y(x, y)}, \\ y' - y = \Delta y &= \frac{-g(x, y)\, f'_x(x, y) + f(x, y)\, g'_x(x, y)}{f'_x(x,\ y)\ g'_y(x, y) - g'_x(x, y)\, f'_y(x, y)}. \end{aligned} \tag{2.106}$$

Dacă se fac următoarele înlocuiri:

$$\begin{aligned} x' &= x^{k+1}, \quad x = x^k \\ y' &= y^{k+1}, \quad y = y^k \end{aligned} \quad \text{şi} \quad J = \det \begin{bmatrix} f'_x(x, y) & f'_y(x, y) \\ g'_x(x, y) & g'_y(x, y) \end{bmatrix},$$

atunci (2.139) devine

$$\left.\begin{aligned} x^{k+1} &= x^k + \frac{g(x^k, y^k)\, f'_y(x^k, y^k) - f(x^k, y^k)\, g'_y(x^k, y^k)}{J} \\ y^{k+1} &= y^k + \frac{f(x^k, y^k)\, g'_x(x^k, y^k) - g(x^k, y^k) f'_x(x^k, y^k)}{J} \end{aligned}\right\} \tag{2.107}$$

pentru $k = 0,1,2,\dots$

2.3.3. *Alte metode pentru rezolvarea sistemelor neliniare*

Afară de metoda lui Newton, pentru sisteme se mai utilizează destul de des și metoda gradientului [14].

● Considerînd din nou sistemul (2.81), se poate rezolva ecuația i pentru x_i sau explicit în funcție de x_j, sau numeric. Aceasta este *metoda lui Jacobi*, metodă definită prin

$$\left.\begin{aligned} &f_1(x_1^{[k+1)}, \; x_2^{(k)}, \ldots, x_n^{(k)}) = 0 \\ &f_2(x_1^{(k)}, \; x_2^{(k+1)}, \ldots, x_n^{(k)}) = 0 \\ &\cdots\cdots\cdots\cdots\cdots \\ &f_n(x_1^{(k)}, \; x_2^{(k)}, \ldots, x_n^{(k+1)}) = 0 \end{aligned}\right\} \tag{2.108}$$

Această metodă converge către soluția exactă pentru o alegere destul de bună a aproximației inițiale $(x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})$ dacă raza spectrată a matricei

$$\mathbf{A} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix} \tag{2.109}$$

este subunitară. Raza spectrală a matricei $\mathbf{A}$ este valoarea maximă absolută a valorilor proprii, care sînt soluțiile λ ale ecuației polinomiale

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0 \tag{2.110}$$

unde $\mathbf{I}$ este matricea unitate.

Exemplu. Fie sistemul

$$\left.\begin{aligned} x^3 - 3xy^2 + 1 &= 0 \\ 3x^2y - y &= 0. \end{aligned}\right\}. \tag{2.111}$$

Folosind metoda lui Jacobi, se obține

$$x = \sqrt{\frac{y^3}{3y}}, \quad y = \sqrt{\frac{x^3 + 1}{3x}},$$

$$x_{k+1} = \sqrt{\frac{y_k^2}{3}}, \quad y_{k+1} = \sqrt{\frac{(x_k)^3 + 1}{3x_k}}, \quad k = 0,1,2,\ldots$$

schema iterativă, care pentru două valori de start $x = 1$, $y = 1$ permit obținerea unei soluții care converge către soluția exactă.

● *Metoda Gauss-Seidel* utilizează iterațiile disponibile la fiecare pas. Astfel (2.108) devine :

$$f_1(x_1^{(k+1)}, x_2^{(k)}, \ldots, x_n^{(k)}) = 0 \text{ (rezolvată pentru } x^{(k+1)}),$$

$$f_2(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_n^{(k)}) = 0 \text{ (rezolvată pentru } x^{(k+1)}); \quad (2.112)$$

$$f_n(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_n^{(k+1)}) = 0 \text{ (rezolvată pentru } x^{(k+1)}).$$

● *Metoda suprarelaxărilor succesive* implică utilizarea unui factor de relaxare. Procesul iterativ se prezintă sub forma

$$x_i^{(k+1)} = x_i^{(k)} + \omega(\bar{x}_i^{(k+1)} - x_i^{(k)}), \quad (2.113)$$

unde pentru $i = 1, 2, \ldots, n$ se determină $\bar{x}^{(k+1)}$ prin

$$f_i(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, \bar{x}_i^{(k+1)}, x_{i+1}^{(k)}, \ldots, x_n^{(k)}) = 0. \quad (2.114)$$

În mod frecvent se poate mări rapiditatea convergenței față de metoda lui Gauss-Seidel, printr-o alegere adecvată a lui ω.

În fig. 2.14 este prezentată diagrama logică, urmată de programul 2.5 în FORTRAN pentru rezolvarea sistemului

$$\left.\begin{aligned} f(x,y) &\equiv x^3 - 3xy^2 - 2x + 2 = 0 \\ g(x,y) &\equiv 3x^2y - y^3 - 2y = 0 \end{aligned}\right\},$$

folosind valorile de start $x_0 = y_0 = 1$.

De asemenea este prezentat programul 2.6 în FORTRAN și diagrama logică (fig. 2.15) pentru rezolvarea sistemului (2.111), folosind metoda lui Jacobi, Gauss-Seidel, metoda relaxării pentru $\omega = 1{,}2$.

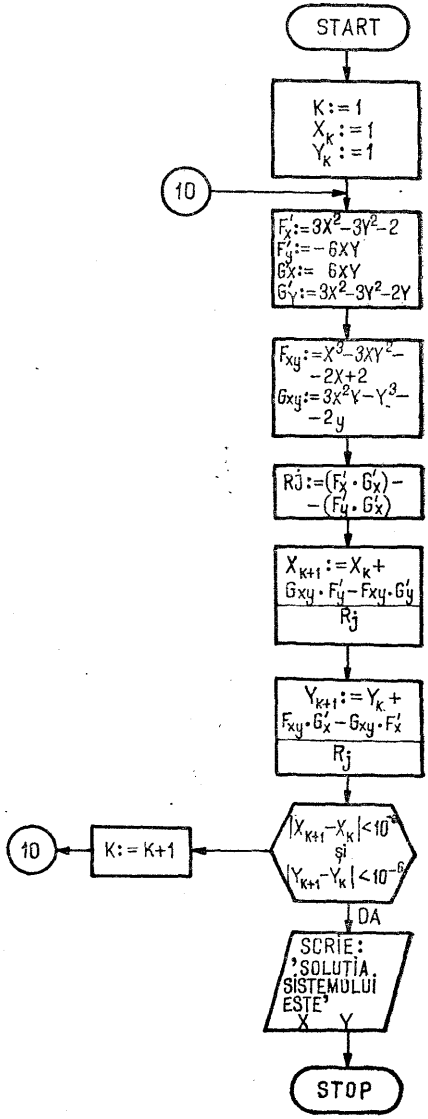

START
K:=1
X_K:=1
Y_K:=1
10
F'_x:=3X²-3Y²-2
F'_y:=-6XY
G'_x:=6XY
G'_y:=3X²-3Y²-2Y
Fxy:=X³-3XY²-2X+2
Gxy:=3X²Y-Y³-2y
Rj:=(F'_x·G'_x)-(F'_y·G'_x)
X_{K+1}:=X_K+(Gxy·F'_y-Fxy·G'_y)/Rj
Y_{K+1}:=Y_K+(Fxy·G'_x-Gxy·F'_x)/Rj
$|X_{K+1}-X_K|<10^{-6}$ și $|Y_{K+1}-Y_K|<10^{-6}$
K:=K+1
10
DA
SCRIE: 'SOLUȚIA SISTEMULUI ESTE' X Y
STOP

Fig. 2.14

```
      DIMENSION X(200),Y(200)
      K=1
      X(K)=1.
      Y(K)=1.
10    FDX=3*X(K)**2-3*Y(K)**2-2
      FDY=-6*X(K)*Y(K)
      GDX=6*X(K)*Y(K)
      GDY=3*X(K)**2-3*Y(K)**2-2
      FXY=X(K)**3-3*X(K)*Y(K)**2-2*X(K)+7
      GXY=3*X(K)**2*Y(K)-Y(K)**3-2*Y(K)
      RJ=(FDX*GDY)-(FDY*GDX)
      X(K+1)=X(K)+((GXY*FDY)-(FXY*GDY))/RJ
      Y(K+1)=Y(K)+((FXY*GDX)-(GXY*FDX))/RJ
      IF((ABS(X(K+1)-X(K)).LT.1.E-6).AND.(ABS(Y(K+1)-Y(K)).LT.1.E-6))
     * GO TO 11
      K=K+1
      GO TO 10
11    WRITE(108,100) X(K+1),Y(K+1),K
100   FORMAT(' ','SOLUTIA SISTEMULUI: X=',F8.3,' Y=',F8.3,/,
     *       'NUMARUL DE ITERATII=',I5)
      STOP
      END
```

```
SOLUTIA SISTEMULUI ESTE :X= .885 ; Y= .590
NUMARUL DE ITERATII = 5
```

Programul 2.5

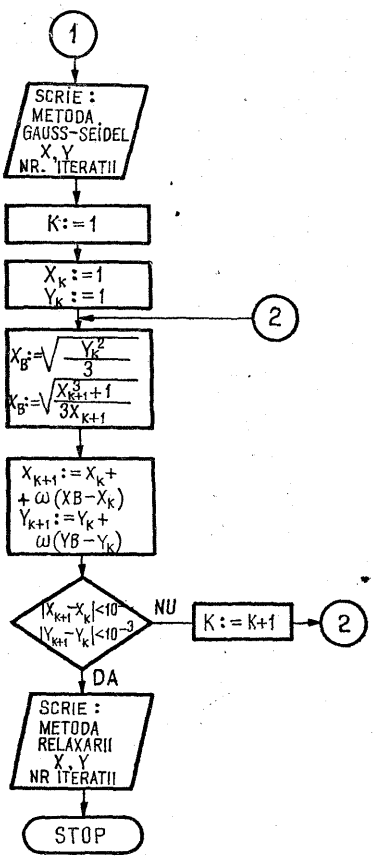


Fig. 2.15

```
C   JACOBI
      DIMENSION X(100),Y(100)
      DATA OMEGA/0.5/
      K=1
      X(K)=1.
      Y(K)=1.
 4    E1=(Y(K)**2)/3
      X(K+1)=SQRT(E1)
      E2=(X(K)**3+1)/(3*X(K))
      Y(K+1)=SQRT(E2)
      IF(ABS(X(K+1)-X(K)).LT.1.E-3) GO TO 6
      IF(ABS(Y(K+1)-Y(K)).LT.1.E-3) GO TO 6
      K=K+1
      GO TO 4
 6    WRITE(108,100) X(K+1),Y(K+1),K
 100  FORMAT(' ','METODA JACOBI       X=',F6.3,' Y=',F6.3,
     *       '     NUMARUL DE ITERATII =',I3)
C   GAUSS-SEIDEL
      K=1
      X(K)=1.
      Y(K)=1.
 10   E1=(Y(K)**2)/3
      X(K+1)=SQRT(E1)
      E2=(X(K+1)**3+1)/(3*X(K+1))
      Y(K+1)=SQRT(E2)
      IF(ABS(X(K+1)-X(K)).LT.1.E-3) GO TO 8
      IF(ABS(Y(K+1)-Y(K)).LT.1.E-3) GO TO 8
      K=K+1
      GO TO 10
 8    WRITE(108,101) X(K+1),Y(K+1),K
 101  FORMAT(' ','METODA GAUSS-SEIDEL X=',F6.3,' Y=',F6.3,
     *       '     NUMARUL DE ITERATII =',I3)
C  RELAXARII
      K=1
      Y(K)=1.
      X(K)=1.
 14   XB1=(Y(K)**2)/3
      XB=SQRT(XB1)
      YB1=(XB**3+1)/(3*XB)
      YB=SQRT(YB1)
      Y(K+1)=Y(K)+OMEGA*(YB-Y(K))
      X(K+1)=X(K)+OMEGA*(XB-X(K))
      IF(ABS(Y(K+1)-Y(K)).LT.1.E-3) GO TO 12
      IF(ABS(X(K+1)-X(K)).LT.1.E-3) GO TO 12
      K=K+1
      GO TO 14
 12   WRITE(108,102) X(K+1),Y(K+1),K
 102  FORMAT(' ','METODA RELAXARII    X=',F6.3,' Y=',F6.3,
     *       '     NUMARUL DE ITERATII =',I3)
      STOP
      END
          LINK
          RUN


 METODA JACOBI       X= .498 Y=  .868  NUMARUL DE INTERATII = 7
 METODA GAUSS-SEIDEL X=  .500 Y=  .866  NUMARUL DE INTERATII = 6
 METODA RELAXARII    X=  .511 Y=  .866  NUMARUL DE INTERATII = 6

          EOJ
```

Programul 2.6

2.4. Metode pentru determinarea rădăcinilor ecuaţiilor polinomiale

Destul de multe aplicaţii practice conduc la rezolvarea ecuaţiilor polinomiale de forma

$$P(x) \equiv a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \ldots + a_n = 0, \quad (2.115)$$

unde coeficienții a_i, $i = 0, 1, 2, \ldots, n$, sînt numere date, care pot fi reale sau complexe. Ecuația (2.115) are n rădăcini, nu neapărat distincte, unele din ele pot fi complexe (sau toate) chiar cînd coeficienții a_i sînt reali, dacă coeficienții sînt reali, rădăcinile complexe fiind sub formă de perechi conjugate, $a \pm ib$.

Majoritatea metodelor descrise în paragrafele precedente pentru rezolvarea ecuațiilor pot fi de asemenea utilizate la rezolvarea ecuațiilor polinomiale, aceste metode trebuind uneori modificate; în plus există [42, 1, 7] o serie de metode care se aplică în principal sau exclusiv la rezolvarea ecuațiilor polinomiale.

În acest paragraf se vor prezenta unele din metodele folosite la rezolvarea ecuațiilor polinomiale.

În rezolvarea ecuațiilor polinomiale se disting două etape:

— *separarea rădăcinilor*, care constă în determinarea unei vecinătăți V în care ecuația polinomială $P(x) = 0$ să aibă o soluție unică;

— *determinarea aproximativă a soluției* și evaluarea erorii considerînd că separarea deja s-a efectuat [128, 15].

Înainte de parcurgerea acestor două etape se vor prezenta o serie de proprietăți generale ale polinoamelor fără demonstrație [75, 67].

● Două polinoame $P(x)$ și $Q(x)$ sînt egale pentru orice x, dacă gradele și coeficienții lor sînt egali respectiv.

● Dacă $P(x)$ este un polinom de gradul n, atunci pentru orice α există un polinom unic $Q(x)$ astfel că

$$P(x) = (x - \alpha)Q(x) + P(\alpha) \tag{2.116}$$

dacă $n > 1$, gradul lui $Q(x)$ este $n - 1$, altfel $Q(x) = 0$.

● Dacă $P(x)$ este un polinom de gradul $n \geqslant 1$ și dacă $P(\alpha) = 0$, atunci există un polinom unic de gradul $n - 1$ astfel că

$$P(x) = (x - \alpha)Q(x). \tag{2.117}$$

de obicei în acest caz polinomul $Q(x)$ se numește polinom de reducere.

● Fie $P(x)$ un polinom de grad $n \geqslant 1$; α este o rădăcină a lui $P(x)$ de ordinul de multiplicitate m dacă și numai dacă

$$P(\alpha)=P'(\alpha)= \ldots = P^{(m-1)}(\alpha)=0 \text{ și } P^{(m)}(\alpha) \neq 0. \quad (2.118)$$

● Fie $P(x)$ un polinom care are s zerouri distincte $r_1, r_2, \ldots, r_s$ cu ordinul de multiplicitate $k_1, k_2, \ldots, k_s$ respectiv. Atunci există un polinom unic $V(x)$ astfel că

$$P(x) = (x - r_1)^{k_1}(x - r_2)^{k_2} \ldots (x - r_s)^{k_s} V(x). \quad (2.119)$$

2.4.1. *Localizarea rădăcinilor ecuațiilor polinomiale*

În cazul în care coeficienții polinomului (2.115) sînt reali, o informație asupra numărului rădăcinilor reale poate fi obținută cu regula lui Descartes privind semnul, iar o informație mult mai precisă cu ajutorul șirului lui Sturm, respectiv al șirului lui Rolle.

● *Metoda șirului lui Rolle.* Presupunem că $P : X \to Y$, $X \subset R$ și $Y \subset R$, satisface condițiile teoremei lui Rolle: între două rădăcini reale consecutive ale derivatei $P'(x)$ există cel mult o rădăcină a polinomului $P(x)$. Fie $x_1 < x_2 < \ldots < x_k$ rădăcinile ecuației polinomiale $P'(x) = 0$, așezate în ordine crescătoare. În acest caz se poate forma șirul lui Rolle:

$$P(-\infty), P(x_1), P(x_2), \ldots, P(x_k), P(+\infty). \quad (2.120)$$

Datorită consecinței enunțate, în fiecare interval

$$(-\infty, x_1), (x_1, x_2), \ldots, (x_i, x_{i+1}), \ldots, (x_k, +\infty) \quad (2.121)$$

se află cel mult o rădăcină reală a polinomului $P(x)$, numai dacă la capătul intervalului polinomul ia valori de semn contrar. Rezultă că ecuația polinomială $P(x) = 0$ are un număr de rădăcini reale egal cu numărul variațiilor de semn prezente în șirul lui Rolle.

În cazul în care $P : [a, b] \to R$, șirul lui Rolle are forma

$$P(a), P(x_1), P(x_2), \ldots, P(x_k), P(b). \quad (2.122)$$

Această metodă prezintă un avantaj deosebit din punct de vedere teoretic, dar din punct de vedere practic are un dezavantaj, deoarece pentru construirea şirului lui Rolle este necesară rezolvarea ecuaţiei $P'(x) = 0$, rezolvare care uneori este la fel de dificilă ca şi rezolvarea ecuaţiei $P(x) = 0$, apărînd situaţii cînd metodele exacte de rezolvare nu se pot utiliza [42].

● *Metoda şirului lui Sturm.* Fie $P : X \to Y$, $X \subset R$, $Y \subset R$. Presupunem că P este continuă şi derivabilă pe $[a, b]$, $a, b \in X$ cu $a < b$.

Definiţie. Se numeşte *şir al lui Sturm* asociat lui P un şir de polinoame $P_0, P_1, P_2, \ldots, P_m$, continue pe $[a, b]$, care satisfac condiţiile:

a) $P_0(x) = P(x)$;

b) $P_m(x) \neq 0$ pentru $x \in [a, b]$;

c) dacă $P_i(x) = 0$ pentru $1 \leqslant i \leqslant m - 1$ şi $x \in [a, b]$, atunci $P_{i-1}(x) \cdot P_{i+1}(x) < 0$;

d) dacă $P_0(x) = 0$, pentru $x \in (a, b)$, atunci $P_0'(x) \cdot P_1(x) > 0$.

Utilizîndu-se această definiţie, se poate enunţa:

Teorema 1. *Dacă aplicaţia P cu derivata continuă pe $[a, b]$ şi $P(a) = 0$, $P(b) = 0$, atunci P admite un şir al lui Sturm $P_0, P_1, P_2, \ldots, P_m$, iar numărul rădăcinilor reale ale ecuaţiei polinomiale $P(x) = 0$ în intervalul (a, b) este egal cu diferenţa dintre numărul variaţiilor de semn în şirul de valori numerice*

$$P_0(a), P_1(a), \ldots, P_m(a)$$

şi numărul variaţiilor de semn în şirul de valori numerice

$$P_0(b), P_1(b), \ldots, P_m(b).$$

Se poate observa cu uşurinţă că teorema permite atît numărarea rădăcinilor reale ale polinomului $P(x) = 0$ dintr-un interval, precum şi separarea acestor rădăcini.

Dacă se cunoaşte limita inferioară rădăcinilor l şi limita superioară L, se poate aplica teorema pentru numere cuprinse între l şi L.

În cazul în care $P(x)$ este un polinom cu coeficienți reali, fără rădăcini multiple, numărul rădăcinilor reale ale ecuației polinomiale $P(x) = 0$ rezultă din :

Teorema 2. *Dacă* $P = P_0$, $P_1 = P'$, $P_2, \ldots, P_m$ *sînt* $m + 1$ *polinoame în care* P_m *este o constantă diferită de zero,* P_{i+1} *este restul împărțirii lui* P_{i-1} *la* P_i, *cu semn schimbat, atunci numărul rădăcinilor reale ale ecuației* $P(x) = 0$ *este egal cu diferența dintre numărul variațiilor de semn în șirurile de numere*

$$\begin{matrix} P_0(-\infty), P_1(-\infty), P_2(-\infty), \ldots, P_m(-\infty) \\ \text{și} \\ P_0(+\infty), P_1(+\infty), P_2(+\infty), \ldots, P_m(+\infty). \end{matrix} \tag{2.123}$$

Demonstrație. Trebuie arătat că șirul de polinoame P_0, P_1, ... , P_m îndeplinește condiția de a fi șir al lui Sturm. Din ipoteza că $P(x)$ nu are rădăcini multiple rezultă că $P(x)$ și $P'(x)$ sînt prime între ele, adică c.m.m.d.c al lor este o constantă diferită de zero. Fie această constantă chiar P_m.

Pentru determinarea c.m.m.d.c. se folosește algoritmul lui Euclid. Algoritmul lui Euclid aplicat polinoamelor $P_0 = P$ și $P_1 = P'$ conduce la :

$$\left.\begin{aligned} P_0(x) &= P_1(x)\, Q_1(x) - P_2(x) \\ P_1(x) &= P_2(x)\, Q_2(x) - P_3(x) \\ P_2(x) &= P_3(x)\, Q_3(x) - P_4(x) \\ &\cdots\cdots\cdots\cdots\cdots \\ P_{i-1}(x) &= P_i(x)\, Q_i(x) - P_{i+1}(x) \\ &\cdots\cdots\cdots\cdots\cdots \\ P_{m-2}(x) &= P_{m-1}(x)\, Q_{m-1}(x) - P_m(x) \end{aligned}\right\} \tag{2.124}$$

Șirul (2.124) a fost construit ca în enunțul teoremei. În acest caz avem :

a) $P_0(x) = P(x)$ prin alegere.

b) $P_m(x) \neq 0$, prin ipoteză.

c) Dacă $P_i(x) = 0$, $1 \leqslant i \leqslant m-1$, atunci din relaţia $P_{i-1}(x) = P_i(x)\,Q_i(x) - P_{i+1}(x)$ rezultă $P_{i-1}(x) = -P_{i+1}(x)$, adică $P_{i-1}(x)P_{i+1}(x) = -P_{i+1}^2(x) < 0$.

d) A mai rămas de demonstrat că două polinoame consecutive din şir nu se anulează simultan. Pentru a evidenţia acest lucru se foloseşte relaţia

$$P_i(x) = P_{i+1}(x)\,Q_{i+1}(x) - P_{i+2}(x). \tag{2.125}$$

Dacă $P_i(x) = P_{i+1}(x) = 0$, rezultă $P_{i+2}(x) = 0$, iar de aici $P_{i+1}(x) = P_{i+2}(x) = 0$, ar rezulta $P_{i+3}(x) = 0$ ş.a.m.d., în final ajungîndu-se ca $P_{m-1}(x) = P_m(x) = 0$, fapt ce contrazice ipoteza că $P_m(x) \neq 0$.

Dacă $P_0(x) = 0$, atunci $P'(x) \cdot P_1(x) = [P'(x)]^2 > 0$, deoarece $P(x) = P_0(x)$ şi $P'(x) = P_1(x)$ nu au rădăcini comune, din presupunerea iniţială.

În cazul în care $P(x)$ are rădăcini multiple, adică $P(x)$ şi $P'(x)$ au rădăcini comune, c.m.m.d.c. al lor fiind $P_m(x)$, polinom de grad $n \geqslant 1$, atunci

$$f_0(x) = \frac{P_0(x)}{P_m(x)}, \quad f_1(x) = \frac{P_1(x)}{P_m(x)},$$

$$f_2(x) = \frac{P_2(x)}{P_m(x)}, \ldots, f_m(x) = \frac{P_m(x)}{P_m(x)} = 1.$$

Exemplu. Se consideră polinomul

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

Atunci

$$P_0(x) = P(x) = x^3 - x^2 + x + 1; \qquad P_1(x) = P'(x) = 3x^2 - 2x + 1.$$

Folosind relaţia iterativă

$$P_{k-2}(x) = P_{k-1}(x)\,Q_{k-1}(x) - P_k(x),$$

se obţine

$$P_2(x) = -\frac{4}{9}x - \frac{10}{9}$$

deoarece prin împărțirea lui $P_0(x)$ la $P_1(x)$ se obține relația

$$P_0(x) = \left(\frac{1}{3}x - \frac{1}{9}\right)P_1(x) - \left(-\frac{4}{9}x - \frac{10}{9}\right)$$

iar

$$P_1(x) = \left(-\frac{27}{4}x + \frac{171}{8}\right)P_2(x) - \left(-\frac{99}{4}\right),$$

de unde rezultă $P_3(x) = -\frac{99}{4}$.

În final rezultă următorul șir al lui Sturm pentru cazul considerat:

$$\left.\begin{aligned} P_0(x) &= x^3 - x^2 + x + 1 \\ P_1(x) &= 3x^2 - 2x + 1 \\ P_2(x) &= -\frac{4}{9}x - \frac{10}{9} \\ P_3(x) &= -\frac{99}{4} \end{aligned}\right\}.$$

Se calculează valorile numerice pentru intervalul $[-2, 2]$ în punctele $-2, -1, 0, 1, 2$, rezultatele trecîndu-se în tabel.

x	P_0	P_1	P_2	P_3	Număr variații semn	Numărul rădăcinilor reale
-2	-13	$+17$	$-\frac{2}{9}$	$-\frac{99}{4}$	2	
-1	-2	$+6$	$-\frac{2}{3}$	$-\frac{99}{4}$	2	
0	$+1$	$+1$	$-\frac{10}{9}$	$-\frac{99}{4}$	1	$x_1 \in (-1,0)$
1	$+3$	$+2$	$-\frac{14}{9}$	$-\frac{99}{4}$	1	
2	$+7$	$+9$	-2	$-\frac{99}{4}$	1	

Din analiza tabelului se vede că pentru intervalul ales a fost depistată o rădăcină reală în intervalul $(-1, 0)$. În continuare se poate aplica metoda bisecției pentru aproximarea rădăcinii polinomului, rădăcină care se găsește în intervalul $(-1, 0)$. De asemenea tabelul arată că în intervalele $(-2, -1)$ și $(0, 2)$ polinomul considerat nu are rădăcini reale.

La folosirea șirului lui Sturm, trebuie acordată o atenție deosebită erorilor de rotunjire introduse de calculator, erori care pot afecta semnul șirului de valori $P_i(x)$, $i = 0, 1, 2, \ldots, m$.

APROXIMARE ȘI INTERPOLARE

3.1. Introducere

În foarte multe aplicații practice apare necesitatea evaluării aproximative a unei funcții $f: [a, b] \to R$. În funcție de natura aplicației funcția $f(x)$ poate fi definită în diverse moduri:

a) Sub forma unui tabel în care se cunoaște valoarea funcției pentru anumite valori ale argumentului

$$\begin{array}{c|c} x & x_1 \quad x_2 \quad \ldots \quad x_n \\ \hline f(x) & f(x_1)\ f(x_2) \ldots f(x_n) \end{array}$$

Aceste valori tabelate pot fi caracterizate de un anume grad de precizie ca în cazul tabelelor logaritmice, sau valorile pot fi rezultatul unor observații sau măsurări experimentale, care în general sînt afectate de erori.

b) Sub forma uneia sau a mai multor formule explicite, de exemplu

$$f(x) = \cos x + 3,$$

$$f(x) = \begin{cases} 1 + x^2, & x \geqslant 0, \\ 5, & -2 \leqslant x < 0, \\ 2x + \cos x, & x < -1. \end{cases}$$

c) Sub forma unei serii, de exemplu

$$f(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

d) Sub forma unui algoritm numeric, de exemplu algoritm finit $f_0 = a_0$, $f_k = xf_{k-1} + a_k$, $k_k = 1, 2, \dots, n$, unde $a_0, a_1, \dots, a_n$ sînt date, sau algoritm infinit

$$f(x) = \lim_{n \to \infty} y_n,$$

unde

$$y_0 = (1 + x)/2, \quad x \geqslant 0,$$

$$y_{n+1} = \frac{1}{2}\left(y_n + \frac{x}{y_n}\right), \quad n = 0, 1, 2, \dots$$

e) Sub forma de soluţii ale unei ecuaţii diferenţiale.

Fie $f: [a, b] \to R$. Se pune problema determinării unei funcţii F, care să aproximeze funcţia f în intervalul $[a, b]$. Se recurge la această aproximare în două cazuri: 1) cînd nu se cunoaşte expresia analitică a lui f, dar se cunosc valorile sale într-un număr finit de puncte, cazurile a) şi d); 2) cînd expresia analitică a lui f este destul de complicată şi cu ajutorul acesteia calculele sînt destul de dificile, cazurile b), c), e).

Pentru evaluarea lui $f(x)$ se caută o altă funcţie $F(x)$ relativ simplă astfel ca pentru orice valoare a lui x valoarea lui $F(x)$ să fie suficient de aproape de valoarea lui $f(x)$. Dacă funcţia $F(x)$ se alege dintr-o anumită clasă de funcţii, de exemplu din clasa polinoamelor de grad n sau mai mic, pentru un anume n, atunci trebuie ca $F(x)$ să ia aceeaşi valoare cu $f(x)$ pentru anumite valori ale lui x. Aceste valori ale lui x sînt adesea referite ca *puncte de interpolare*. De asemenea se poate cere ca anumite derivate ale lui $F(x)$ să ia aceleaşi valori cu valorile derivatelor corespunzătoare ale lui $f(x)$ în anumite puncte de interpolare. Se poate arăta că dacă $F(x)$ este suma a $n + 1$ termeni ai seriei Taylor pentru $f(x)$ în punctul

$x = a$, atunci $F(x)$ poate fi considerat ca un polinom de interpolare pentru $f(x)$ de grad n sau mai mic, deoarece

$$F^{(k)}_{(a)} = f^k(a),\ k = 0, 1, \ldots, n. \tag{3.1}$$

S-a spus despre F că trebuie să fie o funcție simplă, adică ușor de evaluat, diferențiat, integrat. În multe situații această funcție aproximativă se prezintă sub forma unui polinom algebric, deoarece pe intervale de lungime mică curba $y = f(x)$ poate fi aproximată bine cu ajutorul acestuia.

Fie $M = \{f \mid f : [a, b] \to \boldsymbol{R}\}$ un spațiu liniar și să presupunem că printre funcțiile ce fac parte din M există k funcții $\varphi_0(x)$, $\varphi_1(x), \ldots, \varphi_k(x)$ liniar independente (funcțiile φ_i sînt liniar independente dacă din

$$c_0\varphi_0(x) + c_1\varphi_1(x) + \ldots + c_k\varphi_k(x) = 0$$

rezultă $c_0 = c_1 = \ldots = c_k$, oricare ar fi k). Aproximarea unei funcții oarecare f din M, care nu este simplă (se încadrează în unul din cazurile a, b, c, d, e) se face printr-o combinație liniară de un număr finit m, dat dinainte, de funcții liniar independente, adică prin

$$F_m(x) = c_0\varphi_0(x) + c_1\varphi_1(x) + \ldots + c_m\varphi_m(x) = \sum_{k=0}^{m} c_k\varphi_k(x). \tag{3.3}$$

Vom numi funcția (3.3) polinom generalizat, iar aproximarea funcției f în acest caz se face prin polinoame generalizate.

Foarte frecvent în procesul de aproximare a funcțiilor se iau ca funcții liniar independente funcțiile

$$1,\ x,\ x^2,\ x^3, \ldots,\ x^n, \ldots$$

Funcțiile din șir sînt simple, ușor calculabile. În acest caz $F_m(x)$ va fi un polinom algebric

$$F_m(x) \equiv P_m(x) = c_0 + c_1x + c_2x^2 + \ldots + c_mx^m. \tag{3.4}$$

Un alt șir de funcții liniar independente este șirul de funcții trigonometrice

$$1, \cos x, \sin x, \cos 2x, \sin 2x, \ldots, \cos kx, \sin kx, \ldots,$$

iar aproximarea se face prin polinoame trigonometrice de forma

$$T_m(x) = a_0 + a_1\cos x + b_1\sin x + a_2\cos 2x + b_2\sin 2x + \ldots$$
$$\ldots + a_m\cos mx + b_m\sin mx.$$

Din construcția funcției F_m se vede că nu este suficient să cunoaștem funcțiile liniar indepedente, ea depinde și de coeficienții $c_0, c_1, \ldots, c_m$ care trebuie determinați. Pentru determinarea acestor coeficienți, să presupunem că M se poate organiza ca un spațiu metric, adică putem defini pe M o funcție, care să măsoare distanța dintre două funcții din M. Dacă $M = C^0[a, b]$, mulțimea funcțiilor continue, definite pe $[a, b]$, putem defini distanța dintre două funcții f și g din M prin

$$d(f, g) = \max_{x \in [a,b]} |f(x) - g(x)|. \tag{3.5}$$

Dacă aproximarea lui f se face cu ajutorul acestei distanțe, se obține aproximarea uniformă, iar $d(f, g)$ are semnificația de eroare absolută în metrică a lui $g(x)$. Dacă în locul lui $g(x)$ în (3.5) se va pune $F_m(x)$ dată de (3.4), atunci

$$d(f, F_m) = \Phi(c_0, c_1, \ldots, c_m). \tag{3.6}$$

De aici, un criteriu pentru determinarea coeficienților $c_0, c_1, \ldots, c_m$ ar fi ca distanța dintre f și F să fie minimă.

În locul distanței (3.5) putem considera

$$d(f, g) = \left\{\int_a^b [f(x) - g(x)]^2 \, dx\right\}^{1/2} \tag{3.7}$$

sau

$$d(f, g) = \left\{\int_a^b p(x)[f(x) - g(x)]^2 \, dx\right\}^{1/2}, \tag{3.8}$$

unde $p(x) > 0$ şi poartă denumirea de funcţie pondere. Aproximarea datorită acestor distanţe se numeşte aproximare în medie pătratică, respectiv medie pătratică ponderată.

De remarcat că distanţele (3.7) şi (3.8) nu au semnificaţia de eroare absolută în metrică a lui $g(x)$, de aceea le vom numi *cvasimetrice*, la fel se întîmplă şi în cazul următoarei măsuri:

$$d(f, g) = \left\{ \sum_{i=0}^{n} [f(x_i) - g(x_i)]^2 \right\}^{1/2}, \qquad (3.9)$$

$x_0, x_1, \ldots, x_n$ fiind $n + 1$ puncte cunoscute din intervalul $[a, b]$.

Metrica (3.8) în care $g(x)$ se înlocuieşte cu $F_m(x)$ definită de (3.4), $m < n$, conduce tot la o aproximare în medie pătratică, metoda de aproximare purtînd denumirea de metoda celor mai mici pătrate a lui Gauss. Această metodă este utilizată în prelucrarea matematică a datelor experimentale. Putem de asemenea considera şi metrica

$$d(f, g) = \sum_{i=0}^{n} |f(x_i) - g(x_i)|. \qquad (3.10)$$

Considerînd distanţele definite de (3.9), respectiv (3.10), în care $g(x)$ se înlocuieşte cu $F_m(x)$, definită de (3.4), în cazul $m = n$, şi impunînd condiţia de minim, se ajunge la relaţiile

$$f(x_i) = F_n(x_i), \quad i = 0, 1, \ldots, n. \qquad (3.11)$$

Aproximarea în cazul acesta poartă denumirea de *aproximare prin interpolare*, iar polinomul generalizat $F_n(x)$, care satisface condiţia (3.11), *polinomul de interpolare*.

Funcţiile matematice sînt adesea descrise în formă tabelară, adică pentru un set de valori ale variabilei independente $x_1, x_2, \ldots, x_n$ sînt date valorile corespunzătoare ale funcţiei $f(x_1), f(x_2), \ldots, f(x_n)$. Exemple de astfel de funcţii sînt funcţiile trigonometrice, exponenţiale, loga-

ritmice. Procesul de găsire a unei curbe care să treacă prin punctele date în cazul determinării valorilor lui $f(x)$ pentru valorile lui x neexplicitate în tabel se numeşte interpolare.

3.2. Interpolarea grafică şi liniară

Una din metodele cele mai comune ale interpolării este metoda grafică. Această metodă constă în desenarea unui grafic continuu cu ajutorul valorilor din tabel, utilizînd un instrument corespunzător şi ţinînd seama de forma curbei pe intervalul considerat.

De exemplu, fie

Timp, s	0	60	120	180	240	300
v, km/s	0,0000	0,1824	0,4747	0,7502	1,3851	3,2229

Din fig. 3.1 se vede că graficul trece prin toate punctele date în tabel. Dacă timpul este $t = 150$ s, viteza poate fi interpolată ca 0,61 km/s.

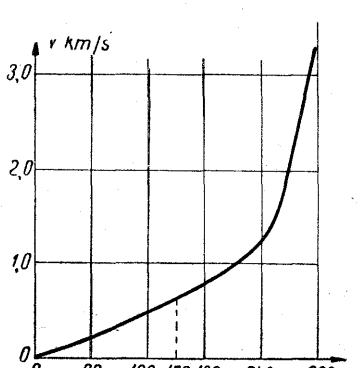


Fig. 3.1

● *Interpolarea liniară* a fost introdusă prima dată în calculul funcţiilor trigonometrice şi al logaritmilor.

Dacă din tabela funcției $\sin x$ se consideră următoarele valori:

x, în grade	0°	10°	20°	30°	40°
$f(x) = \sin x$	0,0000	0,17365	0,34202	0,50000	0,64279

atunci sin22° se calculează utilizînd interpolarea liniară astfel:

$$\frac{\sin 22° - 0{,}34202}{0{,}50000 - 0{,}34202} = \frac{22 - 20}{30 - 20},$$ rezultînd $\sin 22° = 0{,}37362$.

Valoarea exactă a lui sin 22° este 0,37461, de unde rezultă o eroare la a treia cifră cînd este utilizată interpolarea liniară.

Formula generală de interpolare liniară se poate obține geometric prin utilizarea asemănării triunghiurilor din fig. 3.2.

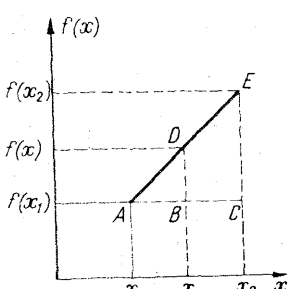


Fig. 3.2.

Triunghiul ABD este asemenea triunghiului ACE, de unde rezultă

$$\frac{DB}{EC} = \frac{BA}{CA} \quad \text{sau} \quad \frac{f(x) - f(x_1)}{f(x_2) - f(x_1)} = \frac{x - x_1}{x_2 - x_1};$$

după explicitarea lui $f(x)$ rezultă

$$f(x) = f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1). \qquad (3.12)$$

O altă metodă de obținere a formulei (3.12) este folosind ecuația dreptei sub forma

$$f(x) = a_1 x + a_2,$$

unde coeficienții a_1 și a_2 se pot determina prin condiția ca dreapta considerată să treacă prin punctele $(x_1, f(x_1))$ și $(x_2, f(x_2))$:

$$f(x_1) = a_1 x_1 + a_2, \quad f(x_2) = a_1 x_2 + a_2.$$

Rezolvînd acest sistem de ecuații în necunoscutele a_1 și a_2, rezultă

$$a_1 = \frac{f(x_2) - f(x_1)}{x_2 - x_1} \quad \text{și} \quad a_2 = f(x_1) - \frac{f(x_2) - f(x_1)}{x_2 - x_1} x_1.$$

Funcția de interpolare este atunci

$$f(x) = \frac{f(x_2) - f(x_1)}{x_2 - x_1} x + f(x_1) - \frac{f(x_2) - f(x_1)}{x_2 - x_1} x_1,$$

expresie ce se poate ordona sub forma (3.12). Relația (3.12) se poate pune sub forma

$$f(x) = f(x_1) \frac{x_2 - x}{x_2 - x_1} + f(x_2) \frac{x - x_1}{x_2 - x_1}$$

care după folosirea notațiilor

$$a_0(x) = \frac{x_2 - x}{x_2 - x_1} \quad \text{și} \quad a_1(x) = \frac{x - x_1}{x_2 - x_1}$$

devine

$$f(x) = a_0(x) f(x_1) + a_1(x) f(x_2).$$

Dacă x satisface condiția $a \leqslant x \leqslant b$, atunci $a_0(x)$ și $a_1(x)$ sînt funcții nenegative și

$$a_0(x) + a_1(x) = 1.$$

Evident, dacă $f(x)$ este o funcție liniară, atunci procesul de interpolare liniară este exact; în acest caz $f''(x)=0$.

În general este clar că diferența $f(x) - f^*(x)$ [unde $f^*(x)$ este funcția exactă] pe intervalul $[a, b]$ depinde de curbura funcției $f(x)$ și deci depinde de valoarea lui $f''(x)$.

3.2.1. *Convergența și precizia metodei de interpolare liniară*

Fie f funcția ce trebuie aproximată, $f \in C[a, b]$. Pentru un M întreg se construiește funcția $F(M; x)$ liniară pe intervale, care coincide cu $f(x)$ în $M + 1$ puncte de interpolare

$$x_k = a + \frac{k(b-a)}{M}, \quad k = 0, 1, 2, \ldots, M.$$

În fiecare subinterval $[x_k, x_{k+1}]$ se determină $F(M; x)$ prin interpolare liniară. Astfel pentru $x \in [x_k, x_{k+1}]$, (3.12) se poate scrie astfel:

$$F(M; x) = f(x_k) + \frac{x - x_k}{x_{k+1} - x_k} [f(x_{k+1}) - f(x_k)]. \quad (3.13)$$

Se urmărește a se arăta că

$$\lim_{M \to \infty} F(M; x) = f(x) \quad (3.14)$$

uniform pentru $x \in [a, b]$. Din (3.13) se poate scrie

$$F(M; x) = a_0(x) f(x_k) + a_1(x) f(x_{k+1}),$$

unde $a_0(x) = \frac{x_{k+1} - x}{x_{k+1} - x_k}$, $a_1 = \frac{x - x_k}{x_{k+1} - x_k}$ și $a_0(x) + a_1(x) = 1$. Din aceste relații rezultă

$$f(x) - F(M; x) = f(x) \; [a_0(x) + a_1(x)] - a_0(x) f(x_k) - a_1(x) f(x_{k+1}) = a_0(x)[f(x) - f(x_k)] + a_1(x)[f(x) - f(x_{k+1})].$$

Deoarece $f(x) \in C[a, b]$, ea este uniform continuă și rezultă că $|f(x) - f(x_k)|$ și $|f(x) - f(x_{k+1})|$ pot fi făcute arbitrar de mici prin alegerea lui M destul de mare. Prin urmare,

$$|f(x) - F(M; x)| \leqslant$$

$$\leqslant |a_0(x)|\,|f(x) - f(x_k)| + |a_1(x)|\,|f(x) - f(x_{k+1})| \leqslant$$

$$\leqslant \max\,(|f(x) - f(x_k)|, |f(x) - f(x_{k+1})|), \qquad (3.15)$$

deoarece $a_0(x) \geqslant 0$, $a_1(x) \geqslant 0$ și $a_0(x) + a_1(x) = 1$ continuă să se păstreze. De observat că pentru convergență nu este necesar ca punctele de interpolare să fie plasate exact la mijloc între două puncte consecutive date. Pentru o funcție continuă definită pe un interval închis se poate obține precizia dorită utilizînd interpolarea liniară, arătînd că se utilizează suficiente puncte de interpolare. Este suficient a considera precizia funcției liniare $F(x)$ din (3.12) care coincide cu $f(x)$ în două puncte a și b; astfel se poate demonstra [1,28]:

Teoremă. *Fie* $f(x) \in C^1[a, b]$ *și* $f \in D^2(a, b)$. *Dacă* $F(x)$ *este dată prin* (3.12), *atunci pentru orice* $c \in (a, b)$ *se poate scrie*

$$f(x) - F(x) = f(x) - f(a) - \frac{x - a}{b - a}[f(a) - f(b)].$$

Mai mult, dacă pentru $x \in (a, b)$ *avem* $|f''(x)| \leqslant M_2$, *atunci*

$$|f(x) - F(x)| \leqslant \frac{(b - a)^2}{8} M_2$$

pentru orice x *din* $[a, b]$.

Demonstrație. Se introduc funcțiile

$$R(x) = f(x) - F(x) \text{ și } P(x) = \frac{R(x)}{(x - a)(x - b)}.$$

Se observă că $P(x)$ nu este definită pentru $x = a$, $x = b$, cu toate acestea prin regula lui l'Hospital se obține

$$\lim_{x \to a+} P(x) = \frac{R'(a)}{a - b}, \quad \lim_{x \to b-} P(x) = \frac{R'(b)}{b - a}.$$

Dacă se extinde definiția lui $P(x)$ prin

$$P(a) = \frac{R'(a)}{a - b}. \quad P(b) = \frac{R'(b)}{b - a},$$

se obține o funcție care este continuă în $[a, b]$. Fie $x \in (a,b)$. Se consideră funcția

$$\Phi(z) = \Phi(z\,;\, x) = f(z) - F(z) - (z - a)(z - b)P(x).$$

Evident, pentru x fixat, $\Phi(z\,;\, x)$ este o funcție continuă de z în I și are prima derivată continuă pe I. Mai mult, $\Phi''(z\,;\, x)$ există pentru $z \in (a, b)$. Deci $\Phi(a\,;\, x) = \Phi(b\,;\, x) = \Phi(x\,;\, x) = 0$. Rezultă din teorema lui Rolle că există c_1 și c_2 cu $c_1 \in (a, x)$, $c_2 \in (x, b)$, astfel că

$$\Phi'(c_1\,;\, x) = \Phi'(c_2\,;\, x) = 0.$$

Aplicînd din nou teorema lui Rolle la $\Phi'(z\,;\, x)$, se găsește că există c în intervalul (c_1, c_2) și $\Phi''(c\,;\, x) = 0$ astfel că

$$\Phi''(c\,;\, x) = f''(c) - F''(c) - 2P(x) = f''(c) - 2P(x) = 0.$$

Deci $F'''(c) = 0$. Prin urmare, avem $P(x) = \frac{f''(c)}{2}$ și

$$f(x) - F(x) = R(x) = (x - a)(x - b)P(x).$$

Se observă că în I funcția $(x - a)(x - b)$ are o valoare maximă absolută pentru $x = (a + b)/2$:

$$\max_{a \leqslant x \leqslant b} |(x - a)(x - b)| = |(x - a)(x - b)|_{x=(a+b)/2} =$$

$$= \frac{(b - a)^2}{4},$$

de unde rezultă

$$f(x) - F(x) = \frac{(x-a)(x-b)}{2} f''(c)$$

și dacă $|f''(x)| \leqslant M_2$, atunci

$$|f(x) - F(x)| \leqslant \frac{(b-a)^2}{8} M_2.$$

Astfel teorema este demonstrată.

3.3. Interpolare polinomială

O altă metodă mai precisă de interpolare este interpolarea polinomială a funcțiilor.

Dacă se dau $n+1$ puncte în care valorile funcției sînt cunoscute, se pune problema determinării unui polinom de gradul n care să treacă prin cele $n+1$ puncte, acest polinom numindu-se polinom de interpolare. În general polinomul $P_n(x)$ de grad cel mult n poate aproxima funcția $f(x)$ în intervalul $a \leqslant x \leqslant b$, dacă o anumită măsură (distanță) a derivatelor polinomului față de funcție pe acest interval sînt destul de mici.

3.3.1. *Interpolare Lagrange*

Se dau valorile lui $f(x)$ în $n+1$ puncte $x_0, x_1, \ldots, x_n$ și se dorește determinarea unui polinom $F(x)$ de gradul n sau mai mic astfel că

$$F(x_i) = f(x_i), \qquad i = 0, 1, \ldots, n. \tag{3.16}$$

Se va arăta că polinomul $F(x)$ există și este unic.

Teoremă. *Dacă $x_0, \ldots, x_n$ sînt distincte, atunci pentru orice $y_0, y_1, \ldots, y_n$ există un polinom $F(x)$ de grad $\leqslant n$, astfel că*

$$F(x_i) = y_i, \qquad i = 0, 1, \ldots, n. \tag{3.17}$$

Demostrație. Se definește $F(x)$ în felul următor :

$$F(x) = F_0(x) + F_1(x) + \ldots + F_n(x), \qquad (3.18)$$

unde

$$F_i(x) = \left(\prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \right) y_i, \qquad i = 0, 1, \ldots, n. \qquad (3.19)$$

Evident $F(x)$ este polinom de grad $\leqslant n$ care satisface (3.17). Pentru a demonstra unicitatea se consideră orice alt polinom $G(x)$ de grad $\leqslant n$ care satisface (3.17). Atunci $H(x) = F(x) - G(x)$ este un polinom de grad $\leqslant n$ care trece prin $n + 1$ puncte ale lui $f(x)$, deci rezultă $H(x) = 0$ și $G(x) = F(x)$.

Exemplu. Fie

$$x_0 = 0, \qquad x_1 = 1, \qquad x_2 = 2,$$

$$y_0 = 1, \qquad y_1 = 2, \qquad y_2 = 4.$$

Atunci

$$F(x) = F_0(x) + F_1(x) + F_2(x)$$

și din (3.19) rezultă

$$F(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_1)} y_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} y_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} y_2.$$

Se observă că $F(x)$ este de grad $\leqslant 2$ și $F(x_0) = y_0$, $F(x_1) = y_1$, $F(x_2) = y_2$. Substituind valorile x_i și y_i, se obține

$$F(x) = \frac{(x - 1)(x - 2)}{(0 - 1)(0 - 2)} 1 + \frac{(x - 0)(x - 2)}{(1 - 0)(1 - 2)} 2 + \frac{(x - 0)(x - 1)}{(2 - 0)(2 - 1)} 4 =$$

$$= \frac{1}{2}(x - 1)(x - 2) - 2x(x - 2) + 2x(x - 1) = \frac{1}{2} x^2 + \frac{1}{2} x + 1.$$

Se poate verifica ușor că $F(0) = 1$, $F(1) = 2$, $F(2) = 4$.

● *Metoda coeficienților nedeterminați.* O altă metodă pentru determinarea polinomului de interpolare se bazează pe utilizarea variantei coeficienților nedeterminați.

Se alege orice valoare convenabilă pentru α și fie

$$F(x) = a_0(x - \alpha)^n + a_1(x - \alpha)^{n-1} + \ldots + a_{n-1}(x - \alpha) + a_n$$

Se urmărește determinarea coeficienților $a_0, a_1, \ldots, a_n$ astfel ca (3.17) să fie satisfăcută. Rezultă sistemul de ecuații algebrice liniare

$$\left.\begin{aligned} a_0(x_0 - \alpha)^n + a_1(x_0 - \alpha)^{n-1} + \ldots + a_n &= y_0 \\ a_0(x_1 - \alpha)^n + a_1(x_1 - \alpha)^{n-1} + \ldots + a_n &= y_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_0(x_n - \alpha)^n + a_1(x_n - \alpha)^{n-1} + \ldots + a_n &= y_n \end{aligned}\right\} \quad (3.20)$$

Sistemul (3.20) de $n + 1$ ecuații cu $n + 1$ necunoscute are soluție unică, dacă determinantul sistemului

$$D = \begin{vmatrix} (x_0 - \alpha)^n & (x_0 - \alpha)^{n+1} & \ldots & (x - \alpha) & 1 \\ (x_1 - \alpha)^n & (x_1 - \alpha)^{n-1} & \ldots & (x - \alpha) & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ (x_n - \alpha)^n & (x_n - \alpha)^{n-1} & \ldots & (x_n - \alpha) & 1 \end{vmatrix} \neq 0. \quad (3.21)$$

Determinantul D este un determinant Vandermonde a cărui valoare este dată de relația

$$D = \prod_{\substack{i,j=0 \\ i<j}}^{n} (x_i - x_j). \quad (3.22)$$

Se va verifica acest lucru pentru cazul $\alpha = 0$, $n = 3$. Atunci

$$D = \begin{vmatrix} x_0^3 & x_0^2 & x_0 & 1 \\ x_1^3 & x_1^2 & x_1 & 1 \\ x_2^3 & x_2^2 & x_2 & 1 \\ x_3^3 & x_3^2 & x_3 & 1 \end{vmatrix} =$$

$$= (x_0 - x_1)(x_0 - x_2)(x_0 - x_3)(x_1 - x_2)(x_1 - x_3)(x_2 - x_3).$$

Deoarece x_i sînt distincţi, rezultă $D \neq 0$ şi a_k pot fi unic determinaţi.

Să aplicăm această metodă la exemplul de mai sus pentru $\alpha = 0$. Atunci $F(x) = a_0x^2 + a_1x + a_2$ şi $a_2 = 1$, $a_0 + a_1 + a_2 = 2$, $4a_0 + 2a_1 + a_2 = 4$, care după rezolvare dau $a_0 = 1/2$, $a_1 = 1/2$, $a_2 = 1$, obţinîndu-se

$$F(x) = \frac{1}{2}x^2 + \frac{1}{2}x + 1.$$

● *Metoda valorilor nedeterminate.* Formula (3.18) poate fi scrisă sub forma

$$F(x) = \sum_{i=0}^{n} a_i(x)\, y_i = \sum_{i=0}^{n} a_i(x)\, F(x_i), \tag{3.23}$$

unde polinoamele $a_i(x)$ au gradul $\leqslant n$. Din (3.19) avem

$$a_i(x) = \prod_{j=0}^{n} \frac{x - x_j}{x_i - x_j}, \quad i = 0, \quad 1, \ldots, n. \tag{3.24}$$

Pentru o valoare a lui x dată, funcţiile $a_0(x), a_1(x), \ldots, a_n(x)$ pot fi considerate drept *coeficienţi*. Astfel $F(x)$ este o combinaţie liniară de $F(x_i)$ cu valorile $a_0(x), a_1(x), \ldots, a_n(x)$. Evident valorile pot fi determinate din (3.24) şi $F(x)$ poate fi evaluată fără determinare a coeficienţilor a_0, $a_1, \ldots, a_n$ ai polinomului $F(x)$. În cazul interpolării lui Lagrange simplă, valorile sînt date explicit prin (3.24). Totuşi există situaţii cînd astfel de expresii ca în (3.24) nu sînt date explicit [1]. Cu toate acestea din (3.24) pentru un x dat se pot determina valorile prin rezolvarea unui sistem liniar de ecuaţii algebrice similar cu cel întîlnit în metoda coeficienţilor nedeterminaţi.

În cazul metodei valorilor nedeterminate se presupune că $F(x)$ are forma

$$F(x) = \sum_{i=0}^{n} a_i(x) F(x_i) \tag{3.25}$$

şi se aleg $a_i(x)$ astfel ca ecuaţia să fie satisfăcută pentru cazul particular $F(x) = 1$, $F(x) = x - \alpha$, $F(x) = (x-\alpha)^2, \ldots$ $\ldots$, $F(x) = (x - \alpha)^n$. Dacă (3.25) este exactă pentru astfel de funcţii, atunci pentru orice coeficienţi $a_0, a_1, \ldots, a_n$ avem

$$c_0(x - \alpha)^n + c_1(x - \alpha)^{n-1} + \ldots + c_n =$$

$$= \sum_{i=0}^{n} a_i(x)\{c_0(x - \alpha)^n + c_1(x-\alpha)^{n-1} + \ldots + c_{n-1}(x-\alpha) + c_n\}, \tag{3.26}$$

deoarece (3.25) va fi exactă pentru orice polinom de grad $\leqslant n$ în $x - \alpha$.

Datorită faptului că orice polinom de grad $\leqslant n$ în $x - \alpha$ este un polinom de grad $\leqslant n$ în $x = (x - \alpha) - \alpha$, rezultă că (3.25) va fi exactă pentru orice polinom de grad $\leqslant n$ în x.

Cerinţa ca (3.25) să aibă sens pentru cazurile particulare $F(x) = 1$, $F(x) = x - \alpha$ implică

$$\begin{aligned}
1 &= a_0(x) + a_1(x) + \ldots + a_n(x),\\
x-\alpha &= a_0(x)(x_0 - \alpha) + a_1(x)(x_1 - \alpha) + \ldots + a_n(x)(x_n - \alpha),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
(x-\alpha)^n &= a_0(x)(x_0-\alpha)^n + a_1(x)(x_1-\alpha)^n + \ldots + a_n(x)(x_n-\alpha)^n.
\end{aligned} \tag{3.27}$$

Pentru un x dat se pot determina valorile $a_i(x)$ dacă determinantul D al sistemului este $\neq 0$:

$$a_k(x) = \frac{D_k}{D}, \quad k = 0, 1, 2, \ldots, n,$$

unde

$$\left.\begin{aligned}
D &= (-1)^k \prod_{\substack{i,j=0\\ i\neq k,\ j\neq k}}^{n} (x_i - x_j) \prod_{\substack{j=0\\ j\neq k}}^{n} (x_k - x_j)\\
D_k &= (-1)^k \prod_{\substack{i,j=0\\ i\neq j\\ i\neq k,\ j\neq k}}^{n} (x_i - x_j) \prod_{\substack{j=0\\ j\neq k}}^{n} (x - x_j)
\end{aligned}\right\} \tag{3.28}$$

Să se aplice această procedură cînd $n = 2$ și $F(0) = 1$, $F(1) = 2$, $F(2) = 4$, unde $x = 1/2$.

Fie $a_i(1/2) = a_i$, $i = 0,1,2$. Atunci avem pentru $\alpha = 0$

$$\left.\begin{aligned} 1 &= a_0 + a_1 + a_2 \\ \frac{1}{2} &= 0\cdot a_0 + 1\cdot a_1 + 2a_2 \\ \left(\frac{1}{2}\right)^2 &= 0\cdot a_0 + 1\cdot a_1 + 4\cdot a_2 \end{aligned}\right\}$$

După rezolvare rezultă $a_0 = \frac{3}{8}$, $a_1 = \frac{3}{4}$, $a_2 = -\frac{1}{8}$,

$$F\left(\frac{1}{2}\right) = \frac{3}{8}F(0) + \frac{3}{4}F(1) - \frac{1}{8}F(2) = \frac{3}{8}\cdot 1 + \frac{3}{4}\cdot 2 - \frac{1}{8}\cdot 4 = \frac{11}{8},$$

$F(x) = \frac{1}{2}x^2 + \frac{1}{2}x + 1$, pentru $x = \frac{1}{2}$ rezultă

$$F\left(\frac{1}{2}\right) = \frac{1}{2}\cdot\frac{1}{4} + \frac{1}{2}\cdot\frac{1}{2} + 1 = \frac{1}{8} + \frac{1}{4} + 1 = \frac{11}{8}.$$

Deci se obține același rezultat [128].

● *Metoda Aitken.* Polinomul de interpolare al lui Lagrange poate fi evaluat utilizînd o serie de interpolări liniare prin metoda lui Aitken. Se va ilustra metoda considerată în cinci puncte de interpolare, folosindu-se tabelul

$$\begin{array}{lllll} f(x_0) & & & & \\ & I_{0,1} & & & \\ f(x_1) & & I_{0,1,2} & & \\ & I_{0,2} & & I_{0,1,2,3} & \\ f(x_2) & & I_{0,1,3} & & I_{0,1,2,3,4} \\ & I_{0,3} & & I_{0,1,2,4} & \\ f(x_3) & & I_{0,1,4} & & \\ & I_{0,4} & & & \\ f(x_4) & & & & \end{array} \qquad (3.29)$$

Aici avem

$$I_{0,j} = I_{0,j}(x) = \frac{x_j - x}{x_j - x_0} f(x_0) + \frac{x - x_0}{x_i - x_0} f(x_j), \quad j = 1, 2, 3, 4,$$

$$I_{0,1,j} = I_{0,1,j}(x) = \frac{x_j - x}{x_j - x_1} I_{0,1} + \frac{x - x_1}{x_j - x_1} I_{0,j}, \quad j = 2, 3, 4,$$

$$I_{0,1,2,j} = I_{0,1,2,j}(x) = \frac{x_j - x}{x_j - x_2} I_{0,1,2} + \frac{x - x_2}{x_j - x_2} I_{0,1,j}, \quad j = 3, 4,$$

$$I_{0,1,2,3,4} = I_{0,1,2,3,4}(x) = \frac{x_4 - x}{x_4 - x_3} I_{0,1,2,3} + \frac{x - x_3}{x_4 - x_3} I_{0;1;2;4}. \tag{3.30}$$

Este ușor de aratat că

$$I_{0,j}(x)_k = f(x_k), \quad k = 0, j.$$

În continuare se arătă că

$$I_{0,1,j}(x) = f(x_k), \quad k = 0, 1, j.$$

Dar

$$I_{0,1,j}(x_0) = \frac{x_j - x_0}{x_j - x_1} I_{0,1}(x) + \frac{x_0 - x_1}{x_j - x_1} I_{0,1}(x_0) = f(x_0),$$

$$I_{0,j}(x_1) = I_{0,1}(x) = f(x_1),$$

$$I_{0;1;j}(x_1) = I_{0;1}(x_1) = f(x_1),$$

$$I_{0;1,j}(x_j) = I_{0;j}(x_j) = f(x_j), \quad j = 2, 3, 4.$$

În același fel se poate arăta că

$$I_{0;1;2,j}(x_k) = f(x_k), \quad k = 0, 1, 2, j, \tag{3.31}$$
$$I_{0,1,2,3,4}(x_k) = f(x_k), \quad k = 0, 1, 2, 3, 4.$$

Deci $I_{0,1,2,3,4}(x)$ este evident un polinom de gradul patru sau mai mic; rezultă că acesta este la fel ca polinomul de interpolare al lui Lagrange.

Exemplu. Pentru $f(0) = 1$, $f(1) = 2$, $f(2) = 4$ avem

$$I_{0,1} = \frac{1-x}{1-0}(1) + \frac{x-0}{1-0}(2) = 1 + x,$$

$$I_{0,2} = \frac{2-x}{2-0}(1) + \frac{x-0}{2-0}(4) = 1 + \frac{3}{2}x,$$

$$I_{0,1,2} = \frac{2-x}{2-1}(1+x) + \frac{x-1}{2-1}\left(1 + \frac{3}{2}x\right) = 1 + \frac{1}{2}x + \frac{1}{2}x^2,$$

care este același polinom obținut și prin interpolarea Lagrange [128].

3.3.2. *Convergența și precizia în cazul interpolării Lagrange*

Presupunînd că se dă o funcție $f(x)$ care este continuă pe intervalul $I = [a, b]$, se va genera o secvență de funcții bazate pe utilizarea interpolării Lagrange și se va studia convergența acestei secvențe.

Se divide intervalul I în M subintervale de lungime h, considerîndu-se punctele de interpolare

$$x_k = a + kh,\ k = 0, 1, 2, \ldots, M, \tag{3.32}$$

unde $h = \frac{b-a}{M}$. Se selectează un întreg n și pentru fiecare $M \geqslant n$ se utilizează interpolarea lui Lagrange în $n + 1$ puncte în felul următor: în orice subinterval

$$I_k = [x_{k-1}, x_k],\quad k = 1, 2, \ldots, M, \tag{3.33}$$

se utilizează interpolarea lui Lagrange în $n + 1$ puncte bazate pe x_{k+1}, x_k și $n - 1$ puncte adiționale cît mai apropiate de x.

Astfel, dacă $n = 3$, se va utiliza x_{k-2}, x_{k-1}, x_k, x_{k+1} dacă $x \in I_k$; caz excepție, dacă $x \in I_1$, se va utiliza x_0, x_1, x_2, x_3. De asemenea, dacă $x \in I_M$, se va utiliza x_{M-3}, x_{M-2}, x_{M-1}, x_M. Dîndu-se x, se etichetează cele $n+1$ puncte de interpolare care sînt utilizate cu $t_0, t_1, \ldots, t_n$, unde $t_0 < t_1 < \ldots < t_n$. Evident se poate construi o secvență compusă din funcțiile

$$F_M(x) = \sum_{j=0}^{n} a_j(x) f(t_j) \tag{3.34}$$

prin formula de interpolare Lagrange, unde

$$a_j(x) = \prod_{\substack{s=0 \\ s \neq j}}^{n} \frac{x - t_s}{t_j - t_s}. \tag{3.35}$$

Deoarece

$$\sum_{j=0}^{n} a_j(x) = 1, \tag{3.36}$$

avem

$$f(x) - F_M(x) = f(x) \sum_{j=0}^{n} a_j(x) - \sum_{j=0}^{n} a_j(x) f(t_j) =$$

$$= \sum_{j=0}^{n} a_j(x)[f(x) - f(t_j)] \tag{3.37}$$

și

$$|f(x) - F_M(x)| \leqslant K(n) \max_{0 \leqslant j \leqslant n} |f(x) - f(t_j)|,$$

unde $K(n)$ sînt constante dependente de n care sînt mărginite pentru $\sum_{j=0}^{n} |a_j(x)|$. Presupunem existența lui $K(n)$. Din continuitatea uniformă a lui $f(x)$ rezultă

$$\max_{0 \leqslant j \leqslant n} |f(x) - f(t_j)| \to 0 \tag{3.38}$$

cînd $M \to \infty$, uniform pe $[a, b]$, și deci

$$\lim_{M\to\infty} |f(x) - F_M(x)| = 0 \tag{3.39}$$

uniform pe $[a, b]$.

Se va arăta că constantele $K(n)$ există. Fiecare factor al numitorului din (3.35) este cel mult h, de aceea numitorul este cel mult h^n. Fiecare factor al numitorului este cel mult nh. Deoarece numărătorul este cel mult nh, avem $|a_j(x)| \leqslant n^n$ pentru fiecare j. Prin urmare

$$K(n) = \sum_{j=0}^{n} |a_j(x)| \leqslant (n+1)\, n^n. \tag{3.40}$$

În continuare se va considera precizia polinomului de interpolare.

Teoremă. *Fie* $f(x) \in C^n[x_0, x_n]$ *și* $f(x) \in D^{(n+1)}(x_0, x_n)$. *Fie* $x_0, x_1, \ldots, x_n$ *numere distincte astfel că* $x_0 < x_1 < \ldots < x_n$ *și* $F(x)$ *polinom unic de grad* $\leqslant n$, *astfel că*

$$F(x_i) = f(x_i), \quad i = 0, 1, 2, \ldots, n. \tag{3.41}$$

Dacă $x \in [x_0, x_n]$, *avem*

$$f(x) - F(x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_n)}{(n+1)!} f^{(n+1)}(c) \tag{3.42}$$

pentru orice $c \in (x_0, x_n)$.

Pentru fiecare $k = 0, 1, 2, \ldots$ se definește

$$M_k = \max_{x\in I} |f^k(x)|, \tag{3.43}$$

unde $I = [x_0, x_n]$. Atunci avem :

Corolar. *Ținînd seama de ipotezele teoremei, dacă* $f^{(n+1)}(x)$ *este continuă pe* I *și dacă* $x_1 - x_0 = x_2 - x_1 = \ldots = x_n - x_{n-1} = h$, *atunci*

$$|f(x) - F(x)| \leqslant \frac{h^{n+1}}{4(n+1)} M_{n+1}. \tag{3.44}$$

Demonstrație. Este ușor de arătat că cea mai mare valoare a expresiei $|(x-x_0)(x-x_1)\ldots(x-x_n)|$ este presupusă în unul din intervalele $x_0<x<x_1$ și $x_{n-1}<x<x_n$. Mai mult, valoarea maximă a acestei expresii satisface inegalitatea

$$\max_{x_0\leqslant x\leqslant x_1}|(x-x_0)(x-x_1)|\max_{x_0<x\leqslant x_1}|(x-x_2)(x-x_3)\ldots(x-x_n)\leqslant$$

$$\leqslant\frac{n!}{4}h^{n+1}. \tag{3.45}$$

Corolar. *Cu ipotezele corolarului de mai sus avem*

$$|f(x)-F(x)|\leqslant\frac{\sqrt{3}}{27}h^3M_3$$

dacă $n=2$ *și*

$$|f(x)-F(x)|\leqslant\frac{h^4}{24}M_4 \tag{3.46}$$

dacă $n=3$.

Demonstrație. Pentru cazul $n=2$, dacă $y=x-x_1$, avem

$$Q_2(x)=(x-x_0)(x-x_1)(x-x_2)=y(y^2-h^2).$$

Evident $Q_2'(x)$ tinde către zero pentru $y=\pm\frac{h}{\sqrt{3}}$, adică pentru $x=x_1\pm\frac{h}{\sqrt{3}}$ se obține

$$|Q_2|_{max}=\left|Q_2\left(x_1\pm\frac{h}{\sqrt{3}}\right)\right|=\frac{2\sqrt{3}}{9}h^3.$$

Pentru cazul $n=3$, se ia $y=x-(x_1+x_2)/2$ și avem

$$Q_3(x)=(x-x_0)(x-x_1)(x-x_2)(x-x_3)=$$

$$=y^4-\frac{5}{2}h^2y^2+\frac{9}{16}h^4.$$

Evident $Q_3'(x)$ tinde la zero pentru $y \to \pm \frac{1}{2}\sqrt{5h}$, adică pentru $x = \frac{x_1 + x_2}{2}$ și $\frac{x_1 + x_2}{2} \pm \frac{1}{2}\sqrt{5h}$ avem

$$\left|Q_3\left(\frac{x_1 + x_2}{2}\right)\right| = \frac{9}{16}h^4, \quad \left|Q_3\left(\frac{x_1 + x_2}{2} \pm \frac{1}{2}\sqrt{5h}\right)\right| = h^4.$$

Evident că valoarea cea mai mare a expresiei $|(x-x_0)(x-x_1)(x-x_2)(x-x_3)|$ apare în intervalele $x_0 \leqslant x \leqslant x_1$ și $x_2 \leqslant x \leqslant x_3$. Aceasta sugerează faptul că, ori de cîte ori este posibil, se va alege pentru x două puncte de interpolare mai mari ca x și două mai mici ca x.

Exemplu. Dacă $f(x)$ este tabelată pentru $x = 0(0,1)1$ și se cere $f(0,37)$ prin intermediul unui polinom al lui Lagrange cu patru puncte de interpolare, atunci se vor utiliza ca puncte de interpolare $\{0,2; 0,3; 0,4; 0,5\}$, în loc de punctele $\{0,3; 0,4; 0,5; 0,6\}$.

3.4. Interpolarea în intervale egale

Dacă punctele $x_0, x_1, \ldots, x_n$ satisfac relația

$$h = x_1 - x_0 = x_2 - x_1 = \ldots = x_n - x_{n-1}, \tag{3.47}$$

polinomul de interpolare Lagrange se poate scrie sub forma

$$F(x) = \sum_{k=0}^{n} a_k(x) f(x_k), \tag{3.48}$$

unde

$$a_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{u - j}{k - j}, \quad k = 0, 1, \ldots, n, \quad u = \frac{x - x_0}{h}. \tag{3.49}$$

În cadrul acestui paragraf se vor introduce o serie de diferențe finite. Aceste diferențe finite sînt:

$\Delta f(x_i)$ diferența la dreapta,
$\nabla f(x_i)$ diferența la stînga,
$\delta f(x_i)$ diferență centrată.

Aceste trei diferențe au interpretarea dată în fig. 3.3.

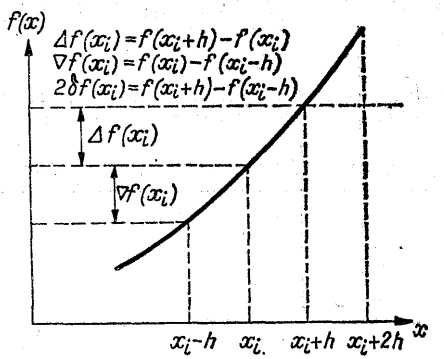


Fig. 3.3

Dacă punctele de interpolare sînt echidistante, atunci calculele în procesul de interpolare sînt mult simplificate dacă se folosesc formule ce implică diferențe finite.

Dacă $\Delta f(x) = f(x + h) - f(x)$, atunci

$$\Delta^2 f(x) = \Delta(\Delta f(x)) = \Delta(f(x + h) - f(x)) =$$
$$= f(x + 2h) - 2f(x + h) + f(x),$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\Delta^n f(x) = \Delta(\Delta^{n-1} f(x)). \qquad (3.50)$$

Utilizînd aceste relații, se poate construi următorul tabel cu diferențe la dreapta:

x	$f(x)$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$	$\Delta^4 f(x)$	$\Delta^5 f(x)$
x_0	$f(x_0)$					
		$\Delta f(x_0)$				
x_1	$f(x_1)$		$\Delta^2 f(x_0)$			
		$\Delta f(x_1)$		$\Delta^3 f(x_0)$		
x_2	$f(x_2)$		$\Delta^2 f(x_1)$		$\Delta^4 f(x_0)$	
		$\Delta f(x_2)$		$\Delta^3 f(x_1)$		$\Delta^5 f(x)$
x_3	$f(x_3)$		$\Delta^2 f(x_2)$		$\Delta^4 f(x_1)$	
		$\Delta f(x_3)$		$\Delta^3 f(x_2)$		
x_4	$f(x_4)$		$\Delta^2 f(x_3)$			
		$\Delta f(x_4)$				
x_5	$f(x_5)$					

(3.51)

În fiecare caz, elementul $\Delta^k f(x_p)$ este obținut prin scăderea lui $\Delta^{k-1} f(x_p)$ din $\Delta^{k-1} f(x_{p+1})$, unde $\Delta^k f(x_p) = f(x_p)$ pentru $k = 0$.

Exemplu. Să se construiască tabelul cu diferențe la dreapta pentru funcția $f(x) = x^3$, x luînd valorile întregi de la unul la cinci cu pasul 1.

x	$f(x)$	$\Delta f(x)$	$\Delta^2 f(x)$	$\Delta^3 f(x)$	$\Delta^4 f(x)$	$\Delta^5 f(x)$
0	0					
		1				
1	1		6			
		7		6		
2	8		12		0	
		19		6		0
3	27		18		0	
		37		6		
4	64		24			
		61				
5	125					

Teoremă. *Dacă $f(x)$ este un polinom de grad $\leqslant n$, atunci $\Delta^{n+1} f(x) \equiv 0$.*

Demonstrație. Operatorul diferență Δ este liniar în sensul că

$$\begin{aligned} \Delta(f(x) + g(x)) &= \Delta f(x) + \Delta g(x), \\ \Delta(cf(x)) &= c\,\Delta f(x). \end{aligned} \tag{3.52}$$

Prin urmare, dacă $f(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n$, atunci

$$\Delta f(x) = a_0 \Delta(x^n) + a_1 \Delta(x^{n-1}) + \ldots + a_n \Delta(a_n),$$

$$\Delta(x^k) = (x + h)^k - x^k = khx^{k-1} + \ldots + h^k,$$

care este un polinom de gradul $k - 1$. Prin urmare, $\Delta f(x)$ este un polinom de gradul $n - 1$, iar $\Delta^2 f(x)$ este un polinom de gradul $n - 2$. În final, $\Delta^n f(x)$ este un polinom de gradul zero (adică o constantă) și $\Delta^{n+1} f(x) = 0$.

Se poate arăta că

$$\Delta^k f(x) = f(x + kh) - \binom{k}{1} f(x - (k-1)h) +$$

$$+ \binom{k}{2} f(x+(k-2)h) + \ldots + (-1)^s \binom{k}{s} f(x + (k-s)h) + \ldots \tag{3.53}$$

unde coeficienții binomiali $\binom{k}{s}$ sînt dați prin

$$\binom{k}{s} = \frac{k(k-1)\ldots(k-s+1)}{s!} = \frac{k!}{s!(k-s)!} \tag{3.54}$$

pentru k și s întregi și pozitivi.

Evident că cele prezentate anterior sînt adevărate pentru $k = 1$, deoarece $\Delta f(x) = f(x + h) - f(x)$. Presupunînd că este adevărat pentru k, considerăm

$$\Delta^{k+1} f(x) = \Delta(\Delta^k f(x)) = \Delta f(x + kh) - \binom{k}{1} \Delta f(x+(k-1)h) +$$

$$+ \binom{k}{2} \Delta f(x+(k-2)h) + \ldots + (-1)^s \binom{k}{s} \Delta f(x+(k-s)h) + \ldots$$

$$\ldots + (-1)^k \Delta f(x) = f(x+(k+1)h) - f(x+kh) - \binom{k}{1} [f(x+kh) -$$

$$- f(x+(k-1)h)] + \binom{k}{2} [f(x+(k-1)h) - f(x+(k-2)h] + \ldots$$

$$\ldots + (-1)^s \binom{k}{s} [f(x + (k-s+1)h) - f(x+(k-s)h)] + \ldots$$

$$+ \ldots + (-1)^k f(x + h) + (-1)^{k+1} f(x). \tag{3.55}$$

Deoarece $\binom{k}{s}+\binom{k}{s-1}=\binom{k+1}{s}$, atunci are loc relația (3.53).

Se poate utiliza relația (3.53) pentru localizarea erorilor în tabelele cu diferențe. Astfel, dacă o eroare ε este comisă într-o valoare tabelată, această eroare va fi propagată în forma din tabel.

