# 1.3. Erori în calculele numerice

Prof.dr.ing. Gabriela Ciuprina

Universitatea "Politehnica" Bucureşti, Facultatea de Inginerie Electrică,
Departamentul de Electrotehnică

Suport didactic pentru disciplina *Metode numerice*, 2017-2018



Notes



## Cuprins





Notes

## Eroarea absolută şi marginea ei

Fie:
$\mathbf{x} \in \mathbb{R}^n$ - valoarea exactă a unei mărimi;
$\bar{\mathbf{x}}$ - valoarea aproximativă.
Eroarea absolută $\mathbf{e_x} \in \mathbb{R}^n$:

$$\mathbf{e_x} = \bar{\mathbf{x}} - \mathbf{x}. \tag{1}$$

Marginea erorii absolute $a_x \in \mathbb{R}$:

$$\|\mathbf{e_x}\| \leq a_x. \tag{2}$$

Dacă $n = 1$ rezultă

$$\bar{x} - a_x \leq x \leq \bar{x} + a_x. \tag{3}$$

Echivalentă cu: $x \in [\bar{x} - a_x, \bar{x} + a_x]$.
Scrisă pe scurt ca:

$$"x = \bar{x} \pm a_x". \tag{4}$$



Notes



## Eroarea relativă şi marginea ei

Eroarea relativă $\varepsilon_x \in \mathbb{R}^n$:

$$\varepsilon_\mathbf{x} = \frac{\mathbf{e_x}}{\|\mathbf{x}\|}. \tag{5}$$

Marginea erorii relative $r_x \in \mathbb{R}$

$$\|\varepsilon_\mathbf{x}\| \leq r_x. \tag{6}$$

Cel mai adesea, $r_x$ se exprimă în procente.
Scriere pe scurt:

$$"\mathbf{x} = \bar{\mathbf{x}} \pm r_x\%". \tag{7}$$



Notes

## Exemplu: $\pi$

$x = 3.1415\ldots$
$\bar{x} = 3.14$
$e_x = -0.0015\ldots$
$a_x = 0.0016$
$\varepsilon_x = -0.0015\ldots/3.1415\ldots$
$r_x = 0.0016/3 \leq 0.0006 = 0.06\%.$

$$\pi = 3.14 \pm 0.0016 \quad \text{sau} \quad \pi = 3.14 \pm 0.06\%.$$



Notes



## Concluzii

**Relaţia "$\mathbf{x} = \bar{\mathbf{x}} \pm a_x$"**

unde $\mathbf{x}, \bar{\mathbf{x}} \in \mathbb{R}^n$ şi $a_x \in \mathbb{R}$ se interpretează astfel:

$$(\exists)\mathbf{e_x} \in \mathbb{R}^n, \|\mathbf{e_x}\| \leq a_x, \quad \text{astfel încât} \quad \bar{\mathbf{x}} = \mathbf{x} + \mathbf{e_x}, \tag{8}$$

**Relaţia "$\mathbf{x} = \bar{\mathbf{x}} \pm r_x\%$"**

unde $\mathbf{x}, \bar{\mathbf{x}} \in \mathbb{R}^n$, $r_x\% = 100 r_x$ şi $r_x \in \mathbb{R}$ se interpretează astfel:

$$(\exists)\varepsilon_{\mathbf{x}} \in \mathbb{R}^n, \|\varepsilon_{\mathbf{x}}\| \leq r_x, \quad \text{astfel încât} \quad \bar{\mathbf{x}} = \mathbf{x} + \|\mathbf{x}\|\varepsilon_{\mathbf{x}}. \tag{9}$$

În cazul unei mărimi scalare ($n = 1$), relaţia (9) se scrie

$$\bar{x} = x(1 \pm \varepsilon_x), \tag{10}$$

semnul plus corespunzând unei valori $x$ pozitive, iar semnul minus uneia negative.



Notes

## Tipuri de erori

În funcţie de tipul cauzelor care le generează:

1. Erori de rotunjire - datorate reprezentării finite a numerelor reale;
2. Erori de trunchiere - datorate reprezentării finite a algoritmului;
3. Erori inerente - datorate reprezentării imprecise a datelor de intrare.



