Universitatea din Craiova

Facultatea de Ştiinţe

Specializarea: Matematică

Cicul I - Licență - ANUL II

Analiză numerică

Autor:

Prof. Dr. Sorin Micu

2026

Cursul 1

Noţiuni introductive de teoria erorilor

1.1 Surse de eroare

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

- *Erori de problemă (de modelare)*: Aceaste erori apar datorită faptului că modelul matematic ales pentru rezolvarea problemei nu descrie complet procesul real cercetat. Modelarea matematică a unui proces real presupune o serie de simplificări care reduc complexitatea lui şi permit formularea cantitativă printr-un obiect matematic (ecuaţie, sistem de ecuaţii, integrală, ecuaţie diferenţială, ecuaţie cu derivate parţiale, punct de minim al unei funcţionale, etc.)

- *Erori în datele de intrare*: Deseori procesul de modelare matematică se bazează pe rezultatele unor experimente, pe anumite seturi de mărimi numerice obţinute în urma măsurărilor (de exemplu: distanţa, masa, viteza) şi care nu sunt exacte. De asemenea, erori apar şi în urma folosirii unor constante aproximative(precum: g, π).

- *Erori de metodă:* se datorează metodelor matematice de aproximare (metode numerice) folosite. Este importantă cunoaşterea mărimii şi a evoluţiei acestor erori pentru a se asigura precizia dorită în fiecare caz.

- *Erori de rotunjire:* apar ca urmare a capacităţii finite de reţinere a datelor în calculator.

Erorile de modelare, în datele de intrare şi de rotunjire, care nu pot fi evitate, nu trebuie să crescă în urma tratamentului numeric ulterior prin metoda de aproximare. Această proprietate este cunoscută sub numele de *stabilitate* şi reprezintă unul dintre obiectivele principale în conceperea unei metode numerice de aproximare.

Fiind dat un număr real x şi o aproximare a sa $\widehat{x}$, se va numi **eroare absolută**

$$E_{abs}(\widehat{x}) = |x - \widehat{x}| \tag{1.1}$$

şi **eroare relativă**

$$E_{rel}(\widehat{x}) = \frac{|x - \widehat{x}|}{|x|}. \tag{1.2}$$

În calculul ştiinţific, unde răspunsurile la diferite probleme pot varia foarte mult în magnitudine, este de mai mare interes eroarea relativă decât eroarea absolută.

1.2 Reprezentarea numerelor reale în calculator (Standardul IEEE 754 din 1985 şi 2008)

Pentru reprezentarea numărului $x \in \mathbb{R}^*$ în calculator conform standardului IEEE 754 se parcurg următoarele etape:

1. Se scrie x în baza doi.
2. Se pune x în virgulă mobilă (flotantă) normalizată

   $$x = \pm d_0.d_1d_2... \times 2^p, \tag{1.3}$$

   p fiind puterea iar $d_1d_2...$ partea fracţionară (zecimală). Se consideră $d_0 \neq 0$ (sau, echivalent, $d_0 = 1$).
3. Semnul numărului x se reţine într-un bit (bit de semn): este 0 dacă $x > 0$ şi 1 dacă $x < 0$. Se acceptă atât $+0$ cât şi -0 dar ele sunt egale.
4. Partea zecimală a lui x (semnificantul, mantisa) se reţine în 23 biţi în precizie simplă sau în 52 biţi în precizie dublă. Fiecare cifră a părţii zecimale se reţine într-un bit.

   În mantisă vom avea ceva de tipul
   $$1.\tilde{d}_1\tilde{d}_2...\tilde{d}_r$$
   unde $r = 23$ în precizie simplă sau $r = 52$ în precizie dublă. În precizie simplă, pentru numerele normalizate, avem că:

   -cea mai mică mantisă este $1.0...00_2 = 1$.

   -cea mai mare mantisă este $1.11...1_2 = 1 + \frac{1}{2} + \frac{1}{2^2} + ... + \frac{1}{2^{23}} = 2 - \frac{1}{2^{23}}$.

   Cifra 1 situată in faţa punctului zecimal nu se reţine în calculator ("hidden bit").
5. Puterea p a lui x se reţine în 8 biţi în precizie simplă sau în 11 biţi în precizie dublă. Se reprezină de fapt $E+p$ (exponentul) unde E este un număr întreg ce depinde de reprezentare ("bias"="deviaţie", "deplasare").

   De exemplu, în precizie simplă $E = 1111111_2 = 2^7 - 1 = 127$ iar în precizie dublă $E = 1111111111_2 = 2^{10} - 1 = 1023$. În precizie simplă, pentru numere normalizate, vom avea:

   -cel mai mare exponent este $11111111_2 - 1 = 254$ ceea ce dă o putere maximă $p = 127$.

-cel mai mic exponent este $0+1=1$ ceea ce dă o putere minimă $p=-126$.

Să menţionăm că folosirea unui exponent deviat (biased) face să nu mai fie necesară reţinerea semnului puterii. Acest lucru conduce la o mai rapidă comparare a numerelor.

6. Valorile extreme ale exponentului (255 şi 0 în precizie simplă, 2047 şi 0 în precizie dublă) sunt rezervate pentru ∞ şi NaN, respectiv numere nenormalizate între care 0. Numerele nenormalizate sunt numere foarte mici, cu o putere mai mică decât cea mai mică putere reprezentabilă, şi prin urmare a căror reprezentare va conţine obligatoriu zerouri pe primele poziţii ale mantisei (deci nu verifică regula de reprezentare dată mai sus). Prin urmare, numerele nenormalizate sunt numere de forma $\pm 0.d_1d_2...d_{23}\times 2^{-126}$ cu $0\le d_i\le 1$. Cum spuneam, între aceste numere se află şi 0.

Precizie simplă: (32 biţi)

Semnul	Puterea+127	Mantisa
1 bit	8 biţi	23 biţi

Precizie dublă: (64 biţi)

Semnul	Puterea+1023	Mantisa
1 bit	11 biţi	52 biţi

Remarcaţi că se pot reprezenta în calculator în precizie simplă numere atât foarte mari, de ordin 2^{128}, cât şi numere foarte mici, de ordin 2^{-126}. Nu toate numerele între aceste două extreme sunt însă reprezentate în calculator. Chiar numere simple în sistemul zecimal nu sunt reprezentate exact în calculator. De exemplu,

$$x=0.1_{10}=0.0(0011)_2=1.1(0011)_2\times 2^{-4}$$

număr care are, în baza doi, o infinite de zecimale şi din care se reţin, în precizie simplă,

$$fl(x)=1.10011001100110011001101\times 2^{-4}.$$

Vom nota prin $\mathcal{F}$ mulţimea numerelor reprezentate în calculator în precizie simplă . Ea cuprinde toate numerele de forma $\pm 1.d_1d_2...d_{23}\times 2^p$ cu $-126\le p\le 127$ şi $0\le d_i\le 1$ (numerele normalizate) dar şi ∞, NaN şi numerele nenormalizate.

Dat un număr real $x=\pm 1.d_1d_2...d_pd_{p+1}...\times 2^p$ în virgulă flotantă normalizată, numerele din $\mathcal{F}$ cele mai apropiate de el vor fi

$$\pm 1.d_1d_2...d_{23}\times 2^p,\quad \pm\left(1.d_1d_2...d_{23}+\frac{1}{2^{23}}\right)\times 2^p.$$

Vom nota prin $fl(x)$ reprezentarea lui x în calculator care, în precizie simplă, este

$$fl(x)=\begin{cases}\begin{cases}1.d_1d_2...d_{23}\times 2^p & \text{dacă } d_{24}=0 \text{ (aproximare prin lipsă)}\\ \left(1.d_1d_2...d_{23}+\frac{1}{2^{23}}\right)\times 2^p & \text{dacă } d_{24}=1 \text{ (aproximare prin adaos)}\end{cases} & \text{dacă } x>0\\ \begin{cases}-\left(1.d_1d_2...d_{23}+\frac{1}{2^{23}}\right)\times 2^p & \text{dacă } d_{24}=0 \text{ (aproximare prin lipsă)}\\ -1.d_1d_2...d_{23}\times 2^p & \text{dacă } d_{24}=1 \text{ (aproximare prin adaos)}\end{cases} & \text{dacă } x<0.\end{cases}$$

Regula de mai sus are o excepţie, când x este situat exact la jumătatea distanţei între cele două aproximări posibile. În acest caz, $x = \pm 1.d_1d_2...d_{23}100... \times 2^p$, şi se ia

$$fl(x) = \begin{cases} \left\{ \begin{array}{ll} 1.d_1d_2...d_{22}d_{23} \times 2^p & \text{dacă } d_{23} = 0 \text{ (aproximare prin lipsă)} \\ \left(1.d_1d_2...d_{23} + \frac{1}{2^{23}}\right) \times 2^p & \text{dacă } d_{23} = 1 \text{ (aproximare prin adaos)} \end{array} \right\} \text{dacă } x > 0 \\ \left\{ \begin{array}{ll} -\left(1.d_1d_2...d_{23} + \frac{1}{2^{23}}\right) \times 2^p & \text{dacă } d_{23} = 1 \text{ (aproximare prin lipsă)} \\ -1.d_1d_2...d_{23} \times 2^p & \text{dacă } d_{23} = 0 \text{ (aproximare prin adaos)} \end{array} \right\} \text{dacă } x < 0. \end{cases}$$

Propoziţia 1.1. *Eroarea relativă comisă la reprezentarea numerelor în calculator în precizie simplă este mai mică decât*

$$\frac{1}{2^{24}} \approx 5.96 \times 10^{-8}. \tag{1.4}$$

Proof. Analiza fiind asemănătoare, vom considera doar cazul $x > 0$. Studiem două cazuri:

- Aproximare prin lipsă: $x \geq fl(x)$.
$$0 \leq \frac{x - fl(x)}{x} < 0.00...0d_{24}d_{25}... \leq 0.00...0011... = \frac{1}{2^{25}} + \frac{1}{2^{26}} + ... = \frac{1}{2^{24}}.$$
- Aproximare prin adaos: $x \leq fl(x)$.
$$0 \leq \frac{fl(x) - x}{x} < \frac{1}{2^{23}} - 0.00...0d_{24}d_{25}... \leq \frac{1}{2^{23}} - \frac{1}{2^{24}} = \frac{1}{2^{24}}.$$

□

Analog se obţine că

Propoziţia 1.2. *Eroarea relativă comisă la reprezentarea numerelor în calculator în precizie dublă este mai mică decât*

$$\frac{1}{2^{53}} \approx 1.11 \times 10^{-16}. \tag{1.5}$$

Am arătat că eroarea relativă la reprezentarea numerelor în calculator în precizie simplă este mai mică decât $\frac{1}{2^{24}} \approx 5.96 \times 10^{-8}$. Se pune întrebarea dacă această valoare este efectiv atinsă. Un răspuns parţial este dat la finalul următoarelor consideraţii.