Exemplu

$$
\begin{array}{lccccccc}
x_0\ f_0 & & & & & & & \\
\quad \Delta f_0 & & & & & & & \\
x_1\ f_1 & \Delta^2 f_0 & & & & & & \\
\quad \Delta f_1 & & \Delta^3 f_0 & & & & & \\
x_2\ f_2 & \Delta^2 f_1 & & \Delta^4 f_0+\varepsilon & & & & \\
\quad \Delta f_2 & & \Delta^3 f_1+\varepsilon & & \Delta^5 f_0-5\varepsilon & & & \\
x_3\ f_3 & \Delta^2 f_2+\varepsilon & & \Delta^4 f_1-4\varepsilon & & \Delta^6 f_0+15\varepsilon & & \\
\quad \Delta f_3+\varepsilon & & \Delta^3 f_2-3\varepsilon & & \Delta^5 f_1+10\varepsilon & & \Delta^7 f_0-35\varepsilon & \\
x_4\ f_4+\varepsilon & \Delta^2 f_3-2\varepsilon & & \Delta^4 f_2+6\varepsilon & & \Delta^6 f_1-20\varepsilon & & \Delta^8 f_0+70 \\
\quad \Delta f_4-\varepsilon & & \Delta^3 f_3+3\varepsilon & & \Delta^5 f_2-10\varepsilon & & \Delta^7 f_1+35\varepsilon & \\
x_5\ f_5 & \Delta^2 f_4+\varepsilon & & \Delta^4 f_3+4\varepsilon & & \Delta^6 f_2+15\varepsilon & & \\
\quad \Delta f_5 & & \Delta^3 f_4+\varepsilon & & \Delta^5 f_3+5\varepsilon & & & \\
x_6\ f_6 & \Delta^2 f_5 & & \Delta^4 f_4+\varepsilon & & & & \\
\quad \Delta f_6 & & \Delta^3 f_5 & & & & & \\
x_7\ f_7 & \Delta^2 f_6 & & & & & & \\
\quad \Delta f_7 & & & & & & & \\
x_8\ f_8 & & & & & & &
\end{array}
$$

Coeficienții termenului eroare sînt în fiecare coloană coeficienții dezvoltări binomiale:

$$
\begin{array}{cccccccccc}
f & \Delta f & \Delta^2 f & \Delta^3 f & \Delta^4 f & \Delta^5 f & \Delta^6 f & \Delta^7 f & \Delta^8 f & \\
\hline
 & & & & +\varepsilon & & & & & \\
 & & & +\varepsilon & & -5\varepsilon & & & & \\
 & & +\varepsilon & & -4\varepsilon & & +15\varepsilon & & & \\
 & +\varepsilon & & -3\varepsilon & & +10\varepsilon & & -35\varepsilon & & \\
+\varepsilon & & -2\varepsilon & & +6\varepsilon & & -20\varepsilon & & +70\varepsilon & (3.56) \\
 & -\varepsilon & & +3\varepsilon & & -10\varepsilon & & +35\varepsilon & & \\
 & & +\varepsilon & & -4\varepsilon & & +15\varepsilon & & & \\
 & & & -\varepsilon & & +5\varepsilon & & & & \\
 & & & & +\varepsilon & & & & &
\end{array}
$$

Eroarea maximă apare în dreapta intrării valorii funcției care a fost afectată de eroarea ε.

3.4.1. *Formula de interpolare Gregory-Newton*

Dacă se introduce operatorul de translație $Ef(x) = f(x + h)$ și operatorul identic sau unitar $If(x) = f(x)$, atunci

$$\Delta f(x) = Ef(x) - f(x) = (E - I)f(x).$$

Evident $E^2f(x) = E[Ef(x)] = E[f(x+h)] = f(x + 2h)$ etc. Astfel se scrie

$$\Delta = E - I \text{ și } \Delta^k f(x) = (E - I)^k f(x). \tag{3.57}$$

Dezvoltînd binomul $(E - I)^k$, se obține

$$(E - I)^k = E^k - \binom{k}{1} E^{k-1} + \binom{k}{2} E^{k-2} + \ldots + (-1)^k I.$$

Deoarece $E = I + \Delta$, avem

$$E^k = I + \binom{k}{1}\Delta + \binom{k}{2}\Delta^2 + \ldots + \binom{k}{k}\Delta^k \tag{3.58}$$

și

$$f(x+kh) = f(x) + \binom{k}{1}\Delta f(x) + \binom{k}{2}\Delta^2 f(x) + \ldots + \binom{k}{k}\Delta^k f(x). \tag{3.59}$$

O demonstrație riguroasă poate fi dată prin inducție. Rezultatul este evident pentru $k = 1$, deoarece $f(x + h) = f(x) + \Delta f(x)$. Presupunînd că este adevărat pentru k, atunci

$$f[x + (k + 1)h] = f(x + kh) + \Delta f(x + kh) =$$

$$= \left[f(x) + \binom{k}{1}\Delta f(x) + \binom{k}{2}\Delta^2 f(x) + \ldots + \binom{k}{k}\Delta^k f(x)\right] +$$

$$+\left[\Delta f(x)+\binom{k}{1}\Delta^2 f(x)+\ldots+\binom{k}{k-1}\Delta^k f(x)\right]+$$

$$+\binom{k}{k}\Delta^{k+1}f(x). \tag{3.60}$$

Considerăm acum problema interpolării funcției $f(x)$ bazate pe valorile $x_0, x_1, \ldots, x_n$. Evident, valorile $f(x_0)$, $f(x_1), \ldots, f(x_n)$ pot fi utilizate la determinarea lui $\Delta f(x_0)$, $\Delta^2 f(x_0), \ldots, \Delta^n f(x_0)$. Considerăm acum funcția

$$F(x) = f(x_0) + \binom{u}{1}\Delta f(x_0) + \binom{u}{2}\Delta^2 f(x_0) + \ldots +$$

$$\ldots + \binom{u}{n}\Delta^n f(x_0), \tag{3.61}$$

unde $u = \dfrac{x - x_0}{h}$ și pentru orice s întreg pozitiv

$$\binom{u}{s} = \frac{u(u-1)\ldots(u-s+1)}{s!}.$$

Evident, prin (3.58) avem $F(x_k) = f(x_k)$, $k = 0, 1, 2, \ldots, n$. Mai mult, $F(x)$ este un polinom în x de grad $\leqslant n$. Deci $F(x)$ este același polinom cu polinomul de interpolare al lui Lagrange.

Exemplu. Pentru $f(0) = 1$, $f(1) = 2$, $f(2) = 4$ se construiește tabelul cu diferențe

n	x	$f(x)$	$\Delta f(x)$	$\Delta^2 f(x)$
0	0	1		
			1	
1	1	2		1
			2	
2	2	4		

Deoarece $h = 1$, $x_0 = 0$, avem $u = (x - x_0)/h = x$ și (3.61) devine

$$F(x) = f(x_0) + \binom{x}{1}\Delta f(x_0) + \binom{x}{2}\Delta^2 f(x_0) + \ldots + \binom{x}{n}\Delta^n f(x_0)$$

iar pentru exemplul considerat avem

$$F(x) = f(x_0) + x\Delta f(x_0) + \frac{x(x-1)}{2!}\Delta^2 f(x_0) = 1 + x \cdot 1 + \frac{x(x-1)}{2} \cdot 1 =$$

$$= 1 + \frac{1}{2!}x + \frac{1}{2}x^2,$$

care este același cu polinomul lui Lagrange obținut în celelalte exemple. În practică nu se exprimă $F(x)$ după puterile lui x. De exemplu, dacă se cere calculul lui $F\left(\frac{1}{2}\right)$, se va calcula

$$F\left(\frac{1}{2}\right) = f(x_0) + \frac{1}{2}\Delta f(x_0) + \frac{\frac{1}{2}\left(\frac{1}{2} - 1\right)}{2}\Delta^2 f(x_0) =$$

$$= 1 + \frac{1}{2} \cdot 1 + \left(-\frac{1}{8}\right) \cdot 1 = \frac{11}{8} = 1{,}375.$$

3.4.2. *Formula de interpolare cu diferențe centrate*

Dacă este necesar un calcul de interpolare extensiv, se va folosi formula de interpolare cu diferențe centrate, [4, 23], mult mai mult decît formula bazată pe polinomul Gregory-Newton.

Se presupune că $f(x)$ este dată pe valori spațiale echidistante numite x_0, x_1, $x_2, \ldots, x_n$, unde (3.47) are loc. Operatorul diferență centrată δ se definește astfel:

$$\delta f(x) = f(x + h/2) - f(x - h/2). \qquad (3.62)$$

Evident $\delta f(x)$ este definită numai pentru $x_{k+1/2}$, $k = 0, 1, \ldots, n-1$. În mod similar

$$\delta^k f(x) = \delta(\delta^{k-1} f(x)).$$

În continuare se prezintă un tabel cu diferențe centrate

k	x	$f(x)$	$\delta f(x)$	$\delta^2 f(x)$	$\delta^3 f(x)$	$\delta^4 f(x)$	$\delta^5 f(x)$
0	x_0	$f(x_0)$					
			$\delta f(x_{1/2})$				
1	x_1	$f(x_1)$		$\delta^2 f(x_1)$			
			$\delta f(x_{3/2})$		$\delta^3 f(x_{3/2})$		
2	x_2	$f(x_2)$		$\delta^2 f(x_2)$		$\delta^4 f(x_2)$	
			$\delta f(x_{5/2})$		$\delta^3 f(x_{5/2})$		$\delta^5 f(x_{5/2})$
3	x_3	$f(x_3)$		$\delta^2 f(x_3)$		$\delta^4 f(x_3)$	
			$\delta f(x_{7/2})$		$\delta^3 f(x_{7/2})$		
4	x_4	$f(x_4)$		$\delta^2 f(x_4)$			
			$\delta f(x_{9/2})$				
5	x_5	$f(x_5)$					

(3.63)

Se observă că elementele de pe aceeași linie au același indice.

Se vor prezenta două formule de interpolare cu diferențe centrate : formula Stirling și formula Bessel. Pentru a decide pe care să o folosim pentru x dat, la început se va găsi argumentul $\bar{x}$ în tabel care este apropiat de x și se calculează $u = \dfrac{x - \bar{x}}{h}$. Dacă $|u| \leqslant \dfrac{1}{4}$, se va utiliza formula Stirling, etichetîndu-se $\bar{x} = x_0$. Atunci se consideră punctele de interpolare x_{-p}, $x_{-(p-1)}$, ..., x_{-1}, x_0, x_1, ..., x_{p-1}, x_p pentru orice $p \geqslant 0$. Formula Stirling de interpolare este

$$F(x) = f(x_0) + u_1 N_1 f(x_0) + \frac{u^2}{2!}\delta^2 f(x_0) + \frac{u(u^2-1)}{3!} N_3 f(x_0) + \ldots$$

$$\ldots + \frac{u(u^2-1)(u^2-2)\ldots(u^2-(p-1)^2)}{(2p-1)!} N_{2p-1} f(x_0) +$$

$$+ \frac{u^2(u^2-1)(u^2-2)^2\ldots(u-(p-1)^2)}{(2p)!}\delta^{2p} f(x_0), \qquad (3.64)$$

unde

$$N_k f(x_0) = \frac{1}{2}\left[\delta^k f(x_{1/2}) + \delta^k f(x_{-1/2})\right], \quad k = 1,3,5. \qquad (3.65)$$

Exemplu. Se consideră formula pentru $p = 2$

h	x	$f(x)$	$\delta f(x)$	$\delta^2 f(x)$	$\delta^3 f(x)$	$\delta^4 f(x)$
-2	x_{-2}	$f(x_{-2})$				
			$\delta f(x_{-3/2})$			
-1	x_{-1}	$f(x_{-1})$		$\delta^2 f(x_{-1})$		
			$\delta f(x_{-1/2})$		$\delta^3 f(x_{-1/2})$	
0	x_0	$f(x_0)$		$\delta^2 f(x_0)$		$\delta^4 f(x_0)$
			$\delta f(x_{1/2})$		$\delta^3 f(x_{1/2})$	
1	x_1	$f(x_1)$		$\delta^2 f(x_1)$		
			$\delta f(x_{3/2})$			
2	x_2	$f(x_2)$				

Valorile încercuite din tabel sînt folosite în formula de interpolare Stirling.

Dacă $\frac{1}{4} < |u| \leqslant \frac{1}{2}$, se utilizează formula de interpolare Bessel. Se notează punctele de interpolare apropiată de x cu x_0 și x_1, unde $x_0 < x_1$. Evident $x \in [x_0, x_1]$. Se consideră punctele de interpolare x_{-p}, $x_{-(p-1)}, \ldots, x_{-1}$, x_0, $x_1, \ldots, x_{p+1}$. Formula de interpolare Bessel este dată prin

$$F(x) = N_0 f(x_{1/2}) + v\delta f(x_{1/2}) + \frac{v^2 - 1/4}{2!} N_2 f(x_{1/2}) +$$

$$+ \frac{v(v^2 - 1/4)}{3!} \delta^3 f(x_{1/2}) + \ldots$$

$$\ldots + \frac{\left(v^2 - \frac{1}{4}\right)\left(v^2 - \frac{9}{4}\right)\ldots[v^2 - (2p-1)^2/4]}{(2p)!} N_{2p} f(x_{1/2}) +$$

$$+ \frac{v(v^2 - 1/4)(v^2 - 9/4)\ldots[v^2 - (2p-1)^2/4]}{(2p+1)!} \delta^{2p+1} f(x_{1/2}),$$

unde

$$v = u - 1/2,$$

$$\mathcal{N}_k f(x_{1/2}) = \frac{1}{2}\,[\delta^k f(x_0) + \delta^k f(x_1)], \qquad k=0,2,4. \tag{3.67}$$

Se va ilustra cu un exemplu tabelul cu diferențe pentru cazul $p = 2$.

k	x	$f(x)$	$\delta f(x)$	$\delta^2 f(x)$	$\delta^3 f(x)$	$\delta^4 f(x)$	$\delta^5 f(x)$
-2	x_{-2}	$f(x_{-2})$					
			$\delta f(x_{-3/2})$				
-1	x_{-1}	$f(x_{-1})$		$\delta^2 f(x_{-1})$			
			$\delta f(x_{-1/2})$		$\delta^3 f(x_{-1/2})$		
0	x_0	$f(x_0)$		$\delta^2 f(x_0)$		$\delta^4 f(x_0)$	
			$\delta f(x_{1/2})$		$\delta^3 f(x_{1/2})$		$\delta^5 f(x_{1/2})$
1	x_1	$f(x_1)$		$\delta^2 f(x_1)$		$\delta^4 f(x_1)$	
			$\delta f(x_{3/2})$		$\delta^3 f(x_{3/2})$		
2	x_2	$f(x_2)$		$\delta^2 f(x_2)$			
			$\delta f(x_{5/2})$				
3	x_3	$f(x_3)$					

Valorile care sînt încercuite sînt acelea utilizate în formulele de interpolare.

Presupunînd că se dă următorul tabel de valori:

x	0	5	10	15	20
$f(x)$	0	0,08716	0.17365	0,25882	0,34202

și se dorește evaluarea lui $f(12)$, atunci $\bar{x} = 10$ și $u=2/5=0{,}40$. Aceasta înseamnă că formula Bessel se va utiliza cu $x_0 = 10$, $x_1 = 15$. Se poate lua $p = 0$ și avem doar două puncte de interpolare, sau se poate lua $p = 1$ și $x_{-1} = 5$, $x_0 = 10$, $x_1 = 15$, $x_2 = 20$. Numărul v va fi egal cu $-0{,}1$.

Dacă se caută $f(9)$, atunci $\bar{x} = 10$ și $u=-0{,}2$. Deci se va utiliza formula Stirling cu unul din următoarele seturi de puncte:

$p = 0$: $x_0 = 10$

$p = 1$: $x_{-1} = 5$ $x_0 = 10$ $x_1 = 15$

$p = 2$: $x_{-2} = 0$ $x_{-1} = 5$ $x_0 = 10$ $x_1 = 15$ $x_2 = 20$.

Formulele de interpolare cu diferențe centrate au numeroase avantaje în comparație cu formulele Gregory-Newton. Este adevărat că interpolarea pentru același set de puncte fixat va conduce la același rezultat, nu are importanță care formulă este utilizată. Totuși în cazul formulei Gregory-Newton, valoarea lui u va fi suficient de mare în general, dacă punctele de interpolare sînt aproximativ simetric plasate relativ la argumentul x pentru care se caută $f(x)$. Aceasta înseamnă că termenii vor tinde la zero mai încet decît în cazul formulelor cu diferențe centrate, unde u (sau v) este mic cînd punctele de interpolare sînt plasate simetric față de x. Dacă se aleg puncte de interpolare astfel ca u utilizat în formula Gregory-Newton va fi mic, atunci valoarea lui x va fi aproape de sfîrșitul intervalului acoperit de punctele de interpolare. Aceasta conduce la o eroare mare, deoarece expresia

$$Q_n(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$$

care apare în formula erorii are un extrem destul de mare in intervalul $x_i \leqslant x \leqslant x_{i+1}$, aproape de sfîrșitul intervalului decît de mijlocul intevalului. Astfel eroarea de acoperire este dată prin expresia

$$\frac{(x - x_0)(x - x_1)\ldots(x - x_n)}{(n+1)!} f^{(n+1)}(c) \qquad (3.68)$$

și va fi normal destul de mare pentru formula Gregory-Newton, unde u este mic, decît pentru formulele cu diferențe centrate utilizate cu seturi diferite ale lui x_i, unde u (sau v) este mic.

● *Estimarea preciziei de interpolare.* Se consideră precizia polinomului de interpolare $F(x)$ ca o reprezentare a lui $f(x)$ cînd $F(x)$ este un polinom de gradul n sau mai mic, polinom care coincide cu $f(x)$ în $n+1$ puncte $x_k = x_0 + kh$, $k = 0, 1, \ldots, n$. Din cele prezentate s-a văzut că pentru diverse valori ale lui n se poate obține o mărginire bună pentru $|Q_n(x)|$ în intervalul $x_0 \leqslant x \leqslant x_n$, unde

$$Q_n(x) = \prod_{k=0}^{n} (x - x_k). \qquad (3.69)$$

Cea mai dificilă problemă este estimarea unei margini pentru cele $n + 1$ derivate ale lui $f(x)$.

Dacă $f(x)$ este un polinom de gradul $n + 1$ sau mai mic

$$f(x) = a_0 x^{n+1} + a_1 x^n + \ldots + a_{n+1},$$

atunci ca o demonstrație a teoremei privind liniaritatea operatorilor diferență avem

$$\Delta^{n+1} f(x) = h^{n+1}(n + 1)\,! \, a_0 = h^{n+1} f^{(n+1)}(x) \qquad (3.70)$$

și deci

$$f^{(n+1)}(x) = \frac{\Delta^{n+1} f(x)}{h^{n+1}}. \qquad (3.71)$$

Astfel în acest caz se poate determina $f^{(n+1)}(x)$ exact prin raportul diferenței de ordinul $n + 1$ a funcției $f(x)$ cu h^{n+1}.

În cazul general, dacă $f(x) \in C^{(n+1)}$ cu $x_0 \leqslant x \leqslant x_0 + h$, pentru un anumit $h > 0$, se poate arată [128, 24], că

$$\lim_{h \to 0} \frac{\Delta^{n+1} f(x_0)}{h^{n+1}} = f^{(n+1)}(x_0). \qquad (3.72)$$

Trebuie multă atenție la funcțiile tabelate obținute în urma măsurărilor unde există variații foarte mari ale funcției între punctele de interpolare [29, 25].

3.5. Interpolarea hermitiană

Într-o serie de aplicații practice pentru aproximarea unor funcții se cere ca funcția de aproximare să coincidă cu funcția de aproximat în punctele de interpolare x_0, $x_1, \ldots, x_n$, precum și derivatele funcției de aproximare (de un anumit ordin) să coincidă cu derivatele funcției de aproximat (de un anumit ordin) în anumite puncte

din punctele de interpolare date. O astfel de interpolare care impune condiții atît funcției de aproximare cît și derivatelor acesteia este întîlnită în literatură sub denumirea de *interpolare hermitiană.*

În cazul interpolării hermitiene, dacă funcția de interpolare se notează cu $H(x)$, aceasta trebuie să îndeplinească următoarele condiții :

$$H(x_k) = f(x_k),\ k = 0, 1, 2, \ldots, n, \tag{3.73}$$

și pentru fiecare k, astfel ca $\alpha_k \geqslant 1$, să aibă loc relațiile

$$H^{(i)}(x_k) = f^{(i)}(x_k),\ i = 1, 2, \ldots, \alpha_k. \tag{3.74}$$

Se pune problema găsirii unui polinom unic $H(x)$ de grad m sau mai mic, unde

$$m = n + \sum_{i=0}^{n} \alpha_i$$

care să satisfacă condițiile (3.73) și (3.74). Problema se reduce la găsirea a $m + 1$ coeficienți $c_0, c_1, c_2, \ldots, c_m$ astfel ca cele $m + 1$ condiții date prin (3.73) și (3.74) să aibă loc pentru

$$H(x) = \sum_{i=0}^{m} c_i (x - \beta)^{m-i}, \tag{3.75}$$

unde β este o constantă oarecare.

Dacă se pune condiția că polinomul $H(x)$ dat prin (3.75) să îndeplinească condițiile (3.73) și (3.74), rezultă un sistem de $m + 1$ ecuații liniare cu $m + 1$ necunoscute $c_0, c_1, \ldots, c_m$, Astfel polinomul $H(x)$ poate fi găsit prin metoda coeficienților nedeterminați.

Dacă se consideră cazul în care α_k este zero sau unu (adică lui $H(x)$ i se cere doar ca derivata sa de ordinul întîi să coincidă cu derivata lui $f(x)$ în anumite puncte de

interpolare), atunci o formă explicită pentru $H(x)$ arată astfel [128, 37]:

$$H(x) = \sum_{k=0}^{n} \frac{P_k(x)}{P_k(x_k)} \left\{ \left[1 - (x - x_k) \frac{P_k'(x_k)}{P_k(x_k)} \right] f(x_k) + \right.$$

$$\left. + \alpha_k (x - x_k) f'(x_k) \right\}, \tag{3.76}$$

unde

$$P_k(x) = \prod_{\substack{i=1 \\ i \neq k}}^{n} (x - x_i)^{\alpha_i + 1} \tag{3.77}$$

Exemplu. Se consideră $n = 1$, $\alpha_0 = \alpha_1 = 1$, $x_0 = a$ și $x_1 = b$. Atunci din (3.77) rezultă

$$\left.\begin{aligned} P_1 &= (x - x_1)^{\alpha_1 - 1} = (x - x_1)^2 = (x - b)^2 \\ P_0 &= (x - x_0)^{\alpha_0 + 1} = (x - x_0)^2 = (x - a)^2 \end{aligned}\right\}. \tag{3.78}$$

Polinomul $H(x)$ devine

$$H(x) = \frac{(x-b)^2}{(a-b)^2} \left\{ \left[1 - (x - a) \frac{2(a-b)}{(a-b)^2} \right] f(a) + (x - a) f'(a) \right\} +$$

$$+ \frac{(x - a)^2}{(b - a)^2} \left\{ \left[1 - (x - b) \frac{2(b - a)}{(b - a)^2} \right] f(b) + (x - b) f'(b) \right\} =$$

$$= \frac{(x - b)^2}{(a - b)^2} \left\{ \left[1 - \frac{2(x - a)}{b - a} \right] f(a) + (x - a) f'(a) \right\} +$$

$$+ \frac{(x - a)^2}{(b - a)^2} \left\{ \left[1 - \frac{2(x - b)}{b - a} \right] f(b) + (x - b) f'(b) \right\} \tag{3.79}$$

Se observă că $H(x)$ este un polinom de gradul trei sau mai mic și

$$H(a) = f(a); \quad H(b) = f(b); \quad H'(a) = f'(a) \quad \text{și} \quad H'(b) = f'(b).$$

Se aplică metoda coeficienților nedeterminați pentru găsirea lui c_0, c_1, c_2, c_3, unde $H(x)$ are forma dată de (3.75):

$$\left.\begin{aligned} H(x) &= c_0(x-\beta)^3 + c_1(x-\beta)^2 + c_2(x-\beta) + c_3 \\ H'(x) &= 3c_0(x-\beta)^2 + 2c_1(x-\beta) + c_2 \end{aligned}\right\} \tag{3.80}$$

Pentru cazul considerat $x_0 = a$, $x_1 = b$, rezultă patru ecuații:

$$\left.\begin{aligned} H(a) &= f(a) = c_0(a-\beta)^3 + c_1(a-\beta)^2 + c_2(a-\beta) + c_3 \\ H(b) &= f(b) = c_0(b-\beta)^3 + c_1(a-\beta)^2 + c_2(b-\beta) + c_2 \\ H'(a) &= f'(a) = 3c_0(a-\beta)^2 + 2c_1(a-\beta) + c_2 \\ H'(b) &= f'(b) = 3c_0(b-\beta)^2 + 2c_1(b-\beta) + c_2 \end{aligned}\right\} \tag{3.81}$$

Constanta β fiind oarecare, pentru simplificarea sistemului (3.81) se consideră $\beta = a$ și $b - a = h$; astfel sistemul devine

$$\left.\begin{aligned} c_3 &= f(a) \\ c_0h^3 + c_1h^2 + c_2h + c_3 &= f(b) \\ c_2 &= f'(a) \\ 3c_0h^2 + 2c_1h + c_2 &= f'(b) \end{aligned}\right\} \tag{3.82}$$

Sistemul (3.82) după rezolvare conduce la următoarele soluții:

$$\begin{aligned} c_0 &= \frac{2}{h^3}[f(a) - f(b)] + \frac{1}{h^2}[f'(a) + f'(b)], \\ c_1 &= \frac{3}{h^3}[f(b) - f(a)] - \frac{1}{h}[2f'(a) + f'(b)], \\ c_2 &= f'(a), \qquad c_3 = f(a). \end{aligned} \tag{3.83}$$

Dacă coeficienții c_0, c_1, c_2, c_3 astfel aflați se introduc în expresia lui $H(x)$ dată în (3.80) pentru $\beta = a$, rezultă expresia lui $H(x)$ dată în (3.79).

Polinomul $H(x)$ se mai poate reprezenta şi sub forma [128, 48]

$$H(x) = \sum_{k=0}^{n} [t_k(x)f(x_k) + t_k^*(x)f'(x_k)], \qquad (3.84)$$

unde $t_k(x)$ şi $t_k^*(x)$ sînt mărimi nedeterminate care există şi sînt unice pentru $k = 0, 1, \ldots, n$. În cazul interpolării hermitiene prin această metodă pentru condiţiile definite în (3.73) şi (3.74) există mărimile

$$t_{k,j}(x),\ k = 0, 1, \ldots, n,\ j = 0, 1, \ldots, \alpha_k, \qquad (3.85)$$

astfel că

$$H(x) = \sum_{j=0}^{\alpha_0} t_{0,j}(x)f^{(j)}(x_0) + \sum_{j=0}^{\alpha_1} t_{1,j}(x)f^{(j)}(x_1) + \ldots$$

$$\ldots + \sum_{j=0}^{\alpha_n} t_{n,j}(x)f^{(j)}(x_n) = \sum_{k=0}^{n}\sum_{j=0}^{\alpha_k} t_{k,j}(x)f^{(j)}(x_k). \qquad (3.86)$$

Metode pentru determinarea mărimilor $t_{k,j}(x)$ din (3.85) sînt date în [128, 59], rezultînd că $H(x)$ din (3.86) este unic determinat.

Pentru a evidenţia eroarea în cazul interpolării hermitiene, se procedează asemănător ca la polinomul lui Lagrange de interpolare. Se poate arăta [67, 75] că

$$f(x) - H(x) = \prod_{j=0}^{n} (x - x_j)^{\alpha_j+1} T(x), \qquad (3.87)$$

unde $T(x)$ este o funcţie continuă şi

$$T(x) = \frac{1}{(k+1)!} f^{(k+1)}(\zeta), \qquad (3.88)$$

ζ aparţinînd intervalului acoperit de punctele de interpolare. În acest caz [128, 81].

$$f(x) - H(x) = \frac{1}{(k+1)!} \prod_{j=0}^{n} (x - x_j)^{\alpha_j+1} f^{(k+1)}(\zeta). \qquad (3.89)$$

Se poate arăta că pentru cazul analizat $n=1$, $\alpha_0=\alpha_1=1$, $x_0 = a$, $x_1 = b$ și $H(x)$ determinat prin relațiile (3.78) — — (3.83), eroarea are următoarea expresie :

$$f(x) - H(x) = \frac{(x - x_0)^2 (x - x_1)^2}{24} f^4(\zeta) =$$

$$= \frac{(x - a)^2 (x - b)^2}{24} f^4(\zeta) \qquad (3.90)$$

unde $\zeta \in [x_0, x_1] = [a, b]$.

3.6. Interpolarea inversă

Se consideră o funcție dată sub forma

x	x_0	x_1	x_2	x_3	x_4
$f(x)$	y_0	y_1	y_2	y_3	y_4

(3.91)

De această dată se pune problema invers : se cere determinarea unei valori x astfel că pentru o valoare a lui y, notată cu y_k, să aibă loc relația

$$I(x) = y_k. \qquad (3.92)$$

Înterpolarea inversă este utilizată frecvent. Se pune problema cercetării tabelului prin care este definită funcția pentru găsirea a două puncte de interpolare care se vor nota cu x_0 și x_1 astfel ca să aibă loc relația

$$[f(x_0) - y_k][f(x_1) - y_k] \leqslant 0. \qquad (3.93)$$

Se construiește funcția liniară, de interpolare inversă $I(x)$, astfel [128, 98] :

$$I(x) = \frac{x_1 - x}{x_1 - x_0} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1) = y_k. \qquad (3.94)$$

Din (3.94) se exprimă x:

$$x = \left(\frac{f(x_1) - y_k}{f(x_1) - f(x_0)}\right) x_0 + \left(\frac{y_k - f(x_0)}{f(x_1) - f(x_0)}\right) x_1. \quad (3.95)$$

Această valoare a lui x reprezintă valoarea argumentului pentru care funcția are valoarea y_k.

3.7. Aproximarea funcțiilor prin polinoame

În cadrul acestui paragraf se consideră reprezentarea unei funcții $f(x)$ (care este definită și continuă pe un interval) printr-un polinom $T_n(x)$ de grad n sau mai mic. În [81, 42] se prezintă diverse criterii pentru măsura calității procesului de aproximare a lui $f(x)$ prin $T_n(x)$.

Se pune problema construirii unui polinom $T_n(x)$ care să minimizeze relația

$$\max_{a \leqslant x \leqslant b} | T_n(x) - f(x) |, \quad (3.96)$$

numită *norma uniformă* a lui $T_n - f$ și este notată prin $\| T_n - f \|$ sau $\| T_n - f \|_\infty$.

O altă măsură utilizată la aprecierea aproximării este norma definită prin

$$\left\{ \int_a^b (T_n(x) - f(x))^2 \, dx \right\}^{1/2} \quad (3.97)$$

care este de obicei notată prin $\| T_n - f \|_2$.

Pentru construcția lui $T_n(x)$ se vor enunța două propoziții necesare:

● Dacă $f(x)$ este continuă pe $[a, b]$, atunci pentru un n întreg există un polinom unic determinat $t_n(x)$ de grad $\leqslant n$, astfel că

$$\varepsilon_n = \max_{a \leqslant x \leqslant b} | t_n(x) - f(x) | \leqslant \max_{a \leqslant x \leqslant b} | T_n(x) - f(x) | \quad (3.98)$$

pentru orice polinom $T_n(x)$ de grad $\leqslant n$.

Mai mult, există un set de $n+2$ puncte $x_0, x_1, \ldots, x_{n+1}$ din $[a, b]$, unde $x_0 < x_1 < \ldots < x_{n+1}$, astfel că

$$t_n(x_i) - f(x_i) = (-1)^i \varepsilon_n, \quad i = 0, 1, \ldots, n, \tag{3.99}$$

sau

$$t_n(x_i) - f(x_i) = (-1)^{i+1} \varepsilon_n, \quad i = 0, 1, \ldots, n.$$

Se poate demonstra că $\varepsilon_n \to 0$ atunci cînd $n \to \infty$ [42, 128].

● Dacă f este o funcție continuă pe intervalul $[a, b]$, atunci dîndu-se un $\varepsilon > 0$, există un polinom $T_n(x)$ astfel că pentru $x \in [a, b]$ avem

$$|T_n(x) - f(x)| < \varepsilon. \tag{3.100}$$

Demonstrația pentru relația (3.100) se poate găsi în [128, 46].

În [42, 128] se arată că în general nu există un algoritm finit pentru găsirea polinomului de cea mai bună aproximație $t_n(x)$.

Dacă se înlocuiește intervalul $[a, b]$ printr-o mulțime de puncte M, din intervalul $[a, b]$, atunci într-un număr finit de etape se poate găsi polinomul $t_n(x)$ de grad $\leqslant n$, astfel că

$$\max_{x \in M} |t_n(x) - f(x)| \leqslant \max_{x \in M} |T_n(x) - f(x)| \tag{3.101}$$

pentru orice polinom $T_n(x)$ de grad $\leqslant n$.

În [128, 59] se prezintă un algoritm pentru găsirea lui $t_n(x)$.

3.7.1. *Aproximare polinomială prin metoda celor mai mici pătrate*

Se pune problema aproximării unei funcții $f(x)$ continue pe intervalul $[a, b]$ prin polinomul $T_n(x)$ de grad $\leqslant n$, în așa fel ca norma l_2, dată prin

$$\|T_n - f\|_2 = \left\{ \int_a^b (T_n(x) - f(x))^2 \mathrm{d}x \right\}^{1/2}, \tag{3.102}$$

să fie minimă pentru toate polinoamele $T_n(x)$ de grad $\leqslant n$.

● În locul intervalului continuu $a \leqslant x \leqslant b$, se va considera o mulțime finită de puncte $M : x_1, x_2, \ldots, x_N$ și se urmărește minimizarea normei l_2 :

$$\| T_n - f \|_2 = \left[\sum_{i=1}^{N} (T_n(x_i) - f(x_i))^2 \right]^{1/2}. \qquad (3.103)$$

Se definește produsul intern a două funcții f și g prin (f, g) dat sub forma

$$(f, g) = \begin{cases} \int f(x) g(x) \, dx & \text{pentru } x \in [a, b] \\ & \text{(cazul continuu),} \\ \sum_{i=1}^{n} f(x_i)\, g(x_i) & \text{pentru } x_i \in M,\ i = 1, 2, \ldots, N, \\ & \text{(cazul discret).} \end{cases} \qquad (3.104)$$

În continuare se va construi o familie de polinoame $t_0(x), t_1(x), \ldots$, cu gradul indicat de indicele inferior, polinoame care sînt ortogonale în sensul că

$$(t_i, t_j) = 0, \text{ dacă } i \neq j. \qquad (3.105)$$

Se observă că dacă $N > n$, atunci

$$(t_i, t_j) = \| t_i \|^2 > 0, \quad i = 0, 1, 2, \ldots, n, \qquad (3.106)$$

pentru că altfel $t_i(x) = 0$ pe $[a, b]$ sau pe mulțimea discretă M.

Dacă un polinom de grad $\leqslant n$ este nul în mai mult decît n puncte, acesta trebuie să fie identic nul. De asemenea în cazul intervalului continuu $[a, b]$

$$(t, t) = \| t \|_2^2 > 0 \qquad (3.107)$$

afară de cazul cînd $t_i(x) = 0$ pe $[a, b]$. Relația (3.107) are sens și în cazul discret afară de cazul cînd $t(x) = 0$ pe mulțimea discretă M.

În concluzie, pentru $N > n$, dacă $t(x)$ este un polinom de grad $\leqslant n$, atunci sau are loc relația (3.107) sau altfel $t(x) \equiv 0$ nu numai pentru mulțimea M, dar pentru orice x.

Construcția [128, 42] șirului de polinoame $t_0(x), t_1(x), \ldots$ se realizează cu ajutorul următoarelor relații de recurență :

$$\left.\begin{aligned} t_0(x) &= 1 \\ t_1(x) &= xt_0(x) - \frac{(xt_0, t_0)}{(t_0, t_0)} t_0(x) = x - \frac{(x, 1)}{(1, 1)} \\ t_{k+1}(x) &= xt_k(x) - a_k t_k(x) - b_k t_{k-1}(x), \quad k = 1, 2, \ldots \end{aligned}\right\} \tag{3.108}$$

unde

$$a_k = \frac{(xt_k, t_k)}{(t_k, t_k)}, \qquad b_k = \frac{(t_k, t_k)}{(t_{k-1}, t_{k-1})}. \tag{3.109}$$

În continuare se va arăta că polinoamele construite prin relația de recurență (3.108) satisfac următoarea relație :

$$(t_{k-3}, t_{k+1}) = (t_{k-4}, t_{k+1}) = \ldots = (t_1, t_{k+1}) = (t_0, t_{k+1}) = 0, \tag{3.110}$$

deci (3.105) rezultă prin inducție. Se arată foarte ușor că $(t_0, t_1) = 0$.

Presupunînd că (3.105) are sens pentru orice $i, j \leqslant k$, este ușor de arătat că $(t_k, t_{k+1}) = 0$, sau mai mult

$$\begin{aligned} (t_{k-1}, t_{k+1}) &= (t_{k-1}, xt_k(x) - a_k t_k(x) - b_k t_{k-1}(x)) = \\ &= (t_{k-1}, xt_k) - a_k(t_{k-1}, t_k) - b_k(t_{k-1}, t_{k-1}) = \\ &= (xt_{k-1}, t_k) - b_k(t_{k-1}, t_{k-1}), \end{aligned} \tag{3.111}$$

deoarece $(t_{k-1}, t_k) = 0$. Dar din (3.108) pentru $k \geqslant 2$

$$xt_{k-1}(x) = t_k(x) + a_{k-1} t_{k-1}(x) + b_{k-1} t_{k-2}(x), \tag{3.112}$$

astfel că folosind (3.112) și (3.105), rezultă

$$\begin{aligned} (xt_{k-1}, t_k) &= (t_k + a_{k-1} t_{k-1} + b_{k-1} t_{k-2}, t_k) = \\ &= (t_k, t_k) + a_{k-1}(t_{k-1}, t_k) + b_{k-1}(t_{k-2}, t_k) = (t_k, t_k). \end{aligned} \tag{3.113}$$

Introducînd (3.113) în (3.111), se obține

$$(t_{k-1},\ t_{k+1}) = (t_k, t_k) - b_k(t_{k-1},\ t_{k-1}). \qquad (3.114)$$

Folosind expresia lui b_k dată în (3.109), (3.114) devine

$$(t_{k-1},\ t_{k+1}) = (t_k,\ t_k) - \frac{(t_k,\ t_k)}{(t_{k-1},\ t_{k-1})}(t_{k-1},\ t_{k-1}) =$$

$$= (t_k,\ t_k) - (t_k,\ t_k) = 0. \qquad (3.115)$$

Deci $(t_{k-1},\ t_{k+1}) = 0$ pentru $k \geqslant 2$.

Dacă $k = 1$, din (3.108) rezultă

$$xt_{k-1}(x) = xt_0(x) = t_1(x) + \frac{(xt_0,\ t_0)}{(t_0,\ t_0)} = t_1(x) + \frac{(x,\ 1)}{(1,\ 1)}\,t_0(x) \qquad (3.116)$$

și

$$(xt_0,\ t_1) = \left(t_1 + \frac{(x,\ 1)}{(1,\ 1)}\,t_0,\ t_1\right) = (t_1,\ t_1) + \frac{(x,\ 1)}{(1,\ 1)}\,(t_0,\ t_1) = (t_1,\ t_1),$$

deoarece $(1,\ t_1) = 0$. Astfel pentru $k = 1$ (3.115) devine $(t_0,\ t_2) = 0$ și, prin urmare,

$$(t_{k-1},\ t_{k+1}) = 0 \text{ pentru } k \geqslant 1. \qquad (3.117)$$

În continuare se va evalua produsul intern (t_{k-2}, t_{k+1}). Folosind relațiile (3.108), rezultă

$$(t_{k-2},\ t_{k+1}) = (t_{k-2},\ xt_k - a_k t_k - b_k t_{k-1}) =$$

$$= (t_{k-2},\ xt_k) - a_k(t_{k-2},\ t_k) - b_k(t_{k-2},\ t_{k-1}) = (xt_{k-2},\ t_k). \qquad (3.118)$$

Dacă $k \geqslant 3$, atunci din (3.118) rezultă

$$xt_{k-2} = t_{k-1} + a_{k-2}\,t_{k-2} + b_{k-2}\,t_{k-2} \qquad (3.119)$$

și dacă se înlocuiește în partea dreaptă a relației (3.118), se obține

$$(t_{k-2}, t_{k+1}) = (xt_{k-2}, t_k) = (t_{k-1} + a_{k-2}t_{k-2} + b_{k-2}t_{k-3}, t_k) =$$

$$= (t_{k-1}, t_k) + a_{k-2}(t_{k-2}, t_k) + b_{k-2}(t_{k-3}, t_k) = 0. \qquad (3.120)$$

În concluzie se vede că relația (3.110) are loc pentru toate valorile lui k, deci (3.105) rezultă prin inducție.

Exemplu (caz continuu). Se cere aproximarea funcției $f(x) = x^3 - 1$ printr-un polinom de grad $\leqslant 2$ pe intervalul $a \leqslant x \leqslant b$, unde $a = 0$, $b = 1$. În acest caz se găsesc polinoamele cu ajutorul relației (3.108) :

$$t_0(x) = 1,$$

$$t_1(x) = xt_0 - \frac{(xt_0, t_1)}{(t_0, t_0)} t_0(x) = x - \frac{(x, 1)}{(1, 1)} = x - \frac{\int_0^1 x\,dx}{\int_0^1 dx} = x - \frac{1}{2},$$

$$t_2(x) = xt_1(x) - a_1 t_1(x) - b_1 t_0(x),$$

$$a_1 = \frac{(xt_1, t_1)}{(t_1, t_1)} = \frac{\left(x^2 - \frac{1}{2}x, x - \frac{1}{2}\right)}{\left(x - \frac{1}{2}, x - \frac{1}{2}\right)} =$$

$$= \frac{\int_0^1 x^2 dx - \frac{1}{2}\int_0^1 x^2 dx + \frac{1}{4}\int_0^1 x\,dx}{\int_0^1 x^2\,dx - \int_0^1 x\,dx + \frac{1}{4}\int_0^1 dx} = \frac{\frac{x^4}{4}\Big|_0^1 - \frac{x^3}{3}\Big|_0^1 + \frac{1}{4}\frac{x^2}{2}\Big|_0^1}{\frac{x^3}{3}\Big|_0^1 - \frac{x^2}{2}\Big|_0^1 + \frac{1}{4}x\Big|_0^1} =$$

$$= \frac{\frac{1}{4} - \frac{1}{3} + \frac{1}{8}}{\frac{1}{3} - \frac{1}{2} + \frac{1}{4}} = \frac{\frac{1}{24}}{\frac{1}{12}} = \frac{1}{2},$$

$$b_1 = \frac{(t_1, t_1)}{(t_0, t_0)} = \frac{\left(x - \frac{1}{2}, x - \frac{1}{2}\right)}{(1,1)} = \frac{\frac{1}{12}}{1} = \frac{1}{12}.$$

Deci pentru $a_1 = \frac{5}{2}$ și $b_1 = \frac{1}{12}$ rezultă

$$t_2(x) = x\left(x - \frac{1}{2}\right) - \frac{1}{2}\left(x - \frac{1}{2}\right) - \frac{1}{12} \cdot 1 =$$

$$= x^2 - \frac{1}{2}x - \frac{1}{2}x + \frac{1}{4} - \frac{1}{12} = x^2 - x + \frac{1}{6}.$$

La fel se calculează și $t_3(x)$ cu ajutorul formulelor

$$t_3(x) = xt_2(x) - a_2t_2(x) - b_2t_1(x),$$

$$a_2 = \frac{(xt_2,\ t_2)}{(t_2,\ t_2)} \text{ și } b_2 = \frac{(t_2,\ t_2)}{(t_1,\ t_1)}.$$

În rezumat, se obține șirul de polinoame $t_0,\ t_1,\ t_2,\ t_3, \ldots$ Pentru aplicația considerată sînt suficiente polinoamele

$$t_0(x) = 1,\quad t_1(x) = x - \frac{1}{2},\quad t_2(x) = x^2 - x + \frac{1}{6}. \qquad (3.121)$$

Se poate arăta că are loc relația $(t_0,\ t_1) = (t_0,\ t_2) = (t_1,\ t_2) = 0$ pentru domeniul considerat (0,1), deci polinoamele sînt ortogonale.

Fiecare polinom $t_k(x)$, $k = 0, 1, 2, \ldots$, este un polinom de gradul k, cu coeficientul lui x^k egal cu unitatea. Datorită acestui fapt x^k se poate exprima ca o combinație liniară de polinoame :

$$t_0(x),\ t_1(x),\ t_2(x), \ldots, t_k(x).$$

Prin urmare, pentru orice polinom $T_n(x)$ de grad $\leqslant n$ avem

$$T_n(x) = \sum_{k=0}^{n} \lambda_k t_k(x). \qquad (3.122)$$

În continuare se urmărește alegerea constantelor λ_k astfel ca să minimizeze expresia

$$I = \| T_n(x) - f(x) \|_2^2 = \left\| \sum_{k=0}^{n} \lambda_k t_k(x) - f(x) \right\|_2^2. \qquad (3.123)$$

Ținînd seama de relațiile (3.106) și (3.122), se poate scrie I sub formă dezvoltată:

$$I = (T_n(x) - f(x),\ T_n(x) - f(x)) =$$

$$= \left(\sum_{k=0}^{n} \lambda_k t_k(x) - f(x),\ \sum_{k=0}^{n} \lambda_k t_k(x) - f(x) \right) =$$

$$= \sum_{k=0}^{n} \lambda_k^2 (t_k, t_k) - 2 \sum_{k=0}^{n} \lambda_k (t_k, f) + (f, f). \qquad (3.124)$$

Pentru a determina constantele λ_k care să minimizeze expresia lui I, se derivează I în raport cu λ_k:

$$\left.\begin{aligned} \frac{\partial I}{\partial \lambda_k} &= 2[\lambda_k(t_k, t_k) - (t_k, f)], \\ \frac{\partial^2 I}{\lambda \partial_k \partial \lambda_p} &= \begin{cases} 2(t_k, t_k), & \text{dacă } k = p, \\ 0, & \text{dacă } k \neq p. \end{cases} \end{aligned}\right\} \qquad (3.125)$$

Prin urmare, I admite un minim absolut dacă

$$\frac{\partial I}{\partial \lambda_k} = 0 \text{ sau } \lambda_k = \frac{(t_k, f)}{(t_k, t_k)}, \quad k = 0, 1, 2, \ldots, n. \qquad (3.126)$$

Introducînd valorile lui λ_k din (3.126) în expresia lui I din (3.124), rezultă că pentru λ_k ales avem

$$I = \sum_{k=0}^{n} \frac{(t_k, f)^2}{(t_k, t_k)^2} (t_k, t_k) - 2 \sum_{k=0}^{n} \frac{(t_k, f)}{(t_k, t_k)} (t_k, f) + (f, f),$$

de unde, după executarea simplificărilor, rezultă

$$I = (f, f) - \sum_{k=0}^{n} \frac{(t_k, f)}{(t_k, t_k)}. \qquad (3.127)$$

Pentru exemplul considerat, dacă se folosesc polinoamele t_k, determinate în (3.121) și faptul că s-a dorit aproximarea funcției printr-un polinom $T_2(x)$ de grad $\leqslant 2$, atunci

$$T_2(x) = \sum_{k=0}^{2} \lambda_k t_k = \lambda_0 t_0(x) + \lambda_1 t_1(x) + \lambda_2 t_2(x) =$$

$$= \lambda_0 + \lambda_1 \left(x - \frac{1}{2}\right) + \lambda_2 \left(x^2 - x + \frac{1}{6}\right).$$

Funcția considerată pentru a fi aproximată este $f(x) = x^3 - 1$, $x \in [0,1]$.

Constantele λ_k se determină cu ajutorul relațiilor (3.126), deci pentru exemplul considerat rezultă:

$$\lambda_0 = \frac{(t_0, f)}{(t_0, t_0)} = \frac{(1, x^3 - 1)}{(1, 1)} = \frac{\int_0^1 (x^3 - 1)\, dx}{\int_0^1 dx} = \frac{-\frac{3}{4}}{1} = -\frac{3}{4},$$

$$\lambda_1 = \frac{(t_1, f)}{(t_1, t_1)} = \frac{\left(x - \frac{1}{2},\ x^3 - 1\right)}{\left(x - \frac{1}{2},\ x - \frac{1}{2}\right)} = \frac{\int_0^1 (x^3 - 1)\left(x - \frac{1}{2}\right) dx}{\int_0^1 \left(x - \frac{1}{2}\right)^2 dx} = \frac{9}{10},$$

$$\lambda_2 = \frac{(t_2, f)}{(t_2, t_2)} = \frac{\left(x^2 - x + \frac{1}{6},\ x^3 - 1\right)}{\left(x^2 - x + \frac{1}{6},\ x^2 - x + \frac{1}{6}\right)} =$$

$$= \frac{\int_0^1 (x^3 - 1)\left(x^2 - x + \frac{1}{6}\right) dx}{\int_0^1 \left(x^2 - x + \frac{1}{6}\right)^2 dx} = \frac{1}{1290}.$$

Cu aceste valori polinomul de cea mai bună aproximare este

$$T_2(x) = -\frac{3}{4} + \frac{9}{10}\left(x - \frac{1}{2}\right) + \frac{1}{1290}\left(x^2 - x + \frac{1}{6}\right), \qquad (3.128)$$

$$I = \|T_2(x) - f(x)\|_2^2 = \int_0^1 [T_2(x) - (x^3 - 1)]^2 \, dx =$$

$$= (f, f) - \sum_{k=0}^{2} \frac{(t_k, f)}{(t_k, t_k)} = (f, f) - \frac{(t_0, f)^2}{(t_0, t_0)} - \frac{(t_1, f)^2}{(t_1, t_1)} - \frac{(t_2, f)^2}{(t_2, t_2)} =$$

$$= \frac{9}{14} - \frac{9}{16} - \frac{27}{400} - \frac{1}{15\,480} \approx 0{,}0012.$$

În practică este mult mai bine să se calculeze $t_k(x)$ cu ajutorul relațiilor de recurență (3.108).

● *Cazul discret.* Pentru cazul discret cînd M este o mulțime finită de puncte din intervalul $[a, b]$ se procedează în felul următor. Fie

$$T_n(x) = \sum_{i=0}^{n} c_i x^i. \tag{3.129}$$

Se aleg coeficienții c_i astfel ca

$$I = \| T_n - f \|_2^2 \tag{3.130}$$

să fie minim. Aceasta conduce la ecuațiile normale

$$\sum_{j=0}^{n} a_{i,j} \, c_j = b_i, \quad i = 0, 1, \ldots, n, \tag{3.131}$$

unde

$$a_{i,j} = (x^i, x^j), \quad b_i = (x^i, f), \quad i, j = 0, 1, 2, \ldots, n. \tag{3.132}$$

Pentru $n = 2$ (3.131) devine

$$\left.\begin{aligned} a_{0,0}c_0 + a_{0,1}c_1 + a_{0,2}c_2 &= b_0 \\ a_{1,0}c_0 + a_{1,1}c_1 + a_{1,2}c_2 &= b_1 \\ a_{2,0}c_0 + a_{2,1}c_1 + a_{2,2}c_2 &= b_2 \end{aligned}\right\}. \tag{3.133}$$

Ţinînd seama de (3.132), sistemul (3.133) devine

$$\left.\begin{aligned}(1,1)c_0 + (1,x)c_1 + (1,x^2)c_2 &= b_0\\ (x,1)c_0 + (x,x)c_1 + (x,x^2)c_2 &= b_1\\ (x^2,1)c_0 + (x^2,x)c_1 + (x^2,x^2)c_2 &= b_2\end{aligned}\right\} \tag{3.134}$$

Matricea asociată sistemului (3.133), respectiv (3.134) este o matrice pozitiv definită şi nesingulară:

$$\mathbf{A} = \begin{bmatrix} a_{0,0} & a_{0,1} & a_{0,2}\\ a_{1,0} & a_{1,1} & a_{1,2}\\ a_{2,0} & a_{2,1} & a_{2,2}\end{bmatrix} = \begin{bmatrix} (1,1) & (1,x) & (1,x^2)\\ (x,1) & (x,x) & (x,x^2)\\ (x^2,1) & (x^2,x) & (x^2,x^2)\end{bmatrix} \tag{3.135}$$

Rezultă că sistemul (3.134) are soluţie unică. Mai mult matricea **A** este foarte bine condiţionată, fapt ce conduce la obţinerea unei soluţii precise. În cazul în care punctele sînt uniform distribuite în intervalul $[0,1]$, matricea **A** este apropiată de matricea lui Hilbert.

CALCUL NUMERIC MATRICEAL

4.1. Introducere

În foarte multe domenii ca : economie, fizică, geofizică, analiză şi sinteza reţelelor electrice, cristalografie, structuri inginereşti, mecanică, aeronautică etc., apar probleme liniare, care implică în rezolvarea lor calcule numerice matriceale. Datorită acestui fapt, în cadrul acestui capitol se vor prezenta o serie de aspecte teoretice privind matricele, elemente necesare analizei numerice şi utilizării calculatoarelor în rezolvarea problemelor liniare care implică calcule matriceale. Acest lucru apare în mod natural pentru rezolvarea unor probleme din algebra liniară ca : rezolvarea sistemelor de ecuaţii algebrice, calculul valorilor şi vectorilor proprii pentru o matrice dată. În plus, probleme de calcul matriceal apar şi la rezolvarea ecuaţiilor neliniare, ecuaţiilor diferenţiale ordinare şi a celor cu derivate parţiale, teoria arpoximării etc., în care metodele de rezolvare conduc în final la rezolvarea unor probleme din algebra liniară cu ajutorul calculului matriceal.

În esenţă algebra liniară este un studiu asupra transformărilor liniare ale spaţiilor vectoriale abstracte. Avînd în vedere natura fizică a elementelor ce constituie o matrice rezultată din aplicaţii practice, se vor analiza transfor-

mările liniare pe spațiile R^n și C^n, presupunînd că sînt cunoscute proprietățile numerelor reale și complexe. Urmărindu-se îndeosebi caracterul aplicativ și de calcul propriu-zis, o serie de aspecte teoretice vor fi doar enunțate, indicîndu-se bibliografie adecvată privind demonstrațiile aferente.

Pentru început se va considera C^n spațiu vectorial n dimensional peste corpul numerelor complexe C, de vectori coloană $\mathbf{x}$, unde vectorul $\mathbf{x}$, transpusul său $\mathbf{x}^T$ și conjugatul transpus $\mathbf{x}^H$ se prezintă astfel :

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_k \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}, \quad \mathbf{x}^T = [x_1,\ x_2,\ \ldots,\ x_k,\ \ldots\ x_n],$$

$$\mathbf{x}^H = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \ldots, \bar{x}_n]$$

cu $x_i \in C$, $i = 1, 2, \ldots, n$. Prin R^n se înțelege spațiul vectorial n dimensional peste corpul numerelor reale R, format din vectorii coloană $\mathbf{x}$ cu componentele $x_1, x_2, \ldots, x_k, \ldots, x_n$ pentru $\mathbf{x} \in R^n$, $\mathbf{x}^T$ reprezentînd vectorul transpus care este un vector linie $(x_1,\ x_2, \ldots, x_k, \ldots, x_n)$ cu $x_i \in R$ pentru $i = 1, 2, \ldots, n$.

Pentru a evidenția modul în care aceste probleme ale algebrei liniare apar din sistemele fizice, în continuare se vor prezenta cîteva exemple.

Exemple. 1. Se consideră rețeaua din fig. 4.1 care este alimentată cu doi curenți $I = 3\text{A}$ și $I' = 5\text{A}$. Dacă se scriu legile lui Kirchoff, rezultă următorul sistem de ecuații pentru determinarea celor zece curenți $I_1,\ I_2, \ldots, I_{10}$ din cele zece laturi ale rețelei considerate :

$$I_1 + I_2 + I_3 = I, \quad I_8 + I_9 + I_{10} = I', \quad -I_1 + I_4 - I_6 = 0,$$

$$-I_3 + I_5 - I_9 = 0, \quad I_6 + I_7 - I_{10} = 0, \quad -R_7 I_7 + R_8 I_8 - R_{10} I_{10} = 0,$$

$$-R_5 I_5 + R_8 I_8 - R_9 I_9 = 0, \qquad R_2 I_2 - R_3 I_3 - R_5 I_5 = 0,$$

$$-R_1 I_1 + R_2 I_2 - R_4 I_4 = 0, \qquad -R_4 I_4 - R_6 I_6 + R_7 I_7 = 0.$$

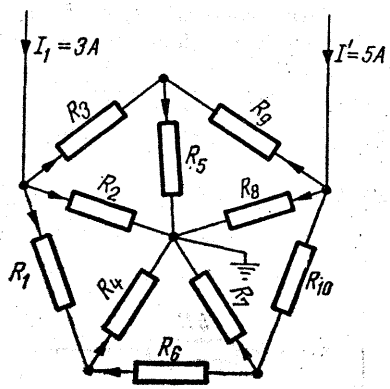


Fig. 4.1.

Acest sistem se poate scrie sub formă matriceală astfel:

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -R_7 & R_8 & 0 & -R_{10} \\ 0 & 0 & 0 & 0 & -R_5 & 0 & 0 & R_8 & -R_9 & 0 \\ 0 & R_2 & -R_3 & 0 & -R_5 & 0 & 0 & 0 & 0 & 0 \\ -R_1 & R_2 & 0 & -R_4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -R_4 & 0 & -R_6 & R_7 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \\ I_8 \\ I_9 \\ I_{10} \end{pmatrix} = \begin{pmatrix} I \\ I' \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

2. Sistemul fizic este format dintr-o rețea electrică (v. fig. 4.2), are un număr finit de rezistențe și se numește sistem cu parametrii distribuiți. Pentru modelarea matematică se aplică legea lui Ohm și faptul că suma tensiunilor în fiecare ochi de circuit este egală cu zero. Astfel se

obțin cinci ecuații cu cinci necunoscute I_1 I_2, I_3, I_4, I_5, care reprezintă curenții de contur. Sistemul de ecuații are următoarea formă :

$$E - 2RI_1 - 8R(I_1 - I_4) - 4R(I_1 - I_2) = 0,$$
$$4E - 4R(I_2 - I_1) - 2R(I_2 - I_3) - 6RI_2 = 0,$$
$$2E - 2R(I_3 - I_2) - 10R(I_3 - I_4) - 2R(I_3 - I_5) = 0,$$
$$3E - 8R(I_4 - I_1) - 10R(I_4 - I_3) - 4R(I_4 - I_5) = 0,$$
$$- 5RI_5 - 2R(I_5 - I_3) - 4R(I_5 - I_4) = 0.$$

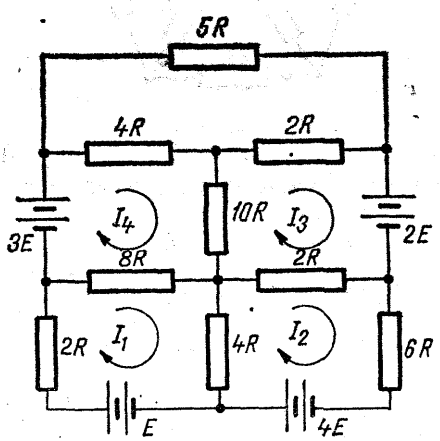


Fig. 4.2.

După ordonare și scriere sub formă matriceală sistemul devine

$$\begin{pmatrix} 14R & -4R & 0 & 8R & 0 \\ -4R & 12R & -2R & 0 & 0 \\ 0 & -2R & 14R & -10R & -2R \\ -8R & 0 & -10R & 22R & -4R \\ 0 & 0 & -2R & -4R & 11R \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{pmatrix} = \begin{pmatrix} E \\ 4E \\ 2E \\ 3E \\ 0 \end{pmatrix}$$

3. Acest exemplu urmărește să evidențieze rolul calculului matriceal la studiul și analiza structurilor. Prin structură se înțelege un sistem care

are funcţia de a transmite sarcinile [107]. Există diverse metode pentru măsurarea forţelor aplicate unei structuri sau a deplasărilor pe care le suferă structura în m puncte situate pe anumite direcţii. În acest sens apare necesitatea introducerii unui sistem de coordonate pentru identificarea măsurărilor. Sistemul de coordonate folosit pentru măsurarea forţelor şi deplasărilor aplicate structurii pot servi şi la măsurarea vitezelor şi acceleraţiilor pentru anumite puncte ale structurii. Măsurările efectuate asupra unei structuri pot fi puse sub forma a doi vectori: vectorul forţelor şi vectorul deplasărilor:

$$\mathbf{F}^T = [F_1, F_2, \ldots, F_n], \qquad \mathbf{D}^T = [d_1, d_2, \ldots, d_n].$$

Dacă se consideră relaţia dintre vectorul forţelor **F** şi vectorul deplasărilor **D** pentru o structură considerată, se disting următoarele moduri de comportare pentru structură:

— elastic, dacă aceasta revine la configuraţia iniţială după aplicarea unei sarcini şi înlăturarea sarcinii respective;

— neelastică, dacă structura nu revine la configuraţia iniţială după înlăturarea sarcinii care a fost aplicată;

— liniară, dacă graficul $\mathbf{D} = f(\mathbf{F})$ conduce la o curbă liniară;

— neliniară dacă graficul $\mathbf{D} = f(\mathbf{F})$ conduce la o curbă neliniară.

În cazul aplicării mai multor forţe unei structuri, există posibilitatea măsurării deplasărilor şi deformaţiilor interne, prin metoda superpoziţiei, efectuîndu-se măsurările pentru fiecare forţă în parte, după care se face sumarea algebrică a rezultatelor.

Caracteristicile anumitor structuri sînt rigiditatea şi flexibilitatea care sînt materializate prin doi coeficienţi. Aceşti coeficienţi şi sistemul de coordonate introdus pot să ajute la caracterizarea şi analiza unei structuri.

Se consideră o structură în două coordonate, formată din două resoarte cu constantele de rigiditate α şi β şi coordonatele 1 şi 2 (fig. 4.3). Se pune problema caracterizării acestei structuri cînd se aplică forţele F_1 şi F_2 cărora le corespund deplasările d_1 şi d_2 în punctele de coordonate 1 şi 2. Pentru aceasta se aplică metoda superpoziţiei. Folosind metoda superpo-

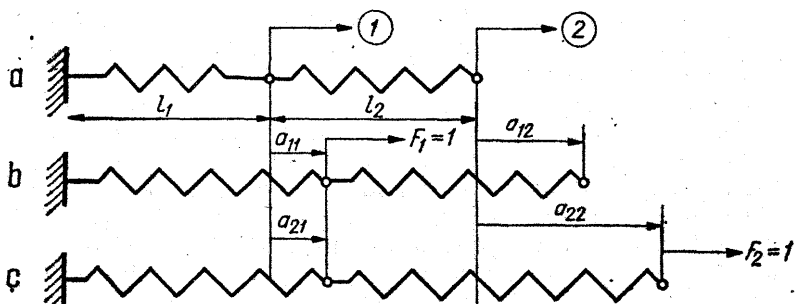


Fig. 4.3.

ziției, se aplică la început forța $F_1 = 1$, celelalte forțe F_i fiind egale cu zero; atunci se generează a_{i1} ($i = 1,2$), deplasările rezultate din aplicarea lui $F_1 = 1$ sînt $a_{i1}F_1$ ($i = 1,2$). Se observă în fig. 4.3,b generarea coeficienților a_{11} și a_{21}, care reprezintă:

a_{11} — deplasarea coordonatei 1 datorită lui F_1;

a_{21} — deplasarea coordonatei 2 datorită lui F_1.

În continuare se aplică structurii forța F_2 (fig. 4.3, b) cu $F_2 = 1$, obținîndu-se coeficienții de deplasare a_{12} și a_{22}.

Pentru determinarea deplasărilor d_1 și d_2 cauzate de forțele F_1 și F_2 prin acțiunea simultană, se adună deplasările datorate lui F_1 cu deplasările datorate lui F_2, obținîndu-se

$$d_1 = a_{11}F_1 + a_{12}F_2, \qquad d_2 = a_{21}F_1 + a_{22}F_2$$

sau

$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \end{bmatrix}, \qquad \mathbf{D} = \mathbf{A}\,\mathbf{F},$$

unde matricea coeficienților este determinată astfel:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{\alpha} & \frac{1}{\alpha} \\ \frac{1}{\alpha} & \frac{1}{\alpha+\beta} \end{bmatrix}.$$

Matricea A se numește matricea de elasticitate a structurii date în fig. 4.3.

4.2. Spațiile vectoriale R^n și C^n

Avînd în vedere natura numerică a aplicațiilor fizice considerate se va acorda o atenție deosebită spațiilor vectoriale R^n și C^n.

Pentru orice n întreg și pozitiv R^n este spațiul vectorial real n-dimensional al vectorilor coloană $\mathbf{x}$ cu componentele $x_1, x_2, \ldots x_i, \ldots, x_n$, unde $x_i \in R$; R^n se mai numește și spațiul n-dimensional al coordonatelor reale. De asemenea, pentru orice n întreg și pozitiv, C^n reprezintă spațiul vectorial complex n-dimensional al vectorilor coloană $\mathbf{x}$ cu componentele $x_1, x_2, \ldots, x_n$, unde $x_i \in C$; C^n se mai numește și spațiul n-dimensional al coordonatelor complexe. Dacă $\mathbf{x}$ și $\mathbf{y}$ sînt vectori coloană din R^n sau C^n, atunci ecuația $\mathbf{x} = \mathbf{y}$ este echivalentă cu sistemul de

ecuații $x_i = y_i$, pentru $i = 1, 2, \ldots, n$, care reprezintă egalitatea componentelor celor doi vectori, de unde se vede că doi vectori sînt egali dacă și numai dacă componentele lor sînt egale. Pentru a defini un nou vector, este suficient a specifica cum sînt formate componentele sale.

Să considerăm spațiile vectoriale R^n și C^n pe care sînt definite două operații: adunarea dintre vectori și înmulțirea cu un scalar pentru vectori coloană din R^n și C^n. Fie $\mathbf{x}, \mathbf{y}, \mathbf{z}$ vectori coloană din R^n:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_k \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_k \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ z_k \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{bmatrix}.$$

Atunci adunarea între doi vectori din R^n este definită astfel:

$$\mathbf{x} + \mathbf{y} = \mathbf{z} \text{ dacă și numai dacă } x_i + y_i = z_i \qquad (4.1)$$

pentru $i = 1, 2, \ldots, n$; $x_i, y_i, z_i \in R$.

Dacă $\mathbf{x}, \mathbf{y}, \mathbf{z} \in C^n$, adunarea vectorilor în C^n este definită astfel:

$$\mathbf{x} + \mathbf{y} = \mathbf{z} \text{ dacă și numai dacă } x_i + y_i = z_i \qquad (4.2)$$

pentru $i = 1, 2, \ldots, n$ și $x_i, y_i, z_i \in C$.

Primele relații din (4.1) și (4.2) reprezintă adunarea vectorilor, a doua relație din (4.1) reprezintă adunarea numerelor reale, iar a doua relație din (4.2) reprezintă adunarea de numere complexe.

Dacă $k_1 \in R^n$ și $\mathbf{x} \in R^n$ iar $k_2 \in C$ și $\mathbf{y} \in C^n$, atunci

$$k_1\mathbf{x} = \begin{bmatrix} k_1x_1 \\ k_1x_2 \\ \cdot \\ \cdot \\ \cdot \\ k_1x_n \end{bmatrix} \quad \text{și} \quad k_2\mathbf{y} = \begin{bmatrix} k_2y_1 \\ k_2y_2 \\ \cdot \\ \cdot \\ \cdot \\ k_ny_n \end{bmatrix}, \tag{4.3}$$

unde $x_i \in R$ și $y_i \in C$ pentru $i = 1, 2, \ldots, n$.

Partea dreaptă a primei relații din (4.3) reprezintă produs de numere reale (k_1, $x_i \in R$), iar partea dreaptă din cea de-a doua relație din (4.3) reprezintă produs de numere complexe (k_2, $y_i \in C$). Considerînd pe $n = 2$, adunarea a doi vectori și înmulțirea cu un scalar se pot reprezenta grafic ca în fig. 4.4.

Considerînd spațiile vectoriale R^n și C^n peste cîmpurile R, respectiv C, unde adunarea vectorială și înmulțirea cu un scalar au fost definite prin (4.1) — (4.3), avem :

$$\begin{aligned}
&1.\ \mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}; \\
&2.\ (\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z}); \\
&3.\ \mathbf{x} + \mathbf{0} = \mathbf{x}; \\
&4.\ \mathbf{x} + (-\mathbf{x}) = \mathbf{0}; \\
&5.\ (k_1k_2)\mathbf{x} = k_1(k_2\mathbf{x}); \\
&6.\ (k_1 + k_2)\mathbf{x} = k_1\mathbf{x} + k_2\mathbf{x}; \\
&7.\ k_1(\mathbf{x} + \mathbf{y}) = k_1\mathbf{x} + k_1\mathbf{y}; \\
&8.\ 1 \cdot \mathbf{x} = \mathbf{x},
\end{aligned} \tag{4.4}$$

unde $\mathbf{x}, \mathbf{y} \in R^n$, $k_1, k_2 \in R$, iar $\mathbf{0}$ și $-\mathbf{x}$ sînt doi vectori coloană care au forma

$$\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \quad -\mathbf{x} = \begin{bmatrix} -x_1 \\ -x_2 \\ \cdot \\ \cdot \\ \cdot \\ -x_n \end{bmatrix}, \tag{4.5}$$

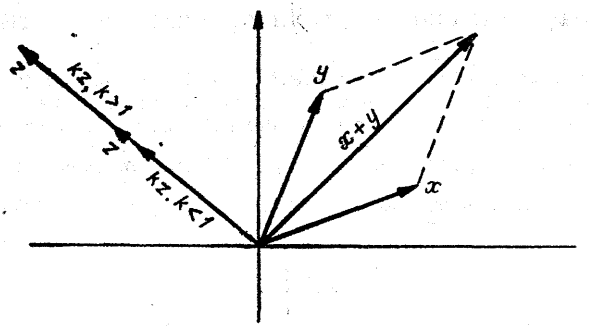


Fig. 4.4.

sau $\mathbf{x}, \mathbf{y} \in C^n$, $k_1, k_2 \in C$, iar $\mathbf{0}$ şi $-\mathbf{x}$ sînt doi vectori coloană din C^n ale căror componente sînt numere complexe, din C scrise sub forma (a, b):

$$\mathbf{0} = \begin{bmatrix} (0,0) \\ (0,0) \\ \cdot \\ \cdot \\ \cdot \\ (0,0) \end{bmatrix}, \quad -\mathbf{x} = \begin{bmatrix} (-a_1, -b_1) \\ (-a_2, -b_2) \\ \cdot \\ \cdot \\ \cdot \\ (-a_n, -b_n) \end{bmatrix} \tag{4.6}$$

cu $x_k = a_k + ib_k$, $k = 1, \ldots, n$, $a_k, b_k \in R$, $x_k \in C$.

4.2.1. *Spaţiul real* R^n

Noţiunea de vector real n-dimensional din spaţiul R^n este o generalizare naturală a reprezentării punctelor din spaţiul R^n prin coordonate carteziene. În acest caz R^n poate fi definit ca

$$R^n = \{ n\text{-uple } \mathbf{x} : \mathbf{x}_i \in R \}.$$

S-a arătat că R^n este un spaţiu vectorial, deci n-uplurile pot fi numite vectori.

Justificarea acestei terminologii este faptul că există o corespondenţă biunivocă între vectorii $\mathbf{x} \in R^n$ şi punctele

din spațiul euclidian n dimensional cu coordonatele $x_1, x_2, \ldots, x_n$.

Pentru exemplificare se consideră spațiul euclidian tridimensional, unde un punct A are coordonatele x_1, x_2, x_3 (fig. 4.5). Pentru fiecare astfel de punct există un segment de dreaptă unic din originea O (0, 0, 0) la punctul A, notat prin $\overrightarrow{OA}$ și reciproc. Datorită corespondenței între A și $\overrightarrow{OA}$, precum și corespondenței între $A(x_1, x_2, x_3)$ și vectorul tridimensional

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

acesta poate fi identificat cu A și $\overrightarrow{OA}$.

Prin introducerea operației de adunare și înmulțire cu un scalar (fig. 4.4) a segmentelor de dreaptă, se poate arăta că mulțimea segmentelor de dreaptă $\overrightarrow{OA}$ constituie un spațiu vectorial abstract.

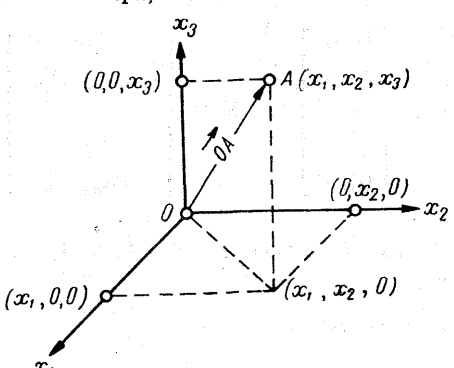


Fig. 4.5.

Prin analogie se poate spune același lucru pentru R^n, punctul $A(x_1, x_2, \ldots, x_n)$ și segmentul $\overrightarrow{OA}$ din originea O (0, 0, . . . ,0) la punctul P fiind identificate cu vectorul n dimensional

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

datorită izomorfismului care există între cele trei mulțimi : {punctele A}, segmentele de dreaptă $\{\overrightarrow{OA}\}$ și {vectorii $\mathbf{x}$}.

4.2.2. *Combinații liniare*

Fie $\mathbf{x}^{(i)} \in R^n$; în acest caz vectorul $\mathbf{x}^{(i)}$ se poate scrie sub forma

$$\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \cdot \\ \cdot \\ \cdot \\ x_n^{(i)} \end{bmatrix}.$$

Fie $c_1, c_2, \ldots, c_p \in R$ și $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)} \in R^n$

$$\mathbf{x} = c_1\mathbf{x}^{(1)} + c_2\mathbf{x}^{(2)} + \ldots + c_p\mathbf{x}^{(p)}; \tag{4.7}$$

vectorul $\mathbf{x}$ este o combinație liniară a vectorilor $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)}$, iar scalarii $c_1, c_2, \ldots, c_p$ sînt numiți coeficienții combinației liniare. Dacă $c_i = 0$ pentru $i = 1, 2, \ldots, p$, combinația liniară (4.7) se numește banală, altfel este nebanală. Dacă se consideră ecuația

$$\begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix} + \ldots + x_n \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix}, \tag{4.8}$$

aceasta se mai poate scrie

$$\mathbf{x} = x_1\mathbf{e}^{(1)} + x_2\mathbf{e}^{(2)} + \ldots + x_n\mathbf{e}^{(n)}, \tag{4.9}$$

unde $\mathbf{e}^{(i)}$ sînt vectori din R^n ale căror componente sînt $e_{(j)}^{i} = \delta_{ij}$, $j = 1, 2, \ldots, n$. Vectorii $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}\}$ se numesc vectorii unitate din R^n.

Din (4.9) se vede că $\mathbf{x} \in R^n$ este o combinație liniară a vectorilor unitate $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}\}$ din R^n.

De asemenea se observă că dacă $\mathbf{x} = \mathbf{0}$ (este vectorul nul), combinația liniară (4.9) este banală, altfel este nebanală.

● Vectorii distincți $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)} \in R^n$ sînt liniar dependenți dacă și numai dacă există o combinație liniară nebanală între ei egală cu vectorul $\mathbf{0}$, adică dacă și numai dacă

$$c_1\mathbf{x}^{(1)} + c_2\,\mathbf{x}^{(2)} + \ldots + c_p\,\mathbf{x}^{(p)} = \mathbf{0} \qquad (4.10)$$

unde $c_i \neq 0$ pentru cel puțin o valoare a lui $i = 1, 2, \ldots, p$.

● Dacă combinația liniară (4.10) este banală, atunci vectorii $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)}$ sînt liniar independenți. Se observă că orice combinație liniară care conține vectorul nul $\mathbf{0}$ este liniar dependentă pentru că $c\mathbf{0} = \mathbf{0}$ pentru orice $c \neq \mathbf{0}$ și combinația $c\mathbf{0} = \mathbf{0}$ este nebanală. Datorită acestui fapt se poate afirma că orice sistem de vectori din R^n care conține vectorul nul $\mathbf{0}$ este un sistem de vectori liniar dependenți. Dacă se consideră sistemul de vectori $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)}\}$, atunci

$$c\mathbf{0} \neq 0 \cdot \sum_{i=1}^{p} \mathbf{x}^{(i)} = \mathbf{0}, \qquad (4.11)$$

dacă $c \neq 0$; combinația liniară (4.11) este banală.