Notes



## Cifre semnificative

Reprezentarea unui număr real în baza 10:

$$\bar{x} = f \cdot 10^n. \tag{11}$$

unde $0.1 \le |f| < 1$.
Cifrele părţii fracţionare se numesc cifre semnificative.
Exemple:
$3.14 = 0.314 \cdot 10^1$
$-0.007856 = -0.7856 \cdot 10^{-2}$.



Notes

## Rotunjirea afectează reprezentarea numerelor reale

$$\bar{x} = \overbrace{0.\underbrace{***\cdots*}_{k \text{ cifre}}}^{f} \cdot 10^n, \tag{12}$$

$$x = 0.\underbrace{***\cdots*}_{k \text{ cifre}}\#\#\#\cdots \cdot 10^n, \tag{13}$$

$$e_x = \bar{x} - x = -0.\underbrace{000\cdots0}_{k \text{ cifre}}\#\#\#\cdots \cdot 10^n = -0.\#\#\#\cdots \cdot 10^{n-k}, \tag{14}$$





## Rotunjirea afectează reprezentarea numerelor reale

$$\varepsilon_x = \frac{e_x}{x} = \frac{-0.\#\#\#\cdots \cdot 10^{n-k}}{0.\underbrace{***\cdots*}_{k \text{ cifre}}\#\#\#\cdots \cdot 10^n} = -\frac{0.\#\#\#\cdots}{0.***\cdots}10^{-k} \tag{15}$$

$$|\varepsilon_x| \leq \frac{1}{0.1}10^{-k} = 10^{-k+1}. \tag{16}$$

Marginea erorii relative de rotunjire a unui sistem de calcul depinde doar de numărul de cifre semnificative ce pot fi memorate. Pentru un sistem de calcul ce lucrează cu $k$ cifre semnificative, marginea erorii relative de rotunjire este $10^{-k+1}$.



Notes

Notes

## Rotunjirea afectează calculele

Adunarea a două numere reale
Intuitiv: pp. $k = 3$, $x_1 + x_2 = ?$
$x_1 = 3.73 = 0.373 \cdot 10^1$
$x_2 = 0.006 = 6 \cdot 10^{-3}$
$x_2 = 6 \cdot 10^{-4} \cdot 10^1 = 0.0006 \cdot 10^1 = 0.000 \cdot 10^1$
Rezultat: $x_1 + x_2 = x_1$.



Notes



## Zeroul maşinii

Zeroul (acurateţea, precizia,"epsilon-ul") maşinii = cel mai mic `eps` pentru care `1 + eps > 1`.

- $(\forall) a < \texttt{eps}$, $1 + a = 1$ (în calculator)
- în mod uzual $\texttt{eps} = 2.22 \cdot 10^{-16}$.
- Matlab: `eps`
- Scilab `%eps`.
- Zeroul maşinii nu trebuie confundat cu cel mai mic număr reprezentabil în calculator şi care, în mod uzual are valoarea $2.23 \cdot 10^{-308}$.

Consecinţă: adunarea numerelor reale în calculator nu este asociativă.



Notes

## Determinarea eps într-un mediu de programare

```
funcţie zeroul_maşinii ()
real eps
eps = 1
cât timp (1 + eps > 1)
        eps = eps/2
•
eps = eps*2
întoarce eps
```



Notes



## Exemplu

$$f(x) = f(x_0) + \frac{x - x_0}{1!} f'(x_0) + \frac{(x - x_0)^2}{2!} f''(x_0) + \cdots \tag{17}$$

sinus, $x_0 = 0$:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}. \tag{18}$$

