LABORATOR 1 - ELEMENTE DE TEORIA ERORILOR ŞI ARITMETICA ˆIN VIRGULĂ FLOTANTĂ.

1. Surse şi clasificarea erorilor

Estimarea preciziei rezultatelor obţinute reprezintă un aspect foarte important al Analizei Numerice (AN). Se disting mai multe tipuri de erori, care pot limita această precizie:

- erori în datele de intrare;
- erori de rotunjire;
- erori de aproximare.

Erorile în datele de intrare sunt în afara controlului calculelor. Aceste erori nu constituie obiectul AN. Ele se pot datora inexactităţii inerente a măsurătorilor fizice, introducerii eronate a datelor în calculator etc.

Erorile de rotunjire apar atunci când se operează cu numere a căror reprezentare conţine un număr finit de cifre, aşa cum se întâmplă, de exemplu, la reprezentarea numerelor în calculator.

Al treilea tip de erori, cele de aproximare, sunt datorate metodei numerice utilizate. În general, este necesar un număr mare de operaţii pentru a ajunge la soluţia exactă, chiar infinit uneori, iar calculele sunt oprite în funcţie de un anumit criteriu, după un număr finit de paşi, atunci când se ajunge la o precizie acceptabilă (deci conştientizată).

2. Aritmetica în virgulă mobilă

2.1. **Numere în virgulă mobilă.** Calculatorul poate lucra cu două categorii de numere: întregi şi fracţionare. Reprezentarea internă a acestor numere este *în virgulă fixă* şi *în virgulă mobilă (flotantă)*. Pentru cazul reprezentării în virgulă fixă nu se pune problema aproximării, deoarece rezultatul adunării, scăderii sau înmulţirii numerelor întregi este tot un număr întreg, care se reprezintă în calculator dacă valoarea sa nu depăşeşte posibilităţile de reprezentare a numerelor întregi în calculatorul respectiv. Astfel, dacă notăm cu a baza de reprezentare, iar cu t precizia maşinii (numărul de cifre în baza a care se pot reprezenta pe un cuvânt sau multiplu de cuvânt de memorie), numerele întregi reprezentabile formează mulţimea:

$$I := \{x \in \mathbb{Z} \mid -a^{t-1} \leq x \leq a^{t-1} - 1\}.$$

Pentru $a = 2$ şi $t = 16$ cifre binare, $I = \{-2^{15}, ..., 2^{15} - 1\}$. Dacă $t = 32$, domeniul de reprezentare este $I = \{-2^{31}, ..., 2^{31} - 1\}$.

Problema erorilor de rotunjire se pune atunci când calculatorul operează cu numere fracţionare, reprezentate în virgulă mobilă. Mulţimea F a numerelor cu virgulă flotantă este caracterizată de patru parametri: baza b, precizia maşinii t şi

intervalul exponenţilor $[L, U]$. Fiecare număr x cu virgulă flotantă, care aparţine lui F, are reprezentarea:

$$x = \pm \left(\frac{c_1}{b} + \frac{c_2}{b^2} + ... + \frac{c_t}{b^t} \right) \cdot b^e,$$

unde numerele naturale $c_1, ..., c_t$, numite *cifre*, satisfac $0 \leq c_i \leq b-1, i = 1, 2, ..., t$, iar numărul întreg e, numit *exponent*, satisface $L \leq e \leq U$. Numărul

$$f = \left(\frac{c_1}{b} + \frac{c_2}{b^2} + ... + \frac{c_t}{b^t} \right)$$

se numeşte *partea fracţionară* sau *mantisă*. Dacă pentru $x \in F, x \neq 0$, este adevărată relaţia $c_1 \neq 0$, sistemul de reprezentare în virgulă mobilă se numeşte *normalizat.* Se remarcă faptul că mulţimea F nu este un continuu, ea fiind chiar finită, având în cazul reprezentării normalizate $2(b-1)b^{t-1}(U-L+1)+1$ elemente, care nu sunt reprezentate uniform.