● Vectorii nenuli $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)} \in R^n$ sînt liniar dependenți dacă și numai dacă unul din vectorii $\mathbf{x}^{(k)}$, $k = 1, 2, \ldots, p$, este o combinație liniară a celorlalți. Această afirmație scoate în evidență că în cazul unei combinații liniare dependente, există cel puțin doi coeficienți diferiți de zero, din

$$\sum_{i=1}^{p} c_i\,\mathbf{x}^{(i)} = 0 \qquad (4.12)$$

și dacă $c_k \neq 0$ este unul din acești coeficienți, (4.12) se

poate scrie sub forma

$$\mathbf{x}^{(k)} = \left[\frac{c_1}{c_k}\right]\mathbf{x}^{(1)} + \ldots + \left[\frac{c_{k-1}}{c_k}\right]\mathbf{x}^{(k-1)} +$$

$$+ \left[\frac{c_{k-1}}{c_k}\right]\mathbf{x}^{(k+1)} + \ldots + \left[\frac{c_p}{c_k}\right]\mathbf{x}^{(p)}.$$

● Fie $\mathbf{x} \in R^n$; acest vector este liniar dependent față de sistemul de vectori $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)} \in R^n$ dacă și numai dacă poate fi exprimat ca o combinație liniară a sistemului de vectori $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)}$.

● Sistemul de vectori $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)} \in R^n$ este liniar independent, deoarece se poate scrie relația

$$\mathbf{0} = \sum_{i=1}^{n} c_i\, \mathbf{e}^{(i)}$$

sau dezvoltat

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \ldots + c_n \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \qquad (4.13)$$

După efectuarea operațiilor de înmulțire cu o constantă și o adunare a vectorilor rezultă:

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix},$$

această ultimă relație avînd loc dacă și numai dacă $c_i = 0$, pentru $i = 1, 2, \ldots, n$.

● Sistemul de vectori unitate $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}\} \in R^n$ formează o bază pentru R^n, deoarece ei sînt liniar independenți.

Un spaţiu vectorial este finit dimensional dacă el are o bază finită, adică are o bază formată dintr-un număr finit de vectori. Din relaţiile (4.13) rezultă că R^n este un spaţiu vectorial finit dimensional. Deci, într-un spaţiu finit dimensional, toate bazele conţin acelaşi număr de vectori. Deoarece sistemul de vectori $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}\}$ formează o bază în R^n, rezultă că orice bază în R^n conţine exact n vectori (deoarece toate bazele au dimensiunea spaţiului). Aceste elemente conduc la afirmaţia că dimensiunea unui spaţiu finit dimensional este egală cu numărul vectorilor din bazele sale, de unde rezultă că $\dim R^n = n$.

● Dacă sistemul de vectori $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(n)}\} \in R^n$ constituie o bază pentru R^n, atunci orice vector $\mathbf{y} \in R^n$ se exprimă în mod unic ca o combinaţie liniară a sistemului de vectori $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(n)}\}$.

Demonstraţia rezultă imediat. Se consideră că $\mathbf{y}$ are două reprezentări în funcţie de baza considerată:

$$\mathbf{y} = \sum_{i=1}^{n} a_i \mathbf{y}^{(i)} \quad \text{şi} \quad \mathbf{y} = \sum_{i=1}^{n} b_i \mathbf{y}^{(i)}.$$

Atunci

$$\mathbf{0} = \sum_{i=1}^{n} a_i \mathbf{y}^{(i)} - \sum_{i=1}^{n} b_i \mathbf{y}^{(i)} = \sum_{i=1}^{n} (a_i - b_i)\mathbf{y}^{(i)}.$$

Deoarece vectorii $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(n)}$ sînt liniar independenţi, rezultă $a_i = b_i$, $i = 1, 2, \ldots, n$. Deci oricare ar fi vectorul $\mathbf{x} \in R^n$, acesta are o exprimare unică în funcţie de vectorii bazei spaţiului R^n.

4.2.3. *Legătura între coordonate şi bazele ordonate*

S-a arătat anterior cum componentele lui $\mathbf{x}$ sînt coeficienţii combinaţiei liniare prin care $\mathbf{x}$ se exprimă în funcţie de vectorii bazei $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}\}$.

● O bază ordonată pentru un spaţiu vectorial este o mulţime ordonată de vectori (adică în ordinea axelor de coordonate) care formează o bază pentru spaţiul vectorial.

Presupunem că se consideră baza ordonată $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$ pentru R^n. Se ştie că orice vector $\mathbf{x} \in R^n$ poate fi exprimat ca o combinaţie liniară unică, faţa de vectorii din baza considerată, astfel:

$$\mathbf{x} = \sum_{i=1}^{n} a_i \mathbf{v}^{(i)}. \tag{4.14}$$

Se observă că în acest caz se poate asocia vectorului $\mathbf{x} \in R^n$ un n-uplu ordonat unic, de forma

$$\mathbf{x} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, \tag{4.15}$$

unde $a_1, a_2, \ldots, a_n$ sînt coordonatele lui $\mathbf{x}$ cu privire la baza ordonată considerată, sau componentele lui $\mathbf{x}$ cu privire la baza ordonată.

De obicei, în cazul cînd $\mathbf{x} \in R^n$, se scrie

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

fără a se menţiona că $x_1, x_2, \ldots, x_n$ sînt componentele lui $\mathbf{x} \in R^n$ cu privire la baza ordonată $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}\} \in R^n$. Datorită faptului că R^n conţine o mulţime de baze ordonate, altele decît baza unitară (sau baza canonică), trebuie specificat, cînd se defineşte un vector, la ce bază se referă. În acest sens se consideră că baza $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}\}$ este bază naturală a spaţiului R^n, componentele lui $\mathbf{x} \in R^n$, din (4.8), se numesc componentele naturale, iar componentele lui $\mathbf{x}$ date prin (4.15) reprezintă componentele vectorului $\mathbf{x}$ cu privire la baza $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$.

● Dacă V este o bază ordonată de vectori $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$ pentru R^n, atunci există o funcţie coordonată

$\mathscr{F}_{\mathbf{V}} : R^n \to R^n$ astfel că $\mathbf{x} \in R^n$ implică

$$\mathbf{x} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

dacă şi numai dacă

$$\mathbf{x} = a_1\mathbf{v}^{(1)} + a_2\,\mathbf{v}^{(2)} + \ldots + a_n\mathbf{v}^{(n)}.$$

Exemplu. Fie $\mathbf{x} \in R^n$ un vector de componente naturale

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

Atunci

$$\mathbf{x} = 1 \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 2 \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 3 \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = 1\mathbf{e}^{(1)} + 2\mathbf{e}^{(2)} + 3\mathbf{e}^{(3)},$$

deci 1, 2, 3 sînt componentele lui $\mathbf{x}$ cu privire la baza naturală $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \mathbf{e}^{(3)}\} \in R^3$.

De asemenea vectorul $\mathbf{x}$ se poate scrie sub forma

$$\mathbf{x} = 1 \cdot \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix} + (-2) \cdot \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} + 2 \cdot \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = 1\mathbf{v}^{(1)} + (-2)\mathbf{v}^{(2)} + 2\mathbf{v}^{(3)},$$

unde $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}\} \in R^3$ constituie o bază pentru R^3. În acest caz componentele 1, -2, 2 reprezintă coordonatele vectorului

$$\mathbf{x} \to \begin{bmatrix} 1 \\ -2 \\ 2 \end{bmatrix}$$

cu privire la baza ordonată $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}\}$ formată din vectorii

$$\left\{\begin{bmatrix}1\\2\\5\end{bmatrix}\ \begin{bmatrix}1\\1\\2\end{bmatrix}\ \begin{bmatrix}2\\1\\1\end{bmatrix}\right\}.$$

4.3. Transformări liniare

Fie aplicația $\mathscr{A}: R^n \to R^n$, care asociază unui vector $\mathbf{x} \in R^n$ un vector unic $\mathbf{y}$ din același spațiu (sau alt spațiu). În acest paragraf se vor prezenta aplicațiile sau transformările liniare.

Considerăm două spații vectoriale U și V peste același cîmp E și fie $\mathscr{A}$ o aplicație a lui U în $V (\mathscr{A}: U \to V)$ astfel că, oricare ar fi $\mathbf{u} \in U$, există un vector unic $\mathbf{v} \in V$, unde $\mathscr{A}\mathbf{u} = \mathbf{v}$.

Dacă pentru $\mathbf{u}, \mathbf{u}^{(1)}, \mathbf{u}^{(2)} \in U$ și $k \in E$ aplicația $\mathscr{A}$ satisface relațiile

$$\left.\begin{aligned}\mathscr{A}(\mathbf{u}^{(1)} + \mathbf{u}^{(2)}) &= \mathscr{A}\,\mathbf{u}^{(1)} + \mathscr{A}\,\mathbf{u}^{(2)}\\ \mathscr{A}(k\,\mathbf{u}) &= k\,\mathscr{A}\,\mathbf{u}\end{aligned}\right\} \tag{4.16}$$

atunci $\mathscr{A}$ este o transformare liniară sau o aplicație liniară a lui U în V.

Denumirea de aplicație liniară se justifică prin faptul că combinațiile liniare se conservă prin aplicațiile liniare, adică dacă se consideră aplicația liniară $\mathscr{A}: U \to V$, atunci pentru $a_1, a_2, \ldots, a_p \in E$ și $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(p)} \in U$ avem

$$\mathscr{A}(a_1\,\mathbf{u}^{(1)} + a_2\,\mathbf{u}^{(2)} + \ldots + a_p\,\mathbf{u}^{(p)}) =$$

$$= a_1\mathscr{A}\,\mathbf{u}^{(1)} + a_2\,\mathscr{A}\mathbf{u}^{(2)} + \ldots + a_p\mathscr{A}\,\mathbf{u}^{(p)}. \tag{4.17}$$

Pentru cele ce vor urma se introduc notațiile:

• $\mathscr{L}(U,V)$ mulțimea aplicațiilor liniare a spațiului vectorial U în spațiul vectorial V, adică $\mathscr{A} \in \mathscr{L}(U, V)$;

● vectorul $\mathbf{v}$ este imaginea lui $\mathbf{u}$ prin $\mathscr{A}$, adică $\mathscr{A}\mathbf{u}=\mathbf{v}$;

● în cazul în care $U = V$, aplicația $\mathscr{A}$ este liniară pe U.

O aplicație liniară este biunivocă dacă pentru două elemente distincte $\mathbf{u}^{(1)}$ și $\mathbf{u}^{(2)}$ din U corespund două elemente $\mathbf{v}^{(1)}$ și $\mathbf{v}^{(2)}$ distincte din V. Se poate arăta că prin aplicațiile liniare combinațiile liniare se conservă, dar nu și dimensiunea lor [128].

Dacă o aplicație liniară este biunivocă, atunci ea este nesingulară. Dacă $\mathscr{A}$ este nesingulară, atunci pentru $\mathbf{u}^{(1)}$, $\mathbf{u}^{(2)} \in U$ cu $\mathbf{u}^{(1)} \neq \mathbf{u}^{(2)}$ și imaginile $\mathbf{v}^{(1)}$, $\mathbf{v}^{(2)} \in V$ sînt distincte, cu $\mathscr{A}\,\mathbf{u}^{(1)}=\mathbf{v}^{(1)}$, $\mathscr{A}\,\mathbf{u}^{(2)} = \mathbf{v}^{(2)}$. Atunci pentru orice vector $\mathbf{v} \in V$ există un vector unic $\mathbf{u} \in U$ pentru care $\mathbf{u}$ este imaginea lui $\mathbf{v}$ prin $\mathscr{A}^{-1}$ $\mathbf{v} = \mathbf{u}$.

Aplicația inversă $\mathscr{A}^{-1} : V \to U$ este liniară, adică [128, 10]

$$\mathscr{A}^{-1} \in \mathscr{L}\ (V,\ U).$$

Din cele prezentate se vede că au loc relațiile

$$\mathscr{A}^{-1}\,(\mathscr{A}\mathbf{u}) = \mathbf{u},\ \mathscr{A}(\mathscr{A}^{-1}\mathbf{v}) = \mathbf{v},\ \mathscr{A}\,(\mathscr{B}\,\mathbf{u}) = (\mathscr{A}\mathscr{B}\mathbf{u}), \qquad (4.18)$$

de unde rezultă că aplicația $\mathscr{A}^{-1}\mathscr{A}$ este aplicația identică pe U iar aplicația $\mathscr{A}\,\mathscr{A}^{-1}$ este aplicația identică pe V.

Ultima relație din (4.18) introduce conceptul de produs a două aplicații, care se execută de la dreapta la stînga. De asemenea are sens și suma a două aplicații [10]: se consideră aplicațiile $\mathscr{A}, \mathscr{M}, \mathscr{N} \in \mathscr{L}(U,V)$ și se presupune că $\mathscr{A}\mathbf{u} = \mathbf{v}^{(1)}$, $\mathscr{M}\,\mathbf{u} = \mathbf{v}^{(2)}$ și $\mathscr{N}\mathbf{u} = \mathbf{v}^{(3)}$. Atunci se poate afirma că $\mathscr{N} = \mathscr{A} + \mathscr{M}$ dacă și numai dacă $\mathbf{v}^{(3)} = \mathbf{v}^{(1)} + \mathbf{v}^{(2)}$, oricare ar fi $\mathbf{u} \in U$. Are sens și produsul $k\mathscr{A}$, unde $k \in E$. Fie $\mathbf{u} \in U$ și $\mathscr{A}\mathbf{u} = \mathbf{v}^{(1)}$ iar $\mathscr{B}\mathbf{u} = \mathbf{v}^{(2)}$. Atunci se poate afirma că $\mathscr{B} = k\mathscr{A}$ dacă și numai dacă $\mathbf{v}^{(2)} = k\mathbf{v}^{(1)}$, oricare ar fi $\mathbf{u} \in U$ și $\mathscr{A}\ \mathscr{B} \in \mathscr{L}\ (U,V)$. Din cele prezentate se vede că se poate face următoarea afirmație:

● Mulțimea aplicațiilor $\mathscr{L}(U,V)$ este un spațiu vectorial prin operația de adunare și înmulțire cu un scalar definite astfel:

1) $\mathscr{A},\mathscr{B} \in \mathscr{L}\ (U,V) \Rightarrow (\mathscr{A} + \mathscr{B})\mathbf{u} = \mathscr{A}\mathbf{u} + \mathscr{B}\mathbf{u}$;

2) $k \in E$ şi $\mathscr{A} \in \mathscr{L}(U,V) \Rightarrow (k\mathscr{A})\mathbf{u} = k\mathscr{A}\mathbf{u}$ (4.19)

pentru orice $\mathbf{u} \in U$ [10, 119].

Dacă aplicaţia $\mathscr{A} \in \mathscr{L}(U, V)$, atunci următoarele afirmaţii sînt echivalente:

— $\mathscr{A}$ este nesingulară;

— $\mathscr{A}$ are o inversă, $\mathscr{A}^{-1}$;

— rang $\mathscr{A}$ = dim U = dim V;

— dacă $\{\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(n)}\}$ este o bază vectorială pentru U, atunci $\{\mathscr{A}\mathbf{u}^{(1)}, \mathscr{A}\mathbf{u}^{(2)}, \ldots, \mathscr{A}\mathbf{u}^{(n)}\}$ este o bază vectorială pentru V.

4.3.1. *Coordonate şi matrice*

Fie $\mathscr{A} \in \mathscr{L}(U,V)$, unde U este un spaţiu vectorial p dimensional şi V un spaţiu vectorial, q dimensional. Deci U şi V au bazele vectoriale ordonate $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)}\}$, respectiv $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(q)}\}$.

În continuare se va arăta cum se poate utiliza matricea A în cadrul unui algoritm pentru calculul coordonatelor vectorului $\mathbf{v} \in V$, dacă $\mathbf{v}$ este imaginea lui $\mathbf{u} \in U$ prin $\mathscr{A}$. Dacă $\mathbf{u} \in U$ este un vector arbitrar, atunci acesta se poate scrie ca o combinaţie liniară de vectorii bazei ordonate a spaţiului vectorial U, adică

$$\mathbf{u} = u_1\mathbf{x}^{(1)} + u_2\mathbf{x}^{(2)} + \ldots + u_p\mathbf{x}^{(p)}. \tag{4.20}$$

Astfel se poate asocia vectorului $\mathbf{u}$ coordonatele sale relativ la baza $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)}\}$:

$$\mathbf{u} \to \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{bmatrix}. \tag{4.21}$$

Dacă $\mathscr{A}\mathbf{u} = \mathbf{v} \in V$ şi $\mathbf{v}$ se exprimă în funcţie de baza ordonată din V, atunci din (4.20) rezultă

$$\mathbf{v} = \mathscr{A}u = \mathscr{A}(u_1\mathbf{x}^{(1)} + u_2\mathbf{x}^{(2)} + \ldots + u_p\mathbf{x}^{(p)}) =$$

$$= u_1 \mathcal{A}\,\mathbf{x}^{(1)} + u_2 \mathcal{A}\mathbf{x}^{(2)} + \ldots + u_p\, \mathcal{A}\,\mathbf{x}^{(p)} = \qquad (4.22)$$

$$= u_1\mathbf{z}^{(1)} + u_2\mathbf{z}^{(2)} + \ldots + u_p\,\mathbf{z}^{(p)}.$$

Din (4.22) se vede că vectorii bazei ordonate $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)}\}$ din U au ca imagine prin $\mathcal{A}$ vectorii $\{\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(p)}\}$ din V, vectori ce se pot exprima sub formă de combinaţii liniare de vectorii bazei ordonate din V. În acest caz relaţia (4.22) devine

$$\mathbf{v} = \mathcal{A}\mathbf{u} = u_1(a_{11}\mathbf{y}^{(1)} + a_{21}\mathbf{y}^{(2)} + \ldots + a_{q1}\mathbf{y}^{(q)}) + u_2(a_{12}\mathbf{y}^{(1)} +$$

$$+ a_{22}\,\mathbf{y}^{(2)} + \ldots + a_{q2}\mathbf{y}^{(q)}) + \ldots + u_p(a_{1p}\mathbf{y}^{(1)} + a_{2p}\mathbf{y}^{(2)} + \ldots + a_{qp}\mathbf{y}^{(q)}) =$$

$$= (u_1a_{11} + u_2a_{12} + \ldots + u_pa_{1p})\mathbf{y}^{(1)} + (u_1a_{21} + u_2a_{22} + \ldots$$

$$\ldots + u_pa_{2p})\mathbf{y}^{(2)} + \ldots + (u_1a_{q1} + u_2a_{q2} + \ldots + u_pa_{qp})\mathbf{y}^{(q)} =$$

$$= v_1\mathbf{y}^{(1)} + v_2\mathbf{y}^{(2)} + \ldots + v_q\,\mathbf{y}^{(q)} = \mathbf{v}.$$

Din (4.21) se vede că $\mathbf{u} \in U$ are coordonate relativ la baza $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)}\}$, la fel va avea şi $\mathbf{v} \in V$ relativ la baza $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(q)}\}$ şi atunci se poate realiza corespondenţă

$$\mathbf{v} \to \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_q \end{bmatrix}. \qquad (4.24)$$

Din (4.24) se vede că dacă sînt date componentele $u_1, u_2, \ldots, u_p$ ale vectorului $\mathbf{u} \in U$, atunci se pot calcula componentele $v_1, v_2, \ldots, v_q$ ale vectorului $\mathbf{v} \in V$ cu ajutorul sistemului de ecuaţii

$$\left.\begin{aligned} v_1 &= u_1a_{11} + u_2a_{12} + \ldots + u_pa_{1p} \\ v_2 &= u_1a_{21} + u_2a_{22} + \ldots + u_pa_{2p} \\ &\cdots\cdots\cdots\cdots\cdots \\ v_q &= u_1a_{q1} + u_2a_{q2} + \ldots + u_pa_{qp} \end{aligned}\right\} \qquad (4.25)$$

Sistemul de ecuaţii (4.25) mai poate fi scris şi sub forma

$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_q \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{q1} & a_{q2} & \cdots & a_{qp} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{bmatrix}. \tag{4.26}$$

Dacă se introduc notaţiile

$$\mathbf{u}_x = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{bmatrix}, \quad \mathbf{v}_y = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_q \end{bmatrix}, \tag{4.27}$$

reprezentînd coordonatele vectorilor $\mathbf{u}$ şi $\mathbf{v}$ relativ la bazele $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(p)}\} \in U$, respectiv $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(q)}\} \in V$, atunci (4.26) se poate scrie sub forma

$$\mathbf{v}_y = A\mathbf{u}_x, \tag{4.28}$$

unde A este matricea asociată transformării liniare $\mathscr{A} \in \mathscr{L}(U,V)$ relativ la cele două baze ordonate de vectori considerate. Relaţia (4.28) permite calculul vectorului $\mathbf{v}_y$ cu ajutorul vectorului $\mathbf{u}_x$. Se observă că matricea A are dimensiunea $p \times q$, unde p şi q sînt dimensiunile lui U, respectiv V. Coloana i din A reprezintă coordonatele vectorului imagine din V, $\mathscr{A}\mathbf{x}^{(i)} \in V$, relativ la baza ordonată $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(q)}\}$. Evident $\mathscr{A}$ este complet determinată prin alegerea bazelor în U şi V.

Avînd în vedere natura aplicaţiilor din practică, atenţia se îndreaptă la spaţiile vectoriale R^n şi C^n şi de aceea foarte des se utilizează baza naturală $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}\}$.

4.4. Produsul intern în R^n şi C^n

Fie $\mathbf{x}, \mathbf{y} \in R^n$. Atunci

$$(\mathbf{x}, \mathbf{y}) = x_1y_1 + x_2y_2 + \ldots + x_ny_n, \ (\mathbf{x}, \mathbf{y}) \in R, \tag{4.29}$$

se numeşte produsul scalar al vectorilor $\mathbf{x}$ şi $\mathbf{y}$. Produsul scalar $(\mathbf{x}, \mathbf{y}) \in R$ poate conduce la următoarea interpretare geometrică: dacă vectorii $\mathbf{x}$ şi $\mathbf{y}$ corespund vectorilor $\overrightarrow{OP}$ şi $\overrightarrow{OQ}$ introduşi în acest capitol, atunci pentru $\mathbf{x}, \mathbf{y} \in R^3$, produsul $(\mathbf{x}, \mathbf{y})$ se poate exprima astfel prin relaţia:

$$(\mathbf{x}, \mathbf{y}) = \sqrt{x_1^2 + x_2^2 + x_3^2}\ \sqrt{y_1^2 + y_2^2 + y_3^2}\ \cos\theta, \qquad (4.30)$$

unde θ este unghiul dintre segmentele de dreaptă orientate $\overrightarrow{OP}$ şi $\overrightarrow{OQ}$.

Din (4.30) se vede că dacă $\mathbf{x} \perp \mathbf{y}$, rezultă $\theta = 90°$ şi $(\mathbf{x}, \mathbf{y}) = 0$.

Dacă $\mathbf{x}, \mathbf{y} \in R^n$, atunci

$$(\mathbf{x}, \mathbf{y}) = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n \qquad (4.31)$$

se numeşte produsul intern al vectorilor $\mathbf{x}$ şi $\mathbf{y}$; în acest caz $\mathbf{x}$ şi $\mathbf{y} \in R^n$ sînt ortogonali dacă şi numai dacă $(\mathbf{x}, \mathbf{y}) = 0$. Din această afirmaţie rezultă că vectorul nul $\mathbf{0}$ este ortogonal pe orice vector din spaţiu.

Produsul intern (scalar) poate fi generalizat pe orice spaţiu vectorial real. Fie V spaţiu vectorial pe R. Atunci produsul intern este o funcţie care asociază fiecărei perechi ordonate de vectori $\mathbf{x}, \mathbf{y} \in V$ un număr real $(\mathbf{x}, \mathbf{y})$ care satisface următoarele proprietăţi:

1) $(\mathbf{x}, \mathbf{x}) > \mathbf{0}$, dacă $\mathbf{x} \neq \mathbf{0}$;

2) $(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})$, dacă $\mathbf{x}, \mathbf{y} \in V$;

3) $(\mathbf{x} + \mathbf{y}, \mathbf{z}) = (\mathbf{x}, \mathbf{z}) + (\mathbf{y}, \mathbf{z})$, dacă $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$; (4.32)

4) $\left.\begin{aligned}(k\mathbf{x}, \mathbf{y}) &= k(\mathbf{x}, \mathbf{y}) \\ (\mathbf{x}, k\mathbf{y}) &= k(\mathbf{x}, \mathbf{y})\end{aligned}\right\}$, dacă $k \in R$ şi $\mathbf{x}, \mathbf{y} \in V$.

De asemenea dacă $\mathbf{x} = \mathbf{0}$, rezultă $(\mathbf{x}, \mathbf{x}) = \mathbf{0}$, respectiv $(\mathbf{0}, \mathbf{y}) = (\mathbf{y}, \mathbf{0}) = \mathbf{0}$.

Un spațiu vectorial V împreună cu produsul intern definit prin (4.32) este numit spațiu euclidian. Ținînd seama de interpretarea geometrică din R^3, o metodă de a defini lungimea unui vector în R^n este utilizarea rădăcinii pătrate a produsului intern

$$\sqrt{(\mathbf{x}, \mathbf{x})} = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}. \tag{4.33}$$

În anumite situații este preferabil în R^n și C^n să folosim produsul intern și să tratăm vectorii ca matrice constînd dintr-o singură linie sau o singură coloană. Se consideră două matrice cu elemente din R^n și se formează produsul matricelor de dimensiune $1 \times n$ și $n \times 1$, obținîndu-se

$$[a_{11}\ a_{12} \ldots a_{1n}] \begin{bmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{n1} \end{bmatrix} = [c_{11}], \tag{4.34}$$

unde

$$c_{11} = a_{11}\, b_{11} + a_{12}\, b_{21} + \ldots + a_{1n}\, b_{n1}.$$

Ținînd seama de notațiile din 4.1 ($\mathbf{x}$ și $\mathbf{x}^T$), se vede din (4.34) că

$$(\mathbf{x}, \mathbf{y}) = \mathbf{y}^T\, \mathbf{x} \quad \text{și} \quad (\mathbf{x}, \mathbf{y}) = \det\, (\mathbf{y}^T\, \mathbf{x}).$$

Pentru definirea produsului intern în C^n se introduc următoarele notații:

● Dacă $c = a + ib$, atunci $\bar{c} = a - ib$ (complex conjugatul).

● $c\bar{c} = a^2 + b^2$ și $c\bar{c} = |\, c\, |^2$.

● Dacă $b = 0$, atunci $c \in R$, de unde rezultă că corpul numerelor reale este o mulțime a lui C. În acest sens se vor prezenta o serie de proprietăți:

● $0 \neq c \in C \Rightarrow 0 < c\,\bar{c} \in R$;

● $a \in R \Leftrightarrow a = \bar{a}$;

$\bullet\ A \in M_R^{q\times p} \Leftrightarrow A = \bar{A}$;

$\bullet\ c \in C \Rightarrow (c + \bar{c}) \in R$;

$\bullet\ A \in M_C^{q\times p} \Rightarrow (A + \bar{A}) \in M_R^{q\times p}$ (4.35)

$\bullet\ cd \in C^{2p} \Rightarrow \overline{cd} = \bar{c}\bar{d}$;

$\bullet\ A,B$ conformabile $\Rightarrow \overline{AB} = \bar{A}\bar{B}$;

$\bullet\ c \in C \Rightarrow \bar{\bar{c}} = c$;

$\bullet\ A \in M_C^{qp} \Rightarrow \overline{\bar{A}} = A$;

$\bullet\ c,d \in C \Rightarrow \overline{c+d} = \bar{c} + \bar{d}$;

$\bullet\ A,B \in M_C^{qp} \Rightarrow \overline{A+B} = \bar{A} + \bar{B}$.

Fie V un spațiu vectorial peste C. Atunci produsul intern pe V este o funcție care asociază fiecărei perechi ordonate de vectori $x, y \in V$ un număr complex (x, y) satisfăcînd proprietățile:

1) $0 < (\mathbf{x},\mathbf{x}) \in R$, dacă $\mathbf{x} \neq \mathbf{0}$;

2) $(\mathbf{x},\mathbf{y}) = \overline{(\mathbf{y},\mathbf{x})}$, dacă $\mathbf{x},\mathbf{y} \in V$; (4.36)

3) $(\mathbf{x}+\mathbf{y},\mathbf{z}) = (\mathbf{x},\mathbf{z}) + (\mathbf{y},\mathbf{z})$, dacă $\mathbf{x},\mathbf{y},\mathbf{z} \in V$;

4) $\left.\begin{aligned}(c\mathbf{x},\ \mathbf{y}) &= \bar{c}(\mathbf{x},\mathbf{y})\\ (\mathbf{x},\ c\mathbf{y}) &= \bar{c}(\mathbf{x},\mathbf{y})\end{aligned}\right\}$, dacă $c \in C$ și $\mathbf{x},\mathbf{y} \in V$.

Folosind relațiile din (4.36), se poate arăta că $\mathbf{x} = \mathbf{0}$ implică $(\mathbf{x},\mathbf{y})=0$ și pentru $\mathbf{y} \in V$ rezultă $(\mathbf{0},\mathbf{y}) = 0$. Produsul intern pentru $\mathbf{x},\mathbf{y} \in C^n$ este dat de relația

$$(\mathbf{x},\mathbf{y}) = x_1\bar{y}_1 + x_2\bar{y}_2 + \ldots + x_n\bar{y}_n \qquad (4.37)$$

în sensul relațiilor din (4.36).

De asemenea dacă $\mathbf{x}, \mathbf{y} \in C^n$, atunci $\mathbf{x}$ și $\mathbf{y}$ sînt ortogonali dacă $(\mathbf{x}, \mathbf{y}) = 0$, iar vectorul nul $\mathbf{0}$ este ortogonal pe orice vector din C^n. Orice spațiu vectorial V împreună cu produsul intern definit prin (4.36) se numește spațiu unitar, ca o consecință, C^n împreună cu produsl intern (4.37) este un spațiu unitar [42, 17].

Se observă că un mod pentru a defini lungimea unui vector $\mathbf{x} \in C^n$ este de a utiliza valoarea pozitivă a rădăcinii pătrate a produsului intern:

$$\sqrt{(\mathbf{x},\mathbf{x})} = \sqrt{|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2}. \tag{4.38}$$

Dacă $A \in M_C^{qp}$ cu $A = (a_{ij})$, atunci A^H este matricea obținută din A prin luarea transpusei lui $\bar{A}$, adică

$$A^H = (\overline{a_{ji}}), \tag{4.39}$$

iar A^H se numește complex conjugata transpusei lui A sau conjugata hermitiană a lui A.

Fie $\mathbf{x} \in C^n$. Dacă

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad x_i \in C,\ i = 1,2,\ldots,n \Rightarrow \mathbf{x}^H = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n], \tag{4.40}$$

atunci produsul intern poate fi scris sub forma

$$(\mathbf{x},\mathbf{y}) = \mathbf{y}^H\mathbf{x}. \tag{4.41}$$

În cazul în care se consideră două matrice A și B, dacă $\mathbf{a}$ este linia i din A și $\mathbf{b}$ este coloana j din B, atunci pentru produsul matriceal cu matrice reale sau complexe $c_{ij} = \mathbf{a}^T\mathbf{b}$; dacă matricele $A, B \in R^n$, atunci $c_{ij} = (\mathbf{b}, \mathbf{a})$, iar dacă $A, B \in C^n$, atunci $c_{ij} \neq (\mathbf{b}, \mathbf{a})$ pentru că $(\mathbf{b}, \mathbf{a}) = \mathbf{a}^H\mathbf{b}$ și nu cu $\mathbf{a}^T\mathbf{b}$ în cazul complex, fapt care impune atenție pentru înlăturarea confuziilor.

4.5. Tipuri speciale de matrice; proprietăţi

În acest paragraf se vor prezenta în special matricele din $M_R^{p\times q}$ şi $M_C^{p\times q}$, ţinînd seama de natura elementelor matricelor ce apar în aplicaţiile curente din practică. În acest sens se poate afirma că matricea A este :

- o matrice complexă hermitiană, dacă $A \in M_C^{p\times q}$ şi are proprietatea $A^H = A^T$;
- o matrice simetrică reală, dacă $A \in M_R^{n\times n}$ şi are proprietatea $A^H = A^T$, în cazul complex al matricei hermitiene $(a_{ij}) = (\overline{a_{ji}})$, iar în cazul real, $(a_{ij}) = (a_{ji})$.

Exemple

$$A = \begin{bmatrix} 3 & 1-\mathrm{i} \\ 1+\mathrm{i} & 5 \end{bmatrix}, \quad A^H = \begin{bmatrix} 3 & 1-\mathrm{i} \\ 1+\mathrm{i} & 5 \end{bmatrix} \Rightarrow A = A^H.$$

$$A = \begin{bmatrix} 1 & 5 \\ 5 & 7 \end{bmatrix}, \quad A^T = \begin{bmatrix} 1 & 5 \\ 5 & 7 \end{bmatrix} \Rightarrow A = A^T.$$

În cazul matricelor complexe şi pentru $k \in C$, au loc următoarele relaţii :

1) $(A^H)^H = A$, 3) $(\bar{k}A)^H = \bar{k}A^H$,

2) $(A + B)^H = A^H + B^H$, 4) $(AB)^H = B^H A^H$.

Dacă matricele sînt reale şi $k \in R$, atunci

1) $(A^T)^T = A^T$, 3) $(kA)^T = kA^T$,

2) $(A + B)^T = A^T + B^T$, 4) $(AB)^T = B^T A^T$,

cu menţiunea că matricele considerate permit ca operaţiile de mai sus să aibă sens, din punct de vedere al dimensiunilor lor.

- Dacă $A \in M_C^{n\times n} \Rightarrow AA^H$ este hermitiană.
- Dacă $A \in M_R^{n\times n} \Rightarrow AA^T$ este simetrică.

● Dacă A admite A^{-1} şi B admite B^{-1}, atunci AB este nesingulară. Cu alte cuvinte:

$$\det(AB) \neq 0 \Leftrightarrow \det A \neq 0 \text{ şi } \det B \neq 0.$$

De asemenea, au loc relaţiile:

1) $(A^{-1})^{-1}=A$, 2) $(kA)^{-1}=(1/k)A^{-1}$, e) $(AB)^{-1} = B^{-1}A^{-1}$

pentru $k \in R$ şi $A, B \in M_R^{n\times n}$ nesingulare.

● Dacă $A \in M_R^{n\times n}$, se defineşte $A^0 = I$ şi $A^1 = A$ şi dacă $n > 1$, se defineşte:

1) $A^n = \underbrace{AA \ldots A}_{n}$, 3) $(A^n)^m = A^{nm}$,

2) $A^m A^n = A^{m+n}$, 4) $A^m A^n = A^n A^m$,

● Dacă $A \in M_C^{nn}$ este nesingulară, atunci

$$(A^{-1})^H = (A^H)^{-1} = A^{-H} \text{ şi } A^{-H} A^H = I.$$

● Dacă $A \in M_R^{n\times n}$ este nesingulară, atunci

$$(A^{-1})^T = (A^T)^{-1} = A^{-T} \text{ şi } A^{-T}A^T = I.$$

● Dacă $A \in M_R^{n\times n}$ este nesingulară şi dacă $A^T = A^{-1}$, atunci A se numeşte matrice ortogonală.

● Dacă A este ortogonală, atunci $AA^T = I$.

Se observă că produsul matriceal AA^T implică produsul intern, unde vectorii sînt liniile lui A; dacă aceste linii sînt vectorii $\{\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(n)}\}$, atunci

$$AA^T = \begin{bmatrix} (\mathbf{a}^{(1)}, \mathbf{a}^{(1)}) & (\mathbf{a}^{(1)}, \mathbf{a}^{(2)}) & \ldots & (\mathbf{a}^{(1)}, \mathbf{a}^{(n)}) \\ (\mathbf{a}^{(2)}, \mathbf{a}^{(1)}) & (\mathbf{a}^{(2)}, \mathbf{a}^{(2)}) & \ldots & (\mathbf{a}^{(2)}, \mathbf{a}^{(n)}) \\ \ldots & \ldots & \ldots & \ldots \\ (\mathbf{a}^{(n)}, \mathbf{a}^{(1)}) & (\mathbf{a}^{(n)}, \mathbf{a}^{(2)}) & \ldots & (\mathbf{a}^{(n)}, \mathbf{a}^{(n)}) \end{bmatrix} = (\delta_{ij}). \quad (4.42)$$

Altfel spus, liniile lui A sînt mutual ortogonale, în plus fiecare linie a lui A are o lungime unitară, vectorii cu lungime unitară fiind numiți vectori normalizați.

Un sistem de vectori $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\}$ normalizat și mutual ortogonal se numește *sistem de vectori ortonormal*. Dacă A este o matrice ortogonală, adică $A^T = A^{-1}$, prin amplificare cu A la stînga sau la dreapta rezultă relația $AA^T = A^TA = I$, obținîndu-se următoarele rezultate:

● Dacă $A \in M_R^{n\times n}$ și este ortogonală, atunci atît liniile cît și coloanele formează sisteme ortonormale.

● Dacă $A \in M_C^{nn}$ și A este nesingulară, iar $A^H = A^{-1}$, atunci A este matrice unitară, de unde prin amplificare cu A la stînga sau la dreapta se obține relația $AA^H = A^HA = I$. Se observă că matricea ortogonală poate fi considerată a fi un caz special al matricei unitare.

● Dacă $A \in M_C^{n\times n}$ este unitară, atunci liniile și coloanele ei formează sisteme de vectori ortonormali.

După această prezentare se poate sintetiza următoarele proprietăți:

● Dacă $A, B, I \in M_C^{n\times n}$, atunci:

1) I este unitară;

2) dacă A este unitară, atunci A^H este unitară;

3) două A și B sînt unitare, atunci AB este unitară;

4) dacă A este unitară, atunci $|\det A| = 1$.

● Pentru $A, B, I \in M_R^{n\times n}$, avem:

1) I este ortogonală;

2) dacă A este ortogonală, și A^T este ortogonală;

3) dacă A și B sînt ortogonale, atunci și AB este ortogonală;

4) dacă A este ortogonală, atunci $\det A = \pm 1$.

Clasa matricelor ortogonale conține așa-numitele matrice elementare de rotație, care diferă de matricele iden-

tice numai prin patru elemente :

$$R(i,j) = \begin{array}{c} \\ \\ \\ \\ \\ i \\ \\ \\ \\ \\ j \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} & & & i & & j & \\ 1 & & & \cdot & & \cdot & \\ & \ddots & & \cdot & & \cdot & \\ & & \ddots & \cdot & & \cdot & \\ & & & 1 & & \cdot & \\ \cdots & & & r_{ii} & \cdots & r_{ij} & \cdots \\ & & & & 1 & \cdot & \\ & & & & & \ddots & \\ & & & & & \cdot & \\ & & & & & 1 & \\ \cdots & & & r_{ji} & \cdots & r_{jj} & \cdots \\ & & & & & & 1 \end{array} \\ \left.\begin{array}{l} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array}\right\} - \end{array} \begin{array}{l} \cos\theta = r_{ii}, \\ \sin\theta = r_{ij}, \\ \sin\theta = r_{ji}, \\ \cos\theta = r_{jj}, \end{array} \tag{4.43}$$

unde r_{ii}, r_{ij}, r_{ji}, $r_{jj} \in R$.

● Există o clasă de matrice care comută cu conjugata sa hermitiană; această clasă conține matricele simetrice reale și matricele complexe hermitiene, ca matrice speciale; ele sînt numite matrice normale. Dacă $A \in M_C^{n\times n}$ și $A^H A = AA^H$, atunci A se numește matrice normală.

● O altă clasă importantă de matrice este clasa matricelor de permutație. O matrice $P \in M_R^{n\times n}$ ale cărei coloane sînt vectorii unitate $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}\}$, dar nu neapărat în ordine, se numește matrice de permutare. De exemplu :

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{4.44}$$

este o matrice de permutare în R^{44}.

O matrice A înmulțită cu P la dreapta (stînga) dă o matrice A' cu coloanele (liniile) permutate. De exemplu:

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} a_{21} & a_{22} & a_{23} & a_{24} \\ a_{41} & a_{42} & a_{43} & a_{44} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{11} & a_{12} & a_{13} & a_{14} \end{bmatrix}$$

● O matrice de permutare $P \in M_{\mathrm{P}}^{n \pm n}$ este unitară, adică $P^H = P^{-1}$.

4.6. Operații între matrice și vectori

S-a introdus în paragrafele precedente produsul între o matrice și un vector precum și produsul intern între vectori, elemente ce se vor folosi în prezentarea formelor cvadratice și hermitiene.

Considerîndu-se spațiile unitare C^m și C^n și matricele $A \in C^{m \times n}$ și $B \in C^{nm}$, dacă $\mathbf{x} \in C^m$ și $\mathbf{y} \in C^n$, atunci:

$$1)\ (\mathbf{x}, A\mathbf{y}) = (A^H\mathbf{x}, \mathbf{y}) \quad 2)\ (B\mathbf{x}, \mathbf{y}) = (\mathbf{x}, B^H\mathbf{y}), \qquad (4.45)$$

iar pentru spațiile euclidiene R^m și R^n împreună cu $A \in R^{m \times n}$ și $B \in R^{n \times m}$, dacă $\mathbf{x} \in R^m$ și $\mathbf{y} \in R^n$, atunci au loc relațiile:

$$1)\ (\mathbf{x}, A\mathbf{y}) = (A^T\mathbf{x}, \mathbf{x}), \quad 2)\ (B\mathbf{x}, \mathbf{y}) = (\mathbf{x}, B^T\mathbf{y}). \qquad (4.46)$$

Expresia $(B\mathbf{x}, \mathbf{y})$ din (4.46) apare destul de frecvent în calcule și este numită forma biliniară în $x_1, x_2, \ldots, x_m$ și $y_1, y_2, \ldots, y_n$, aceasta poate fi scrisă dezvoltat astfel:

$$(B\mathbf{x}, \mathbf{y}) = \mathbf{y}^T B\mathbf{x} = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} b_{ij}\, x_j \right) y_i. \qquad (4.47)$$

În cazul cînd $B \in R^{nn}$ și $\mathbf{x} = \mathbf{y}$, rezultă

$$(B\mathbf{x}, \mathbf{y}) = \mathbf{x}^T B\mathbf{x} = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} b_{ij}\, x_j \right) x_i. \tag{4.48}$$

Exemplu. Pentru $n = 2$ se poate scrie

$$L = \mathbf{x}^T A\mathbf{x} = [x_1,\ x_2] \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \tag{4.49}$$

$$= a_{11}x_1^2 + a_{22}x_2^2 + a_{12}x_1x_2 + a_{21}x_2x_1.$$

Scalarul a_{ij} este coeficientul termenului x_ix_j, dar $x_ix_j = x_jx_i$, deci $a_{12}x_1x_2 + a_{21}x_2x_1 = (a_{12} + a_{21})x_1x_2$.

Dacă $a_{12} \neq a_{21}$, atunci se calculează $\hat{a}_{12}$ și $\hat{a}_{21}$ ca o medie aritmetică astfel:

$$\hat{a}_{12} = \hat{a}_{21} = \frac{a_{12} + a_{21}}{2} \Rightarrow \hat{A} = \begin{bmatrix} a_{11} & \hat{a}_{12} \\ \hat{a}_{21} & a_{22} \end{bmatrix}.$$

unde A este o matrice simetrică. În acest caz forma biliniară L are expresia

$$L = \mathbf{x}^T\hat{A}\mathbf{x} = [x_1x_2] \begin{bmatrix} a_{11} & \hat{a}_{12} \\ \hat{a}_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{4.50}$$

Din analiza relațiilor (4.49) și (4.50) se vede că expresia $\mathbf{x}^TA\mathbf{x}$ cu A nesimetrică este egală cu expresia $\mathbf{x}^T\hat{A}\mathbf{x}$ cu $\hat{A}$ simetrică, lucru valabil și pentru n întreg și arbitrar.

Această afirmație se poate generaliza dacă $L = \mathbf{x}^TA\mathbf{x}$, cu $A \in R^{n\times n}$, arbitrară, rezultă $\mathbf{x}^TA\mathbf{x} = \mathbf{x}^T\hat{A}\mathbf{x}$, unde $\hat{A} = \frac{1}{2}(A^T + A)$.

● Dacă $\mathbf{x} \in R^n$ și $A \in R^{n\times n}$, unde A este simetrică, atunci expresia

$$L = \mathbf{x}^TA\mathbf{x} = (A\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}\, x_i\, x_j \tag{4.51}$$

se numește *formă pătratică* în $x_1, x_2, \ldots, x_n$, iar matricea simetrică A se numește *matricea formei pătratice*.

● Dacă $\mathbf{x} \in C^n$ și $A \in C^{n \times n}$, unde A este hermitiană, atunci expresia

$$L = \mathbf{x}^H A\mathbf{x} = (A\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_j \bar{x}_i \qquad (4.52)$$

se numește *formă hermitiană* în $x_1, x_2, \ldots, x_n$. Matricea hermitiană A se numește *matricea formei hermitiene*. Se observă că se pot scrie relațiile

$$(A\mathbf{x}, \mathbf{x}) = \mathbf{x}^T A\mathbf{x}, \qquad (A\mathbf{x}, \mathbf{x}) = \mathbf{x}^H A\mathbf{x} \qquad (4.53)$$

în spațiul euclidian, respectiv în spațiul unitar cu matricea A nesimetrică și A respectiv nehermitiană.

Relația (4.53) permite definirea formei produsului intern, unde forma pătratică și forma hermitiană sînt cazuri particulare.

● Fie $M \in C^{n \times n}$ și $\mathbf{x} \in C^n$; atunci

$$L = \mathbf{x}^H M\ \mathbf{x} = (M\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij} x_j \bar{x}_i \qquad (4.54)$$

formează un produs intern în $x_1, x_2, \ldots, x_n$, unde matricea M este numită matricea formei produsului intern. Cînd M este simetrică și $\mathbf{x}$ real, (4.54) se reduce la (4.51), iar cînd M este hermitiană și $\mathbf{x}$ complex, (4.54) se reduce la (4.52).

● Dacă matricea A este hermitiană, atunci forma hermitiană

$$\mathbf{x}^H A\mathbf{x} = \mathbf{x}^H A^H \mathbf{x} = (\mathbf{x}^H A\mathbf{x})^H = \overline{\mathbf{x}^H A\mathbf{x}} \qquad (4.55)$$

implică faptul că $\mathbf{x}^H A\mathbf{x}$ este un număr real, deci forma hermitiană

$$\mathbf{x}^H A\mathbf{x} = (A\mathbf{x}, \mathbf{x}) \qquad (4.56)$$

este un număr real pentru orice $\mathbf{x} \in C^n$.

● Dacă $A \in C^{n\times n}$, atunci $\mathbf{x}^H A\mathbf{x} = (A\mathbf{x}, \mathbf{x}) \in R$ pentru orice $\mathbf{x} \in C^n$ dacă și numai dacă A este hermitiană.

● Dacă forma hermitiană $(A\mathbf{x}, \mathbf{x}) = \mathbf{x}^H A\mathbf{x}$ este pozitivă pentru $\mathbf{x} \in C^n$ cu $\mathbf{x} = \mathbf{0}$ și $A \in C^{n\times n}$, atunci ea se numește o formă hermitiană pozitiv definită și matricea hermitiană A se numește matrice hermitiană pozitiv definită.

● Dacă forma pătratică $(A\mathbf{x}, \mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ este pozitivă pentru orice $\mathbf{x} \in R^n$ cu $\mathbf{x} \neq \mathbf{0}$, atunci aceasta se numește formă pătratică pozitiv definită și matricea $A \in R^{n\times n}$ se numește matrice simetrică pozitiv definită.

● Dacă A este o matrice reală simetrică pozitiv definită, atunci [42, 50] există o matrice nesingulară $B \in R^{n\times n}$ astfel că

$$A = B^T B. \tag{4.57}$$

● Dacă $A \in R^{n\times n}$ este simetrică și produsul $(A\mathbf{x}, \mathbf{x})$ este pozitiv pentru $\mathbf{x} \in R^n$ cu $\mathbf{x} \neq \mathbf{0}$, atunci $(A\mathbf{x}, \mathbf{x})$ este pozitiv pentru orice $\mathbf{x} \in C^n$ cu $\mathbf{x} \neq \mathbf{0}$. Pentru demonstrație se folosește relația (4.57) astfel:

$$(A\mathbf{x}, \mathbf{x}) = (B^T B\mathbf{x}, \mathbf{x}) = (B\mathbf{x}, B\mathbf{x}) = (\mathbf{y}, \mathbf{y}), \tag{4.58}$$

dar $(\mathbf{y}, \mathbf{y})$ este pozitiv pentru orice $\mathbf{x} \in C^n$ cu $\mathbf{x} \neq \mathbf{0}$.

Se poate arăta că pentru $\mathbf{x} \in R^n$ este posibil ca $\mathbf{x}^T A\mathbf{x}$ să fie reală pentru orice $\mathbf{x} \in R^n$, chiar dacă A este o matrice complexă nehermitiană.

Exemplu. Fie

$$L = [x_1\ x_2]\begin{bmatrix} 1 & 3-\mathrm{i} \\ 1+\mathrm{i} & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + 3x_2^2 + (3-\mathrm{i})x_1x_2 + (1+\mathrm{i})x_2x_1 = \\ = x_1^2 + 3x_2^2 + 3x_1x_2 + x_2x_1.$$

Se vede că expresia L este reală pentru orice $\mathbf{x} \in R^n$ și $A \in C^{n\times n}$ nehermitiană.

În spațiul euclidian, dacă $A \in R^{n\times n}$ este nesimetrică, $\mathbf{x}^T A\mathbf{x}$ este totdeauna reală.

Deoarece $A \in C^{n\times n}$ include $A \in R^{n\times n}$ ca un caz particular, se poate afirma că $A \in C^{n\times n}$ și este reală, pozitiv definită

dacă şi numai dacă au loc relaţiile :

1) $A^T + A$ este matrice reală;

2) $A^T + A$ este pozitiv definită. (4.59)

Pentru demonstraţie se vede că dacă $A = (B + iC)$, unde $B, C \in R^{n \times n}$ şi $\mathbf{x} \in R^n$, rezultă

$$L = \mathbf{x}^T A\mathbf{x} = \mathbf{x}^T(B + iC)\mathbf{x} = \mathbf{x}^T B\mathbf{x} + i\mathbf{x}^T C\mathbf{x} =$$

$$= \frac{1}{2}[\mathbf{x}^T(B^T + B)\mathbf{x} + i\mathbf{x}^T(C^T + C)\mathbf{x}]. \quad (4.60)$$

Expresia L din (4.60) este reală dacă şi numai dacă $\mathbf{x}^T(C^T + C)\mathbf{x} = 0$ pentru orice $\mathbf{x} \in R$ şi aceasta este adevărat dacă şi numai dacă $C^T + C = 0$.

Din relaţia (4.60) se vede că $\mathbf{x}^T A\mathbf{x}$ este pozitiv definită dacă şi numai dacă $B^T + B$ este pozitiv definită, deoarece

$$A^T + A = (B^T + B) + i(C^T + C) = B^T + B. \quad (4.61)$$

Dacă se consideră o matrice arbitrară $A \in C^{n \times n}$, s-a arătat că AA^H este hermitiană şi că forma hermitiană

$$\mathbf{x}^H AA^H\mathbf{x} = (AA^H\mathbf{x}, \mathbf{x}) = (A^H\mathbf{x}, A^H\mathbf{x}) = (\mathbf{y}, \mathbf{y}) \geqslant 0 \quad (4.62)$$

pentru orice $\mathbf{x}$. Dacă are loc şi egalitatea cu zero, atunci $A^H\mathbf{x} = \mathbf{0}$, pentru A nesingulară implică $\mathbf{x} = 0$, elemente ce conduc la următoarea afirmaţie :

● Pentru orice matrice $A \in C^{n \times n}$, matricea AA^H este hermitiană şi nenegativ definită (cazul în care inegalitatea din (4.62) este şi egală cu zero); dacă A este nesingulară, atunci AA^H este pozitiv definită [cazul în care inegalitatea din (4.62) este strict mai mare ca zero].

4.7. Grafuri şi matrice

O serie de proprietăţi ale matricelor se pot interpreta cu ajutorul grafului asociat [119, 100, 42]. În acest sens se vor introduce cîteva noţiuni elementare din teoria grafurilor.

Se consideră matricea $A=(a_{ij})$ cu $A\in R^{n\times n}$ sau $A\in C^{n\times n}$ și n puncte distincte $P_1, P_2, \ldots, P_n$ din plan pe care le vom numi noduri. Pentru orice element a_{ij} al matricei diferit de zero se unesc punctele P_i și P_j printr-un arc cu sens de la P_i la P_j (fig. 4.6). În acest fel se poate asocia oricărei matrice A de dimensiune $n\times n$ un graf direct finit $G(A)$.

Fig. 4.6.

Pentru exemplificare se consideră matricele

$$A=\begin{bmatrix}0 & 0 & 1 & 1\\ 0 & 1 & 1 & 1\\ 1 & 0 & 1 & 0\\ 1 & 1 & 0 & 1\end{bmatrix},\qquad B=\begin{bmatrix}1 & 1\\ & \\ 0 & 1\end{bmatrix},$$

ale căror grafuri directe $G(A)$ sînt date în fig. 4.7.

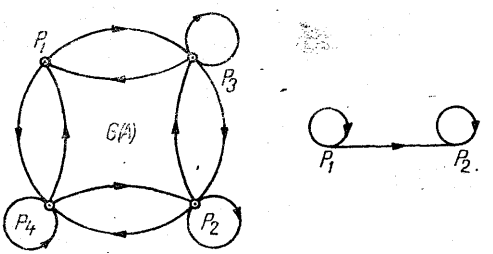


Fig. 4.7

Un graf este tare conex dacă pentru orice pereche de puncte P_i, P_j există o cale de legătură directă ce leagă P_i cu P_j astfel : $\overrightarrow{P_iP_{i_1}}, \overrightarrow{P_{i_1}P_{i_2}}, \ldots, \overrightarrow{P_{i_{m-2}}P_{i_{m-1}}}, \overrightarrow{P_{i_{m-1}}P_{i_m}}=j$, o astfel de cale are lungimea $l=m$ (pentru cazul considerat). Din analiza lui $G(A)$, și din definiția grafului tare conex se vede că $G(A)$ este tare conex, în timp ce $G(B)$ este slab conex deoarece nu există cale de legătură de la P_2 la P_1.

În continuare se va pune în evidență legătura între matricea ireductibilă și graf tare conex.

● O matrice A de dimensiune n reală sau complexă este reductibilă dacă există o matrice de permutare P [definită în (4.44)] astfel că

$$\tilde{A} = PA\,P^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \tag{4.63}$$

unde matricele A_{11} și A_{22} sînt matrice pătrate. În cazul în care nu există o astfel de matrice P, atunci A este ireductibilă.

Noțiunile de matrice reductibilă, respectiv ireductibilă se mai întîlnesc în literatură sub denumirea de matrice decompozabilă, respectiv nedecompozabilă.

În cazul în care se pune problema rezolvării unui sistem de forma $\tilde{A}\mathbf{y} = \mathbf{b}$, unde $\tilde{A} = PAP^T$ este o matrice partiționată, atunci vectorii $\mathbf{y}$ și $\mathbf{b}$ se pot partiționa în mod similar, așa că ecuația matriceală $A\mathbf{x} = \mathbf{b}$ se poate scrie sub forma

$$\left.\begin{aligned} A_{11}\mathbf{x}_1 + A_{12}\mathbf{x}_2 &= \mathbf{b}_1 \\ A_{22}\mathbf{x}_2 &= \mathbf{b}_2 \end{aligned}\right\}, \qquad \mathbf{y} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \text{ și } \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}. \tag{4.64}$$

Prin rezolvarea celei de-a doua ecuații din (4.64) se obține vectorul soluție $\mathbf{x}_2$, care reprezintă o parte din componentele lui $\mathbf{y}$, în urma partiționării.

Introducînd $\mathbf{x}_2$ în prima ecuație din (4.64) se determină vectorul $\mathbf{x}_1$ care conține celelalte componente ale vectorului $\mathbf{y}$. În acest fel se poate rezolva ecuația $A\mathbf{y} = \mathbf{b}$ prin reducerea ei la două ecuații matriceale de ordin inferior.

● O matrice A reală sau complexă este reductibilă dacă și numai dacă pentru orice doi indici distincți $1 \leqslant i, j \leqslant n$ există o secvență de elemente nenule ale lui A de forma

$$\{a_{i,i_1},\ a_{i_1 i_2}, \ldots, a_{i_m, j}\}. \tag{4.65}$$

Din analiza acestei definiții și din definiția grafului tare conex rezultă:

● O matrice A reală sau complexă este ireductibilă dacă și numai dacă graful său asociat este tare conex.

● Dacă se consideră o matrice tridiagonală de forma

$$A = \begin{bmatrix} a_1 & c_1 & & & & \\ b_2 & a_2 & c_2 & & 0 & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & \ddots & c_{n-1} \\ & & & & b_n & a_n \end{bmatrix} \tag{4.66}$$

și i se asociază graful $G(A)$ corespunzător, dat în fig. 4.8, cu presupunerea că $c_1, \ldots, c_{n-1}$ și $b_2, \ldots, b_n$ sînt toate nenule, se observă că graful asociat este tare conex și matricea A este ireductibilă.

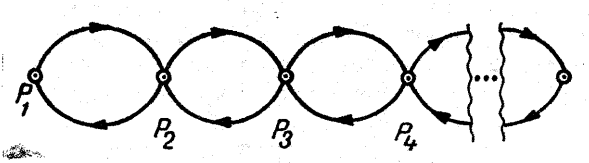


Fig. 4.8

Din cele prezentate se vede că graful direct asociat unei matrice ne oferă o metodă geometrică de analiză a unei matrice dacă aceasta este reductibilă sau ireductibilă.

Fie $A(a_{ij})$ o matrice din $R^{n\times n}$, pozitiv definită și ireductibilă. Atunci :

● Matricea A va avea o valoare proprie $\lambda > 0$ egală cu raza spectrală $\rho(A)$.

● Raza spectrală $\rho(A)$ crește cînd orice element al lui A crește.

● Razei spectrale $\rho(A)$ îi corespunde un vector propriu $\mathbf{x} > 0$.

Dacă matricea A considerată anterior are k valori proprii, al căror modul este egal cu $\rho(A)$, atunci :

● pentru $k = 1$, A este o matrice primitivă,

● pentru $k > 1$, A este o matrice ciclică de index k; în acest caz există k valori proprii ale matricei A de forma

$$\lambda_j = \rho(A)\ \exp(\mathrm{i}\ \varphi_j),\ 0 \leqslant j \leqslant k-1,$$

unde argumentele φ_j se ordonează astfel:

$$0 = \varphi_0 \leqslant \varphi_1 \leqslant \ldots \leqslant \varphi_{k-1} \leqslant 2\pi. \tag{4.67}$$

Din relațiile (4.66) și (4.67) rezultă că valorile proprii sînt dispuse în planul complex pe un cerc de rază egală cu raza spectrală $\rho(A)$ a matricei A, defazate între ele prin unghiul $\frac{2\pi}{k}$.

Graful direct asociat unei matrice [100, 86] este utilizat pentru a determina dacă o matrice A din $R^{n\times n}$, pozitivă și ireductibilă, este primitivă sau ciclică de un anumit indice k. În sensul utilizării grafurilor asociate matricelor pentru a determina dacă o matrice este primitivă sau ciclică sînt necesare următoarele noțiuni:

● Grafurile directe pentru puterile matricei $A(A \geqslant 0)$ se pot deduce din graful direct al lui $G(A)$ asociat matricei A, deoarece graful direct pentru matricea A^k, $k > 1$, se poate obține direct considerînd toate drumurile lui $G(A)$ de lungime egală cu k.

● Pentru drumul $\overrightarrow{P_iP_{s_1}}, \overrightarrow{P_{s_1}P_{s_2}}, \ldots, \overrightarrow{P_{s_{k-1}}P_{s_k}} = j$ determinat din graful lui $G(A)$ (drum de lungime $l = k$) se poate conecta direct nodul P_i cu nodul P_j printr-un segment egal cu unitatea în direcția P_j, în scopul obținerii grafului $G(A^k)$ asociat matricei A^k, fără a fi necesară ridicarea la putere a matricei A.

● Dacă A este primitivă, atunci pentru graful direct $G(A^k)$ cu k suficient de mare, fiecare nod P_i se leagă de fiecare nod P_j printr-un arc de lungime unitatea.

● Dacă A este ireductibilă și ciclică de indice $s > 1$, atunci fiecare graf $G(A^{ks})$, $k > 1$, este o reuniune de s subgrafuri directe *tare conexe*.

Fie matricea $A \in R^{4\times 4}$ și matricea A^2 corespunzătoare:

$$A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 2 \\ 1 & 3 & 0 & 0 \\ 2 & 3 & 0 & 0 \end{bmatrix} \text{ și } A^2 = \begin{bmatrix} 3 & 6 & 0 & 0 \\ 5 & 9 & 0 & 0 \\ 0 & 0 & 4 & 7 \\ 0 & 0 & 5 & 8 \end{bmatrix} \tag{4.68}$$

Se consideră grafurile asociate $G(A)$ și $G(A^2)$ date în fig. 4.9. Matricea A este ciclică de index 2; de asemenea se vede că graful $G(A^2)$ asociat matricei A^2 este o reuniune de două subgrafuri disjuncte *tare conexe*.

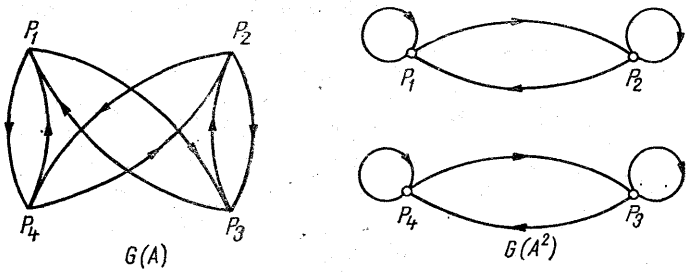


Fig. 4.9

Se observă că graful $G(A^2)$ se poate obține direct din $G(A)$ dacă se unesc între ele punctele P_i cu P_j, puncte pentru care lungimea drumului în $G(A)$ este $l = 2$. Folosind metoda de a obține graful lui A^k din graful matricei A, în mod direct, prin unirea în $G(A^k)$ numai a punctelor P_i cu P_j pentru care lungimea drumului de legătură în $G(A)$ este $l = k$, se pot da o serie de metode pentru a determina dacă o matrice ireductibilă, nenegativă A este primitivă sau ciclică de un anumit index $k > 1$ [86,42].

Teoremă. *Fie* $A = (a_{ij}) \geqslant 0$, $A \in R^{n \times n}$, *ireductibilă cu* $G(A)$ *graful direct asociat. Pentru fiecare nod* P_i *al lui* $G(A)$ *se consideră toate drumurile închise ce leagă pe* P_i *cu el însuși, notîndu-se lungimea acestor drumuri cu* l_i *și mulțimea acestor lungimi* l_i *cu* L_i; *fie* d_i *cel mai mare divizor comun al tuturor lungimilor* l_i *ce formează mulțimea* L_i,

$$d_i = \text{c.m.m.d.c. } \{l_i\},\ 1 \leqslant i \leqslant n,$$

Atunci dacă A *este primitivă,* $d_1 = d_2 = \ldots = d_n = d$, *unde* $d = 1$; *dacă* A *este ciclică de index* d, $d_1 = d_2 = \ldots = d_n = d$, $d > 1$.

Demonstrația teoremei se găsește în [86].

Pentru evidențierea elementelor teoremei se consideră un exemplu, Fie matricele A și B de forma

$$A = \begin{bmatrix} 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 2 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 1 & 5 & 0 & 0 \end{bmatrix},$$

iar grafurile $G(A)$ și $G(B)$ se dau în fig. 4.10. Din analiza grafului $G(A)$ se vede că

$$l_1 = \overrightarrow{P_1P_3} + \overrightarrow{P_3P_2} + \overrightarrow{P_2P_4} + \overrightarrow{P_4P_1} = 4 \text{ unități}$$

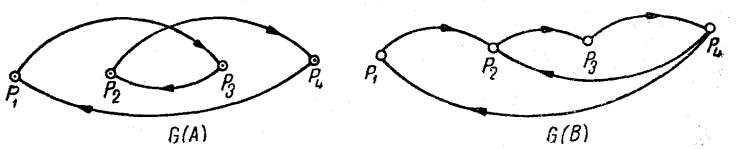


Fig. 4.10

sau, calculînd în continuare,

$$d_1 = \text{c.m.m.d.c. } \{4, 8, 12, \ldots\} = 4,$$

de unde rezultă că matricea A este ciclică de index 4. Din analiza grafului $G(B)$ asociat matricei B se vede că

$$l_1 = \overrightarrow{P_1P_2} + \overrightarrow{P_2P_3} + \overrightarrow{P_3P_4} + \overrightarrow{P_4P_1} = 4,$$

$$l_2 = \overrightarrow{P_2P_3} + \overrightarrow{P_3P_4} + \overrightarrow{P_4P_2} = 3 \text{ sau } l_2 = \overrightarrow{P_2P_3} + \overrightarrow{P_3P_4} + \overrightarrow{P_4P_1} + \overrightarrow{P_1P_2} = 4,$$

$$l_4 = \overrightarrow{P_4P_2} + \overrightarrow{P_2P_3} + \overrightarrow{P_3P_4} = 3 \text{ sau } l_4 = \overrightarrow{P_4P_1} + \overrightarrow{P_1P_2} + \overrightarrow{P_2P_3} + \overrightarrow{P_3P_4} = 4.$$

Drumurile de lungime cea mai scurtă apar la legarea lui P_2 și P_4 cu ele însele, rezultînd lungimile 3 și 4; atunci

$$d_1 = \text{c.m.m.d.c. } \{3, 4, \ldots\} = 1.$$

Deoarece $d_1 = 1$, rezultă că matricea B este primitivă. Fie matricele C și D date sub forma

$$C = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \qquad D = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{bmatrix}$$

și grafurile $G(C)$ și $G(D)$ asociate date în fig. 4.11. Din analiza grafului $G(C)$ se vede că legarea lui P_1 cu el însuși se realizează printr-un drum de lungime unitatea, deci $d_1 = 1$, ceea ce implică faptul că C este o matrice primitivă. Din analiza grafului $G(D)$ se vede că acesta este format din două subgrafuri disjuncte tare conexe, fapt care face ca matricea D să fie o matrice reductibilă.

Dacă graful G este un graf direct finit *tare conex*, atunci G este:

- un graf ciclic de index $d > 1$;
- un graf primitiv dacă c.m.m.d.c. al tuturor lungimilor pentru drumurile cele mai scurte ce leagă un punct

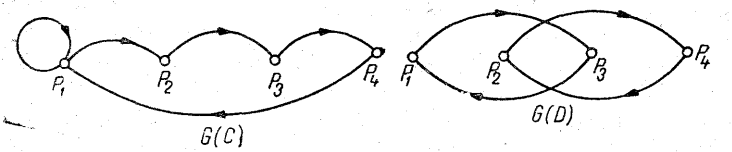


Fig. 4.11

al grafului cu el însuși este respeciv $d > 1$ sau $d = 1$. Amănunte privind noțiunea de graf direct ciclic și graf finit primitiv se pot găsi în [86, 42, 100, 119].

Unei matrice A i se poate asocia [86] un graf direct $G(A)$ de tipul al doilea. Dacă $a_{ij} \neq 0$, atunci arcul din nodul P_i în P_j va fi notat prin două săgeți dacă $j > i$, altfel cu o săgeată (fig. 4.12, a) iar graful matricei A arată

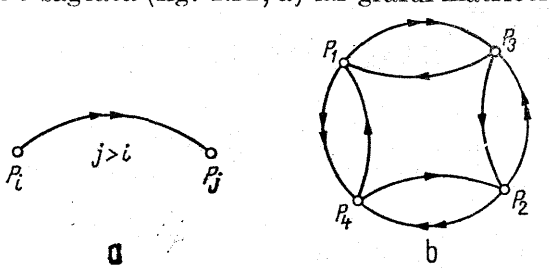


Fig. 4.12

ca în fig. 4.12, b. Arcele cu două săgeți se numesc arce de legătură majoră iar arcele cu o săgeată se numesc arce de legătură minoră.

● Matricea A este *consistent ordonată* numai dacă drumul cel mai scurt între un punct P_i al lui $G(A)$ și el însuși are un număr egal de arce minore și majore [86]. Graful direct de tipul doi permite o verificare a matricei dacă este consistent ordonată.

4.8. Norme vectoriale și norme matriceale

O normă vectorială pe R^n este o funcție aplicată pe R^n cu valori în R^+; aceste numere reale pozitive măsoară într-un anumit sens dimensiunea vectorului din R^n. De

asemenea în aplicații este adesea necesar să considerăm mărimea sau lungimea unui vector din C^n. Se urmărește asocierea unui număr pozitiv unic fiecărui vector, așa cum se asociază fiecărui număr complex $z \in C$ un modul $|z| \in R$. Pentru realizarea acestui lucru se introduce noțiunea de normă. Datorită faptului că există un număr destul de mare de norme vectoriale, în continuare se vor prezenta un număr limitat de norme, care se întîlnesc mai frecvent în aplicații presupunînd că se lucrează cu vectorii din C^n.

● Norma vectorială $\| \cdot \|_\beta$ este o funcție nenegativă pe spațiul C^n cu următoarele proprietăți:

1) $\| \mathbf{x} \|_\beta > 0$ dacă $\mathbf{x} \neq \mathbf{0}$;

2) $\| a\mathbf{x} \|_\beta = |a| \, \| \mathbf{x} \|_\beta$ pentru orice $a \in C$ și orice $\mathbf{x} \in C^n$;

3) $\| \mathbf{x} + \mathbf{y} \|_\beta \leqslant \| \mathbf{x} \|_\beta + \| \mathbf{y} \|_\beta$ pentru orice vectori $\mathbf{x}, \mathbf{y} \in C_n$.

(4.69)

Se va arăta în continuare că există un număr destul de mare de norme vectoriale care satisfac relațiile din (4.69).

Cîteva proprietăți ale normei $\| \cdot \|_\beta$ introduse:

● $\| \mathbf{x} \|_\beta = 0$ dacă și numai dacă $\mathbf{x} = \mathbf{0}$.

● Dacă $\mathbf{x}, \mathbf{y} \in C^n$, orice normă vectorială satisface relația

$$| \, \| \mathbf{x} \|_\beta - \| \mathbf{y} \|_\beta \, | \leqslant \| \mathbf{x} - \mathbf{y} \|_\beta.$$

Dacă se utilizează relația 3) din (4.69), se poate scrie

$$\| \mathbf{x} \|_\beta = \| (\mathbf{x} - \mathbf{y}) + \mathbf{y} \|_\beta \leqslant \| \mathbf{x} - \mathbf{y} \|_\beta + \| \mathbf{y} \|_\beta. \quad (4.70)$$

Relația (4.70) permite scrierea următoarei relații:

$$\| \mathbf{x} \|_\beta - \| \mathbf{y} \|_\beta \leqslant \| \mathbf{x} - \mathbf{y} \|_\beta$$

și în mod similar se obține

$$\| \mathbf{y} \|_\beta - \| \mathbf{x} \|_\beta \leqslant \| \mathbf{y} - \mathbf{x} \|_\beta.$$

Proprietatea 2) din (4.69) pentru $a = -1$ permite egalarea părților din dreapta a inegalităților (4.70):

$$| \, \| \mathbf{x} \|_\beta - \| \mathbf{y} \|_\beta \, | \leqslant \| \mathbf{x} - \mathbf{y} \|_\beta.$$

Datorită faptului că o matrice $A \in C^{n\times n}$ poate fi considerată ca un vector n^2 dimensional, se vor folosi aceleași proprietăți în definirea normei matriceale.

● Norma matriceală $\|\cdot\|_\alpha$ este o funcție nenegativă pe spațiul C^{n^2} cu următoarele proprietăți:

1) $\|A\|_\alpha > 0$, dacă $A \neq 0$.

2) $\|aA\|_\alpha = |a|\,\|A\|_\alpha$ pentru orice $a \in C$ și $A \in C^{n\times n}$.

3) $\|A + B\|_\alpha \leqslant \|A\|_\alpha + \|B\|_\alpha$ pentru orice $A, B \in C^{n\times n}$.

4) $\|AB\|_\alpha \leqslant \|A\|_\alpha \|B\|_\beta$ pentru orice $A, B \in C^{n\times n}$.

(4.71)

Relația 3) poartă numele de inegalitatea triunghiului iar relația 4) de inegalitatea produsului. Există destul de multe norme matriceale care satisfac relațiile din (4.71). Unele proprietăți ale normei matriceale astfel introduse sînt:

● $\|A\|_\alpha = 0$ dacă și numai dacă $A = 0$ și $A \in C^{n\times n}$.

● Pentru $A, B \in C^{n\times n}$ și orice normă matriceală, are loc relația

$$|\,\|A\|_\alpha - \|B\|_\alpha| \leqslant \|A - B\|_\alpha.$$

● Dacă $A \in C^{n\times n}$ reprezintă o aplicație liniară a oricărui vector $\mathbf{x} \in C^n$ în vectorul $\mathbf{y} \in C^n$ prin $\mathbf{y} = A\mathbf{x}$, atunci pentru orice normă $\|\cdot\|_\beta$ se poate defini funcția

$$\sup_{\mathbf{x}\neq 0} \frac{\|\mathbf{y}\|_\beta}{\|\mathbf{x}\|_\beta} = \sup_{\mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\|\mathbf{x}\|_\beta = 1} \|A\mathbf{x}\|_\beta . \qquad (4.72)$$

Aceasta este o funcție de A și se poate arăta că ea satisface cele patru relații cerute de norma matriceală. Astfel se poate scrie relația

$$\|A\|_\gamma = \sup \frac{\|A\mathbf{x}\|_\beta}{\|\mathbf{x}\|_\beta}, \quad \text{unde } \gamma = \gamma(\beta). \qquad (4.73)$$

Dîndu-se orice normă vectorială $\|\cdot\|$, norma matriceală este determinată prin relația (4.73). Aceasta conduce la conceptul de subordonare al normei matriceale.

● Norma matriceală $\|\cdot\|_\gamma$ definită prin (4.73) este numită normă subordonată normei vectoriale $\|\cdot\|_\beta$ corespunzătoare.

● Pentru orice $A \in C^{n\times n}$ și $\mathbf{x} \in C^n$ este satisfăcută relația

$$\|A\mathbf{x}\|_\beta \leqslant \|A\|_\gamma \|\mathbf{x}\|_\beta. \tag{4.74}$$

Datorită faptului că norma matriceală aleasă este subordonată normei vectoriale, (4.74) rezultă direct din (4.73) [86, 124].

● Norma matriceală $\|A\|_\gamma$, și norma vectorială $\|\mathbf{x}\|_\beta$ pentru care are loc inegalitatea (4.74), oricare ar fi A și $\mathbf{x}$, se numesc *norme consistente* sau compatibile [42, 12]. Norma vectorială $\|\cdot\|_\beta$ și norma matriceală subordonată ei $\|\cdot\|_\gamma$ sînt totdeauna consistente. Astfel cel puțin o normă matriceală este consistentă cu o normă vectorială dată și invers.

● Dacă $\mathbf{x} \in C^n$ este un vector arbitrar și $C = [\mathbf{x}, \mathbf{0}, \ldots \ldots, \mathbf{0}]$ este o matrice din $C^{n\times n}$, atunci

$$\|\mathbf{x}\|_\beta = \|C\|_\gamma \tag{4.75}$$

definește o normă vectorială $\|\cdot\|_\beta$ consistentă cu norma matriceală $\|\cdot\|_\gamma$.

● Fie $A \in C^{n\times n}$ și $\mathbf{x} \in C^n$ oarecare; atunci, folosind și relația (4.75), se obține

$$\|A\mathbf{x}\|_\beta = \|[A\mathbf{x}, \mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}]\|_\gamma = \|AB\|_\gamma \leqslant \|A\|_\gamma \|B\|_\gamma = \|A\|_\gamma \|\mathbf{x}\|_\beta,$$

de unde rezultă condiția de consistență (4.74).

Exemple de norme vectoriale. Cel mai frecvent utilizată este:

$$\|\mathbf{x}\|_p = \begin{cases} \sqrt{\sum_{i=1}^{n} |x_i|^p}, & p = 1, 2, 3, \ldots \\ \max_i |x_i|, & p \to \infty. \end{cases} \tag{4.76}$$

Dintre normele vectoriale mai mult utilizate în aplicații sînt normele l_1, l_2, și l_∞ care sînt definite astfel:

$$\begin{aligned} \|\mathbf{x}\|_1 &= \sum_{j=1}^{n} |x_j| && (l_1 - \text{norma sumă}), \\ \|\mathbf{x}\|_2 &= \left(\sum_{j=1}^{n} |x_j|^2\right)^{1/2} && (l_2 - \text{norma euclidiană}), \\ \|\mathbf{x}\|_p &= \left(\sum_{j=1}^{n} |x_j|^p\right)^{1/p} && p \geqslant 1 \quad (l_p - \text{norma } p), \\ \|\mathbf{x}\|_\infty &= \max_j |x_j| && (l_\infty - \text{norma maximă}). \end{aligned} \tag{4.77}$$

Din (4.77) se vede că normele l_1, l_2 sînt cazuri particulare ale clasei generale de norme l_p, unde $p \in [1, \infty)$. De la penultima normă dată în (4.77) se poate trece la ultima prin trecere la limită:

$$\lim_{p \to \infty} \|\mathbf{x}\|_p = \|\mathbf{x}\|_\infty.$$

O altă clasă importantă de norme sînt normele eliptice:

$$\|\mathbf{x}\|_e = (\mathbf{x}^T, B\mathbf{x})^{1/2}, \tag{4.78}$$

unde B este o matrice reală simetrică pozitiv definită. Norma eliptică este valabilă pe C^n dacă $\mathbf{x}^T$ este înlocuit prin $\mathbf{x}^H$ și matricea B este hermitiană și pozitiv definită.