Definiţia 1.2.1. *Numărul $\varepsilon > 0$ se numeşte* **zeroul maşinii** *dacă este cel mai mare număr pozitiv reprezentabil în maşină cu proprietatea că $1 + \varepsilon$ se reprezintă prin 1.*

Propoziţia 1.3. *Zeroul maşinii este $\frac{1}{2^{24}} \approx 5.96 \times 10^{-8}$ în precizie simplă şi $\frac{1}{2^{53}} \approx 1.11 \times 10^{-16}$ în precizie dublă.*

Proof. Numărul $1+\frac{1}{2^{24}}$ se reprezintă în bază doi 1.00...01, ultimul 1 fiind pe poziţia 24. Reprezentarea acestui număr (caz excepţional!) va fi prin urmare 1. Pe de altă parte orice număr mai mare decât $\frac{1}{2^{24}}$ reprezentabil în calculator face ca în primele 24 de zecimale să mai apară un 1. □

Observaţia 1.1. *Propoziţia precedentă ne arată că numărul $1 + \varepsilon$ se reprezintă în calculator cu o eroare absolută egală cu $\frac{1}{2^{24}}$. Eroarea relativă în acest caz va fi $\frac{1}{2^{24}(1+\varepsilon)} = \frac{1}{2^{24}+1} < \frac{1}{2^{24}}$.*

1.3 Studiul unor cazuri particulare

EXEMPLUL NUMERIC 1: Cum este reprezentat numărul 0.1 în precizie simplă? Care este eroarea la reprezentarea lui în precizie simplă? Vedeţi Figura 1.1.

1. Avem următoarea scriere în bază 2:

$$0.1_{10} = 0.0(0011)_2 = 1.1(0011)_2 \times 2^{-4}$$

2. Avem următoarea aproximare:

$$fl(0.1) = 1.10011001100110011001101 \times 2^{-4}$$

$$\begin{aligned} fl(0.1) &= \frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^8} + \frac{1}{2^9} + \frac{1}{2^{12}} + \frac{1}{2^{13}} + \frac{1}{2^{16}} + \frac{1}{2^{17}} + \frac{1}{2^{20}} + \frac{1}{2^{21}} + \frac{1}{2^{24}} + \frac{1}{2^{25}} + \frac{1}{2^{27}} \\ &= \frac{3}{2^5} + \frac{3}{2^9} + \frac{3}{2^{13}} + \frac{3}{2^{17}} + \frac{3}{2^{21}} + \frac{3}{2^{25}} + \frac{1}{2^{27}} = 3 \times \frac{\frac{1}{2^{25}}\frac{1}{2^4} - \frac{1}{2^5}}{\frac{1}{2^4} - 1} + \frac{1}{2^{27}} \\ &= 3 \times \frac{2^{24} - 1}{2^{25}(2^4 - 1)} + \frac{1}{2^{27}} = \frac{2^{26} \times 3 + 3}{2^{27} \times 15} = 0.1 + \frac{1}{2^{27} \times 5} = 0.1 + \frac{1}{671,088,640} \\ &= 0.1 + 0.000000001490116119384765625 \end{aligned}$$

3. Avem următoarea reprezentare în precizie simplă:

$$0.1 \to 0\ 01111011\ 10011001100110011001101$$

4. Eroarea de aproximare:

$$fl(0.1) - 0.1 = 0.000000001490116119384765625 \approx 1.5 \times 10^{-9}$$

$$\frac{|fl(0.1) - 0.1|}{0.1} \approx 1.5 \times 10^{-8} < 5.96 \times 10^{-8}$$

EXEMPLUL NUMERIC 2: Următorul calcul este făcut în MATLAB (vedeţi Figura 1.2):

$$1 - 3 * (4/3 - 1) = 2.220446049250313e - 016.$$

Care este explicaţia? Observaţi că eroarea la calculul numărului $3 * (4/3 - 1)$ este mai mare dacât eroarea relativă maximă în precizie dublă!

Cu un calcul exact, rezultatul operaţiilor $1 - 3 * (4/3 - 1)$ ar fi zero. Lucrând cu virgulă mobilă, obținem $\epsilon-$ul maşinii în precizie dublă. De ce? Se pare că singura aproximare are loc la împărțirea 4/3. Rezultatul nu poate fi exact 4/3 (decât în acel computer trinar rusesc numit SETUN, construit în 1958 la Universitatea de Stat din Moscova sub conducerea lui Serghei Sobolev și Nikolay Brusentsov). În consecință, valoarea stocată în $a = 4/3$ este apropiată de, dar nu exact egală cu, 4/3. Scăderea $b = a - 1$ produce o cantitate b al cărei ultimul bit este 0. Aceasta înseamnă că înmulțirea $3 * b$ se poate face fără nicio rotunjire. Valoarea stocată în $c = 3 * b$ nu este exact egală cu 1, și astfel valoarea stocată în e nu este 0. Înainte de standardul IEEE, acest cod a fost folosit ca o modalitate rapidă de a estima nivelul erorii de rotunjire pe diferite computere.

1. Avem următoarea scriere în bază 2:
$$\left(\frac{4}{3}\right)_{10} = 1.(01)_2$$
Într-adevăr, avem că: $0.(01)_2 = \frac{1}{2^2} + \frac{1}{2^4} + \cdots = \frac{1}{4}\frac{1}{1-\frac{1}{4}} = \frac{1}{3}$.

2. Avem următoarele aproximări în precizie dublă:
$$db(4/3) = 1.0101010101010101010101010101010101010101010101010101010101$$

$$db(4/3-1) = 1.0101010101010101010101010101010101010101010101010101010100 \times 2^{-2}$$
$$= \frac{1}{2^2} + \frac{1}{2^4} + \cdots + \frac{1}{2^{52}} = \frac{1}{4}\frac{\frac{1}{2^{52}}-1}{\frac{1}{4}-1} = \frac{1}{3}\left(1 - \frac{1}{2^{52}}\right)$$

$$db(3*(4/3-1)) = 1 - \frac{1}{2^{52}} = \frac{4503599627370495}{4503599627370496}$$
$$= 0.99999999999999977795539507496869191527366638183593 75$$

$$db(1-3*(4/3-1)) = \frac{1}{2^{52}} = \frac{1}{4503599627370496}$$
$$= 2.22044604925031308084726333618164062 5e-16$$

EXEMPLUL NUMERIC 3: Următoarea testare făcut în MATLAB oferă un răspuns negativ (vedeţi Figura 1.3):
$$0.7 + 0.1 == 0.8$$
$$\text{ans} = 0$$
Care este explicaţia?

Avem următoarele scrieri în bază 2:

$A = 0.7$

$A \to +1.0110011001100110011001100110011001100110011001100110011001100110_2 \times 2^{-1}$

$A \to 0\ 01111111110\ 0110011001100110011001100110011001100110011001100110$

$A \to 0.69999999999999995559107901499373838305473327636718 75$

$B = 0.1$

$B \to +1.1001100110011001100110011001100110011001100110011001100110011010_2 \times 2^{-4}$

$B \to 0\ 01111111011\ 1001100110011001100110011001100110011001100110011010$

$B \to 0.10000000000000000555111512312578270211815834045410156 25$

$C = 0.8$

$C \to +1.1001100110011001100110011001100110011001100110011001100110011101_2 \times 2^{-1}$

$C \to 0\ 01111111110\ 1001100110011001100110011001100110011001100110011001100110$**10**

$C \to 0.80000000000000004440892098500626161694526672363281 25$

Pentru adunarea a doua numere in format double se parcurg urmatoarele etape:

1. Se rescrie numărul mai mic astfel încât exponentul său să se potrivească cu exponentul numărului mai mare.
2. Se aduna mantisele
3. Se normalizeaza suma
4. Se rotunjeste suma

Tinand cont de aceste reguli se obtine:

$A + B \rightarrow 0\ 01111111110\ 1001100110011001100110011001100110011001100110011001100110011001$**01**

$A + B \rightarrow 0.79999999999999993338661852249060757458209991455078125$

```
>> a=0.1

a =

    0.1000

>> format long
>> a

a =

   0.10000000000000

>> a=single(0.1)

a =

   0.1000000

>> sprintf('%0.17f',a)

ans =

0.10000000149011612

>> num2strexact(a)

ans =

0.100000001490116119384765625
```

Figura 1.1: Studiul unor cazuri particulare. EXEMPLUL NUMERIC 1



Figura 1.2: Studiul unor cazuri particulare. EXEMPLUL NUMERIC 2

1.4 Epsilonul mașinii (opţional)

Epsilonul mașinii reprezintă o limită superioară a erorii relative de aproximare cauzată de rotunjirile în aritmetica cu virgulă mobilă. Această valoare are o importanță fundamentală în analiza numerică a operațiilor aritmetice din computer. Cantitatea se mai numește și *macheps*, având simbolul grecesc ε.

