Metode numerice (notiţe curs)

– UPB 2020

I. Introducere

Determinarea practică a soluţiilor matematice al unor probleme ştiinţifice şi inginereşti necesită, în cele mai multe cazuri, construirea unui algoritm pentru găsirea lor, adaptat posibilităţilor de lucru ale sistemului de calcul numeric.

Astfel, **algoritmul** reprezintă totalitatea raţionamentelor matematice şi logice, efectuate pas cu pas pentru rezolvarea problemei, deci o reţetă generală, precis formulată, care conduce la rezultate concrete într-un interval finit de timp.

Formularea matematică (modelul) a unei probleme fizice sau inginereşti conţine inevitabil aproximări datorate:

- înţelegerii incomplete a fenomenelor naturale sau a naturii aleatoare ale acestora;
- descrierii matematice exclusive a trăsăturilor esenţiale ale procesului fizic, prin neglijarea deliberată a detaliilor legate de efectele secundare

Erorilor asociate acestor aproximări li se adaugă cele rezultate din prelucrarea datelor cu ajutorul unui sistem de calcul numeric. Acesta poate realiza numai un număr limitat de operaţii aritmetice simple (adunare, scădere, împărţire, înmulţire, ridicare la putere) cu numere finite raţionale. Operaţii importante ca derivarea, integrarea, evaluarea seriilor infinite nu pot fi implementate direct pe un calculator numeric. Acesta este compus din registre de calcul şi memorii cu capacitate limitată, încât este imposibil să se reprezinte cantităţile infinit mici sau infinit mari, ori un interval finit de numere reale.

Un algoritm adaptat posibilităţilor calculatorului numeric, utilizând numai operaţii aritmetice şi unele operaţii logice se spune că utilizează **Metode Numerice**.

Lucrarea tratează, în continuare, prin această prismă, principalele capitole matematice utilizate în inginerie.

II. Erori

Eroarea absolută, e_x, este definită ca fiind diferenţa dintre valoarea adevărată (exactă), x şi valoarea aproximativă , $\bar{x}$ (determinată prin calcul sau prin măsurare):

$$e_x = x - \bar{x} \qquad \text{(II.1)}$$

Dintre cele 3 mărimi numai valoarea aproximativă este cunoscută de obicei, uneori cunoscându-se şi marginea erorii (modulul maxim probabil).

Eroarea relativă este eroarea absolută împărţită la valoarea aproximativă (prin convenţie), pentru că aceasta se cunoaşte în general.

$$\varepsilon_x = e_x / \bar{x} \qquad \text{(II.2)}$$

II.1 Erori iniţiale

Erorile iniţiale sunt erori din datele iniţiale, cauzate de incertitudinea măsurătorilor, din greşeli (grosolane) sau din reprezentarea unei valori transcendente sau iraţionale printr-un număr finit de cifre.

Exemple:

a) o măsurătoare fizică nu poate fi exactă; instrumentul folosit transferă precizia sa mărimii măsurate; în cazul dimensiunilor liniare, un şubler indică cu certitudine 0,1 mm, un micrometru sau un comparator 0,01 mm, iar un traductor capacitiv ajunge la o rezoluţie nanometrică. Numărul de cifre semnificative corecte variază deci în funcţie de precizia măsurătorii, recomandându-se limitelor toleranţei mărimii.
b) multe numere nu pot fi reprezentate printr-un număr finit de cifre. Dacă exemplele lui π sau $\sqrt{2}$ sunt evidente, o simplă fracţie poate să nu aibă o reprezentare exactă, în funcţie de baza de numeraţie folosită. În sistemul zecimal, numărul 1/10 este 0,1, în timp ce în sistemul binar el este 0,0001100110011……., deci o reprezentare infinită imposibilă în calculator. Deci, dacă se face suma a 10 numere reprezentând aproximaţii binare ale lui 0,1, nu se va obţine 1,0.

II.2 Erori de trunchiere

O procedură numerică poate introduce ea însăşi erori. De exemplu, pentru calculul sinusului unui unghi (în radiani) se poate utiliza seria Maclaurin:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

Aceasta este infinită şi nu se pot utiliza toţi termenii seriei în calcul. Limitarea expresiei la un număr finit de termeni conduce la o eroare reprezentată de termenii omişi, numită eroare de trunchiere.