Pentru orice normă, mulțimea $\{\mathbf{x} : \|\mathbf{x}\| \leqslant 1\}$ sau suprafața $\{\mathbf{x} : \|\mathbf{x}\| = 1\}$ este sfera unitate. Pentru normele unitate introduse în (4.77) și (4.78) se poate da reprezentarea geometrică din fig. 4.13.

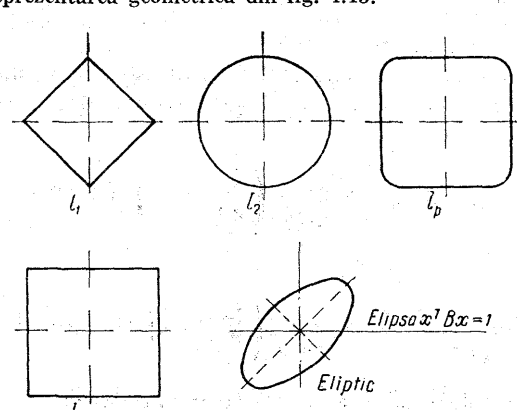


Fig. 4.13

Între normele definite anterior au loc [42] următoarele inegalităţi şi proprietăţi :

$$\bullet\ \|\mathbf{x}\|_\infty \leqslant \|\mathbf{x}\|_2 \leqslant \|\mathbf{x}\|_1, \qquad \|\mathbf{x}\|_\infty \leqslant \|\mathbf{x}\|_2 \leqslant \sqrt{n}\|\mathbf{x}\|_\infty, \tag{4.79}$$

$$\|\mathbf{x}\|_\infty \leqslant \|\mathbf{x}\|_1 \leqslant n\|\mathbf{x}\|_\infty, \qquad n^{-1/2}\|\mathbf{x}\|_1 \leqslant \|\mathbf{x}\|_2 \leqslant \|\mathbf{x}\|_1.$$

● Orice normă vectorială este o funcţie continuă de componentele vectorului considerat.

● Convergenţa într-o anumită normă implică convergenţa în orice altă normă, ca o consecinţă a teoremei de echivalenţă a normei.

● Fie $\|\cdot\|_\alpha$ şi $\|\cdot\|_\beta$ două norme oarecare pe R^n (sau C^n); atunci există două constante $M \geqslant m \geqslant 0$, astfel că

$$m\|\mathbf{x}\|_\alpha \leqslant \|\mathbf{x}\|_\beta \leqslant M\|\mathbf{x}\|_\alpha, \quad M > m > 0, \tag{4.80}$$

oricare ar fi $\mathbf{x}$.

Pentru a demonstra relaţia (4.80) este suficient să presupunem că $\|\cdot\|_\beta$ este norma l_2. Atunci au loc două relaţii

$$a_1\|\mathbf{x}\|_\alpha \leqslant \|\mathbf{x}\|_2 \leqslant a_2\|\mathbf{x}\|_\alpha, \quad a_1'\|\mathbf{x}\|_\beta \leqslant \|\mathbf{x}\|_2 \leqslant a_2'\|\mathbf{x}\|_\beta,$$

de unde se vede că $m = a_1/a_2'$ şi $M = a_2/a_1'$.

● Fie $\|\cdot\|$ o normă arbitrară pe R^n (C^n) şi A o matrice arbitrară nesingulară din $R^{n\times n}$($C^{n\times n}$). Atunci $\|\mathbf{x}\|' = \|A\mathbf{x}\|$ defineşte o normă pe $R^n(C^n)$.

Tipurile de norme matriceale consistente (compatibile) cu normele vectoriale (4.77) pot fi definite cu ajutorul relaţiilor :

$$\|A\|_1 = \sup_{\|x\|_1=1} \|A\mathbf{x}\|_1 = \max_{1\leqslant j\leqslant n}\left(\sum_{i=1}^{n} |a_{ij}|\right) \text{ (normă sumă-coloană)}, \tag{4.81}$$

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|A\mathbf{x}\|_\infty = \max_{1\leqslant i\leqslant n}\left(\sum_{j=1}^{n} |a_{ij}|\right) \text{ (normă sumă-linie)},$$

$$\|A\|_2 = \max_{\|x\|_2=1} \|A\mathbf{x}\|_2 = \sqrt{\lambda_1} \text{ (normă spectrală)}. \tag{4.81}$$

Pentru claritate se prezintă dezvoltat $\|A\|_1$ și $\|A\|_\infty$, fapt care impune examinarea dezvoltată a normelor $\|Ax\|_1$ și $\|Ax\|_\infty$. Deci

$$Ax = \begin{bmatrix} a_{11}\, a_{12} \dots a_{1n} \\ a_{21}\, a_{22} \dots a_{2n} \\ \vdots \\ a_{i1}\, a_{i2} \dots a_{in} \\ \vdots \\ a_{n1}\, a_{n2} \dots a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} a_{1j}x_j \\ \sum_{j=1}^{n} a_{2j}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{ij}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{nj}x_j \end{bmatrix}. \quad (4.82)$$

Considerăm norma l_1, $\|A\|_1$ pentru orice $\mathbf{x} \in R^n$:

$$\|Ax\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \leqslant \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|\,|x_j| =$$

$$= \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}| \leqslant \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |a_{ij}|\, \|\mathbf{x}\|_1. \quad (4.83)$$

Fie ca

$$\max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} |a_{ij}| = \sum_{i=1}^{n} |a_{ik}|$$

și e^k este vectorul unitate iar $a^{(k)}$ este vectorul format din coloana k a matricei A, deci rezultă

$$\|Ae^{(k)}\|_1 = \|a_k\| = \sum_{i=1}^{n} |a_{ik}|.$$

Dacă se consideră $\|A\|_\infty$, atunci pentru orice $\mathbf{x} \in R^n$

$$\|A\mathbf{x}\|_\infty = \sum_{j=1}^{n} \left| \sum_{i=1}^{n} a_{ij} x_j \right| \leqslant \sum_{j=1}^{n} \sum_{i=1}^{n} |a_{ij}|\,|x_j| =$$

$$= \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}| \leqslant \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |a_{ij}|\, \|\mathbf{x}\|_\infty .$$

Dacă pentru $i = k$ are loc egalitatea

$$\max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{kj}|,$$

atunci $\|A\|_\infty$ este descrisă ca suma elementelor liniei k din matricea A, sumă care este maximă în valoare absolută față de sumele realizate cu elementele celorlalte linii. Datorită acestui fapt $\|A\|_\infty$ se mai numește și *norma dată de suma elementelor unei linii*, iar norma $\|A\|_1$ ca *norma dată de suma elementelor unei coloane*. Pentru a vedea ce devine $\|A\|_2$ se folosește faptul că

$$\|\mathbf{x}\|_2 = \sqrt{(\mathbf{x}, \mathbf{x})} = \sqrt{\mathbf{x}^H \mathbf{x}} = \sqrt{\sum_{i=1}^{n} |x_i|^2}. \qquad (4.84)$$

Deoarece $A\mathbf{x}$ este un vector, folosind relația (4.84), rezultă

$$\|A\|_2 = \sup_{\|\mathbf{x}\|_2 = 1} \|A\mathbf{x}\|_2 = \sup_{\|\mathbf{x}\|_2 = 1} \sqrt{(A\mathbf{x}, A\mathbf{x})} = \sup_{\|\mathbf{x}\|_2 = 1} \sqrt{\mathbf{x}^H A^H A \mathbf{x}}\,.$$

Dar matricea $A^H A$ este hermitiană și nenegativ definită. Atunci valorile sale proprii sînt reale și nenegative $0 \leqslant \lambda_n \leqslant \ldots \leqslant \lambda_2 \leqslant \lambda_1$. De asemenea, $A^H A$ are un sistem de vectori proprii complet și ortonormal $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$ care formează o bază pentru spațiul vectorial C^n. Atunci orice vector $\mathbf{x}$ cu $\|\mathbf{x}\|_2 = 1$, dar arbitrar, se poate exprima ca o combinație liniară de vectori proprii care constituie o bază pe C^n, adică

$$\mathbf{x} = \sum_{j=1}^{n} a_j \mathbf{v}^{(j)}, \qquad (4.85)$$

de unde, deoarece $(\mathbf{x}, \mathbf{x}) = 1$, rezultă $\sum_{j=1}^{n} |a_j|^2 = 1$. Astfel

$$A^H A\mathbf{x} = A^H A\left[\sum_{j=1}^{n} a_j \mathbf{v}^{(j)}\right] = \sum_{j=1}^{n} a_j A^H A\mathbf{v}^{(j)} =$$

$$= \sum_{j=1}^{n} a_j \lambda_i \mathbf{v}^{(j)} \tag{4.86}$$

Din faptul că vectorii $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$ sînt ortonormali și din relațiile (4.85) și (4.86) rezultă

$$(A\mathbf{x}, A\mathbf{x}) = \mathbf{x}^H(A^H A\mathbf{x}) = \sum_{j=1}^{n} |a_j|^2 \lambda_j \leqslant \lambda_1 \sum_{j=1}^{n} |a_j|^2 = \lambda_1.$$

Dacă $\mathbf{x}$ este considerat a fi vectorul $\mathbf{v}^{(1)}$ corespunzător valorii proprii maxime $[\lambda_1 = \max_i (\lambda_i)]$, atunci

$$(A\mathbf{v}^{(1)}, A\mathbf{v}^{(1)}) = (\mathbf{v}^{(1)}, A^H A\mathbf{v}^{(1)}) = (\mathbf{v}^{(1)}, \lambda_1 \mathbf{v}^{(1)}) = \lambda_1(\mathbf{v}^{(1)}, \mathbf{v}^{(1)}) = \lambda_1$$

și astfel are loc egalitatea.

Deoarece $\|A\|_2$ este maximul din $\sqrt{(A\mathbf{x}, A\mathbf{x})}$ pentru $\|\mathbf{x}\|_2 = 1$, rezultă că

$$\|A\|_2 = \sqrt{\lambda_1}, \tag{4.87}$$

adică $\|A\|_2$ se poate descrie ca rădăcina pătrată pozitivă din valoarea proprie cea mai mare a matricei $A^H A$; de asemenea $\|A\|_2$ se mai denumește și norma spectrală a lui A [42, 108].

● Norma matriceală are o interpretare geometrică, adică $\|A\|_2$ este lungimea maximă a unui vector după transformarea prin A.

Figura 4.14 indică o reprezentare pentru R^2, $\|A\|_2$ în cazul euclidian este lungimea axei mari a elipsei $\{A\mathbf{x} : \|\mathbf{x}\|_2 = 1\}$.

Norma matriceală euclidiană este definită [108] prin relaţia

$$\|A\|_E = \left(\sum_{j=1}^{n} \sum_{i=1}^{n} |a_{ij}|^2 \right)^{1/2}. \tag{4.88}$$

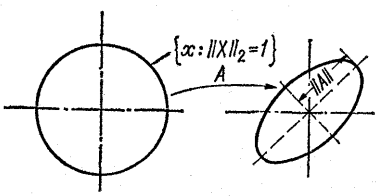


Fig. 4.13.

$\|A\|_E$ este o normă matriccală consistentă cu $\|\mathbf{x}\|_2$ (norma vectorială euclidiană), dar $\|A\|_E$ nu este subordonată normei $\|\mathbf{x}\|_2$. De asemenea se poate afirma că $\|A\|_E$ nu este subordonată la vreo normă vectorială. Pentru a arăta acest lucru, presupunem că există norma vectorială $\|\mathbf{x}\|_\beta$, astfel că $\|A\| = \sup_{\|\mathbf{x}\|_\beta = 1} \|A\mathbf{x}\|$. Fie $A = I$; atunci $\|I\|_E = \sup_{\|\mathbf{x}\|_\beta = 1} \|I\mathbf{x}\| = 1$, dar $\|I\|_E = \sqrt{n}$, de unde rezultă contradicţia.

Faptul că norma matriceală euclidiană nu este norma care se subordonează normei vectoriale euclidiene are o serie de consecinţe:

● În foarte multe lucrări se utilizează mai mult $\|A\|_2$ decît $\|A\|_E$, deoarece $\|A\|_2$ este subordonată normei $\|\mathbf{x}\|_2$.

● În calcule numerice şi analiza erorilor se utilizează $\|A\|_E$ fiind preferată faţă de $\|A\|_2$ pentru că:

a) $\|A\|_E$ este mai uşor de calculat decît $\|A\|_2$;

b) $\|A\|_E$ pentru $A = (a_{ij})$ este aceeaşi cu norma lui $|A| = (|a_{ij}|)$.

Normele $\|A\|_E$ şi $\|A\|_2$ sînt ambele consistente cu $\|\mathbf{x}\|_2$ şi uneori $\|A\|_E$ este o aproximaţie acceptabilă a lui $\|A\|_2$.

Norma euclidiană satisface [50, 128] inegalitățile

1) $\|A\|_2 \leqslant \|A\|_E \leqslant \sqrt{n}\,\|A\|_2$;

2) $\|A\|_2 \leqslant \||A|\|_2 \leqslant \||A|\|_E = \|A\|_E \leqslant \sqrt{n}\,\|A\|_2$.

Fie λ o valoare proprie a lui $A \in C^{n\times n}$; atunci $|\lambda| \leqslant \|A\|_\alpha$ pentru orice normă matriceală (raza spectrală a matricei A este mărginită de norma lui A).

Folosind ecuația valorilor și vectorilor proprii $\lambda\mathbf{x} = A\mathbf{x}$ și alegînd norma matriceală $\|\cdot\|_\alpha$ care este consistentă cu norma vectorială $\|\cdot\|_\beta$, se poate scrie

$$|\lambda|\,\|\mathbf{x}\|_\beta \leqslant \|A\|_\alpha \|\mathbf{x}\|_\beta,$$

de unde rezultă $|\lambda| \leqslant \|A\|_\alpha$. Deci pentru orice valoare proprie, raza spectrală a lui A este mărginită pentru fiecare normă a lui A.

Folosindu-se aplicația liniară $\mathbf{y} = A\mathbf{x}$ a unui vector $\mathbf{x}$ din U într-un vector $\mathbf{y}$ din V, se poate da o generalizare a normei matriceale:

$$\|A\|_\alpha = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{y}\|_\beta}{\|\mathbf{x}\|_\beta} = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_\beta}{\|\mathbf{x}\|_\beta}$$

pentru o normă vectorială arbitrară $\|\cdot\|_\alpha$. Se observă că $\|\cdot\|_\alpha$ este normă vectorială pentru U și V. Dacă, pe de altă parte, se alege o normă vectorială $\|\cdot\|_s$ pe U și o normă vectorială $\|\cdot\|_r$ pe V, atunci se poate defini o normă matriceală generală astfel:

$$\|A\|_{s,r} = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{y}\|_r}{\|\mathbf{x}\|_s} = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_r}{\|\mathbf{x}\|_s},$$

care este echivalentă cu $\|A\|_{r,s} = \sup_{\|\mathbf{x}\|_r=1} \|A\mathbf{x}\|_s$. Pentru cazul în care $r = s$, se obține norma matriceală subordonată ca un caz particular. Schema logică din fig. 4.15 și programul 4.1 scris în FORTRAN prezintă un exemplu de

calcul al normelor $\|A\|_1$, $\|B\|_1$; $\|A\|_E$ $\|B\|_E$ şi $\|A\|_\infty$, $\|B\|_\infty$, unde

$$A = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 1 \\ -1 & -1 & 2 \end{bmatrix}$$

folosind relaţiile din (4.77).

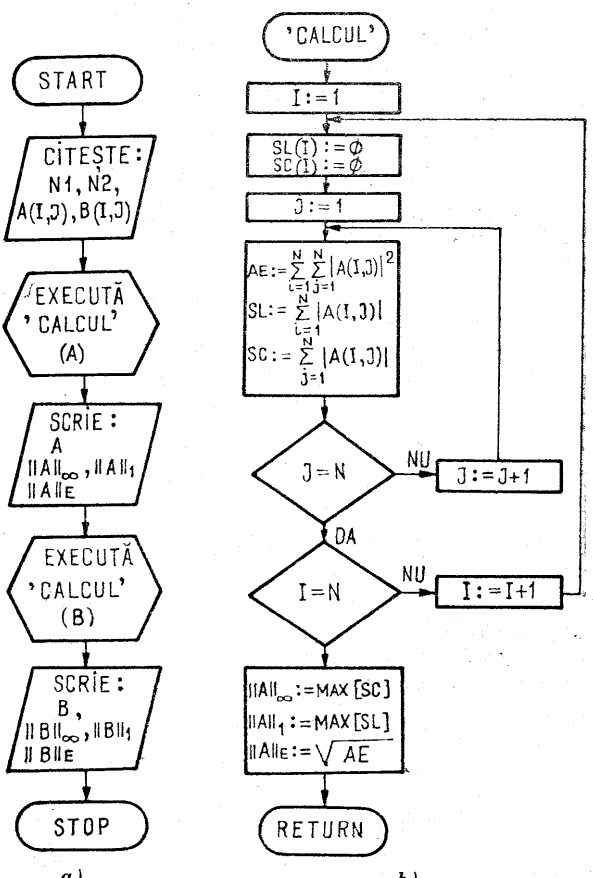


Fig. 4.15

```
C   CALCULUL NORMELOR MATRICILOR : A , B
      DIMENSION A(2,2),B(3,3)
      REAL MSL,MSC
      DATA MSL/-1000./,MSC/-1000./
      READ(105,100) N1,((A(I,J),J=1,2),I=1,2),N2,((B(I,J),J=1,3),I=1,3)
100   FORMAT(I5,4(F5.0),I5,9(F5.0))
      CALL CALCUL (N1,A,AE,MSL,MSC)
      WRITE(108,101) AE,MSL,MSC
101   FORMAT(' ',' NORMELE MATRICII A SINT : ',2(F5.2,','),F5.2)
      CALL CALCUL(N2,B,AE,MSL,MSC)
      WRITE(108,102) AE,MSL,MSC
102   FORMAT(//,' NORMELE MATRICII B SINT : ',2(F5.2,','),F5.2)
      STOP
      END

      SUBROUTINE CALCUL(N,A,AE,MSL,MSC)
      DIMENSION A(N,N),SL(3),SC(3)
      INTEGER SL,SC
      REAL MSL,MSC
      AE=0
      DO 20 I=1,N
      SL(I)=0
      SC(I)=0
      DO 20 J=1,N
      AE=AE+(ABS(A(I,J)))**2
      SL(I)=SL(I)+ABS(A(I,J))
20    SC(I)=SC(I)+ABS(A(J,I))
      AE=SQRT(AE)
      DO 21 I=1,N
21    IF(SL(I).GT.MSL) MSL=SL(I)
      IF(SC(I).GT.MSC) MSC=SC(I)
      RETURN
      END
          LINK
          RUN

        NORMELE MATRICII A SINT : 5.48 , 7.00 , 5.00
        NORMELE MATRICII B SINT : 3.00 , 7.00 , 5.00

          EOJ
```

Programul 4.1

4.9. Convergența vectorială și matriceală

Convergența vectorială și matriceală apare în foarte multe probleme de analiză a erorilor, a stabilității și consistenței metodelor numerice de calcul.

● Un șir de vectori $\{\mathbf{x}^{(i)}\} \in C^n$ este convergent către vectorul $\mathbf{x}$ cînd i tinde la ∞, adică

$$\mathbf{x}^{(i)} \to \mathbf{x} \text{ pentru } i \to \infty \text{ sau } \lim_{i\to\infty} \mathbf{x}^{(i)} = \mathbf{x} \qquad (4.89)$$

dacă și numai dacă $x_k^{(i)} \to x_k$ cînd $i \to \infty$ pentru orice k, adică cele n șiruri formate din componentele șirului de vectori $\mathbf{x}^{(i)}$ să aibă ca limită componentele vectorului $\mathbf{x}$ (se mai numește și convergența componentelor).

Şirul de matrice din C^{nn}

$$\{A^{(k)}\}_{k\in N} = A^{(1)}, A^{(2)}, \ldots, A^{(k)}, \ldots$$

converge către matricea $A \in C^{nn}$, cînd $k \to \infty$, adică $A^{(k)} \to A$ $(k \to \infty)$ sau $\lim\limits_{k\to\infty} A^{(k)} = A$ dacă şi numai dacă are loc pentru orice i şi j, adică dacă cele n^2 şiruri formate cu elementele a_{ij} ale şirului de matrice $\{A^{(k)}\}_{k\in N}$ au ca limită cele n^2 elemente ale matricei A (convergenţa este la nivel de element).

● Dacă se consideră seria infinită de vectori din C^n,

$$\sum_{i=1}^{\infty} \mathbf{x}^{(i)} = \mathbf{x}^{(1)} + \mathbf{x}^{(2)} + \ldots + \mathbf{x}^{(i)} + \ldots \qquad (4.90)$$

împreună cu şirul sumelor parţiale

$$\begin{aligned} S^{(1)} &= \mathbf{x}^{(1)}, \\ S^{(2)} &= \mathbf{x}^{(1)} + \mathbf{x}^{(2)}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ S^{(i)} &= \mathbf{x}^{(1)} + \mathbf{x}^{(2)} + \ldots + \mathbf{x}^{(i)}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \end{aligned}$$

atunci se poate defini convergenţa unei serii infinite de vectori astfel: seria infinită de vectori (4.90) converge către vectorul $\mathbf{S}$, dacă şi numai dacă şirul sumelor parţiale $\{\mathbf{s}^{(i)}\}$ converge la $\mathbf{s}$, adică

$$\mathbf{S}^{(i)} \to \mathbf{S} \quad (i \to \infty) \text{ sau } \sum_{i=1}^{\infty} \mathbf{x}^{(i)} = \mathbf{S}.$$

● Dacă se consideră seria infinită de matrice din C^{nn}

$$\sum_{k=1}^{\infty} A^{(k)} = A^{(1)} + A^{(2)} + \ldots + A^{(k)} + \ldots \qquad (4.91)$$

împreună cu şirul sumelor parţiale

$$S^{(1)} = A^{(1)},$$
$$S^{(2)} = A^{(1)} + A^{(2)},$$
$$\cdots\cdots\cdots\cdots$$
$$S^{(k)} = A^{(1)} + A^{(2)} + \ldots + A^{(k)},$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$

atunci seria (4.91) converge către matricea $S \in C^{nn}$ dacă şi numai dacă şirul sumelor parţiale $\{S^{(k)}\}_{k \in N}$ converge către S, adică

$$\text{sau} \quad S^k \to S \ (k \to \infty) \quad \text{sau} \quad \sum_{k=1}^{\infty} A^{(k)} = S.$$

Cu ajutorul normei vectoriale şi matriceale se poate prezenta o serie de criterii de convergenţă către matricea nulă 0, respectiv vectorul nul $\mathbf{0}$.

● Dacă şirul vectorial $\{\mathbf{x}^{(i)}\}_{i \in N}$ converge către vectorul nul $\mathbf{0}$ ($\mathbf{x}^{(i)} \to \mathbf{0}$, $i \to \infty$), atunci $\|\mathbf{x}^{(i)}\|_\beta \to 0$ pentru orice normă vectorială $\|\cdot\|_\beta$ definită în 4.8. Adesea această convergenţă este referită ca o convergenţă cu privire la norma $-\beta$ sau convergenţă în norma $-\beta$.

Pentru a demonstra convergenţa către vectorul nul $\mathbf{0}$ se consideră vectorul $\mathbf{x}^{(i)}$, exprimat sub forma unei combinaţii liniare de baza unitară $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \ldots, \mathbf{e}^{(n)}\}$ astfel:

$$\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_k^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix} = x_1^{(i)}\mathbf{e}^{(1)} + x_2^{(i)}\mathbf{e}^{(2)} + \ldots + x_n^{(i)}\mathbf{e}^{(n)} = \sum_{k=1}^{n} x_k^{(i)}\mathbf{e}^{(k)}. \tag{4.92}$$

Aplicînd orice normă vectorială definită în 4.8, rezultă

$$\|\mathbf{x}^{(i)}\|_\beta = \left\| \sum_{k=1}^{n} x_k^{(i)}\mathbf{e}^{(k)} \right\|_\beta \leqslant \sum_{k=1}^{n} |x_k^{(i)}| \cdot \|\mathbf{e}^{(k)}\|_\beta \leqslant p \max_k |\mathbf{x}_k^{(i)}|, \tag{4.93}$$

unde p este o constantă (pentru norma considerată):

$$p = \sum_{k=1}^{n} \| \mathbf{e}^{(k)} \|_\beta .$$

Presupunînd că $\mathbf{x}^{(i)} \to \mathbf{0}$ cînd $i \to \infty$, înseamnă că pentru orice $k = 1, 2, \ldots, n$, $x_k^{(i)} \to 0$; în consecinţă $\max_k |x_k^{(i)}| \to 0$, cînd $i \to \infty$. Atunci din (4.93) se vede că

$$\| \mathbf{x}^{(i)} \|_\beta \to 0, \quad i \to \infty. \tag{4.94}$$

● Dacă şirul $\{\mathbf{x}^{(i)}\} \to \mathbf{0}$ $(i \to \infty)$, atunci pentru orice normă vectorială $\| \cdot \|_\beta$ rezultă

$$\| \mathbf{y} + \mathbf{x}^{(i)} \|_\beta \to \| \mathbf{y} \|_\beta, \quad i \to \infty. \tag{4.95}$$

Relaţia (4.95) pune în evidenţă proprietatea de continuitate a normei.

● Fie $\| \cdot \|_\beta$ orice normă vectorială definită pe spaţiul vectorial n dimensional C^n; atunci există două constante pozitive m şi M, independente de $\mathbf{x}$ astfel că

$$m \max_k |x_k| \leqslant \| \mathbf{x} \|_\beta \leqslant M \max_k |x_k|$$

sau, altfel scris,

$$m \| \mathbf{x} \|_\infty \leqslant \| \mathbf{x} \|_\beta \leqslant M \| \mathbf{x} \|_\infty . \tag{4.96}$$

Relaţia (4.96) evidenţiază o relaţie de comparare a normelor vectoriale.

● Convergenţa componentelor în cazul şirului de vectori este echivalentă cu convergenţa în orice normă vectorială.

● Dacă şirul de matrice $\{A^{(k)}\}_{k \in N}$ converge către matricea nulă $[A^{(k)} \to 0, \ k \to \infty]$, atunci $\| A^{(k)} \|_\alpha \to 0$ pentru orice normă matriceală $\| \cdot \|_\alpha$. Această convergenţă este referită în [127, 124], drept convergentă cu privire la norma α sau convergenţa în norma $-\alpha$.

● Dacă $\{A^{(k)}\}_{k\in N}$ este un șir de matrice astfel că $A^{(k)} \to 0$, $k \to \infty$, atunci pentru orice normă matriceală $\|\cdot\|_\alpha$ are loc relația

$$\|A^{(k)} + B\|_\alpha \to \|B\|_\alpha,\ k \to \infty. \tag{4.97}$$

Această relație evidențiază continuitatea normei matriceale.

● Fie $\|\cdot\|_\alpha$ orice normă matriceală definită pe C^{nn}; există două constante pozitive p și q independente de matricea A, astfel că $p \max\limits_{i,j} |a_{ij}| \leqslant \|A\|_\alpha \leqslant q \max\limits_{i,j} |a_{ij}|$ pentru orice $A \in C^{nn}$.

Convergența la nivelul elementelor unei matrice [108] este echivalentă cu convergența normei matriceale.

În foarte multe aplicații apar probleme de convergență în general nu numai către vectorul nul sau matricea nulă, fapt pentru care în continuare se vor enunța o serie de definiții privind convergența în general.

● Vectorul $\mathbf{x}^{(i)} \to \mathbf{x}$ cînd $i \to \infty$, dacă și numai dacă $\mathbf{x}^{(i)} - \mathbf{x} \to \mathbf{0}$.

● Vectorul $\mathbf{x}^{(i)} \to \mathbf{x}$, cînd $i \to \infty$, dacă și numai dacă $\|\mathbf{x}^{(i)} - \mathbf{x}\|_\beta \to 0$ pentru orice normă vectorială $\|\cdot\|_\beta$.

● Dacă $\mathbf{x}^{(i)} \to \mathbf{x}$, cînd $i \to \infty$, atunci $\|\mathbf{x}^{(i)}\|_\beta \to \|\mathbf{x}\|_\beta$ pentru orice normă vectorială.

● Matricea $A^{(k)} \to A$, cînd $k \to \infty$, dacă și numai dacă $A^{(k)} - A \to 0$.

● $A^{(k)} \to A$, cînd $k \to \infty$, dacă și numai dacă $\|A^{(k)} - A\|_\beta \to 0$ pentru orice normă matriceală $\|\cdot\|_\beta$.

● Dacă $A^{(k)} \to A$, cînd $k \to \infty$, atunci $\|A^{(k)}\|_\beta \to \|A\|_\beta$ pentru orice normă matriceală.

Demonstrațiile relativ la convergența în general se găsesc în [124, 106, 86, 108].

METODE DE CALCUL PENTRU REZOLVAREA SISTEMELOR DE ECUAȚII LINIARE

5.1. Introducere

5.1.1. *Generalități*

Rezolvarea sistemelor algebrice liniare și operațiile de calcul numeric matriceal (evaluarea determinanților, inversarea matriceală, calculul valorilor și vectorilor proprii) sînt incluse în domeniul algebrei liniare. Experiența arată că în diverse procese de calcul algebra liniară este implicată în procentul de 70% în problemele științifice. În acest sens se pot da cîteva exemple:

- Problemele care depind de un număr finit de grade de libertate, cazuri continue reprezentate prin ecuații diferențiale ordinare sau ecuații cu derivate parțiale sînt în mod comun transformate, cu ajutorul diferențelor finite, în sisteme de ecuații liniare.
- Aproximarea problemelor neliniare sînt frecvent soluționate prin procese de liniarizare, prin urmare, din nou se apelează la domeniul algebrei liniare.
- Programarea liniară, domeniu ce se ocupă cu minimizarea costurilor și eforturilor, a unor fenomene, implică rezolvarea unor sisteme de ecuații algebrice liniare.
- Foarte multe probleme inginerești din domeniul rețelelor electrice, analiza structurilor, proiectarea clădirilor, vapoarelor, avioanelor, transportul lichidelor și gazelor prin conducte etc. necesită pentru soluționarea rezolvarea unor sisteme de ecuații algebrice liniare.

Exemple 1. Se consideră structura din fig. 5.1, structură încărcată conform desenului :

p este sarcină uniform distribuită (unități de forță/lungime);
F_1, F_2 sînt două forțe laterale (unități de forță);
l este lungimea elementelor structurii;
I reprezintă momentul de inerție ale elementelor structurii (se consideră că este același pentru toate elementele structurii);
E reprezintă modulul de elasticitate al materialului din care este confecționată structura.

După încărcarea structurii cu sarcina distribuită p și a forțelor F_1 și F_2 se determină unghiurile de rotație φ_1 și deplasările orizontale δ_1 și δ_2.

În fig. 5.1 se prezintă forma structurii (punctat) după acțiunea sarcinilor p și F_1, F_2.

Există numeroase metode pentru rezolvarea acestei probleme, una din ele este *metoda pantei de deflexie* [32], care în final conduce la următorul sistem de ecuații algebrice, scris sub formă matriceală :

$$\begin{bmatrix} 4 & 1 & 0 & 1 & 0 & -\frac{3}{l} & \frac{3}{l} \\ 1 & 4 & 0 & 0 & 1 & -\frac{3}{l} & \frac{3}{l} \\ 0 & 0 & 4 & 1 & 0 & 0 & -\frac{3}{l} \\ 1 & 0 & 1 & 4 & 1 & -\frac{3}{l} & 0 \\ 0 & 1 & 0 & 1 & 6 & -\frac{3}{l} & 0 \\ 1 & 1 & 0 & 1 & 1 & -\frac{4}{l} & \frac{4}{l} \\ 0 & 0 & 1 & 1 & 1 & 0 & \frac{6}{l} \end{bmatrix} \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \delta_1 \\ \delta_2 \end{bmatrix} = \begin{bmatrix} \frac{pl^3}{24EI} \\ -\frac{Pl^3}{24EI} \\ -\frac{pl^3}{24EI} \\ 0 \\ -\frac{pl^3}{24EI} \\ \frac{F_1 l^2}{6EI} \\ \frac{(F_1+F_2)l^2}{6EI} \end{bmatrix}. \tag{5.1}$$

Rezolvarea acestui sistem conduce la determinarea unghiurilor de rotație φ_i ($i = 1,2,3,4,5$) și a deformărilor δ_1 și δ_2.

2. Se consideră un circuit electric format din cinci rezistențe R_1, R_2, R_3, R_4, R_5 și două surse de tensiune E_1, E_2. Se pune problema determinării curenților din laturi, indicați în fig. 5.2.

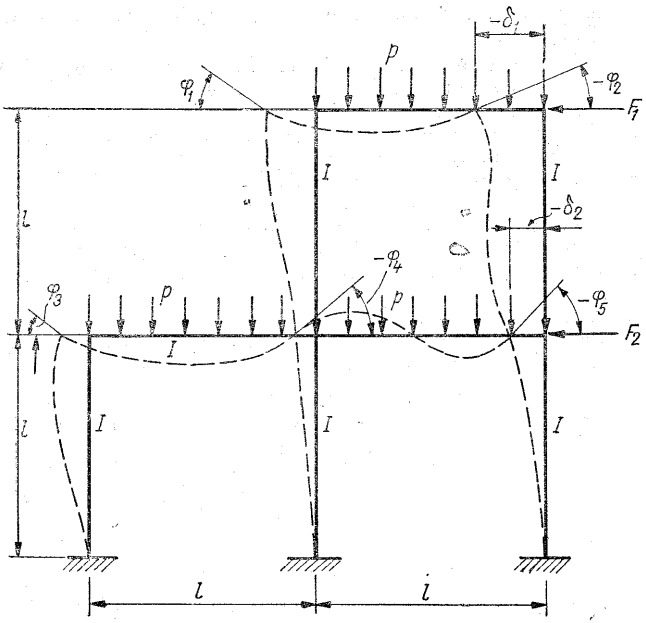


Fig. 5.1.

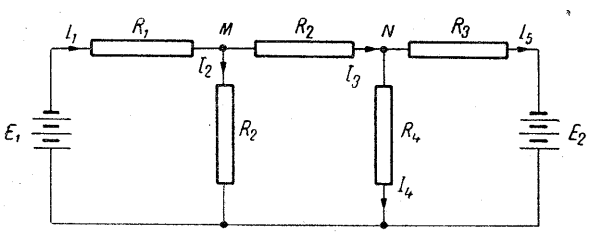


Fig. 5.2.

În acest sens se scriu ecuațiile lui Kirchoff pe cele trei ochiuri, precum și ecuația curenților în cele două noduri M și N, obținîndu-se un sistem de cinci ecuații cu cinci necunoscute, scris sub formă matriceală astfel:

$$\begin{bmatrix} R_1 & R_2 & 0 & 0 & 0 \\ 0 & -R_2 & R_3 & R_4 & 0 \\ 0 & 0 & 0 & -R_4 & R_5 \\ 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 \end{bmatrix} \cdot \begin{bmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{bmatrix} = \begin{bmatrix} E_1 \\ 0 \\ -E_2 \\ 0 \\ 0 \end{bmatrix} \cdot \qquad (5.2)$$

3. Se consideră aripa unui avion care este solicitată pe partea superioară de o forță a vîntului dată sub formă $F \sin \alpha t$, forță distribuită ca în fig. 5.3. În acest caz α reprezintă frecvența sarcinii aplicate. Se introduc următoarele mărimi :

$m_1, \ldots, m_n$ sînt masele concentrate respectiv în punctele $1,2,\ldots,n$;

$y_i = A_i \sin \alpha t$ este ecuația de mișcare a fiecărei mase m_i ($i = 1,2, \ldots$ $\ldots, n$), unde t reprezintă timpul, ω este frecvența în radiani/s, iar y_i deplasarea masei m_i.

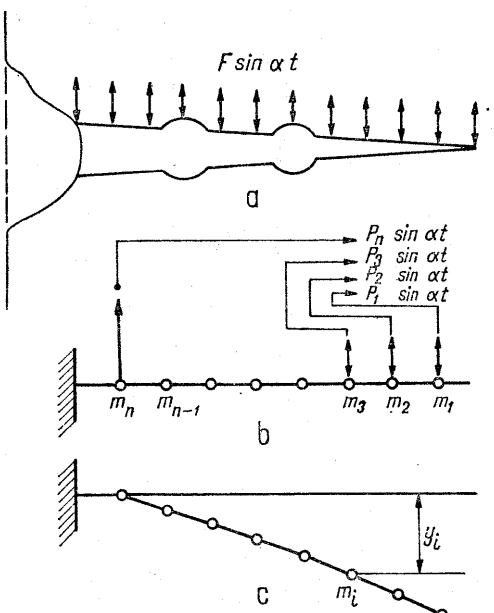


Fig. 5.3.

$a_{ij} (i = j = 1,2,\ldots,n)$ sînt coeficienți de influență,

$k = \dfrac{1}{\alpha^2}$, unde α este frecvența,

A_i amplitudinea mișcării pentru m_i, $i = 1,2,\ldots,n$.

Cu aceste elemente introduse și folosind legile de rezistență materialelor și legea lui Newton, rezultă următorul sistem de ecuații, care guvernează fenomenul considerat :

$$\begin{bmatrix} a_{11}m_1-k_1 & a_{12}m_2 & a_{13}m_3 & \dots & a_{1n}m_n \\ a_{21}m_1 & a_{22}m_2-k & a_{23}m_3 & \dots & a_{2n}m_n \\ a_{31}m_1 & a_{32}m_2 & a_{33}m_3-k & \dots & a_{3n}m_n \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ a_{n1}m_1 & a_{n2}m_2 & a_{n3}m_3 & \dots & a_{nn}m_n-k \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \cdot \\ \cdot \\ \cdot \\ A_n \end{bmatrix} =$$

$$= k \begin{bmatrix} \sum_{j=1}^{n} a_{1j}F_j \\ \sum_{j=1}^{n} a_{2j}F_j \\ \sum_{j=1}^{n} a_{3j}F_j \\ \cdot \\ \cdot \\ \cdot \\ \sum_{j=1}^{n} a_{nj}F_j \end{bmatrix} \qquad (5.3)$$

Sistemul (5.3) poate fi rezolvat în raport cu necunoscutele A_i, $i = 1,2,\dots,n$. Se știe că dacă frecvența α ia o valoare egală cu una din valorile frecvenței naturale ale aripei, atunci amplitudinile vor fi infinit de mari. Astfel de fenomen se numește *rezonanță*.

Fenomenele de vibrații pot fi întîlnite în multe alte sisteme astfel ca : vapoare, poduri, clădiri, mașini electrice etc., și multe fenomene din aceste domenii pot fi analizate într-o manieră asemănătoare, care în final conduce la rezolvarea unui sistem de ecuații algebrice. În general matricele asociate aplicațiilor de tipul dat în exemplele precedente prezintă o serie de caracteristici specifice ca structură și conținut.

5.1.2. *Sisteme de ecuații, interpretări geometrice*

Cuvîntul „liniar" este de obicei luat în sensul că variabilele apar la puterea întîi în fiecare termen al funcției. Astfel :

$$f(x, y) = a_1x + b_1y \text{ și } g(x, y) = b_1x + b_2y \qquad (5.4)$$

sînt două forme liniare.

În mod frecvent sînt utilizate pe scară largă următoarele două proprietăți ale funcțiilor liniare:

$$f(x_1 + x_3, x_2 + x_4) = f(x_1, x_2) + f(x_3, x_4),$$

$$cf(x_1, x_2) = f(cx_1, cx_2). \tag{5.5}$$

Aceste proprietăți de liniaritate ale unei funcții sau ecuații sînt utilizate în metodele de eliminare la rezolvarea numerică a sistemelor de ecuații liniare (în sensul că ecuațiile sistemului pot fi amplificate cu constante, combinate cu ajutorul operațiilor elementare, în scopul obținerii unei forme finale mult simplificate).

• Fie $A \in R^{nn}$, $\mathbf{b} \in R^n$ și vectorul necunoscut $\mathbf{x} \in R^n$; atunci un sistem de ecuații liniare se scrie sub una din formele:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned}\right\},$$

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}, \tag{5.6}$$

$$A\mathbf{x} = \mathbf{b},$$

sau

$$\sum_{i=1}^{n} a_{ij}x_j = b_i, \quad i = 1,2,\ldots,n.$$

• Un sistem de ecuații algebrice (5.6) este *consistent* dacă are cel puțin o soluție și este *inconsistent* dacă nu are nici o soluție [47, 86].

Pentru a pune în evidență consistența și neconsistența sistemelor de ecuații algebrice liniare se consideră următorul exemplu:

Fie $A \in R^{22}$, $b \in R^2$ și următoarele trei sisteme de ecuații:

$$\text{a)} \quad \begin{bmatrix} 2 & 2 \\ 4 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \end{bmatrix}, \qquad A: \begin{bmatrix} 5/2 \\ 1/2 \end{bmatrix} \to \begin{bmatrix} 6 \\ 8 \end{bmatrix} \in R^2;$$

$$\text{b)} \quad \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \end{bmatrix}, \qquad A: \exists\, \mathbf{x} \in R^2 \to \begin{bmatrix} 6 \\ 0 \end{bmatrix} \in R^2; \qquad (5.7)$$

$$\text{c)} \quad \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \end{bmatrix}, \qquad A: \forall\, \mathbf{x} \in R^2 \to \begin{bmatrix} 6 \\ 6 \end{bmatrix} \in R^2.$$

În fig. 5.4 se dă reprezentare geometrică a celor trei sisteme. Sistemul a) este reprezentat cu ajutorul a două drepte care se intersectează într-un singur punct $P(5/2,1/2)$ și este reprezentarea geometrică a soluției; deci sistemul a) este consistent și vectorul $\mathbf{x}^T = [5/2,\ 1/2]$ este o soluție unică a sistemului a).

Sistemul b) reprezintă două drepte paralele care nu au nici un punct de intersecție, fapt care demonstrează că sistemul b) este neconsistent.

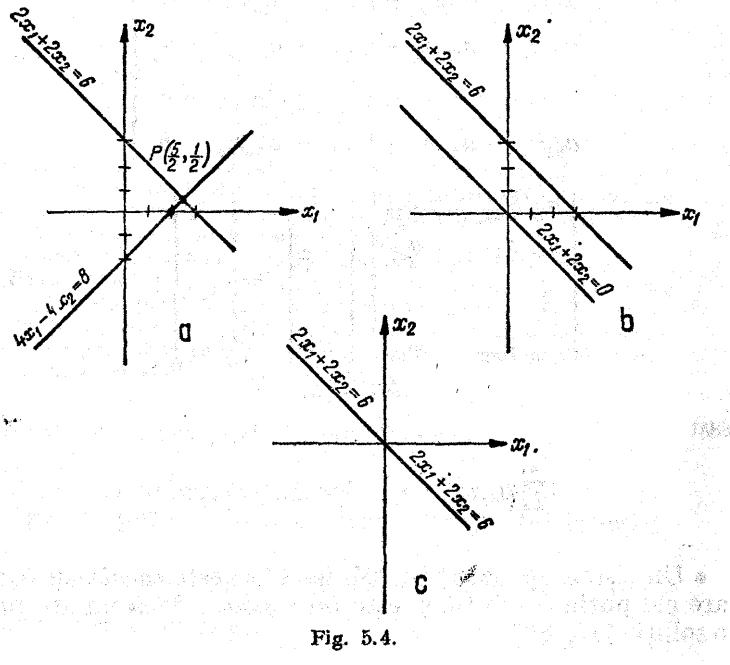


Fig. 5.4.

Sistemul c) reprezintă două drepte confundate (deoarece cele două ecuații sînt identice), deci ele se intersectează într-o infinitate de puncte, fapt care evidențiază consistența sistemului c).

În rezumat sistemul a) are o soluție unică, sistemul b) nu are nici o soluție, iar sistemul c) are o infinitate de soluții.

Orice sistem de ecuații algebrice care se va considera în continuare va avea o soluție, nici o soluție sau o infinitate de soluții (nu există alte posibilități).

• Sistemele de ecuații se pot clasifica și după vectorul $\mathbf{b}$ din (5.6) în :

a) *Sisteme omogene.* Dacă $\mathbf{b} = \mathbf{0}$, sistemul (5.6) este omogen.

Orice sistem omogen de forma $A\mathbf{x} = \mathbf{0}$ cu $A \in R^{nn}$, $\mathbf{x} \in R^n$, este un sistem consistent, deoarece are soluția $\mathbf{x} = \mathbf{0}$, neinteresantă.

Un sistem omogen are o soluție nebanală dacă și numai dacă $\det A = 0$, adică dacă A este singulară. Altfel sistemul nu admite o soluție unică (afară de soluția banală), soluția depinde de cel puțin un parametru, iar în practică este adesea necesar să se calculeze una sau mai multe valori ale unui parametru, care apare în coeficienți.

b) *Sisteme neomogene* dacă vectorul $\mathbf{b} \neq \mathbf{0}$. Dacă $A \in R^{nn}$ și $\mathbf{b} \in R^n$, atunci sistemul (5.6) are o soluție unică pentru orice $\mathbf{b} \neq \mathbf{0}$, dacă și numai dacă sistemul omogen $A\mathbf{x} = \mathbf{0}$ nu are altă soluție decît soluția banală (adică A este nesingulară).

În cazul sistemelor neomogene unde numărul ecuațiilor diferă de numărul necunoscutelor, se urmărește determinarea vectorului $\mathbf{x}$, care are ca imagine prin transformarea A vectorul $\mathbf{b}$.

Exemplu. Fie $A\mathbf{x} = \mathbf{b}$, $A \in R^{23}$, $\mathbf{b} \in R^2$:

$$\begin{bmatrix} 2 & -2 & -2 \\ 2 & 2 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 4 \end{bmatrix}; \qquad (5.8)$$

se urmărește determinarea vectorului $\mathbf{x} \in R^3$ care are ca imagine, prin transformarea A, vectorul $\begin{bmatrix} 0 \\ 4 \end{bmatrix} \in R^2$.

Sistemul (5.8) se poate scrie sub forma dezvoltată :

$$\left.\begin{aligned} 2x_1 - 2x_2 - 2x_3 &= 0 \\ 2x_1 + 2x_2 - 2x_3 &= 4 \end{aligned}\right\}.$$

Aceste ecuații reprezintă două plane P_1 și P_2 în R^3 (fig. 5.5). Cele două plane se intersectează după o dreaptă MN. Orice punct ce aparține dreptei MN reprezintă o soluție pentru sistemul (5.8), adică o *simplă infinitate de soluții.*

În cazul în care soluțiile unui anumit sistem de ecuații corespund punctelor unui plan, atunci sistemul considerat are o *dublă infinitate de soluții.*

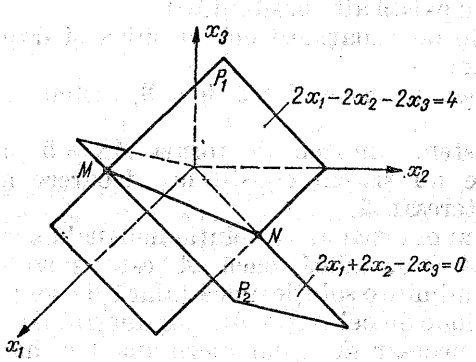


Fig. 5.5.

5.1.3. *Unicitatea și existența soluției unui sistem de ecuații*

Din exemplele și interpretarea geometrică, dată pentru sistemele de ecuații algebrice, rezultă necesitatea unui criteriu pentru evidențierea existenței soluției sistemului $\mathcal{A}\mathbf{x} = \mathbf{b}$, iar cînd soluția există, este necesar un criteriu pentru a arăta unicitatea ei.

Matricea $\mathcal{A}$ din (5.6) se numește matricea coeficient a sistemului, vectorul $\mathbf{b}$ se numește vectorul constant al sistemului.

Matricea de ordinul $n \times (n+1)$ formată din $\mathcal{A}$ și $\mathbf{b}$, $[\mathcal{A}, \mathbf{b}]$ se numește matricea bordată a sistemului (5.6), adică

$$[\mathcal{A}, \mathbf{b}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \ldots a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} \ldots a_{2n} & b_2 \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & a_{n3} \ldots a_{nn} & b_n \end{bmatrix}. \tag{5.9}$$

• Dacă $A \in R^{nn}$, $\mathbf{b} \in R^n$, sistemul de ecuații (5.6) este consistent dacă și numai dacă rang A=rang $[A, \mathbf{b}]$ (adică sistemul are cel puțin o soluție, deci soluția există).

• Dacă $A \in R^{nn}$, $\mathbf{b} \in R^n$ și rang A = rang $[A, \mathbf{b}] = k$, atunci:

— dacă $k = n$, $A\mathbf{x} = \mathbf{b}$ are o soluție unică;

— dacă $k < n$, $A\mathbf{x} = \mathbf{b}$ are $n - k$ familii de soluții parametrice.

Aceste două propoziții pot servi drept criterii pentru a arăta existența, respectiv unicitatea soluției unui sistem de tipul (5.6).

În concluzie, pentru un sistem (5.6) există următoarele situații: nu are soluție, are o soluție unică, are o infinitate de soluții.

Dacă sistemul are o infinitate de soluții, atunci diferența între două soluții oarecare aparține spațiului nul al matricei A, notat prin $K(A)$. În acest sens se poate afirma că sistemul (5.6) are o soluție unică dacă și numai dacă spațiul nul al matricei A conține doar vectorul nul, adică

$$K(A) = \{\mathbf{0}\}. \tag{5.10}$$

În cazul în care soluția sistemului (5.6) există și este unică, aceasta poate fi exprimată în două forme simple: cu ajutorul regulii lui Cramer și al matricei inverse, forme care sînt echivalente.

• Fie $A\mathbf{x} = \mathbf{b}$, cu $A \in R^{nn}$ și $\mathbf{b} \in R^n$, sistem care are o soluție unică. Atunci soluția sistemului este

$$x_1 = \frac{D_1}{D}, \quad x_2 = \frac{D_2}{D}, \ldots, x_j = \frac{D_j}{D}, \ldots, x_n = \frac{D_n}{D}, \tag{5.11}$$

unde determinanții D_i, $i = 1, 2, \ldots, n$, și D au forma

$$D_i = \det \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1,i-1} & b_1 & a_{1,i+1} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2,i-1} & b_2 & a_{2,i+1} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{n,i-1} & b_n & a_{n,i+1} & \ldots & a_{nn} \end{bmatrix} \text{ și } D = \det A. \tag{5.12}$$

• O altă formă simplă de exprimare a soluției se poate face cu ajutorul matricei inverse A^{-1}, astfel:

$$A\mathbf{x} = \mathbf{b}, A^{-1}A\mathbf{x} = A^{-1}\mathbf{b}, I\mathbf{x} = A^{-1}\mathbf{b}, \mathbf{x} = A^{-1}\mathbf{b}. \quad (5.13)$$

Considerînd că $A^{-1} = (c_{ij})$, $i,j = 1,2,\ldots,n$, există, ultima relație matriceală din (5.13) se poate scrie dezvoltată sub forma

$$x_i = \sum_{j=1}^{n} c_{ij} b_j, \; i = 1, 2,\ldots,n. \quad (5.14)$$

Relația (5.14) permite calculul tuturor componentelor vectorului necunoscut $\mathbf{x}^T = (x_1, x_2,\ldots, x_n)$. Considerînd elementele c_{ij} ale matricei A^{-1} cunoscute, se observă din (5.14) că pentru fiecare componentă x_i a vectorului $\mathbf{x}$ sînt necesare n operații de înmulțire și $n-1$ operații de adunare. Pentru calculul unei componente x_i a vectorului $\mathbf{x}$ sînt necesare $n+(n-1)$ operații elementare, deci pentru calculul vectorului $\mathbf{x} \in R^n$ sînt necesare n^2 operații de înmulțire și $n(n-1)$ operații de adunare.

5.1.4. *Condiționarea numerică a sistemelor liniare*

Fie sistemul de ecuații

$$A\mathbf{x} = \mathbf{b}, \quad (5.15)$$

unde A este o matrice nesingulară. Atunci sistemul are o soluție unică

$$\mathbf{x} = A^{-1}\mathbf{b}. \quad (5.16)$$

Trebuie avut în vedere faptul că în sistemele obținute din aplicațiile fizice numerele ce constituie matricea A și vectorul $\mathbf{b}$ sînt afectate de erori datorită măsurărilor sau erori de calcul datorită faptului că se lucrează cu un calculator electronic, care are o lungime a cuvîntului fixă (de exemplu numărul 1/7 nu poate fi reprezentat exact deoarece reprezentarea lui în binar are o infinitate de biți).

Datorită acestui fapt se pune problema care este toleranța rezultatelor, adică care este maximul de eroare admis în cadrul soluției. Apar următoarele elemente de discuție.

a) Dacă A este cunoscută exact și nu este afectată de erori, dar vectorul $\mathbf{b}$ este afectat, atunci se lucrează cu $\mathbf{b} + \delta\,\mathbf{b}$ și atunci vectorul soluție $\mathbf{x}$ este $\mathbf{x} + \delta\,\mathbf{x}$ iar sistemul (5.15) devine

$$A(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b} \text{ sau } A\mathbf{x} + A\delta\mathbf{x} = \mathbf{b} + \delta\mathbf{b}. \quad (5.17)$$

Folosindu-se relația (5.15), rezultă că toleranța din soluție este

$$\delta\mathbf{x} = A^{-1}\,\delta\mathbf{b} \text{ sau } \|\delta\mathbf{x}\| \leqslant \|A^{-1}\|\,\|\delta\mathbf{b}\|. \quad (5.18)$$

Egalitatea în a doua relație din (5.18) este posibilă pentru anumiți vectori $\delta\mathbf{b}$.

Aplicînd norma egalității (5.15), rezultă

$$\|\mathbf{b}\| \leqslant \|A\|\,\|\mathbf{x}\|. \quad (5.19)$$

Dacă se multiplică a doua relație din (5.18) prin $\|\mathbf{b}\|$, rezultă

$$\|\delta\mathbf{x}\|\,\|\mathbf{b}\| \leqslant \|A\|\,\|A^{-1}\|\,\|\mathbf{x}\|\,\|\delta\mathbf{b}\|. \quad (5.20)$$

Presupunînd că $\mathbf{b} \neq \mathbf{0}$, rezultă din (5.20) următoarea inegalitate:

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \|A\|\,\|A^{-1}\|\,\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}. \quad (5.21)$$

unde A^{-1} este matricea inversă a lui A, care a fost obținută prin calcule cu ajutorul unui calculator (deci $AA^{-1} \neq I$).

În acest sens pentru o matrice nesingulară A pozitiv definită se definește numărul de condiționare A astfel:

$$\text{cond.}\,(A) = \|A\|\,\|A^{-1}\| = \lambda_1/\lambda_n \geqslant 1, \quad (5.22)$$

unde λ_1 și λ_n sînt valoarea proprie maximă, respectiv minimă a lui A. În acest caz (5.21) se scrie sub forma

$$\frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \text{cond. } (A) \cdot \frac{\|\delta \mathbf{b}\|}{\|\mathbf{b}\|}, \tag{5.23}$$

unde $\frac{\|\delta \mathbf{b}\|}{\|\mathbf{b}\|}$ măsoară incertitudinea relativă existentă în vectorul $\mathbf{b}$ (dacă elementele lui $\mathbf{b}$ sînt cunoscute cu trei cifre semnificative, atunci $\|\delta \mathbf{b}\| / \|\mathbf{b}\|$ este aproximativ 10^{-3} sau 10^{-4}); $\frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|}$ măsoară incertitudinea relativă existentă în vectorul $\mathbf{x}$, care este determinată de incertitudinea existentă în vectorul $\mathbf{b}$.

b) Dacă și matricea A și vectorul $\mathbf{b}$ sînt afectați de erori, atunci sistemul (5.15) se scrie sub forma

$$(A + \delta A)(\mathbf{x} + \delta \mathbf{x}) = \mathbf{b} + \delta \mathbf{b}. \tag{5.24}$$

În acest caz eroarea existentă în vectorul $\mathbf{x}$ este $\delta \mathbf{x}$, exprimată prin relația

$$\delta \mathbf{x} = (A + \delta A)^{-1}(\delta \mathbf{b} - \delta A \mathbf{x}). \tag{5.25}$$

c) Dacă $\mathbf{b}$ este cunoscut exact iar matricea A este afectată de erori, atunci sistemul (5.15) devine $(A + \delta A)(\mathbf{x} + \delta \mathbf{x}) = \mathbf{b}$, de unde rezultă

$$\mathbf{x} + \delta \mathbf{x} = (A + \delta A)^{-1}\mathbf{b}. \tag{5.26}$$

Dacă se înlocuiește $\mathbf{x}$ prin expresia (5.16), relația (5.26) devine $A^{-1}\mathbf{b} + \delta \mathbf{x} = (A + \delta A)^{-1}\mathbf{b}$. După ordonare rezultă

$$\delta \mathbf{x} = [(A + \delta A)^{-1} - A^{-1}]\mathbf{b}. \tag{5.27}$$

În cazul în care se introduce notația $A + \delta A = C$ și se folosește [127, 52] identitatea

$$C^{-1} - A^{-1} = A^{-1}(A - C)C^{-1},$$

atunci (5.27) devine

$$\delta \mathbf{x} = A^{-1}(A - A - \delta A)(A + \delta A)^{-1}\mathbf{b} = -A^{-1}(\delta A)(A + \delta A)^{-1}\mathbf{b}$$

iar folosind (5.27), rezultă

$$\delta \mathbf{x} = -A^{-1}(\delta A)(\mathbf{x} + \delta \mathbf{x}). \tag{5.28}$$

Aplicînd norma relației (5.28), se obține

$$\|\delta \mathbf{x}\| \leqslant \|A^{-1}\| \, \|\delta A\| \, \|\mathbf{x} + \delta \mathbf{x}\|$$

sau

$$\frac{\|\delta \mathbf{x}\|}{\|\mathbf{x} + \delta \mathbf{x}\|} \leqslant \text{cond.}\,(A)\,\frac{\|\delta_A\|}{\|A\|}. \tag{5.29}$$

Din (5.29) se vede că incertitudinea din vectorul soluție $\mathbf{x}$ relativ la $\mathbf{x} + \delta \mathbf{x}$ este mărginită de incertitudinea relativă a matricei A, înmulțită cu numărul de condiționare a lui A [cond. (A)]. Deci, dacă erorile mici în cadrul coeficienților lui A și $\mathbf{b}$, sau în procesul de calcul au un efect redus asupra vectorului soluție, un astfel de sistem este *bine condiționat*, iar dacă efectul este considerabil, un astfel de sistem este *slab condiționat*.

Din analiza efectuată se vede că [relațiile (5.18) — —(5.29)] pentru mici variații în elementele matricei A (introduse prin matricea perturbațiilor δA) sau mici variații în elementele vectorului $\mathbf{b}$ (introduse prin vectorul perturbațiilor $\delta \mathbf{b}$) sau mici variații atît relativ la elementele lui A cît și relativ la componentele lui $\mathbf{b}$ se vor produce mici schimbări în valorile vectorului soluție exactă $\mathbf{x}$, dacă numărul de condiționare al matricei A este mic.

Dacă matricea A din sistemul (5.15) este nesingulară, atunci

$$\text{cond.}\,(A) = \|A\|_\alpha \, \|A^{-1}\|_\alpha \tag{5.30}$$

se numește numărul de condiționare pentru problema liniară $A\mathbf{x} = \mathbf{b}$, unde $\|\cdot\|$ este o normă matriceală oarecare.

Dacă $\|\cdot\|_\alpha$ este norma spectrală l_2, atunci

$$\text{cond. } (A) = \|A\|_2 \ \|A^{-1}\|_2 = \sqrt{\frac{\lambda_1}{\lambda_n}}, \qquad (5.31)$$

unde λ_1 și λ_n sînt valorile proprii maximă, respectiv minimă a matricei A, pozitiv definită.

În concluzie, dacă numărul de condiționare cond. (A) este mare, atunci perturbații reduse în A și/sau $\mathbf{b}$ pot introduce perturbații mari în $\mathbf{x}$ (problema liniară este slab *condiționată*). Pe de altă parte, dacă numărul de condiționare al matricei A este mic, atunci perturbații reduse în A și/sau $\mathbf{b}$ conduc la perturbații reduse în vectorul soluției $\mathbf{x}$ (problema liniară este *bine condiționată*).

Aceste elemente servesc la corectarea soluției obținute din calculul, la alegerea metodei de calcul (ținînd seama de numărul și ordinea operațiilor de executat), la modul de reprezentare a informației numerice în calculator (în virgulă fixă, virgulă mobilă sau zecimal) la tipul de aritmetică cerut în programul de calcul precum și la precizia impusă calculelor.

Pentru obținerea unor rezultate cît mai precise algoritmii de calcul trebuie să prezinte o *stabilitate numerică*, calitate care impune schimbarea liniilor matricei A și introducerea unui nou set de necunoscute $(x'_1, x'_2, \ldots, x'_n)$, pentru a permite și schimbarea coloanelor între ele în cadrul matricei A. Aceste schimbări se fac înaintea elaborării algoritmului de calcul sau în timpul desfășurării algoritmului de calcul. Un algoritm de calcul nestabil numeric poate conduce la rezultate eronate.

Este adevărat că metoda eliminării a lui Gauss cu pivotare parțială (adică pe parcursul derulării algoritmului) este mult mai stabilă decît algoritmul lui Gauss de eliminare fără pivotare parțială, cel puțin pentru o clasă de probleme. Pentru probleme liniare în care matricea A este reală, simetrică și pozitiv definită, pivotarea parțială nu este singura necesitate pentru stabilitatea numerică [1, 10, 14].

5.1.5. *Scalarea ecuațiilor și necunoscutelor în cadrul sistemelor*

În cadrul sistemelor de ecuații liniare, date în (5.1)—(5.3), se poate vedea în mod frecvent că necunoscutele

x_j (componentele vectorului necunoscut x) și componentele vectorului $\mathbf{b}$ (b_j, $j = 1,2,\ldots,n$), date sub formă numerică, au o semnificație fizică. De exemplu, curenți, respectiv tensiuni (5.2) sau deplasări măsurate în m sau forțe măsurate în N. Deoarece astfel de unități fizice sînt arbitrare și pot diferi foarte mult între ele din punctul de vedere al ordinului de mărime, în aceste cazuri se va face o substituire $x_j = 1{,}0^n x'_j$; atunci fiecare termen $a_{i,j}\, x^j$ va fi înlocuit prin $1{,}0^n a_{i,j} x'_j$ și astfel coloana j a matricei A va fi înmulțită cu $1{,}0^n$.

Se presupune în general că x_j se înlocuiește prin

$$x_j = n_j^{(2)} x'_j,\ j = 1,\ 2,\ldots,n. \tag{5.32}$$

Fie N_2 o matrice diagonală nesingulară de forma

$$N_2 = \begin{bmatrix} n_1^{(2)} & & & \\ & n_2^{(2)} & & 0 \\ & & \ddots & \\ & 0 & & n_n^{(2)} \end{bmatrix}. \tag{5.33}$$

Această substituție are forma

$$\mathbf{x} = N_2 \mathbf{x}'. \tag{5.34}$$

În mod asemănător se consideră matricea N_1, diagonală și nesingulară, avînd forma

$$N_1 = \begin{bmatrix} n_1^{(1)} & & & \\ & n_2^{(1)} & & 0 \\ & & \ddots & \\ & 0 & & n_n^{(1)} \end{bmatrix}, \tag{5.35}$$

care este folosită pentru a realiza substituția

$$\mathbf{b} = N_1 \mathbf{b}' \tag{5.36}$$

pentru vectorul $\mathbf{b}$ din partea dreaptă a sistemului de ecuații $A\mathbf{x}=\mathbf{b}$. Cu ajutorul acestor substituții (5.34) și (5.36) sistemul $A\mathbf{x} = \mathbf{b}$ devine

$$AN_2\mathbf{x}' = N_1\mathbf{b}' \text{ sau } N_1^{-1}AN_2\mathbf{x}' = \mathbf{b}'. \tag{5.37}$$

În urma celor două schimbări de variabile se obține un nou sistem liniar de ecuații (5.37), a cărui matrice este $A' = N_1^{-1}AN_2$, iar partea dreaptă este $\mathbf{b}' = N_1^{-1}\mathbf{b}$.

• Matricea A' este diagonal echivalentă cu matricea A, dacă există matricele diagonale nesingulare N_1 și N_2, astfel ca să aibă loc relația

$$A' = N_1^{-1}AN_2; \tag{5.38}$$

atunci $a'_{i,j} = [n_i^{(1)}]^{-1}\, a_{i,j} n_i^{(2)}$.

Echivalența între matricele A' și A este o echivalență într-un sens particular utilizat în cadrul teoriei matriceale (datorită faptului că matricele N_1, N_2 sînt matrice speciale nesingulare).

Se observă cu ușurință că produsul AN_2 reprezintă matricea A cu coloanele înmulțite prin constantele $n_1^{(2)}$, $n_2^{(2)}, \ldots, n_n^{(2)}$ iar produsul $N^{-1}A$ este matricea A cu liniile sale înmulțite cu constantele $1/n_2^{(1)}, \ldots, 1/n_n^{(1)}$. Matricea $A' = N_1^{-1}AN_2$ este rezultatul efectuării celor două operații de înmulțire a coloanelor și liniilor matricei A. Datorită proprietății de asociativitate a produsului matriceal, nu are importanța care înmulțire se execută înainte (la nivel de coloană sau la nivel de linie).

Matricea A' este echivalenta B-scalată cu matricea A dacă $A' = N_1^{-1}AN_2$, unde N_1 și N_2 sînt matrice diagonale nesingulare, ale căror elemente diagonale sînt toate puteri de numere întregi reprezentați în virgulă mobilă folosind baza de reprezentare B.

Avînd în vedere importanța pe care o joacă numărul de condiționare al unei matrice (cond. (A)) pentru precizie și stabilitatea numerică a algoritmului folosit la rezolvarea sistemului liniar $A\mathbf{x} = \mathbf{b}$, se pune întrebarea care este legătura dintre cond. (A) și cond. (A')? În acest sens se pune următoarea problemă: dîndu-se o matrice A, ce matrice N_1 și N_2 trebuie alese ca în final cond. $(A') =$ $=$ cond. $(N_1^{-1}AN_2)$ să fie minim [46, 126]?

Pentru realizarea acestui lucru mai este necesar ca matricea A să se echilibreze înainte de a se trece la rezolvarea sistemului. O matrice este *echilibrată* dacă atît liniile ei cît și coloanele au aproximativ aceeași lungime într-o normă oarecare. Astfel rezultă:

- Matricea A este echilibrată la nivel de linii (relativ la norma $\|\mathbf{x}\|_\infty$) dacă are loc relația

$$B^{-1} \leqslant \max_{1 \leqslant j \leqslant n} \|a_{i,j}\| \leqslant 1, \; i = 1, 2, \ldots, n, \qquad (5.39)$$

unde B este baza de reprezentare a numerelor în virgulă mobilă.

- Matricea A este echilibrată la nivel de coloană (relativ la norma $\|\mathbf{x}\|_\infty$) dacă are loc relația

$$B^{-1} \leqslant \max_{1 < i \leqslant n} |a_{i,j}| \leqslant 1, \qquad j = 1, 2, \ldots, n. \qquad (5.40)$$

- Matricea A este echilibrată în ambele sensuri (la nivel de linie și coloană) dacă au loc relațiile (5.39) și (5.40) simultan.

Pentru unele matrice forma echilibrată [46, 128] nu este unică (obținîndu-se două matrice A' și A'', una care rezultă prin echilibrarea întîi la nivel de linie și apoi de coloană, respectiv prin echilibrare întîi la nivel de coloană și apoi la nivel de linie). Se observă că în cazurile în care A' diferă de A'', sistemele de ecuații asociate celor două matrice vor avea numere de condiționare diferite [cond. $(A') \neq$ $\neq$ cond. (A'')] și se vor alege pivoți diferiți pentru metoda de eliminare a lui Gauss. Această afirmație atrage atenția ca problema scalării matricei A să se facă cu multă atenție. Se poate afirma că nu există o soluție practică satisfăcătoare pentru scalarea sistemelor liniare, metodă de scalare care să fie bună pentru matrice arbitrare și norme matriceale arbitrare.

5.1.6. *Clasificarea metodelor de rezolvare a sistemelor*

Metodele de rezolvare a sistemelor de ecuații liniare pot fi grupate în trei clase:

- metode bazate pe calculul determinanților;

• metode bazate pe eliminare (în general atribuite lui Gauss);

• metode iterative.

Din punctul de vedere al calculelor efectuate cu ajutorul calculatorului, prima clasă bazată pe calculul determinanţilor este neeficientă, datorită numărului foarte mare de operaţii pe care le implică (care nu este foarte important numai din punctul de vedere al timpului de execuţie), în plus precizia calculului este afectată dacă numărul operaţiilor este foarte mare şi eroarea de rotunjire se acumulează. Datorită acestui fapt în continuare aceste metode bazate pe evaluarea determinanţilor nu se vor mai prezenta.

La alegerea unei metode de calcul pentru o anume aplicaţie şi un sistem algebric dat trebuie avute în vedere o serie de criterii, cum ar fi :

• Care este numărul de operaţii aritmetice necesare pentru aplicaţia respectivă?

• Care va fi precizia rezultatelor finale?

• Cum poate fi testată precizia calculelor prin verificări intermediare?

Pentru o aplicaţie dată se alege metoda de calcul numeric care poate satisface în măsura cea mai mare toate cele trei deziderate.

În continuare se vor prezenta două clase de metode :

• Clasa metodelor directe (sau metode exacte), metode în care o secvenţă de operaţii se execută o singură dată, iar rezultatele obţinute sînt o aproximaţie a rezultatului exact. Aceste metode permit onţinerea soluţiei sistemului considerat, făcînd abstracţie de erorile de rotunjire şi trunchiere, folosind un număr finit de operaţii elementare.

• Clasa metodelor indirecte (sau metode iterative). Acestea permit găsirea soluţiei printr-un proces de aproximări succesive. O aceeaşi secvenţă de operaţii (mai redusă ca la metodele directe) este repetată de mai multe ori, obţinîndu-se o soluţie din ce în ce mai bună în sensul preciziei (convergenţa procesului). În cadrul acestei clase de metode, soluţia sistemului se obţine ca *limita unui şir* de vectori (vectori care reprezintă soluţia pentru diversele iteraţii efectuate). În cadrul metodelor indirecte se pune

problema alegerii acelei metode, care este convenabilă din punctul de vedere al vitezei de convergență, pentru o alegere adecvată a aproximației inițiale (a valorii de start).

5.2. Metode directe pentru rezolvarea sistemelor de ecuații liniare

Metodele directe pentru rezolvarea sistemelor cele mai frecvent utilizate sînt metodele bazate pe procesul de eliminare sau descompunerea matricei A. În ambele cazuri sistemul inițial trece prin diverse forme, dar toate acestea forme trebuie să fie echivalente cu forma inițială.

Pentru un sistem de ecuații $A\mathbf{x} = \mathbf{b}$, cu o matrice densă, ale cărei elemente sînt stocate în memoria calculatorului, se consideră că metoda de eliminare introdusă de Gauss reprezintă algoritmul cel mai bun atît în cea ce privește timpul de execuție cît și precizia care caracterizează soluția. Metoda de eliminare are o largă aplicabilitate la rezolvarea sistemelor de ecuații algebrice precum și la calculul inversei unei matrice.

5.2.1. *Metoda de eliminare a lui Gauss*

Pentru prezentarea metodei lui Gauss se consideră următorul sistem :

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned}\right\}, \qquad (5.41)$$

unde

$$A = \begin{bmatrix} a_{11}\, a_{12} \ldots a_{1n} \\ a_{21}\, a_{22} \ldots a_{2n} \\ \cdot \\ \cdot \\ \cdot \\ a_{n1}\, a_{n2} \ldots a_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix},$$

sau sub formă matriceală

$$A\mathbf{x} = \mathbf{b}. \tag{5.41'}$$

După operațiile de scalare, condiționare și echilibrare ale sistemului dat în (5.41), operații prezentate în 5.1, se aplică metoda de eliminare a lui Gauss care constă din următorul procedeu.

În primul pas, se folosește prima ecuație a sistemului la eliminarea necunoscutei x_1 din celelalte $n-1$ ecuații, obținîndu-se un nou sistem

$$A^{(1)}\ \mathbf{x} = \mathbf{b}^{(1)}. \tag{5.42}$$

În al doilea pas se folosește a doua ecuație din sistemul (5.42) pentru eliminarea necunoscutei x_2 din ultimele $n-2$ ecuații ale sistemului (4.52), obținîndu-se un nou sistem

$$A^{(2)}\ \mathbf{x} = \mathbf{b}^{(2)}. \tag{5.43}$$

În urma procesului de eliminare a necunoscutelor au loc o serie de transformări asupra elementelor matricei A și a vectorului coloană $\mathbf{b}$. Pentru a pune în evidență aceste transformări, înainte de eliminarea necunoscutei x_k, se vede că sistemul echivalent obținut după eliminarea necunoscutelor $x_1, x_2, \ldots, x_{k-1}$ poate fi scris sub forma

$$A^{(k)}\mathbf{x} = \mathbf{b}^{(k)}, \quad k = 1, 2, \ldots, n, \tag{5.44}$$

k reprezentînd numărul transformărilor aplicate sistemului inițial,

$$A^{(k)} = (a_{i,j}^{(k)}), \quad \mathbf{b}^{(k)} = \begin{bmatrix} b_1^{(k)} \\ b_2^{(k)} \\ \cdot \\ \cdot \\ \cdot \\ b_n^{(k)} \end{bmatrix}. \tag{5.45}$$

Pentru $k = 1$, rezultă $A^{(1)} = A$, $\mathbf{b}^{(1)} = \mathbf{b}$, iar elementele date în (5.45) pentru $k = 2, 3, \ldots, n$ se calculează cu ajutorul relațiilor

$$a_{i,j}^{(k)} = \begin{cases} a_{i,j}^{(k-1)}, & i \leqslant k-1, \\ 0, & i \geqslant k,\ j \leqslant k-1, \\ a_{i,j}^{(k-1)} - \dfrac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} a_{k-1,j}^{(k-1)}, & i \geqslant k,\ j \geqslant k, \end{cases} \tag{5.46}$$

$$b_i^{(k)} = \begin{cases} b_i^{(k-1)}, & i \leqslant k-1, \\ b_i^{(k-1)} - \dfrac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} b_{k-1}^{(k-1)}, & i \geqslant k. \end{cases}$$

Relațiile (5.46) constituie efectul înmulțirii ecuației $k-1$ a sistemului $A^{(k-1)}\mathbf{x} = \mathbf{b}^{(k-1)}$ prin raportul $a_{i,k-1}^{(k-1)}/a_{k-1,k-1}^{(k-1)}$ și scăderea ecuației astfel obținute din toate ecuațiile i, pentru orice $i \geqslant k$. În acest fel variabila x_{k-1} este eliminată din ultimele $n-k+1$ ecuații ale sistemului. În această etapă sistemul nou obținut arată astfel:

$$\left[\begin{array}{cccc|ccc} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \ldots & a_{1,k-1}^{(1)} & a_{1,k}^{(1)} \ldots a_{1,n}^{(1)} \\ & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & a_{2,k-1}^{(2)} & a_{2,k}^{(2)} \ldots a_{2,n}^{(2)} \\ & & a_{33}^{(3)} & \ldots & a_{3,k-1}^{(3)} & a_{3,k}^{(3)} \ldots a_{3,n}^{(3)} \\ & & & \ldots & \ldots & \ldots \\ & & & & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} \ldots a_{k-1,n}^{(k-1)} \\ & & & & \ldots & \ldots \\ & & & & & a_{k,k}^{(k)} \ldots a_{k,n}^{(k)} \\ & & & & & a_{k+1,k}^{(k)} \ldots a_{k+1,n}^{(k)} \\ & & & & & \vdots \\ & & & & & a_{n,k}^{(k)} \ldots a_{n,n}^{(k)} \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{k-1} \\ \ldots \\ x_k \\ x_{k+1} \\ \vdots \\ x_n \end{array}\right] = \left[\begin{array}{c} b_1^{(1)} \\ b_2^{(1)} \\ b_3^{(3)} \\ \vdots \\ b_{k-1}^{(k-1)} \\ \ldots \\ b_k^{(k)} \\ b_{k+1}^{(k)} \\ \vdots \\ b_n^{(k)} \end{array}\right]$$

sau sub formă matriceală $A_{(k)}\,\mathbf{x} = \mathbf{b}^{(k)}$. (5.47)

În tot acest proces de calcul s-a presupus că $a_{k.k}^{(k)} \neq 0$, $k = 1, 2, \ldots, n$. Pentru $k = n$, înseamnă că necunoscuta x_{n-1} a fost eliminată din ultima ecuație, obținîndu-se un sistem sub formă triunghiulară

$$\left.\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \ldots + a_{1,k}^{(1)}x_k + \ldots + a_{1n}^{(1)}x_n &= b_1^{1)} \\ a_{222}^{(k2)}x_2 + \ldots + a_2^{()}x_k + \ldots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{kh}^{(k)}x_k + \ldots + a_{kn}^{(k)}x_n &= b_k^{(k)} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{nn}^{(n)}x_n &= b_n^{(n)} \end{aligned}\right\} \quad (5.48)$$

sau

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \ldots & a_{1,n-1}^{(1)} & a_{1,n}^{(1)} \\ & a_{22}^{(2)} & \ldots & a_{2,n-1}^{(2)} & a_{2,n}^{(2)} \\ & & \ldots\ldots\ldots & & \\ & & & a_{n-1,n-1}^{(n-1)} & a_{n-1,n}^{n-1} \\ & & & & a_{n,n}^{(n)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_{n-1}^{(n-1)} \\ b_n^{(n)} \end{bmatrix} \quad (5.49)$$

respectiv sub formă matriceală

$$A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}, \quad (5.50)$$

unde $A^{(n)}$ este o matrice superior triunghiulară. Sistemul $A^{(n)}\,\mathbf{x} = \mathbf{b}^{(n)}$ este echivalent cu sistemul $A\mathbf{x}=\mathbf{b}$ [42, 128], adică admit aceeași soluție într-o anumită vecinătate din spațiul soluției $(x_1, x_2, \ldots, x_n)$, matricea $A^{(n)}$ și vectorul $\mathbf{b}^{(n)}$ au fost obținute din A și $\mathbf{b}$ prin diverse transformări elementare.

