

## ERORI ÎN CALCULUL NUMERIC

### 1. NUMERE APROXIMATIVE

#### EROAREA ABSOLUTĂ ȘI RELATIVĂ

Numărul  $a$  se numește aproximare a numărului  $A$  dacă valorile lor se deosebesc neînsemnat și  $a$  poate înlocui  $A$  în calcule numerice. Dacă  $a < A$  vom numi  $a$  aproximare prin lipsă, dacă  $a > A$  – aproximare prin adaos. De exemplu pentru  $\pi$  numărul 3.14 va fi aproximare prin lipsă, iar 3.142 – aproximare prin adaos. Într-adevăr  $3.14 < \pi < 3.142$ .

Prin eroare  $\Delta a$  se înțelege diferența  $A - a$  (uneori și  $a - A$ ). În dependență de raportul  $a$  și  $A$ ,  $\Delta a$  poate fi negativ sau pozitiv. Pentru a obține numărul exact  $A$  se adăugă la valoarea erorii  $\Delta a$

$$A = a + \Delta a$$

De multe ori este cunoscută valoarea aproximativă  $a$  fără a se cunoaște semnul erorii. În aceste cazuri se utilizează *eroarea absolută*, care se definește în felul următor:

**Eroare absolută**  $\Delta$  a valorii aproximative  $a$  se consideră modulul diferenței dintre valoarea exactă  $A$  și valoarea aproximativă  $a$ .

$$\Delta = |A - a|$$

Eroarea absolută nu este un indice suficient pentru estimarea exactității calculelor sau măsurătorilor. Fie că în urma măsurării lungimii a două bare au fost obținute rezultatele:

1. 100 m  $\pm$  0.5cm
2. 18 cm  $\pm$  0.2 cm

Cu toate că valoarea absolută a erorii în primul caz este mai mare, este evident că prima măsurătoare a fost realizată mult mai exact decât a doua. Pentru a determina calitatea măsurătorii (calculului) se raportează mărimea erorii absolute la o unitate de lungime (sau, în caz general, la o unitate de măsură respectivă).

**Eroare relativă**  $\delta$  a valorii aproximative se consideră raportul dintre eroarea absolută  $\Delta$  și modulul numărului exact  $A$  ( $A \neq 0$ )

$$\delta = \Delta / |A|$$

**Exemplu:** O bară are lungimea de 100 cm. În urma unei măsurări a fost stabilită o valoare a lungimii egală cu 101 cm. Distanța dintre punctele A și B este de 3000 m. În urma măsurării s-a obținut distanța de 2997 m. Se cere să se determine eroarea relativă a fiecărei măsurări și măsurarea mai exactă.

#### Rezolvare:

pentru prima măsurare

$$\Delta = |100 - 101| = 1 \text{ cm}$$

$$\delta = \frac{1}{100} = 0,01$$

pentru măsurarea a doua

$$\Delta = |3000 - 2997| = 3 \text{ m}$$

$$\delta = \frac{3}{3000} = 0,001$$

Deoarece eroarea relativă a celei de a doua măsurări este mai mică, această măsurare este mai exactă.

### Întrebări și exerciții

1. Explicați noțiunea de eroare. Dați exemple de apariție a erorilor în situații reale.
2. Definiți noțiunea de eroare absolută. Exemplificați.
3. Definiți noțiunea de eroare relativă. Exemplificați. Cum se va modifica formula de calcul a erorii relative, dacă eroarea trebuie indicată în % de la valoarea exactă a mărimii cercetate?
4. O bară cu lungimea exactă de 100 cm a fost măsurată cu o eroare absolută de 2 cm. Care sunt valorile posibile, obținute în procesul de măsurare?
5. Volumul exact al unui vas este de 20l. Măsurările de volum au fost efectuate cu o eroare relativă de 0.001. Care sunt valorile posibile ale volumului măsurat?

## 2. SURSELE ERORILOR DE CALCUL

Erorile, care apar în timpul rezolvării problemelor pot proveni din diferite surse. Cunoașterea curselor de apariție a erorilor permite ocolirea lor și minimizarea efectului cumulativ al erorilor. Cele mai des întâlnite tipuri de erori sunt:

- Erori de problemă
- Erori de metodă
- Erori ale datelor de intrare
- Erori de aproximare
- Erori de rotunjire.

**Erori de problemă.** Această categorie de erori apare în situațiile când modelul matematic ales pentru rezolvarea problemei nu descrie complet procesul real cercetat. Astfel, în exemplul 2 (1.2 Modelul matematic și modelarea matematică) pentru a construi modelul matematic al sistemului *bilă-arc* s-a utilizat ecuația oscilațiilor armonice pentru *deformații mici* și *în lipsa forței de frecare*. Prin urmare, rezultatul obținut prin utilizarea formulei va fi diferit de cel exact, diferența fiind cu atât mai semnificativă cu cât e mai mare forța de frecare în sistemul real și deformarea arcului.