Există două definiții predominante. În analiza numerică, epsilonul mașinii depinde de tipul de rotunjire utilizat și este numit și rotunjire unitară, care are simbolul aldin roman u. Cu toate acestea, după o definiție mai puțin formală, dar utilizată pe scară largă, epsilonul mașinii este independent de metoda de rotunjire și poate fi echivalent cu u sau $2u$.

Rotunjirea este o procedură de alegere a reprezentării într-un format în virgulă mobilă a unui număr real. Pentru un format și o procedură de rotunjire, epsilonul mașinii este eroarea relativă maximă a procedurii de rotunjire alese.

Pentru un format în virgulă mobilă toate numerele cu același exponent, e, au spațierea, $b^{e-(p-1)}$. Spațierea se modifică la numerele care sunt de forma b^e; distanța între două numere consecutive mai mari decat b^e este de b ori mai mare decât distanța între două numere consecutive mai mici

```
2/16/24 9:22 PM                MATLAB Command Window                    1 of 1

>> format long
>> b=0.1

b =

   0.10000000000000

>> num2strexact(b)

ans =

0.1000000000000000055511151231257827021181583404541015625

>> a=0.7

a =

   0.70000000000000

>> num2strexact(a)

ans =

0.69999999999999995559107901499373830305473327636718750

>> c=0.8

c =

   0.80000000000000

>> num2strexact(c)

ans =

0.80000000000000004440892098500626161694526672363281250

>> d=a+b

d =

   0.80000000000000

>> num2strexact(d)

ans =

0.79999999999999993338661852249060757458209991455078125
```

Figura 1.3: Studiul unor cazuri particulare. EXEMPLUL NUMERIC 3

decat b^e.

Deoarece epsilonul mașinii este o limită pentru eroarea relativă, este suficient să luăm în considerare numerele cu exponent $e = 0$. De asemenea, este suficient să luăm în considerare numerele pozitive. Pentru tipul obișnuit de rotunjire la cel mai apropiat număr, eroarea de rotunjire absolută este cel mult jumătate din distanța între două numere consecutive, adică

$$\frac{1}{2}b^{-(p-1)}.$$

Această valoare este cel mai mare numărător posibil pentru eroarea relativă. Numitorul în eroarea relativă este numărul rotunjit, care ar trebui să fie cât mai mic posibil pentru a face eroarea relativă mare. Prin urmare, cea mai mare eroare relativă are loc atunci când rotunjirea este aplicată numerelor de forma $1 + a$ unde a este între 0 și $\frac{1}{2}b^{-(p-1)}$. Toate aceste numere se rotunjesc la 1 cu eroare relativă $\frac{a}{1+a}$. Maximul apare atunci când a se află la capătul superior al intervalului său. $1 + a$ din numitor este neglijabil în comparație cu numărător, așa că este lăsat deoparte din motive de oportunitate și deci $\frac{1}{2}b^{-(p-1)}$ este luat ca epsilonul mașinii. După cum s-a arătat aici, eroarea relativă este cea mai mare pentru numerele care se rotunjesc la 1, așa că epsilonul mașinii este numit și rotunjire unitară, ceea ce înseamnă aproximativ “eroarea maximă care poate apărea

prin rotunjirea la valoarea unitară 1".

Distanța maximă dintre un număr normalizat în virgulă mobilă, x, și un număr normalizat adiacent este $2\varepsilon|x|$.

Standardul IEEE nu definește termenii *epsilonul mașinii* și *unit roundoff*, așa că sunt utilizate definiții diferite ale acestor termeni, ceea ce poate provoca o oarecare confuzie. Definiția formală a epsilonului mașinii data mai sus este cea folosită de prof. James Demmel în notitele sale de curs si pachetul de algebră liniară LAPACK.

O definiție alternativă este mult mai răspândită în afara mediului academic: epsilonul mașinii este diferența dintre 1 și următorul număr normalizat scris in virgulă mobilă.

Prin această definiție, ε este egal cu unitatea de pe ultimul loc sau unitatea de cea mai mică precizie (ulp), adică valoarea pe care o reprezintă cifra cea mai puțin semnificativă dacă este 1. Astfel,

$$\varepsilon = b^{-(p-1)},$$

unde b este baza sistemului cu virgulă mobilă și p este precizia

Pe de alta parte, *unit roundoff* u va fi egala cu

$$u = \frac{\varepsilon}{2},$$

presupunând modul de aproximare este rotunjire la cel mai apropiat numar.

Prevalența acestei definiții este înrădăcinată prin utilizarea sa în standardul ISO C pentru constantele referitoare la tipurile în virgulă mobilă. Este, de asemenea, utilizata pe scară largă în software-ul de calcul științific și în literatura numerică și de calcul.

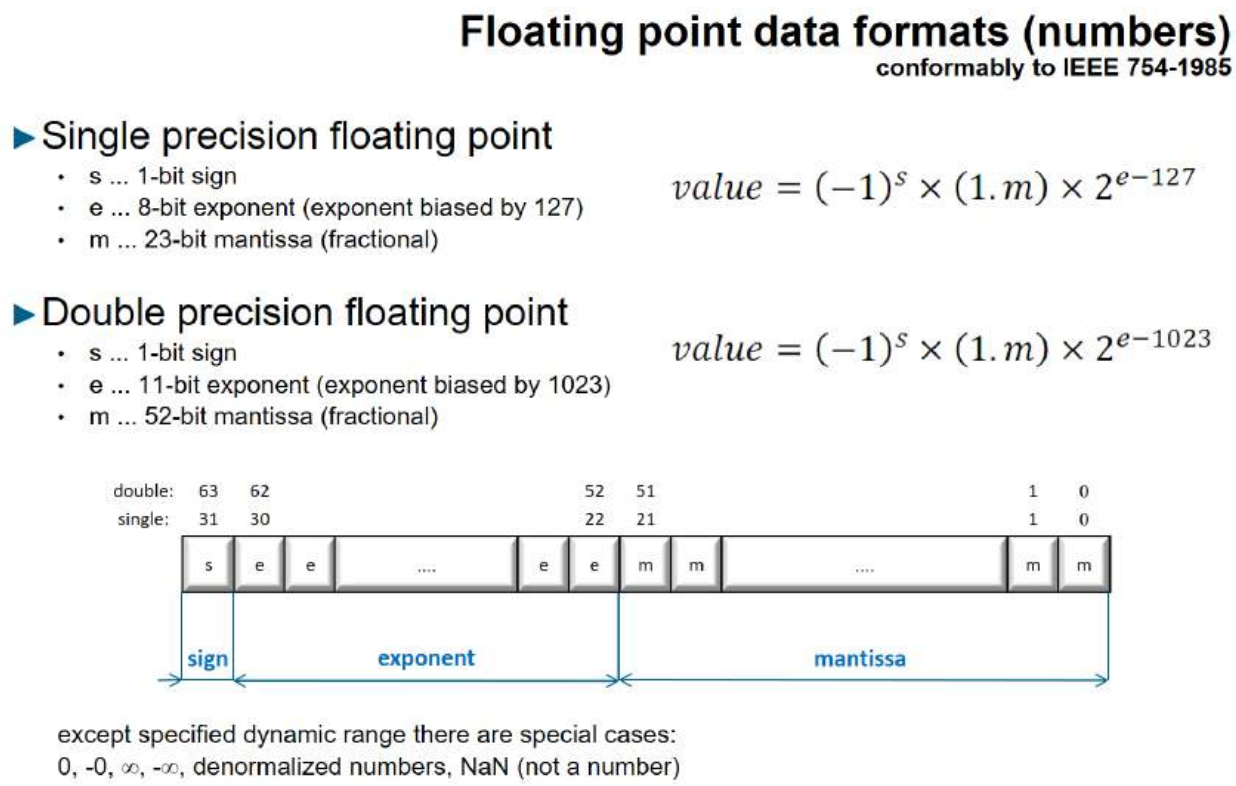


Figura 1.4: Formatele binare in IEEE 754

Cursul 2

Rezolvarea ecuaţiilor neliniare (I)

2.1 Introducere

Una dintre cele mai vechi şi necesare probleme matematice este aceea de determinare a rădacinilor ecuaţiilor. Pentru unele ecuaţii simple, cum ar fi ecuaţiile algebrice de grad mai mic sau egal cu patru, s-au găsit formule exacte pentru determinarea soluţiilor. În aceeaşi categorie intră şi unele ecuaţii trigonometrice sau exponenţiale.

Pentru cea mai mare parte a ecuaţiilor însă nu se dispune de formule exacte. De exemplu,

$$\frac{1}{2}\cos(x) = x, \quad \arctan(x) = x, \quad e^x - 5x^2 = 0,$$

$$(x - 0.6)(x - 1.3)^2(x - 2)^3 + 0.1234\ln(x) = 0, \quad \tan(x) = x.$$

În acest caz nu avem altă opţiune decât să încercăm să aproximăm rădăcinile acestor ecuaţii. Ca procedeu general, pornim cu o aproximare iniţială (pe care o putem lua la întâmplare şi care poate fi foarte proastă) şi folosim o formulă dată de ecuaţie pentru a obţine o a doua aproximare. Aceeaşi formulă cu a doua aproximare ne dă cea de a treia aproximare s.a.m.d. Aceasta este o schemă iterativă (deoarece pentru calculul celei de-a doua aproximări se foloseşte prima, pentru calculul celei de-a treia se foloseşte a doua aproximare s.a.m.d). Vom opri acest proces când am atins precizia dorită.

Schema iterativă ideală ar trebui să ofere un şir de aproximări care să conveargă (aproximări din ce în ce mai bune) cu rapiditate (viteză) şi indiferent de valoarea aleasă iniţial (siguranţă). În practică toate aceste lucruri se întâlnesc extrem de rar.

Nu exisă o reţetă care să garanteze întotdeauna reuşita. Putem însă stabili unele principii utile:

- Este foarte folositoare localizarea în prealabil a rădăcinilor. Acest lucru se realizează prin studiu matematic sau grafic.
- Se încearcă găsirea unei aproximaţii iniţiale cât mai apropiate de rădăcina căutată. Acest lucru poate fi uşurat de studiul anterior sau de folosirea unei metode care converge (eventual lent) pentru orice aproximare iniţială.
- Dacă suntem suficient de aproape de rădăcină vom dispune de metode care converg suficient de rapid.