$$\bar{s} = \sum_{k=0}^{n} (-1)^k \frac{x^{2k+1}}{(2k+1)!}. \tag{19}$$

$$|e_s| = |\bar{s} - s| \le \frac{x^{2n+1}}{(2n+1)!}. \tag{20}$$



Notes

## Algoritm cu controlul erorii de trunchiere

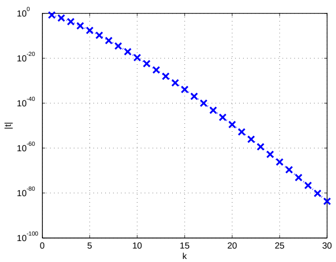
```
funcţie sinus(x, e)
; întoarce valoarea funcţiei sinus in punctul x
; prin trunchierea seriei Taylor dezvoltată in 0
real x              ; punctul în care se va evalua funcţia sin
real e              ; eroarea de trunchiere impusă
real t, s
întreg k
t = x
s = t
k = 0
cât timp (|t| > e)
    k = k + 1
    t = (−1) ∗ t ∗ x^2/((2k)(2k+1))
    s = s + t
•
intoarce s
```



Notes

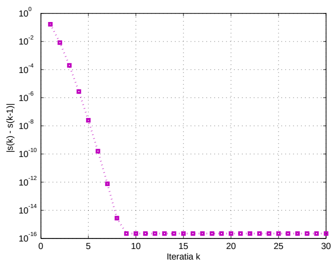


## Rezultate numerice




Modulul termenului curent al dezvoltării în serie

Taylor a funcţiei sinus.




Modulul diferenţei dintre sume parţiale consecutive

la dezvoltarea în serie Taylor a funcţiei sinus.



Notes

## Efectul perturbaţiilor datelor de intrare

$$y = f(x_1, x_2, \dots, x_n). \tag{21}$$

$$\mathrm{d}y = \frac{\partial f}{\partial x_1}\mathrm{d}x_1 + \frac{\partial f}{\partial x_2}\mathrm{d}x_2 + \dots \frac{\partial f}{\partial x_n}\mathrm{d}x_n. \tag{22}$$

$$\Delta y \approx \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 + \dots \frac{\partial f}{\partial x_n}\Delta x_n. \tag{23}$$

$$\Delta x_k = \bar{x}_k - x_k = e_{x_k}, \tag{24}$$



Notes



## Eroarea absolută a rezultatului şi marginea ei

$e_y = \bar{y} - y = \Delta y$:

$$e_y = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k} e_{x_k}. \tag{25}$$

$$\left| \sum_{k=1}^{n} \frac{\partial f}{\partial x_k} e_{x_k} \right| \leq \sum_{k=1}^{n} \left| \frac{\partial f}{\partial x_k} e_{x_k} \right| = \sum_{k=1}^{n} \left| \frac{\partial f}{\partial x_k} \right| |e_{x_k}| \leq \sum_{k=1}^{n} \left| \frac{\partial f}{\partial x_k} \right| a_{x_k}, \tag{26}$$

unde $|e_{x_k}| \leq a_{x_k}$.

Marginea erorii absolute a rezultatului

$$a_y = \sum_{k=1}^{n} \left| \frac{\partial f}{\partial x_k} \right| a_{x_k}. \tag{27}$$



Notes

## Eroarea relativă a rezultatului şi marginea ei

$$\varepsilon_y = e_y/|y|$$

$$\varepsilon_y = \frac{\sum_{k=1}^{n} \frac{\partial f}{\partial x_k} e_{x_k}}{|y|} = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k} \frac{e_{x_k}}{|y|} = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k} \frac{|x_k|}{|y|} \varepsilon_{x_k}. \tag{28}$$

Marginea erorii relative a rezultatului

$$r_y = \sum_{k=1}^{n} \left| \frac{\partial (\ln f)}{\partial x_k} \right| |x_k| r_{x_k}. \tag{29}$$