Sistemul triunghiular obținut (5.49) este sistemul cel mai simplu din șirul de transformări; pentru rezolvarea lui se folosește procesul de eliminare inversă, descris astfel:

$$\begin{aligned} x_n &= b_n^{(n)}/a_{nn}^{(n)}; \\ x_{n-1} &= (b_{n-1}^{(n-1)} - a_{n-1.n-1}^{(n-1)}x_n)/(a_{n-1.n-1)}^{(n-1)}); \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_j &= (b_j^{(j)} - a_{jn}^{(j)}x_n - \ldots - a_{j,j+1}^{(j)}x_{j+1})/a_{jj}^{(j)}; \\ & j = n-2, n-1, \ldots, 3, 2, 1. \end{aligned} \quad (5.51)$$

Această metodă înlocuieşte sistemul general de ecuaţii $Ax = b$ cu un sistem triunghiular echivalent $A^{(n)}x = b^{(n)}$, care este foarte simplu de rezolvat prin metoda substituţiei inverse.

Procedura de găsire a sistemului echivalent (5.50) presupune calculul şirului de matrice $A^{(1)}, A^{(2)}, \ldots, A^{(k)}, \ldots, A^{(n)}$ [unde $A^{(n)}$ este matrice superior triunghiulară] şi al şirului de vectori $b^{(1)}, b^{(2)}, \ldots, b^{(k)}, \ldots, b^{(n)}$. Vom prezenta procedura de construcţie a celor două şiruri (şirul de matrice, repsectiv şirul de vectori).

Pentru eliminarea lui x_1 din ecuaţiile numerotate cu $i = 2, \ldots, n-1$, se adună prima ecuaţie înmulţită cu constanta m_{i1}, la toate ecuaţiile $i = 2, 3, \ldots, n$, unde

$$m_{1i} = -a_{i1}^{(1)}/a_{11}^{(1)}, \quad i \geqslant 2. \tag{5.52}$$

După această operaţie rezultă sistemul

$$A^{(2)}\,\mathbf{x} = b^{(2)}, \tag{5.53}$$

sistem în care se găseşte x_1 doar în prima ecuaţie şi lipseşte din celelalte $n-1$ ecuaţii.

Este uşor de verificat că matricea $M^{(1)}$

$$M^{(1)} = \begin{bmatrix} 1 & 0 \\ \hline m_{21} & \\ m_{31} & I_{n-1} \\ \vdots & \\ m_{n,1} & \end{bmatrix} \tag{5.54}$$

realizează primul pas al metodei eliminării. Astfel $M^{(1)}A^{(1)}\mathbf{x} = M^{(1)}b^{(1)}$ devine

$$A^{(2)}\,\mathbf{x} = b^{(2)}. \tag{5.55}$$

Se poate arăta că sistemul (5.41) este echivalent cu sistemul (5.55). În acest sens s-a presupus că (5.41) are o

soluție unică, deci $A^{(1)}$ este nesingulară, respectiv $\det A^{(1)} \neq 0$. Din construcția matricei $M^{(1)}$ se vede că $\det M^{(1)} \neq 1$. Deoarece

$$A^{(2)} = M^{(1)}A^{(1)} \Rightarrow \det A^{(2)} = \det M^{(1)} \cdot \det A^{(1)} = \det A^{(1)}, \tag{5.56}$$

deci $A^{(2)}$ este nesingulară și sistemul (5.41) are soluție unică.

Din cele două relații matriceale date în (5.55) se vede că soluția lui (5.41) satisface și a doua relație din (5.55), deci cele două soluții unice ale lui (5.41) și (5.55) sînt identice. În pasul k se elimină x_k din cele $n-k$ ecuații. Pentru aceasta se înmulțește ecuația pivotată k prin constantele m_{ik} și se adună la ecuațiile unde $k < i \leqslant n$,

$$m_{ik} = -\, a_{ik}^{(k)} / a_{kk}^{(k)}. \tag{5.57}$$

După această operație rezultă sistemul în care a fost eliminată necunoscuta x_k din ecuațiile $k < i \leqslant n$, sistem care are forma

$$A^{(k+1)}\,\mathbf{x} = \mathbf{b}^{(k+1)}. \tag{5.58}$$

Se observă că matricea

$$M^{(k)} = \left[\begin{array}{c|c} I_{k-1} & 0 \\ \hline & \begin{array}{ccccc} 1 & & & & \\ m_{k+1,\,k} & 1 & & & 0 \\ m_{k+2,\,k} & 0 & 1 & & \\ \cdot & \cdot & \cdot & \cdot & \\ m_{rk} & \cdot & \cdot & \cdot & 0 \quad 1 \end{array} \\ \hline 0 & \end{array}\right] \tag{5.59}$$

realizează eliminarea lui x_k. Astfel $M^{(k)}A^{(k)}\,\mathbf{x} = M^{(k)}\mathbf{b}$ devine

$$A^{(k+1)}\mathbf{x} = \mathbf{b}^{(k+1)}. \tag{5.60}$$

Din cele descrise se vede că procesul de triunghiularizare al matricei A se poate realiza cu ajutorul produsului de matrice

$$M = M^{(n-1)}M^{(n-2)} \ldots M^{(2)}\,M^{(1)}, \tag{5.61}$$

care, înmulțită cu sistemul inițial, dă

$$MA^{(1)}\mathbf{x} = M\mathbf{b}^{(1)} \text{ sau } A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}. \tag{5.62}$$

Sistemul (5.62) este echivalent cu (5.41), fapt care rezultă din relațiile $A^{(n)} = MA^{(1)}$. Dar

$$\det A^{(n)} = [\det M] \det [A^{(1)}],$$

unde

$$\det M = \prod_{k=1}^{n-1} \det M^{(k)} = 1,$$

avînd în vedere forma matricelor $M^{(k)}$, pentru $k = 1, 2, \ldots$ $\ldots, n-1$. Deci

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

deoarece $A^{(n)}$ este triunghiulară.

În cazul în care unul din pivoții $a_{i,i}$ este egal cu zero, se face o schimbare a ecuațiilor între ele cu ajutorul matricelor de permutare [128, 32, 10, 50, 108].

5.2.2. *Metoda Gauss-Jordan*

Această metodă constituie o formă modificată a metodei eliminării, introduse de Gauss în 1823. Metoda constă în a transforma un sistem de ecuații liniare cu o matrice pătrată într-un sistem echivalent cu sistemul inițial considerat, ce are ca matrice chiar matricea unitate. Din cele prezentate în 5.2.1, se vede că matricele $M^{(1)}$, $M^{(2)}$, $\ldots$ $\ldots, M^{(k)}$ au forma

$$M^{(1)} = \left[\begin{array}{c|c} 1 & 0 \\ \hline m_{21} & \\ m_{31} & I_{n-1} \\ \vdots & \\ m_{n1} & \end{array}\right],$$

$$M^{(2)} = \begin{bmatrix} 1 & m_{12} & 0 \ldots 0 \\ 0 & 1 & 0 \ldots 0 \\ 0 & m_{32} & 1 \ldots 0 \\ \cdot & \cdot & \cdot \\ 0 & m_{n2} & 0 \ldots 1 \end{bmatrix}, \ldots, M^{(k)} = \left[\begin{array}{c|c|c} & m_{1k} & \\ I_{k-1} & m_{2k} & 0 \\ & m_{k-1,\,k} & \\ \hline & 1 & \\ \hline 0 & m_{k+1,\,k} & I_{n-k} \\ & m_{n,k} & \end{array}\right] \tag{5.63}$$

unde

$$m_{i1} = -\,a_{i1}^{(1)}/a_{11}^{(1)}\,;\; m_{i2} = -\,(a_{i2}^{(2)}/a_{22}^{(2)})\,;\; m_{ik} = -\,(a_{ik}^{(k)}/a_{k,k}^{(k)}) \tag{5.64}$$

$$i = 2,\ldots,n\,;\quad i=1,3,\ldots,n\,;\quad i=1,2,\ldots,k-1,k+1,\ \ldots,\ n.$$

Se observă că

$$M^{(n)}\,M^{(n-1)}\,\ldots\,M^{(k)}\,\ldots M^{(2)}\,M^{(1)}\,A^{(1)} = I, \tag{5.65}$$

$$M^{(n)}M^{(n-1)}\,\ldots\,M^{(k)}\,\ldots\,M^{(2)}\,M^{(1)}\,\mathbf{b} = \mathbf{b}^{(n)}.$$

Dacă $M = M^{(n)}\,M^{(n-1)}\,\ldots\,M^{(k)}\,\ldots\,M^{(2)}\,M^{(1)}$, atunci ecuaţia $MA\mathbf{x} = M\mathbf{b}$ devine

$$I\mathbf{x} = \mathbf{b}^{(n)} \tag{5.66}$$

sau dezvoltat

$$\begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} b_1^{(n)} \\ b_2^{(n)} \\ \cdot \\ \cdot \\ b_n^{(n)} \end{bmatrix}. \tag{5.67}$$

Algoritmul de calcul prin care se ajunge la sistemul echivalent (5.67) se poate prezenta cu ajutorul relaţiilor următoare :

$$a_{ij}^{(1)} = a_{ij},$$

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}}\,a_{k-1,j}^{(k-1)},\quad i \neq k-1,\ j \geqslant k-1, \tag{5.68}$$

$$a_{k-1,j}^{(k)} = a_{k-1,j}^{(k-1)},\ j \geqslant k-1,$$

$$a_{i,j}^{(k)} = a_{ij}^{(k-1)},\ j < k-1,$$

iar asupra componentelor lui $\mathbf{b}$,

$$b_i^{(1)} = b_i,$$

$$b_i^{(k)} = b_i^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} b_{k-1}^{(k-1)}, \quad i \neq k-1, \tag{5.69}$$

$$b_{k-1}^{(k)} = b_{k-1}^{(k-1)}, \qquad k = 2, 3, \ldots, n.$$

Soluția sistemului în cazul metodei Gauss-Jordan este dată de relațiile

$$x_i = (b_i^{(n)})/(a_{ii}^{(n)}) \text{ sau } x_i = b_i^{(n)}, \; i = 1, 2, \ldots, n. \tag{5.70}$$

5.2.3. *Metoda lui Gauss prin descompunerea matricei A*

Dacă se face apel la relațiile (5.62)

$$MA^{(1)}\,\mathbf{x} = M\mathbf{b}^{(1)} \text{ sau } A^{(n)}\,\mathbf{x} = \mathbf{b}^{(n)},$$

se observă că

$$MA^{(1)} = A^{(n)} \text{ sau } A = A^{(1)} = M^{-1}A^{(n)}. \tag{5.71}$$

Ținînd seama de (5.61), avem

$$M^{-1} = [M^{(1)}]^{-1}[M^{(2)}]^{-1} \ldots [M^{n-1}]^{-1}; \tag{5.72}$$

unde

$$[M^{(k)}]^{-1} = \begin{bmatrix} I_{k-1} & 0 & & \\ & 1 & & 0 \\ & -m_{k+1,k} & 1 & \\ 0 & -m_{k+2,k} & 0 & 1 \\ & \ldots & & \\ & -m_{n,k} & \ldots & 1 \end{bmatrix}, \tag{5.73}$$

$$M^{-1} = \begin{bmatrix} 1 & & & & \\ -m_{21} & 1 & & 0 & \\ -m_{31} & -m_{32} & 1 & & \\ -m_{41} & -m_{42} & -m_{43} & 1 & \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ -m_{n1} & -m_{n2} & -m_{n3} & -m_{n4} & \ldots 1 \end{bmatrix}.$$

Se observă că M^{-1} este o matrice inferior triunghiulară care conține pe diagonala principală numai unu.

Din metoda lui Gauss se vede că $A^{(n)}$ este o matrice superior triunghiulară. În acest caz, dacă toți pivoții sînt diferiți de zero în procesul de formare al matricei $A^{(n)}$ din $A^{(1)}$, atunci $A^{(1)} = A$ poate fi descompusă într-un produs de două matrice: una inferior triunghiulară (cu unu pe diagonala principală) și alta superior triunghiulară.

Teoremă. *O matrice $A \in R^{n\times n}$ poate fi scrisă sub forma unui produs TS de două matrice, unde T este inferior triunghiulară și S superior triunghiulară dacă*

$$\det [a_{11}] \neq 0, \det \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0, \ldots, \det A \neq 0.$$

Descompunerea este unică dacă elementele lui T sau S de pe diagonala principală sînt specificate astfel:

$$A = TS = \begin{bmatrix} a_{11} & a_{12} \ldots a_{1n} \\ a_{21} & a_{22} \ldots a_{2n} \\ \ldots & \ldots \\ a_{n1} & a_{n2} \ldots a_{nn} \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & & & \\ t_{21} & 1 & & \\ t_{31} & t_{32} & 1 & 0 \\ \ldots & \ldots & & \\ t_{n1} & t_{n2} \ldots & t_{n,n-1} & 1 \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} & s_{13} \ldots & s_{1n} \\ & s_{22} & s_{23} \ldots & s_{2n} \\ & & s_{33} \ldots & s_{3n} \\ & 0 & \ldots & \\ & & & s_{nn} \end{bmatrix} \cdot \qquad (5.74)$$

Dacă se execută produsul dintre matricele T și S din (5.75) și se face identificarea, rezultă n^2 ecuații neliniare cu n^2 necunoscute. Calculîndu-se componentele primei linii a matricei produs TS, rezultă

$$s_{11} = a_{11},\ s_{12} = a_{12}, \ldots,\ s_{1n} = a_{1n}, \qquad (5.75)$$

de unde se vede că prima linie a matricei S coincide cu prima linie a matricei A. Dacă se calculează elementele

primei coloane (sub diagonala principală), se obțin ecuațiile

$$t_{21}s_{11} = a_{21};\ t_{31}s_{11} = a_{31}, \ldots, t_{n1}s_{11} = a_{n1}. \qquad (5.76)$$

Se observă că pentru $s_{11} = a_{11} \neq 0$, rezultă din aceste ecuații elementele primei coloane a matricei T.

În general, dacă se calculează elementele liniei k, din matricea S, începînd cu elementul de pe diagonală principală, rezultă ecuațiile următoare și deoarece t_{kj} și s_{jk} sînt cunoscute, pentru $j \leqslant k-1$, se poate rezolva sistemul și găsi elementele lui S din linia k:

$$\left.\begin{array}{l|l}
\sum_{j=1}^{k-1} t_{kj}s_{jk} + s_{kk} = a_{kk} & s_{kk} = a_{kk} - \sum_{j=1}^{k-1} t_{kj}s_{jk} \\
\sum_{j=1}^{k-1} t_{kj}s_{j,k+1} + s_{k,k+1} = a_{k,k+1} & s_{k,k+1} = a_{k,k+1} - \sum_{j=1}^{k-1} t_{kj}s_{j,k+1} \\
\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots \\
\sum_{j=1}^{k-1} t_{kj}s_{jn} + s_{kn} = a_{k,n} & s_{kn} = s_{kn} - \sum_{j=1}^{k-1} t_{kj}s_{jn}
\end{array}\right\} \qquad (5.77)$$

În mod asemănător se calculează elementele coloanei k a matricei T (de sub diagonala principală), obținîndu-se ecuațiile următoare, iar dacă $s_{kk} \neq 0$, atunci sistemul de ecuații permite găsirea elementelor coloanei k a matricei T:

$$\left.\begin{array}{l|l}
\sum_{j=1}^{k-1} t_{k+1,j}s_{jk} + t_{k+1,k}s_{kk} = a_{k+1,k} & t_{k+1,k} = \frac{1}{s_{kk}}\left(a_{k+1,k} - \sum_{j=1}^{k-1} t_{k+1,j}s_{jk}\right) \\
\sum_{j=1}^{k-1} t_{k+2,j}s_{jk} + t_{k+2,k}s_{kk} = a_{k+2,k} & t_{k+2,k} = \frac{1}{s_{kk}}\left(a_{k+2,k} - \sum_{j=1}^{k-1} t_{k+2,j}s_{jk}\right) \\
\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots \\
\sum_{j=1}^{k-1} t_{nj}s_{jk} + t_{nk}s_{kk} = a_{nk} & t_{nk} = \frac{1}{s_{kk}}\left(a_{nk} - \sum_{j=1}^{k-1} t_{nj}s_{jk}\right)
\end{array}\right\} \qquad (5.78)$$

În mod asemănător s-ar desfășura calculele dacă S ar fi avut pe diagonala principală unitatea.

Exemplu. Fie ($n = 3$)

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & & 0 \\ t_{21} & 1 & \\ t_{31} & t_{32} & 1 \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ & s_{22} & s_{23} \\ 0 & & s_{33} \end{bmatrix} = T\,S. \tag{5.79}$$

După efectuarea produsului TS și identificarea, rezultă $n^2 = 9$ ecuații neliniare cu nouă necunoscute. Primele trei ecuații arată că elementele primei linii a lui S coincid cu elementele primei linii a matricei A :

$$s_{11} = a_{11}, \quad s_{12} = a_{12}, \quad s_{13} = a_{13}.$$

Pentru calculul primei coloane a matricei T se folosesc următoarele ecuații ($s_{11} = a_{11} \neq 0$ prin ipoteză) :

$$\left.\begin{aligned} t_{21}s_{11} &= a_{21} \\ t_{31}s_{11} &= a_{31} \end{aligned}\right| \quad \begin{aligned} t_{21} &= \frac{a_{21}}{s_{11}} = \frac{a_{21}}{a_{11}} \\ t_{31} &= \frac{a_{31}}{s_{11}} = \frac{a_{31}}{a_{11}} \end{aligned}$$

Din următoarele ecuații se pot calcula termenii s_{22} și s_{23} :

$$t_{21}s_{12} + s_{22} = a_{22}, \; s_{22} = a_{22} - t_{21}s_{12} = a_{22} - \frac{a_{21}a_{12}}{a_{11}},$$

$$t_{21}s_{13} + s_{23} = a_{23}, \; s_{23} = a_{23} - t_{21}s_{13} = a_{23} - \frac{a_{21}a_{13}}{a_{11}}.$$

Au mai rămas de determinat elementele t_{32} și s_{33}, care sînt necunoscutele din următoarele ecuații :

$$t_{31}s_{12} + t_{32}s_{22} = a_{32}, \qquad t_{32} = \frac{a_{32}a_{11} - a_{31}a_{12}}{a_{11}a_{22} - a_{21}a_{12}}.$$

$$t_{31}s_{13} + t_{32}s_{23} + s_{33} = a_{33}, \quad s_{33} = a_{33} - t_{31}s_{13} - t_{32}s_{23}.$$

În acest fel au fost determinate toate elementele matricelor T și S.

Analizînd expresiile coeficienţilor matricelor T şi S şi ţinînd seama de relaţiile (5.46) şi (5.47) şi (5.64), se observă că

$$A = A^{(1)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} \end{bmatrix} = \begin{bmatrix} 1 & & 0 \\ -m_{21} & 1 & \\ -m_{31} & -m_{32} & 1 \end{bmatrix} \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} \\ & a_{22}^{(2)} & a_{23}^{(2)} \\ 0 & & a_{33} \end{bmatrix} \tag{5.80}$$

În urma exemplului considerat se vede că

$$A = A^{(1)} = TS, \text{ unde } T = M^{-1} \text{ şi } S = A^{(n)}. \tag{5.81}$$

În acest caz se poate afirma că metoda lui Gauss de eliminare este echivalentă cu descompunerea matricei $A = TS$, cu T avînd pe diagonală unitatea.

În urma descompunerii matricei A în T şi S sistemul $A\mathbf{x} = \mathbf{b}$ se rezolvă astfel:

$$A\mathbf{x} = \mathbf{b},\ A = TS,\ TS\mathbf{x} = \mathbf{b}, \tag{5.82}$$

unde sistemul $TS\mathbf{x} = \mathbf{b}$ este echivalent cu două sisteme triunghiulare $T\mathbf{y} = \mathbf{b}$, $S\mathbf{x} = \mathbf{y}$. Sistemul $T\mathbf{y} = \mathbf{b}$ se poate rezolva prin substituţie directă, iar $S\mathbf{x} = \mathbf{y}$ prin substituţie inversă:

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ t_{21} & 1 & 0 & \dots & 0 \\ t_{31} & t_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ t_{n1} & t_{n2} & t_{n3} & \dots & t_{n,n-1}\ 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix},$$

$$\begin{bmatrix} s_{11} & s_{12} & s_{13} & \dots & s_{1n} \\ & s_{22} & s_{23} & \dots & s_{2n} \\ & & s_{33} & \dots & s_{3n} \\ & 0 & \dots & \dots & \dots \\ & & & & s_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix},$$

obținîndu-se următoarele relații pentru determinarea necunoscutelor $(y_1, y_2, \ldots, y_n)$ și a necunoscutelor $(x_1, x_2, \ldots, x_n)$:

$$y_1 = b_1,$$

$$y_k = b_k - \sum_{i=1}^{k-1} t_{ki} y_i,\ i = 2, \ldots, n,$$

$$x_n = s_{nn}/y_n, \tag{5.83}$$

$$x_k = \frac{1}{s_{k,k}} \left(y_k - \sum_{i=n}^{k+1} s_{k,i} x_i \right),\ i = n-1, n-2, \ldots, 3, 2, 1.$$

Metoda de descompunere a matricei A într-un produs de două matrice triunghiulare este exemplificată prin schema logică din fig. 5.6 și programul 5.1.

5.2.4. *Alte variante ale metodei lui Gauss de eliminare*

Există foarte multe metode pentru rezolvarea sistemelor liniare care diferă foarte puțin de metoda lui Gauss de eliminare; diferența dintre aceste metode și metoda lui Gauss nu se referă la numărul de operații, ci la reducerea spațiului de memorie. În continuare se vor prezenta cîteva variante ale metodei lui Gauss.

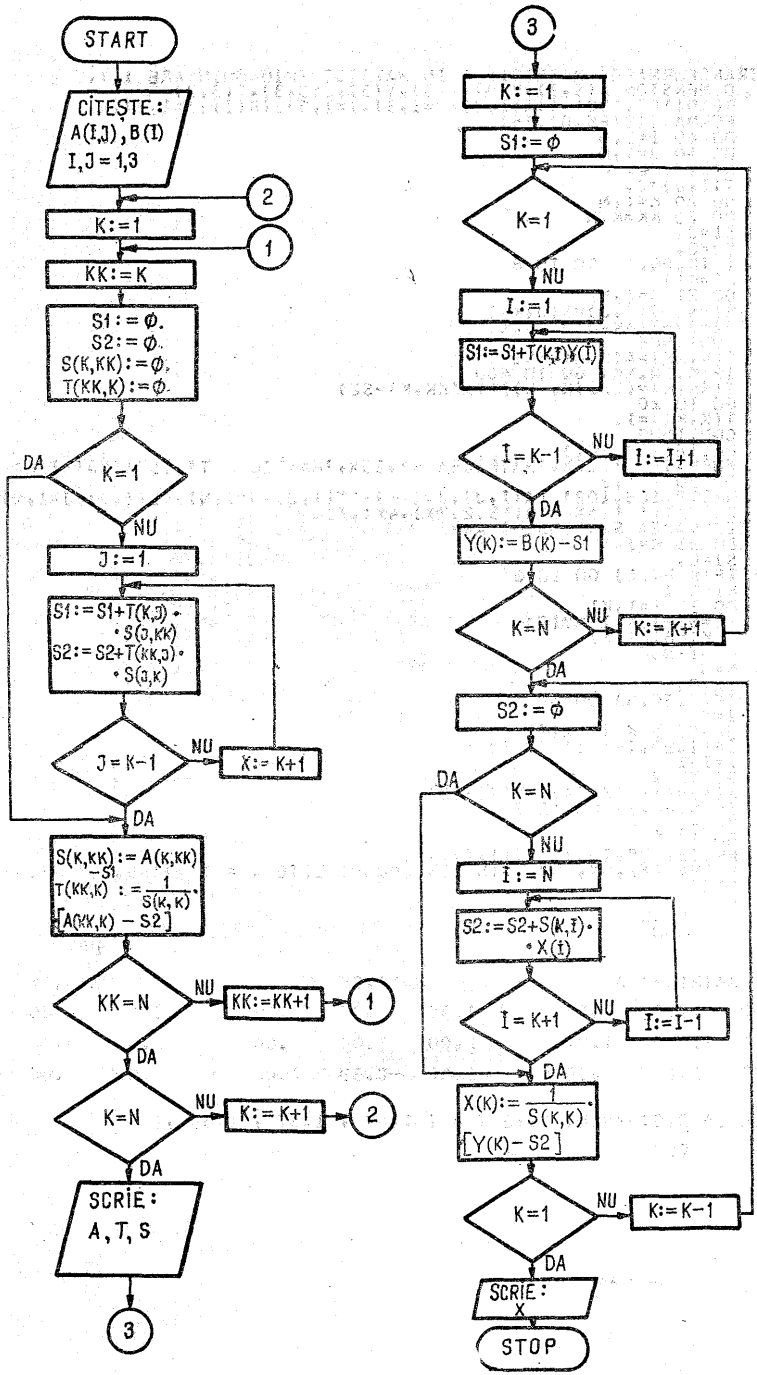
În metoda lui Gauss de eliminare se cerea executarea următoarelor calcule:

$$M^{(1)}A,\ M^{(2)}M^{(1)}A,\ M^{(3)}M^{(2)}M^{(1)}A, \ldots$$

$$M^{(1)}\mathbf{b},\ M^{(2)}M^{(1)}\mathbf{b},\ M^{(3)}M^{(2)}M^{(1)}\mathbf{b}, \ldots$$

și memorarea elementelor tuturor acestor linii în toate matricele reduse care au fost schimbate prin operații numerice.

● *Metoda lui Crout.* Această metodă este aplicabilă dacă liniile și coloanele sînt astfel aranjate că în metoda lui Gauss nu se cere nici o schimbare. În general pivotul nu va fi elementul maxim (deci erorile pot crește rapid în


START
CITEȘTE:
A(I,J), B(I)
I,J = 1,3
2
K:=1
1
KK:=K
S1:=φ.
S2:=φ.
S(K,KK):=φ.
T(KK,K):=φ.
DA
K=1
NU
J:=1
S1:=S1+T(K,J)·
·S(J,KK)
S2:=S2+T(KK,J)·
·S(J,K)
J=K-1
NU
K:=K+1
DA
S(K,KK):=A(K,KK)
-S1
T(KK,K):=1/S(K,K)·
[A(KK,K)-S2]
KK=N
NU
KK:=KK+1
1
DA
K=N
NU
K:=K+1
2
DA
SCRIE:
A, T, S
3
3
K:=1
S1:=φ
K=1
NU
I:=1
S1:=S1+T(K,I)·Y(I)
I=K-1
NU
I:=I+1
DA
Y(K):=B(K)-S1
K=N
NU
K:=K+1
DA
S2:=φ
DA
K=N
NU
I:=N
S2:=S2+S(K,I)·
·X(I)
I=K+1
NU
I:=I-1
DA
X(K):=1/S(K,K)·
[Y(K)-S2]
K=1
NU
K:=K-1
DA
SCRIE:
X
STOP

```
C   TRANSFORMAREA MATRICII A IN MATRICI TRIUNGHIULARE : T,S
      DIMENSION A(3,3),B(3),X(3),Y(3),S(3,3),T(3,3)
      READ(105,100) ((A(I,J),J=1,3),I=1,3),(B(I),I=1,3),N
  100 FORMAT(12(F5.0),I5)
      DO 40 I=1,N
      DO 40 J=1,N
      S(I,J)=0.
   40 T(I,J)=0.
      DO 20 K=1,N
      DO 20 KK=K,N
      S1=0.
      S2=0.
      IF(K.EQ.1) GO TO 8
      K1=K-1
      DO 21 J=1,K1
      S1=S1+T(K,J)*S(J,KK)
   21 S2=S2+T(KK,J)*S(J,K)
    8 CONTINUE
      S(K,KK)=A(K,KK)-S1
      IF(K.EQ.KK) GO TO 400
      T(KK,K)=(1/S(K, K))*(A(KK,K)-S2)
      GO TO 20
  400 T(K,KK)=1.
   20 CONTINUE
      WRITE(108,101)
  101 FORMAT(' ',10X,'MATRICEA A',15X,'MATRICEA T',15X,'MATRICEA S',/
      DO 30 I=1,3
   30 WRITE(108,102) (A(I,J),J=1,N),(T(I,J),J=1,N),(S(I,J),J=1,N)
  102 FORMAT(' ',5X,3(3(F5.2,2X),4X),/)
C   REZOLVAREA SISTEMULUI AX=B
      DO 31 K=1,N
      S1=0.
      IF(K.EQ.1) GO TO 6
      K1=K-1
      DO 32 I=1,K1
   32 S1=S1+T(K,I)*Y(I)
    6 Y(K)=B(K)-S1
   31 CONTINUE
      K=N
    5 S2=0.
      IF(K.EQ.N) GO TO 4
      I=N
    2 S2=S2+S(K,I)*X(I)
      IF(I.EQ.(K+1)) GO TO 4
      I=I-1
      GO TO 2
    4 X(K)=(1/S(K,K))*(Y(K)-S2)
      IF(K.EQ.1) GO TO 7
      K=K-1
      GO TO 5
    7 WRITE(108,104) (X(I),I=1,3)
  104 FORMAT(//,6X,'SOLUTIA SISTEMULUI ESTE X = (',3(F5.2,' ,'),')')
      STOP
      END
          LINK
          RUN

      MATRICEA A                MATRICEA T                MATRICEA S
  1.00   1.00   2.00       1.00    .00    .00       1.00   1.00   2.00
  2.00  -1.00   1.00       2.00   1.00    .00        .00  -3.00  -3.00
  1.00   2.00    .00       1.00  -0.33   1.00        .00    .00  -3.00

  SOLUTIA SISTEMULUI ESTE X = ( 1.67 , 1.67 , 1.33 ,)

          EOJ
```

```
      DIMENSION A(4,4),B(4),X1(4),X2(4),ER(4),SC(4),SL(4),X3(4)
      DATA X1/0.,0.,0.,0./,VMC/-1000./,VML/-1000./,N/4/,ITER/0/,OM/0.5/
      DATA EPS/1.E-6/
      READ(105,100) ((A(I,J),J=1,4),I=1,4),(B(I),I=1,4)
  100 FORMAT(20(F4.2))
      N=4
      DO 20 I=1,N
      SC(I)=0.
      SL(I)=0.
      DO 20 J=1,N
      IF(I.EQ.J) GO TO 20
      SL(I)=SL(I)+(A(I,J)/A(I,I))
      SC(I)=SC(I)+A(J,I)/A(J,J)
   20 CONTINUE
      DO 21 I=1,N
      IF(SL(I).GT.VML) VML=SL(I)
      IF(SC(I).GT.VMC) VMC=SC(I)
   21 CONTINUE
      IF((VMC.LT.1.).AND.(VML.LT.1.)) GO TO 600
      WRITE(108,400) VML,VMC
  400 FORMAT(' ',5X,'MATRICEA P NU ESTE CONVERGENTA ',/,
     *       5X,'NORMELE MATRICII P SINT : ',2(F5.2))
      GO TO 700
  600 CONTINUE
      IF(VMC-VML)22,22,23
   22 VMIN=1/VMC
      GO TO 24
   23 VMIN=1/VML
   24 R=ALOG10(VMIN)
C    MERODA JACOBI
      CALL JACOBI(A,B,X1,X2,ER,N,ITER,EPS)
      WRITE(108,102) (X2(I),I=1,N),ITER
  102 FORMAT(' ',5X,'METODA JACOBI       : X = (',3(F8.3,' ,'),F8.3,
     *       ' )    NR. ITERATII =',I3)
C    METODA GAUSS-SEIDEL
      CALL GAUSS(A,B,X1,X2,ER,N,ITER,EPS)
      WRITE(108,109) (X2(I),I=1,N),ITER
  109 FORMAT(' ',5X,'METODA GAUSS-SEIDEL : X = (',3(F8.3,' ,'),F8.3,
     *       ' )    NR. ITERATII =',I3)
C    METODA RELAXARII
      CALL RELAX(A,B,X1,X2,X3,ER,N,ITER,OM,EPS)
      WRITE(108,111) (X3(I),I=1,N),ITER
  111 FORMAT(' ',5X,'METODA RELAXARII    : X = (',3(F8.3,' ,'),F8.3,
     *       ' )    NR. ITERATII =',I3)
  700 STOP
      END
      SUBROUTINE JACOBI(A,B,X1,X2,ER,N,ITER,EPS)
      DIMENSION A(4,4),B(4),X1(4),X2(4),ER(4)
   30 DO 25 I=1,N
      S=0.
      DO 26 J=1,N
      IF(J.EQ.I) GO TO 26
      S=S+A(I,J)*X1(J)
   26 CONTINUE
      X2(I)=(1/A(I,I))*(B(I)-S)
   25 ER(I)=X2(I)-X1(I)
      DO 27 I=1,N
      IF(ABS(ER(I)).LT.EPS) GO TO 28
   27 CONTINUE
      DO 29 I=1,N
   29 X1(I)=X2(I)
      ITER=ITER+1
      GO TO 30
   28 RETURN
      END
      SUBROUTINE GAUSS(A,B,X1,X2,ER,N,ITER,EPS)
      DIMENSION A(4,4),B(4),X1(4),X2(4),ER(4)
      ITER=0
      DO 45 I=1,N
      X1(I)=0.
      ER(I)=0.
   45 X2(I)=0.
   47 DO 40 I=1,N
      S1=0.
      S2=0.
      DO 41 J=1,N
      IF(J-I) 42,41,43
   42 S1=S1+A(I,J)*X2(J)
      GO TO 41
   43 S2=S2+A(I,J)*X1(J)
   41 CONTINUE
      X2(I)=(1/A(I,I))*(B(I)-S1-S2)
   40 ER(I)=X2(I)-X1(I)
      DO 44 I=1,N
```

Programul 5.1

metoda lui Crout). Metoda lui Crout este construită, cu scopul de a reduce numărul rezultatelor intermediare care trebuie memorate. Deci metoda este foarte bună pentru calculatoare cu memorie redusă.

Metoda lui Crout poate fi modificată să utilizeze ca pivot elementul maxim de pe coloană, prin folosirea schimbării liniilor. Această procedură de eliminare compactă se bazează pe faptul că numai elementele $a_{ij}^{(k)}$ (din metoda lui Gauss), pentru care $j \geqslant i$ și $i \leqslant k$, sînt necesare în substituția inversă finală. În acest sens se folosește o metodă recursivă pentru definirea coloanelor matricei T (matrice inferior triunghiulară) și a liniilor matricei S (matrice superior triunghiulară).

Această metodă exactă de rezolvare a sistemelor liniare are în vedere faptul că orice matrice A nesingulară poate fi descompusă într-un produs de două matrice T și S, introducînd un algoritm de calcul pentru elementele celor două matrice T și S în funcție de elementele matricei A date.

Din cele prezentate în paragraful precedent se știe că $A = T\,S$, sau sub formă dezvoltată

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & & & & \\ t_{21} & 1 & & 0 & \\ t_{31} & t_{32} & 1 & & \\ \dots & \dots & \dots & & \\ t_{n1} & t_{n2} & t_{n3} & \dots & 1 \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} & s_{13} & \dots & s_{1n} \\ & s_{22} & s_{23} & \dots & s_{2n} \\ & & s_{33} & \dots & s_{3n} \\ & 0 & & & \\ & & & & s_{nn} \end{bmatrix}.$$

În urma efectuării produsului matriceal și a procesului de identificare se obțin relațiile

$$\left.\begin{aligned} s_{kj} &= a_{kj} - \sum_{p=1}^{k-1} t_{kp} s_{pj}, && \text{dacă } k \leqslant j, \\ t_{ik} &= \frac{1}{s_{kk}} \left(a_{ik} - \sum_{p=1}^{k-1} t_{ip} s_{pk} \right), && \text{dacă } i > k. \end{aligned}\right\} \qquad (5.84)$$

Relaţiile (5.84) pentru $k \geqslant 2$ permit determinarea la început a elementelor liniei k din matricea S şi după aceea determinarea elementelor coloanei k din matricea T, presupunînd că se cunosc elementele primei linii din S şi elementele primei coloane din T.

În metoda lui Crout elementele primei linii din matricea S sînt $s_{1j} = a_{1j}$, $j = 1, 2, \ldots, n$, iar elementele primei coloane din T sînt $t_{i1} = a_{i1}/a_{11}$, $i = 2, \ldots, n$. Dacă se defineşte $s_{1,n+1} = b_1$ şi $s_{i,n+1} = b_i$, $i = 2, \ldots, n$, şi se utilizează formula (5.84) pentru $j = n + 1$, se găseşte o coloană a matricei S ($s_{i,n+1}$) care este tocmai vectorul $\mathbf{b}^{(n)}$ obţinut în urma transformărilor vectorului $\mathbf{b}$.

În acest moment se cunoaşte matricea S, se cunoaşte $\mathbf{b}^{(n)}$; deci pentru găsirea soluţiei sistemului

$$A\mathbf{x} = \mathbf{b} \text{ sau } TS\mathbf{x} = T\mathbf{b}^{(n)}$$

este suficient a se rezolva sistemul (deoarece T este nesingulară)

$$S\mathbf{x} = \mathbf{b}^{(n)},$$

unde S reprezintă primele n coloane ale matricei de elemente (s_{ij}), $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$, iar coloana $n+1$ este tocmai vectorul $\mathbf{b}^{(n)}$. Soluţia se obţine prin eliminarea inversă, obţinîndu-se $x_n, x_{n-1}, \ldots, x_2, x_1$:

$$x_n = \frac{b_n^{(n)}}{s_{nn}}, \quad x_i = \frac{1}{s_{ii}} \left(b_i^{(i)} - \sum_{j=i+1}^{n} s_{ij} x_j \right),$$

$$i = n - 1, \; n - 2, \ldots, 2, \; 1.$$

Această metodă se aplică în general pentru sisteme de dimensiune redusă sau în cazul cînd eroarea de calcul poate fi testată şi nu are o creştere foarte mare.

● *Metoda lui Doolittle.* Această metodă este strîns legată de metoda de descompunere a matricei A într-un produs TS de matrice triunghiulare.

Se consideră pentru simplificare $n = 4$. Atunci procesul de eliminare pentru patru ecuații conduce la matricea

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ a_{21} & a_{22} & a_{23} & a_{24} & b_2 \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \\ a_{41} & a_{42} & a_{43} & a_{44} & b_4 \end{bmatrix} \longrightarrow \begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} & c_1 \\ & s_{22} & s_{23} & s_{24} & c_2 \\ 0 & & s_{33} & s_{34} & c_3 \\ & & & s_{44} & c_4 \end{bmatrix}, \tag{5.85}$$

unde a_{ij} sînt elementele matricei A, iar b_i $(i = 1,2,3,4)$ sînt componentele vectorului $\mathbf{b}$. În urma unor transformări elementare se obține o matrice superior triunghiulară S și vectorul $\mathbf{c}$, de componente $c_i (i = 1,2,3,4)$. Deci $A = MS$ sau

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} 1 & & & 0 \\ -m_{21} & 1 & & \\ -m_{21} & -m_{32} & 1 & \\ -m_{41} & -m_{42} & -m_{43} & 1 \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ & s_{22} & s_{23} & s_{24} \\ 0 & & s_{33} & s_{34} \\ & & & s_{44} \end{bmatrix}, \tag{5.86}$$

unde M este o matrice de constante de multiplicare care se determină la fel ca în metoda lui Gauss de eliminare.

Se observă că vectorul final $\mathbf{c}$ este obținut din ecuația

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} 1 & & & 0 \\ -m_{21} & 1 & & \\ -m_{31} & -m_{32} & 1 & \\ -m_{41} & -m_{42} & -m_{43} & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}, \quad \mathbf{b} = M\mathbf{c}.$$

Dacă $A = MS$, $\mathbf{b} = M\mathbf{c}$, sistemul inițial se reduce la un sistem superior triunghiular astfel:

$$A\mathbf{x} = \mathbf{b}, \quad MS\mathbf{x} = M\mathbf{c}, \quad S\mathbf{x} = \mathbf{c}.$$

Deoarece matricea M este nesingulară, vectorul $\mathbf{x}$ se obține din ultima ecuație matriceală prin eliminare inversă.

● *Metoda factorizării directe.* Relația (5.84) folosită în metoda lui Crout oferă posibilitatea unui studiu mai general privind descompunerea matricei $A = TS$ în care elementele de pe diagonala lui T nu sînt neapărat unitatea.

Exemplu. Pentru $n = 3$ avem

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} t_{11} & & 0 \\ t_{21} & t_{22} & \\ t_{31} & t_{32} & t_{33} \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ & s_{22} & s_{23} \\ 0 & & s_{33} \end{bmatrix}.$$

După efectuarea produsului în partea dreaptă se obține

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} t_{11}s_{11} & t_{11}s_{12} & t_{11}s_{13} \\ t_{21}s_{11} & t_{21}s_{12} + t_{22}s_{22} & t_{21}s_{13} + t_{22}s_{23} \\ t_{31}s_{11} & t_{31}s_{12} + t_{32}s_{22} & t_{31}s_{13} + t_{32}s_{23} + t_{33}s_{33} \end{bmatrix}.$$

În urma procesului de identificare se obțin :

$$\begin{array}{lll} a_{11} = t_{11}s_{11} & a_{12} = t_{11}\,s_{12} & a_{13} = t_{11}\,s_{13} \\ a_{21} = t_{21}s_{11} & a_{22} = t_{21}\,s_{12} + t_{22}\,s_{22} & a_{23} = t_{21}\,s_{13} + t_{22}\,s_{23} \\ a_{31} = t_{31}s_{11} & a_{23} = t_{31}\,s_{13} + t_{32}\,s_{22} & a_{33} = t_{31}\,s_{13} + t_{32}\,s_{23} + t_{33}\,s_{33} \end{array}$$

Se observă că

$$t_{22}\,s_{22} = a_{22} - t_{21}\,s_{12}, \qquad s_{23} = \frac{1}{t_{22}}(a_{23} - t_{21}\,s_{13}),$$

$$t_{33}\,s_{33} = a_{33} - t_{31}\,s_{13} - t_{32}\,s_{23}, \qquad t_{32} = \frac{1}{s_{22}}(a_{32} - t_{31}s_{12}).$$

Rezultă relațiile :

$$t_{kk}\,s_{kk} = a_{kk} - \sum_{p=1}^{k} t_{kp}\,s_{pk}, \qquad k \geqslant 2. \tag{5.87}$$

Restul elementelor matricei S se exprimă astfel :

$$s_{kj} = \frac{1}{t_{kk}}\left(a_{kj} - \sum_{p=1}^{k-1} t_{kp}\,s_{pj}\right), \qquad 2 \leqslant k \leqslant j. \tag{5.88}$$

Se observă că elementele matricei T care nu sînt pe diagonala principală se exprimă astfel:

$$t_{ik} = \frac{1}{s_{kk}} \left(a_{ik} - \sum_{p=1}^{k-1} t_{ip}\, s_{pk} \right), \qquad 2 \leqslant k < i. \tag{5.89}$$

Pentru $k = 1$, $i = 2, 3, \ldots$, relaţia (5.89) permite determinarea elementelor primei coloane a matricei T dacă se cunoaşte s_{11}. De asemenea pentru $k = 1$, $j = 2,3,\ldots$ dacă se renunţă la Σ, relaţia (5.88) serveşte la determinarea elementelor primei linii a matricei S, dacă se cunoaşte elementul t_{11}. Ecuaţia (5.87) determină produsul $t_{kk}s_{kk}$ în funcţie de elementele din liniile precedente ale lui S şi coloanele precedente ale matricei T.

De îndată ce elementele t_{kk} şi s_{kk} sînt alese astfel ca să satisfacă ecuaţia (5.87), se pot utiliza (5.88) şi (5.89) pentru determinarea elementelor care au mai rămas necunoscute în linia, respectiv coloana k.

Dacă produsul $t_{kk}\, s_{kk} = 0$, factorizarea matricei A nu este posibilă, afară de cazul cînd toate parantezele din (5.88) şi (5.89) sînt nule, pentru $j > k$, respectiv $i > k$.

Dacă A este nesingulară, atunci se utilizează ca pivot elementul maxim de pe coloană, obţinîndu-se elementele pivot într-o secvenţă de forma $(i_1, 1), (i_2, 2), \ldots, (i_n, n)$, deci are loc o schimbare a liniilor între ele înainte de a începe procesul de descompunere a matricei A. În rezumat, dacă A este nesingulară, o decompunere triunghiulară $A = TS$ este posibil să nu se poată realiza, dar o permutare a liniilor lui A poate duce la

$$B = P^T A = TS, \tag{5.90}$$

unde $P = (p_{rs})$ şi

$$p_{rs} = \begin{cases} 0, & r \neq i_s, \\ 1, & r = i_s. \end{cases}$$

Matricea P poate fi găsită astfel ca

$$|t_{kk}| \geqslant |t_{ik}|, \quad i > k; \; k = 1, 2, \ldots, n-1. \tag{5.91}$$

O alegere simetrică $t_{kk} = s_{kk}$ poate conduce la numere imaginare dacă partea dreaptă din (5.87) este negativă.

În mod similar cu metoda lui Crout se poate considera vectorul $\mathbf{b}$ ca o coloană adițională a lui A, $a_{i,n+1} = b_i$ și utilizîndu-se relația (5.88) pentru $j = n + 1$, se găsesc componentele $b_i^{(n)} = s_{i,n+1}$ astfel că

$$S\mathbf{x} = T^{-1}\mathbf{b} = \mathbf{b}^{(n)}.$$

Prin procesul de eliminare inversă se determină soluția sistemului $x_n, x_{n-1}, \ldots, x_2, x_1$ cu ajutorul relațiilor:

$$x_n = b_n^{(n)}/s_{nn}, \quad x_i = \frac{1}{s_{ii}}\left(b_i^{(i)} - \sum_{j=i+1}^{n} s_{ij}\, x_i\right), \qquad (5.92)$$

$$i = n - 1,\ n - 2, \ldots, 2, 1.$$

Această metodă a factorizării este importantă pentru că generează o serie de metode de rezolvare a sistemelor de ecuații algebrice, metode care folosesc anumite particularități ale matricei A din sistemul considerat.

● *Metoda Cholesky* (*metoda rădăcinii pătrate*). Este o metodă exactă de rezolvare a sistemelor de ecuații algebrice liniare de forma $A\mathbf{x} = \mathbf{b}$ și urmărește punerea matricei A sub forma unui produs de două matrice inferior, respectiv superior triunghiulare.

Dacă A este nesingulară, atunci condiția necesară și suficientă pentru ca descompunerea matricei A să fie posibilă este ca minorii principali să îndeplinească condițiile

$$a_{11} \neq 0 \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \neq 0, \ldots, |A_{n-1,\,n-1}| \neq 0. \qquad (5.93)$$

Această metodă se aplică îndeosebi sistemelor în care matricea A este simetrică și pozitiv definită. În acest caz matricea A se poate pune sub forma

$$TS = A \text{ sau } TT^T = A, \qquad (5.94)$$

unde T este o matrice inferior triunghiulară, iar S este transpusa sa ($S = T^T$). Relaţia (5.94) se poate scrie dezvoltat sub forma

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} =$$

$$= \begin{bmatrix} t_{11} & & & & \\ t_{21} & t_{22} & & 0 & \\ \dots & \dots & \dots & & \\ t_{n1} & t_{n2} & t_{n3} & \dots & t_{nn} \end{bmatrix} \begin{bmatrix} t_{11} & t_{21} & t_{31} & \dots & t_{n1} \\ & t_{22} & t_{32} & \dots & t_{n2} \\ & 0 & \dots & \dots & \dots \\ & & & & t_{nn} \end{bmatrix}. \qquad (5.95)$$

Dacă se efectuează produsele primei linii a matricei T cu coloanele matricei T^T şi se face identificarea cu elementele din prima linie a matricei A, rezultă ecuaţiile

$$t_{11}^2 = a_{11},\ t_{11}t_{21} = a_{12},\ t_{11}t_{31} = a_{13}, \dots, t_{11}t_{n1} = a_{1n}.$$

Din aceste ecuaţii rezultă elementele primei linii din matricea T^T.

În urma produsului liniei a doua a matricei T cu toate coloanele matricei T^T şi a identificării cu elementele liniei a doua din matricea A, rezultă ecuaţiile

$$t_{21}t_{11} = a_{21}, t_{21}^2 + t_{22}^2 = a_{22}, t_{21}t_{31} + t_{22}t_{32} = a_{23}, \dots$$

$$\dots,\ t_{21}t_{n1} + t_{22}\,t_{n3} = a_{2n},$$

de unde rezultă elementele liniei a doua din matricea T^T ş.a.m.d.

În cazul produsului dintre linia i a matricei T cu toate coloanele matricei T^T şi a identificării cu elementele liniei i din matricea A, rezultă ecuaţiile

$$\begin{aligned} &t_{i1}t_{j1} + t_{i2}t_{j2} + \dots + t_{ii}t_{ji} = a_{ij}, \qquad i < j \qquad (5.96) \\ &t_{i1}^2 + t_{i2}^2 + \dots + t_{i,i-1}^2 + t_{ii}^2 = a_{ii}, \quad i > 1. \end{aligned}$$

Matricea A a fost considerată pozitiv definită, înseamnă că $a_{ii} > 0$; în plus matricea A a fost considerată simetrică (adică $a_{ij} = a_{ji}$). Deci elementele matricei T se pot calcula cu următoarele relații :

$$t_{11} = a_{11},\ t_{i1} = \frac{a_{1j}}{a_{11}},\quad i,j = 2,\ldots,n,$$

$$t_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} t_{i,k}^2} \qquad i > 1, \tag{5.97}$$

$$t_{ji} = \frac{1}{t_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} t_{ik}\, t_{jk}\right), \qquad j > i.$$

Cu ajutorul relațiilor (5.98) se pot calcula toate elementele matricei T^T, iar descompunerea $A = TT^T$ conduce la rezolvarea a două sisteme de ecuații algebrice liniare cu matrice triunghiulare. Sistemul $A\mathbf{x} = \mathbf{b}$ se descompune în două sisteme

$$T\mathbf{x} = \mathbf{y},\ T^T\mathbf{y} = \mathbf{b}, \tag{5.98}$$

sisteme care se descompun în forma scalară astfel :

$$\left.\begin{aligned} t_{11}y_1 + t_{21}y_2 + \ldots + t_{n1}y_n &= b_1 \\ t_{22}y_2 + \ldots + t_{n2}y_n &= b_2 \\ \ldots\ldots\ldots\ldots & \\ t_{nn}y_n &= b_n \end{aligned}\right\} \quad \left.\begin{aligned} t_{11}x_1 &= y_1 \\ t_{21}x_1 + t_{22}x_2 &= y_2 \\ \ldots\ldots\ldots & \\ t_{n1}x_1 + t_{n2}x_2 + \ldots + t_{nn}x_n &= y_n \end{aligned}\right\} \tag{5.99}$$

Din primul sistem se determină $y_n, y_{n-1}, \ldots, y_2, y_1$ cu ajutorul relațiilor

$$y_n = \frac{b_n}{t_{nn}},\ y_i = \frac{1}{t_{ii}}\left(b_i - \sum_{k=i+1}^{n} t_{ki}\, x_k\right),\ i = n-1, n-2,\ldots,1,$$

iar din al doilea sistem se determină necunoscutele $x_1, x_2, \ldots, x_n$ prin intermediul relațiilor

$$x_1 = \frac{y_1}{t_{11}},\ x_i = \frac{1}{t_{11}}\left(y_i - \sum_{k=1}^{n} t_{ik}\, x_k\right),\ i = 2,\ldots,n.$$

5.2.5. *Metode de rezolvare a sistemelor de ecuații liniare în cazul matricelor de formă specială*

De foarte multe ori se întîlnesc în aplicații sisteme de ecuații algebrice liniare, a căror matrice A are o formă specială.

● *Matricea A are forma tridiagonală sau jacobiană,* dacă elementele matricei sînt nule pentru $|i-j| > 1$, unde i reprezintă indicele de linie iar j indicele de coloană, adică

$$A = \begin{bmatrix} a_1 & c_1 & & & & & \\ b_2 & a_2 & c_2 & & & 0 & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & & & \\ & 0 & & b_{n-1} & a_{n-1} & c_{n-1} & \\ & & & & b_n & a_n & \end{bmatrix}. \quad (5.100)$$

În cazul în care matricea A poate fi factorizată sub forma a două matrice bidiagonale de forma

$$A = TS = \begin{bmatrix} t_1 & & & & & 0 & \\ p_2 & t_2 & & & & & \\ & p_3 & t_3 & & & & \\ & & \ddots & \ddots & & & \\ & & & \ddots & \ddots & & \\ & 0 & & p_{n-1} & t_{n-1} & & \\ & & & & p_n & t_n & \end{bmatrix} \begin{bmatrix} 1 & s_1 & & & & 0 & \\ & 1 & s_2 & & & & \\ & & 1 & s_3 & & & \\ & & & \ddots & \ddots & & \\ & & & & \ddots & & \\ & 0 & & & & s_{n-1} & \\ & & & & & 1 & \end{bmatrix}, \quad (5.101)$$

elementele matricelor T și S se determină cu ajutorul relațiilor :

$$\begin{aligned} &t_1 = a_1, && s_1 = c_1/t_1 = c_1/a_1, \\ &t_i = a_i - p_i\, s_{i-1}, && i = 2, 3, \ldots, n, \\ &s_i = c_i/t_i, && i = 2, 3, \ldots, n-1, \\ &p_i = b_i, && i = 2, 3, \ldots, n. \end{aligned} \quad (5.102)$$

În cazul în care $t_i \neq 0$ pentru $i = 1, 2, \ldots, n$, factorizarea lui A este posibilă și elementele matricei T și S se determină cu relațiile date în (5.102).

În cazul în care matricea A este o matrice a sistemului $A\mathbf{x} = \mathbf{b}$, și au fost determinate matricele T și S, atunci $T\mathbf{b}^{(n)} = \mathbf{b}$, unde

$$b_1^{(n)} = b_1/t_1, \; b_i^{(n)} = (b_i - p_i s_{i-1})/t_i, \quad i = 2, \ldots, n. \tag{5.103}$$

În final rezolvarea sistemului $A\mathbf{x} = \mathbf{b}$ se reduce la rezolvarea sistemului $S\mathbf{x} = \mathbf{b}^{(n)}$, ale cărui soluții sînt date de relațiile

$$x_n = b_n^{(n)}, \; x_i = b_i^{(n)} - s_i x_{i+1}, \; i = n-1, n-2, \ldots, 2, 1. \tag{5.104}$$

● *Matricea A este o matrice bloc tridiagonală.* La rezolvarea numerică a ecuațiilor cu derivate parțiale și în cazul ecuațiilor integrale se întîlnesc destul de frecvent matrice bloc tridiagonale de forma :

$$A = \begin{bmatrix} A_1 & C_1 & & & 0 \\ B_2 & A_2 & C_2 & & \\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & \ddots & C_{n-1} \\ & & & B_n & A_n \end{bmatrix}, \tag{5.105}$$

unde A_i este matrice pătrată de ordinul m_i, iar B_i trebuie să fie matrice de dimensiune $m_i \times m_{i-1}$ și C_i matrice de ordinul $m_i \times m_{i+1}$. În cazul în care toate valorile m_i sînt egale cu m, atunci toate submatricele sînt pătrate de ordinul m. Ordinul matricei A este $\sum_{i=1}^{n} m_i$, iar pentru $m_i = m$ ordinul matricei A este $m \times n$.

Un sistem de ecuații care are o matrice de forma (5.105) se poate rezolva printr-un procedeu analog ca în cazul factorizării matricei Jacobi.

Fie sistemul de forma $A\mathbf{x} = \mathbf{b}$, unde matricea A este de forma (5.105) iar

$$\mathbf{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ x^{(n)} \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b^{(1)} \\ b^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ b^{(n)} \end{bmatrix},$$

fiecare $x^{(i)}$, $b^{(i)}$ are m_i componente din vectorii coloană $\mathbf{x}$ și $\mathbf{b}$. Componentele vectorului $\mathbf{x}$ sînt grupate în subsisteme $x^{(i)}$, care sînt eliminate ca în metoda lui Gauss. Atunci

$$A = TS = \begin{bmatrix} T_1 & & & & \\ P_2 & T_2 & & \mathbf{0} & \\ & P_3 & T_3 & & \\ & & \cdot & \cdot & \\ \mathbf{0} & & \cdot & \cdot & \\ & & & P_n & T_n \end{bmatrix} \begin{bmatrix} I_1 & S_1 & & & \\ & I_2 & S_2 & & \mathbf{0} \\ & & I_3 & S_3 & \\ & & & \cdot & \cdot \\ \mathbf{0} & & & \cdot & S_{n-1} \\ & & & & I_n \end{bmatrix}, \tag{5.106}$$

unde I_i sînt matrice unitate de ordinul m_j, T_j sînt matrice pătrate de ordinul m_j și S_j sînt matrice cu m_j linii și m_{j+1} coloane.

Elementele (matrice) ale matricelor bloc cu două diagonale T și S se determină cu ajutorul relațiilor matriceale :

$$\begin{aligned} &T_1 = A_1, && S_1 = A_1^{-1} C_1; \\ &T_i = A_i - B_i S_{i-1}, && i = 2, 3, \ldots, n, \\ &S_i = A_i^{-1} C_i, && i = 2, 3, \ldots, n-1. \end{aligned} \tag{5.107}$$

Din definiția produsului a două matrice apare evident că matricele S_i sînt de ordinul $m_i \times m_{i+1}$ și produsele $B_i S_{i-1}$ și deci A_i sînt matrice pătrate de ordinul m_i. În

acest fel sistemul $A\mathbf{x} = \mathbf{b}$ este echivalent cu $T\mathbf{y} = \mathbf{b}$, $S\mathbf{x} = \mathbf{y}$ sau, sub formă dezvoltată, se poate scrie astfel:

$$\begin{bmatrix} T_1 & & & \\ P_2 & T_2 & 0 & \\ & \ddots & \ddots & \\ 0 & & P_n & T_n \end{bmatrix} \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{b}^{(1)} \\ \mathbf{b}^{(2)} \\ \vdots \\ \mathbf{b}^{(n)} \end{bmatrix};$$

$$\begin{bmatrix} I_1 & S_1 & & \\ & I_2 & S_2 & 0 \\ & & \ddots & \ddots \\ 0 & & & S_{n-1} \\ & & & I_n \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \vdots \\ \mathbf{x}^{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ \vdots \\ \mathbf{y}^{(n)} \end{bmatrix} \tag{5.108}$$

Primul sistem din (5.108) se poate scrie sub forma

$$T_1\mathbf{y}^{(1)} = \mathbf{b}^{(1)}, \quad T_i\mathbf{y}^{(i)} + P_i\,\mathbf{y}^{(i-1)} = \mathbf{b}^{(i)}, \quad i = 2,\ldots,n, \tag{5.109}$$

iar al doilea sistem se poate scrie astfel:

$$I_i\mathbf{x}^{(i)} + S_i\mathbf{x}^{i+1} = \mathbf{y}^{(i)}, \quad i = 1, 2,\ldots,n-1, \tag{5.110}$$
$$I_n\mathbf{x}^{(n)} = \mathbf{y}^{(n)}.$$

Din relațiile matriceale (5.109) și (5.110) rezultă

$$\mathbf{y}^{(1)} = T_1^{-1}\mathbf{b}^{(1)}, \; \mathbf{y}^{(i)} = A_i^{-1}\,(\mathbf{b}^{(i)} - B_i\,\mathbf{y}^{(i-1)}), \; i = 2, 3,\ldots,n, \tag{5.111}$$

și

$$\mathbf{x}^{(n)} = \mathbf{y}^{(n)}, \mathbf{x}^{(i)} = \mathbf{y}^{(i)} - S_i\mathbf{x}^{(i+1)}, i = n-1, n-2,\ldots,2,1.$$

● *Matricea A este o matrice cu elemente numere complexe.* Dacă elementele lui A și $\mathbf{b}$ sînt complexe, soluția sistemului $A\mathbf{x} = \mathbf{b}$ va fi în general formată din numere

complexe. Soluția în acest caz poate fi găsită prin oricare din metodele prezentate. Dacă A este o matrice complexă, atunci

$$A = B + \mathrm{i}C, \mathbf{b} = \mathbf{c} + \mathrm{i}\mathbf{d}, \mathbf{x} = \mathbf{y} + \mathrm{i}\mathbf{z}. \qquad (5.112)$$

În acest caz sistemul de ecuații $A\mathbf{x} = \mathbf{b}$ devine

$$(B + \mathrm{i}C)(\mathbf{y} + \mathrm{i}\mathbf{z}) = \mathbf{c} + \mathrm{i}\mathbf{d}$$

iar după efectuarea calculelor se obțin două ecuații matriceale reale

$$B\mathbf{y} - C\mathbf{z} = \mathbf{c}, \; C\mathbf{y} + B\mathbf{z} = \mathbf{d}. \qquad (5.113)$$

Sistemul de ecuații matriceale se poate scrie matriceal partiționat sub forma

$$\left[\begin{array}{c|c} B & -C \\ \hline C & B \end{array}\right] \left[\begin{array}{c} \mathbf{y} \\ \hline \mathbf{z} \end{array}\right] = \left[\begin{array}{c} \mathbf{c} \\ \hline \mathbf{d} \end{array}\right] \qquad (5.114)$$

Matricea sistemului (5.114) este de ordinul $2n$, iar numărul de operații aritmetice implicat în rezolvare este aproximativ de opt ori mai mare decît în cazul real pentru o matrice de ordinul n.

Exemplu. Folosindu-se teoria clasică a algebrei numerelor complexe, se poate face o analogie între calcul cu numere de forma $a + ib$, și matrice de tipul $Ia + \beta J$, unde

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ și } J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \qquad (5.115)$$

Pentru că $I^2 = I$, $J^2 = -I$, $IJ = JI = J$, avem

$$(aI + bJ)^2 = (a^2 - b^2)I + 2abJ, \qquad (a + \mathrm{i}b)^2 = a^2 - b^2 + 2ab\mathrm{i}$$

$$(aI + bJ)^{-1} = (a^2 + b^2)^{-1}(aI - bJ), \; (a + b\mathrm{i})^{-1} = (a^2 + b^2)^{-1}(a - b\mathrm{i}).$$

Din aceste exemple se observă legătura dintre algebra numerelor complexe $a + b\mathrm{i}$ și algebra matricelor de tipul $aI + bJ$.

Se observă că se poate înlocui fiecare element $a_{rs} + ib_{rs}$ al matricei A printr-o matrice de forma

$$a_{rs} I + b_{rs} J = a_{rs} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + b_{rs} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} a_{rs} & b_{rs} \\ -b_{rs} & a_{rs} \end{bmatrix} \cdot \quad (5.116)$$

Dacă se consideră sistemul de ecuații

$$\left.\begin{aligned} (3 + 5i)x + (1 + i)y &= -9 + 19i \\ (2 + 4i)x - (1 + 3i)y &= -25 + 17i \end{aligned}\right\}$$

care are soluția $x = x_1 + iy_1 = 2 + 4i$, $y = x_2 + iy_2 = 1 - 4i$, acesta poate fi scris matricial astfel:

$$\begin{bmatrix} 3 + 5i & 1 & +i \\ 2 + 4i & -1 & -3i \end{bmatrix} \begin{bmatrix} x_1 + iy_1 \\ x_2 + iy_2 \end{bmatrix} = \begin{bmatrix} -9 + 19i \\ -25 + 17i \end{bmatrix}.$$

Folosind (5.114), sistemul considerat devine

$$\left[\begin{array}{cc:cc} 3 & 1 & -5 & -1 \\ 2 & -1 & -4 & 3 \\ \hdashline 5 & 1 & 3 & 1 \\ 4 & -3 & 2 & -1 \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ \hdashline y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} -9 \\ -25 \\ \hdashline 19 \\ 17 \end{bmatrix},$$

ce se poate rezolva prin oricare din metodele prezentate anterior. Dacă se folosesc relațiile (5.116), orice element al matricei A se poate înlocui printr-o matrice de 2×2, iar sistemul devine

$$\begin{bmatrix} 3 & 5 & 1 & 1 \\ -5 & 3 & -1 & 1 \\ 2 & 4 & -1 & -3 \\ -4 & 2 & 3 & -1 \end{bmatrix} \begin{bmatrix} x_1 & y_1 \\ -y_1 & x_1 \\ x_2 & y_2 \\ -y_2 & x_2 \end{bmatrix} = \begin{bmatrix} -9 & 19 \\ -19 & -9 \\ -25 & 17 \\ -17 & -25 \end{bmatrix}.$$

După efectuarea produselor, rezultă două sisteme echivalente:

$$\left.\begin{aligned} 3x_1 - 5y_1 + x_2 - y_2 &= -9 \\ -5x_1 - 3y_1 - x_2 - y_2 &= -19 \\ 2x_1 - 4y_1 - x_2 + 3y_2 &= -25 \\ -4x_1 - 2y_1 + 3x_2 + y_2 &= -17 \end{aligned}\right\} \qquad \left.\begin{aligned} 3y_1 + 5x_1 + y_2 + x_2 &= 19 \\ -5y_1 + 3x_1 - y_2 + x_2 &= -9 \\ 2y_1 + 4x_1 - y_2 - 3x_2 &= 17 \\ -4y_1 + 2x_1 + 3y_2 - x_2 &= -25 \end{aligned}\right\}$$

care după rezolvare conduc la aceeași soluție.

5.2.6. *Analiza comparativă a metodelor*

La începutul acestui capitol au fost prezentate o serie de elemente privind condiționarea sistemelor, eroarea metodelor și precizia soluției. În cadrul acestui paragraf se va acorda atenție numărului de operații aritmetice și necesarului de memorie.

● Metoda lui Gauss de eliminare folosită la rezolvarea unui sistem de n ecuații $A\mathbf{x} = \mathbf{b}$, prin care A este redusă la o matrice superior triunghiulară cu ajutorul eliminării directe, iar aceleași operații sînt efectuate și asupra vectorului $\mathbf{b}$ [32, 47], implică $n\left(\frac{1}{3}n^2 + \frac{1}{2}n - \frac{5}{6}\right)$ operații de înmulțire și $n\left(\frac{1}{3}n^2 - \frac{1}{3}\right)$ operații de adunare.

Pentru substituția inversă necesară calculării necunoscutelor $x_n, x_{n-1}, \ldots, x_2, x_1$ sînt necesare $n\left(\frac{1}{3}n^2 + n - \frac{1}{3}\right)$ operații de înmulțire și $n\left(\frac{1}{3}n^2 + \frac{1}{2}n - \frac{5}{6}\right)$ adunări.

Relativ la necesarul de memorie este evident faptul că nu este necesară o nouă zonă de memorie pentru sistemul redus, coeficienții sistemului redus se scriu peste coeficienții sistemului precedent, iar multiplicatorii pot fi scriși la adresele din memorie unde au fost scriși coeficienții ce au fost eliminați.

În cazul în care se folosesc schimbări de linii sau coloane pentru găsirea pivotului maxim, se execută aceleași operații dar apare necesitatea declarării unei zone de memorie suplimentară pentru realizarea schimbului de linii sau coloane, calculul unor determinanți și memorarea unor informații suplimentare.

● Metoda Gauss-Jordan, care transformă matricea A într-o matrice unitate, implică pentru găsirea soluției $\frac{1}{2}n^3 + n^2 - \frac{1}{2}n$ operații de înmulțire și $\frac{1}{2}n^3 - \frac{1}{2}n$ operații de adunare.

• Metoda Gauss de eliminare prin descompunerea matricei $A = TS$, $T\mathbf{b} = \mathbf{y}$, $S\mathbf{x} = \mathbf{y}$, implică $\frac{1}{3} n^3 + n^2 - \frac{1}{2} n$ operații de înmulțire și $\frac{1}{3} n^3 + \frac{1}{2} n^2 - \frac{5}{6} n$ operații de adunare.

• Pentru calculul matricei T din $A = TT^T$ sînt necesare $\frac{1}{6} n^3 + \frac{1}{2} n^2 - \frac{2}{3} n$ operații de înmulțire și $\frac{1}{6} n^3 - \frac{1}{6} n$ operații de adunare iar pentru rezolvarea sistemului $T\mathbf{y} = \mathbf{b}$, $T^T\mathbf{x} = \mathbf{y}$ sînt necesare $\frac{1}{2} n(n + 1)$ operații de înmulțire și $\frac{1}{2} n(n - 1)$ operații de adunare.

• În cazul matricei A tridiagonale, $A = TS$, sînt necesare $5n-4$ operații elementare.

• În cazul cînd matricea A este o matrice bloc tridiagonală, cu matricele de dimensiunile date, sînt necesare în total $(3n-2)(m^3 + m^2)$ operații elementare.

Se *pot face următoarele observații privind metodele prezentate*:

— Metoda eliminării și substituția inversă sînt întotdeauna metode destul de rapide.

— Metoda Gauss-Jordan este considerată lentă față de celelalte metode, dar programarea este mult mai simplă, deoarece nu implică eliminarea inversă.

— Pentru rezolvarea unui sistem de ecuații $A\mathbf{x} = \mathbf{b}$, eliminarea compactă cu schimbarea liniilor sau coloanelor este superioară tuturor metodelor cînd produsul scalar poate fi acumulat precis, deoarece nici calculele nici necesarul de memorie nu sînt mult mai mari.

— În cazul matricei A simetrice, descompunerea $A = TT^T$ are un *avantaj semnificativ*, deoarece chiar dacă $|a_{rs}| < 1$, atunci orice $|s_{rs}| < 1$ și nu apare problema scalării, în plus calculul produsului TT^T duce la $A + \delta A$, unde $|(\delta A)_{rs}|$ nu depășește o singură eroare de rotunjire dacă produsele scalare sînt acumulate corect.

În cazul sistemelor slab condiționate (unde erorile de rotunjire au efect serios), odată ce soluția a fost obținută prin oricare din metodele prezentate, se pune problema îmbunătățirii ei printr-un proces iterativ. Prima etapă în cadrul acestui proces iterativ este calculul vectorului rezidual

$$\mathbf{r}^{(1)} = \mathbf{b} - A\mathbf{x}^{(1)}, \tag{5.117}$$

unde $\mathbf{x}^{(1)}$ este soluția calculată cu ajutorul calculatorului. Componentele vectorului $\mathbf{r}^{(1)}$ sînt relativ mici, în sensul, că $\mathbf{x}^{(1)}$ are anumite cifre corecte în fiecare componentă.

Dacă există erori în soluția calculată $\mathbf{x}^{(1)}$, atunci diferență dintre $\mathbf{x}^{(1)}$ și vectorul $\mathbf{x}$ (soluția exactă) reprezintă vectorul eroare

$$\mathbf{e}^{(1)} = \mathbf{x} - \mathbf{x}^{(1)}. \tag{5.118}$$

Amplificînd vectorul eroare $\mathbf{e}^{(1)}$ cu matricea A, rezultă

$$A\mathbf{e}^{(1)} = A(\mathbf{x} - \mathbf{x}^{(1)}) = A\mathbf{x} - A\mathbf{x}^{(1)} = \mathbf{b} - A\mathbf{x}^{(1)} = \mathbf{r}^{(1)}$$

sau

$$A\mathbf{e}^{(1)} = \mathbf{r}^{(1)}. \tag{5.119}$$

Din (5.119) se observă că vectorul eroare este soluția sistemului liniar de matrice A și partea dreaptă vectorul rezidual $\mathbf{r}^{(1)}$. Din (5.118) se vede că soluția exactă se obține din

$$\mathbf{x} = \mathbf{x}^{(1)} + \mathbf{e}^{(1)}, \tag{5.120}$$

unde $\mathbf{x}^{(1)}$ este soluția calculată iar $\mathbf{e}^{(1)}$ este soluția sistemului (5.119).

De asemenea trebuie menționat că nu se poate rezolva (5.119) și obține $\mathbf{e}^{(1)}$ ci o aproximație a lui $\mathbf{e}^{(1)}$ [chiar dacă $\mathbf{r}^{(1)}$ este calculat exact], pentru că dacă s-ar putea rezolva ecuația $A\mathbf{x} = \mathbf{b}$, discuția s-ar fi încheiată.

Dacă se notează soluția calculată a sistemului (5.119) prin $\boldsymbol{\varepsilon}^{(1)}$ care este o aproximație a soluției exacte $\mathbf{e}^{(1)}$, atunci se poate scrie relația (5.120) sub forma

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \boldsymbol{\varepsilon}^{(1)}, \tag{5.121}$$

unde $\mathbf{x}^{(2)}$ reprezintă o soluție îmbunătățită a lui $\mathbf{x}$.

În general, dacă se dispune de soluția $\mathbf{x}^{(i)}$ și se dorește calculul unei soluții îmbunătățite $\mathbf{x}^{(i+1)}$, se procedează în felul următor :

— se calculează cu o precizie cît mai mare

$$\mathbf{r}^{(i)} = \mathbf{b} - A\mathbf{x}^{(i)}; \tag{5.122}$$

— cu vectorul $\mathbf{r}^{(i)}$ calculat se rezolvă sistemul

$$A\mathbf{e}^{(i)} = \mathbf{r}^{(i)}, \tag{5.123}$$

obținîndu-se o valoare aproximativă $\boldsymbol{\varepsilon}^{(i)}$ pentru soluția exactă $\mathbf{e}^{(i)}$;

— se calculează soluția îmbunătățită $\mathbf{x}^{(i+1)}$ prin intermediul relației

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} + \boldsymbol{\varepsilon}^{(i)}. \tag{5.124}$$

Se observă că algoritmul de îmbunătățire a soluției obținute prin metodele exacte implică la fiecare etapă rezolvarea sistemului de ecuații (5.123) în care apare matricea A. Dacă la determinarea lui $\mathbf{x}^{(1)}$ s-a folosit algoritmul de descompunere a lui A în TS și matricele T și S sînt în memorie, atunci operațiile de calcul pentru rezolvarea sistemului (5.123) se reduc mult dacă se folosește doar matricea T și algoritmul de descompunere a ales ca matricea T să fie inferior triunghiulară, dar unitară.

5.3. Metode iterative pentru rezolvarea sistemelor algebrice liniare

Metodele de eliminare și metodele iterative folosite la rezolvarea sistemelor de ecuații de forma $A\mathbf{x} = \mathbf{b}$ au o serie de avantaje și de dezavantaje.

Metodele de eliminare (metodele directe) au avantajul că necesită un număr fix de operații elementare pentru un număr de ecuații dat și au dezavantajul că acumulează

erori de rotunjire, care conduc la o eroare relativă destul de mare în soluție. Eroarea de rotunjire poate fi minimizată prin reordonarea ecuațiilor sistemului după fiecare etapă (pentru menținerea elementelor maximale pe diagonală), operație destul de anevoioasă pentru sistemele mari.

Metodele iterative au dezavantajul că apare posibilitatea de neconvergență (alegerea unei valori de strat necorespunzătoare sau a unui sistem slab condiționat), acestea implică un număr mare de operații elementare chiar atunci cînd converg destul de rapid, de asemenea se pot acumula erori de rotunjire destul de însemnate.

Metodele iterative pentru rezolvarea sistemului $A\mathbf{x}=\mathbf{b}$ constă din alegerea unui vector inițial $\mathbf{x}^{(0)}$ pentru vectorul soluție și cu ajutorul unui algoritm de calcul iterativ se determină un șir de soluții aproximative, $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$,...,$\mathbf{x}^{(n)}$,... care în principiu trebuie să conveargă către soluția, exactă a sistemului. Aceste metode sînt iterative în sensul că se trunchiază șirul infinit de operații în momentul cînd este atinsă precizia dorită, pe cînd metodele directe sînt finite deoarece implică un număr finit de operații și bine determinat de dimensiunea sistemului și de metoda aleasă.

Unul din avantajele importante ale metodelor iterative este faptul că eroarea de rotunjire și chiar cea de trunchiere pot fi eliminate. Unele metode iterative pot fi utilizate la îmbunătățirea soluției obținute prin metode directe (5.2.6). În general metodele iterative sînt folosite la sistemele la care convergența este rapidă, matricele sînt mari dar conțin un mare număr de zerouri, pentru care metodele de eliminare sînt laborioase și necesită un spațiu mare de memorie.