Să notăm însă că aceşti paşi nu pot fi realizaţi întotdeauna. Studiul cazurilor dificile depinde de cunoştinţele, experienţa şi uneori intuiţia programatorului. Analiza numerică este ştiinţă dar şi artă!

2.2 Metoda aproximaţiilor succesive

Fie o funcţie reală $g : [a, b] \to \mathbb{R}$.

Definiţia 2.1. *Numărul $x^* \in [a, b]$ se numeşte* **punct fix** *al funcţiei g dacă $g(x^*) = x^*$.*

Teorema 2.1. *Presupunem că funcţia $g : [a, b] \to \mathbb{R}$ verifică proprietăţile:*

1. *$g([a, b]) \subseteq [a, b]$,*
2. *există $q \in (0, 1)$ astfel încât $|g(x) - g(y)| \leq q|x - y|$ (spunem că g este o $q-$contracţie).*

Există un singur punct fix $x^* \in [a, b]$ al lui g în $[a, b]$.

Proof.

- Existenţa: Construim funcţia $h : [a, b] \to \mathbb{R}$, $h(x) = g(x) - x$.
  Funcţia h este continuă pe $[a, b]$ şi $h(a) = g(a) - a \geq 0$, $h(b) = g(b) - b \leq 0$. Prin urmare există un $x^* \in [a, b]$ astfel încât $h(x^*) = 0$. Demonstraţia existenţei punctului fix al lui g se încheie.
- Unicitatea: Fie x^* şi y^* două puncte fixe ale lui g. Avem că
  $$|x^* - y^*| = |g(x^*) - g(y^*)| \leq q|x^* - y^*|$$
  şi prin urmare
  $$(1 - q)|x^* - y^*| \leq 0.$$
  Rezultă că $x^* = y^*$ şi demonstraţia unicităţii se încheie.

□

Fie $x^0 \in [a, b]$ arbitrar şi construim şirul definit prin recurenţă:

$$x^{n+1} = g(x^n), \quad n \geq 0. \tag{2.1}$$

După cum vom vedea mai jos, şirul $(x^n)_{n\geq 0}$ generat de (2.1) va converge în anumite condiţii la un punct fix al funcţiei g. Metoda aceasta de aproximare a punctelor fixe se numeşte *metoda aproximaţiilor succesive* sau *metoda iterativă de punct fix.*

Teorema 2.2. *Dacă funcţia g verifică cele două proprietăţi din Teorema 2.1 atunci şirul dat de (2.1) converge la x^*, unicul punct fix al lui g în $[a, b]$. În plus, avem că*

$$|x^n - x^*| \leq \frac{q^n}{1 - q}|x^1 - x^0|, \quad n \geq 1 \tag{2.2}$$

$$|x^n - x^*| \leq \frac{q}{1 - q}|x^n - x^{n-1}|, \quad n \geq 1. \tag{2.3}$$

Proof. Avem următoarele relaţii

$$|x^k - x^*| = |g(x^{k-1}) - g(x^*)| \leq q|x^{k-1} - x^*| \leq ... \leq q^k|x^0 - x^*|.$$

Rezultă că şirul $(x^k)_{k\geq 0}$ este convergent la x^*.

Pentru demonstrarea relaţiilor (2.2) şi (2.3) să notăm, în primul rând că

$$|x^{k+1} - x^k| = |g(x^k) - g(x^{k-1})| \leq q|x^k - x^{k-1}| \leq ... \leq q^k|x^1 - x^0|,$$
$$|x^{n+k} - x^{n+k-1}| = |g(x^{n+k-1}) - g(x^{n+k-2})| \leq q|x^{n+k-1} - x^{n+k-2}| \leq ... \leq q^k|x^n - x^{n-1}|.$$

Deducem că

$$\begin{aligned}|x^{n+p} - x^n| &= |x^{n+p} - x^{n+p-1} + x^{n+p-1} - x^{n+p-2} + ... + x^{n+1} - x^n| \leq \\ &\leq |x^{n+p} - x^{n+p-1}| + |x^{n+p-1} - x^{n+p-2}| + ... + |x^{n+1} - x^n| \leq \\ &\leq q^{n+p-1}|x^1 - x^0| + q^{n+p-2}|x^1 - x^0| + ... + q^n|x^1 - x^0| = \\ &= \frac{q^{n+p} - q^n}{q-1}|x^1 - x^0| \leq \frac{q^n}{1-q}|x^1 - x^0|.\end{aligned}$$

Facând p să tindă la infinit se obţine (2.2).

Pe de altă parte

$$\begin{aligned}|x^{n+p} - x^n| &= |x^{n+p} - x^{n+p-1} + x^{n+p-1} - x^{n+p-2} + ... + x^{n+1} - x^n| \leq \\ &\leq |x^{n+p} - x^{n+p-1}| + |x^{n+p-1} - x^{n+p-2}| + ... + |x^{n+1} - x^n| \leq \\ &\leq q^p|x^n - x^{n-1}| + q^{p-1}|x^n - x^{n-1}| + ... + q|x^n - x^{n-1}| = \\ &= \frac{q^{p+1} - q}{q-1}|x^n - x^{n-1}| \leq \frac{q}{1-q}|x^n - x^{n-1}|.\end{aligned}$$

Facând p să tindă la infinit se obţine (2.3). □

Observaţia 2.1. *Prima evaluare a erorii (2.2) ne arată că viteza de convergenţă a metodei aproximaţiilor succesive este cea cu care q^n converge la zero. De asemenea, să notăm că formula (2.2) ne oferă o metodă de estimare a numărului de iteraţii necesare obţinerii unei precizii date: fixând o eroare ε, la pasul n avem $|x^n - x^*| \leq \varepsilon$ dacă*

$$\frac{q^n}{1-q}|x^1 - x^0| \leq \varepsilon \Leftrightarrow n \geq \log \frac{\varepsilon(1-q)}{|x^1 - x^0|} / \log q. \tag{2.4}$$

Observaţia 2.2. *Evaluarea (2.3) ne dă un criteriu de oprire a iteraţiilor. Dată o eroare $\varepsilon > 0$, vom opri algoritmul dacă dacă $|x^n - x^{n-1}| < \varepsilon$ (caz în care eroarea va fi mai mică decât $\varepsilon\frac{q}{1-q}$).*

Observaţia 2.3. *Dacă g este derivabilă în $[a,b]$ şi $|g'(x)| \leq q < 1$ pentru orice $x \in [a,b]$, atunci g este o $q-$contracţie pe $[a,b]$. Într-adevăr, din Teorema lui Lagrange, pentru orice $x, y \in [a,b]$ există $\xi \in [a,b]$ astfel încât*

$$|g(x) - g(y)| = |g'(\xi)|\,|x - y| \leq q|x - y|.$$

Prin urmare, g este o contracţie pe $[a,b]$.

Observaţia 2.4. *Dacă g este o $q-$contracţie pe $[x^0 - r, x^0 + r]$, unde $x^0 \in \mathbb{R}$ iar $r > 0$, astfel încât*

$$|g(x^0) - x^0| \leq (1-q)r$$

atunci g are un punct fix unic $x^* \in [x^0 - r, x^0 + r]$ iar şirul $(x^n)_{n\geq 0}$ dat de (2.1) converge la x^*.

Conform Teoremei 2.2 este suficient să arătăm că $g([x^0 - r, x^0 + r]) \subseteq [x^0 - r, x^0 + r]$. Fie $x \in [x^0 - r, x^0 + r]$. Avem că

$$|g(x) - x^0| \leq |g(x) - g(x^0)| + |g(x^0) - x^0| \leq q|x - x^0| + (1-q)r \leq qr + (1-q)r = r.$$

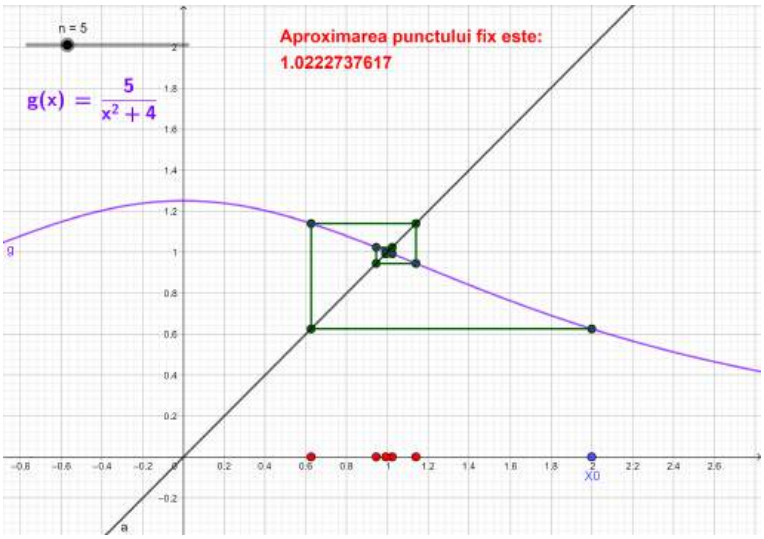


Figura 2.1: Evoluţia iteraţiilor în Exemplul 1

2.3 Exemplul 1

Considerăm ecuaţia

$$x^3 + 4x - 5 = 0$$

care are o unică rădăcină în $\mathbb{R}$, $x^* = 1$.