Notes



## Cazuri particulare: +, -

| **Erori** | **Adunare** $y = x_1 + x_2$ | **Scădere** $y = x_1 - x_2$ |
|---|---|---|
| Eroare absolută: $e_y =$ | $e_{x_1} + e_{x_2}$ | $e_{x_1} - e_{x_2}$ |
| majorată de: $a_y =$ | $a_{x_1} + a_{x_2}$ | $a_{x_1} + a_{x_2}$ |
| Eroare relativă: $\varepsilon_y =$ | $\left\lvert\frac{x_1}{x_1+x_2}\right\rvert \varepsilon_{x_1} + \left\lvert\frac{x_2}{x_1+x_2}\right\rvert \varepsilon_{x_2}$ | $\left\lvert\frac{x_1}{x_1-x_2}\right\rvert \varepsilon_{x_1} - \left\lvert\frac{x_2}{x_1-x_2}\right\rvert \varepsilon_{x_2}$ |
| majorată de $r_y =$ | $\left\lvert\frac{x_1}{x_1+x_2}\right\rvert r_{x_1} + \left\lvert\frac{x_2}{x_1+x_2}\right\rvert r_{x_2}$ | $\left\lvert\frac{x_1}{x_1-x_2}\right\rvert r_{x_1} + \left\lvert\frac{x_2}{x_1-x_2}\right\rvert r_{x_2}$ |

Erorile rezultatului adunării şi scăderii a două numere reale în funcţie de erorile datelor de intrare.

NB! La adunare şi scădere marginile erorilor absolute se adună.

- Adunarea este o operaţie bine condiţionată.
- Scăderea este o operaţie prost condiţionată.



Notes

# Exemplu

$x_1 = 1.23 \pm 1\%$ , $x_2 = 1.22 \pm 1\%$

- Scădere:
  $r = |1.23/0.01 \cdot 1/100 + 1.22/0.01 \cdot 1/100 =$
  $1.23 + 1.22 = 2.45 = 245\%$
  $x_1 - x_2 = 0.01 \pm 245\%$.
- Adunare:
  $r = |1.23/2.45 \cdot 1/100 + 1.22/2.45 \cdot 1/100 \approx$
  $0.5 \cdot 1/100 + 0.5 \cdot 1/100 = 1/100 = 1\%$.
  $x_1 + x_2 = 2.45 \pm 1\%$.



Notes



# Cazuri particulare: *, /

| **Erori** | **Înmulţire** $y = x_1 x_2$ | **Împărţire** $y = \frac{x_1}{x_2}$ |
|---|---|---|
| Eroare absolută: $e_y =$ | $x_2 e_{x_1} + x_1 e_{x_2}$ | $\frac{1}{x_2} e_{x_1} - \frac{x_1}{x_2^2} e_{x_2}$ |
| majorată de: $a_y =$ | $\|x_2\| a_{x_1} + \|x_1\| a_{x_2}$ | $\frac{1}{\|x_2\|} a_{x_1} + \frac{\|x_1\|}{x_2^2} a_{x_2}$ |
| Eroare relativă: $\varepsilon_y =$ | $\varepsilon_{x_1} + \varepsilon_{x_2}$ | $\varepsilon_{x_1} - \varepsilon_{x_2}$ |
| majorată de $r_y =$ | $r_{x_1} + r_{x_2}$ | $r_{x_1} + r_{x_2}$ |

Erorile rezultatului înmulţirii şi împărţirii a două numere reale în funcţie de erorile datelor de intrare.

NB! La înmulţire şi împărţire marginile erorilor relative se adună.

- Înmulţirea şi împărţirea sunt operaţii bine condiţionate.



Notes

## Scăderea trebuie evitată

$ax^2 + bx + c = 0$
$x_{1,2} = (-b \pm \sqrt{b^2 - 4ac})/(2a)$
Ce se întâmplă dacă $b > 0$ şi $b^2 \gg 4ac$ ?
Ce avantaj are următorul cod?

<u>`dacă`</u> $b > 0$
 $x1 = (-b - \sqrt{b^2 - 4ac})/(2a)$
<u>`altfel`</u>
 $x1 = (-b + \sqrt{b^2 - 4ac})/(2a)$
•
$x2 = c/(a * x1)$



Notes



## Extragerea radicalului

$y = \sqrt{x}$

$$e_y = \frac{df}{dx} e_x = \frac{1}{2\sqrt{x}} e_x, \tag{30}$$

$$\varepsilon_y = \frac{e_y}{y} = \frac{1}{2\sqrt{x}\sqrt{x}} e_x = \frac{e_x}{2x} = \frac{\varepsilon_x}{2}. \tag{31}$$

Dar rotunjirea nu poate fi ignorata!