Exemplul 1. *Dacă F este normalizată, pentru $b = 2, t = 3, L = -1, U = 2, F$ are 33 de elemente:*

$0, \pm.100 \cdot 2^e, \pm.101 \cdot 2^e, \pm.110 \cdot 2^e, \pm.111 \cdot 2^e,$ *cu* $e \in \{-1, 0, 1, 2\}$.

Figura 1 *Elementele pozitive ale mulţimii F*

.

În figura de mai sus s-au reprezentat numerele pozitive ale mulţimii F.

Mai remarcăm faptul că numerele nenule reprezentabile satisfac relaţia: $m \leq |x| \leq M$, unde $m = b^{L-1}$ se numeşte cel mai mic număr pozitiv reprezentabil, iar $M = b^U \left(1 - \frac{1}{b^t}\right)$ - cel mai mare număr pozitiv reprezentabil. În exemplul nostru $m = \frac{1}{4}, M = \frac{7}{2}$.

De asemenea, se observă faptul că numerele mulţimii F sunt mai "dese" în apropierea originii (puterile lui b scad) şi mai "rare" spre extremitatea lui F (puterile lui b cresc).

Exemplul 2. *Pentru $b = 2$ şi 32 de poziţii binare pentru un număr real, reprezentarea se realizează în* ***simplă precizie****, iar pentru 64 de poziţii binare pentru un număr real, reprezentarea se realizează în* ***dublă precizie.***

După standardul IEEE (the "Institute of Electrical and Electronics Engineers" Inc. USA), pentru $b = 2$ rezultă $t = 24, L = -126$ şi $U = 127$ pentru reprezentarea în simplă precizie şi $t = 53, L = -1022$ şi $U = 1023$ pentru reprezentarea în dublă precizie. Rezultă, de asemenea, pentru reprezentarea în simplă precizie, $m \approxeq 10^{-38}$ şi $M \approxeq 10^{38}$, iar pentru reprezentarea în dublă precizie, $m \approxeq 10^{-308}$ şi $M \approxeq 10^{308}$.

2.2. **Reprezentarea aproximativă a numerelor. Scheme de rotunjire.** Fie x un număr care nu depăşeşte marginile mulţimii F; în calculator acest număr este reprezentat de numărul cu virgulă mobilă notat $fl(x)$; în aceste condiţii, spunem că x este reprezentat aproximativ (rotunjit).

Orice număr x ce satisface relaţia $m \leq |x| \leq M$ se poate scrie sub forma:

$$x = f \cdot b^e + g \cdot b^{e-t}.$$

În cazul fracţiilor normalizate, sunt îndeplinite relaţiile:

$$\frac{1}{b} \leq |f| < 1, 0 \leq |g| < 1.$$

Există mai multe tipuri de rotunjire, şi anume:

(a) *rotunjirea prin tăiere (trunchiere):*

$fl(x)$ este cel mai apropiat element din F faţă de x, cu proprietatea $|fl(x)| \leq |x|$. În notaţia de mai sus, $fl(x) = f \cdot b^e$.

Exemplul 3. *Dacă* $b = 10, t = 4$ *şi* $x = 12945.734$, *putem scrie:*

$$x = 0.12945734 \cdot 10^5 = 0.1294 \cdot 10^5 + 0.5734 \cdot 10.$$

În acest caz, $fl(x) = 0.1294 \cdot 10^5 = 12940 \neq x$.

(b) *rotunjirea uniformă (metoda cifrei pare):*

La rotunjirea uniformă, $fl(x)$ are următoarea expresie:

$$fl(x) = \begin{cases} f \cdot b^e, \text{ dacă } |g| < 0.\overline{\frac{b}{2}} \\ f \cdot b^e \pm b^{e-t}, \text{ dacă } |g| > 0.\overline{\frac{b}{2}} \\ f \cdot b^e \pm b^{e-t}, \text{ dacă } |g| = 0.\overline{\frac{b}{2}}, \text{ ultima cifră } f \text{ - impară} \\ f \cdot b^e, \text{ dacă } |g| = 0.\overline{\frac{b}{2}}, \text{ ultima cifră } f \text{ - pară} \end{cases}.$$

În această expresie, semnul "+" se consideră pentru $f > 0$ şi semnul "-" se consideră pentru $f < 0$. Această modalitate de reprezentare e adoptată şi de standardul IEEE.