O vom aproxima cu metoda aproximaţiilor succesive considerând algoritmul

$$x^0 = 2, \quad x^{n+1} = g(x^n),$$

unde

$$g(x) = \frac{5}{x^2 + 4}.$$

- Funcţia g verifică ambele condiţii din Teorema 2.2 pe intervalul $[0, 2]$:

$$g([0,2]) = [g(2), g(0)] = \left[\frac{5}{8}, \frac{5}{4}\right] \subset [0,2]$$

$$|g'(x)| = \left|-\frac{10x}{(x^2+4)^2}\right| := \tilde{g} \leq q < 1.$$

$$\tilde{g}'(x) = \frac{30x^2 - 40}{(x^2+4)^3} \quad \Rightarrow \quad \max_{x\in[0,2]} |g'(x)| = |g'(\sqrt{4/3})| := q.$$

- Primele trei iteraţii sunt date de:

$$x^0 = 2, \; x^1 = \frac{5}{8} = 0.625, \; x^2 = \frac{320}{281} = 1.13879...$$

$$x^3 = 0.943958..., \; x^4 = 1.02227..., \; x^5 = 0.99107...$$

Cursul 3

Metodele de bracketing

Fie funcţia reală $f : [a, b] \to \mathbb{R}$ cu următoarele proprietăţi:

- f este continuă în $[a, b]$
- $f(a)f(b) < 0$

În ipotezele de mai sus rezultă că există cel puţin o rădăcină a funcţiei f în (a, b). Următoarele două metode îşi propun să construiască un şir descrescător de intervale $[a^n, b^n]$ (brackets) care conţin o rădăcină a lui f. Dacă lungimea acestor intervale tinde la zero, obţinem aproximaţii din ce în ce mai bune ale unei rădăcini a lui f. Chiar dacă lungimea intervalelor nu tinde la zero, putem aproxima cu succes o rădăcină a lui f printr-unul dintre capetele intervalelor.

3.1 Metoda bisecţiei

Fie funcţia reală $f : [a, b] \to \mathbb{R}$ continuă şi având proprietatea că $f(a)f(b) < 0$. Rezultă că există cel puţin o rădăcină $x^* \in [a, b]$. Să presupunem că x^* este unica rădăcină în $[a, b]$ a funcţiei f. Vom defini următorul şir:

$$
(3.1)\quad \begin{cases} a^0 = a,\ b^0 = b,\ c^0 = \frac{a^0+b^0}{2} \\ i)\quad \text{Dacă } f(a^n)\, f(c^n) < 0 \text{ atunci } a^{n+1} = a^n,\ b^{n+1} = c^n,\ c^{n+1} = \frac{a^{n+1}+b^{n+1}}{2}, \quad n \geq 0 \\ ii)\quad \text{Dacă } f(a^n)\, f(c^n) > 0 \text{ atunci } a^{n+1} = c^n,\ b^{n+1} = b^n,\ c^{n+1} = \frac{a^{n+1}+b^{n+1}}{2}, \quad n \geq 0 \\ iii)\quad \text{Dacă } f(a^n)\, f(c^n) = 0 \text{ atunci } x^* = c^n. \end{cases}
$$

Rezultatul următor stabileşte convergenţa metodei şi oferă o estimare a erorii la fiecare pas.

Teorema 3.1. *Fie funcţia $f : [a, b] \to \mathbb{R}$ continuă şi cu o singură rădăcină $x^* \in [a, b]$. Presupunem că $f(a)f(b) < 0$ şi construim şirul $(c^n)_{n\geq 0}$ ca în (3.1). Avem că*

$$
(3.2)\qquad |x^* - c^n| \leq \frac{b-a}{2^{n+1}}, \quad n \geq 0.
$$

Proof. Prin inducţie demonstrăm că rădăcina x^* aparţine intervalului $[a^n, b^n]$ care are lungimea $\frac{b-a}{2^n}$. De aici ar rezulta că

$$|x^* - c^n| = \left| x^* - \frac{a^n + b^n}{2} \right| \leq \frac{1}{2}|b^n - a^n| = \frac{b-a}{2^{n+1}}.$$

Pentru $n = 0$ avem că x^* aparţine intervalului $[a^0, b^0]$ care are lungimea $\frac{b-a}{2^0} = b - a$.
Presupunem că proprietatea este adevărată pentru n şi o demonstrăm pentru $n + 1$.
În cazul i) rădăcina este în $(a^n, c^n) = (a^{n+1}, b^{n+1})$ care are lungimea $\frac{b-a}{2^{n+1}}$.
În cazul ii) rădăcina este în $(c^n, b^n) = (a^{n+1}, b^{n+1})$ care are lungimea $\frac{b-a}{2^{n+1}}$.
În cazul iii) rădăcina coincide cu c^n şi prin urmare este în $[a^n, c^n] = [a^{n+1}, b^{n+1}]$. Demonstraţia se încheie. □

3.2 Regula Falsi

Vom defini din aproape în aproape următorul şir:

$$(3.3) \quad \begin{cases} a^0 = a,\ b^0 = b, \\ c^0 = \text{ intersecţia dreptei ce trece prin punctele } (a^0, f(a^0)) \text{ şi } (b^0, f(b^0)) \text{ cu axa } Ox \\ \\ i) \quad \text{Dacă } f(a^n)\, f(c^n) < 0 \text{ atunci } a^{n+1} = a^n, \quad b^{n+1} = c^n \\ c^{n+1} = \text{ intersecţia dreptei ce trece prin punctele } (a^n, f(a^n)) \text{ şi } (c^n, f(c^n)) \text{ cu axa } Ox \\ \\ ii) \quad \text{Dacă } f(a^n)\, f(c^n) > 0 \text{ atunci } a^{n+1} = c^n, \quad b^{n+1} = b^n \\ c^{n+1} = \text{ intersecţia dreptei ce trece prin punctele } (c^n, f(c^n)) \text{ şi } (b^n, f(b^n)) \text{ cu axa } Ox \\ \\ iii) \quad \text{Dacă } f(c^n) = 0 \text{ atunci } x^* = c^n. \end{cases}$$

Reamintim că ecuaţia dreptei ce trece prin punctele (x_1, y_1) şi (x_2, y_2) este dată de:

$$(3.4) \qquad y - y_2 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_2).$$

Pentru găsirea lui c^0 intersectăm dreptele

$$\begin{cases} y - f(b) = \frac{f(b)-f(a)}{b-a}(x - b) \\ y = 0 \end{cases}$$

şi obţinem formula

$$(3.5) \qquad c^0 = b - \frac{b-a}{f(b) - f(a)} f(b) = \frac{f(b)}{f(b) - f(a)} a - \frac{f(a)}{f(b) - f(a)} b.$$

Să observăm că dacă $f(a) < 0 < f(b)$ expresia de mai sus este o combinaţie convexă a numerelor a şi b, deci c^0 va fi între ele.

Dacă vom fi în cazul iii) al algoritmului de mai sus, avem rădăcina exactă şi ne oprim. Vom presupune în cele ce urmează că nu vom întâlni acest caz. Algoritmul continuă cu determinarea

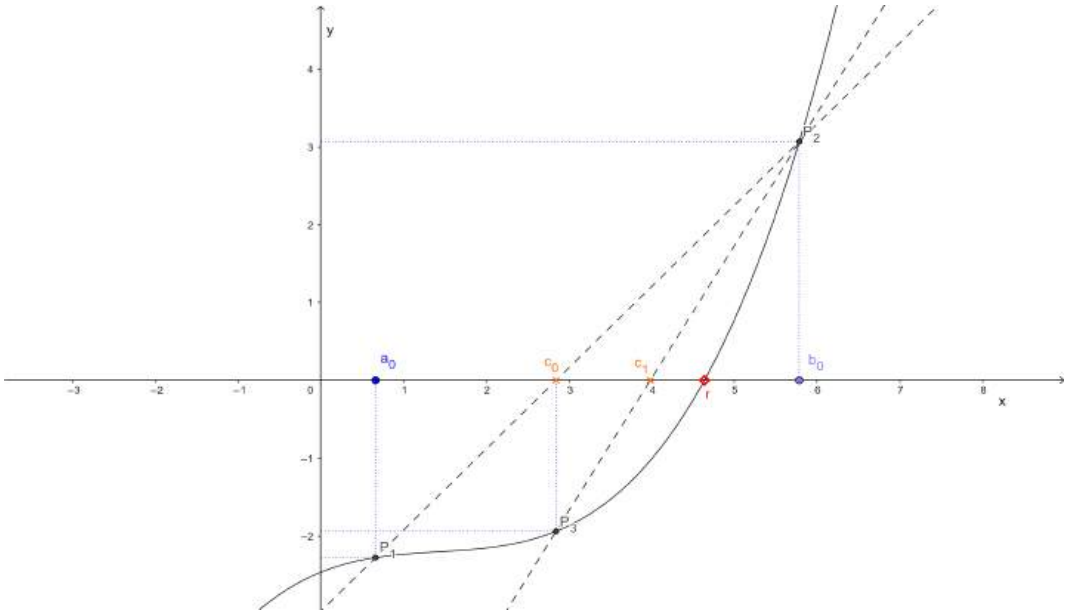


Figura 3.1: Metoda Regula Falsi

intervalelor succesive $[a^n, b^n]$. Fiecare interval $[a^n, b^n]$ va conţine o rădăcină a lui f, pentru $n \geq 0$. Să observăm că, la fiecare iteraţie, unul dintre capetele intervalului $[a^n, b^n]$ se schimbă, fiind înlocuit de noul punct c^n găsit la iteraţia respectivă. În anumite condiţii extremitatea din dreapta a tuturor intervalelor poate rămâne punctul b.