**Erori de metodă.** Este o categorie de erori generată de imposibilitatea determinării unei metode exacte de rezolvare a problemei. În acest caz problema inițială este rezolvată prin o metodă euristică, care poate genera diferențe destul de mari între rezultatul obținut și cel exact.

Un exemplu elocvent este utilizarea metodei *Greedy* pentru rezolvarea *problemei rucsacului*., care are următoarea formulare:

*Este dat un rucsac de volum  $S$  și  $n$  obiecte cu volumele  $v_i, i=1..n$  și costurile  $c_i, i=1..n$ . Să se pună în rucsac obiecte din setul propus astfel, încât costul total al obiectelor puse în rucsac să fie cât mai mare posibil.*

Dacă pentru rezolvarea problemei este folosită metoda *Greedy* în majoritatea cazurilor soluția obținută va fi mai rea decât cea exactă. Astfel, pentru un set din cinci obiecte cu volumele 5, 7, 13, 20, 10, prețurile respectiv de 4, 8, 15, 23, 5, și un rucsac cu volumul de 30 unități prin metoda *Greedy* se obține soluția cu valoarea 27 formată din obiectele cu indicii 3, 2, 1. Pentru același set de date inițiale, soluția determinată prin metoda *trierii* sau *reluării* are valoarea 31, fiind formată din obiectele cu indicii 2 și 4.

**Erori ale datelor de intrare.** Deseori procesul de modelare matematică se bazează pe rezultatele unor experiențe, adică pe careva seturi de mărimi numerice, obținute în urma măsurărilor. Aceste mărimi nu sunt exacte. (ex.: distanța, masa, viteza). În cele ce urmează se va cerceta, cum pot influența erorile inițiale (a datelor de intrare) rezultatele finale.

**Exemplul 1:** fie că se cercetează un model descris de funcția  $f(x) = x^2 + x + 1$  pe segmentul  $[0,1]$  ( $x \in [0,1]$ )

Este cunoscut că  $x$  se calculează cu o exactitate de 0,01 (adică eroarea absolută la determinarea lui  $x$  nu depășește 0,01). Dacă  $z$  este valoarea exactă a argumentului, atunci  $|z - x| \leq 0,01$ .

Vom încerca să determinăm în ce măsură influențează eroarea la măsurarea inițială a valorilor lui  $x$  rezultatul calculului:

$$|f(x) - f(z)| = |x^2 + x + 1 - z^2 - z - 1| = |x^2 + x - z^2 - z| = |(x - z)(x + z) + x - z| = |(x - z)(x + z + 1)|$$

Deoarece  $x + z + 1 \leq 3$  și  $|z - x| \leq 0,01$ , rezultă

$$|f(x) - f(z)| \leq 3 \times 0,01.$$

Diferența dintre valoarea funcției de argument exact ( $z$ ) și valoarea funcției de argument măsurat ( $x$ ) este o mărime constantă. De aici rezultă că erorile de calcul nu depind de  $x$  ci numai de exactitatea cu care acesta este măsurat.

Se consideră că funcția  $f(x): I \rightarrow R$  posedă proprietatea Lipschits dacă  $\exists$  o constantă  $m > 0$  astfel încât:  $|f(x) - f(z)| \leq m \times |z - x|$ ,  $\forall x, z \in I$ .

Se consideră că determinarea valorilor funcției  $f(x)$  este stabilă la erori, dacă  $f(x)$  posedă proprietatea Lipschits.

În alți termeni, stabilitatea funcției la erori presupune variații mici ale valorilor funcției la variații mici ale argumentului.

**Erori de aproximare.** Este o categorie de erori generate de anumite definiții și noțiuni matematice. Prezența lor este acceptată, în special în problemele care folosesc noțiunea de limită, convergență etc. Apariția este motivată de însăși structura definiției care conține elemente de aproximare.

**Exemplul 2:** Se consideră că șirul  $\{x_n\}$  este convergent și are limita  $x$  dacă pentru orice  $\varepsilon > 0$ ,  $\varepsilon \rightarrow 0$ , există un rang  $n(n_\varepsilon)$  astfel încât  $|x - x_n| < \varepsilon \quad \forall n > n_\varepsilon$ .

Din punct de vedere al analizei numerice  $\varepsilon$  indică precizia cu care  $x_n$  îl aproximează pe  $x$ . Fie  $\varepsilon = 0,001$ . Dacă șirul  $\{x_n\}$  este convergent și are limita  $x$ , înseamnă că  $\exists n_\varepsilon(n_{0,001})$  astfel încât  $\forall n > n_\varepsilon \quad |x - x_n| < 0,001$ .