Exemplul 4. *Se consideră* $b = 10, t = 4$ *şi o rotunjire uniformă.*

(i) $x = 12944.9942 = 0.1294 \cdot 10^5 + 0.4994 \cdot 10^{5-4}$.

Se observă că $|g| < 0.5$, *deci* $fl(x) = 0.1294 \cdot 10^5 = 12940 \neq x$.

(ii) $x = 129551 = 0.1295 \cdot 10^6 + 0.51 \cdot 10^{6-4}$.

Se observă că $|g| > 0.5$, *deci* $fl(x) = 0.1295 \cdot 10^6 + 10^{6-4} = 129600 \neq x$.

(iii) $x = 1297.5 = 0.12975 \cdot 10^4 = 0.1297 \cdot 10^4 + 0.5 \cdot 10^{4-4}$.

Se observă că $|g| = 0.5$ *şi ultima cifră a lui* f *este impară, deci* $fl(x) = 0.1297 \cdot 10^4 + 10^{4-4} = 1298 \neq x$.

(iv) $x = 1296.5 = 0.12965 \cdot 10^4 = 0.1296 \cdot 10^4 + 0.5 \cdot 10^{4-4}$.

Se observă că $|g| = 0.5$ *şi ultima cifră a lui* f *este pară, deci* $fl(x) = 0.1296 \cdot 10^4 = 1296 \neq x$.

Având în vedere reprezentarea aproximativă a numerelor, se pot defini următoarele două tipuri de erori:

(a) *eroarea absolută,* notată e_x :

$$e_x = |x - fl(x)|;$$

(b) *eroarea relativă,* notată ε_x :

$$\varepsilon_x = \frac{e_x}{|x|} \approxeq \frac{e_x}{|fl(x)|} = \frac{|x - fl(x)|}{|fl(x)|}.$$

În expresia anterioară, $|x|$ s-a aproximat prin $|fl(x)|$ deoarece, în general, valoarea x nu se cunoaşte în sensul că nu se reprezintă exact în calculator. Se demonstrează faptul că eroarea relativă are valoarea cea mai mare atunci când f are în modul valoarea cea mai mică, iar g valoarea cea mai mare în modul, adică $|f| = \frac{1}{b}, |g| = 1$. Avem:

$$\varepsilon_x \lessapprox \frac{|x - fl(x)|}{|fl(x)|} = \frac{|f \cdot b^e + g \cdot b^{e-t} - fl(x)|}{|fl(x)|} \leq k \cdot \frac{1 \cdot b^{e-t}}{\left(\frac{1}{b}\right) \cdot b^e} = k \cdot b^{1-t},$$

unde b^{1-t} este o mărime specifică maşinii de calcul, ea caracterizând precizia relativă de reprezentare. Se observă imediat că la rotunjirea prin tăiere, $k = 1$.

2.3. **Operaţii elementare în virgulă mobilă. 1. Adunarea**

Oricare ar fi două numere x şi y, pentru care există $fl(x)$ şi $fl(y)$, numărului $x + y$ i se asociază numărul $fl(x + y)$, obţinut în felul următor:

(i) se reprezintă intern numerele x şi y prin $fl(x)$ şi, respectiv, $fl(y)$;

(ii) dacă numerele au exponent diferit, atunci numărul cu exponent mai mic se aduce la o formă în care exponentul să fie egal cu cel al celuilalt termen, operaţie numită *denormalizare*. Acest lucru se realizează prin deplasarea mantisei spre dreapta, inserând zerouri după virgulă;

(iii) se adună mantisele şi din rezultat se păstrează t cifre;

(iv) dacă este necesar, se normalizează rezultatul.

Observaţii:

1. Consecinţa principală a acestui mod de definire este aceea că, spre deosebire de aritmetica reală, adunarea nu este asociativă.