Fie $A\mathbf{x}=\mathbf{b}$ un sistem liniar de n ecuații algebrice cu n necunoscute, neomogene și $\mathbf{x}^{(0)} \in R^n$ un vector oarecare. Se presupune că plecînd de la $\mathbf{x}^{(0)}$, se construiește un șir de vectori $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$,..., pe baza unei formule recurente de forma

$$\mathbf{x}^{(k+1)} = F_k(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}), \qquad (5.125)$$

unde F_k este o funcție vectorială nu neapărat liniară de argumentele sale, în general ea depinde de matricea A a

sistemului considerat. Dacă şirul de vectori definit recurent de relaţia (5.125) tinde către soluţia $\mathbf{x} = A^{-1}\mathbf{b}$ a sistemului considerat, se poate afirma că s-a construit o metodă iterativă (de aproximaţii succesive) pentru rezolvarea sistemului, sau că şirul definit este un şir de iterare. Dacă funcţia F_k este liniară pentru orice k, procesul de aproximaţii succesive este liniar; cînd F_k depinde de un singur vector, procesul de aproximare este de ordinul întîi sau staţionar.

O mare clasă de metode iterative pot fi definite astfel. Fie sistemul de rezolvat

$$A\mathbf{x} = \mathbf{b}, \tag{5.126}$$

unde det $|A| \neq 0$. Atunci matricea A poate fi descompusă într-un număr infinit de moduri sub forma $A = B - C$, unde B şi C sînt matrice de acelaşi ordin cu A. Sistemul (5.126) se poate scrie sub forma

$$(B - C)\mathbf{x} = \mathbf{b} \text{ sau } B\mathbf{x} = C\mathbf{x} + \mathbf{b}. \tag{5.127}$$

Dacă se foloseşte o valoare de start $\mathbf{x}^{(0)}$, se poate calcula un şir de vectori $\{\mathbf{x}^{(k)}\}_{k\in N}$ cu ajutorul formulei iterative

$$B\mathbf{x}^{(k+1)} = C\mathbf{x}^{(k)} + \mathbf{b}, \quad k = 0, 1, 2, \ldots \tag{5.128}$$

La descompunerea matricei A se ţine seama de îndeplinirea următoarelor cerinţe: det $B \neq 0$ şi sistemul $B\mathbf{y} = \mathbf{b}$ se rezolvă destul de uşor.

În cazul în care se cere un grad de precizie mai mare, şirul de vectori se va calcula cu o formă echivalentă a relaţiei (5.128), înlocuindu-se matricea C prin $B - A$, iar (5.128) devine

$$B\mathbf{x}^{(k)} = (B - A)\mathbf{x}^{k-1} + \mathbf{b} \text{ sau } B(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) = \mathbf{b} - A\mathbf{x}^{k-1}. \tag{5.129}$$

Pentru analiza convergenţei şirului de vectori $\{\mathbf{x}^{(k)}\}_{k\in N}$, format din soluţiile aproximative (obţinute iterativ),

către vectorul soluției exactă $\mathbf{x}$, se impune introducerea unei matrice $P = B^{-1}C$ și a vectorului eroarea

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}, \quad k = 0, 1, 2, \ldots \tag{5.130}$$

Dacă se face diferența dintre ecuațiile matriceale

$$B\mathbf{x} = C\mathbf{x} + b, \; B\mathbf{x}^{(k)} = C\mathbf{x}^{(k-1)} + \mathbf{b},$$

rezultă

$$B(\mathbf{x}^{(k)} - \mathbf{x}) = C(\mathbf{x}^{(k-1)} - \mathbf{x}) \text{ sau } B\mathbf{e}^{(k)} = C\mathbf{e}^{(k-1)}. \tag{5.131}$$

S-a presupus că B este nesingulară, deci există B^{-1}, atunci (5.131) devine

$$\mathbf{e}^{(k)} = B^{-1}C\mathbf{e}^{(k-1)} = P\mathbf{e}^{(k-1)} = P^2\mathbf{e}^{(k-2)} = \ldots = P^{(k)}\mathbf{e}^{(0)},$$
$$k = 1, 2, \ldots, \tag{5.132}$$

unde $\mathbf{e}^{(0)}$ este vectorul eroare inițială, ales arbitrar.

După introducerea acestor elemente, o condiție suficientă pentru convergența șirului $\{x^{(k)}\}_{k \in N} \to \mathbf{x}$ cînd $k \to \infty$ este

$$\lim_{k \to \infty} \mathbf{e}^{(k)} = \mathbf{o}, \text{ adică } \lim_{k \to \infty} P^{(k)} = 0. \tag{5.133}$$

Această condiție este și necesară dacă metoda converge pentru orice vector $\mathbf{e}^{(0)}$.

Teoremă. *Matricea P este convergentă, adică*

$$\lim_{k \to \infty} P^k = 0,$$

dacă și numai dacă toate valorile proprii ale matricei P sînt în modul subunitare [*sau raza spectrală* $\rho(P) < 1$, *unde* $\rho(P) = \max_i |\lambda_i|$, λ_i, $i = 1, 2, \ldots, n$, *fiind valorile proprii ale matricei P*].

Verificarea condiției din teoremă este dificilă din punct de vedere practic, deoarece problema determinării valorilor

proprii ale matricei P este mai laborioasă decît rezolvarea unui sistem de ecuații algebrice liniare. Datorită acestui fapt se va folosi o condiție mult mai convenabilă și anume că matricea $P = (p_{ij})$ este convergentă dacă sau norma matriceală $\|\cdot\|_\infty$ sau $\|\cdot\|_1$ este subunitară, adică

$$\|P\|_\infty = \max_i \sum_{j=1}^{n} P_{ij} < 1, \quad \|P\| = \max_j \sum_{i=1}^{n} p_{ij} < 1. \quad (5.134)$$

Pentru a justifica afirmațiile anterioare, fie $\mathbf{x} \in R^n$ fixat, $P \in R^{nn}$, $\mathbf{c} \in R^n$ astfel că

$$\bar{\mathbf{x}} = P\bar{\mathbf{x}} + \mathbf{c}. \quad (5.135)$$

Dacă $\|P\|_{\infty,1} < 1$, atunci $\lim_{k\to\infty} \mathbf{x}^{(k)} = \bar{\mathbf{x}}$, unde $\mathbf{x}^{(k)}$ este definit de relațiile

$$\mathbf{x}^1 = \mathbf{x}, \ \mathbf{x}^{(k+1)} = P\mathbf{x}^{(k)} + \mathbf{c}, \ (\forall)\mathbf{x} \in R^n. \quad (5.136)$$

Făcîndu-se diferența dintre relațiile (5.136) și (5.135), rezultă

$$\bar{\mathbf{x}} - x^{(k+1)} = P\bar{\mathbf{x}} + \mathbf{c} - (P\mathbf{x}^{(k)} + \mathbf{c}) = P(\bar{\mathbf{x}} - \mathbf{x}^{(k)}) =$$
$$= P[P\bar{\mathbf{x}} + \mathbf{c} - (P\mathbf{x}^{(k-1)} + \mathbf{c}) = P^2[\bar{\mathbf{x}} - \mathbf{x}^{(k-1)}] = \ldots$$
$$\ldots = P^{(k)}[\bar{\mathbf{x}} - \mathbf{x}^{(1)}]. \quad (5.137)$$

Dacă se aplică norma vectorială în stînga și norma matriceală în dreapta și se folosesc proprietățile normei și legătura dintre norma vectorială și norma matriceală, rezultă

$$\|\bar{\mathbf{x}} - \mathbf{x}^{(k+1)}\| \leqslant \|P^k(\bar{\mathbf{x}} - \mathbf{x}^{(1)})\| \leqslant \|P^k\| \, \|\bar{\mathbf{x}} - \mathbf{x}^{(1)}\| \leqslant$$
$$\leqslant (\|P\|)^k \|\mathbf{x} - \mathbf{x}^{(1)}\|.$$

Din această ultimă relație se vede că dacă $\|P\| < 1$, atunci $(\|P\|)^k \to 0$, pentru $k \to \infty$ și deci $\|\bar{\mathbf{x}} - \mathbf{x}^{(k+1)}\| \to 0$ pentru $k \to \infty$, adică $\mathbf{x}^{(k+1)} \to \bar{\mathbf{x}}$ pentru $k \to \infty$.

Revenind la schemele iterative (5.128)—(5.132) şi presupunînd că sînt convergente, se poate introduce rata de convergenţă R a schemei iterative definite astfel :

$$R = - \lg \rho(P). \tag{5.138}$$

Semnificaţia acestei mărimi este uşor de evidenţiat, dacă se foloseşte relaţia $\rho(P) = \|P\|$ (pentru mulţimea normelor naturale).

Dacă se cunoaşte vectorul eroarea iniţială $\mathbf{e}^{(0)}$, (5.132) permite estimarea în funcţie de orice normă naturală

$$\|\mathbf{e}^{(k)}\| \leqslant \|P\|^k \|e^{(0)}\|. \tag{5.139}$$

Pentru un $\varepsilon > 0$ există o anumită normă astfel că

$$\|\mathbf{e}^{(k)}\| \leqslant [\rho(P) + \varepsilon]^k \|\mathbf{e}^{(0)}\|. \tag{5.140}$$

Din nou folosind relaţia (5.133), dacă vectorul eroare $\mathbf{e}^{(0)}$ este vectorul propriu al matricei P, corespunzător valorii proprii maxime, atunci

$$\|\mathbf{e}^{(k)}\| = [\rho(P)]^k \|\mathbf{e}^{(0)}\|. \tag{5.141}$$

Să presupunem că se cere reducerea amplitudinii erorii printr-un factor 10^{-p}, $p > 0$. Din (5.140) se vede că într-o anumită normă amplitudinea erorii este redusă printr-un factor apropiat de $[\rho(P)]^k$.

Numărul de iteraţii cerut este acea valoare a lui k pentru care

$$[\rho(P)]^k \leqslant 10^{-n}. \tag{5.142}$$

Se logaritmează relaţia (5.142) şi se ţine seama de (5.138); rezultă

$$k \lg [\rho(P)] \leqslant - n, \quad k \leqslant \frac{n}{- \lg [\rho(P)]} = \frac{n}{R}, \tag{5.143}$$

de unde se vede că numărul de iteraţii k necesar reducerii erorii iniţiale prin 10^{-p} este invers proporţional cu rata de convergenţă.

După această prezentare generală se va trece la analiza cîtorva din metodele iterative mai frecvent întîlnite în practică, eficiente din punct de vedere al numărului de iterații, necesarului de memorie și diminuării erorilor [108, 87, 67, 66, 42].

5.3.1. *Metoda iterativă Jacobi*

Această metodă se mai întîlnește într-o serie de lucrări și sub denumirea de metoda *iterațiilor simultane.*

Fie sistemul $A\mathbf{x} = \mathbf{b}$, unde $A \in R^{n\times n}$, $\mathbf{x} \in R^n$, și se pune problema găsirii vectorului soluție $\mathbf{x}$. Vectorul soluție $\mathbf{x}$ există și este unic dacă și numai dacă A este nesingulară; acesta poate fi explicitat prin relația $\mathbf{x} = A^{-1}\mathbf{b}$.

Presupunînd că matricea A este nesingulară și că elementele $a_{ii} \neq 0$, $i = 1, 2, \ldots, n$, matricea A poate fi descompusă sub forma

$$A = -T + D - S = D - C, \tag{5.144}$$

unde

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix},$$

$$-T = \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 \\ a_{21} & 0 & \ldots & 0 & 0 \\ a_{31} & a_{32} & 0 \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{n,n-1} & 0 \end{bmatrix}, \tag{5.145}$$

$$D = \begin{bmatrix} a_{11} & 0 & 0 \ldots & 0 \\ 0 & a_{22} & 0 \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 \ldots & a_{nn} \end{bmatrix}, \quad -S = \begin{bmatrix} 0 & a_{12} & a_{13} \ldots a_{1n} \\ 0 & 0 & a_{23} \ldots a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 \ldots a_{n,n-1} \\ 0 & 0 & 0 \ldots 0 \end{bmatrix};$$

A este matrice pătrată, D matrice diagonală ($a_{ii} \neq 0$, $i = 1, 2, \ldots, n$), $-T$ matrice strict inferior triunghiulară ($a_{ij} = 0$, $i \leqslant j$ şi $a_{ij} \neq 0$, $i > j$); $-S$ matrice strict superior triunghiulară ($a_{ij} \neq 0$, $i < j$ şi $a_{ij} = 0$, $i \geqslant j$).

Înlocuind expresia matricei A din (5.144) în $A\mathbf{x} = \mathbf{b}$, rezultă

$$(-T + D - S)\mathbf{x} = \mathbf{b} \text{ sau } D\mathbf{x} = (T + S)\mathbf{x} + \mathbf{b}, \quad (5.146)$$

de unde rezultă o metodă iterativă

$$D\mathbf{x}^{(k+1)} = (T + S)\mathbf{x}^{(k)} + \mathbf{b}, \quad k = 0, 1, 2, \ldots, \quad (5.147)$$

sau dezvoltat

$$a_{ii}x_i^{(k+1)} = b_i - \sum_{j=1}^{n} a_{ij}x_j^{(k)}, \quad i = 1, 2, \ldots, n. \quad (5.148)$$

Relaţia (5.147) mai poate fi scrisă matriceal sub forma

$$\mathbf{x}^{(k+1)} = D^{-1}(T + S)\mathbf{x}^{(k)} + D^{-1}\mathbf{b}, \quad k = 0, 1, 2, \ldots, \quad (5.149)$$

sau dezvoltat

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{n} a_{ij}x_j^{(k)}\right], \quad i = 1, 2, \ldots, n.$$

Matricea $D^{-1}(T + S)$ din (5.149) poartă numele de matrice Jacobi asociată matricei A.

În sensul realizării unei analize privind rata de convergenţă pentru metoda Jacobi se introduce notaţia

$$P = D^{-1}(T + S) = D^{-1}C. \quad (5.150)$$

În acest fel schema iterativă Jacobi din (5.149) se scrie sub forma

$$\mathbf{x}^{(k)} = P\mathbf{x}^{(k-1)} + D^{-1}\mathbf{b}. \quad (5.151)$$

Pentru analiza convergenţei şirului $\{\mathbf{x}^{(k)}\}_{k \in N}$ către soluţia exactă $\mathbf{x}$ a sistemului $A\mathbf{x} = \mathbf{b}$, este necesară analiza vectorului eroare $\mathbf{e}^{(k)}$ care apare în metoda iterativă Jacobi:

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}, \qquad k = 0, 1, 2, \ldots$$

Se consideră relația (5.146) amplificată la stînga cu D^{-1}:

$$\mathbf{x} = D^{-1}(T+S)\mathbf{x} + D^{-1}\mathbf{b} \text{ sau } \mathbf{x} = P\mathbf{x} + D^{-1}\mathbf{b}. \quad (5.152)$$

Dacă se face diferența dintre (5.151) și (5.152), rezultă

$$\mathbf{x}^{(k)} - \mathbf{x} = P\,[\mathbf{x}^{(k-1)} - \mathbf{x}].$$

Se pune în evidență vectorul eroare în cazul metodei iterative Jacobi:

$$\mathbf{e}^{(k)} = P\mathbf{e}^{(k-1)} = P^2\mathbf{e}^{(k-2)} = \ldots = P^{(k)}\mathbf{e}^{(0)}, k = 0, 1, 2, \ldots \quad (5.153)$$

Aplicînd o normă naturală ultimei relații, se obține

$$\|\mathbf{e}^{(k)}\| \leqslant \|P\|^k \|\mathbf{e}^{(0)}\|. \quad (5.154)$$

Se demonstrează [128, 42] că pentru fiecare normă matriceală de ordinul n și fiecare $\varepsilon > 0$ există o normă naturală $\|P\|$ astfel că

$$\rho(P) \leqslant \|P\| \leqslant \rho(P) + \varepsilon. \quad (5.155)$$

Folosind relația (5.155) în (5.154), se obține

$$\|\mathbf{e}^{(k)}\| \leqslant [\rho(P) + \varepsilon]^k \|\mathbf{e}^{(0)}\|. \quad (5.156)$$

Dacă vectorul eroare inițială $\mathbf{e}^{(0)}$ este vectorul propriu al matricei P, asociat valorii proprii maxime pentru matricea P, atunci (5.156) se poate scrie sub forma

$$\|\mathbf{e}^{(k)}\| = [\rho(P)]^k \|\mathbf{e}^{(0)}\|. \quad (5.157)$$

Dacă se cere reducerea amplitudinii erorii prin factorul 10^{-p}, $p > 0$, din (5.156) se vede că, pentru anumite norme, amplitudinea erorii se reduce printr-un factor apropiat de $[\rho(P)]^k$, unde k este numărul de iterații necesar, pentru care are loc relația

$$[\rho(P)]^k \leqslant 10^{-p}, \text{ iar } 0 \leqslant \rho(P) < 1. \quad (5.158)$$

Dacă se logaritmează în baza zece relația (5.158), se obține

$$k \lg [\rho(P)] \leqslant -p \text{ sau } k \geqslant \frac{p}{-\lg [\rho(P)]}.$$

Folosind (5.138), rata de convergență a metodei iterative are expresia

$$R = -\lg \rho(P) = \lg \frac{1}{\rho(P)},$$

de unde rezultă $k \geqslant \frac{p}{R}$, relație dintre numărul de iterații k, ordinul de precizie p și rata de convergență.

Ținînd seama de expresia matricei P, în cazul metodei Jacobi, se poate scrie

$$\|P\|_\infty = \max_i \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{a_{ij}}{a_{ii}} < 1 \quad \text{și} \quad \|P\|_1 = \max_j \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{a_{ij}}{a_{ii}} < 1. \tag{5.160}$$

Aceste două teste asupra matricei P sînt ușor de verificat. Dacă relațiile (5.160) sînt îndeplinite, matricea P este convergentă (deci $\rho(P) \leqslant \|P\| < 1$), obținîndu-se o limită inferioară pentru rata de convergență:

$$R = \lg \frac{1}{\rho(P)} \geqslant \lg \frac{1}{\min (\|P\|_\infty, \|P\|_1)}. \tag{5.161}$$

Din relația (5.159) se vede că numărul de iterații pentru metoda Jacobi este de n^2 operații.

Pentru a reduce eroarea inițială cu 10^{-p} sînt necesare k iterații, unde $k \geqslant \frac{p}{R}$. Atunci avem $\frac{p \times n^2}{R}$ operații în total.

Pentru ca metoda Jacobi să aibă o eficiență apropiată de a metodelor directe, este necesar ca pentru un factor de precizie impus 10^{-p} rata de convergență a metodei să

satisfacă relația $\frac{p}{R} \leqslant \frac{n}{3}$. Ordinul de precizie p a fost ales astfel ca soluția iterativă obținută prin metoda iterativă Jacobi să fie comparabilă ca acuratețe cu soluția obținută prin metoda directă, folosind același număr de cifre în calculul soluției.

Din (5.151) se vede că metoda Jacobi necesită rezervarea a două spații de memorie, unul pentru componentele vectorului $\mathbf{x}^{(k+1)}$ și altul pentru componentele vectorului $\mathbf{x}^{(k)}$. În această metodă pentru calculul componentelor vectorului soluție $\mathbf{x}^{(k+1)}$ se folosesc în exclusivitate componentele vectorului soluție $\mathbf{x}^{(k)}$ de la iterația k.

5.3.2. *Metoda Gauss-Seidel*

Această metodă mai este întîlnită și sub denumirea de metoda iterațiilor succesive. Metoda Gauss-Seidel este o modificare a metodei Jacobi. La calculul componentelor $x_i^{(k+1)}$ se folosesc toate componentele $\mathbf{x}^{(k+1)}$ cu $j < i$; $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, n-1$, componente care sînt cunoscute din calculele precedente efectuate în cadrul iterației de ordinul $k+1$. În concluzie, componentele vectorului $x^{(k+1)}$ sînt utilizate succesiv în ordinea în care acestea au fost obținute, fapt ce se vede și din relația

$$a_{ii}\, x_i^{(k+1)} = b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)}, \qquad (5.162)$$

$$i = 1, 2, \ldots, n, \quad k = 0, 1, 2, \ldots$$

Relația (5.162) se scrie sub formă matriceală astfel:

$$(D - T)\mathbf{x}^{(k+1)} = S\mathbf{x}^{(k)} + \mathbf{b}, \quad k \geqslant 0, \qquad (5.163)$$

unde D, $-T$ și $-S$ sînt: matricea diagonală, matricea strict inferior triunghiulară și matricea strict superior triunghiulară definite în (5.145). Matricea $D - T$ este inferior triunghiulară și, din presupunerea făcută asupra lui D, este nesingulară [există $(D-T)^{-1}$].

Dacă se amplifică (5.163) cu $(D-T)^{-1}$ la stînga, atunci

$$\mathbf{x}^{(k+1)} = (D - T)^{-1} S\mathbf{x}^{(k)} + (D - T)^{-1}\mathbf{b}, \quad k = 0, 1, \ldots, \tag{5.164}$$

unde $P = (D - T)^{-1} S$ se numește *matricea Gauss-Seidel* asociată matricei A din sistemul $A\mathbf{x} = \mathbf{b}$.

Pentru analiza convergenței metodei Gauss-Seidel se scrie relația (5.162) sub forma

$$x_i^{(k)} = \frac{b_i}{a_{ii}} - \sum_{j=1}^{i-1} \frac{a_{ij}\, x_j^{(k+1)}}{a_{ii}} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(k-1)},$$

$$i = 1, 2, \ldots, n, \quad k = 1, 2, \ldots,$$

respectiv

$$x_i = \frac{b_i}{a_{ii}} - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j.$$

Făcînd diferența acestor utime două relații, rezultă

$$x_i^{(k)} - x_i = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} (x_j^{(k)} - x_j) - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} (x_j^{(k-1)} - x_j). \tag{5.165}$$

Dacă se introduc componentele vectorului eroare sub forma

$$e_i^{(k)} = x_i^{(k)} - x_i, \quad k = 0, 1, 2, \ldots, \quad i = 1, 2, \ldots, n, \tag{5.166}$$

relația (5.165) devine

$$\mathbf{e}_i^{(k)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} e_j^{(k)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} e_j^{(k-1)}, \tag{5.167}$$

$$i = 1, 2, \ldots, n, \quad k = 1, 2, \ldots$$

Lemă. *Fie vectorul eroare* $\mathbf{e}^{(k)}$, $k = 1, 2, \ldots$, *definit prin* (5.167) *cu* $\mathbf{e}^{(0)}$ *arbitrar. Definim norma maximă* $\|\cdot\|_\infty$ *și constantele* c_i *prin*

$$\|\mathbf{e}^{(k)}\|_\infty = \max_j |\, e_j^{(k)}|, \quad c_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right|. \tag{5.168}$$

Dacă

$$c = \max_i \; c_i < 1, \tag{5.169}$$

atunci

$$\| \mathbf{e}^{(k)} \|_\infty \leqslant c^{(k)} \| \mathbf{e}^{(0)} \|_\infty \text{ şi } \mathbf{e}^{(k)} \to \mathbf{0} \text{ cînd } k \to \infty.$$

Demonstraţia rezultă din inegalitatea

$$\| \mathbf{e}^{(k)} \|_\infty \leqslant c^k \| \mathbf{e}^{(k-1)} \|_\infty, \; k = 1, 2, \ldots, \tag{5.170}$$

care se va demonstra prin inducţie. Dacă se scrie relaţia (5.167) pentru $i = 1$, se obţine

$$e_1^k = - \sum_{j=2}^{n} \frac{a_{1j}}{a_{11}} \, e_1^{(k-1)}. \tag{5.171}$$

Aplicînd modulul şi utilizînd relaţiile (5.167), rezultă

$$| e_1^{(k)} | \leqslant \sum_{j=2}^{n} \left| \frac{a_{1j}}{a_{11}} \right| \, | e_1^{(k-1)} | \leqslant \| \mathbf{e}^{(k-1)} \|_\infty \sum_{j=2}^{n} \frac{a_{1j}}{a_{11}} = \| \mathbf{e}^{(k-1)} \|_\infty c_1 \leqslant$$

$$\leqslant r_1 \| \mathbf{e}^{(k-1)} \|_\infty. \tag{5.172}$$

Presupunînd că $| e_j^{(k)} | \leqslant c \, \| \mathbf{e}^{(k-1)} \|_\infty$, pentru $j = 1, 2, \ldots, i-1$ şi $c < 1$, se aplică modulul relaţiei (5.167), obţinîndu-se

$$| e_i^{(k)} | \leqslant \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| | e_j^{(k)} | + \sum_{j=i+1}^{n} \left| \frac{a_{ix}}{a_{ii}} \right| | e_j^{(k-1)} | \leqslant$$

$$\leqslant \| \mathbf{e}^{(k-1)} \|_\infty c \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| + \| \mathbf{e}^{(k-1)} \| \sum_{j=i+1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \leqslant \tag{5.173}$$

$$\leqslant \| \mathbf{e}^{(k-1)} \|_\infty \left(\sum_{j=1}^{i-1} c \left| \frac{a_{ij}}{a_{ii}} \right| + \sum_{j=i+1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \right) \leqslant$$

$$\leqslant \| \mathbf{e}^{(k-1)} \|_\infty \sum_{j=1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| = c_i \| \mathbf{e}^{(k-1)} \|_\infty \leqslant c \| \mathbf{e}^{(k-1)} \|_\infty,$$

sau, dacă se consideră extremele inegalităţii, rezultă

$$\| \mathbf{e}^{(k)} \|_\infty \leqslant c \| \mathbf{e}^{(k-1)} \|_\infty \leqslant \ldots \leqslant c^k \| \mathbf{e}^{(0)} \|_\infty. \tag{5.174}$$

Dacă $c < 1$, $\mathbf{e}^{(k)} \to \mathbf{0}$ cînd $k \to \infty$.

După cum se vede pentru metoda Gauss-Seidel testul privind convergența este destul de simplu. Metoda Gauss-Seidel converge mult mai repede decît metoda Jacobi iar necesarul de memorie este mai mic, fiind suficient un singur spațiu pentru vectorul soluție, spațiu care în etapa de iterație k va memora componentele $x_i^{(k)}$ ce au fost calculate și componentele $x_j^{(k-1)}$ pentru $j > i$, $j = i + 1, i + 2, \ldots, n$.

5.3.3. *Metoda relaxărilor succesive*

Această metodă iterativă folosită la rezolvarea sistemelor algebrice liniare este destul de asemănătoare cu metoda Gauss-Seidel.

La început se aplică metoda Gauss-Seidel pentru calculul componentelor vectorului soluție $\mathbf{x}$, astfel:

$$a_{ii}\bar{x}_i^{(k+1)} = -\sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)}, \qquad (5.175)$$

$$k = 0, 1, 2, \ldots, n, \quad i = 1, 2, \ldots, n.$$

În metoda relaxărilor succesive se consideră, în etapa iterativă $k + 1$, următoarele componente pentru vectorul soluție $\mathbf{x}$:

$$x_i^{(k+1)} = x_i^{(k)} + \omega\,(\bar{x}_i^{(k+1)} - x_i^{(k)}), \qquad (5.176)$$

$$k = 0, 1, 2, \ldots, \quad i = 1, 2, \ldots, n.$$

Constata ω se numește factor de relaxare și cînd $\omega > 1$, se realizează o suprarelaxare, iar cînd $\omega < 1$, se realizează o subrelaxare; de asemenea se observă că

$$\omega = \frac{x_i^{(k+1)} - x_i^{(k)}}{\bar{x}_i^{(k+1)} - x_i^{(k)}}, \quad 0 \leqslant \omega \leqslant 1. \qquad (5.177)$$

Dacă $\omega = 1$, se vede că $x_i^{(k+1)} = \bar{x}_i^{(k+1)}$, deci metoda relaxărilor succesive se reduce la metoda Gauss-Seidel. Relația (5.176) se mai scrie și sub forma

$$x_i^{(k+1)} = (1 - \omega)\, x_i^{(k)} + \omega \bar{x}_i^{(k+1)}, \qquad (5.178)$$

unde $x_i^{(k+1)}$ reprezintă componentele vectorului soluție calculate prin metoda relaxărilor succesive, iar $\bar{x}_i^{(k+1)}$ reprezintă componentele vectorului soluție calculate prin metoda Gauss-Seidel. Cu alte cuvinte, fiecare etapă iterativă în metoda relaxărilor succesive constă din două faze:

— aplicarea metodei Gauss-Seidel pentru determinarea vectorului soluție $\bar{\mathbf{x}}^{(k+1)}$,

— îmbunătățirea soluției $\bar{\mathbf{x}}^{(k+1)}$ cu ajutorul relației (5.178).

Dacă se introduce $\bar{x}_i^{(k+1)}$ dat în (5.175) în (5.178), se obține

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)}\right]$$

sau

$$a_{ii}\, x_i^{(k+1)} + \omega \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} = (1 - \omega)\, a_{ii} x_i^{(k)} -$$

$$- \omega \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + \omega\, b_i. \qquad (5.179)$$

Folosind descompunerea matricei $A = -T + D - S$, relația (5.179) se scrie sub forma

$$(D - T\omega)\mathbf{x}^{(k+1)} = [(1 - \omega)D + \omega S]\mathbf{x}^{(k)} + \omega\, \mathbf{b}. \qquad (5.180)$$

Cu presupunerea făcută privind existența matricei inverse $(D - T\omega)^{-1}$, relația (5.180) devine

$$\mathbf{x}^{(k+1)} = (D - \omega T)^{-1}[(1 - \omega)D + \omega S]\, \mathbf{x}^{(k)} + \omega\, (D - \omega T)^{-1}\mathbf{b} \qquad (5.181)$$

sau pentru $P = (D - \omega T)^{-1}\, [(1 - \omega)D + \omega S]$ se obține

$$\mathbf{x}^{(k+1)} = P\mathbf{x}^{(k)} + \omega\, (D - \omega T)^{-1}\mathbf{b}. \qquad (5.182)$$

Se observă că pentru $\omega = 1$ se obține din (5.181) metoda Gauss-Seidel.

Din relația (5.179) se vede că și această metodă necesită doar un singur spațiu de memorie pentru vectorul soluție $\mathbf{x}$, pe timpul efectuării procesului iterativ.

În încheierea acestui capitol sînt date o diagramă logică (fig. 5.7) și programul 5.2 corespunzător, construit sub forma a trei rutine pentru metodele Jacobi, Gauss-Seidel și a relaxării.

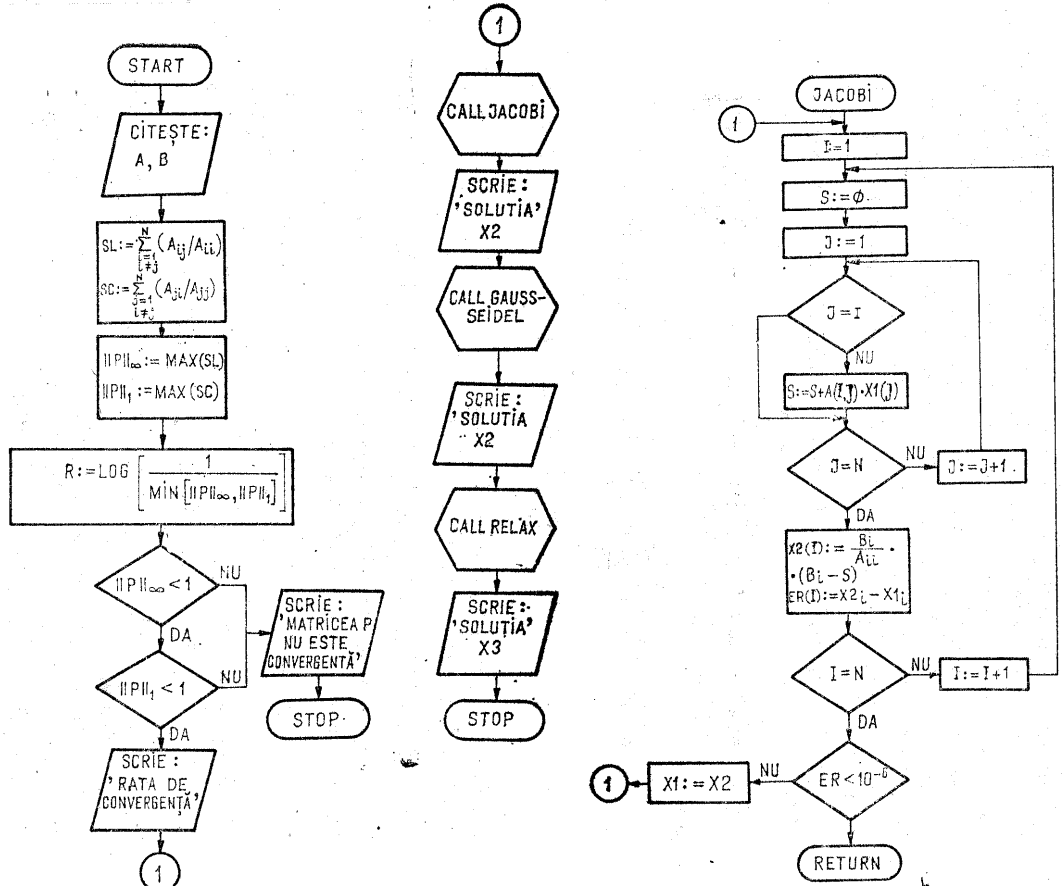
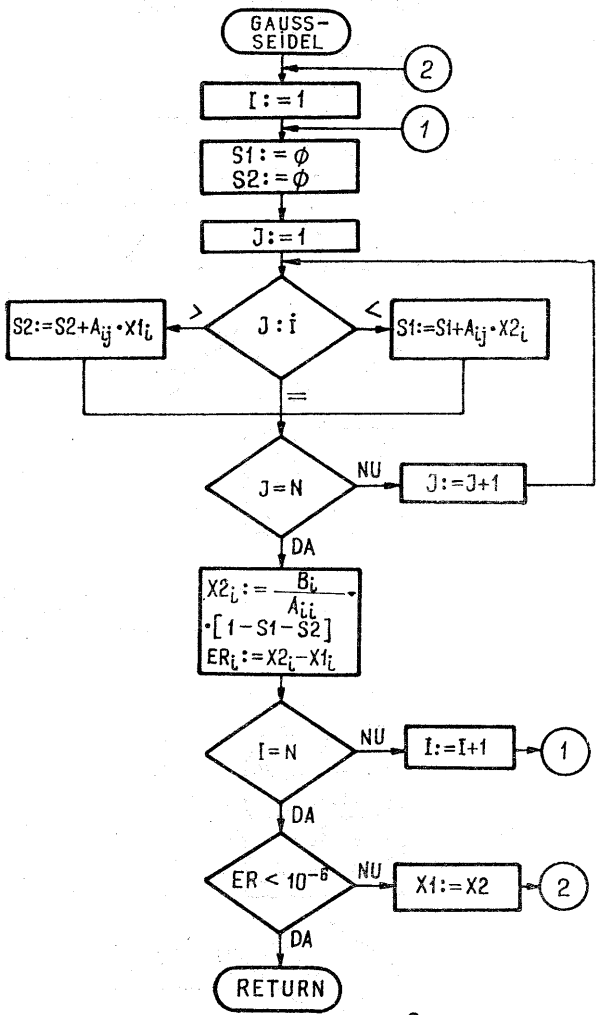
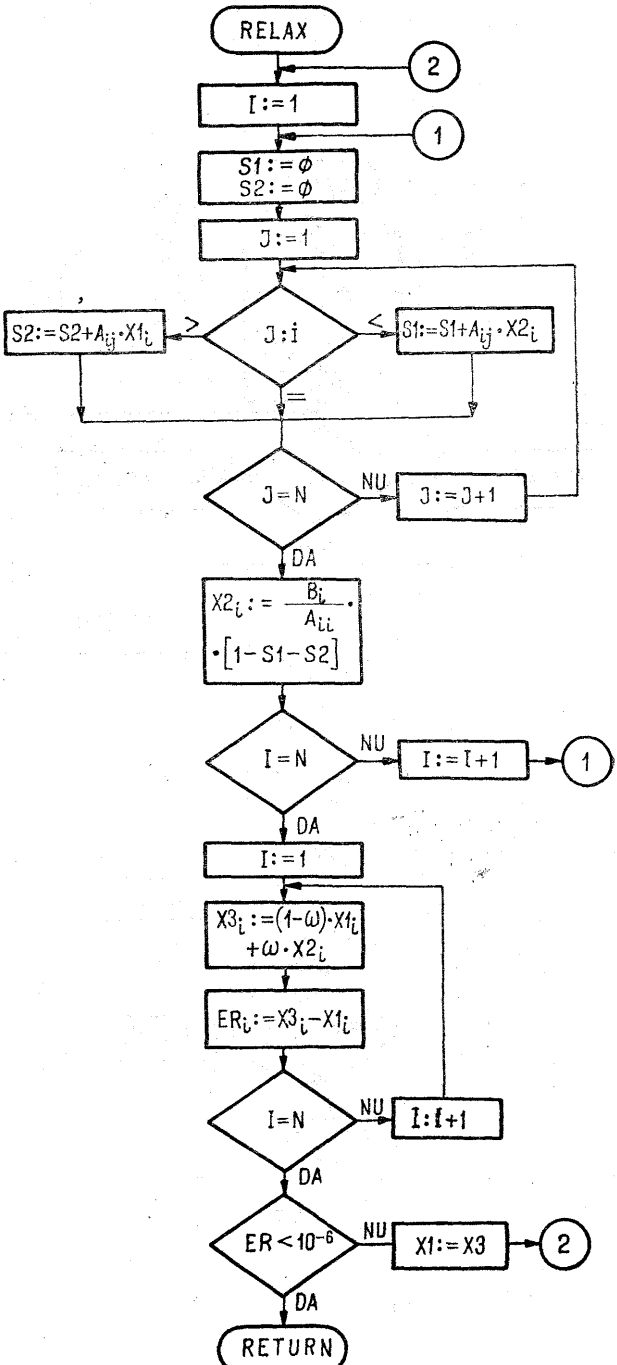

START
CITEȘTE: A, B
SL:=Σ (Aij/Aii)
SC:=Σ (Aji/Ajj)
||P||∞:= MAX(SL)
||P||1 := MAX (SC)
R:=LOG [1 / MIN[||P||∞, ||P||1]]
||P||∞ < 1
NU
DA
SCRIE: 'MATRICEA P NU ESTE CONVERGENTĂ'
STOP
||P||1 < 1
NU
DA
SCRIE: 'RATA DE CONVERGENȚĂ'
1
1
CALL JACOBI
SCRIE: 'SOLUȚIA' X2
CALL GAUSS-SEIDEL
SCRIE: 'SOLUȚIA X2
CALL RELAX
SCRIE: 'SOLUȚIA' X3
STOP
JACOBI
1
I=1
S:=ϕ
J:=1
J=I
NU
S:=S+A(I,J)·X1(J)
J=N
NU
J:=J+1
DA
X2(I):= -Bi/Aii · ·(Bi-S)
ER(I):=X2i-X1i
I=N
NU
I:=I+1
DA
1
X1:=X2
NU
ER<10^-6
RETURN

Fig. 6.7.


GAUSS-
SEIDEL
2
I:=1
1
S1:=ϕ
S2:=ϕ
J:=1
S2:=S2+Aij·X1i
>
J : I
<
S1:=S1+Aij·X2i
=
J=N
NU
J:=J+1
DA
X2i:=Bi/Aii
·[1−S1−S2]
ERi:=X2i−X1i
I=N
NU
I:=I+1
1
DA
ER < 10−6
NU
X1:=X2
2
DA
RETURN

C

Fig. 5.7


RELAX
2
I:=1
1
S1:=ϕ
S2:=ϕ
J:=1
S2:=S2+Aij·X1i
>
J:i
<
S1:=S1+Aij·X2i
=
J=N
NU
J:=J+1
DA
X2i:= Bi/Aii ·
·[1-S1-S2]
I=N
NU
I:=I+1
1
DA
I:=1
X3i:=(1-ω)·X1i
+ω·X2i
ERi:=X3i-X1i
I=N
NU
I:I+1
DA
ER<10-6
NU
X1:=X3
2
DA
RETURN

```
      IF(ABS(ER(I)).LT.EPS) GO TO 48
 44   CONTINUE
      DO 46 I=1,N
 46   X1(I)=X2(I)
      ITER=ITER+1
      GO TO 47
 48   RETURN
      END
      SUBROUTINE RELAX(A,B,X1,X2,X3,ER,N,ITER,OM,EPS)
      DIMENSION A(4,4),B(4),X1(4),X2(4),X3(4),ER(4)
      ITER=0
      DO 50 I=1,N
      X1(I)=0.
      X3(I)=0.
      ER(I)=0.
 50   X2(I)=0.
 60   DO 51 I=1,N
      S1=0.
      S2=0.
      DO 52 J=1,N
      IF(J-I) 53,52,54
 53   S1=S1+A(I,J)*X2(J)
      GO TO 52
 54   S2=S2+A(I,J)*X1(J)
 52   CONTINUE
 51   X2(I)=(B(I)/A(I,I))*(1-S1-S2)
C   RELAXARE
      DO 56 I=1,N
      X3(I)=(1-OM)*X1(I)+OM*X2(I)
      ER(I)=X3(I)-X1(I)
 56   CONTINUE
      DO 59 I=1,N
      IF(ABS(ER(I)).LT.EPS) GO TO 58
 59   CONTINUE
      DO 57 I=1,N
 57   X1(I)=X3(I)
      GO TO 60
 58   RETURN
      END
          LINK
          RUN

 METODA JACOBI        : X=( 1.000, 1.000, 1.000. 1.000) NR. INTERATII=16
 METODA GAUSS-SEIDEL  : X=( 1.000, 1.000, 1.000. 1.000) NR. INTERATII=14
 METODA RELAXARII     : X=( 1.000, 1.000, 1.000. 1.000) NR. INTERATII=13
          EOJ
```

Programul 5.2

VALORI ȘI VECTORI PROPRII. METODE DE CALCUL

6.1. Introducere

Problema valorilor și vectorilor proprii apare într-o mare varietate de aplicații și se scrie sub forma $A\mathbf{x} = \lambda\mathbf{x}$, sau, mai general, $A\mathbf{x} = \lambda B\mathbf{x}$; se pot menționa domenii ca vibrația corpurilor elastice, difuziunea multigrup în sectoarele nucleare, sisteme oscilatorii etc.

În studiul vibrațiilor sau alte domenii este adesea necesară rezolvarea unor sisteme de ecuații liniare având forma

$$\begin{bmatrix} a_{11}-\lambda & a_{12} & a_{13} & \dots a_{1n} \\ a_{21} & a_{22}-\lambda & a_{23} & \dots a_{2n} \\ a_{31} & a_{32} & a_{33}-\lambda & \dots a_{3n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots a_{nn}-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \mathbf{0} \qquad (6.1)$$

sau, altfel scris,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \lambda & 0 & 0 & \dots & 0 \\ 0 & \lambda & 0 & \dots & 0 \\ 0 & 0 & \lambda & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}. \qquad (6.2)$$

În general sistemul (6.2) este exprimat matriceal sub forma

$$A\mathbf{x} = \lambda I\mathbf{x} \text{ sau } A\mathbf{x} = \lambda\mathbf{x}, \qquad (6.3)$$

unde A este o matrice simetrică reală, $\mathbf{x}$ vectorul variabilelor independente iar λ un parametru scalar denumit valoare caracteristică sau valoare proprie. În cazul considerat parametrul λ reprezintă frecvența naturală în vibrațiile sistemului. Apare problema determinării parametrului λ și a vectorului $\mathbf{x}$ corespunzător, vector cunoscut sub denumirea de vector caracteristic sau vector propriu. Pentru a evidenția semnificația fizică a valorilor și vectorilor proprii se consideră următoarele exemple.

Exemple. 1. Studiul vibrațiilor libere ale unui sistem constînd din trei mase (fig. 6.1) conduce la ecuațiile mișcării și la formularea unei probleme de valori proprii :

$$\begin{bmatrix} 2k - 2m\omega^2 & -k & 0 \\ -k & 2k - 4m\omega^2 & -k \\ 0 & -k & 2k - 6m\omega^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{0}, \qquad (6.4)$$

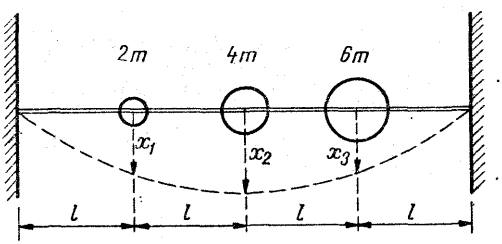


Fig. 6.1

unde ω reprezintă frecvența, $\mathbf{x}$ — vectorul deplasare și $k = s/l$ (s-tensiunea în bară, $4l$ — lungimea barei). Dacă se realizează substituția $\lambda = m\,\omega^2/k$, (6.4) devine

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \lambda \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \qquad (6.5)$$

2. Se consideră sistemul fizic din fig. 6.2, al cărui model matematic este sistemul de ecuații diferențiale ordinare :

$$\left.\begin{aligned} m_1 \frac{\mathrm{d}x_1^2}{\mathrm{d}t^2} &= -k_1 x_1 + k_2(x_2 - x_1) \\ m_2 \frac{\mathrm{d}x_2^2}{\mathrm{d}t^2} &= -k_2(x_2 - x_1) - k_3 x_2 \end{aligned}\right\} \qquad (6.6)$$

unde m_1, m_2 sînt cele două mase legate între ele prin trei resoarte de coeficient de elasticitate k_1, k_2, k_3, x_1, x_2 deplasările pe orizontală față de starea de echilibru, iar t este timpul.

Pentru oscilațiile naturale sistemul va oscila la o frecvență unică ω_n, obținîndu-se oscilații sinusoidale de amplitudine y și un unghi de defazaj θ. În acest caz expresiile celor două deplasări sînt :

$$x_1 = y_1 \sin(\omega_n t - \theta), \quad x_2 = y_2 \sin(\omega_n t - \theta). \tag{6.7}$$

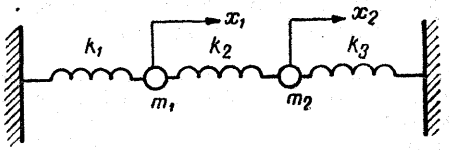


Fig. 6.2

Dacă se înlocuiesc relațiile din (6.7) în (6.6), după derivare și ordonare rezultă sistemul

$$\left.\begin{aligned} -m_1\omega_n^2 y_1 - k_1 y_1 + k_2 y_1 - k_2 y_2 &= 0 \\ -m_2\omega_n^2 y_2 + k_2 y_2 + k_3 y_2 - k_2 y_1 &= 0. \end{aligned}\right\} \tag{6.8}$$

Pentru $m_1 = m_2 = m$, $k_1 = k_2 = k_3 = k$ și frecvența adimensională $\lambda = m\,\omega_n^2/k$, sistemul (6.8) devine

$$\left.\begin{aligned} (2-\lambda)y_1 - y_2 &= 0 \\ -y_1 + (2-\lambda)y_2 &= 0 \end{aligned}\right\} \tag{6.9}$$

Sistemul omogen în necunoscutele y_1, y_2 are soluția banală $y_1 = y_2 = 0$ pentru orice valoare λ, fapt neinteresant din punct de vedere fizic. Astfel se impune condiția ca $\det A \neq 0$ (A fiind matricea sistemului), rezultînd ecuația polinomială în λ

$$P_2(\lambda) \equiv \det \begin{bmatrix} 2-\lambda & -1 \\ -1 & 2-\lambda \end{bmatrix} \equiv \lambda^2 - 4\lambda + 3 = 0\,;\ \lambda_1 = 1,\ \lambda_2 = 3.$$

Gradul polinomului caracteristic $P_2(\lambda)$ reprezintă numărul gradelor de libertate ale sistemului fizic, precum și numărul valorilor proprii. Fiecărei valori proprii îi corespunde cel puțin un vector propriu, în cazul de față există doi vectori proprii :

$$\mathbf{y}^{(1)} = \begin{bmatrix} y_1^1 \\ y_2^1 \end{bmatrix} \to \lambda_1, \quad \mathbf{y}^{(2)} = \begin{bmatrix} y_1^2 \\ y_2^2 \end{bmatrix} \to \lambda_2. \tag{6.10}$$

Din punct de vedere fizic λ_1 și λ_2 reprezintă frecvențe naturale în forma adimensională pentru sistemul considerat :

$$\lambda_1 = \frac{m\omega_1^2}{k} = 1, \quad \omega_1 = \sqrt{k/m},$$

$$\lambda_2 = \frac{m\omega_2^2}{k} = 3, \quad \omega_2 = \sqrt{3k/m}. \tag{6.11}$$

Dacă se rezolvă sistemul (6.9) în y_1^1 și y_2^1 pentru $\lambda_1 = 1$ și în y_1^2, y_2^2 pentru $\lambda = 3$, rezultă vectorii proprii $y^{(1)}$ și $y^{(2)}$:

$$\mathbf{y}^{(1)} = \begin{bmatrix} y_1^1 \\ y_2^1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \rightarrow \lambda_1; \quad \mathbf{y}^{(2)} = \begin{bmatrix} y_2^1 \\ y_2^2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \rightarrow \lambda_2. \tag{6.12}$$

Din (6.11) se vede că ω_1 este frecvența cea mai joasă iar ω_2 este frecvența cea mai înaltă pentru sistemul considerat.

Se observă din (6.12) că la frecvența ω_1 cele două mase se deplasează la fel și în aceeași direcție, iar la frecvența ω_2 cele două mase se deplasează cu aceeași mărime dar în sensuri opuse. Cele două exemple prezentate au menirea să evidențieze faptul că valorile proprii și vectorii proprii descriu modul de comportare al unui sistem fizic considerat, reprezentînd o serie de mărimi ce descriu comportarea sistemului.

În afară de importanță pe care o au valorile proprii în cadrul analizei comportării sistemelor fizice, cunoașterea valorilor și vectorilor proprii ai unei matrice poate fi foarte utilă pentru simplificarea calculelor matriceale, determinarea soluției particulare a unui sistem de ecuații diferențiale, analiza convergenței metodelor iterative utilizate la rezolvarea sistemelor de ecuații algebrice liniare etc. În acest sens se vor considera cîteva cazuri particulare.

Fie $A \in M_R^{n\times n}$ o matrice ce are n vectori proprii liniar independenți $(\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n)$, care corespund valorilor proprii $\lambda_1, \lambda_2, \ldots, \lambda_n$ ale matricei A. Orice vector $\mathbf{z} \in R^n$ poate fi exprimat în mod unic prin baza B, formată cu vectorii proprii $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$ liniar independenți, adică

$$\mathbf{z} = a_1\mathbf{x}^1 + a_2\mathbf{x}^2 + \ldots + a_n\mathbf{x}^n = \sum_{i=1}^{n} a_i\mathbf{x}^i, \tag{6.13}$$

unde coeficienții a_i reprezintă componentele vectorului z cu privire la baza $B = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$.

Datorită faptului că relația (6.8) se poate scrie sub forma

$$A\mathbf{x}^i = \lambda_i \mathbf{x}^i, \quad i = 1, 2, \ldots, n, \tag{6.14}$$

și dacă se amplifică (6.13) cu matricea A de k ori, se obține relația

$$\begin{aligned} A^k\mathbf{z} &= A^k(a_1\mathbf{x}^1 + a_2\mathbf{x}^2 + \ldots + a_n\mathbf{x}^n) = \\ &= A^{k-1}(a_1 A\mathbf{x}^1 + a_2 A\mathbf{x}^2 + \ldots + a_n A\mathbf{x}^n) = \\ &= A^{k-1}(a_1\lambda_1\mathbf{x}^1 + a_2\lambda_2\mathbf{x}^2 + \ldots + a_n\lambda_n\mathbf{x}^n) = \\ &= A^{k-2}(a_1\lambda_1^2\mathbf{x}^1 + a^2\lambda_2^2\mathbf{x}^2 + \ldots + a_n\lambda_n^2\mathbf{x}^n) = \\ &= a_1\lambda_1^k\mathbf{x}^1 + a_2\lambda_2^k\mathbf{x}^2 + \ldots + a_n\lambda_n^k\mathbf{x}^n. \end{aligned} \tag{6.15}$$

În cazul în care între valorile proprii există relația

$$|\lambda_1| > |\lambda_i|, \; i = 2, 3, \ldots, n \text{ și } a_1 \neq 0, \tag{6.16}$$

atunci (6.15) se poate scrie sub forma

$$\begin{aligned} A^n\mathbf{z} = \lambda_1^k \Bigg[a_1\mathbf{x}^1 + \left(\frac{\lambda_2}{\lambda_1}\right)^k a_2\mathbf{x}^2 + \left(\frac{\lambda_3}{\lambda_1}\right)^k a_3\mathbf{x}^3 + \ldots \\ \ldots + \left(\frac{\lambda_n}{\lambda_1}\right)^k a_n\mathbf{x}^n \Bigg] \approx a_1\mathbf{x}^1\lambda_1^k, \end{aligned} \tag{6.17}$$

de unde se vede că vectorul $A^k\mathbf{z}$ se găsește pe aceeași direcție cu vectorul propriu $\mathbf{x}^1$, corespunzător valorii proprii λ_1. Dacă se cunoaște valoarea proprie maximă a matricei A și vectorul propriu asociat, precum și componentele vectorului z cu privire la baza formată cu vectorii proprii ai matricei A $[B = (\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n)]$, atunci se poate determina A^k din relația (6.17). Operația de ridicare la puterea k a unei matrice implică un număr considerabil de operații și spațiu de memorie, care pentru anumite dimensiuni ale lui A este imposibilă pe calculatoare medii.

Exemplul 3. Fie matricea $A \in M_R^{2\times 2}$, avînd elementele și vectorii proprii :

$$A = \begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix} \text{ și } \mathbf{x}^1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mathbf{x}^2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Se poate arăta că relația (6.3) pentru acest caz devine

$$A\mathbf{x}^1 = 4\mathbf{x}^1 \text{ și } A\mathbf{x}^2 = 6\mathbf{x}^2, \quad \lambda_1 = 4 \text{ și } \lambda_2 = 6.$$

Se observă că 4 și 6 reprezintă valorile proprii ale matricei A considerate, asociate vectorilor proprii $\mathbf{x}^1$ și $\mathbf{x}^2$.

Orice vector $\mathbf{z} \in R^2$ poate fi exprimat în mod unic prin baza $B = \{\mathbf{x}^1, \mathbf{x}^2\}$ sub forma

$$\mathbf{z} = a_1\mathbf{x}^1 + a_2\mathbf{x}^2, \text{ unde } \begin{matrix} a_1 = \frac{1}{2}(z_1 + z_2) \\ a_2 = \frac{1}{2}(z_1 - z_2) \end{matrix}, \quad \mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

Pentru acest caz (6.17) devine

$$A^k\mathbf{z} = 4^k a_1\mathbf{x}^1 + 6^k a_2\mathbf{x}^2 = 6^k\left[a_1\left(\frac{4}{6}\right)^k \mathbf{x}^1 + a_2\mathbf{x}^2\right] \approx 6^k a_2\mathbf{x}^2.$$

În concluzie valorile și vectorii proprii au un mare rol în simplificarea calculelor matriceale, reducînd numărul de operații, necesarul de memorie aferentă, timpul de execuție pe calculator etc.

Fie $\mathbf{y}(t) \in R^n$ un vector care în formă transpusă se prezintă astfel : $[\mathbf{y}(t)]^T = [y_1(t), y_2(t), \ldots, y_n(t)]$, avînd n funcții componente dependente de scalarul t. Derivata vectorului $y(t)$ este

$$\left[\frac{dy}{dt}\right]^T = \left[\frac{dy^1}{dt}, \frac{dy_2}{dt}, \ldots, \frac{dy}{dt^n}\right]. \tag{6.18}$$

Atunci sistemul de ecuații diferențiale cu coeficienți constanți se scrie sub forma

$$\frac{d\mathbf{y}(t)}{dt} = A\mathbf{y}(t), \quad \mathbf{y}(0) \neq \mathbf{0}. \tag{6.19}$$

Soluția particulară a acestui sistem poate fi obținută cu ajutorul valorilor și vectorilor proprii ai matricei A.

Folosind relația $A\mathbf{x} = \lambda\mathbf{x}$ și alegînd ca soluție

$$\mathbf{y}(t) = \mathrm{e}^{\lambda t}\,\mathbf{x}, \tag{6.20}$$

se obține după derivare

$$\frac{\mathrm{d}\mathbf{y}(t)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{e}^{\lambda t}\,\mathbf{x}) = \lambda\mathrm{e}^{\lambda t}\,\mathbf{x} = \lambda\mathbf{y}(t) \tag{6.21}$$

și după amplificare a vectorului $\mathbf{y}(t)$ cu A rezultă

$$A\mathbf{y}(t) = A\mathrm{e}^{\lambda t}\,\mathbf{x} = \mathrm{e}^{\lambda t}\,A\mathbf{x} = \mathrm{e}^{\lambda t}\,\lambda\mathbf{x} = \lambda\mathrm{e}^{\lambda t}\,\mathbf{x} = \lambda\mathbf{y}(t). \tag{6.22}$$

Din analiza relațiilor (6.21) și (6.22) se vede că acestea sînt egale, adică se verifică (6.19) pentru soluția (6.20), soluție care poate fi găsită direct, dacă se cunosc valorile și vectorii proprii λ, respectiv $\mathbf{x}$ ai matricei A.

Pentru o problemă generală cu valori inițiale (6.19) se știe că soluția $\mathbf{y}(t) = \exp(\lambda t)\mathbf{y}(0)$ aduce o cantitate de informație redusă deoarece nu se vede dacă soluția tinde la zero, oscilează sau devine nemărginită cînd $t \to \infty$. Datorită acestui fapt se va căuta soluția particulară a sistemului (6.19) de forma (6.20) cu vectorul $\mathbf{x} = \text{const} \neq \mathbf{0}$.

Dacă se introduce soluția din (6.20) în (6.19) și după derivare se simplifică cu $\mathrm{e}^{\lambda t}$, rezultă $\lambda\mathrm{e}^{\lambda t}\,\mathbf{x} = A\mathrm{e}^{\lambda t}\,\mathbf{x}$, deci $\lambda\mathbf{x} = A\mathbf{x}$, de unde

$$(\lambda I - A)\mathbf{x} = 0, \quad \mathbf{x} \neq \mathbf{0}. \tag{6.23}$$

Ultima ecuație, numită și ecuația caracteristică, va fi rezolvată pentru acele valori reale sau complexe ale lui λ, pentru care ea are soluții $\mathbf{x} \neq \mathbf{0}$. Pentru existența unor soluții $\mathbf{x} \neq \mathbf{0}$, condiția necesară și suficientă este ca

$$\det\,(\lambda I - A) = 0. \tag{6.24}$$

Exemplul 4. Fie A din (6.19) de forma

$$A = \begin{bmatrix} -5 & 3 \\ -6 & 1 \end{bmatrix}.$$

Atunci (6.24) devine

$$\begin{bmatrix} \lambda+5 & -3 \\ 6 & \lambda-1 \end{bmatrix} = \lambda^2+4\lambda+13, \qquad \begin{matrix} \lambda_1 = -2+3i, \\ \lambda_2 = -2-3i. \end{matrix} \tag{6.25}$$

Vectorul propriu $\mathbf{x}^1$ corespunzător valorii proprii λ_1 se poate găsi cu ajutorul ecuației

$$(\lambda_1 I - A)\,\mathbf{x}^1 = \begin{bmatrix} 3+3i & -3 \\ 6 & -3+3i \end{bmatrix} \begin{bmatrix} x_1^1 \\ x_2^1 \end{bmatrix} = \mathbf{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$\left.\begin{matrix} (3+3i)x_1^1 - 3x_2^1 = 0 \\ 6x_1^1 - (3-3i)\,x_2^1 = 0 \end{matrix}\right\}; \quad \mathbf{x}^1 = \begin{bmatrix} x_1^1 \\ x_2^1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1+i \end{bmatrix} \neq 0,$$

iar vectorul propriu $\mathbf{x}^2$ corespunzător valorii proprii λ_2 se găsește cu ajutorul ecuației

$$(\lambda_2 I - A)\mathbf{x}^2 = \begin{bmatrix} 3-3i & -3 \\ 6 & -3-3i \end{bmatrix} \begin{bmatrix} x_2^1 \\ x_2^2 \end{bmatrix} = \mathbf{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$\left.\begin{matrix} (3-3i)\,x_2^1 - 3x_2^2 = 0 \\ 6\,x_2^1 - (3+3i)\,x_2^2 = 0 \end{matrix}\right\}; \quad \mathbf{x}^2 = \begin{bmatrix} x_1^2 \\ x_2^2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1-i \end{bmatrix} \neq 0.$$

Folosind relația (6.20), soluțiile sistemului considerat se pot scrie sub forma

$$\mathbf{y}_1(t) = \mathbf{x}^1 e^{\lambda_1 t} = \begin{bmatrix} 1 \\ 1+i \end{bmatrix} e^{-(2+3i)} \tag{6.26}$$

$$\mathbf{y}_2(t) = \mathbf{x}^2 e^{\lambda_2 t} = \begin{bmatrix} 1 \\ 1-i \end{bmatrix} e^{(-2-3i)}.$$

Dacă se consideră drept condiție inițială

$$\mathbf{y}(0) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \tag{6.27}$$

se observă că nici una din soluțiile (6.26) nu satisface această condiție inițială ; astfel se caută un vector $\mathbf{y}(t)$ ca o combinație liniară de forma

$$\mathbf{y}(t) = c_1\mathbf{y}_1(t) + c_2\mathbf{y}_2(t). \tag{6.28}$$

Pentru orice c_1 și c_2 combinația liniară (6.28) reprezintă o soluție a ecuației diferențiale omogene $\mathbf{y}'(t) = A\mathbf{y}(t)$ (scrisă sub formă matriceală). Coeficienții c_1 și c_2 ai combinației liniare se vor determina cu ajutorul condiției inițiale (6.27) :

$$\mathbf{y}(0) = c_1\mathbf{y}_1(0) + c_2\mathbf{y}_2(0)$$

sau

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ 1+\mathrm{i} \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ 1-\mathrm{i} \end{bmatrix};$$

$$\left.\begin{aligned} c_1 + c_2 &= 1 \\ (1+\mathrm{i})c_1 + (1-\mathrm{i})c_2 &= -1 \end{aligned}\right\} \Rightarrow \begin{aligned} c_1 &= \frac{1+2\mathrm{i}}{2}, \\ c_2 &= \frac{1-2\mathrm{i}}{2}. \end{aligned}$$

După determinarea constantelor c_1 și c_2 soluția dată în (6.28) are urm-ă toarea formă :

$$\mathbf{y}(t) = \frac{1+2\mathrm{i}}{2}\begin{bmatrix} 1 \\ 1+\mathrm{i} \end{bmatrix} e^{(-2+3\mathrm{i})t} + \frac{1-2\mathrm{i}}{2}\begin{bmatrix} 1 \\ 1-\mathrm{i} \end{bmatrix} e^{(-2-3\mathrm{i})t} =$$

$$= \begin{bmatrix} \frac{1+2\mathrm{i}}{2} \\ \frac{-1+3\mathrm{i}}{2} \end{bmatrix} e^{(-2+3\mathrm{i})t} + \begin{bmatrix} \frac{1-2\mathrm{i}}{2} \\ \frac{-1-3\mathrm{i}}{2} \end{bmatrix} e^{(-2-3\mathrm{i})t} =$$

$$= \begin{bmatrix} \frac{1+2\mathrm{i}}{2} \\ \frac{-1+3\mathrm{i}}{2} \end{bmatrix} e^{-2t}\, e^{\mathrm{i}3t} + \begin{bmatrix} \frac{1-2\mathrm{i}}{2} \\ \frac{-1-3\mathrm{i}}{2} \end{bmatrix} e^{-2t}\, e^{-\mathrm{i}3t} =$$

$$= e^{-2t}\left\{\begin{bmatrix} \frac{1+2\mathrm{i}}{2} \\ \frac{-1+3\mathrm{i}}{2} \end{bmatrix} (\cos 3t + \mathrm{i}\sin 3t) + \begin{bmatrix} \frac{1-2\mathrm{i}}{2} \\ \frac{-1-3\mathrm{i}}{2} \end{bmatrix} (\cos 3t - \mathrm{i}\sin 3t)\right\} =$$

$$= e^{-2t}\left|\cos 3t \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \mathrm{i}\sin 3t \begin{bmatrix} 2\mathrm{i} \\ 3\mathrm{i} \end{bmatrix}\right| = e^{-2t}\begin{bmatrix} \cos 3t - 2\sin 3t \\ -\cos 3t - 3\sin 3t \end{bmatrix}$$

sau

$$\mathbf{y}(t) = e^{-2t}\begin{bmatrix} \cos 3t & -2\sin 3t \\ -\cos 3t & -3\sin 3t \end{bmatrix} \Rightarrow \begin{cases} y_1(t) = e^{-2t}(\cos 3t - 2\sin 3t), \\ y_2(t) = -e^{-2t}(\cos 3t + 3\sin 3t). \end{cases} \tag{6.28}$$

Soluția prezentată în (6.20) oferă informații suficiente privind comportarea soluției în timp.

În exemplele prezentate s-a urmărit evidențierea semnificației pe care o pot avea valorile proprii și vectorii proprii ai unei matrice A din cadrul modelelor matematice cele mai diverse.

6.2. Valori și vectori proprii. Proprietăți

Fie $\mathscr{A} \in \mathscr{L}(C^n, C^n)$ și matricea A asociată aplicației liniare relativ la o bază ordonată. Se consideră o bază ordonată în C^n și se va lucra cu vectorii coordonate relativ la această bază. Dacă vectorul $\mathbf{x}$ trece în $\mathbf{y}$ prin aplicația $\mathscr{A}$, atunci se poate scrie

$$A\mathbf{x} = \mathbf{y}. \tag{6.29}$$

Relația (6.29) conduce la intrebarea : există sau nu vectori în C care sînt invarianți prin aplicația considerată?

Dacă V este un spațiu vectorial peste R^n și dacă $A \in M_R^{n\times n}$, vectorul $\mathbf{x} \in V$ este un invariant față de aplicația $\mathscr{A}$ prin matricea A dacă și numai dacă

$$\mathbf{x} \overset{A}{\to} \lambda\mathbf{x}, \ \lambda \in R. \tag{6.30}$$

Se poate arăta că λ nu depinde de reprezentarea matriceală dată, deoarece fiecare reprezentare matriceală depinde de alegerea bazei pentru spațiul V iar invarianța (sau neinvarianța) unui vector în urma aplicării matricei A este independentă de alegerea bazei. De asemenea se poate pune întrebarea dacă există un vector $\mathbf{x} \in R^n$, astfel ca $\mathbf{y}$ din (6.29) să aibă expresia $\mathbf{y} = \lambda\mathbf{x}$, de unde (6.29) devine

$$A\mathbf{x} = \lambda\mathbf{x}, \ \mathbf{x} \neq \mathbf{0}. \tag{6.31}$$

Din (6.31) se vede că vectorul $\mathbf{x} \in R^n$ are proprietatea de a rămîne invariant ca direcție, după aplicarea matricei A.

Fie $\mathbf{x} \in R^3$, $\mathbf{x} \neq \mathbf{0}$, şi $A \in M_R^{3\times 3}$ o matrice de forma

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}; \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Aplicînd matricea A vectorului $\mathbf{x}$, se obţine un vector $\mathbf{y}$, adică $A\mathbf{x}=\mathbf{y}$; dacă $\mathbf{y} = \lambda\mathbf{x}$, $\lambda \in R$, atunci $\mathbf{x}$ este un vector din R^3 care este invariant ca direcţie după transformarea lui $\mathbf{x}$ prin matricea A.

Astfel se poate spune că vectorii $\mathbf{x}$ şi $A\mathbf{x}$ au aceeaşi direcţie (se suprapun), dacă $\lambda > 1$ $|A\mathbf{x}| > |\mathbf{x}|$, iar dacă $\lambda < 1$, $|A\mathbf{x}| < |\mathbf{x}|$.

În fig. 6.3 se dă o interpretare geometrică pentru cele prezentate. Axele sistemului sînt x_1, x_2, x_3 şi vectorul $\mathbf{x}$ este invariant ca direcţie în urma aplicării matricei A.

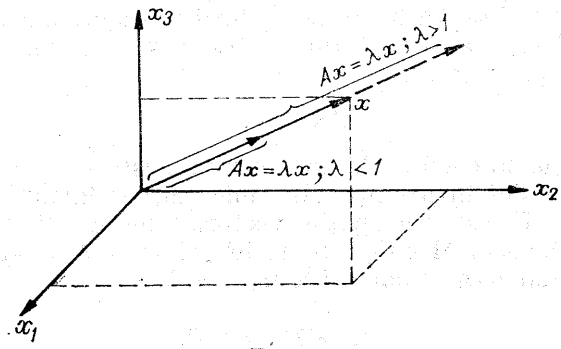


Fig. 6.3

Ecuaţia (6.31) se scrie sub forma $(A - \lambda I)\mathbf{x} = \mathbf{0}$ sau dezvoltat astfel :

$$\begin{bmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (6.32)$$

rezultînd un sistem de ecuaţii omogen care are soluţie nebanală numai dacă

$$\det(A - \lambda I) = 0. \quad (6.33)$$

Se vede că $\det(A - \lambda I)$ este un polinom de gradul n în λ, adică

$$\det(A - \lambda I) = a_n\lambda^n + a_{n-1}\lambda^{n-1} + a_{n-2}\lambda^{n-2} + \ldots + a_1\lambda + a_0. \tag{6.34}$$

Pentru determinarea vectorilor $\mathbf{x} \neq \mathbf{0}$, care să fie invarianți [să satisfacă relația (6.31)] trebuie determinate valorile λ care satisfac ecuația polinomială (6.34), care are n rădăcini $\lambda_1, \lambda_2, \ldots, \lambda_n$. Relația (6.34) se numește polinomul caracteristic al matricei A, iar (6.33) ecuația caracteristică a matricei A.

● Dacă $\det(A - \lambda I) = 0$ este ecuația caracteristică a matricei $A \in M_R^{n\times n}$, atunci cele n rădăcini $\lambda_1, \lambda_2, \ldots, \lambda_n$ ale acestei ecuații se numesc valorile proprii (valori latente, valori caracteristice) ale matricei A.

● Dacă $A \in M_R^{n\times n}$ și dacă $\lambda_1, \lambda_2, \ldots, \lambda_n$ sînt valori proprii ale matricei A, atunci orice soluție nenulă $\mathbf{x}^i$ a ecuației

$$(A - \lambda_i I)\mathbf{x}^i = \mathbf{0}, \text{ sau } A\mathbf{x}^i = \lambda_i\mathbf{x}^i,\ i = 1, 2, \ldots, n, \tag{6.35}$$

este numit vector propriu (vector caracteristic sau latent) corespunzător valorii proprii λ_i. Toți vectorii soluție $\mathbf{x}^i \in R^n$ din (6.35) au proprietatea că

$$\mathbf{x}^{(i)} \overset{A}{\rightarrow} \lambda_i\mathbf{x}^i,\ i = 1, 2, \ldots, n, \tag{6.35}$$

adică prin aplicația $\mathscr{A}$ de matrice A ei rămîn invarianți ca direcție.

Prin definiție vectorul nul $\mathbf{0}$ nu poate fi un vector propriu, dar nimic nu exclude ca o valoare proprie, de exemplu $\lambda_k = 0$; evident vectorii proprii, corespunzător $\mathbf{x}^k$, sînt vectori nenuli din spațiul nul al matricei A. Prin spațiul nul al matricei A se înțelege sistemul de vectori $S(A)$ definit astfel:

$$S(A) = \{\mathbf{v} \in R^n : A\mathbf{v} = \mathbf{0}\}. \tag{6.37}$$

În continuare se vor prezenta o serie de definiții și teoreme privind valorile și vectorii proprii.

● Fie matricea $A \in M_R^{n\times n}$. Atunci spectrul radial $\rho(A)$ al matricei A este dat de $|\lambda_1|$, unde λ_1 este valoarea pro-

prie a lui A, de valoare absolută maximă, adică

$$\rho(A) = |\lambda_1|, \quad |\lambda_1| > |\lambda_i|, \quad i = 2, \ldots, n. \tag{6.38}$$

● Urma unei matrice $A \in M_R^{n\times n}$ este suma elementelor de pe diagonală principală, adică

$$\text{urma } A = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \ldots + a_{nn}. \tag{6.39}$$

● Pentru o matrice $A \in M_R^{n\times n}$ superior triunghiulară sau inferior triunghiulară

$$\begin{bmatrix} a_{11} & 0 & 0 & \ldots & 0 \\ a_{21} & a_{22} & 0 & \ldots & 0 \\ a_{31} & a_{32} & a_{33} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & a_{nn} \end{bmatrix}, \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1n} \\ 0 & a_{22} & a_{23} & \ldots & a_{2n} \\ 0 & 0 & a_{33} & \ldots & a_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & a_{nn} \end{bmatrix} \tag{6.40}$$

valorile proprii sînt

$$\lambda_i = a_{ii}, \quad i = 1, 2, \ldots, n. \tag{6.41}$$

Ţinînd seama de (6.39),

$$\text{urma } A = \sum_{i=1}^{n} \lambda_i = \lambda_1 + \lambda_2 + \ldots + \lambda_n. \tag{6.42}$$

● Folosind cele enunţate, se vede că o matrice A are drept valoare proprie $\lambda = 0$, dacă şi numai dacă ea este singulară.

● Valorile proprii ale unei matrice diagonale $D = (d_{ii})$ sînt date de relaţia

$$\lambda_i = d_{ii}, \quad i = 1, 2, \ldots, n. \tag{6.43}$$

Teoremă. *Dacă $A \in M_R^{n\times n}$ este o matrice nesingulară (există A^{-1}) şi dacă $A\mathbf{x} = \lambda\mathbf{x}$, unde $\lambda \in R$ şi $\mathbf{x} \in R^n$, atunci*

$$A^{-1}\mathbf{x} = \frac{1}{\lambda}\mathbf{x}. \tag{6.44}$$

Demonstrație. Se consideră ecuația $A\mathbf{x} = \lambda\mathbf{x}$, care după amplificare cu A^{-1} la stînga conduce la

$$A^{-1}A\mathbf{x} = A^{-1}\lambda\mathbf{x} \quad \text{sau} \quad \mathbf{x} = \lambda A^{-1}\mathbf{x};$$

după împărțirea cu λ ($\lambda \neq 0$ pentru că există A^{-1}), se obține

$$\frac{1}{\lambda}\mathbf{x} = A^{-1}\mathbf{x} \quad \text{sau} \quad A^{-1}\mathbf{x} = \frac{1}{\lambda}\mathbf{x}.$$

Dacă se scriu două ecuații

$$A\mathbf{x} = \lambda\mathbf{x}, \quad A^{-1}\mathbf{x} = \frac{1}{\lambda}\mathbf{x}, \tag{6.45}$$

se pot enunța următoarele :

— Dacă A este nesingulară (există A^{-1}), atunci A și A^{-1} au același sistem de vectori proprii.

— Valorile proprii ale matricei A^{-1} constituie inversele valorilor proprii ale matricei A.

— Valoarea proprie de modul maxim a lui A conduce la găsirea valorii proprii de modul minim a lui A^{-1}.

Din punct de vedere teoretic se poate considera că un bun algoritm pentru găsirea valorilor proprii ale unei matrice $A \in M_R^{n\times n}$ este acela care găsește zerourile polinomului caracteristic asociat (6.34).

Exemplu [128]. Fie polinomul

$$P_{20}(x) = \prod_{i=1}^{20} (x - i),$$

care are drept zerouri pe $x_i = i$, $i = 1,2,\ldots,20$, și polinomul $T_{20}(x)$ obținut din $P_{20}(x)$ astfel :

$$T_{20}(x) = P_{20}(x) - 2^{-23}\, x^{19}.$$

Acest ultim polinom $T_{20}(x)$ are zece zerouri reale și celelalte zerouri sînt cinci perechi de numere complex conjugate. De la un polinom cu toate rădăcinile reale s-a ajuns la un polinom $T_{20}(x)$ cu cinci perechi de rădăcini complex conjugate. Acestea impun calculul coeficienților $a_0, a_1, \ldots, a_n$ ai polinomului caracteristic cu o precizie foarte mare.

Avînd în vedere faptul că acești coeficienți ai polinomului caracteristic se calculează cu ajutorul elementelor lui A, este posibil prin procesele de rotunjire să se introducă o serie de erori de ordinul 2^{-23}, care ne plasează în cadrul altei probleme, lucru evident dacă analizăm polinoamele $P_{20}(x)$ și $T_{20}(x)$. Din această analiză se vede că determinarea valorilor proprii cu ajutorul polinomului caracteristic (6.34) este apreciată din punct de vedere

teoretic, dar practic trebuie căutate alte metode pentru determinarea valorilor proprii ale matricei A [42, 128, 51, 93].

În acest sens foarte frecvent [94, 59] zerourile polinomului caracteristic se găsesc folosind valorile proprii ale unei matrice asociate matricei A, utilizînd transformările similare. Pentru găsirea valorilor proprii ale matricei A, se face o transformare a matricei A, obținîndu-se în final o matrice B similară cu A, dar de o formă mult mai simplă cu mai multe elemente nule deasupra și sub diagonala principală.