Deşi nu se poate calcula precis, eroarea de trunchiere se poate estima prin metode specifice fiecărui algoritm, limitele acesteia fiind puse în evidenţă la fiecare problemă analizată.

II.3 Erori de calcul

II.3.1. Calculul în virgulă mobilă

La reprezentarea numerelor în dispozitive cu capacitate limitată de memorare şi prelucrare este specifică tăierea cifrelor nesemnificative de la sfârşitul acestora, care poate apărea atât la introducerea numărului, cât şi în urma efectuării unor operaţii aritmetice. În unele cazuri particulare, eroarea rezultată prin tăiere poate avea valori importante.

Rezolvarea limitării tehnice a reprezentării numerelor reale în calculator o constituie utilizarea numerelor în virgulă mobilă. Acestea sunt compuse dintr-o fracţie numită mantisă şi un întreg numit exponent sau caracteristică.

$$N=m\ b^c \qquad \text{(II.3)}$$

Unde: m = mantisă, c = exponent, b = baza de numeraţie.

Exemplu: În baza 10, numărul N=36,82 se va rescrie $N=0,3682\cdot10^{-2}$. El este normalizat pentru că prima cifră după virgulă a mantisei este diferită de zero, adică:

$$\frac{1}{10} \le |m| < 1 \qquad \text{(II.4)}$$

Procedeul de adunare a două numere reale în virgulă mobilă va fi ilustrat pe un calculator ipotetic, zecimal cu patru poziţii afectate mantisei pentru numerele a) $x_1=165,2$; $x_2= 0,21$ şi respectiv b) $x_3=26,31$; $x_4=-19,76$.

Etapele procedurii sunt:

1) normalizarea numerelor:

a) $x_1=0,1652\cdot10^3$; $x_2=0,2100\cdot10^0$

b) $x_3=0,2631\cdot10^2$; $x_4=-0,1976\cdot10^2$

2) compararea exponenţilor:

a) $c_1=3 > c_2=0$

b) $c_3=c_4=2$

3) de-normalizarea numărului cu exponent mai mic prin deplasarea mantisei spre dreapta virgulei, cu atâtea poziţii cât reprezintă diferenţa exponenţilor:

a) $x_2=0,2100\cdot10^0= 0,0002\cdot10^3$

Prin de-normalizare se pot pierde cifre semnificative ale numărului.

4) adunare propriu-zisă

a) $S_{12}=x_1+x_2=0,1652\cdot10^3+0,0002\cdot10^3=0,1654\cdot10^3$

b) $S_{34}=x_3+x_4=0,2632\cdot10^2+(-0,1976)\cdot10^2=0,0655\cdot10^2$

5) re-normalizarea rezultatului (dacă este cazul)

b) $S_{34}=0,6550\cdot10^1$

În exemplul a, suma este afectată de tăierea efectuată la pasul 3, iar în exemplul b ultima cifră (0) nu mai are nici o semnificaţie, ea apărând în procesul deplasării. Aceste erori se propagă în următoarele operaţii, lor adăugându-li-se altele, specifice fiecărei etape. Tăierea ultimei cifre este evitată la unele sisteme de calcul printr-o poziţie pentru aşa-zisa cifră de gardă (în exemplul dat o a cincea cifră în mantisa calculatorului ipotetic), permiţând salvarea unei cifre din mantisa deplasată spre dreapta.

Probleme deosebite la calculul în virgulă mobilă mai pot apare în următoarele cazuri:

a) scăderea a două numere aproximativ egale cu exponent minim conduce la depăşire inferioară în virgulă mobilă (*underflow*).

Exemplu: convenindu-se că exponentul minim reprezentabil în calculatorul ipotetic este -615, atunci: $0,1459\cdot10^{-615}-0,1024\cdot10^{-615}=0,0425\cdot10^{-615}=0,4250\cdot10^{-616}$.

Diverse sisteme de calcul rezolvă diferit această problemă, dar, de obicei, atribuie rezultatului valoarea zero în virgulă mobilă, ceea ce poate conduce la rezultate foarte ciudate cu un algoritm aparent corect.

b) adunarea a două numere cu exponent maxim poate conduce la depăşire superioară în virgulă mobilă (*overflow*).

Exemplu: Dacă exponentul maxi reprezentabil în calculatorul ipotetic este 615, atunci $0,9999\cdot10^{615}+0,9999\cdot10^{615}=1,9998\cdot10^{615}=0,1999\cdot10^{616}$

De obicei, depăşirea exponentului maxim reprezentabil în calculator este semnalată prin întreruperea execuţiei programului.