În procesul de calcul al limitei se utilizează o aproximare consecutivă, care se apropie tot mai mult de limita exactă, fără a o atinge. Stoparea procesului de calcul are loc atunci când deviația (eroarea) devine mai mică decât eroarea maximă admisibilă ( $\varepsilon$ ).

**Exemplul 3:** Aproximarea valorilor termenilor șirului  $1/n$  de limita lui 0.

|         |                |
|---------|----------------|
| n=1     | 1/n =1         |
| n=2     | 1/n =0,5       |
| ...     |                |
| n= 501  | 1/n = 0,001996 |
| n= 502  | 1/n = 0,001992 |
| n= 503  | 1/n = 0,001988 |
| ...     |                |
| n= 999  | 1/n = 0,001001 |
| n= 1000 | 1/n = 0,001000 |
| n= 1001 | 1/n = 0,000999 |
| ...     |                |

De reținut, că rezultatele obținute prin calcul numeric sunt acceptate doar ca rezultate cantitative, și nu pot servi drept demonstrație pentru careva afirmații matematice.

**Erori de rotunjire.** Este un tip aparte de erori, generate de faptul că în procesul prelucrării mărimilor numerice în calculator ele pot fi păstrate doar cu un anumit număr de semne zecimale după virgulă. Exemplu poate servi constanta  $\pi$ ; valorile funcțiilor trigonometrice etc.

Fie  $A = \{a_1, a_2, a_3, \dots, a_n\}$  mulțimea tuturor numerelor care pot fi reprezentate în calculator. (Se consideră că  $a_1 < a_2 < \dots < a_n$ )

Vom cerceta oricare din numerele  $(a_i + a_{i+1})/2$ . Evident el lipsește în mulțimea inițială  $A$ . În cazul apariției unei asemenea situații de calcul apare eroarea legată de înlocuirea rezultatului real prin numărul cel mai apropiat de el din mulțimea  $A$ . Această procedură este numită rotunjire.

*Funcția de rotunjire* aplicată la calculatoare se definește în felul următor:  $\forall x \in \mathbb{R}, a_i, a_{i+1} \in A$

$$rd(x) = a_i \text{ dacă } x \in [a_i, (a_i + a_{i+1})/2)$$

$$rd(x) = a_{i+1} \text{ dacă } x \in [(a_i + a_{i+1})/2, a_{i+1}]$$

Utilizarea funcției de rotunjire generează abateri de la legile de bază ale operațiilor aritmetice: adunarea, înmulțirea nu mai sunt asociative, distributive.

**Exemplul 4:** Fie că în procesul de calcul se poate opera cu cel mult 3 cifre după virgulă. În acest caz pentru numerele

$$a=0,2334 \quad b=0,2331 \quad c=0,233 \text{ se obține}$$

$$rd(rd(a+b)+c) = rd(rd(0,4665)+0,233) = rd(0,467+0,233) = 0,7$$

$$rd(a+rd(b+c)) = rd(0,2334+rd(0,4661)) = rd(0,2334+0,466) = rd(0,6994) = 0,699$$

În cazul rotunjirii rezultatelor de calcul erorile nu depășesc după modul valoarea de  $0,5 \cdot 10^{-n}$ , (aceste erori se numesc erori de rotunjire absolute) unde  $n$  este numărul de semne semnificative (care pot fi percepute) în procedura de calcul. Erorile de rotunjire pot fi atât pozitive cât și negative. În cazul alternării lor are loc procesul de compensare, în rezultat eroarea finală nu crește odată cu numărul de calcule.

## Întrebări și exerciții

1. Identificați principalele surse de erori. Exemplificați
2. Dați exemple de erori, generate de imposibilitatea formulării exacte a problemei. Motivați imposibilitatea formulării exacte.
3. Utilizarea metodei Greedy în rezolvarea problemelor de cele mai multe ori duce la obținerea unui soluții diferite de soluția optimă. De ce?
4. Care funcții posedă proprietatea de stabilitate a calculului valorilor față de erori?
5. Este oare stabil față de erori calculul valorilor funcției liniare? Dar al funcției  $y = \sqrt{x}, \quad x \in [1,2]$ ?
6. Explicați noțiunea de eroare de aproximare.

7. Soluția unei probleme se calculează iterativ după formula
 
$$x_0 = 0$$

$$x_{i+1} = \sqrt{2 + x_i}$$

Va atinge oare șirul soluțiilor calculate valoarea soluției exacte? Stabiliți experimental

sau analitic soluția exactă. După câte iterații eroarea absolută a soluției calculate va deveni mai mică de 0,0001?

8. Din care motiv rezultatele obținute prin metode numerice nu pot servi drept demonstrații a afirmațiilor matematice?
9. Care este motivul apariției erorilor de rotunjire? Exemplificați.
10. Scrieți un program care determină produsul și câtul numerelor 1,00000001 și 0,999999999. Analizați rezultatele obținute. Explicați cauzele apariției erorii.