6.3. Reducerea matricelor prin transformări similare

Procesul de reducere a matricelor prin transformări similare este sugerat de necesitatea simplificării metodelor de calcul și obținerea unor rezultate mai precise.

Presupunînd că $\lambda_1, \lambda_2, \ldots, \lambda_n$ sînt valorile proprii ale matricei $A \in M_R^{n\times n}$, pentru fiecare valoare proprie există cel puțin un vector propriu și astfel se pot scrie n ecuații :

$$\left.\begin{aligned} A\mathbf{x}^1 &= \lambda_1 \mathbf{x}^1 \\ A\mathbf{x}^2 &= \lambda_2 \mathbf{x}^2 \\ &\cdots\cdots \\ A\mathbf{x}^n &= \lambda_n \mathbf{x}^n \end{aligned}\right\} \quad \text{sau } A\mathbf{x}^i = \lambda_i \mathbf{x}^i,\ i = 1, 2, \ldots, n. \tag{6.46}$$

Dacă se notează cu X matricea formată din vectorii proprii $X^T = (\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}_n)^T$, atunci cele n ecuații se scriu sub forma matriceală

$$AX = X\Lambda, \tag{6.47}$$

unde A, X și Λ au expresiile

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}, \quad X = \begin{bmatrix} x_1^1 & x_1^2 & \ldots & x_1^n \\ x_2^1 & x_2^2 & \ldots & x_2^n \\ x_3^1 & x_3^2 & \ldots & x_3^n \\ \cdots & \cdots & \cdots & \cdots \\ x_n^1 & x_n^2 & \ldots & x_n^n \end{bmatrix},$$

$$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & 0 \\ & & \ddots & \\ & 0 & & \lambda_n \end{bmatrix},$$

Matricea X ale cărei coloane sînt vectorii proprii ai matricei A se numește *matricea modală* a lui A, iar ecuația (4.47) se numește *ecuația modală*.

Teorema 2. *Dacă matricea X este formată din n coloane liniar independente (adică vectori proprii $x^1, x^2, \ldots, x^n$ sînt liniar independenți), atunci există X^{-1} și ecuația modală se scrie sub forma*

$$X^{-1}AX = \Lambda = \text{diag.}(\lambda_1, \lambda_2, \ldots, \lambda_n), \qquad (6.48)$$

unde matricea A și Λ sînt două matrice similare.

● O matrice $B \in M_{\Gamma}^{n\times n}$ este similară (echivalentă) cu matricea $A \in M_R^{n\times n}$ dacă și numai dacă există o matrice nesingulară $P \in M_R^{n\times n}$ astfel că

$$PAP^{-1} = B; \qquad (6.49)$$

operația se numește transformare similară a amatricei A.

Din relația (6.48) se mai vede că dacă se cunosc vectorii proprii ai matricei A și ei sînt liniar independenți, rezultă imediat valorile proprii ale matricei A dacă se efectuează produsul matriceal $X^{-1}AX$.

Teorema 3. *Dacă A și B sînt similare, atunci matricele A și B au același polinom caracteristic.*

Demonstrație. Dacă A și B similare, atunci $B = PAP^{-1}$, $\det(B - \lambda I) = \det(PAP^{-1} - \lambda I) = \det[P(A - \lambda I)P^{-1}] =$ $= \det(P) \cdot \det(P^{-1}) \det(A - \lambda I) = \det(A - \lambda I)$ pentru că $\det(P) \cdot \det(P^{-1}) = \det(P \cdot P^{-1}) = \det I = 1$.

Teorema 4. *Dacă $B \in M_R^{n\times n}$ este similară cu $A \in M_R^{n\times n}$, atunci A este similară cu B.*

Pentru a demonstra acest lucru se consideră $Q = P^{-1}$. Atunci $Q^{-1} = P$ și relația (6.49) se scrie

$$Q^{-1}AQ = B \quad \text{sau} \quad A = QBQ^{-1}. \qquad (6.50)$$

● Din relația (6.48) se vede că dacă matricea $A \in M_R^{n\times n}$ are un sistem de n vectori proprii liniar independenți, atunci A este similară cu matricea diagonală Λ, care are pe diagonală principală valorile proprii ale matricei A.

● O matrice $A \in M_R^{n\times n}$, dacă are n vectori proprii liniar independenți, se numește *nedefectivă*, iar dacă are $k < n$ vectori proprii liniar independenți, se numește *defectivă*.

Calculul valorilor proprii pentru matricele defective întîmpină o serie de dificultăți teoretice și practice.

• O matrice dată are o infinitate de vectori proprii. Într-adevăr, dacă $A\mathbf{x} = \lambda\mathbf{x}$, atunci $A(k\mathbf{x}) = \lambda(k\mathbf{x})$ pentru $\forall k \in R$, adică dacă $\mathbf{x}$ este vector propriu și $k\mathbf{x}$ este vector propriu, oricare ar fi $k \in R$, $k \neq 0$. Important pentru o matrice este a determina care este numărul vectorilor săi proprii liniar independenți [10, 15].

Corolarul 1. *Dacă* $A \in M_R^{n\times n}$ *are n valori proprii* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *distincte, atunci A are un sistem de n vectori proprii* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ *liniar independenți și, prin urmare, este nedefectivă.*

În acest caz A este similară și cu matricea diagonală Λ, avînd loc relația

$$PAP^{-1} = \Lambda. \tag{6.51}$$

• Matricele care sînt similare cu matricea diagonală Λ se numesc matrice diagonalizabile.

Teorema 5. *Dacă* $A, B, P \in M_R^{n\times n}$ *cu P nesingulară (există* P^{-1}*) se află în relația*

$$PAP^{-1} = B, \tag{6.52}$$

atunci matricele A și B au aceleași valori proprii.

Demonstrație. Fie λ o valoare proprie a matricei A și $\mathbf{x}$ vectorul propriu asociat. Atunci are loc relația

$$A\mathbf{x} = \lambda\mathbf{x}. \tag{6.53}$$

Fie $\mathbf{y} = P\mathbf{x}$ sau $\mathbf{x} = P^{-1}\mathbf{y}$. Dacă se înlocuiește această expresie a lui $\mathbf{x}$ în (6.53), ecuația devine

$$AP^{-1}\mathbf{y} = \lambda P^{-1}\mathbf{y} \text{ sau } PAP^{-1}\mathbf{y} = \lambda\mathbf{y}. \tag{6.54}$$

Dar folosind relația (6.52), ultima relație din (6.54) se scrie

$$B\mathbf{y} = \lambda\mathbf{y}, \tag{6.55}$$

de unde rezultă că λ este și o valoare proprie a matricei B. Într-o manieră asemănătoare se poate arăta că dacă λ este o valoare proprie a lui B iar B și A sînt similare, atunci λ este valoare proprie și pentru matricea A. Dacă se analizează relațiile (6.53) și (6.55), se vede că matricele A și B au aceleași valori proprii λ, A avînd vectorii proprii $\mathbf{x}$ iar B vectorii proprii $\mathbf{y}$ (între $\mathbf{x}$ și $\mathbf{y}$ existînd relațiile $\mathbf{y} = P\mathbf{x}$, $\mathbf{x} = P^{-1}\mathbf{y}$).

În concluzie matricele similare au aceleași valori proprii iar vectorii lor proprii sînt obținuți cu ajutorul transformării liniare de matrice P nesingulară [42, 71, 86].

Reciproca teoremei 5 nu are sens pentru că două matrice A și B pot avea aceleași valori proprii dar nu există

transformări similare cu ajutorul cărora să se poată trece de la o matrice la alta.

Exemplu. Două matrice $A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$, $B = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$ au aceleași valori proprii dar nu există transformări similare prin care se poate transforma matricea A în matrice B.

Din cele prezentate se vede importanța procesului de diagonalizare a matricelor, pentru că dacă se poate găsi o transformare similară care să diagonalizeze matricea considerată, atunci valorile proprii se vor găsi pe diagonala principală a matricei obținute. Procesul de diagonalizare a matricelor nu este ușor de caracterizat, dar matricele de permutare, matrice ortogonale — matrice elementare de rotație, normale, unitare joacă un rol important în procesul de diagonalizare.

- O matrice pătrată A este simetrică dacă $A^T = A$.
- O matrice $A \in M_C^{n\times n}$ este hermitiană dacă $A^H = A$, sau A este hermitiană dacă și numai dacă $\bar{a}_{ij} = a_{ji}$. Dar această condiție implică $\bar{a}_{ii} = a_{ii}$ (deci elementele de pe diagonala principală ale unei matrice hermitiene trebuie să fie reale).
- O matrice simetrică este hermitiană totdeauna, dar o matrice hermitiană este simetrică numai dacă este reală.
- Dacă $A \in M_C^{n\times n}$ este hermitiană, atunci pentru orice $\mathbf{x} \in C^n$, expresia $\mathbf{x}^H A\mathbf{x}$ este reală, fapt ce rezultă din

$$(\mathbf{x}^H A\mathbf{x})^H = \mathbf{x}^H A^H \mathbf{x} = \mathbf{x}^H A\mathbf{x}. \qquad (6.56)$$

- O matrice hermitiană A este pozitiv definită dacă pentru $\mathbf{x} \neq \mathbf{0}$ rezultă $\mathbf{x}^H A\mathbf{x} > 0$, semipozitiv definită dacă pentru $\mathbf{x} \neq \mathbf{0}$ rezultă $\mathbf{x}^H A\mathbf{x} \geqslant 0$.

Teorema 6. *O matrice $A \in M_C^{n\times n}$ este diagonalizabilă dacă și numai dacă există o matrice hermitiană P, pozitiv definită astfel ca $PAP^{-1} = B$, unde B este o matrice normală.*

Din această teoremă se vede că dacă $B = A$ și $P = I$, relația din teoremă este adevărată și dacă $B \in M_C^{n\times n}$ este normală, atunci este diagonalizabilă.

Corolarul 2. *Dacă $A \in M_R^{n\times n}$ este simetrică și $B \in M_C^{n\times n}$ este hermitiană, atunci A și B sînt matrice diagonalizabile.*

În concluzie matricele normale sînt diagonalizabile și matricele simetrice reale și complexe hermitiene sînt exemple de matrice normale.

Teorema 7. *O matrice $A \in M_R^{n\times n}$ sau $A \in M_C^{n\times n}$ este diagonalizabilă dacă și numai dacă este nedefectivă* [86, 94, 108].

Dacă A este defectivă, ea are $k < n$ vectori liniar independenți, fapt care face ca vectorii proprii ai matricei A în acest caz să nu poată constitui o bază pentru C^n. Apar inconveniente pentru destule aplicații care se pot simplifica dacă se folosesc ca bază în C vectorii proprii ai matricei A.

Trebuie menționat faptul că există o legătură directă între ordinul de multiplicitate al unor valori proprii și numărul vectorilor proprii liniar dependeți. Pentru a studia structura unei matrice $A \in M_C^{n\times n}$, ținînd seama de ordinul de multiplicitate al valorilor proprii și respectiv numărul de vectori liniar independenți, este necesară reducerea matricei A prin transformări similare la forma canonică Jordan, care permite evidențierea faptului dacă o matrice este defectivă sau nedefectivă și determinarea ordinului de multiplicitate al valorilor proprii precum și numărul vectorilor liniar independenți pe care îi posedă.

● Orice matrice de forma $A-\lambda I$ cu $A \in M_\Gamma^{n\times n}$ poate fi transformată într-o matrice diagonală D, avînd forma

$$D = \begin{bmatrix} P_1(\lambda) & & 0 \\ & P_2(\lambda) & \\ & & \ddots \\ 0 & & P_n(\lambda) \end{bmatrix}; \quad PAQ = D, \tag{6.57}$$

unde $P_i(\lambda)$ sînt polinoame în λ cu proprietatea că $P_i(\lambda)$ divide pe $P_{i+1}(\lambda)$, $i = 1, 2, \ldots, n$, iar P și Q sînt matrice care au ca elemente polinoame în λ cu coeficienți din C, iar determinanții lui P și Q sînt diferiți de zero și nu depind de λ. Matricea D astfel definită se numește forma canonică Smith pentru matricea $A-\lambda I$.

● Polinoamele $P_i(\lambda)$ din forma canonică Smith cu $k \leqslant n$ sînt denumite factori invarianți ai lui $(A - \lambda I)$:

$$\begin{aligned}
P_1(\lambda) &= (\lambda - \beta_1)^{k_{11}}(\lambda - \beta_2)^{k_{12}} \ldots (\lambda - \beta_k)^{k_{1k}}, \\
P_2(\lambda) &= (\lambda - \beta_1)^{k_{21}}(\lambda - \beta_2)^{k_{22}} \ldots (\lambda - \beta_k)^{k_{2k}}, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
P_i(\lambda) &= (\lambda - \beta_1)^{k_{i1}}(\lambda - \beta_2)^{k_{i2}} \ldots (\lambda - \beta_k)^{k_{ik}}, \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
P_n(\lambda) &= (\lambda - \beta_1)^{k_{n1}}(\lambda - \beta_2)^{k_{n2}} \ldots (\lambda - \beta_k)^{k_{nk}}.
\end{aligned} \tag{6.58}$$

Datorită faptului că $P_i(\lambda)$ divide pe $P_{i+1}(\lambda)$, exponenții factorilor invarianți satisfac inegalitățile $k_{ij} \leqslant k_{i+1,j}$ pentru $j = 1, 2, \ldots, k$.

Exemplu. Fie matricea $A \in M_\Gamma^{4\times 4}$ căreia i se asociază forma canonică Smith D, avînd expresia

$$D = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & \lambda-\beta_1 & \\ 0 & & & (\lambda-\beta_1)\,(\lambda-\beta_2) \end{bmatrix} \begin{matrix} P_1(\lambda) = 1 = (\lambda-\beta_1)^0 = (\lambda-\beta_2)^0, \\ P_2(\lambda) = 1 = (\lambda-\beta_1)^0 = (\lambda-\beta_2)^0, \\ P_3(\lambda) = (\lambda-\beta_1) = (\lambda-\beta_1)^1(\lambda-\beta_2)^0, \\ P_4(\lambda) = (\lambda-\beta_1)(\lambda-\beta_2) = (\lambda-\beta_1)^1(\lambda-\beta_2)^2 \end{matrix} \tag{6.59}$$

Din exemplul considerat se vede că mulți exponenți sînt egali cu zero. Termenii $(\lambda-\beta_i)^{k_{ij}}$, în care $k_{ij} \neq 0$, se numesc divizori elementari ai matricei A. Astfel, divizorii elementari sînt $\lambda-\beta_1$, $\lambda-\beta_1$ și $(\lambda-\beta_2)^2$; dacă $k_{ij} > 1$, atunci A are un divizor elementar neliniar.

Teorema 8. *Matricele $A, B \in M_\Gamma^{n\times n}$ sînt similare dacă și numai dacă $A - \lambda I$ și $B - \lambda I$ au aceeași formă canonică Smith.*

De aici rezultă și următoarele două corolare [42, 10, 12].

Corolarul 3. *Matricele A, $B \in M_\Gamma^{n\times n}$ sînt similare dacă și numai dacă $A - \lambda I$ și $B - \lambda I$ au aceiași factori invarianți $P_i(\lambda)$, $i = 1,2, \ldots, n$.*

Corolarul 4. *Matricele $A, B \in M_\Gamma^{n\times n}$ sînt similare dacă și numai dacă au aceiași divizori elementari $(\lambda - \beta_j)^{k_{ij}}$.*

Teorema 9. *Dacă $A \in M_\Gamma^{n\times n}$ este similară cu $B \in M_\Gamma^{n\times n}$ și B este similară cu $C \in M_\Gamma^{n\times n}$, atunci A este similară cu C.*

Din teorema 8 și din corolarul 3 rezultă că dacă A, B, C fac parte din aceeași clasă de matrice similare, atunci $A - \lambda I$, $B - \lambda I$, $C - \lambda I$ au aceeași formă canonică Smith și aceiași factori invarianți $P_i(\lambda)$, $i = 1, 2, \ldots, n$, precum și aceiași divizori elementari.

De îndată ce divizorii elementari ai matricei A sînt cunoscuți (în general pentru toate matricele similare cu A), se poate constitui forma matriceală canonică Jordan corespunzătoare matricei A. Pentru realizarea acestui lucru se asociază fiecărui divizor elementar $(\lambda - \beta)^k$ un bloc Jordan de forma

$$J_p = \begin{bmatrix} \beta & 1 & & 0 \\ & \beta & 1 & \\ & & & 1 \\ 0 & & & \beta \end{bmatrix}, \tag{6.60}$$

de dimensiune $k \times k$; dacă divizorul elementar este liniar, adică $k = 1$, atunci J_β conține un singur element.

● *Forma canonică Jordan* este o matrice bloc diagonală ale cărei blocuri de pe diagonala principală sînt blocuri elementare Jordan de forma J_β din (6.60).

Pentru exemplul considerat anterior, la trei divizori elementari le corespund trei blocuri elementare Jordan :

$$J_1 = [\beta_1],\ J_2 = [\beta_1];\quad J_3 = \begin{bmatrix} \beta_2 & 1 \\ 0 & \beta_2 \end{bmatrix},$$

iar

$$J = \begin{bmatrix} J_1 & & 0 \\ & J_2 & \\ 0 & & J_3 \end{bmatrix} \text{ sau } J = \begin{bmatrix} \beta_1 & & & 0 & \\ & \beta_1 & & & \\ & & & \beta_2 & 1 \\ & 0 & & 0 & \beta_2 \end{bmatrix}$$

Cele trei blocuri elementare Jordan de pe diagonala lui J au ordinele $1 \times 1, 1 \times 1, 2 \times 2$. Se observă că J este o matrice superior triunghiulară, deci elementele de pe diagonala principală sînt valorile proprii ale matricei J. De asemenea se observă că β_1 și β_2 sînt valori proprii de ordin de multiplicitate citate doi, dar există o diferență între cele două cazuri de multiplicitate în momentul cînd se examinează vectorii proprii.

Se poate enunța o teoremă și două corolare [128, 42, 119, 93].

Teorema 10. *Dacă* $A \in M_\Gamma^{n\times n}$ *și* J *este matricea canonică Jordan asociată matricei* A, *atunci* $J \in M_\Gamma^{n\times n}$ *este similară cu* A [42].

Metoda de construcție a matricei J arată că J este unică, excepție făcînd doar de posibilele permutări ale blocurilor elementare J_β.

Corolarul 4. *Dacă* $(\lambda - \lambda_1)^{k_1}, (\lambda - \lambda_2)^{k_2}, \ldots, (\lambda - \lambda_i)^{k_i}$ *sînt divizorii elementari ai matricei* $A \in M_\Gamma^{n\times n}$, *unde* $\lambda_1, \lambda_2, \ldots, \lambda_i$ *nu trebuie să fie distincte, atunci*

$$n = k_1 + k_2 + \ldots + k_i.$$

Se observă că dacă $k_1 = k_2 = \ldots = k_i = 1$, atunci $i = n$, cu alte cuvinte dacă toți divizorii elementari sînt liniari, forma canonică Jordan este o matrice diagonală.

Corolarul 5. *Dacă* $A \in M_\Gamma^{n\times n}$ *are toți divizorii elementari liniari, atunci matricea* A *este diagonalizabilă, respec-*

tiv nedefectivă. În cazul în care A are cel puțin un divizor elementar neliniar, ea este defectivă și nu poate fi diagonalizabilă [42, 128].

Teorema 11. *Dacă matricea $A \in M_\Gamma^{n\times n}$ are $k \leqslant n$ divizori elementari, adică k blocuri elementare Jordan, atunci A are un sistem de k factori proprii liniar independenți.*

Această teoremă arată că numărul vectorilor proprii liniar independenți corespunde numărului de blocuri elementare Jordan din forma canonică Jordan.

Teorema 12. *Fie λ o valoare proprie a matricei $A \in M_\Gamma^{n\times n}$ care apare în k divizori elementari ai matricei A (adică în k blocuri elementare din forma canonică Jordan). Atunci există k vectori proprii independenți ai matricei A, care corespunde valorii proprii λ.*

Acești k vectori proprii liniar independenți ai matricei A formează o bază în subspațiul k dimensional V^k; acest subspațiu V^k este un exemplu de subspațiu, numit *subspațiu invariant al aplicației liniare* $\mathscr{A} \in \mathscr{L}(c^n, c^n)$ de matrice asociată A.

● O matrice $A \in M_\Gamma^{n\times n}$ se numește *degenerată* dacă și numai dacă aceeași valoare proprie apare în mai multe blocuri elementare din forma canonică Jordan.

● O matrice $A \in M_\Gamma^{n\times n}$ are cel puțin o valoare proprie multiplă dacă și numai dacă este defectivă sau degenerată sau și defectivă și degenerată.

Exemplu de formă canonică Jordan a unei matrice A poate fi următorul:

$$J = \begin{bmatrix} 3 & 1 & & & & & & \\ 0 & 3 & & & & & 0 & \\ & & 4 & 1 & 0 & & & \\ & & 0 & 4 & 1 & & & \\ & & 0 & 0 & 4 & & & \\ & & & & & 5 & 1 & \\ & 0 & & & & 0 & 5 & \end{bmatrix} = \text{diag.}\,(J_{\lambda_1}^{(2)}, J_{\lambda_2}^{(3)}, J_{\lambda_3}^{(2)}).$$

Matricea A a fost transformată într-o formă canonică Jordan având trei blocuri pe diagonala principală:

$J_{11} = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$, cu valoarea proprie $\lambda_1 = 3$, de ordin de multiplicitate $k_1 = 2$;

$J_{22} = \begin{bmatrix} 4 & 1 & 0 \\ 0 & 4 & 1 \\ 0 & 0 & 4 \end{bmatrix}$, cu valoarea proprie $\lambda_2 = 4$, de ordin de multiplicitate $k_2 = 3$;

$J_{33} = \begin{bmatrix} 5 & 1 \\ 0 & 5 \end{bmatrix}$, cu valoarea proprie $\lambda_3 = 5$, de ordin de multiplicitate $k_3 = 2$.

Trebuie specificat că există situații cînd aceeași valoare proprie apare în blocuri diferite.

Pentru forma canonică Jordan se poate introduce următoarea terminologie:

— Dacă X este matricea care reduce matricea A la forma canonică Jordan (6.60), atunci vectorii $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^{k_1}$ satisfac relațiile

$$A\mathbf{x}^1 = \lambda_1 \mathbf{x}^1, A\mathbf{x}^{i+1} = \lambda_i \mathbf{x}^{i+1} + \mathbf{x}^i, i = 1, 2, \ldots, k_1 - 1.$$

Acest lucru este valabil și pentru $k_2, k_3, \ldots, k_s$.

— Polinoamele $P_i(\lambda) = \det (J_{\lambda_i}^{k_i} - \lambda I) = (\lambda - \lambda_i)^{k_i}$, $i = 1, 2, \ldots, s$, sînt numite divizori elementari ai matricei A. Aceste polinoame divid polinomul caracteristic al matricei A, notat prin $P_A(\lambda)$, ce poate fi scris sub forma $P_A(\lambda) = P_1(\lambda) P_2(\lambda) \ldots P_k(\lambda)$.

— O matrice A este defectivă cînd are divizori elementari, iar $P_i(\lambda)$ neliniari.

— În cazul în care anumite valori proprii ale lui A apar în mai multe blocuri J_{ii}, atunci matricea A se numește degenerată.

O matrice A este degenerată, dacă forma sa canonică Jordan este degenerată.

— Forma canonică Jordan a matricei A este o formă diagonală numai cînd $k_1 = k_2 = \ldots = k_s = 1$, iar în acest caz fiecare polinom caracteristic $P_i(\lambda)$, $i = 1, 2, \ldots, s$ (corespunzător blocului diagonal J_{ii}, $i = 1, 2, \ldots, s$) este liniar.

Teorema 13. *Dacă* $A \in M_\Gamma^{n \times n}$ *este o matrice hermitiană, atunci*:

— *valorile proprii* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *ale lui* A *sînt reale*;

— *vectorii proprii* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ *ai lui* A *sînt distincți și ortogonali*;

— *matricea* A *posedă un sistem complet de vectori ortonormali*;

— *există o transformare similară de matrice* $U \in M_{\Gamma}^{n\times n}$ *astfel că* $UAU^{-1} = \Lambda$, *unde* $\Lambda \in M_{\Gamma}^{n\times n}$ *este diagonală și matricea* U *este unitară.*

● O transformare similară cu ajutorul unei matrice unitate se numește transformare similară unitară.

● O transformare similară cu o matrice ortogonală este denumită transformare similară ortogonală.

● Dacă $A \in M_R^{n\times n}$ este simetrică, atunci există o transformare similară $Q\,A\,Q^{-1} = \Lambda$, unde $Q \in M_R^{n\times n}$ este ortogonală și $\Lambda \in M_R^{n\times n}$ este diagonală. Dacă A este simetrică, atunci valorile proprii $\lambda_i \in R$ pentru $i = 1, 2, \ldots, n$.

Teorema 14. *Matricea* $A \in M_{\Gamma}^{n\times n}$ *este pozitiv definită dacă și numai dacă* A *este hermitiană și toate valorile proprii ale lui* A *sînt pozitive* [53, 86, 100].

Teorema 15. *Fie* $A \in M_{\Gamma}^{n\times n}$. *Atunci există o matrice unitară astfel că* $UAU^H = T$, *unde* T *este o matrice superior triunghiulară și* $t_{ii} = \lambda_i$, $i = 1, 2, \ldots, n$ *sînt valorile proprii ale matricei* A, *iar*

$$\det A = \prod_{i=1}^{n} \lambda_i. \tag{6.61}$$

Teorema 16. *Dacă* $A, B \in M_C^{n\times n}$, *atunci matricele* AB *și* BA *au aceleași valori proprii* [108, 124].

Aceste teoreme, corolare, definiții și exemple au avut drept scop să prezinte o serie de aspecte teoretice privind valorile proprii și vectorii proprii în cazul matricelor de diverse tipuri.

În momentul cînd se cer toate valorile și vectorii proprii ai unei matrice sau un număr suficient de mare, trebuie utilizate metodele directe care în mod efectiv reduc matricea A la o formă mult mai simplă prin transformări similare, transformări care nu schimbă valorile proprii.

6.4. Metode de localizare a valorilor proprii

Într-un număr destul de mare de aplicații este suficient dacă se poate realiza o localizare a valorilor proprii într-un anumit domeniu, informație care poate fi destul de prețioasă.

În acest sens Gerșgorin [42, 86] a dat o serie de criterii materializate cu ajutorul unor teoreme.

Teorema 17. *Fie* $A \in M_{\mathrm{T}}^{n\times n}$ *o matrice. Fiecare valoare proprie a matricei* A *se găsește în cel puțin un disc cu centrul în* a_{ii} *și raza* r_{ij}, *unde*

$$r_i = l_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, 2, \ldots, n,$$

sau

$$r_j = c_j = \sum_{\substack{i=1 \\ i\neq j}}^{n} |a_{ij}|, \quad j = 1, 2, \ldots, n.$$

În [86] se arată că toate valorile proprii se găsesc în reuniunea discurilor L_i, $i = 1, 2, \ldots, n$, unde

$$L_i = \{z : |z - a_{ii}| \leqslant l_i = r_i\}, \quad i = 1, 2, \ldots, n,$$

sau în reuniunea discurilor

$$C_j = \{z : |z - a_{jj}| \leqslant c_j = r_j\}, \quad j = 1, 2, \ldots, n.$$

Pentru a demonstra această afirmație, fie λ o valoare proprie a matricei A. Atunci pentru $\mathbf{x} \neq 0$ se poate scrie ecuația valorilor proprii $A\mathbf{x} = \lambda\mathbf{x}$, care implică

$$\sum_{j=1}^{n} a_{ij}x_j = \lambda x_i, \quad i = 1, 2, \ldots, n.$$

Dacă vectorul $\mathbf{x}$ se normalizează astfel ca $\max\limits_i |x_i| = 1$ și dacă $i = k$ este componenta de modul maxim, atunci

$$x_k(\lambda - a_{kk}) = a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{k,k-1}x_{k-1} + a_{k,k+1}x_{k+1} + \ldots + a_{kn}x_n.$$

De aici

$$\lambda - a_{kk} = \left[a_{k1}\left(\frac{x_1}{x_k}\right) + \ldots + a_{k,k-1}\left(\frac{x_{k-1}}{x_k}\right) + \right.$$

$$\left. + a_{k,k+1}\left(\frac{x_{k+1}}{x_k}\right) + \ldots + a_{kn}\left(\frac{x_n}{x_k}\right)\right],$$

de unde rezultă

$$|\lambda - a_{kk}| \leqslant \sum_{\substack{j=1 \\ j\neq k}}^{n} a_{kj} = r_k. \qquad (6.62)$$

Cu alte cuvinte, $|\lambda - a_{kk}| < r_k$, deci valoarea proprie λ se găsește în discul cu centrul în a_{kk} și raza r_k.

Dacă k discuri Gerșgorin formează un domeniu compact care este izolat de celelalte $n-k$ discuri posibile, atunci acest domeniu va conține k valori proprii ale matricei A.

În cazul în care $D_j \cap D_i = \emptyset$, $i = 1,2,3,\ldots,n$ și $i \neq j$, atunci D_j conține o singură valoare proprie a matricei A.

Dacă în linia i a matricei A există un singur element diferit de zero, $a_{ii} \neq 0$, atunci a_{ii} este o valoare proprie a matricei A.

Mulțimea discurilor

$$D_i = \left\{\lambda : |\lambda - a_{ii}| \leqslant \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|\right\}, \quad i = 1, 2, \ldots, n,$$

sînt discuri în planul complex cu centrul în a_{ii} și de rază $r_i = \sum_{i\neq j} |a_{ij}|$, numite discuri Gerșgorin ale matricei A.

Demonstrația teoremei Gerșgorin dată prin relația finală (6.62) arată nu numai că fiecare valoare proprie a lui A trebuie să se găsească într-un disc Gerșgorin, dar și faptul că dacă componenta x_k a unui vector propriu este maximă, atunci valoarea proprie corespunzătoare trebuie să aparțină discului D_k.

Exemplu. Se consideră matricea A de forma

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 6 & 2 \\ 1 & 1 & 11 \end{bmatrix}.$$

Dacă se calculează suma valorilor absolute pentru linii și coloane, rezultă

$$r_1 = l_1 = \sum_{j=2}^{3} |a_{1j}| = 1 + 1 = 2,\ r_1 = c_1 = \sum_{i=2}^{3} |a_{i1}| = 1 + 1 = 2,$$

$$r_2 = l_2 = \sum_{\substack{j=1 \\ j\neq 2}}^{3} |a_{2j}| = 1 + 2 = 3,\ r_2 = c_2 = \sum_{\substack{i=1 \\ i\neq 2}}^{3} |a_{i2}| = 1 + 1 = 2,$$

$$r_3 = l_3 = \sum_{\substack{j=1 \\ j\neq 3}}^{3} |a_{3j}| = 1 + 1 = 2;\ r_3 = c_3 = \sum_{\substack{i=1 \\ i\neq 3}}^{3} |a_{i3}| = 1 + 2 = 3.$$

Valorile proprii ale matricei A se găsesc în reuniunea domeniilor de formă circulară :

$$\{z : |z - 2| \leqslant l_1\} \cup \{z : |z - 6| \leqslant l_2\} \cup \{z : |z - 11| \leqslant l_3\}.$$

Reprezentarea grafică a acestor domenii este dată în fig. 6.4. Se vede că ultimul disc cu centrul în (11, 0) de rază $r_3 = 2$ este disjunct de celelalte două discuri D_1 și D_2, deci el conține o singură valoare proprie a matricei A, iar reuniunea $D_1 \cup D_2$ conține celelalte două valori proprii.

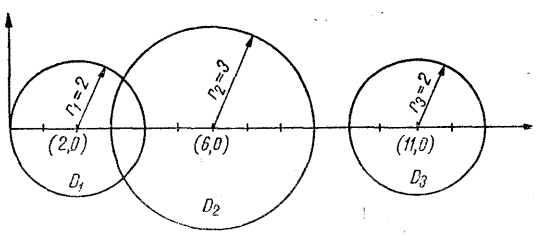


Fig. 6.4

Teorema 18. *Dacă k discuri Gerșgorin ale matricei A sînt disjuncte de celelalte $n-k$, atunci exact k valori proprii se vor găsi în reuniunea celor k discuri* [86].

6.5. Metode de calcul pentru valorile proprii

În general în aplicațiile inginerești există un interes deosebit pentru determinarea valorilor proprii reale și complexe ale unei matrice reale. Analiza metodelor de calcul

ale valorilor proprii pentru matricele reale pot fi extinse la matricele cu elemente complexe. Asemănător rezolvării sistemelor de ecuații, și în cazul determinării valorilor și vectorilor proprii există două tipuri de metode: *metode directe* și *metode indirecte*. Folosirea metodelor directe sau indirecte depinde mai mult de natura soluției cerute decît de forma matricei.

Există aplicații în care se pot cere n valori proprii și n vectori proprii ai unei matrice de ordinul n și alte aplicații în care se cere doar un număr $k < n$ de valori și vectori proprii, precum și aplicații la care se cere valoarea proprie maximă (minimă) în modul. În cazul în care se cere o valoare proprie sau un număr redus de valori proprii $k < n$, se pot aplica metodele indirecte (iterative), iar în cazul în care se cer toate sau aproape toate valorile proprii, este preferabil să se folosească metodele directe care implică transformările similare ce reduc matricea inițială A la o matrice similară B de o formă particulară (reducîndu-se astfel algoritmul de calcul propriu-zis).

Trebuie menționat că are loc și o clasificare a matricelor din punctul de vedere al calculului, pentru că problema calculului valorilor și vectorilor proprii pentru matrice hermitiene este mult mai simplă decît pentru matricele nehermitiene. Această afirmație poate fi justificată în diverse moduri. În acest sens poate fi specificat că orice matrice hermitiană $A \in M_C^{n \times n}$ este diagonalizabilă prin transformări unitare, adică pentru orice A există o matrice unitară P astfel încît

$$P^H A P = \Lambda = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix},$$

unde valorile proprii λ_i, $i = 1, 2, \ldots, n$, sînt reale și matricea A are un sistem complet, ortonormal de vectori proprii (coloanele matricei P). În cazul în care $A = A^H$, matricea A este bine condiționată, cu privire la valorile proprii, în sensul că la variații reduse ale elementelor matricei A apar variații reduse în valorile proprii.

În cazul matricelor nehermitiene trebuie să analizăm dacă matricea A este defectivă sau nu (dacă este defectivă trebuie calculaţi vectorii generalizaţi) şi dacă valorile proprii sînt reale sau complexe (în acest caz trebuie folosită în calcul aritmetica complexă); de asemenea pot apărea o serie de dificultăţi datorită acumulării erorilor de rotunjire, astfel problema poate fi slab condiţionată.

Metodele de calcul al valorilor şi vectorilor proprii, după natura matricei A (hermitiană sau nehermitiană) se prezintă în [128, 42].

6.6. Algoritimi de calcul al valorilor şi vectorilor proprii în cazul matricelor nehermitiene

6.6.1. *Algoritmul puterii directe*

În aplicaţiile care solicită determinarea valorilor proprii maxime şi minime este utilizată metoda puterii.

Presupunem că A are divizori elementari liniari, adică A este nedefectivă şi are n valori proprii distincte care implică un sistem de n vectori proprii liniar independenţi (A este diagonalizabilă). De asemenea se consideră că valorile proprii sînt ordonate:

$$|\lambda_1| > |\lambda_i|, \; i = 1, 2, \ldots, n, \tag{6.63}$$

unde λ_1 este valoarea proprie dominantă.

Fie $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ un sistem de vectori normalizaţi (adică componenta maximă a fiecărui vector este unitatea), liniar independenţi, ei constituind o bază pentru R^n. Atunci orice vector $\mathbf{y} \in R^n$ poate fi scris ca o combinaţie liniară unică (cu $\mathbf{y} \neq \mathbf{0}$), adică

$$\mathbf{y} = c_1\mathbf{x}^1 + c_2\mathbf{x}^2 + \ldots + c_n\mathbf{x}^n .$$

Fie $\mathbf{y} = \mathbf{y}^0$ un vector cu care se construieşte şirul $\mathbf{y}^0, \mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^k, \ldots$ definit astfel:

$$\mathbf{y}^1 = A\mathbf{y}^0 = A(c_1\mathbf{x}^1 + c_2\mathbf{x}^2 + \ldots + c_n\mathbf{x}^n) =$$

$$= c_1\lambda_1\mathbf{x}^1 + c_2\lambda_2\mathbf{x}^2 + \ldots + c_n\lambda_n\mathbf{x}^n ,$$

$$\mathbf{y}^2 = A\mathbf{y}^1 = A^2\mathbf{y}^0 = A^2(c_1\mathbf{x}^1 + c_2\mathbf{x}^2 + \ldots + c_n\mathbf{x}^n) =$$
$$= c_1\lambda_1^2\mathbf{x}^1 + c_2\lambda_2^2\mathbf{x}^2 + \ldots + c_n\lambda_n^2 x^n,$$
. . . . . .
$$\mathbf{y}^k = A\mathbf{y}^{k-1} = A^k\mathbf{y}^0 = A^k(c_1\mathbf{x}^1 + c_2\mathbf{x}^2 + \ldots + c_n\mathbf{x}^n) =$$
$$= c_1\lambda_1^k\mathbf{x}^1 + c_2\lambda_2^k\mathbf{x}^2 + \ldots + c_n\lambda_n^k\mathbf{x}^n,$$
. . . . . .

Dacă se dă factor comun λ_1^k în ultima relație iterativă, rezultă

$$\mathbf{y}^k = A^k\mathbf{y}^0 = \lambda_1^k\left[c_1\mathbf{x}^1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\mathbf{x}^2 + \ldots + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\mathbf{x}^n\right] \tag{6.64}$$

sau

$$\mathbf{y}^k = A^k\mathbf{y}^0 = \lambda_1^k(c_1\mathbf{x}^1 + \mathbf{r}^k).$$

Pentru un k suficient de mare ($k\to\infty$), $\mathbf{r}^k \to \mathbf{0}$ și în acest caz rezultă

$$\mathbf{y}^k = \lambda_1^k c_1\mathbf{x}^1, \text{ respectiv } \mathbf{y}^{k+1} = \lambda_1^{k+1} c_1\mathbf{x}^1,\ c_1 \neq 0. \tag{6.64'}$$

Se știe că operația de împărțire a doi vectori nu are sens, dar se pot împărți componentele vectorilor, astfel că, împărținînd cele două relații la nivel de componente, rezultă

$$\lambda_1 = \frac{y_i^{(k+1)}}{y_i^{(k)}},\ i = 1, 2, \ldots, n, \tag{6.65}$$

relație prin care se poate determina o aproximație pentru valoarea proprie maximă a matricei A. Din (6.65) se vede că vectorul $\boldsymbol{\varepsilon}^k$ are componente destul de mici datorită faptului că $\lambda_i/\lambda_1 < \lambda_{i+1}/\lambda_1$, $\lambda_i/\lambda_1 < 1$ pentru $i \geqslant 2$, adică

$$\mathbf{r}^k = \sum_{i=2}^{n} c_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\mathbf{x}^i \to \mathbf{0},\ k\to\infty. \tag{6.66}$$

Pentru k destul de mare, vectorii $\mathbf{y}^0, \mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^k, \ldots$ sînt aproximații multiple unul a celuilalt, adică ei aproximează vectorii proprii în sensul că

$$\mathbf{y}^{k+1} = A\mathbf{y}^k \approx \lambda_1\mathbf{x}^k,\ k\to\infty, \tag{6.67}$$

de unde se vede că vectorul $\mathbf{y}^{k+1}$ este o aproximație normalizată a vectorului propriu corespunzător valorii proprii

λ_1. Convergența acestui proces iterativ depinde de cît de repede vectorul $\mathbf{r}^k$ tinde la $\mathbf{0}$, adică cît de repede termenii $c_i(\lambda_i/\lambda_1)^k$ din (6.66) tind către zero.

Privind relația (6.63), pot fi considerate două cazuri:

— matricea A are o singură valoare proprie λ_1 de valoare absolută maximă, pozitivă sau negativă:

$$\pm \lambda_1 = |\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|; \tag{6.68}$$

— matricea A are valori proprii dominante sub formă complex conjugată, adică

$$\lambda_1 = a + \mathrm{i}b, \quad \lambda_2 = a - \mathrm{i}b \quad (a, b \in R, \ b \neq 0) \tag{6.69}$$

și

$$|\lambda_1| = |\lambda_2| > |\lambda_3| \geqslant |\lambda_4| \geqslant \ldots \geqslant |\lambda_n|.$$

Mai există și alte cazuri, de exemplu matricea A are două valori proprii $\lambda_1 = +1$ și $\lambda_2 = -1$, sau poate avea șapte valori proprii dominante, fie $\lambda_k = -9 \exp(2k\pi\mathrm{i}/7)$ $k = 1, 2, \ldots, 7)$.

Se observă că în ambele situații translația $\lambda_k \to \lambda_{k+1}$ cauzată prin adunarea cifrei 1 la fiecare element al matricei A conduce la obținerea valorii proprii λ_{k+1} care intră în cazul real (6.68) sau cazul complex (6.69).

În aplicațiile practice pentru matricea A se presupune că ne găsim fie în cazul real (6.68) fie în cazul complex (6.69) dar nu se știe în care anume. Dacă ne găsim în cazul real (6.68), se calculează λ_1, iar dacă ne găsim în cazul complex, se calculează λ_1 și $\lambda_2 = \bar{\lambda}_1$.

Considerînd ca valoare de start vectorul $\mathbf{y} = \mathbf{y}^0$ definit prin

$$y_j^0 = \left(1 + \frac{\pi}{100}\right)^{j-1}, \ j = 1, 2, \ldots, n, \tag{6.70}$$

se poate calcula prima iterație $\mathbf{y}^1 = A\mathbf{y}^0$.

Dacă se folosește metoda celor mai mici pătrate relativ la $\mathbf{y}^1$ și $\mathbf{y}^0$, se determină numărul λ care să minimizeze

$$\|\mathbf{y}^1 - \lambda\mathbf{y}^0\| = (\mathbf{y}^1 - \lambda\mathbf{y}^0, \mathbf{y}^1 - \lambda\mathbf{y}^0) = \sum_{j=1}^{n} (y_j^1 - \lambda y_j^0)^2. \tag{6.71}$$

Derivînd în raport cu λ și explicitînd pe λ, rezultă expresia

$$\lambda = \frac{\sum_{j=1}^{n} y_j^1 y_j^0}{\sum_{k=1}^{n} y_k^0} = \frac{(\mathbf{y}^0, \mathbf{y})}{(\mathbf{y}^0, \mathbf{y}^0)}.$$

Dacă se dă un $\varepsilon > 0$ și dacă $\mathbf{y}^0$, $\mathbf{y}^1$ și λ satisfac relația

$$\|\mathbf{y}^1 - \lambda \mathbf{y}^0\|^2 < \varepsilon^2 \|\mathbf{y}^1\|^2, \tag{6.72}$$

atunci ne găsim în cazul real (6.68) și valoarea proprie $\lambda = \lambda_1$.

Dacă relația (6.72) nu este satisfăcută, se calculează iterația următoare $\mathbf{y}^2 = A\mathbf{y}^1$, folosindu-se numerele a și b care minimizează norma

$$\|\mathbf{y}^2 + a\mathbf{y}^1 + b\mathbf{y}^0\| = \|\mathbf{y}\|^2 + a^2\|\mathbf{y}^1\|^2 + b^2\|\mathbf{y}^0\|^2 + \\ + 2a(\mathbf{y}^2,\mathbf{y}^1) + 2b(\mathbf{y}^2,\mathbf{y}^0) + 2ab(\mathbf{y}^1,\mathbf{y}^0),$$

de unde rezultă condițiile necesare pentru minimizare:

$$\|\mathbf{y}^1\|^2 a + (\mathbf{y}^1, \mathbf{y}^0) b + (\mathbf{y}^2, \mathbf{y}^1) = 0,$$
$$(\mathbf{y}^1, \mathbf{y}^0) a + \|\mathbf{y}^0\|^2 b + (\mathbf{y}^2, \mathbf{y}^0) = 0,$$

obținîndu-se valorile optime pentru a și b:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{-1}{\|\mathbf{y}^0\|^2\|\mathbf{y}^1\|^2 - (\mathbf{y}^1,\mathbf{y}^0)^2} \begin{bmatrix} \|\mathbf{y}^0\|^2 & -(\mathbf{y}^1, \mathbf{y}^0) \\ -(\mathbf{y}^1, \mathbf{y}^0) & \|\mathbf{y}^1\|^2 \end{bmatrix} \begin{bmatrix} (\mathbf{y}^2, \mathbf{y}^1) \\ (\mathbf{y}^2, \mathbf{y}^0) \end{bmatrix}. \tag{6.73}$$

În cazul în care mărimile $\mathbf{y}^0$, $\mathbf{y}^1$, $\mathbf{y}^2$, a, b satisfac inegalitatea

$$\|\mathbf{y}^2 + a\mathbf{y}^1 + b\mathbf{y}^0\| \leqslant \varepsilon^2 \|\mathbf{y}^2\|^2,$$

ne găsim în cazul complex (6.69) cînd rădăcinile proprii de modul maxim sînt λ_1 și $\lambda_2 = \bar{\lambda}_1$.

Dacă testele (6.72) și (6.73) (pentru cazul real, respectiv pentru cazul complex) nu sînt satisfăcute, atunci se consideră procesul iterativ (6.64), care implică o anumită scalare pentru a nu se obține numere prea mari sau prea

mici în cazul reprezentării în virgulă mobilă. Datorită acestui fapt, procedeul de calcul îmbunătățit implică normalizarea vectorilor în fiecare etapă iterativă (un procedeu este de a face cea mai mare componentă a vectorului egală cu unitatea). În acest caz procedeul iterativ arată astfel :

$$\begin{aligned} \mathbf{y}^1 &= A\mathbf{y}^{(0)}, \quad \mathbf{t}^1 = \frac{1}{n_1}\mathbf{y}^1, \\ \mathbf{y}^2 &= A\mathbf{t}^1, \quad \mathbf{t}^2 = \frac{1}{n_2}\mathbf{y}^2, \\ &\cdots\cdots\cdots\cdots \\ \mathbf{y}^k &= A\mathbf{t}^{k-1}, \quad \mathbf{t}^k = \frac{1}{n_k}\mathbf{y}^k, \ldots \end{aligned} \tag{6.74}$$

unde n_k este componenta vectorului $\mathbf{y}^k$ de modul maxim. În acest fel cînd $\mathbf{t}^k \approx \mathbf{x}^1$ se poate presupune că $\mathbf{y}^{k+1} \approx A\mathbf{x}^1 \approx \approx \lambda_1 \mathbf{x}^1$, astfel că elementele vectorului $\mathbf{y}^{k+1}$ de modul maxim satisfac relația $n_{k+1} \approx \lambda_1$. Avantajul algoritmului (6.74) constă în faptul că șirul factorilor de normalizare $n_1, n_2, n_3, \ldots, n_k, \ldots$ converge către valoarea proprie λ_1.

În cazul în care se alege vectorul $\mathbf{y}^0$ astfel ca $c_1 = 0$, pentru $|\lambda_2| > \lambda_i$ $i=3, 4, \ldots, n$, procesele iterative (6.64) și (6.74) converg către valoarea proprie λ_2, de unde rezultă că metoda puterii permite determinarea valorilor proprii intermediare printr-o alegere adecvată a vectorului $\mathbf{y} = \mathbf{y}^0$.

6.6.2. *Algoritmul puterii inverse*

Acest algoritm permite aproximarea oricărei valori proprii, nu neapărat λ_1 sau λ_n. S-a demonstrat că dacă λ este o valoare proprie a matricei nesingulare A și $\mathbf{x}$ vectorul propriu corespunzător, atunci λ^{-1} este valoarea proprie a matricei A^{-1}. corespunzătoare la același vector propriu $\mathbf{x}$.

Pentru simplificare se presupune că $A \in M_R^{n\times n}$ și că A are divizorii elementari liniari, precum și că valorile proprii sînt reale. Algoritmul de calcul presupune alegerea unui vector $\mathbf{y}^0 \neq \mathbf{0}$, $\mathbf{y} \in R^n$ exprimat sub forma unei

combinații liniare de vectorii proprii $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ ai matricei A. Dacă în algoritmul (6.64) se înlocuiește matricea A prin A^{-1} și se folosește $\mathbf{y}=\mathbf{y}^0$, atunci se obține procesul iterativ :

$$\mathbf{y}^1 = A^{-1}\mathbf{y}^0 = A^{-1}(c_1\mathbf{x}^1 + c_2\mathbf{x}^2 + \ldots + c_n x^n) =$$

$$= c_1\mathbf{x}^1\frac{1}{\lambda_1} + c_2\mathbf{x}^2\frac{1}{\lambda_2} + \ldots + c_n\mathbf{x}^n\frac{1}{\lambda_n},$$

$$\mathbf{y}^2 = A^{-1}\mathbf{y}^1 = A^{-1}\left(c_1\mathbf{x}^1\frac{1}{\lambda_1} + c_2\mathbf{x}^2\frac{1}{\lambda_2} + \ldots + c_n\mathbf{x}^n\frac{1}{\lambda_n}\right) =$$

$$= c_1\mathbf{x}^1\frac{1}{\lambda_1^2} + c_2\mathbf{x}^2\frac{1}{\lambda_2^2} + \ldots + c_n\mathbf{x}^n\frac{1}{\lambda_n^2}, \qquad (6.75)$$

. . . . . .

$$\mathbf{y}^k = A^{-1}\mathbf{y}^{k-1} = A^{-1}\left(c_1\mathbf{x}^1\frac{1}{\lambda_1^{k-1}} + \ldots + c_n x^n\frac{1}{\lambda_n^{k-1}}\right) =$$

$$= c_1\mathbf{x}^1\frac{1}{\lambda_1^k} + c_2\mathbf{x}^2\frac{1}{\lambda_2^k} + \ldots + c_n\mathbf{x}^n\frac{1}{\lambda_n^k}.$$

Acest șir de iterații permite determinarea valorii proprii dominante pentru matricea A^{-1} (valoare proprie care reprezintă valoarea proprie minimă pentru matricea A), ținînd seamă de următoarele două ecuații :

$$A\mathbf{x} = \lambda\mathbf{x} \text{ și } A^{-1}\mathbf{x} = \frac{1}{\lambda}\mathbf{x}.$$

Dacă valorile proprii ale matricei A satisfac relația (6.63), unde λ_n este valoarea proprie minimă pentru A și maximă pentru A^{-1}, atunci din (6.75) se obține

$$\mathbf{y}^k = \frac{1}{\lambda_n^k}\left[c_1\mathbf{x}^1\left(\frac{\lambda_n}{\lambda_1}\right)^k + c_2\mathbf{x}^2\left(\frac{\lambda_n}{\lambda_2}\right)^k + \ldots\right.$$

$$\left. \ldots + c_{n-1}\mathbf{x}^{n-1}\left(\frac{\lambda_n}{\lambda_{n-1}}\right)^k + c_n\mathbf{x}^n\right] = \frac{1}{\lambda_n^k}[\mathbf{r}^k + c_n\mathbf{x}^n] \approx \frac{1}{\lambda_n^k}c_n\mathbf{x}^n,$$

respectiv

$$\mathbf{y}^{k+1} = \frac{1}{\lambda_n^{k+1}}[\mathbf{r}^{k+1} + c_n\mathbf{x}^n] \approx \frac{1}{\lambda_n^{k+1}}c_n\mathbf{x}^n,$$

de unde rezultă

$$\lambda_n = \max_i \frac{y_i^k}{y_i^{k+1}}, \quad i = 1, 2, \ldots, n,$$

care este valoarea aproximativă pentru valoarea proprie maximă a matricei A^{-1} și minimă a matricei A. Vectorul $\mathbf{r}^k$ dat prin expresia

$$\mathbf{r}^k = \sum_{i=1}^{n-1} c_i \left(\frac{\lambda_n}{\lambda_i}\right) \mathbf{x}^i$$

tinde la zero după un număr de iterații $k \to \infty$,

6.6.3. *Algoritmul puterii cu deplasarea originii*

În cazul în care se realizează o traslatare a originii cu constanta p în planul complex sau în planul real cu p unități pe axa reală, atunci se poate enunța

Lema 1. *Matricele $A-pI$ și A au același sistem de valori proprii, adică pentru fiecare valoare proprie λ_i a matricei A există valoarea proprie corespunzătoare $\lambda_i - p$ a matricei $A - pI$.*

Demonstrație. Fie $\lambda_1, \lambda_2, \ldots, \lambda_n$ valorile proprii ale matricei A și $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ vectorii proprii corespunzători. În acest caz se poate scrie ecuația $A\mathbf{x}^i = \lambda_i \mathbf{x}^i$, $i = 1, 2, \ldots, n$.

Din ipoteză matricea $A - pI$ are același sistem de vectori proprii ca și matricea A. Atunci se poate scrie

$$(A - pI)\mathbf{x}^i = A\mathbf{x}^i - pI\mathbf{x}^i = \lambda_i \mathbf{x}^i - p\mathbf{x}^i = (\lambda_i - p)\mathbf{x}^i,$$

adică

$$(A - pI)\mathbf{x}^i = (\lambda_i - p)\mathbf{x}^i, \; i = 1, 2, \ldots, n,$$

de unde se vede în mod evident că matricea $(A-pI)$ are valorile proprii $\lambda_i - p$ pentru $i = 1, 2, \ldots, n$.

În cazul în care se introduce în locul matricei A^{-1} din algoritmul puterii inverse (6.75) matricea inversă $(A-pI)^{-1}$, (6.75) devine

$$\begin{aligned}
\mathbf{y}^1 &= (A - pI)^{-1} \mathbf{y}^0, \\
\mathbf{y}^2 &= (A - pI)^{-1} \mathbf{y}^1, \\
&\cdots\cdots\cdots \\
\mathbf{y}^k &= (A - pI)^{-1} \mathbf{y}^{k-1},
\end{aligned}$$

sau sub formă dezvoltată

$$\mathbf{y}^k = (A - pI)^{-1}\mathbf{y}^{k-1} = [(A - pI)^{-1}]^k\,\mathbf{y}^0 =$$
$$= [(A - pI)^{-1}]^k\,(c_1\mathbf{x}^1 + c_2\mathbf{x}^2 + \ldots + c_n\mathbf{x}^n) =$$
$$= c_1\mathbf{x}^1\frac{1}{(\lambda_1 - p)^k} + c_2\frac{1}{(\lambda_2 - p)^k}\mathbf{x}^2 + \ldots + c_n\frac{1}{(\lambda_n - p)^k}\mathbf{x}^n =$$
$$= \frac{1}{\lambda_i - p}\left[c_1\left(\frac{\lambda_i - p}{\lambda_1 - p}\right)^k\mathbf{x}^1 + \ldots\right.$$
$$\ldots + c_{i-1}\left(\frac{\lambda_i - p}{\lambda_{i-1} - p}\right)^k\mathbf{x}^{i-1} + c_i\mathbf{x}^i + c_{i+1}\left(\frac{\lambda_i - p}{\lambda_{i+1} - p}\right)^k x^{i+1} + \ldots$$
$$\left.\ldots + c_n\mathbf{x}^n\left(\frac{\lambda_i - p}{\lambda_n - p}\right)^k\right] = \frac{c_i}{(\lambda_i - p)^k}(\mathbf{x}^i + \mathbf{r}^k).$$

În final se poate scrie

$$\mathbf{y}^k = \frac{c_i}{(\lambda_i - p)^k}(\mathbf{x}^i + \mathbf{r}^k) \approx \frac{c_i}{(\lambda_i - p)^k}\mathbf{x}^i,$$

$$\mathbf{y}^{k+1} = \frac{c_i}{(\lambda_i - p)^{k+1}}(\mathbf{x}^i + \mathbf{r}^k) \approx \frac{c_i}{(\lambda_i - p)^{k+1}}\mathbf{x}^i.$$

Dacă se împart componentele celor doi vectori $\mathbf{y}^{k+1}$ și $\mathbf{y}^k$, se obține

$$\frac{1}{\lambda_i - p} \approx \frac{y_j^{k+1}}{y_j^k},\quad j = 1, 2, \ldots, n. \tag{6.76}$$

Relația (6.76) permite calculul valorii proprii cele mai apropiate de p (unde p este un număr complex din planul complex sau un număr de pe axa reală). De aici rezultă faptul că printr-o alegere judicioasă a lui p, cu ajutorul relației (6.76) se poate determina orice valoare proprie a lui A cu acest algoritm.

Dacă se consideră matricea $A \in M_R^{n\times n}$ care are divizori elementari liniari, atunci toate valorile proprii sînt reale, avînd loc următoarea relație de ordine

$$|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_{n-1}| > \lambda_n.$$

Trebuie acordată atenție alegerii constantei p, valorile dominante ale matricei $A - pI$ vor fi $\lambda_1 - p$ și $\lambda_n - p$.

Dacă se alege $q = \frac{1}{2}(\lambda_2 + \lambda_n)$, se va obține viteza maximă de convergență către valoarea proprie $\lambda_1 - q$ cînd se utilizează matricea $A - pI$ în loc de matricea A (în cadrul algoritmul puterii directe), ca matrice iterativă. De asemenea $q = \frac{1}{2}(\lambda_1 + \lambda_{n-1})$ este o valoare optimă pentru convergența algoritmului către $\lambda_n - p$. În acest sens, dacă se scrie relația iterativă (6.64) pentru $A - pI$ și $\lambda_i - p$, $i = 1, 2, \ldots, n$, rezultă

$$(A - pI)^k \mathbf{y}^0 = (\lambda_1 - p)^k \left[c_1 \mathbf{x}^1 + c_2 \left(\frac{\lambda_2 - p}{\lambda_1 - p} \right)^k \mathbf{x}^2 + \ldots \right.$$

$$\left. \ldots + c_n \left(\frac{\lambda_n - p}{\lambda_1 - p} \right) x^n \right].$$

Convergența [42, 119] este determinată de viteza cu care tinde la zero raportul

$$\left(\frac{\lambda_2 - p}{\lambda_1 - p} \right)^k = \left[\frac{\lambda_2 - \frac{1}{2}(\lambda_2 + \lambda_n)}{\lambda_1 - \frac{1}{2}(\lambda_2 + \lambda_n)} \right]^k = \left(\frac{\lambda_2 - \lambda_n}{2\lambda_1 - \lambda_2 - \lambda_n} \right)^k.$$

În cazul determinării valorii proprii $\lambda_n - p$ convergența este determinată de viteza cu care raportul următor tinde la zero :

$$\left(\frac{\lambda_{n-1} - p}{\lambda_n - p} \right)^k = \left[\frac{\lambda_{n-1} - \frac{1}{2}(\lambda_1 + \lambda_{n-1})}{\lambda_n - \frac{1}{2}(\lambda_1 + \lambda_{n-1})} \right]^k = \left(\frac{\lambda_{n-1} - \lambda_1}{2\lambda_n - \lambda_1 - \lambda_{n-1}} \right)^k.$$

Se observă că prin deplasarea originii cu o constantă p se poate determina valoarea proprie λ_n la fel ca λ_1, iar algoritmul puterii inverse are o serie de avantaje față de metoda puterii directe, datorită vitezei de convergență și a preciziei obținute.

6.6.4. *Algoritmul L—R (left-right→stînga-dreapta)*

Fie matricea $A \in M_R^{n\times n}$, pentru care se construieşte şirul de iteraţii $A = A_1, A_2, A_3, \ldots$, printr-o descompunere triunghiulară în pasul iniţial

$$A_1 = L_1 R_1, \tag{6.77}$$

unde L_1 este o matrice inferior triunghiulară cu elementele $l_{ii} = 1$, $i = 1,2,\ldots,n$, iar R_1 o matrice superior triunghiulară. Dacă matricele L_1 şi R_1 se înmulţesc în ordine inversă, rezultă matricea $A_2 = R_1 L_1$, procesul repetîndu-se cu ajutorul ecuaţiilor

$$A_2 = L_2 R_2,\ R_2 L_2 = A_3, \ldots, A_p = L_p R_p,\ R_p L_p = A_{p+1}, \ldots \tag{6.78}$$

Acest proces iterativ conduce la o transformare similară, deoarece din (6.77) rezultă

$$L_1 = A_1 R_1^{-1}, A_2 = R_1 L_1 = R_1 A_1 R_1^{-1}, \ldots$$
$$\ldots, A_{p+1} = (R_p R_{p-1} \ldots R_1) A_1 (R_p R_{p-1} \ldots R_1)^{-1}$$

şi toate matricele A_p au aceleaşi valori proprii (sînt similare). De asemenea se vede din (6.77) că $R_1 = L^{-1} A_1$; astfel rezultă

$$A_2 = R_1 L_1 = L^{-1} A_1 L_1, \ldots$$
$$\ldots, A_{p+1} = (L_1 L_2 \ldots L_p)^{-1} A_1 (L_1 L_2 \ldots L_p). \tag{6.79}$$

Se observă că $B_p = L_1 L_2 \ldots L_p$ şi $C_p = R_p R_{p-1} \ldots R_1$ sînt matrice cu elemente unitate inferior triunghiulare, respectiv superior triunghiulare, pentru orice p. Datorită faptului că $L_p R_p = R_{p-1} L_{p-1}$, rezultă $B_p C_p = A^p$ şi matricea triunghiulară finală reprezintă descompunerea puterii lui A.

În cazul matricelor de tip bandă care apar în cazul soluţionării ecuaţiilor diferenţiale ordinare, se foloseşte în mod frecvent algoritmul L—R, pentru că se reduce timpul de execuţie şi spaţiul de memorie necesar. Un dezavantaj al metodei L—R este imprecizia care apare la descompunerea matricei generale în matrice triunghiulare. Se impune execuţia descompunerii cu permutarea liniilor matricei A în cazul matricelor triunghiulare astfel ca nici un element al lui L să nu depăşească unitatea, folosin-

du-se efectul alegerii celui mai mare element ca pivot. În acest caz transformările similare corespunzătoare metodei, L—R sînt date prin ecuațiile

$$A_1 = I_1^{-1}L_1R_1,\ A_2 = R_1I_1^{-1}L_1 = I_2'L_2R_2,\ A_3 = R_2I_2'L_2,\ \ldots$$

unde I_r' este matrice pentru permutarea liniilor. Se poate arată că

$$A_2 = R_1A_1R_1^{-1},\ A_3 = (R_2R_1)A_1(R_2R_1)^{-1},\ldots$$

Schimbarea liniilor cu ajutorul matricelor de permutare I_r' permite obținerea unei precizii îmbunătățite.

În final șirul iterativ $A_1, A_2, \ldots, A_p, \ldots$ converge către o matrice superior triunghiulară care are valorile proprii pe diagonala principală. În [93, 108] se prezintă o serie de dificultăți privind implementarea algoritmului L—R, fapt pentru care s-a propus înlocuirea matricei L printr-o matrice unitară Q, obținîndu-se algoritmul Q—R care se va descrie în continuare.

6.6.5. *Algoritmul Q—R*

Algoritmul Q—R are ca obiect descompunerea matricei $A = A_1$ în produsul Q_1R_1, unde R_1 este superior triunghiulară și Q_1 este ortogonală. Șirul de transformări succesive se prezintă într-o manieră similară algoritmului L—R și este definit astfel :

$$A_1 = Q_1R_1,\ A_2 = R_1Q_1 = Q_2R_2,\ A_3 = R_2Q_2 = Q_3R_3, \ldots,$$

obținîndu-se transformările similare

$$A_2 = R_1A_1R_1^{-1},\ A_3 = R_2A_2R_2^{-1},\ A = (R_2R_1)A_1(R_2R_1)^{-1}.$$

În fiecare etapă matricea ortogonală Q_k este construită din produsul de matrice ortogonale simple de tipul celor folosite în metoda Jacobi-Givens. Cu ajutorul acestor matrice simple se elimină un singur element și se prelucrează coloană cu coloană pînă se elimină toate elementele de sub diagonală.

Exemplu. Se înmulțește matricea T_1' cu matricea A :

$$T_1'A = \begin{bmatrix} c & b & 0 \\ -b & c & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11}' & a_{12}' & a_{13}' \\ 0 & a_{22}' & a_{23}' \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = A',$$

unde $c = \cos \alpha$, $b = \sin \alpha$. Se observă că $a'_{12} = -ba_{11} + a_{21}c$; pentru $a'_{21}=0$ rezultă $-ba_{11} + ca_{21} = 0$, astfel că

$$b = \frac{a_{21}}{\sqrt{a_{11}^2 + a_{21}^2}}, \quad c = \frac{a_{11}}{\sqrt{a_{11}^2 + a_{21}^2}}.$$

Dacă se continuă operația de înmulțire a matricei T'_2 și A', rezultă

$$T'_2A' = \begin{bmatrix} c & 0 & b \\ 0 & 1 & 0 \\ -b & 0 & c \end{bmatrix} \begin{bmatrix} a'_{11} & a'_{12} & a'_{13} \\ 0 & a'_{22} & a'_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a''_{11} & a''_{12} & a''_{13} \\ 0 & a'_{22} & a'_{23} \\ 0 & a'_{32} & a'_{33} \end{bmatrix} = A''.$$

Pentru acest caz elementele c și b din T'_2 sînt

$$b = \frac{a_{31}}{\sqrt{a_{11}'^2 + a_{31}'^2}}, \quad c = \frac{a_{11}}{\sqrt{a_{11}'^2 + a_{31}'^2}}.$$

Dacă matricea A'' se înmulțește cu T'_3 care are elementul 1 pe poziția (1,1), se obține

$$T'_3A'' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c & b \\ 0 & -b & c \end{bmatrix} \begin{bmatrix} a''_{11} & a''_{12} & a''_{13} \\ 0 & a'_{22} & a'_{23} \\ 0 & a''_{32} & a''_{33} \end{bmatrix} = \begin{bmatrix} a''_{11} & a''_{12} & a''_{13} \\ 0 & a''_{22} & a''_{23} \\ 0 & 0 & a''_{33} \end{bmatrix}$$

pentru $b = \dfrac{a''_{32}}{\sqrt{a_{22}'^2 + a_{23}'^2}}$ și $c = \dfrac{a''_{22}}{\sqrt{a_{22}'^2 + a_{23}'^2}}$.

Algoritmul Q—R poate fi descris prin iterații astfel:

$$A_i = Q_iR_i,$$

$$A_{i+1} = Q_i^{-1}A_iQ_i = R_iQ_i, \quad i = 1, 2, 3, \ldots$$

Condițiile în care matricea A permite o descompunere unică sînt date de teorema următoare.

Teoremă. *Dacă A este nesingulară, atunci există descompunerea $A = QR$ pentru care Q este unitară și R superior triunghiulară.*

În cazul în care elementele diagonale $r_{ii} \in R^+$, descompunerea este unică. Demonstrația se găsește în [128, 59], unde se arată că metoda de calcul folosită în demonstrația teoremei nu este o procedură de calcul eficientă în cazul

aplicațiilor practice. Implementarea algoritmului Q—R nu este atît de simplă în cazul general [93, 128], algoritmul este practic numai cînd se aplică la matrice de tip Hessenberg (aproape de forma triunghiulară):

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & a_{24} & \dots & a_{2n} \\ & a_{32} & a_{33} & a_{34} & \dots & a_{3n} \\ & & a_{43} & a_{44} & \dots & a_{4n} \\ & \mathbf{0} & & \dots & & \\ & & & & a_{n,n-1} & a_{nn} \end{bmatrix}.$$

Numărul de operații implicate de o etapă în algoritmul Q—R este aproximativ n^3 pentru o matrice completă, în timp ce pentru o matrice Hessenberg este de n^2 operații. În acest sens se recomandă întîi utilizarea unei proceduri prin care matricea este adusă la forma Hessenberg și după aceea să se aplice algoritmul Q—R matricei în formă Hessenberg. O altă recomandare [93] pentru implementarea algoritmului Q—R este de a se utiliza deplasarea originii.

6.6.6. *Reducerea unei matrice la forma Hessenberg*

Din cele prezentate se vede că metodele L—R și Q—R se aplică în condiții mult mai bune dacă matricea A este adusă la forma Hessenberg. Metodele Givens și Householder [42, 128] pot fi folosite pentru a transforma o matrice generală nesimetrică într-o matrice Hessenberg, care poate fi tridiagonală în cazul în care matricea A este simetrică. În general aceste metode implică un număr mare de operații, fapt pentru care se utilizează transformări similare elementare, utilizîndu-se matrice de forma M_r și matrice de permutare I_{rk}. Matricele de tip M_r sînt utilizate la fel ca la metoda lui Gauss de eliminare. În cazul în care un element al matricei A care candidează ca pivot este zero, trebuie schimbată linia r cu linia k pentru obținerea

unui pivot diferit de zero, operație ce se poate efectua cu ajutorul matricei de forma

$$I_{rk} = \begin{array}{c} \\ \\ \\ \\ \\ \\ r \\ \\ \\ \\ \\ k \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccccc} r & & & & & & k & & & & \end{array} \\ \begin{bmatrix} 1 & & & & & & & & & & \\ & 1 & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & \ddots & & & & & & & \\ & & & & \ddots & & & & & & \\ & & & & & 1 & & & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & 0 & \cdots & \cdots & 1 & \cdots & \cdots \\ & & & & & & 1 & & & & \\ & & & & & & & \ddots & & & \\ & & & & & & & & \ddots & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & 1 & \cdots & \cdots & 0 & \cdots & \cdots \\ & & & & & & & & & 1 & \\ & & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{bmatrix} \end{array} ; I_{rk} A I_{rk} = A_1'.$$

În urma produsului matriceal a rezultat matricea A_1', care reprezintă matricea A unde s-a schimbat linia r cu linia k pentru $r < k$.

În scopul triangularizării matricei A se folosesc matricele de forma

$$M_r = \begin{bmatrix} I_{r-1} & \cdots & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & m_{r+1,r} & 1 & \cdots & 0 \\ 0 & \cdots & m_{r+2,r} & 0 & \cdots & 1 \\ & \cdots & \vdots & & & \\ 0 & \cdots & m_{n,r} & 0 & \cdots & 1 \end{bmatrix} \quad r = 1, \ldots, n-1.$$

Matricele $I_{r,k}$ și M_r sînt utilizate împreună pentru a transforma matricea A într-o formă similară cu ea.

Exemplu. Pentru a ilustra modul în care o matrice $A \in M_R^{n \times n}$ se poate reduce la o matrice Hessenberg se consideră $n=4$. Fie matricea $A \in M_R^{4 \times 4}$

de forma

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}.$$

În cazul în care se impune o schimbare de linii și coloane pentru fiecare etapă, presupunînd că $|a_{41}| > |a_{i1}|$ pentru $i = 2$ și 3, este necesară schimbarea liniei 2 cu 4 și a coloanelor 2 cu 4 prin transformări similare :

$$A_1' = I_{24} A I_{24} =$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{14} & a_{13} & a_{12} \\ a_{41} & a_{44} & a_{43} & a_{42} \\ a_{31} & a_{34} & a_{33} & a_{32} \\ a_{21} & a_{24} & a_{23} & a_{22} \end{bmatrix} =$$

$$= \begin{bmatrix} a_{11}' & a_{12}' & a_{13}' & a_{14}' \\ a_{21}' & a_{22}' & a_{23}' & a_{24}' \\ a_{31}' & a_{32}' & a_{33}' & a_{34}' \\ a_{41}' & a_{42}' & a_{43}' & a_{44}' \end{bmatrix} = A_1'.$$

Pentru realizarea formei Hessenberg în cazul matricei A' se vor folosi transformări similare de tipul M pentru anularea elementelor a_{31}' și a_{41}' :

$$B = M_2 A_1' M_2^{-1} =$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & m_{32} & 1 & 0 \\ 0 & m_{42} & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11}' & a_{12}' & a_{13}' & a_{14}' \\ a_{21}' & a_{22}' & a_{23}' & a_{24}' \\ a_{31}' & a_{32}' & a_{33}' & a_{34}' \\ a_{41}' & a_{42}' & a_{43}' & a_{44}' \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -m_{32} & 1 & 0 \\ 0 & -m_{42} & 0 & 1 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ b_{41} & b_{42} & b_{43} & b_{44} \end{bmatrix} = B,$$

unde $m_{32} = -\dfrac{a_{31}'}{a_{21}'}$ și $m_{42} = -\dfrac{a_{41}'}{a_{21}'}$. Valorile pentru m_{32} și m_{42} rezultă din presupunerea făcută ca a_{31}' și a_{41}' să fie nule, de unde se vede că matricea M_2 este cunoscută.

Presupunînd că $|b_{42}| > |b_{32}|$, apare necesară schimbarea ultimelor două linii între ele și a ultimelor două coloane ale matricei B cu ajutorul

matricei I_{34}, astfel rezultînd matricea C de forma

$$C = I_{34}\, B\, I_{34} =$$

$$\begin{bmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&0&1\\ 0&0&1&0 \end{bmatrix}\begin{bmatrix} b_{11}&b_{12}&b_{13}&b_{14}\\ b_{21}&b_{22}&b_{23}&b_{24}\\ 0&b_{32}&b_{33}&b_{34}\\ 0&b_{42}&b_{43}&b_{44} \end{bmatrix}\begin{bmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&0&1\\ 0&0&1&0 \end{bmatrix} = \begin{bmatrix} c_{11}&c_{12}&c_{13}&c_{14}\\ c_{21}&c_{22}&c_{23}&c_{24}\\ 0&c_{32}&c_{33}&c_{34}\\ 0&c_{42}&c_{43}&c_{44} \end{bmatrix} = C.$$

În continuare se folosește matricea M_3 pentru transformarea matricei C; astfel rezultă matricea

$$D = M_3 C\, M_3^{-1} =$$

$$\begin{bmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&m_{43}&1 \end{bmatrix}\begin{bmatrix} c'_{11}&c'_{12}&c'_{13}&c'_{14}\\ c_{21}&c_{22}&c_{23}&c_{24}\\ 0&c_{32}&c_{33}&c_{34}\\ 0&c_{42}&c_{43}&c_{44} \end{bmatrix}\begin{bmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&-m_{43}&1 \end{bmatrix} = \begin{bmatrix} d_{11}&d_{12}&d_{13}&d_{14}\\ d_{21}&d_{22}&d_{23}&d_{24}\\ 0&d_{32}&d_{33}&d_{34}\\ 0&0&d_{43}&d_{44} \end{bmatrix} = D,$$

unde $m_{43} = -\dfrac{c_{42}}{c_{32}}$. Matricea D reprezintă forma Hessenberg a matricei A obținută cu ajutorul transformărilor similare.

În fiecare etapă au fost executate două operații matriceale, obținîndu-se perechile de matrice (A,B), (C,D). Matricele A și D sînt similare, deci au aceleași valori proprii.

Pentru a rezuma cele prezentate, se consideră cazul general, cu matricea A de forma

$$A = \begin{bmatrix} a_{11}&a_{12}&a_{13}&\dots&a_{1n}\\ a_{21}&a_{22}&a_{23}&\dots&a_{2n}\\ a_{31}&a_{32}&a_{33}&\dots&a_{3n}\\ \dots&\dots&\dots&\dots&\dots\\ a_{n1}&a_{n2}&a_{n3}&\dots&a_{nn} \end{bmatrix}.$$

Se examinează mărimea elementelor $a_{21}, a_{31}, \dots, a_{i1}, \dots$ $\dots, a_{n1}$. Fie a_{i1} elementul maxim care se alege drept pivot; în acest caz se va realiza schimbarea liniei 2 cu linia i și coloanei 2 cu coloana i cu ajutorul transformărilor similare. Rezultă

$$A'_1 = I_{2i} A I_{2i}.$$

În continuare are loc procesul de anulare a elementelor a'_{31}, $a'_{41}, \ldots, a'_{n1}$ de pe prima coloană a matricei A'_1 cu ajutorul matricei M_2, prin intermediul relației

$$A_2 = M_2 A'_1 M_2^{-1},$$

unde matricea M_2 are forma

$$M_2 = \begin{bmatrix} 1 & 0 & & & \\ 0 & 1 & & 0 & \\ 0 & m_{32} & 1 & & \\ 0 & m_{42} & 0 & 1 & \\ \cdot & \cdot & \cdot & \cdot & \\ 0 & m_{n2} & 0 & 0 \ldots & 1 \end{bmatrix}; \quad m_{i2} = -\frac{a'_{i1}}{a'_{21}}, i = 3, 4, \ldots, n.$$

Aceste relații constituie prima etapă în care se obține matricea A_2, avînd pe coloana principală elementele $a_{i1} = 0$, $i = 3, 4, \ldots, n$.