Notes

## Superpoziţia erorilor

eroarea relativă într-un calcul aproximativ
=
eroarea relativă produsă de calculul aproximativ cu numere exacte (eroarea de rotunjire)
+
eroarea relativă produsă de calculul exact cu numere aproximative (afectate deci de erori inerente).

$$\bar{y} = y_i(1 + \mathrm{eps}) = y(1 + \varepsilon_y)(1 + \mathrm{eps}) \approx y(1 + \varepsilon_y + \mathrm{eps}),$$

de unde $(\bar{y} - y)/y = \varepsilon_y + \mathrm{eps}$.

$$\varepsilon_{\sqrt{x}} = \frac{\varepsilon_x}{2} + \mathrm{eps}. \tag{32}$$

Eroarea relativă a oricărui rezultat numeric este cel puţin egală cu zeroul maşinii.



Notes



## Condiţionare vs. stabilitate

**Condiţionarea**

se referă la comportarea problemei matematice la perturbaţii ale datelor.

**Stabilitatea**

se referă la comportarea algoritmului la perturbaţii ale datelor.



Notes

## Condiţionare

Problemă matematică $f$ formulată explicit:

$$\text{Fie}\quad f : \mathcal{D} \to \mathcal{X} \quad \text{şi} \quad \mathbf{d} \in \mathcal{D}.$$
$$\text{Să se găsească} \quad \mathbf{x} \in \mathcal{X} \quad \text{astfel încât} \quad f(\mathbf{d}) = \mathbf{x}. \quad (33)$$

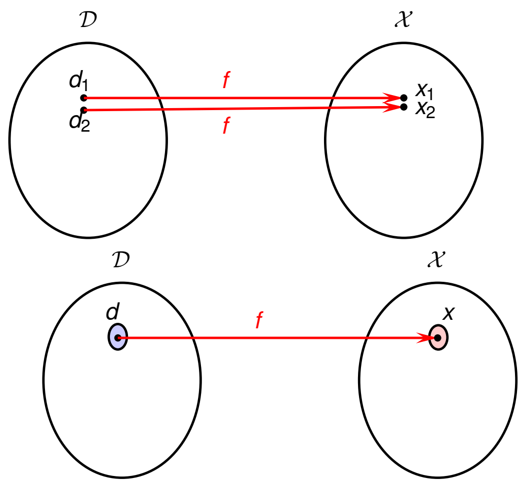
O problemă este bine condiţionată dacă perturbaţii mici ale datelor conduc la perturbaţii mici ale rezultatului.



Notes



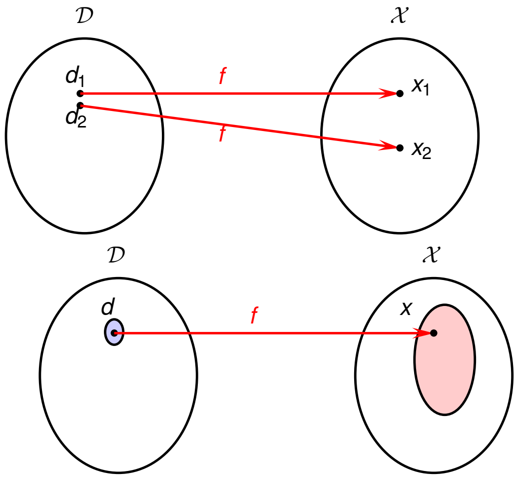
## Reprezentări intuitive - problemă bine condiţionată






Notes

## Reprezentări intuitive - problemă prost condiţionată






Notes



## Condiţionare

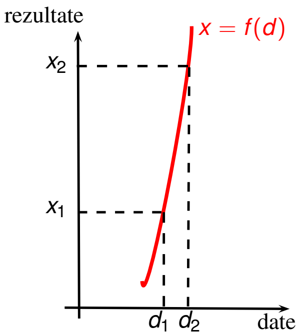
Problemă matematică poate fi formulată şi implicit:

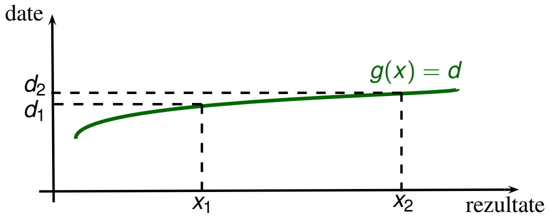
Fie $g : \mathcal{X} \to \mathcal{D}$ şi $\mathbf{d} \in \mathcal{D}$.
Să se găsească $\mathbf{x} \in \mathcal{X}$ astfel încât $g(\mathbf{x}) = \mathbf{d}$. (34)



Notes

## Reprezentări intuitive - problemă prost condiţionată






Notes



## Condiţionare - rezultat important

Se demonstrează că între perturbaţia în date (reziduu) şi perturbaţia în rezultat (eroare) există următoarea relaţie:

$$\|\mathbf{e_x}\| \leq \kappa \|\varepsilon_\mathbf{d}\|, \tag{35}$$

unde $\kappa$ este un scalar numit *număr de condiţionare*, care depinde de problema numerică abordată. *(Vom reveni asupra lui la cursul următor).*



Notes

## Condiţionare - concluzii

- Reziduul nu dă informaţii despre eroare.
- Eroarea şi reziduul sunt legate prin numărul de condiţionare.
- Pentru o problemă cu număr de condiţionare mic, o perturbaţie mică în date va duce la o perturbaţie mică a rezultatului.
- Problemele matematice care au $\kappa$ mare sunt prost condiţionate şi ele nu pot fi rezolvate cu ajutorul calculatorului. Pentru astfel de probleme, trebuie găsită o formulare matematică echivalentă din punct de vedere al rezultatului, dar bine condiţionată.



Notes



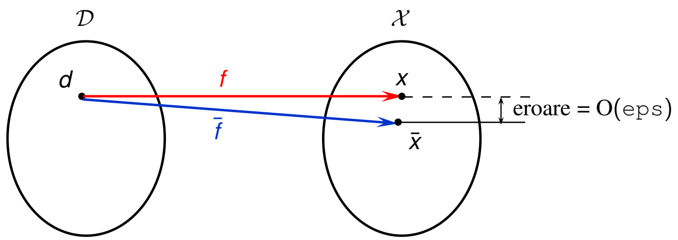
În cele ce urmează vom presupune că problema $f$ este bine condiţionată şi pentru rezolvarea ei a fost conceput un algoritm $\bar{f}$.



Notes

# Acurateţea unui algoritm

Acurateţea unui algoritm se referă la eroarea soluţiei numerice.




Reprezentarea intuitivă a unui algoritm a cărui precizie este ideală.

În mod ideal, un algoritm este precis dacă:

$$\frac{\|\bar{f}(\mathbf{d}) - f(\mathbf{d})\|}{\|f(\mathbf{d})\|} = \mathrm{O}(\mathrm{eps}). \qquad (36)$$

$\bar{f}(\mathbf{d})$ = "rezultatul algoritmului $\bar{f}$ aplicat datelor $\mathbf{d}$".



Notes



# Stabilitatea unui algoritm

Dar, rotunjirea datelor este inevitabilă, erorile se acumulează şi perturbă rezultatul. Este mai util să se ţintească stabilitatea algoritmului.
Stabilitatea unui algoritm se referă la comportarea algoritmului atunci când datele de intrare sunt perturbate.
Un algoritm $\bar{f}$ folosit pentru rezolvarea unei probleme $f$ este stabil dacă

$$\frac{\|\bar{f}(\bar{\mathbf{d}}) - f(\mathbf{d})\|}{\|f(\mathbf{d})\|} = \mathrm{O}(\mathrm{eps}), \qquad (37)$$

pentru $(\forall)\bar{\mathbf{d}}, \mathbf{d}$ care satisfac $\|\bar{\mathbf{d}} - \mathbf{d}\|/\|\mathbf{d}\| = \mathrm{O}(\mathrm{eps})$.
Pe scurt, un algoritm stabil dă raspunsul aproape corect pentru date reprezentate aproape precis.