Lema 3.1. *În următoarele ipoteze asupra funcţiei f:*

- *f este continuă în $[a, b]$*
- *$f(a) < 0 < f(b)$*
- *f este convexă în $[a, b]$*
- *c^n nu este o rădăcină a lui f,*

avem că c^{n+1} va fi dat de intersecţia dreptei ce trece prin punctele $(c^n, f(c^n))$ şi $(b, f(b))$ cu axa Ox şi prin urmare este dat de formula

$$c^{n+1} = \frac{f(b)}{f(b) - f(c^n)} c^n + \frac{f(c^n)}{f(c^n) - f(b)} b \qquad (n \geq 0). \tag{3.6}$$

Proof. Cum în ipotezele date $f(b) > 0$, este suficient să arătăm că

$$f(c^n) \leq 0 \qquad (n \geq 0). \tag{3.7}$$

Pentru $n = 0$ această proprietate este adevărată datorită convexităţii lui f:

$$f(c^0) = f\left(\frac{f(b)}{f(b) - f(a)} a + \frac{f(a)}{f(a) - f(b)} b\right) \leq \frac{f(b)}{f(b) - f(a)} f(a) + \frac{f(a)}{f(a) - f(b)} f(b) = 0.$$

Presupunem că proprietatea este adevărată pentru n şi o demonstrăm pentru $n+1$. Conform formulei (3.6) pentru c^{n+1}, folosind convexitatea lui f, avem că

$$f(c^{n+1}) = f\left(\frac{f(b)}{f(b)-f(c^n)}c^n + \frac{f(c^n)}{f(c^n)-f(b)}b\right) \leq \frac{f(b)}{f(b)-f(c^n)}f(c^n) + \frac{f(c^n)}{f(c^n)-f(b)}f(b) = 0.$$

Deci, $f(c^{n+1}) \leq 0$ şi demonstraţia se încheie. □

Observaţia 3.1. *Dacă a doua condiţie este înlocuită cu*

$$f(b) < 0 < f(a),$$

atunci capătul din stânga, a, va rămâne constant.

Observaţia 3.2. *Să observăm că formula* (3.6) *este echivalentă cu*

$$c^{n+1} = c^n - \frac{f(c^n)}{f(b)-f(c^n)}(b-c^n) \qquad (n \geq 0). \tag{3.8}$$

În condiţiile precedente putem demonstra convergenţa schemei.

Teorema 3.2. *Fie $f : [a,b] \to \mathbb{R}$ o funcţie care verifică ipotezele din Lema 3.1. Dacă $(c^n)_{n\geq 0}$ este şirul dat de regula falsi* (3.6) *atunci f are o rădăcină $x^* \in [a,b]$ şi avem că*

$$\lim_{n\to\infty} c^n = x^*. \tag{3.9}$$

Proof. Din formula (3.6), prin inducţie, deducem că $c^{n+1} < b$:

$$c^{n+1} = \frac{f(b)}{f(b)-f(c^n)}c^n + \frac{f(c^n)}{f(c^n)-f(b)}b < \frac{f(b)}{f(b)-f(c^n)}b + \frac{f(c^n)}{f(c^n)-f(b)}b = b.$$

Pe de alta parte

$$c^{n+1} = \frac{f(b)}{f(b)-f(c^n)}c^n + \frac{f(c^n)}{f(c^n)-f(b)}b \geq \frac{f(b)}{f(b)-f(c^n)}c^n + \frac{f(c^n)}{f(c^n)-f(b)}c^n = c^n.$$

Prin urmare şirul $(c^n)_{n\geq 0}$ este mărginit superior şi crescător, deci convergent la $x^* \in [a,b]$. Trecând la limită în relaţia (3.6) obţinem că

$$x^* = \frac{f(b)}{f(b)-f(x^*)}x^* + \frac{f(x^*)}{f(x^*)-f(b)}b,$$

ceea ce este echivalent cu

$$f(x^*)(b-x^*) = 0.$$

Cum $x^* \neq b$ (deoarece $f(x^*) = \lim_{n\to\infty} f(c^n) \leq 0 < f(b)$) ultima relaţie implică $f(x^*) = 0$. Deci x^* este o rădăcină a lui f în $[a,b]$. □

Cursul 4

Metoda lui Newton

Ne propunem să aproximăm o rădăcină z a ecuaţiei

$$f(x) = 0 \tag{4.1}$$

prin metoda lui Newton (a tangentei). Metoda de aproximare constă în generarea, prin recurenţă, a următorului şir :

$$\begin{cases} x^0 \text{ dat }, \\ x^{n+1} = x^n - \dfrac{f(x^n)}{f'(x^n)}, \quad n \geq 0. \end{cases} \tag{4.2}$$

Obţinerea primei iteraţii cu metoda Newton este ilustrată geometric în Figura 4.1. Fiind dat x^0, se intersectează dreapta tangentă la graficul funcţiei f în punctul $(x^0, f(x^0))$ cu axa orizontală şi de obţine x^1:

$$\begin{cases} y - f(x^0) = f'(x^0)(x - x^0) \\ y = 0. \end{cases}$$

Soluţia sistemului de mai sus este:

$$x^1 = x^0 - \frac{f(x^0)}{f'(x^0)},$$

adică exact ceea ce ne dă formula (4.2) pentru $n = 0$.

În următoarea teoremă vom vedea în ce condiţii metoda lui Newton converge şi care este viteza ei de convergenţă.

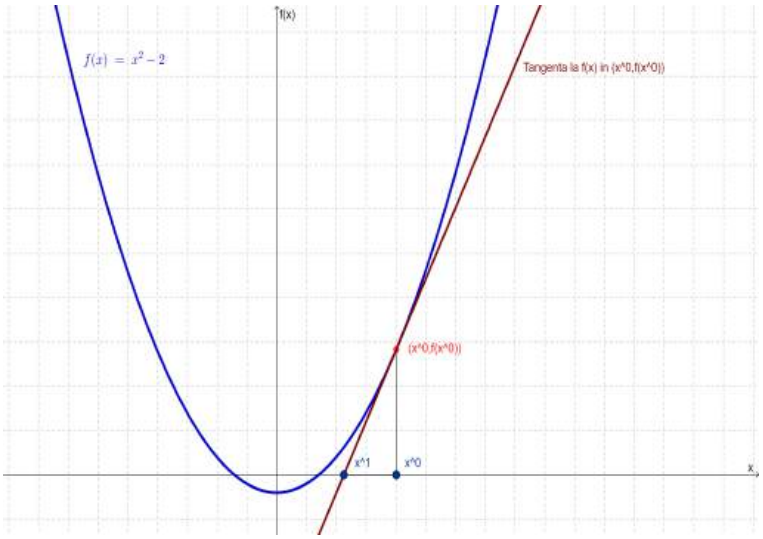


Figura 4.1: Obţinerea primei iteraţii cu metoda Newton

Teorema 4.1. *Fie $f \in C^2[a,b]$ o funcţie cu următoarele proprietăţi:*

i) $f(a)f(b) < 0$;

ii) $|f'(x)| \geq m > 0, \quad \forall x \in [a,b]$;

iii) $|f''(x)| \leq M, \quad \forall x \in [a,b]$.

În aceste condiţii există o unică rădăcină z a lui f în (a,b). Pentru orice $q \in (0,1)$, există o vecinătate $V = (z-r, z+r) \subset [a,b]$ a lui z, astfel încât, pentru orice $x^0 \in V$, şirul (4.2) este definit şi converge la z. În plus,

$$|z - x^n| \leq \frac{2m}{M} q^{2^n}, \quad n \geq 1. \tag{4.3}$$

$$|z - x^n| \leq \frac{1}{m}|f(x^n)| \leq \frac{M}{2m}|x^n - x^{n-1}|^2, \quad n \geq 1. \tag{4.4}$$

Proof.

Etapa 1 (Existenţa şi unicitatea rădăcinii.) Condiţia i) şi continuitatea lui f implică existenţa unei rădăcini $z \in (a,b)$. Cum f' este continuă şi nu se anulează (conform lui ii)), rezultă că are semn constant pe $[a,b]$. Rezultă că f este strict monotonă pe $[a,b]$, de unde deducem unicitatea rădăcinii z.

Etapa 2 (Analiza apriori.) Vom presupune că am găsit x^{n-1} şi vedem în ce condiţii am putea construi x^n. Din formula lui Taylor

$$0 = f(z) = f(x^{n-1}) + f'(x^{n-1})(z - x^{n-1}) + \frac{1}{2}f''(\xi)(z - x^{n-1})^2$$

şi deducem că

$$\left| z - \underbrace{\left(x^{n-1} - \frac{f(x^{n-1})}{f'(x^{n-1})} \right)}_{x^n} \right| \leq \frac{1}{2} \left| \frac{f''(\xi)}{f'(x^{n-1})} \right| |z - x^{n-1}|^2$$

şi prin urmare

$$|z - x^n| \leq \frac{M}{2m} |z - x^{n-1}|^2. \tag{4.5}$$

Etapa 3 (Definim intervalul de referinţă V şi demonstrăm corecta definire a şirului.) Fie acum $r > 0$ astfel încât

1. $V = (z - r, z + r) \subset [a, b]$
2. $\frac{M}{2m} r \leq q < 1$.

Se demonstrează prin inducţie că $(x^n)_{n \geq 0}$ este bine definit şi conţinut în V dacă $x^0 \in V$. Într-adevăr, dacă x^{n-1} a fost construit şi aparţine lui V, atunci $f'(x^{n-1}) \neq 0$ şi x^n poate fi, de asemenea, construit. Din (4.5) deducem că

$$|z - x^n| \leq \frac{M}{2m} |z - x^{n-1}|^2 \leq \frac{M}{2m} r^2 \leq qr < r$$

şi prin urmare $x^n \in V$. De asemenea, se verifică (4.5).

Etapa 4 (Demonstrăm convergenţa şirului şi (4.3).) Prin iterarea lui (4.5) se obţine

$$|z - x^n| \leq \frac{M}{2m} |z - x^{n-1}|^2 \leq \frac{M}{2m} \left(\frac{M}{2m} \right)^2 |z - x^{n-2}|^4 \leq \frac{M}{2m} \left(\frac{M}{2m} \right)^2 \ldots \left(\frac{M}{2m} \right)^{2^{n-1}} |z - x^0|^{2^n} =$$

$$= \left(\frac{M}{2m} \right)^{\frac{2^n - 1}{2-1}} |z - x^0|^{2^n} = \frac{2m}{M} \left(\frac{M}{2m} |z - x^0| \right)^{2^n} \leq \frac{2m}{M} \left(\frac{Mr}{2m} \right)^{2^n} \leq \frac{2m}{M} q^{2^n}.$$

De aici rezultă atât convergenţa şirului $(x^n)_{n \geq 0}$ la z cât şi (4.3).