În etapa k ($k \leqslant n-2$) se dispune de matricea A_k de forma

$$A_k = \begin{bmatrix} b_{11} & b_{12} & \ldots & b_{1,k-2} & b_{1,k-1} & \cdot & b_{1k} & \ldots & b_{1n} \\ b_{21} & b_{22} & \ldots & b_{2,k-2} & b_{2,k-1} & \cdot & b_{2k} & \ldots & b_{2n} \\ 0 & b_{32} & \ldots & b_{3,k-2} & b_{3,k-1} & \cdot & b_{3k} & \ldots & b_{3n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 0 & b_{k,k-1} & \cdot & b_{k,k} & \ldots & b_{kn} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 0 & 0 & \cdot & b_{k+1,k} & \ldots & b_{k+1,n} \\ 0 & 0 & \ldots & 0 & 0 & \cdot & b_{k+2,k} & \ldots & b_{k+2,n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 0 & 0 & \cdot & b_{n,k} & \ldots & b_{nn} \end{bmatrix},$$

unde se vor analiza elementele $b_{k+1,k}, \ldots, b_{nk}$ pentru determinarea pivotului și a matricei de permutare $I_{k+1,r}$ cu $r = k+2, \ldots, n$. După această operație are loc transformarea similară

$$A'_k = I_{k+1,r} A_k I_{k+1,r}.$$

În continuare are loc anularea termenilor $b_{k+2,k}, \ldots, b_{n,k}$ cu ajutorul transformării similare

$$A_{k+1} = M_{k+1} A'_k M_{k+1}^{-1},$$

unde matricea M_{k+1} are forma următoare :

$$M_{k+1}=\begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & \ddots & & 0 & & \\ & & & \ddots & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & m_{k+2,\,k+1} & & \ddots & \\ & & 0 & m_{k+3,\,k+1} & & & \\ & & & \vdots & & & \\ & & & m_{n,\,k+1} & & & 1 \end{bmatrix}; \quad m_{r,\,k+1}=-\frac{b'_{rk}}{b'_{k+1,\,k}},$$

$$r=k+2, k+3, \ldots, n.$$

Aceste etape de calcul pot fi prezentate în general sub forma următorului proces iterativ :

$$A_{k+1}=[M_{k+1}I_{k+1,\,r}]A_k[M_{k+1}I_{k+1,\,r}]^{-1},\ k=1,2,\ldots,n-2, \quad (6.80)$$

în urma căruia are loc transformarea matricei $A=A_1$ într-o matrice A_{n-1} de forma Hessenberg.

În aplicațiile practice apare posibilitatea ca unele elemente sub diagonală a formei Hessenberg să fie nule; în cazul partiționării relativ la aceste elemente, forma Hessenberg A_{n-1} se poate scrie sub formă de blocuri superior triunghiulare, unde fiecare bloc este o matrice Hessenberg cu elemente subdiagonale nenule. În acest fel A_{n-1} se poate scrie sub forma

$$A_{n-1}=\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ & A_{22} & A_{23} \\ 0 & & A_{33} \end{bmatrix},$$

unde A_{11}, A_{22}, A_{33} sînt matrice Hessenberg cu elemente subdiagonale nenule [128, 60, 94].

Se poate arăta că algoritmul Q—R se poate aplica independent fiecărui bloc A_{ii} astfel că dacă $A_{ii}=Q_iR_i$ și $A_{n-1}=QR$, atunci Q este suma directă a matricelor Q_i și blocurile diagonale ale matricei QR sînt tocmai produsele R_iQ_i 119, 107].

Dacă matricea $A=A_1$ are un sistem de valori proprii care satisfac inegalitățile $|\lambda_1|>|\lambda_2|>\ldots>|\lambda_n|>0$, atunci șirul $\{A_i\}$ obținut prin algoritmul Q—R converge în

final către o matrice superior triunghiulară care conține pe diagonala principală valorile proprii.

Algoritmul Q—R se poate aplica și cu deplasarea originii, mai ales în cazul în care valorile proprii distincte au același modul. Astfel de valori proprii distincte se găsesc într-un cerc din planul complex (cu centrul în origine), iar deplasarea originii cu mărimea q poate schimba faptul că modulul valorilor proprii este egal, aceasta metodă de deplasare a originii conducînd la realizarea unui proces convergent.

Dacă se consideră o matrice Hessenberg H neredusă ale cărei valori proprii au module distincte, atunci există o transformare similară astfel că

$$PHP^{-1} = \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & 0 & \\ & & \ddots & \\ & 0 & & \lambda_n \end{bmatrix}. \qquad (6.81)$$

În cazul în care se formează matricea $H-qI$, adică dacă are loc deplasarea originii cu mărimea q, atunci (6.81) devine

$$P(H-qI)P^{-1} = \Lambda - qI = \begin{bmatrix} \lambda_1 - q & & & \\ & \lambda_2 - q & 0 & \\ & & \ddots & \\ & 0 & & \lambda_n - q \end{bmatrix}.$$

În acest fel se va scădea numărul q din fiecare valoare proprie.

În cazul în care se aplică algoritmul Q—R (cu deplasarea originii) matricei $H-qI$, convergența către zero a elementelor subdiagonale $h_{n,n-1}$ depinde de raportul $\left|\frac{\lambda_n - q}{\lambda_{n-1} - q}\right|$, iar dacă se găsește o metodă de alegere a lui q foarte apropiat de λ_n, procesul de convergență poate fi accelerat, dar aici apare o problemă că λ_n nu se cunoaște în momentul alegerii lui q, fapt care face [51, 59, 94] ca să se aleagă diferite valori pentru q în fiecare etapă iterativă.

În această situație pentru H_1, matrice Hessenberg ireductibilă, se formează următorul șir de iterații:

$$H_i - q_i I = Q_i R_i, \quad H_{i+1} = R_i Q_i + q_i I, \quad i = 1, 2, 3, \ldots$$

de unde se observă că H_{i+1} este simetrică cu H_i [42, 127].

6.6.7. *Valorile și vectorii proprii ai matricei Hessenberg*

Pentru rezolvarea completă a problemei considerate trebuie mai întîi determinate valorile și vectorii proprii ai matricei Hessenberg, după care se va opera asupra vectorilor proprii (ai matricei Hessenberg) cu matrice de transformare în scopul obținerii vectorilor proprii ai matricei A. În continuare se vor prezenta două metode pentru rezolvarea problemei propuse.

● Prima metodă constă în reducerea matricei Hessenberg prin transformări similare la o matrice de forma tridiagonală.

Dacă se consideră matricea Hessenberg $D = A_1$ dată în 6.6.6, se verifică ușor că transformarea similară $D_2 = M_2 A_1 M_2^{-1}$, unde

$$M_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & m_3 & m_{24} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}; \quad M_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -m_{23} & -m_{24} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \begin{matrix} m_{23} = d_{13}/d_{12}, \\ m_{24} = d_{14}/d_{12}, \end{matrix}$$

produce zerouri în locurile necesare din prima linie a matricei D_2, fără a afecta zerourile din prima coloană a lui D_2. În final transformarea $D_3 = M_3 D_2 M_3^{-1}$ cu

$$M_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & m_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad M_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -m_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

produce un zero pe ultima poziție din linia a doua, obținîndu-se o reducere a matricei Hessenberg D la formă tridiagonală. În această fază, soluția problemei se poate obține folosind orice metodă de calcul pentru valorile și vectorii proprii ai unei matrice tridiagonale nesimetrice.

● A doua metodă constă în folosirea matricei Hessenberg fără nici o altă transformare, obținîndu-se simultan cîte o valoare proprie și vectorul propriu corespunzător.

Pentru matricea Hessenberg D, dată în (6.6.6), se rezolvă sistemul de ecuații :

$$\left.\begin{aligned}(d_{11}-\lambda)x_1+d_{12}x_2+d_{13}x_3+d_{14}x_4&=0\\ d_{21}x_1+(d_{22}-\lambda)x_2+d_{23}x_3+d_{24}x_4&=0\\ d_{32}x_2+(d_{33}-\lambda)x_3+d_{34}x_4&=0\\ a_{43}x_3+(a_{44}-\lambda)x_4&=0\end{aligned}\right\}$$

Pentru orice valoare generală a lui λ se rezolvă ultimele trei ecuații în succesiune, alegîndu-se $x_4 = 1$ și, folosindu-se substituția inversă, rezultă x_3, x_2, x_1. Dacă se substituie în prima ecuație, se găsește un număr care este proporțional cu $|D - \lambda I|$ pentru o valoare particulară a lui λ. Folosindu-se metoda lui Müller sau alte metode (cap. 2), se pot găsi valorile proprii și vectorii proprii corespunzători.

6.7. Algoritmi de calcul pentru valorile și vectorii proprii în cazul matricelor hermitiene

Trebuie menționat că orice algoritm din cele prezentate (în paragraful 6.6) pentru matrice nehermitiene se poate aplica și matricelor hermitiene.

Fie $A \in M_{\Gamma}^{n\times n}$ o matrice hermitiană, care este diagonalizabilă prin intermediul transformărilor similare unitare, adică pentru orice A hermitiană există o matrice unitară R astfel că

$$R^H A R = \Lambda = \begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & 0 \\ & & \ddots & \\ & 0 & & \\ & & & \lambda_n\end{bmatrix},$$

unde $\lambda_1, \lambda_2, \ldots, \lambda_n$ sînt valori proprii reale ale lui A cărora le corespunde un sistem complet de vectori proprii ortonormali, care reprezintă coloanele matricei R.

Teoremă. *Fie $A \in M_{\Gamma}^{n\times n}$ hermitiană cu valorile proprii $\lambda_1, \lambda_2, \ldots, \lambda_n$. Atunci matricea A are n vectori mutual ortogonali și unitari $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^n$ care satisfac relația $A\mathbf{v}^i = \lambda_i \mathbf{v}^i$, $i = 1,2,\ldots,n$, și $R^H A R = \Lambda$, unde R este o matrice unitară cu coloanele $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^n$ și*

$$\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m) \quad [51, 93].$$

În cazul în care $A = A^H$, matricea A este bine condiționată în sensul că pentru variații de un anumit ordin a coeficienților matricei A au loc variații de același ordin pentru valorile proprii ale matricei A. O matrice reală și simetrică este hermitiană, iar rezultatele obținute pentru matricele reale simetrice se pot extinde la matrice hermitiene complexe. De aici rezultă faptul că pentru determinarea valorilor proprii în cazul matricelor hermitiene apare necesitatea utilizării calculelor aritmetice în real sau complex. În cazul în care nu se dorește folosirea aritmeticei complexe la determinarea valorilor proprii pentru o matrice hermitiană complexă (chiar dacă valorile proprii sînt reale, de obicei vectorii proprii sînt sub formă complexă), se scrie ecuația valorilor proprii $A\mathbf{x} = \lambda\mathbf{x}$ sub formă descompusă, adică pentru $A^H = A$ și $A \in M_C^{n\times n}$ rezultă

$$(B + \mathrm{i}C)(\mathbf{u} + \mathrm{i}\mathbf{v}) = \lambda(\mathbf{u} + \mathrm{i}\mathbf{v}),$$

unde $B \in M_R^{n\times n}$ este simetrică și $C \in M_R^{n\times n}$ cu $C^T = -C$, $\mathbf{u},\mathbf{v} \in R^n$. Dacă se identifică părțile reale și părțile imaginare, rezultă următoarele ecuații matriceale:

$$\begin{aligned} B\mathbf{u} - C\mathbf{v} &= \lambda\mathbf{u} \\ C\mathbf{u} + B\mathbf{v} &= \lambda\mathbf{v} \end{aligned} \quad \text{sau} \quad \begin{bmatrix} B & -C \\ C & B \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}. \qquad (6.82)$$

Ultima relație reprezintă o ecuație matriceală, unde matricea coeficient este reală și aparține lui $M_R^{2n\times 2n}$.

Se observă că dacă matricea hermitiană complexă are n valori proprii $\lambda_1, \lambda_2, \ldots, \lambda_n$, pentru evitarea aritmeticii complexe se ajunge la problema (6.82) care are $2n$ valori proprii: $\lambda_1, \lambda_1, \lambda_2, \ldots \lambda_n, \lambda_n$.

Exemplu. Fie matricea hermitiană complexă

$$A = \begin{bmatrix} 1 & 2+\mathrm{i} \\ 2-\mathrm{i} & -3 \end{bmatrix}, \quad A = B + \mathrm{i}C = \begin{bmatrix} 1 & 2 \\ 2 & -3 \end{bmatrix} + \mathrm{i}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Atunci (6.82) se reduce la

$$\begin{bmatrix} 1 & 2 & 0 & -1 \\ 2 & -2 & 1 & 0 \\ 0 & 1 & 1 & 2 \\ -1 & 0 & 2 & -3 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ v_1 \\ v_2 \end{bmatrix} = \lambda \begin{bmatrix} u_1 \\ u_2 \\ v_1 \\ v_2 \end{bmatrix}.$$

În continuare se vor prezenta cîțiva algoritmi de calcul mai frecvent întîlniți în practică [42, 100, 86].

6.7.1. *Algoritmul lui Jacobi*

Acest algoritm a fost introdus de Jacobi în anul 1846. Algoritmul urmărește transformarea matricei A în matricea Λ similară cu A, folosind o serie de transformări similare.

Pentru a motiva algoritmul lui Jacobi se pleacă de la analiza axelor principale ale unui elipsoid. Un elipsoid cu centrul în origine este reprezentat printr-o ecuație de forma

$$a_{11} x_1^2 + a_{22} x_2^2 + \ldots + a_{nn} x_n^2 + 2 \sum_{i<j}^{n} a_{ij} x_i x_j = 1$$

sau, pentru $a_{ij} = a_{ji}$, $i > j$, rezultă

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j = 1. \tag{6.83}$$

Fie matricea $A \in M_R^{n \times n}$ simetrică. În reprezentarea produsului scalar, ecuația (6.83) ia forma $(A\mathbf{x}, \mathbf{x}) = 1$, deoarece $\mathbf{x}$ are componente reale.

Axa principală a unui elipsoid are direcția unui vector din origine în punctul x de pe elipsoid astfel că vectorul este normal la elipsoid în punctul x. Pentru $n = 2$ se consideră fig. 6.5. O elipsă are două axe principale independente, un elipsoid în R^3 are trei axe principale independente. Din geometria analitică este cunoscut faptul că axele principale sînt sau pot fi alese mutual ortogonale.

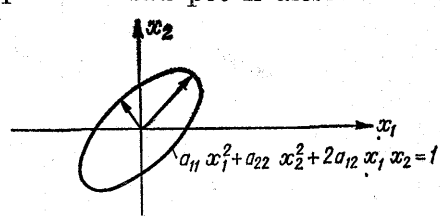


Fig. 6.5

Principalele axe ale unui elipsoid sînt vectorii proprii ai matricei A simetrice reale. Diferențiala $dx = (dx_1, dx_2, \ldots, dx_n)$ pe suprafața $\Phi(x_1, x_2, \ldots, x_n) = \text{const}$ satisface relația

$$d\Phi = \frac{\partial \Phi}{\partial x_1} dx_1 + \frac{\partial \Phi}{\partial x_2} dx_2 + \ldots + \frac{\partial \Phi}{\partial x_n} dx_n = 0.$$

Vectorul grad $\Phi = (\partial \Phi / \partial x_k)$ $(k = 1, 2, \ldots, n)$ fiind normal la toate diferențele $dx = d(x_k)$ în apropierea punctului x, este, prin urmare, un vector normal la suprafața $\Phi = \text{const}$ în punctul x.

Fie $\Phi(x)$ o formă pătratică $\sum\sum a_{ij} x_i x_j$; atunci se poate calcula vectorul normal la elipsoidul $\Phi(x) = 1$:

$$\frac{\partial \Phi}{\partial x_k} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \frac{\partial (x_i x_j)}{\partial x_k}, \quad k = 1, 2, \ldots, n. \tag{6.84}$$

Folosind regula de derivare a unui produs:

$$\frac{\partial (x_i x_j)}{\partial x_k} = \frac{\partial x_i}{\partial x_k} x_j + x_i \frac{\partial x_j}{\partial x_k} = \delta_{ik} x_j + x_i \delta_{jk}. \tag{6.85}$$

Dacă se introduce (6.85) în (6.84), după operația de însumare rezultă

$$\frac{\partial \Phi}{\partial x_k} = \left(\sum_j a_{kj} x_j + \sum_i a_{ik} x_i \right). \tag{6.86}$$

Dacă $a_{ik} = a_{ki}$, atunci (6.86) devine

$$\frac{\partial \Phi}{\partial x_k} = 2 \sum_{j=1}^{n} a_{kj} a_j = 2Ax.$$

Vectorul normal Ax și vectorul radial x sînt reprezentați în fig. 6.6. Vectorul x nu este o axă principală pentru că nu are aceeași direcție cu Ax.

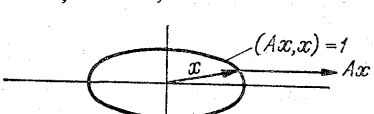


Fig. 6.6

Ecuația care definește axele principale este

$$A\mathbf{x} = \lambda\mathbf{x} \tag{6.87}$$

pentru anumite valori ale scalarului λ, unde $(A\mathbf{x},\mathbf{x}) = 1$. Deoarece $\mathbf{x} \neq \mathbf{0}$, axa principală este un vector propriu al matricei A. Lungimea axei principale asociate cu valoarea proprie λ_1 este $1/\sqrt{\lambda_1}$. Pentru a demonstra acest lucru se consideră produsul scalar în ambele părți ale relației (6.87), adică

$$1 = (A\mathbf{x}, \mathbf{x}) = (\lambda\mathbf{x}, \mathbf{x}) = \lambda \|\mathbf{x}\|^2, \|\mathbf{x}\| = 1/\sqrt{\lambda}.$$

Forma pătratică $(A\mathbf{x}, \mathbf{x}) = 1$ reprezintă ecuația unui elipsoid.

Exemplu. Se consideră elipsa

$$x_1^2 + 2x_1x_2 + 3x_2^2 = 1, \ (A\mathbf{x},\mathbf{x}) = 1,$$

unde

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}; \quad \begin{bmatrix} \lambda - 1 & 1 \\ 1 & \lambda - 3 \end{bmatrix} = \lambda^2 - 4\lambda + 2 = 0, \quad \begin{matrix} \lambda_1 = 2 + \sqrt{2}, \\ \lambda_2 = 2 - \sqrt{2}. \end{matrix}$$

La axa principală $\mathbf{x}^1$ îi corespunde λ_1 :

$$(\lambda_1 I - A)\mathbf{x} = \begin{bmatrix} 1 + \sqrt{2} & -1 \\ -1 & -1 + \sqrt{2} \end{bmatrix} \begin{bmatrix} x_1^1 \\ x_2^1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Axele principale $\mathbf{x}^1$ și $\mathbf{x}^2$ sînt ortogonale :

$$\mathbf{x}^1 = \alpha \begin{bmatrix} 1 + \sqrt{2} \\ 1 \end{bmatrix}, \ \mathbf{x}^2 = \beta \begin{bmatrix} 1 - \sqrt{2} \\ 1 \end{bmatrix},$$

iar constantele α și β pot fi determinate astfel ca $\mathbf{x}^1$ și $\mathbf{x}^2$ să se găsească în elipsa considerată inițial, iar

$$\|\mathbf{x}^1\| = \frac{1}{\sqrt{\lambda_1}} = \frac{1}{\sqrt{2 + \sqrt{2}}}, \ \|\mathbf{x}^2\| = \frac{1}{\sqrt{\lambda_2}} = \frac{1}{\sqrt{2 - \sqrt{2}}}.$$

Se consideră o formă pătratică în două variabile

$$F_2 = a_{11}x_1^2 + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{22}x_2^2 =$$

$$= [x_1, \ x_2]^T \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{x}^T A\mathbf{x}.$$

În cazul în care $F_2 = c$, $c \in R$, aceasta reprezintă ecuația unei conice, de exemplu o elipsă, ca în fig. 6.7.

Metoda Jacobi folosește rotirea elipsei pentru ca noile axe să coincidă cu axele principale ale elipsei. Realizarea produsului de rotație presupune schimbarea de axe:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix} \cdot \begin{bmatrix} t_1 \\ t_2 \end{bmatrix},$$

$$\mathbf{x} = R\mathbf{t}.$$

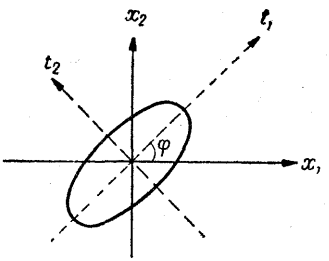


Fig. 6.7

Dacă $A = A^T$, atunci pentru $\mathbf{x} = R\mathbf{t}$ și $\mathbf{x}^T = (R\,\mathbf{t})^T = \mathbf{t}^T(R^T)$ și F_2 devine

$$F_2 = \mathbf{x}^T A\mathbf{x} = \mathbf{t}^T R^T A R\mathbf{t} = \mathbf{t}^T \Lambda\mathbf{t} = \lambda_1 t_1^2 + \lambda_2 t_2^2.$$

Deoarece $R^T = R^{-1}$, atunci

$$R^T A R = \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix},$$

A și Λ fiind două matrice similare.

Efectul acestui proces de rotație este de a anula din forma pătratică F_2 doi termeni $a_{12}x_1x_2$ și $a_{21}x_2x_1$, sau de a reduce matricea A la o matrice diagonală Λ similară cu A, unde Λ are pe diagonala principală valorile proprii ale matricei A considerate. Datorită faptului că $R^T = R^{-1}$, transformarea $R^T A R = \Lambda$ este o transformare similară ortogonală.

Pentru spațiul n dimensional se introduce o matrice de rotație $R(k, p)$ care este o generalizare a matricei din planul bidimensional. Matricea de rotație $R(k, p)$ este definită

astfel :

$$R(k,p) = \begin{array}{c} \\ k \\ \\ p \\ \\ \end{array} \begin{array}{c} \begin{array}{cccccccc} k & & & & p & & & \end{array} \\ \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & \cos\varphi & \cdots & \sin\varphi & & \\ & & & \vdots & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & \\ & & & \sin\varphi & \cdots & \cos\varphi & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{bmatrix} \end{array} \tag{6.88}$$

unde $k < p$ şi $r_{kk} = r_{pp} = \cos\varphi$, $r_{ii} = 1$, $i \neq p$, $i = k$,

$-r_{kp} = r_{pk} = \sin\varphi$, $r_{ij} = 0$ în celelalte cazuri.

Transformare similară ortogonală prin $R(k, p)$ va anula termenii în $x_p x_k$ şi $x_k x_p$ din forma pătratică :

$$F_n = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j =$$

$$= [x_1 x_2 \ldots x_n] \begin{bmatrix} a_{11} & a_{12} \ldots a_{1n} \\ a_{21} & a_{22} \ldots a_{2n} \\ \ldots & \ldots \\ a_{n1} & a_{n2} \ldots a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \mathbf{x}^T A \mathbf{x}. \tag{6.89}$$

Pentru $\mathbf{x} = R(k, p)\mathbf{t}$ şi $\mathbf{x}^T = [R(k, p)\mathbf{t}]^T = \mathbf{t}^T R(k, p)$, atunci (6.89) devine

$$F_n = \mathbf{t}^T R^T(k, p) A R(k, p)\mathbf{t} = \mathbf{t}^T A' \mathbf{t},$$

unde

$$A' = R^T(k, p) A R(k, p). \tag{6.90}$$

Prin intermediul acestei transformări similare ortogonale s-a obținut matricea A' în care elementele a_{pk} și a_{kp} au fost anulate, fapt pentru care (6.90) reprezintă o rotație în planul (k, p).

Relația matriceală (6.90) conduce la următorul algoritm pentru calculul elementelor matricei A' în funcție de elementele matricei A și unghiul de rotație φ (algoritmi ce rezultă în urma multiplicării și identificării):

$$\left.\begin{aligned} a'_{ki} &= \quad a_{ki}\cos\varphi + a_{pi}\sin\varphi \\ a'_{pi} &= -a_{ki}\sin\varphi + a_{pi}\cos\varphi \end{aligned}\right\}$$

$$\left.\begin{aligned} a'_{ik} &= \quad a_{ik}\cos\varphi + a_{ip}\sin\varphi \\ a'_{ip} &= -a_{ik}\sin\varphi + a_{ip}\cos\varphi \end{aligned}\right\} \quad i \neq p,k, \tag{6.91}$$

$$\left.\begin{aligned} a'_{kk} &= a_{kk}\cos^2\varphi + 2a_{kp}\sin\varphi\cos\varphi + a_{pp}\sin^2\varphi \\ a'_{pp} &= a_{pp}\sin^2\varphi - 2a_{pk}\sin\varphi\cos\varphi + a_{pp}\cos^2\varphi \end{aligned}\right\}, \tag{6.92}$$

$$\left.\begin{aligned} a'_{pk} &= a_{pk}\cos 2\varphi + \frac{1}{2}(a_{pp} - a_{kk})\sin 2\varphi \\ a'_{kp} &= a_{kp}\cos 2\varphi + \frac{1}{2}(a_{kk} - a_{pp})\sin 2\varphi \end{aligned}\right\}. \tag{6.93}$$

Datorită faptului că rotația în planul (k, p) are rolul să anuleze coeficienții a_{pk} și a_{kp}, adică $a_{pk} = a_{kp} = 0$, rezultă valoarea lui φ din

$$\operatorname{tg} 2\varphi = \frac{2a_{pk}}{a_{kk} - a_{pp}} \; ; \quad k < p.$$

Pentru $a_{kk} = a_{pp}$, rezultă $\varphi = \frac{\pi}{4}$. Elementele matricei A se calculează cu relațiile (6.91)—(6.93).

Folosind relația (6.90), se poate defini un șir de matrice similare dînd alte valori lui k și p;

$$A_1 = A,\ A_2 = R_1^T A_1 R_1,\ A_3 = R_2^T A_2 R_2, \ldots,\ A_{k+1} = R_k^T A_k R_k, \ldots \tag{6.94}$$

unde în fiecare etapă transformarea similară ortogonală are drept scop să anuleze două elemente simetrice față de diagonala principală.

Rotația plană (k, p) are drept pivot elementul $a_{p\,k}$. Dacă pentru fiecare transformare $R_k^T A R_k$ pivotul rotației plane are valoarea mai mare decît media din afara diagonalei lui A_k, atunci pentru $k \to \infty$ are loc limita [128, 119]:

$$A_k \to \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & 0 \\ & & \ddots & \\ & 0 & & \lambda_n \end{bmatrix}.$$

Toate matricele șirului $A_1, A_2, \ldots, A_k, \ldots$ sînt similare, deci ele au același sistem de valori proprii. Din punct de vedere practic, șirul de iterații (6.94) este continuat pînă cînd suma pătratelor elementelor din afara diagonalei principale a matricei A_{k+1} este mai mică decît un $\varepsilon > 0$; dacă condiția aceasta este indeplinită, atunci elementul de pe diagonala principală aproximează destul de bine valorile proprii pentru matricea A [71].

Convergența șirului (6.94) poate avea loc chiar dacă pivoții nu sînt aleși pe baza valorii lor maxime, dar sînt aleși [60, 93] în următoarea ordine: $(2,1), (3,1), (3,2), (4,1), (4,2), (4,3), \ldots, (n,1)\,(n,2), \ldots, (n, n-1)(2,1), \ldots$, utilizînd $\varphi \leqslant \frac{\pi}{4}$. Alegerea pivoților de rotație în această ordine conduce la algoritmul ciclic Jacobi. În ambele situații trebuie ținut seama de stabilitatea numerică a matricei A [cond. (A)].

Șirul de iterații (6.94) se mai poate scrie și sub forma

$$A_{k+1} = (R_1 R_2 R_3 \ldots R_k)^T A_1 (R_1 R_2 \ldots R_k) = R^T A R, \qquad (6.95)$$

unde $R = R_1 R_2 \ldots R_k$ este produsul primelor k matrice de rotație.

Matricea R este destul de aproape de matricea care are drept coloane vectorii proprii ai matricei A, adică

$$R^T A R \approx \Lambda \Rightarrow A R \approx R \Lambda,$$

obținîndu-se o aproximare a ecuației modale asociate matricei A [100, 94]. Coloanele matricei R aproximează vectorii proprii ai matricei A, coloana r_j reprezintă vectorul propriu $\mathbf{v}^j$, corespunzător valorii proprii λ_j, $j = 1,2,\ldots,n$, fapt pentru care se afirmă [51], că algoritmul Jacobi oferă același ordin de precizie pentru calculul valorilor și vectorilor proprii ai matricei A.

În cadrul acestor procese de rotație, elementele de pe pozițiile (p,k) și (k,p) nu devin zero în urma rotației plane (k,p) dar după un număr de iterații devin nesemnificative (putînd fi considerate egale cu zero) și elementele de pe diagonala principală aproximează destul de bine valorile proprii ale matricei A.

6.7.2. *Algoritmul lui Givens*

În 1954 Givens a propus [51] un algoritm îmbunătățit care prin transformări similare face ca elementele din afara diagonalei principale să devină nule.

Pentru a înțelege etapele algoritmului, se consideră o matrice de ordinul patru, care se rotește la început în planul (2, 3) pentru care $R(2, 3)$ este de forma

$$R(2,3) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\varphi & \sin\varphi & 0 \\ 0 & -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

În urma acestui proces de rotație cu pivotul a_{32}, se anulează elementele a_{31} și a_{13} (în loc să se anuleze elementele a_{32} și a_{23}, scop urmărit la algoritmul Jacobi).

Algoritmul Givens constă în alegerea pivoților din pozițiile $(3, 2)$, $(4, 2)$, $\ldots$, $(n, 2)$ și unghiurile φ_i corespunzătoare pentru (anularea perechilor simetrice din pozițiile $(1, 3)$ și $(3, 1)$, $(4, 1)$, $(1, 4)$,$\ldots$,$(n, 1)$ și $(1, n)$ respectiv. În urma acestui proces de transformare se obține dintr-o matrice simetrică A matricea A', care are $n-2$ zerouri

în prima linie și $n-2$ zerouri în prima coloană:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} \longrightarrow$$

$$\longrightarrow \begin{bmatrix} a'_{11} & a'_{12} & 0 & 0 & \dots & 0 \\ a'_{12} & a'_{22} & a'_{23} & a'_{24} & \dots & a'_{2n} \\ 0 & a'_{32} & a'_{33} & a'_{34} & \dots & a'_{3n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & a'_{n2} & a'_{n3} & a'_{n4} & \dots & a'_{nn} \end{bmatrix} = A'.$$

Dacă în continuare se aleg pivoții din coloana a treia în pozițiile $(4, 3), (5, 3), \dots, (n, 3)$ și unghiurile de rotație φ_i corespunzătoare se anulează, atunci elementele simetrice din pozițiile $(4, 2)$ și $(2, 4)$, $(5, 2)$ și $(2, 5), \dots, (n, 2)$ și $(2, n)$ devin nule, obținîndu-se matricea

$$\begin{bmatrix} a''_{11} & a''_{12} & 0 & 0 & 0 & \dots & 0 \\ a''_{21} & a''_{22} & a''_{23} & 0 & 0 & \dots & 0 \\ 0 & a''_{32} & a''_{33} & a''_{34} & a''_{35} & \dots & a''_{3n} \\ 0 & 0 & a''_{43} & a''_{44} & & & \cdot \\ 0 & 0 & a''_{53} & a''_{54} & & & \cdot \\ \dots & \dots & \dots & \dots & \dots & \dots & \cdot \\ 0 & 0 & a''_{n3} & & a''_{n4} & \dots & a''_{nn} \end{bmatrix}.$$

Matricea A'' are $n-3$ zerouri în coloana a doua și $n-3$ în linia a doua.

Dacă se continuă acest proces în maniera celor două etape prezentate pentru obținerea lui A' și A'', atunci după $(n-2)+(n-3)+\dots+2+1 = \frac{1}{2}(n-1)(n-2)$ rotații plane se obține o matrice simetrică reală tridiago-

nală T de forma

$$\begin{bmatrix} \times & \times & & & & & \\ \times & \times & \times & & & 0 & \\ & \times & \times & \times & & & \\ & & & & \times & \times & \times \\ 0 & & & & & \times & \times \end{bmatrix} = T. \qquad (6.96)$$

Se observă că matricea T obținută prin algoritmul Givens este o matrice tridiagonală care are aceleași valori proprii ca și matricea A, fiind necesară o nouă etapă de calcul pentru determinarea valorilor proprii. Deci, algoritmul Givens constă din două etape : prima are ca obiectiv obținerea unei matrice tridiagonale (operație ce se realizează printr-un număr finit de transformări similare), iar a doua etapă constă în determinarea valorilor proprii pentru o matrice tridiagonală. Algoritmul Givens este mult mai eficient atît din punctul de vedere al numărului de operații cît și al preciziei.

6.7.3. *Algoritm de calcul pentru valorile proprii ale unei matrice tridiagonale simetrice*

Fie matricea tridiagonală (6.96) obținută prin metoda Givens. Atunci polinomul caracteristic corespunzător lui T are expresia

$$Q_n(\lambda) = \det(T - \lambda I) =$$

$$= \det \begin{bmatrix} a_1 - \lambda & b_1 & & & & 0 & \\ b_1 & a_2 - \lambda & b_2 & & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & b_{n-2} & a_{n-1} - \lambda & b_{n-1} & \\ 0 & & & & b_{n-1} & a_n - \lambda & \end{bmatrix}. \qquad (6.97)$$

Se presupune că $b_i \neq 0$, $i = 1, 2, \ldots, n-1$, dar chiar dacă $b_i = 0$ pentru un anumit i, matricea este bloc tridia-

gonală și se poate lucra cu fiecare bloc separat. Se va calcula polinomul caracteristic în mod direct (fără evaluarea coeficienților lui în funcție de elementele matricei). În [42, 100] se prezintă un procedeu de recurență pentru determinarea polinomului caracteristic $Q_n(\lambda)$ asociat matricei T.

Fie $Q_i(\lambda)$ determinantul submatricei formate din primele i linii și coloane ale matricei $T-\lambda I$. Atunci din (6.97) pentru $i = 1, 2, 3$ se obține

$$Q_1(\lambda) = a_1-\lambda,$$

$$Q_2(\lambda) = (a_2-\lambda)Q_1(\lambda)-b_1^2,$$

$$Q_3(\lambda) = (a_3-\lambda)Q_2(\lambda)-b_2^2Q_1(\lambda),$$

rezultînd în final o formulă de recurență. Dacă T este o matrice simetrică reală tridiagonală și dacă $T-\lambda I$ este definită ca în (6.97), atunci pentru $i = 1, 2, \ldots, n$ are loc relația de recurență

$$Q_i(\lambda) = (a_i - \lambda)Q_{i-1}(\lambda) - b_i^2\, Q_{i-2}(\lambda), \qquad (6.98)$$

unde se definește $Q_{-1}(\lambda) \equiv 0$, $Q_0(\lambda) = 1$ și $b_0 = 0$, sau

$$Q_i(\lambda) = \det \begin{bmatrix} a_1-\lambda & b_1 & & & & \\ b_1 & a_2-\lambda & b_2 & & 0 & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & b_{i-2} & \ddots & \ddots & \\ & 0 & & a_{i-1}-\lambda & b_{i-1} & \\ & & & b_{i-1} & a_i-\lambda & \end{bmatrix}.$$

Orice submatrice principală T_i pentru $i = 1, 2, \ldots, n$ din T este simetrică, reală și are valori proprii reale, adică rădăcinile ecuației caracteristice $Q_i(\lambda) = 0$ sînt reale pentru $i = 1, 2, \ldots, n$.

Ecuația caracteristică a matricei T este $Q_n(\lambda) = 0$. Algoritmul urmărește găsirea rădăcinilor pentru $Q_n(\lambda)=0$

fără a calcula coeficienții pentru $Q_n(\lambda)$ în funcție de elementele matricei T.

Teorema lui Givens. *Fie T o matrice tridiagonală, simetrică, reală cu $b_i \neq 0$, $i = 1, 2, \ldots, n$. Atunci rădăcinile fiecărei ecuații $Q_i(\lambda) = 0$ sînt distincte și sînt separate prin rădăcinile lui $Q_{i-1}(\lambda) = 0$.*

Demonstrație. S-a presupus $b_i \neq 0$, $i = 1, 2, \ldots, n-1$. Atunci două polinoame consecutive din șir nu pot avea o rădăcină comună pentru că dacă ar exista un λ_1 astfel ca $Q_{i-2}(\lambda_1) = 0$ și $Q_{i-3}(\lambda_1) = 0$, atunci din (6.98) rezultă că și $Q_{i-1}(\lambda_1) = 0$ ș.a.m.d.; se ajunge la faptul că și $Q_1(\lambda) = 0$, dar $Q_1(\lambda) = 0$ are o singură rădăcină a_1, care nu este rădăcină pentru $Q_2(\lambda)$ deoarece din ipoteza $b_1 \neq 0$.

Pentru a demonstra că rădăcinile lui $Q_{i-1}(\lambda) = 0$ separă rădăcinile lui $Q_i(\lambda) = 0$ se folosește inducția. Este ușor de verificat că a_1, rădăcina ecuației $Q_1(\lambda) = 0$, se găsește întredouă rădăcini distincte ale ecuației $Q_2(\lambda) = 0$. Se presupune că rădăcinile lui $Q_{i-2}(\lambda) = 0$ și $Q_{i-1}(\lambda) = 0$ sînt distincte și că rădăcinile primei ecuații separă rădăcinile ultimei.

Fie $t_1 < t_2 < \ldots < t_{i-1}$ rădăcinile ecuației $Q_{i-1}(\lambda) = 0$. Atunci din formula de recurență (6.98) pentru $k = 1, 2, \ldots, i-1$ are loc relația

$$Q_i(t_k) = b_{i-1}^2 Q_{i-2}(t_k), \quad k = 1, 2, \ldots, i-1,$$

relație care implică faptul că $Q_i(t_k)$ și $Q_{i-2}(t_k)$ au semne opuse. Prin inducția folosită în ipoteza, $Q_{i-2}(\lambda)$ schimbă semnul între t_k și t_{k-1}, $k = 1, 2, \ldots, i-2$, fapt ce arată că $Q_i(\lambda)$ schimbă semnul între t_k și t_{k+1}. Altfel spus, $Q_i(\lambda) = 0$ are o rădăcină între fiecare pereche de rădăcini adiacente ale ecuației $Q_{i-1}(\lambda) = 0$. Deoarece

$$Q_i(\lambda) \to \begin{cases} \infty, & \lambda \to -\infty, \\ (-1)^i \infty, & \lambda \to \infty, \end{cases} \quad i = 1, 2, \ldots, n,$$

rezultă că $Q_i(\lambda) = 0$ are o rădăcină la dreapta lui t_{i-1} și o rădăcină la stînga lui t_1. În cazul în care rădăcinile polinomului $Q_i(\lambda) = 0$ sînt $r_1, r_2, \ldots, r_i$, există relația de ordine

$$t_1 < r_1 < t_2 < r_2 < \ldots < r_{i-1} < t_{i-1} < r_i,$$

adică rădăcinile lui $Q_i(\lambda) = 0$ sînt distincte și separate prin rădăcinile lui $Q_{i-1}(\lambda) = 0$.

În [128, 95] se consideră șirul de polinoame reale $Q_0(x), Q_1(x), Q_2(x), \ldots, Q_m(x)$ cu următoarele două proprietăți relativ la intervalul (a, b) (unde a poate fi $-\infty$ și b poate fi $+\infty$):

1) Pentru orice valoare $x_0 \in (a, b)$, dacă $Q_k(x_0) = 0$, atunci

$$Q_{k-1}(x_0)\, Q_{k+1}(x_0) < 0;$$

2) $Q_0(x) \neq 0$ pentru orice $x \in (a, b)$.

Un șir de polinoame $Q_0(x), Q_1(x), \ldots, Q_m(x)$ cu aceste două proprietăți este un șir Sturm pe intervalul (a, b).

Dacă se notează cu $S(c)$ numărul perechillor de polinoame consecutive din șirul $Q_0(c), Q_1(c), \ldots, Q_n(c)$, care au același semn, atunci se poate enunța următoarea

Teoremă. *Mărimea $S(c)$ reprezintă numărul rădăcinilor lui $Q_n(\lambda) = 0$, care sînt mai mari sau egale cu c.*

Exemplu. Fie $A \in M_R^{3\times 3}$ o matrice care prin intermediul primei etape a algoritmului Givens este transformată într-o formă tridiagonală T, unde

$$T = \begin{bmatrix} a_1 & b_1 & 0 \\ b_1 & a_2 & b_2 \\ 0 & b_2 & a_3 \end{bmatrix}, \quad T - \lambda I = \begin{bmatrix} a_1 - \lambda & b_1 & 0 \\ b_1 & a_2 - \lambda & b_2 \\ 0 & b_2 & a_3 - \lambda \end{bmatrix}.$$

Pentru cazul considerat rezultă următorul șir de polinoame:

$$Q_0(\lambda) = 1, \qquad Q_2(\lambda) = (a_2 - \lambda)Q_1(\lambda) - b_1^2\, Q_0(\lambda),$$

$$Q_1(\lambda) = a_1 - \lambda, \qquad Q_3(\lambda) = (a_3 - \lambda)Q_2(\lambda) - b_2^2\, Q_1(\lambda).$$

În fig. 6.8 sînt reprezentate grafic polinoamele $Q_i(\lambda)$, $i = 0, 1, 2, 3$. Ținînd seama de proprietățile acestor polinoame, se poate construi tabelul 6.1 pentru semnele polinoamelor $Q_i(\lambda)$. Numărul $S(c)$ este egal cu numărul rădăcinilor polinomului $Q_3(\lambda) = 0$, rădăcini ce sînt mai mari sau egale cu

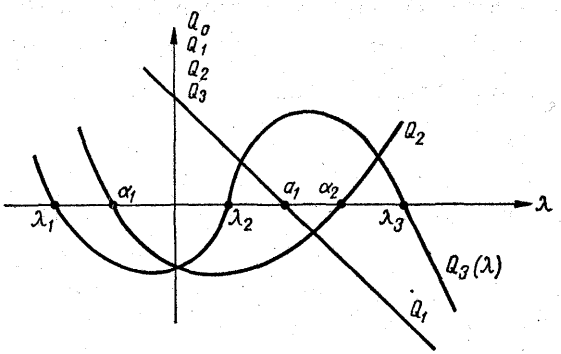


Fig. 6.8

constanta c. Se observă din tabel că pentru $c \in (\lambda_3, +\infty)$ avem $S(c) = 0$, adică pentru $c > \lambda_3$, $Q_3(\lambda) \neq 0$, deci toate rădăcinile pentru $Q_3(\lambda) = 0$ sînt mai mici decît λ_3.

	$(-\infty,\lambda_1]$	$[\lambda_1,\alpha_1]$	$[\alpha_1,\lambda_2]$	$[\lambda_2,a_1]$	$[a_1,\alpha_2]$	$[\alpha_2,\lambda_3]$	$[\lambda_3,+\infty)$
$Q_0(\lambda)=1$	+	+	+	+	+	+	+
$Q_1(\lambda)$	+	+	+	+	−	−	−
$Q_2(\lambda)$	+	+	−	−	−	+	+
$Q_3(\lambda)$	+	−	−	+	+	+	−
	3	2	2	1	1	1	0

Se consideră noțiunea de spectru al unei matrice $S(T) \leqslant \|T\|_p$ pentru orice normă matriceală p.

Este cunoscut faptul [119, 108] că toate valorile proprii ale matricei T similare cu A [rădăcinile polinomului $Q_n(\lambda)=0$] se găsesc în intervalul închis $[-\|T\|_\infty, \|T\|_\infty]$. Deoarece valorile proprii sînt reale și distincte, se poate continua procesul de bisecție pînă cînd se obține un interval care conține o singură valoare proprie.

În momentul în care se cunoaște intervalul în care se află o singură valoare proprie, metoda bisecției poate continua pînă cînd valoarea proprie căutată se determină cu precizia dorită [124, 60].

O altă metodă pentru determinarea valorilor proprii ale unei matrice T de tipul (6.96) folosește teorema Gerșgorin care arată că valorile proprii se găsesc în intervale închise de pe axa reală de forma

$$\lambda_1 \in [a_1 - |b_1|,\ a_1 + |b_1|],$$

$$\lambda_i \in [a_i - |b_{i-1}| - |b_i|,\ a_i + |b_{i-1}| + |b_i|],$$

$$i = 2, 3, \ldots, n-1,$$

$$\lambda_n \in [a_n - |b_{n-1}|,\ a_n + |b_{n-1}|].$$

Există și alte metode de transformare a matricei A într-o matrice tridiagonală T, precum și o serie de programe pentru calcularea efectivă a valorilor și vectorilor proprii [124].

6.7.4. *Algoritmul Givens-Householder*

În 1958 Householder a introdus o metodă de tridiagonalizare a matricelor care necesită aproape jumătate din calculele cerute de algoritmul Givens [60]. În [128] se prezintă o combinație între metoda Householder și Givens sub denumirea de algoritmul Givens-Householder, realizîndu-se o serie de programe pe calculator care determină valorile și vectorii proprii cu o precizie suficient de bună.

Algoritmul Givens utiliza $n-j-1$ rotații plane pentru a crea $n-j-1$ zerouri în linia j și $n-j-1$ zerouri în coloana j a matricei A, în timp ce algoritmul Householder folosește o singură transformare similară ortogonală, creînd același număr de zerouri în aceleași poziții.

Operația de tridiagonalizare prin algoritmul Householder necesită crearea de zerouri în cele $n-2$ linii și coloane (în afara elementelor celor trei diagonale principale); operație ce se realizează prin $n-2$ transformări Householder. Transformările Householder sînt mult mai complexe decît rotațiile plane propuse de Givens, dar metoda Householder necesită doar jumătate din calculele reclamate de algoritmul Givens.

Algoritmul Householder foloseşte transformarea secvenţială :

$$A_k = H_k A_{k-1} H_k, \quad A_0 = A,$$

unde $H \in M_R^{n\times n}$ şi $\mathbf{v} \in R^n$, $H_k = I - 2\,\mathbf{v}^k(\mathbf{v}^k)^T$; $\mathbf{v}\mathbf{v}^T = 1$, $\mathbf{v}\mathbf{v}^T = v_1^2 + v_2^2 + \ldots + v_n^2 = 1$.

Matricea H în formă dezvoltată arată astfel :

$$H = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix} - 2 \begin{bmatrix} v_1^2 & v_1v_2 & v_1v_3 & \ldots & v_1v_n \\ v_2v_1 & v_2^2 & v_2v_3 & \ldots & v_2v_n \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ v_nv_1 & v_nv_2 & v_nv_3 & \ldots & v_n^2 \end{bmatrix} =$$

$$= \begin{bmatrix} 1-2v_1^2 & -2v_1v_2 & -2v_1v_3 & \ldots & -2v_1v_n \\ -2v_1v_2 & 1-2v_2^2 & -2v_2v_3 & \ldots & -2v_2v_n \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ -2v_nv_1 & -2v_nv_2 & -2v_nv_3 & \ldots & 1-2v_n^2 \end{bmatrix}. \tag{6.99}$$

şi are următoarele proprietăţi :

$$\text{a) } H = H^T \text{ şi b) } H = H^{-1} \Rightarrow \text{c) } H^T = H^{-1}.$$

Proprietatea a) a matricei H rezultă din analiza expresiei matricei H. Pentru a pune în evidenţă proprietatea b), se pleacă de la faptul că a) există, adică $H = H^T$; atunci

$$H^TH = H^2 = (I - 2\mathbf{v}\mathbf{v}^T)(I - 2\mathbf{v}\mathbf{v}^T) = I - 4\,\mathbf{v}\mathbf{v}^T + \\ + 4\mathbf{v}(\mathbf{v}^T\mathbf{v})\mathbf{v}^T = I - 4\mathbf{v}\mathbf{v}^T + 4\mathbf{v}\mathbf{v}^T = I.$$

Din această ultimă relaţie se vede că $H^T = H^{-1} = H$, adică H este simetrică şi ortogonală. Se consideră $n-2$ transformări Householder începînd cu matricea simetrică $A \in M_R^{n\times n}$, adică

$$A_1 = A, A_2 = H_2A_1H_2, A_3 = H_3A_2H_3, \ldots, A_{n-1} = \\ = H_{n-1}A_{n-2}H_{n-1}. \tag{6.100}$$

Matricea H_k, pentru $k = 2, 3, \ldots, n-1$, este o matrice de forma

$$H_k = I - 2\mathbf{v}^{(k)}(\mathbf{v}^{(k)})^T, \quad \mathbf{v}^{(k)} = \begin{bmatrix} 0 \\ \vdots \\ v_k \\ v_{k+1} \\ \vdots \\ v_n \end{bmatrix} \text{ și } v_k^2 + v_{k+1}^2 + \ldots$$

$$\ldots + v_n^2 = 1.$$

Din (6.100) se vede că obiectivul algoritmului Householder este de a transforma o matrice simetrică reală $A = A_1$ într-o matrice tridiagonală $A_{n-1} = T$ prin $n-2$ transformări similare ortogonale. Pentru a determina matricea A_k din etapa k trebuie la început calculate componentele $v_k, v_{k+1}, \ldots, v_n$ ale vectorului $\mathbf{v}^{(k)}$ astfel ca transformarea $H_k A_{k-1} H_k$ să anuleze $n-k$ elemente din afara celor trei diagonale în linia $k-1$ și coloana $k-1$.

Pentru mai multă claritate, se consideră o matrice $A \in MR^{3\times 3}$ și $k = 2$. Fie:

$$A = A_1 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad A_2 = H_2 A_1 H_2,$$

unde

$$H_2 = I - 2\mathbf{v}^2(\mathbf{v}^2)^T, \quad \mathbf{v}^2 = \begin{bmatrix} 0 \\ v_2 \\ v_3 \end{bmatrix}, \quad v_2^2 + v_3^2 = 1,$$

$$H_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - 2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & v_2^2 & v_2 v_3 \\ 0 & v_3 v_2 & v_3^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 - 2v_2^2 & -2v_3 v_2 \\ 0 & -2v_3 v_2 & 1 - 2v_3^2 \end{bmatrix};$$

$$A_2 = H_2 A_1 H_2 =$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-2v_2^2 & -2v_2v_3 \\ 0 & -v_3v_2 & 1-2v_3^2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-2v_2^2 & -2v_2v_3 \\ 0 & -2v_3v_2 & 1-2v_3^2 \end{bmatrix} =$$

$$= \begin{bmatrix} a_{11} & a'_{12} & a'_{13} \\ a_{21}(1-2v_2^2) + a_{31}(-2v_2v_3) & a'_{22} & a'_{23} \\ a_{21}(-2v_3v_2) + a_{31}(1-2v_3^2) & a'_{32} & a'_{33} \end{bmatrix}.$$

Dacă se analizează elementele din prima coloană a matricei A_2 obținute, se găsesc următoarele relații:

$$\begin{aligned} a'_{11} &= a_{11}, \\ a'_{21} &= a_{21} - 2v_2(a_{21}v_2 + a_{31}v_3) = a_{21} - 2v_2 S, \\ a'_{31} &= a_{31} - 2v_3(a_{21}v_2 + a_{31}v_3) = a_{31} - 2v_3 S. \end{aligned} \tag{6.101}$$

Fie p_1^2 suma elementelor din prima coloană și de sub diagonala principală, adică

$$p_1^2 = (a'_{21})^2 + (a'_{31})^2 = (a_{21} - 2v_2S) + (a_{31} - 2v_3S) = a_{21}^2 - a_{31}^2 - 4v_2a_{21}S -$$

$$-4v_3a_{31}S + 4v_2^2S^2 + 4v_3^2S^2 = a_{21}^2 + a_{31}^2 - 4S(a_{21}v_2 + a_{31}v_3) + 4S^2(v_2^2 + v_3^2) =$$

$$= a_{21}^2 + a_{31}^2 - 4S^2 + 4S^2 = a_{21}^2 + a_{31}^2.$$

Din această ultimă relație se vede că p_1^2 este același și pentru $A = A_1$ și pentru A_2, adică este un invariant în urma transformării Householder $H_2A_1H_2$.

Dacă în ecuațiile (6.101) se consideră $a'_{31} = 0$, atunci rezultă $a'_{11} = a_{11}$,

$a'_{21} = a_{21} - 2v_2S$, $0 = a_{31} - 2v_3S$ sau $a_{21} - 2v_2S = \pm p_1$, $a_{31} - 2v_3S = 0$.

Ultimele două ecuații permit calculul componentelor v_2 și v_3 iar $p_1 = \sqrt{a_{21}^2 + a_{31}^2}$, rezultînd

$$v_2 = \pm \sqrt{\frac{p_1 \pm p_2}{2p_1}}, \quad v_3 = \pm \frac{a_{31}}{2v_2p_1},$$

alegîndu-se ca semn semnul lui a_{21}.

În mod asemănător se determină componentele vectorului $\mathbf{v}^{(k)}$ pentru matricea $H_k = I - 2\mathbf{v}^k(\mathbf{v}^k)^T$, care intervine în transformarea similară $H_k A_{k-1} H_k$.

În final algoritmul Householder conduce la o matrice tridiagonală T, ale cărei valori proprii se determină ca în 6.7.3.

6.7.5. *Algoritmi pentru calculul vectorilor proprii*

La început se va prezenta un algoritm care permite calculul vectorilor proprii în cazul matricelor hermitiene.

Restrîngem discuția la o matrice simetrică reală $A \in \in M_{\mathrm{T}}^{n\times n}$ care a fost redusă la o matrice T tridiagonală prin metoda Givens sau Givens-Householder, adică

$$T = H^T A H, \; HH^T = 1, \text{ deci } A = HTH^T.$$

Dacă λ este o valoare proprie a matricei T și $\mathbf{y}$ vectorul propriu corespunzător, atunci se poate scrie ecuația

$$T\mathbf{y} = \lambda \mathbf{y}. \tag{6.103}$$

Matricea H introdusă de Householder are proprietatea $H = H^T = H^{-1}$, de unde rezultă că $H^T H = HH^T = I$.

Amplificînd relația (6.103) cu H la stînga, rezultă $HT\mathbf{y} = \lambda H\mathbf{y}$ sau $HTI\mathbf{y} = \lambda H\mathbf{y}$, respectiv $HTH^T H\mathbf{y} = \lambda H\mathbf{y}$. Ținînd seama de (6.102), ultima relație devine

$$A(P\mathbf{y}) = \lambda(P\mathbf{y}); \; A\mathbf{x} = \lambda\mathbf{x}; \; \mathbf{x} = P\mathbf{y},$$

deoarece A și T sînt similare și au aceleași valori proprii. Deci vectorul propriu al matricei A se poate calcula cu ajutorul matricei H și al vectorului propriu $\mathbf{y}$ al matricei T, corespunzător valorii proprii λ, atît pentru matricea A cît și pentru matricea T. Deci dacă se cunosc vectorii proprii $\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^n$ ai matricei T și matricea H, atunci

se pot calcula vectorii proprii $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ ai matricei A, cu ajutorul relației

$$\mathbf{x}^i = P\mathbf{y}^i, \quad i = 1, 2, \ldots, n.$$

Calculul vectorilor proprii ai matricei T (de formă tridiagonală) nu este dificil [42, 100].

În [94, 12] se indică folosirea metodei puterii pentru calculul vectorilor proprii ai matricei $T\mathbf{y}^i$, $i = 1, 2, \ldots, n$, aceasta urmărind o stabilitate a calculelor. Această metodă constă în alegerea unei valori α, care este o aproximare a lui λ cu $\lambda \neq \alpha$ și alegerea unui vector normalizat $\mathbf{y} \in R^n$, formîndu-se șirul

$$\begin{aligned}
&(T - \alpha I)\mathbf{y}^1 = \mathbf{y}^0, \quad \mathbf{z}^1 = \frac{1}{m_1}\mathbf{y}^1, \\
&(T - \alpha I)\mathbf{y}^2 = \mathbf{z}^1, \quad \mathbf{z}^2 = \frac{1}{m_2}\mathbf{y}^2, \qquad (6.104) \\
&\cdots\cdots\cdots\cdots\cdots \\
&(T - \alpha I)\mathbf{y}^k = \mathbf{z}^{k-1}, \quad \mathbf{z}^k = \frac{1}{m_k}\mathbf{y}^k.
\end{aligned}$$

În cadrul procesului iterativ, factorul de normalizare este componenta de modul maxim a vectorului $\mathbf{y}$, adică $|m_k| = \max_i |y_i^k|$. Alegînd α și $\mathbf{y}^0$ corespunzătoare, șirul $\mathbf{z}^k$ converge către $\mathbf{y}$ ($\mathbf{z}^k \to \mathbf{y}$). Dacă $|\lambda - \alpha|$ este foarte apropiată de $|\lambda_i - \alpha|$ pentru un anumit i, sau $\mathbf{y}^0$ are o componentă foarte mică în direcția unui vector propriu $\mathbf{y}^i$, atunci șirul (6.104) converge către vectorul propriu corespunzător, pentru $k = 2$.

Șirul (6.104) implică rezolvarea unui sistem de ecuații algebrice pentru fiecare $k = 1, 2, \ldots, n$, dar avînd în vedere structura matricei T, calculele sînt extrem de simple.

În cazul matricelor nehermitiene, se consideră o matrice Hessenberg H, obținută din matricea $A \in M_{\Gamma}^{m \times n}$ prin transformări similare, adică $H = U^H A U$. Atunci se pot scrie relațiile

$$A = UHU^H, \quad UU^H = I = U^H U. \qquad (6.105)$$

Dacă λ este valoarea proprie a matricei Hessenberg H, obținută din A prin transformări similare, iar $\mathbf{y}$ este vectorul propriu corespunzător, atunci se poate scrie relația

$$H\mathbf{y} = \lambda \mathbf{y}. \tag{6.106}$$

Dacă se amplifică relația (6.106) cu U la stînga, rezultă

$$UH\mathbf{y} = \lambda U\mathbf{y} \rightarrow UHU^H U\mathbf{y} = \lambda U\mathbf{y}.$$

Folosind prima relație din (6.105) în ultima relație din (6.106), se obține

$$AU\mathbf{y} = \lambda U\mathbf{y}, \text{ deci } A\mathbf{x} = \lambda\mathbf{x}, \text{ de unde } \mathbf{x} = U\mathbf{y}.$$

Deci valoarea proprie λ a matricei A are ca vector propriu pe $\mathbf{x} = U\mathbf{y}$. Dacă $\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^n$ sînt vectori proprii ai matricei H similare cu A, atunci vectorii proprii ai matricei A se obțin cu ajutorul relației

$$\mathbf{x}^i = U\mathbf{y}^i, \; i = 1, 2, \ldots, n.$$

Determinarea vectorilor proprii $\mathbf{y}^i$, $i = 1, 2, \ldots, n$, ai matricei Hessenberg H (similară cu matricea A) se poate realiza folosind metoda puterii inverse. Fie β o aproximație a valorii proprii λ cu $\lambda \neq \beta$ și $\mathbf{y}^0 \in C$. Atunci se poate forma șirul

$$\begin{aligned}
&(H - \beta I)\mathbf{y}^1 = \mathbf{y}^0, \quad \mathbf{z}^1 = \frac{1}{m_1}\mathbf{y}^1, \\
&(H - \beta I)\mathbf{y}^2 = \mathbf{z}^1, \quad \mathbf{z}^2 = \frac{1}{m_2}\mathbf{y}^2, \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
&(H - \beta I)\mathbf{y}^k = \mathbf{z}^{k-1}, \quad \mathbf{z}^k = \frac{1}{m_k}\mathbf{y}^k,
\end{aligned} \tag{6.107}$$

unde m este componenta de modul maxim a vectorului $\mathbf{y}$.

Dacă vectorul $\mathbf{y}^0$ are o componentă maximă în direcția lui $\mathbf{y}$, atunci $\mathbf{z}^k \rightarrow \mathbf{y}$.

Rezolvarea sistemelor (6.107) nu prezintă dificultăți deoarece H este o matrice Hessenberg.

O matrice are vectori proprii la stînga și vectori proprii la dreapta [42], adică

$$(\mathbf{y}^j)^T A = \lambda_j (\mathbf{y}^j)^T, \quad j = 1, 2, \ldots, n, \tag{6.108}$$

și dacă se transpune relația (6.108), se obține

$$A^T \mathbf{y}^j = \lambda_j \mathbf{y}^j, \; j = 1, 2, \ldots, n. \tag{6.109}$$

Deci vectorii proprii la stînga ai matricei A coincid cu vectorii proprii la dreapta ai matricei A^T.

Teoremă. *Dacă* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ *sînt vectorii proprii la dreapta ai matricei* A *și* $\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^n$ *sînt vectori proprii la stînga ai matricei* A *pentru* $A \in M_\Gamma^{n \times n}$, *atunci*

$$(\mathbf{y}^j)^T \mathbf{x}^i = 0, \quad \lambda_i \neq \lambda_j.$$

Demonstrație. Dacă $\mathbf{x}^i$ sînt vectori proprii la stînga ai matricei A, atunci are loc relația

$$A\mathbf{x}^i = \lambda_i \mathbf{x}^i, \quad i = 1, 2, \ldots, n. \tag{6.110}$$

Dacă se înmulțește relația (6.108) cu $\mathbf{x}^i$ la dreapta și relația (6.110) cu $(\mathbf{y}^i)^T$, se obțin următoarele două relații:

$$(\mathbf{y}^i)^T A \mathbf{x}^i = \lambda_j (\mathbf{y}^j)^T \mathbf{x}^i, \quad (\mathbf{y}^j)^T A \mathbf{x}^i = (\mathbf{y}^j)^T \lambda_i \mathbf{x}^i,$$

care după scădere conduc la $0 = (\lambda_i - \lambda_j)(\mathbf{y}^j)^T \mathbf{x}^i$, de unde se vede că

$$(\mathbf{y}^j)^T \mathbf{x}^i = \begin{cases} 0 \text{ pentru } i \neq j, \; \lambda_i = \lambda_j, \\ 1 \text{ pentru } i = j, \; \lambda_i \neq \lambda_j. \end{cases}$$

La calculul valorilor și vectorilor proprii ai unei matrice este necesară o analiză în următoarele direcții: dacă problema este corect pusă sau nu, dacă este bine condiționată sau nu, dacă este stabilă sau nu, precum și problema convergenței calculelor și propagarea erorilor [42, 59, 86].

1. Arden, W. B., Astill, N. K. *Numerical algorithms origins and applications.* California, Addison-Wesley, 1969,
2. Ames, W. F. *Nonlinear partial differential equations in engineering.* New York, Academie Press, 1965.
3. Azis, K. A., Mayers, M. A. *Periodic solutions of hyperbolic partial differential equations in a strip.* Trans. Amer., Math. Soc., 1969.
4. Ahlberg, J. H., E. N. Nilson, J. L. Walsh. *The theory of splines and their applications.* New York, Academic Press, 1967.
5. Allen, D. N. de G. *Relaxation Methodes.* New York, McGraw-Hill Book, 1954.
6. Antosiewicz, H. A., Gantschi, W. *Numerical methods in ordinary differential equytions.* New York, McGraw-Hill Book., 1962.
7. Adams, Duane, A. *A stopping criterion of polynomial root finding.* Comm. Assoc. Comput. Mach., **10**, 1967.
8. Ablow, M. G. *A characteristic finit difference methodes for the Wave* equations in two dimensions. SIAM I. Numer Anal, vol. 9, nr. 1, 1972.
9. Bauer, F. L., Fike, C. T. *Norms and exlusion theorems.* Numer Math. 2, 1970.
10. Bellman R. *Introductions to matrix analysis.* New York, McGraw--Hill, 1960.
11. Birkhoff, G., Richards, S. Varga, David Young. *Alternating direction implicit methods.* Advances in computer, vol. 3, New York, Academie Press, 1962.
12. Bronson R. *Matrix methods, an introduction,* New York, Academic Press, 1969.
13. Businger, P. A. *Monitoring the numerical stability of Gaussion elimination,* Numer. Math., **16**, 1971.
14. Berezin, I. S. *Metodi vicisleniia.* Vol. I, II, Moscova,
15. Bruce, A. *Numercial algoritms origins and application.* Mass., Addison Wesley, 1970.
16. Babuska, H. P., Vitasek, E. *Numerical processes in differential equations,* New York, John Wiley, 1966.
17. Bruce, C., Steplеman, R. *A general theory of convergence for numerical methods.* SIAM J. Numer. Anal., vol. 9., nr. 3, 1972.
18. Buchanan, M. *A necessary and sufficient condition for stability of difference schemes for initial value problems,* SIAM J. vol. 4, 1972.
19. Bucur, C. M. *Metode numerice.* Timișoara, Editura Facla, 1973.
20. Berstein, D. L. *Existente theoreme in differential equations.* Princenton Univ. Press, 1960.
21. Beckett, R., Hurt, T. *Numerical calculations and algorithms.* New York, McGrow-Hill Book, Co. 1960.
22. Chartress, B. A. *Automatic controled precision calculations.* J. Assoc. Comput. Math., **13**, 1966.
23. Cherney, E. W. *Introduction to approximation theory.* New York, McGraw-Hill Book, Co., 1966.
24. Cody, W. J. *The influence of machine design on numerical algorithms.* Procedings Spring Joint Computer Conference, 1967.
25. Collatz, L. *Functional analysis and numerical mathematics.* New York, Academic Press, 1966.
26. Courant, R., Hilbert, D. *Methods of mathematical physic.* Vol. II, New York, Interscience Publisher, 1962.

27. Crank, J., Nicolson, P. *A practical method for numerical evaluations of partial differential equations of that conductional.* Proc. Comb. Phil., Soc., **4**, 1947.
28. Creţu, I. *Hidraulica generală în subterană.* Bucureşti Editura didactică şi pedagogică, 1971.
29. Cuculescu, I. *Analiză numerică.* Bucureşti, Editura tehnică, 1967.
30. Carnahan, B., Wilkes, J. *Digital computing and numerical methods.* New York, McGrow-Hill, 1970.
31. Cristescu, R. *Analiză funcţională.* Bucureşti, Editura didactică şi pedagogică, 1965.
32. Deleeuw, L. S., Southworth, W. R. *Digital Computation and Numercial methods.* New York, McGrow-Hill, 1965.
33. Dancea, I. *Programarea calculatoarelor.* Cluj, Editura Dacia, 1973.
34. Demidovici, B., Marcu, I. *Elements de calcul numérique.* Moscova, Mir, 1973.
35. Dimo, P. *Programarea în* FORTRAN. Bucureşti, Editura didactică şi pedagogică, 1971.
36. Diakonov, E. *On the application of disintegrating difference operators.* J. Vicisl. mat. i matem. fizika, **3**, 1963.
37. Dodescu Gh, Racoveanu, N. *Metode de calcul numeric,* Bucureşti, ASE, 1970.
38. Dumitrescu, I., Dodescu, Gh. *Procedee pentru determinarea eforturilor din instalaţiile de foraj ale ţiţeiului în regim dinamic prin utilizarea sistemelor de calcul automat.* Studii şi cercetări de mecanică aplicată, vol. 23, 1974.
39. Dodescu Gh. ş.a. *Minicalculatoare şi aplicaţii.* Vol. 1,2, Bucureşti, Editura tehnică, 1977.
40. Dodescu Gh. ş.a. *Calculatoare electronice şi sisteme de operare.* Bucureşti, Editura didactică şi pedagogică, 1975.
41. Dodescu, Gh. ş.a. *Limbajul* BASIC *şi aplicaţii.* Bucureşti, Editura didactică şi pedagogică, 1978.
42. Dodescu, Gh., Toma, M. *Metode de calcul numeric.* Bucureşti, Editura didactică şi pedagogică, 1976.
43. Ehrlich, L. W. *The block symmetric overreloxation method.* SIAM J., 12, 1968.
44. Evans, W. G., Wallance, F. G. *Simulation using digital computers.* New Jersey, Prentice-Hall, 1967.
45. Forsythe, G. E., Wasow, W. R. *Finite difference methods for partial differential equations.* New York, John Wiley Sons, Inc., 1960.
46. Fox, L., Mayer, D. F. *Computing methodes for scientist and engineering.* Oxford, Clarendon-Press, 1968.
47. Franklin, J. N. *Matrix theory,* New York Printice-Hall, 1968.
48. Forsythe, E. G., Moler B. Cleve. *Computer solution of linear algebraie Systems,* New York, Printice-Hall, 1967.
49. Florence, J. O. *Matrix inversion by Monte Carlo methods.* Math. Methods for digital computers, New York, Wiley, 1960.
50. Fike, T. C. *Computer Evaluation of mathematical Functions.* New Jersey, Prentice-Hall, 1967.
51. Givens, Wallace. *A method of computing eigenvalues and eigenvertors suggested by classical results on symmetric matrices,* Appl, Math. Series, nr. 29, 1953.

52. Golub, Gene H. *The use of Ghebishev matrix palynomials in the interative solution of linear systems compared with the method successive averreloxation.* Doctoral thesis, University of Illinois, 1959.
53. Gregory, R. T. *Computing eigenvalue and eigenvectors of a simmetric matrix on the* ILLIAC. Math. of Comp., 7, 1953.
54. Golube, G. *Use of fort direct methods for the efficient numerical solution of nonseparable eliptic equations.* Rep. STANFORD CS-72, 1972.
55. Gregory, T. R. *The use of residue aritmetic with automatic digital computers,* CNA-27, Texas, 1971.
56. Ghika, A. *Analiză funcțională.* București, Editura Academiei, 1967.
57. Henrici, Peter. *Theoretical and experimental studies on the accumulation of error in the numerical solutions of initial — value problems for systems of ordinary differential equations.* Proc. of the International Conference on Information Procesing, Paris, 1959.
58. Henrici, P. *On the speed of convergence of cyclic and quasiciclic Jacobi methods for computing eingenvalue of Hermitian matrices.* SIAM J., 6,1958.
59. Hildebrand, F. B. *Introduction to numerical analysis.* New York, McGraw-Hill, Book Co., Inc., 1956.
60. Householder, A. S. *Principle of numerical analysis.* New York, McGraw-Hill Book, 10, Inc., 1953.
61. Householder, A. S. *The approximate solution of matrix problems.* J. Assoc. Comput, 5, 1958.
62. Householder, A. S. *Unitary triangularization of a nonsymmetric matrix.* J. Assoc. Comput, 5, 1958.
63. Householder, A. S. *The numerical treatement of a simple nonlinear equation.* New York, McGraw-Hill Book Co., 1970.
64. Howell, J. A., Gregory, R. T. *Solving systems of linear algebric equations using residue arithmetic.* Report TNN—82 Austin, 1969.
65. Halanay, A. ș.a. *Teoria calitativă a sistemelor cu impulsuri.* București, Editura Academiei, 1968.
66. Herbert, B. K. *The numerical solution of parabolic partial differential equations.* Mathematical methods for digital computers, New York, J. Wiley, 1960.
67. Isaacson, E., Herbert, B. K. *Analysis of numerical methods.* New York, J. Wiley, 1966.
68. Iankó, B. *Rezolvarea numerică a sistemelor de ecuații liniare,* București, Editura Academiei, 1961.
69. Ionescu, D. V. *Cuadraturi numerice.* București, Editur atehnică,1957.
70. Ionescu D. V. *Ecuații diferențiale ordinare cu derivate parțiale.* București, Editura didactică și pedagogică, 1965.
71. Kuo, S. S. *Computer applications of numerical methods.* Menlo Park, California, Addison-Wesley, 1971.
72. Keast, P., Mitchell, A. R. *Finite difference solution of the third boundary problem in elliptic and parabolic equations.* Numer. Math., **10**, 1967.
73. Lees, M. *A linear three liveles difference schemes for quasilinear parabolic equations.* Math. of computations, **20**, 1966.
74. Lax, P. D. *Differentiall equations difference equation and matrix theory* Comm. Pure, Appl. Math., vol. 11, 1958.
75. Legras, J. *Méthodes et techniques de l'analyse numérique.* Paris, Dunod, 1971.

76. Martin, R. S., Reinsch, C., Wilkinson, J. H. *Householder's tridiagonalization of a symmetric matrix.* Numer. Math., 11, 1968.
77. Mitchell, A. R. *Computational method in partial differential equations.* New York, Wiley, 1971.
78. Marlin, D. *Iterative methods for linear equations with symmetric positive definite matrix.* Comp., Journal, **4**, 1961.
79. McCracken, D. D., Dorn W. *Metode de calcul și programe, în* FORTRAN (trad. din l. engl.). București Editura tehnică, 1976.
80. Morton, K. W. *The design of difference schemes for studying physical instabilities.* New York, Wiley, 1974.
81. Marinescu, G. *Analiză numerică.* București, Editura Academiei, 1974.
82. Moszynski, K. *Metode numerice de rezolvare a ecuațiilor diferențiale ordinare,* București, Editura tehnică, 1973.
83. Marinescu, G. *Tratat de analiză funcțională.* Vol. I, II, București, Editura Academiei, 1972.
84. Niculescu, S. *Inițiere în* FORTRAN. București, Editura tehnică, 1972.
85. Nicolescu, M. ș.a. *Analiza matematică.* Vol. II, București, Editura didactică și pedagogică, 1963.
86. Ortega, J. M. *Numerical analysis.* New York, Academic Press, 1973.
87. Osborne, M. R. *The numerical solution of a periodic problem subject to a nonlinear boundary condition.* Numer. Math, **10**, 1967.
88. Oroveanu, T. *Mecanica fluidelor viscoase.* București, Editura Academiei, 1967.
89. Oroveanu, T. *Hidraulica și transportul produselor petroliere.* București, Editura didactică și pedagogică, 1966.
90. Ortega J. M., Werner, C. R. *Iterative solution of nonlinear equations in several variables.* New York, Academic Press, 1970.
91. Parlett, B. N. *Singular and invariant matrices under the QR transformation.* Math. of. Comp., **20**, 1966.
92. Parlett, B. N. *Global convergence of the basic QR algorithm on Hessenberg matrices.* Math. of. Comp., **22**, 1968.
93. Poole, W. G. *A geometric convergence theory for the QR, Rayleigh quotient and power iterations.* Comp. Center. Tehn., Report, nr. 41, Berkeley, 1970.
94. Pipes, L. *Matrix methods for engineering.* Pretntice-Hall, Englewood Cliffs, 1963.
95. Ping Chun Vang. *Metode numerice și matriceale în mecanica construcțiloe,* (trad. din l. engl.), București, Editura tehnică, 1970.
96. Popoviciu, E. *Analiză numerică.* Vol. I, Cluj, Editura Dacia, 1967.
97. Richtmyr, R. D., Morton, K. W. *Difference methods for initial-value problems.* New York, Interscience Publishers, 1967.
98. Racoveanu, N. ș.a. *Metode numerice pentru ecuații cu derivate parțiale de tip hiperbolic.* București, Editura tehnică, 1976.
99. Racoveanu N., Dodescu, Gh. *Metode de calcul numeric.* Note de curs pentru uzul studenților din ASE, București.
100. Rubinstein, F. M. *Matrix computers analysis of structures.* New York, Prentice-Hall, 1966.
101. Reddi, M. *Finite element solution of the incompressible lubrification problem.* Transactions of ASME, 1969.
102. Roșculeț, M. N. *Analiză matematică.* Vol. I, II, București, Editura didactică și pedagogică, 1966.
103. Richards, S. V. *Matrix iterative analysis.* New York Prentice-Hall, 1965.

104. Rice, J. R. *Approximation des functions, Théorie lineaire.* Paris, Dunod, 1969.
105. Stewart, G. W. *Incorporating origin shifts into the QR algorithm for symmetric tridiagonal matrices.* Comm. Assoc. Comput., **13**, 1970.
106. Starang, G. *The finite element method and approximation theory Numerical solution of partial differential equations,* New York, Academic Press, 1970.
107. Stoutemeyr, R. D. *PL/I programming for engineering and science* New Jersey, Prentice-Hall, 1972.
108. Stewart, G. W. *Introduction to motrix computation.* New York, Academic Press, 1973.
109. Şchiop, I. A. *Metode aproximative în analiza neliniară.* Bucureşti, Editura Academiei, 1972.
110. Shan, S. Kuo. *Computer applications of numerical methods.* Menlo Park, Prentice-Hall, 1973.
111. Şabac, I. G. *Matematici speciale.* Vol. I,II,Bucureşti, Editura didactică şi pedagogică, 1960.
112. Traub, J. F. *Iterative methods for the solution of equations.* New York, Prentice-Hall, 1965.
113. Tropper, A. M. *An introduction to linear algebra.* New York, Elsevier, 1969.
114. Tihonov, N., Samarski, A. *Odnorodnîe raznostnîe shemi.* În: *vîcisl. matem. i matem. fizica,* **1**, 1963.
115. Teodorescu, N., Olariu, V. *Ecuaţiile fizicii matematice.* Vol. I, II, Bucureşti, Editura didactică şi pedagogică, 1965.
116. Teodorescu, N. *Introducere în fizica matematică.* Bucureşti, Editura tehnică, 1970.
117. Tucker, T. S. *Stability of nonlinear computing schemes.* SIAM, J. Num. Anal., 6, 1969.
118. Varga, R.S. *Factorization and normalization iterative method, Boundary problems in differential equations,* New York Wisconsin Press, 1964.
119. Varga, R. S. *Matrix iterative analysis.* New York, Prentice-Hall, 1962.
120. Varga, R. S. *Functional analysis and approximation theory in numerical analysis.* SIAM, Conference, 1971.
121. Warlick, C. H., David, M. Y. *A priori methods for determination of the optimum relaxation factor for the succesive overrelaxation method* Austin University, 1970.
122. Wilkinson, J. H. *The solution of ill-conditioned linear equations,* Ralstom and Wilf, 1967.
123. Young, D. M. *Second-degree iterative methods for solution of large linear systems.* Jour. of. Approx theory, 1971.
124. Young, D. M. *A bound for the optimum relaxation factor for the succesive overrelaxation method.* Numer, Math., 16, 1971.
125. Young, D. M. *Norms of the succesive overrelaxation method and method.* Austin, University 1969.
126. Young, D. M. *Iterative methods for solving partial differential equations of eliptic type.* Trans. Amer. Teath, 1954.
127. Young, D. M. *A survey of modern numerical analysis.* SIAM, Review, 15, 1973.
128. Young, D. *A survey of numerical mathematics,* Vol. I, II,, New York, Addison Wesley, 1973.