La înmulţire/împărţire mantisele normalizate se înmulţesc/împart, iar exponenţii se adună/scad, normalizarea rezultatului fiind simplă. Depăşirea superioară poate apare atât la înmulţire, dar şi la împărţire prin zero (când împărţitorul poate fi rezultatul unei depăşiri inferioare).

În mod normal, calculatoarele numerice lucrează în baza 2 sau 16, încât reprezentarea numerelor în virgulă mobilă este:

$x=m\cdot2^c$ cu $\frac{1}{2}\le|m|<1$, sau (II.5)

$x=m\cdot16^c$ cu $\frac{1}{16}\le|m|<1$. (II.6)

II.3.2. Erori de rotunjire

Rezultatul oricăreia dintre cele 4 operaţii, într-o reprezentare a numerelor cu t cifre semnificative pentru mantisă, poate fi scris ca:

$x = m_x \cdot 10^c + p_x \cdot 10^{c-t}$,

unde p_x este partea din mantisă asupra căreia se acţionează la rotunjire, care nu este normalizată, deci $0 \le p_x < 1$.

Unele compilatoare pentru limbaje de nivel înalt folosesc tăierea termenului $p_x \cdot 10^{c-t}$, atribuind rezultatului valoarea $m_x \cdot 10^c$.

În acest caz, modulul erorii de rotunjire maxime este:

$$\frac{e_x}{\bar{x}} = \left| \frac{p_x \cdot 10^{c-t}}{m_x \cdot 10^c} \right| \le \frac{1 \cdot 10^{c-t}}{0{,}1 \cdot 10^c} = 10^{-t+1} \qquad \text{(II.7)}$$

În consecinţă, maximul erorii relative de rotunjire prin tăiere nu depinde de mărimea numărului, ci de numărul de cifre semnificative reprezentate în mantisă (t).

Un alt tip de rotunjire uzuală este cea simetrică, aproximarea numărului x, fiind:

$$|\bar{x}| = \begin{cases} |m_x| \cdot 10^c, daca |p_x| < \frac{1}{2} \\ |m_x| \cdot 10^c + 10^{c-t}, daca |p_x| \ge \frac{1}{2} \end{cases} \qquad \text{(II.8)}$$

unde $\bar{x}$ şi m_x sunt de acelaşi semn.

Eroarea absolută maximă este:

$$|e_x| = \begin{cases} |p_x| \cdot 10^{c-t} \le \frac{1}{2} \cdot 10^{c-t}, \; daca |p_x| < \frac{1}{2} \\ |1 - p_x| \cdot 10^{c-t} \le \frac{1}{2} \cdot 10^{c-t}, \; daca |p_x| \ge \frac{1}{2} \end{cases}$$

Valoarea absolută a erorii relative maxime de rotunjire simetrică este:

$$\left| \frac{e_x}{\bar{x}} \right| \le \left| \frac{\frac{1}{2} \cdot 10^{c-t}}{m_x \cdot 10^c} \right| \le \frac{\frac{1}{2} \cdot 10^{c-t}}{0{,}1 \cdot 10^c} = 5 \cdot 10^{-t} \;,$$

deci cel mult jumătate din cea obţinută la rotunjirea prin tăiere.

II.4. Propagarea erorilor

S-a constatat că, în cazuri extreme, precum scăderea a două numere foarte mici, eroarea relativă a diferenţei poate fi destul de mare, propagându-se prin orice operaţie aritmetică ulterioară.

De aceea expresiile erorilor relative şi absolute ale rezultatului unei operaţii în funcţie de erorile operanzilor sunt foarte importante. Notându-se $x = \bar{x} + e_x$ şi $y = \overline{y} + e_y$, se determină erorile absolute pentru fiecare operaţie:

a) adunarea (scăderea)

$$e_{x\pm y} = (x \pm y) - (\bar{x} \pm \bar{y}) = e_x \pm e_y \qquad \text{(II.9)}$$

b) înmulţirea

$$e_{x\cdot y} = x \cdot y - \bar{x} \cdot \bar{y} = (\bar{x} + e_x) \cdot (\bar{y} + e_y) - \bar{x} \cdot \bar{y} = \bar{x} e_y + \bar{y} e_x + e_x \cdot e_y \cong \bar{x} e_y + \bar{y} e_x \quad \text{(II.10)}$$