Etapa 5 (Demonstrăm prima inegalitate din (4.4).) Cum f' are semn constant pe $[a, b]$, pentru orice $x, y \in [a, b]$ avem că:

$$|f(x) - f(y)| = \left| \int_x^y f'(s) \, \mathrm{d}s \right| = \int_x^y \left| f'(s) \right| \, \mathrm{d}s \geq m |y - x|.$$

Aplicân această formulă lui $x = z$ şi $y = x^n$, rezultă imediat că

$$|z - x^n| \leq \frac{1}{m} |f(x^n)|,$$

şi prima parte a lui (4.4) este demonstrată.

Etapa 6 (Demonstrăm a doua inegalitate din (4.4)**.)** Pentru a doua parte a lui (4.4) aplicăm formula lui Taylor astfel

$$f(x^n) = f(x^{n-1}) + f'(x^{n-1})(x^n - x^{n-1}) + \frac{1}{2}f''(\zeta)(x^n - x^{n-1})^2$$

şi deducem că

$$f(x^n) = f'(x^{n-1})\left(f(x^{n-1})/f'(x^{n-1}) + x^n - x^{n-1}\right) + \frac{1}{2}f''(\zeta)(x^n - x^{n-1})^2$$

adică

$$f(x^n) = \frac{1}{2}f''(\zeta)(x^n - x^{n-1})^2.$$

Prin urmare

$$|f(x^n)| \leq \frac{M}{2}|x^n - x^{n-1}|^2$$

şi a doua inegalitate din (4.4) este demonstraă. Demonstraţia teoremei se încheie.

□

Observaţia 4.1. *Prima evaluare a erorii (4.3) ne arată că viteza de convergenţă a metodei Newton este cea cu care q^{2^n} converge la zero. Aceasta este cu mult mai mare decât viteza metodei de punct fix, care era în general q^n. De asemenea, să notăm că formula (4.5) ne arată că, dacă la pasul $n-1$ avem o eroare ϵ, atunci la pasul următor, eroarea este mai mică decât $\frac{M}{2m}\epsilon^2$. Acest lucru implică următorul fapt: dacă la pasul $n-1$ avem calculate k zecimale exacte ale rădăcinii, atunci la pasul următor avem practic $2k$ zecimale exacte. O metodă cu astfel de caracteristici se numeşte de ordin doi.*

Observaţia 4.2. *Evaluările (4.4) ne dau două criterii de oprire a iteraţiilor. Dat $\varepsilon > 0$, vom opri algoritmul dacă $f(x^n) < \varepsilon$ (caz în care eroarea va fi mai mică decât ε/m) sau dacă $|x^n - x^{n-1}| < \varepsilon$ (caz în care eroarea va fi mai mică decât $\varepsilon^2 M/(2m)$).*

Observaţia 4.3. *Metoda Newton este una dintre cele mai folosite metode de aproximare a rădăcinilor ecuaţiilor neliniare datorită vitezei mari de convergentă. Ea are însă şi unele dezavantaje:*

1. *pentru a avea asigurată convergenţa trebuie să luăm aproximarea iniţială suficient de aproape de rădăcină. În demonstraţia teoremei se indică modalitatea de construire a vecinătăţii V. Aceasta poate fi foarte mică iar găsirea ei presupune un studiu matematic detaliat al funcţiei f.*

2. *pentru construirea iteraţiilor este nevoie de calculul derivatei. Acest lucru este cu atât mai laborios dacă dorim rezolvarea unor sisteme de ecuaţii. S-au propus mai multe metode care evită calculul derivatei:*

$$x^0 \in [a,b], \quad x^{n+1} = x^n - af(x^n), \tag{4.6}$$

(Metoda Newton modificată sau a liniilor paralele),

$$x^0, x^1 \in [a,b], \quad x^{n+1} = x^n - \frac{x^n - x^{n-1}}{f(x^n) - f(x^{n-1})}f(x^n), \tag{4.7}$$

(Metoda secantei).

Observaţia 4.4. *Definim funcţia* $g : V \to \mathbb{R}$,

$$g(x) = x - \frac{f(x)}{f'(x)}.$$

Avem că

- g *este de clasă* $C^2(V)$.
- $g(z) = z$.
- $$g'(z) = 1 - \frac{(f'(z))^2 - f(z)f''(z)}{(f'(z))^2} = \frac{f(z)f''(z)}{(f'(z))^2} = 0$$

  şi prin urmare $|g'(x)| \leq q < 1$ *într-o vecinătate a lui* z.

Prin urmare, metoda Newton poate fi privită şi ca o metodă de punct fix pentru funcţia g.

Mai reţinem şi următorul rezultat (A. Ostrowski, *Solution of equations and systems of equations*, Academic Press, New York, 1966, Theorem 7.2):

Teorema 4.2. *Fie* $f \in C^2([a,b])$ *şi* $x^0 \in [a,b]$. *Presupunem că funcţia* f *are următoarele proprietăţi*

i) $f(x^0)f'(x^0) \neq 0$

ii) Intervalul $I_0 = \left[\min\{x^0, x^0 + 2h_0\},\ \max\{x^0, x^0 + 2h_0\}\right]$, *unde* $h_0 = -f(x^0)/f'(x^0)$, *este inclus în* $[a,b]$

iii) $2|h_0|M \leq |f'(x^0)|$, *unde* $M = \max_{x \in I_0} |f''(x)|$.

Atunci ecuaţia (4.1) are o soluţie unică $z \in I_0$ *iar şirul (4.2) dat de metoda Newton este bine definit, inclus în* I_0 *şi converge la* z. *În plus,*

$$|z - x^{n+1}| \leq \frac{M}{2\,|f'(x^n)|}|x^n - x^{n-1}|^2, \quad n \geq 1. \tag{4.8}$$

4.1 Exemplu

Considerăm ecuaţia $x^2 - 2 = 0$ care are o unică rădăcină în $\mathbb{R}_+$, $z = \sqrt{2}$. O vom aproxima cu metoda Newton considerând algoritmul

$$x^0 = 2, \quad x^{n+1} = x^n - \frac{(x^n)^2 - 2}{2x^n} = \frac{(x^n)^2 + 2}{2x^n}.$$

Să calculăm primele iteraţii

$$x^0 = 2,\ x^1 = \frac{3}{2} = 1.5,\ x^2 = \frac{17}{12} = 1.4167,\ x^3 = \frac{577}{408} = 1.4142157.$$

Observăm că s-a obţinut o aproximare foarte bună penrtru $\sqrt{2} = 1.41421356237310...$ cu o eroare mai mică decât 10^{-5} în doar trei iteraţii.

Cursul 5

Rezolvarea ecuaţiilor polinomiale

Una dintre cele mai importante şi mai frecvente probleme matematice este aceea de determinare a rădăcinilor polinoamelor cu coeficienţi reali. Oricare dintre metodele discutate anterior poate fi utilizată în acest scop. Unele dintre aceste metode permit determinarea rădăcinilor complexe (de exemplu metoda Newton, dacă se lucrează cu aproximare iniţială număr complex) altele nu (metoda bisecţiei). Există însă şi metode specifice ecuaţiilor polinomiale (Lobacevski, Bairstow, Laguerre, Müler etc) care pot determina toate rădăcinile, reale sau complexe.

În cele ce urmează vom arăta cum procedeul deflaţiei în combinaţie cu metoda Newton pot fi folosite pentru a aproxima toate rădăcinile reale ale unui polinom

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_1 x + a_0,$$

unde a_i sunt numere reale.

Notăm cu $\mathbb{P}_n[x]$ mulţimea polinoamelor de grad cel mult n cu coeficienţi reali. Vom începe printr-o discuţie a modului în care putem evalua eficient un polinom şi derivata lui într-un punct.

5.1 Evaluarea polinoamelor şi schema lui Horner

Pentru evaluarea polinomului $p(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_1 x + a_0$ în punctul a putem folosi următorul algoritm:

1. $s = a_0$
2. Pentru $i = 1, 2, ..., n$
   - 2.1 $s = s + a_i * a$
   - 2.2 $a = a * a$
3. $p(a) = s$.

Acest algoritm organizează calculele astfel:

$$p(a) = a_0 + (a_1 * a + (a_2 * a * a + (a_3 * a^2 * a... + (a_{n-1} * a^{n-2} * a + (a_n * a^{n-1} * a)...)),$$

şi presupune efectuarea a n adunări şi $2n$ înmulţiri.

Există însă o modalitate mai eficientă de evaluare a polinomului, care se bazează pe următoarea scriere a polinomului p

$$p(a) = a_0 + a * (a_1 + a * (a_2 + ... + a * (a_{n-1} + a_n * a)...)).$$

a_n	a_{n-1}	a_{n-2}	.....	a_1	a_0
a_n	$a_{n-1} + ab_{n-1}$	$a_{n-2} + ab_{n-2}$	.....	$a_1 + ab_1$	$a_0 + ab_0$
b_{n-1}	b_{n-2}	b_{n-3}	.....	b_0	$p(a)$

Tabelul 5.1: Schema lui Horner de determinare a restului şi a câtului împărţirii lui $p(x)$ la $x - a$.