Notes

## Ilustrarea stabilităţii unui algorim - problema

$$\mathbf{Ax} = \mathbf{b}, \quad \text{unde} \quad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

$$\begin{aligned} & & x_2 &= 1 \\ x_1 &+ & x_2 &= 0 \end{aligned} \tag{38}$$

$x_1 = -1$, $x_2 = 1$. $\mathbf{x} = f(\mathbf{d}) = [-1, 1]^T$.
Să considerăm acum că datele au fost perturbate:

$$\bar{\mathbf{A}} = \begin{bmatrix} 10^{-20} & 1 \\ 1 & 1 \end{bmatrix},$$

$$\begin{aligned} 10^{-20}x_1 &+ x_2 &= 1 \\ x_1 &+ x_2 &= 0 \end{aligned} \tag{39}$$

$x_1' = -x_2' = 1/(10^{-20} - 1) \approx -1$. Se poate demonstra că această problemă este bine condiţionată.



Notes



## Ilustrarea stabilităţii unui algorim - algoritmul $\bar{f}_1$

- **Pasul 1:** se înmulţeşte prima ecuaţie a sistemului cu $(-10^{20})$ şi se adună cu a doua, rezultând $x_2$;
- **Pasul 2:** se calculează $x_1$ din prima ecuaţie.

La pasul 1 se ajunge la ecuaţia $(1 - 10^{20})x_2 = -10^{20}$ care, în calculator devine datorită rotunjirilor $-10^{20}x_2 = -10^{20}$, de unde va rezulta $x_2 = 1$, ceea ce este corect.
La pasul 2 ecuaţia de rezolvat devine $10^{-20}x_1 + 1 = 1$, de unde va rezulta $x_1 = 0$, ceea ce este greşit, foarte departe de valoarea adevărată.
Acest algoritm este instabil.



Notes

# Ilustrarea stabilităţii unui algorim - algoritmul $\bar{f}_2$

- **Pasul 1:** se înmulţeşte a doua ecuaţie a sistemului cu $(-10^{-20})$ şi se adună cu prima, rezultând $x_2$;
- **Pasul 2:** se calculează $x_1$ din a doua ecuaţie.

La pasul 1 se ajunge la ecuaţia $(1 - 10^{-20})x_2 = 1$, care în calculator devine $x_2 = 1$.
La pasul 2 ecuaţia de rezolvat este $x_1 + 1 = 0$, de unde $x_1 = -1$, ceea ce este corect.
Algoritmul $\bar{f}_2$ este stabil. Stabilitatea lui este foarte puternică, el a dat raspunsul exact pentru date de intrare aproape precise.



Notes



# Concluzii - estimarea acurateţii unei solutii numerice

1. Se estimează numărul de condiţionare al problemei. Se continuă numai dacă problema matematică este bine condiţionată.
2. Se investighează stabilitatea algoritmului. Cel mai simplu este ca acest lucru să se realizeze experimental, rulându-se algoritmul pentru date perturbate. Dacă dispersia rezultatelor este mare atunci algoritmul este instabil şi trebuie schimbat.
3. Dacă algoritmul este stabil, atunci acurateţea finală (modulul erorii relative) este majorată de produsul dintre numărul de condiţionare şi modulul reziduului relativ.

Despre un algoritm stabil care generează erori mici pentru probleme bine condiţionate se spune că este robust.



Notes

# Lectura obligatorie pentru această săptămână

- Erori - Cap.2 din

  [1] Gabriela Ciuprina, Mihai Rebican, Daniel Ioan - Metode numerice in ingineria electrica - Îndrumar de laborator pentru studenţii facultăţii de Inginerie electrică, Editura Printech, 2013, disponibil la http://mn.lmn.pub.ro/indrumar/IndrumarMN_Printech2013.pdf



Notes

Notes