2. Scăderea se realizează la fel ca adunarea, cu deosebirea că mantisele se scad. De fapt, scăderea reprezintă o adunare în care scăzătorul are semn schimbat.

Exemplul 5. *Se consideră o aritmetică a virgulei mobile cu* $b = 10, t = 3$, *reprezentare normalizată şi rotunjire prin tăiere. Fie calculul:* $0.001 + 1 - 1$. *Asociind primii doi termeni, se ajunge la rezultatul* $fl(fl(10^{-3} + 1) - 1) = 0$, *care este eronat. În cazul asocierii ultimilor doi termeni se obţine* $fl(10^{-3} + fl(1 - 1)) = 10^{-3}$, *care este rezultatul corect.*

Exemplul 6. *Se consideră aritmetică a virgulei mobile cu* $b = 10, t = 3$, *reprezentare normalizată şi rotunjire prin tăiere. Fie calculul:* $1.001 - 1$. *Se obţine:* $fl(1.001 - 1) = 0$, *rezultat eronat.*

Exemplele 5 şi 6 pun în evidenţă două fenomene nedorite şi generatoare de erori, care pot apărea la efectuarea unei adunări în virgulă mobilă:

1. *omiterea catastrofală*: apare atunci când se adună doi termeni şi valoarea absolută a unui termen este mai mică decât precizia de reprezentare a celuilalt termen; în acest caz, rezultatul este dat de termenul cu valoare absolută mai mare (situaţie ilustrată în Exemplul 5).

2. *neutralizarea termenilor*: apare când se adună numere de semne diferite şi cu valori absolute apropiate; în acest caz, în mod eronat, rezultatul este nul (situaţie ilustrată în Exemplul 6).

Precizia calculelor numerice este caracterizată de două mărimi constante a căror valoare este dependentă de tipul maşinii de calcul folosite. Cele două valori menţionate sunt:

- *epsilonul maşină pentru adunare* (notat ε_m^+) - reprezintă cel mai mic număr real reprezentabil care schimbă, prin adunare, unitatea maşinii de calcul: $fl(1+\varepsilon_m^+) > 1$; el are valoarea b^{1-t};

- *epsilonul maşină pentru scădere* (notat ε_m^-) - reprezintă cel mai mic număr real reprezentabil care schimbă, prin scădere, unitatea maşinii de calcul: $fl(1-\varepsilon_m^-) < 1$; el are valoarea b^{-t}.

Exemplul 7. *Valorile celor două constante de maşină, în standardul IEEE, sunt următoarele:*

- pentru reprezentarea în simplă precizie:

$$\varepsilon_m^- = 5.96 \cdot 10^{-8}, \varepsilon_m^+ = 1.19 \cdot 10^{-7};$$

- pentru reprezentarea în dublă precizie:

$$\varepsilon_m^- = 1.11 \cdot 10^{-16}, \varepsilon_m^+ = 2.22 \cdot 10^{-16}.$$

2. Înmulţirea

Oricare ar fi două numere x şi y, pentru care există $fl(x)$ şi $fl(y)$, numărului $x \cdot y$ i se asociază numărul $fl(x \cdot y)$, obţinut în felul următor:

(i) se reprezintă intern numerele x şi y prin $fl(x)$ şi, respectiv, $fl(y)$;

(ii) se înmulţesc fracţiile şi se adună exponenţii;

(iii) din fracţia rezultată se opresc t cifre;

(iv) dacă este necesar, se normalizează rezultatul.

Observaţii:

1. Înmulţirea nu este asociativă.

2. Împărţirea se realizează în aceeaşi manieră ca şi înmulţirea, cu deosebirea că la pasul (ii) mantisele se împart, iar exponenţii se scad.

Exemplul 8. *Se consideră o aritmetică a virgulei mobile cu* $b = 10, t = 3$, *reprezentare normalizată şi rotunjire prin tăiere. Fie* $x = 22.547$ *şi* $y = 0.43936$. *Urmărind etapele descrise anterior, avem:*

(i) $fl(x) = 0.225 \cdot 10^2$, $fl(y) = 0.439 \cdot 10^0$;

(ii) $0.225 \cdot 0.439 = 0.098775$, $10^2 \cdot 10^0 = 10^2$;

(iii) $0.098775 \cdot 10^2 \to 0.098 \cdot 10^2$;

(iv) $fl(x \cdot y) = 0.098 \cdot 10^2 = 9.8$.