c) împărţirea

$$e_{x/y} = x/y - \bar{x}/\bar{y} = (\bar{x} + e_x)/(\bar{y} + e_y) - \bar{x}/\bar{y} = \frac{\bar{x} + e_x}{\bar{y}} \cdot \frac{1}{1 + \frac{e_y}{\bar{y}}} - \frac{\bar{x}}{\bar{y}} =$$

$$= \frac{\bar{x} + e_x}{\bar{y}} \cdot \left[1 - \frac{e_y}{\bar{y}} + \left(\frac{e_y}{\bar{y}}\right)^2 - \ldots\ldots\right] - \frac{\bar{x}}{\bar{y}} \cong \frac{e_x}{\bar{y}} - \frac{\bar{x}}{\bar{y}^2} \cdot e_y \qquad \text{(II.11)}$$

Relaţiile deduse specifică valoarea maximă posibilă a erorii rezultatului operaţiei respective. Cum semnul erorilor operanzilor este rareori cunoscut, rezultă că nu întotdeauna adunarea măreşte eroarea, iar scăderea o micşorează aşa cum s-ar putea deduce la o analiză superficială a expresiilor.

Erorile relative derivă din cele absolute astfel:

a) adunare

$$\frac{e_{x+y}}{\bar{x} + \bar{y}} = \frac{\bar{x}}{\bar{x} + \bar{y}} \cdot \frac{e_x}{\bar{x}} + \frac{\bar{y}}{\bar{x} + \bar{y}} \cdot \frac{e_y}{\bar{y}} \qquad \text{(II.12)}$$

b) scădere

$$\frac{e_{x-y}}{\bar{x} - \bar{y}} = \frac{\bar{x}}{\bar{x} - \bar{y}} \cdot \frac{e_x}{\bar{x}} - \frac{\bar{y}}{\bar{x} - \bar{y}} \cdot \frac{e_y}{\bar{y}} \qquad \text{(II.13)}$$

c) înmulţire

$$\frac{e_{x\cdot y}}{\bar{x} \cdot \bar{y}} = \frac{e_x}{\bar{x}} + \frac{e_y}{\bar{y}} \qquad \text{(II.14)}$$

d) împărţire

$$\frac{e_{x/y}}{\bar{x}/\bar{y}} = \frac{e_x}{\bar{x}} - \frac{e_y}{\bar{y}} \qquad \text{(II.15)}$$

În formulele de propagare erorile pot fi de orice tip, termenii x şi y putând fi obţinuţi în urma unor măsurători (cu erori iniţiale) sau a unor calcule anterioare (cu erori de trunchiere, tăiere sau rotunjire). Chiar şi în cazul operaţiilor cu numere lipsite de erori rezultatul va fi afectat de rotunjire.

Exemplu: u = (x+y)·z, unde x,y,z nu au erori.

În urma efectuării adunării va apare o eroare de rotunjire simetrică:

$$\left|\frac{e_{x+y}}{\bar{x}+\bar{y}}\right| \le \frac{1}{2}\cdot 10^{-t+1}$$

La produs se va adăuga sumei erorilor relative ale factorilor şi eroarea de rotunjire r_m specifică operației.

$$\left|\frac{e_u}{\bar{u}}\right| = \left|\frac{e_{x+y}}{\bar{x}+\bar{y}} + \frac{e_u}{\bar{u}} + r_m\right| \le \left|\frac{e_{x+y}}{\bar{x}+\bar{y}}\right| + |r_m| \le \frac{1}{2}\cdot 10^{-t+1} + \frac{1}{2}\cdot 10^{-t+1} = 10^{-t+1}$$

Unde $e_z=0$.

Marginea erorii absolute a rezultatului va fi:

$$|e_u| \le |\bar{u}|\cdot 10^{-t+1}$$

Reguli pentru precizia calculelor

Succesiunea operaţiilor, adoptată prin algoritmul de calcul, poate avea o influenţă decisivă asupra mărimii erorii rezultatului final. Astfel:

1) Când se adună și/sau scad mai multe numere se va începe cu acelea mai mici
2) Scăderea a două numere aproximativ egale se va evita prin rescrierea expresiei (dacă este posibil)
3) Algoritmul trebuie conceput cu un număr minim de operaţii aritmetice.