3. Propagarea erorilor în calculele numerice

Având în vedere cele prezentate mai sus, putem scrie relaţiile:

$$x = fl(x) \pm e_x, y = fl(y) \pm e_y, \varepsilon_x = \frac{e_x}{|fl(x)|}, \varepsilon_y = \frac{e_y}{|fl(y)|}.$$

Se numeşte *calcul aproximativ* un calcul efectuat într-o aritmetică a virgulei mobile. Vom accepta ca **postulat** următoarea afirmaţie: *eroarea relativă într-un calcul cu numere aproximative este egală cu suma dintre eroarea relativă produsă de calculul exact respectiv cu numere aproximative* (T_1) *şi eroarea relativă produsă de calculul aproximativ cu numerele exacte corespunzătoare* (T_2).

Fie $*$ una din operaţiile descrise anterior. Postulatul de mai sus, descris de relaţia:

$$\varepsilon_{x*y} = T_1 + T_2,$$

permite determinarea modalităţii de propagare a erorilor relative pentru operaţiile în virgulă mobilă, după cum urmează.

Pentru *calculul exact cu numere aproximative (T_1)* avem:

$$\begin{aligned} e_x &\neq 0, e_y \neq 0, fl(x * y) = fl(x) * fl(y), \\ T_1 &= k_1 \cdot \varepsilon_x + k_2 \cdot \varepsilon_y. \end{aligned}$$

Constantele k_1 şi k_2 pot avea valori diferite, după cum urmează:
- pentru adunare:

$$k_1 = \frac{|fl(x)|}{fl(x) + fl(y)}, k_2 = \frac{|fl(y)|}{fl(x) + fl(y)};$$

- pentru scădere:

$$k_1 = \frac{|fl(x)|}{fl(x) - fl(y)}, k_2 = -\frac{|fl(y)|}{fl(x) - fl(y)};$$

- pentru înmulţire:

$$k_1 = k_2 = 1;$$

- pentru împărţire:

$$k_1 = 1, k_2 = -1.$$

Pentru *calculul aproximativ cu numere exacte (T_2)* avem:

$$e_x = 0, e_y = 0, fl(x * y) \neq fl(x) * fl(y)$$

şi atunci:

$$T_2 \leq k \cdot b^{1-t},$$

ca pentru orice număr real care nu se reprezintă exact.

Exemplul 9. *Se consideră următorul polinom:*

$$y(x) = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1,$$

pentru $x \in [0.998, 1.012]$, cu pasul de evaluare 0.0001. Figura de mai jos ilustrează cu linie punctată rezultatul calculelor în dublă precizie (standardul IEEE) utilizând formula de mai sus, iar cu linie continuă sunt reprezentate cele corecte. Acestea din urmă sunt obţinute evaluând polinomul $y(x) = (x-1)^7$.

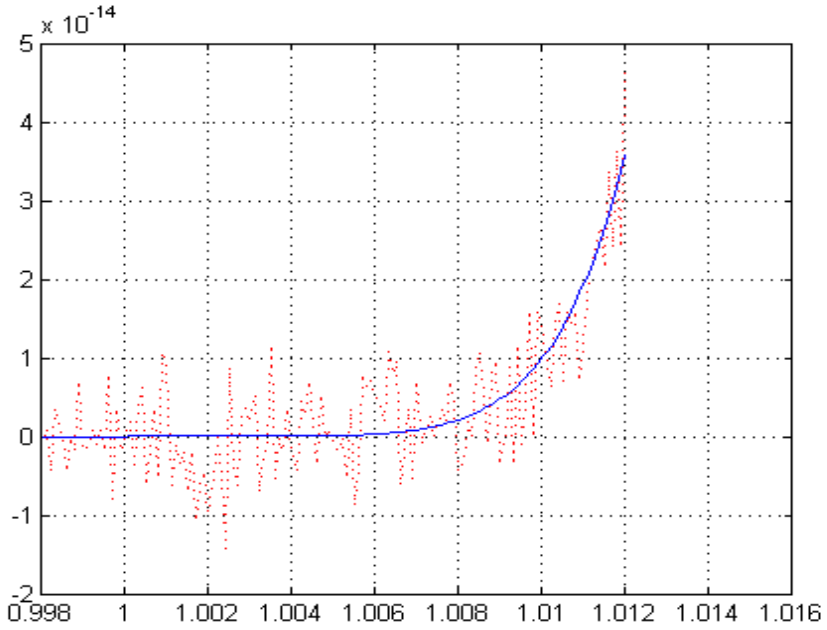

Figura 2 *Ilustrarea propagării erorilor într-un calcul numeric*

Analizând ***Figura 2****, se remarcă faptul că valorile calculate sunt de ordinul* 10^{-14}*, ceea ce evidenţiează erori relativ mici între rezultatele celor două maniere de calcul. Aceste diferenţe se explică prin fenomenele de neutralizarea termenilor şi omitere catastrofală.*

Cele expuse până acum demonstrează faptul că într-un calcul numeric erorile se propagă de la o operaţie la alta. Pe măsură ce numărul operaţiilor dintr-un calcul creşte, pot apărea situaţii în care erorile se acumulează excesiv de mult, fapt care determină obţinerea unei valori total incorecte a rezultatului final. Ca urmare, la întocmirea unui algoritm de calcul şi apoi la implementarea acestuia, utilizatorul trebuie să se asigure că soluţiile nu vor fi afectate de erori care să depăşească anumite limite admisibile.

Ca urmare, se pot formula următoarele reguli generale pentru mărirea preciziei calculelor:

1. când se adună sau se scad numere, este recomandabil să se înceapă cu cele mai mici în valoare absolută, separat pentru cele pozitive şi separat pentru cele negative;

2. dacă este posibil, este recomandabil să se evite scăderea a două numere aproximativ egale; o expresie care conţine o astfel de scădere poate fi rescrisă;

3. o expresie de forma $(a-b)\cdot c$ poate fi rescrisă sub forma $a\cdot c - a\cdot c$, iar o expresie de forma $\dfrac{a-b}{c}$ poate fi rescrisă sub forma $\dfrac{a}{c} - \dfrac{b}{c}$; dacă numerele a şi b sunt aproximativ egale, este recomandabil să se efectueze mai întâi scăderea şi apoi înmulţirea şi împărţirea.

4. dacă regulile generale enunţate mai sus nu se pot aplica, se va urmări minimizarea numărului de operaţii aritmetice implicate.

Exemplul 10. *Se consideră următoarele relaţii de calcul:*

$$\begin{aligned} h &= \frac{1}{2}, x = \frac{2}{3} - h,\ y = \frac{3}{5} - h,\ e = x + x + x - h, \\ f &= y + y + y + y + y - h,\ q = \frac{f}{e}. \end{aligned}$$

Efectuând calculele manual, se obţin rezultatele:

$$h = \frac{1}{2}, x = \frac{1}{6}, y = \frac{1}{10}, e = 0, f = 0, q = ?$$

Implementând aceste relaţii de calcul într-un program scris în MATLAB obţinem rezultatele:

$$e \neq 0, f \neq 0, q = \textit{valoare finită.}$$

Explicaţia acestor rezultate constă în următoarele:

- numărul $\frac{2}{3} = 0.(6)$ *are un număr infinit de cifre în baza de numeraţie zece şi deci nu va fi reprezentat exact;*

- $\frac{3}{5} = 0.6$ *are un număr infinit de cifre în baza de numeraţie doi şi deci nici acesta nu va fi reprezentat exact:* $0.6_{[10]} = 0.(1001)_{[2]}$.

Aceste erori se vor propaga în calculul valorilor e *şi* f *rezultând valori de ordinul epsilonului maşină.*