Algoritmul de calcul corespunzător este următorul:

1. $b_{n-1} = a_n$
2. Pentru $i = n-1, n-2, ..., 1$

   $$b_{i-1} = a_i + b_i * a$$

3. $p(a) = a_0 + b_0 * a$

Acest algoritm presupune doar n adunări şi n înmulţiri şi este de preferat celui anterior. El reprezintă de fapt schema lui Horner de determinare a restului şi a câtului împărţirii lui $p(x)$ la $x - a$, aşa cum este redat în Tabelul 5.1. Să observăm că, în plus, determinăm şi câtul împărţirii lui $p(x)$ la $x - a$:

$$p(x) = (x-a)(b_{n-1}x^{n-1} + b_{n-2}x^{n-2} + ... + b_1x + b_0) + p(a) := (x-a)q(x) + p(a).$$

De asemenea, să notăm că

(5.1)
$$p'(a) = q(a) = b_{n-1}a^{n-1} + b_{n-2}a^{n-2} + ... + b_1a + b_0 = b_0 + a(b_1 + ... + a(b_{n-2} + ab_{n-1})...)$$

şi putem realiza un algoritm eficient de calcul simultan al valorilor $p(a)$ şi $p'(a)$:

1. $p = a_n$
2. $dp = 0$
3. Pentru $i = n-1, n-2, ..., 0$

   3.1 $dp = p + dp * a$

   3.2 $p = a_i + p * a$

5.2 Deflaţia

Pentru determinarea tuturor rădăcinilor lui p se poate folosi metoda Newton în combinaţie cu procedeul **deflaţiei**. Algoritmul constă în următorii paşi:

1. Determinarea unei rădăcini r a lui p. Se poate folosi metoda Newton:
   1.1. Se alege x^0
   1.2. Se generează prin recurenţă şirul
   $$x^{k+1} = x^k - \frac{p(x^k)}{p'(x^k)} \qquad (k \geq 0),$$
   folosind algoritmul de mai sus pentru evaluarea simultană a lui $p(x^k)$ şi $p'(x^k)$.
   1.3 Se foloseşte unul dintre criteriile de oprire a iteraţiilor şi se aproximează r cu ultima iteraţie calculată (să zicem a $n-$a):
   $$r \approx x^n.$$
2. Împărţirea lui p la $x - r$ (de fapt la $x - x^n$) şi determinarea câtului q cu folosirea schemei Horner (deflaţie).
3. Reluarea algoritmului cu q în loc de p.

Probleme care pot apărea:

- Metoda Newton nu converge. În acest caz se încearcă o mai bună localizare a rădăcinii r şi alegerea unei aproximaţii iniţiale x^0 mai aproape de ea. Există tehnici speciale de localizare (Sturm, Gersgorin). De asemenea, se poate aplica la început o metodă mai lentă precum metoda bisecţiei care are proprietăţi mai bune de convergenţă.
- Prin deflaţii succesive se pot introduce erori din ce în ce mai mari. De aceea este convenabil uneori să se înceapă aproximarea celei mai mici rădăcini în modul şi să se continue cu celelalte rădăcini în ordine crescătoare. Odată obţinută o aproximare x^n a unei rădăcini, o metodă de asigurare constă în aplicarea metodei Newton direct polinomului original p (şi nu celui obţinut prin deflaţii) luând ca aproximare iniţială tocmai pe x^n. Acest procedeu se numeşte **rafinare**.

5.3 Garantarea convergenţei metodei lui Newton

Următorul rezultat ne arată că, în anumite cazuri, putem garanta convergenţa metodei lui Newton pentru o largă plajă de posibile valori ale aproximaţiei iniţiale.

Teorema 5.1. *Dacă rădăcinile polinomului $p \in \mathbb{P}_n[x]$ sunt toate reale şi aranjate crescător astfel*

$$r_1 \leq r_2 \leq ... \leq r_n,$$

atunci metoda lui Newton converge pentru orice aproximare iniţială $x^0 > r_n$.

Proof. Fără a afecta generalitatea problemei, putem presupune că $a_n > 0$. Rezultă că

$$p(x) > 0, \quad \forall x > r_n. \tag{5.2}$$

Rădăcinile derivatei p' sunt toate reale, situate între r_1 şi r_n. Aceeaşi proprietate o are şi p''. Ca mai sus, rezultă că

$$p'(x) > 0, \quad p''(x) > 0, \quad \forall x > r_n. \tag{5.3}$$

Fie $x^0 > r_n$ şi

$$x^1 = x^0 - \frac{p(x^0)}{p'(x^0)}.$$

Din (5.2)-(5.3) rezultă că $x^1 < x^0$. De asemenea, cum există ξ între x^0 şi r_n astfel încât

$$0 = p(r_n) = p(x^0) + (r_n - x^0)p'(x^0) + \frac{1}{2}(r_n - x^0)^2 p''(\xi),$$

avem că

$$p(x^0) + (r_n - x^0)p'(x^0) < 0 \Leftrightarrow r_n - \left(x^0 - \frac{p(x^0)}{p'(x^0)}\right) < 0 \Leftrightarrow r_n < x^1.$$

Prin inducţie matematică se demonstrează că şirul $(x^k)_{k\geq 0}$ dat de metoda lui Newton este descrescător şi minorat de r_n. Prin urmare şirul $(x^k)_{k\geq 0}$ este convergent la o valoare $r \geq r_n$. Prin trecere la limită în relaţia de recurenţă

$$x^{k+1} - x^k = \frac{p(x^k)}{p'(x^k)},$$

se obţine că $p(r) = 0$ şi deci $r = r_n$. □

Analog se poate demonstra şi următorul rezultat.

Teorema 5.2. *Dacă rădăcinile polinomului $p \in \mathbb{P}_n[x]$ sunt toate reale şi aranjate crescător astfel*

$$r_1 \leq r_2 \leq ... \leq r_n,$$

atunci metoda lui Newton converge pentru orice aproximare iniţială $x^0 < r_1$.

Teorema de mai sus ne arată că, dacă aproximaţia iniţială este mai mare decât cea mai mare rădăcină a polinomului, atunci avem garantată convergenţa. Cum se presupune că nu cunoaştem cea mai mare rădăcină a polinomului, putem folosi in locul ei un majorant. Următorul rezultat ne spune cum putem calcula uşor un majorant al celei mai mari rădăcini a unui polinom.

Teorema 5.3. *Orice rădăcină r a polinomului $p(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_1 x + a_0$ verifică*

$$|r| \leq 1 + \max_{0\leq k\leq n-1} \left|\frac{a_k}{a_n}\right|. \tag{5.4}$$

Proof. Fie $|x| > 1 + \max_{0\leq k\leq n-1} \left|\frac{a_k}{a_n}\right| := 1 + c$. Avem că

$$|p(x)| \geq |a_n|\,|x|^n - \left(\sum_{0\leq k\leq n-1} |a_k|\,|x|^k\right) = |a_n|\,|x|^n \left(1 - \frac{1}{|a_n|\,|x|^n} \sum_{0\leq k\leq n-1} |a_k|\,|x|^k\right) \geq$$

$$\geq |a_n|\,|x|^n \left(1 - \frac{1}{|a_n|\,|x|^n} \max_{0\leq k\leq n-1} |a_k| \sum_{0\leq k\leq n-1} |x|^k\right) =$$

$$= |a_n|\,|x|^n \left(1 - \frac{|x|^n - 1}{|x|^n\,(|x| - 1)} \frac{1}{|a_n|} \max_{0\leq k\leq n-1} |a_k|\right) \geq$$

$$\geq |a_n|\,|x|^n \left(1 - \frac{|x|^n}{|x|^n\,(|x| - 1)} c\right) = |a_n|\,|x|^n \frac{|x| - 1 - c}{|x| - 1} > 0.$$

Prin urmare, orice $|x| > 1 + c$ nu poate fi rădăcină a lui p. □

Observaţia 5.1. *Prin urmare, dacă ştim că toate rădăcinile unui polinom* $p(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$ *sunt reale şi luăm*

$$x^0 \geq 1 + \max_{0 \leq k \leq n-1} \left| \frac{a_k}{a_n} \right|,$$

atunci şirul dat de metoda Newton converge garantat la cea mai mare rădăcină a lui p.

5.4 Exemplu

Dorim găsirea rădacinilor polinomului $p(x) = 5x^3 - x^2 - 10x + 2 = 0$ folosind metoda lui Newton combinată cu deflaţie. Cum sunt trei rădăcini, vom avea trei etape:

Etapa 1: Formula lui Newton ne dă relaţia

$$x^{n+1} = x^n - \frac{5(x^n)^3 - (x^n)^2 - 10x^n + 2}{15(x^n)^2 - 2x^n - 10} = \frac{10(x^n)^3 - (x^n)^2 - 2}{15(x^n)^2 - 2x^n - 10}.$$

Pentru a determina pe x^0 folosim Teorema 5.3

$$1 + \max_{0 \leq k \leq 2} \left| \frac{a_k}{a_n} \right| = 1 + \max \left\{ \frac{2}{5}, \frac{10}{5}, \frac{1}{5} \right\} = 3,$$

deci putem lua $x^0 = 3$. Rezultă că:

$$x^1 = \frac{259}{119} = 2.17647058823529; \quad x^2 = 1.69944970726437; \quad x^3 = 1.47692799467256;$$

$$x^4 = 1.41836046282858; \quad x^5 = 1.41423365036378; \quad x^6 = 1.41421356284808$$

S-ar părea că $r_3 = \sqrt{2}$(?).Vom încerca să vedem dacă este aşa printr-un calcul exact

$$P(\sqrt{2}) = 10\sqrt{2} - 2 - 10\sqrt{2} + 2 = 0 \text{ (URAAA!)}$$

Am obţinut cea mai mare rădăcină:

$$r_3 = \sqrt{2}.$$

Etapa 2: Întâi facem deflaţie. Folosind schema lui Horner, împărţim pe $p(x)$ la $x - \sqrt{2}$.

5	-1	-10	2
5	$-1 + 5\sqrt{2}$	$-\sqrt{2}$	0
b_2	b_1	b_0	$p(\sqrt{2})$

Tabelul 5.2: Schema lui Horner de determinare a restului şi a câtului împărţirii lui $p(x)$ la $x - \sqrt{2}$.

Obţinem (vezi Tabelul 5.2): $p(x) = (x - \sqrt{2})p_1(x)$ cu

$$p_1(x) = 5x^2 + (-1 + 5\sqrt{2})x - \sqrt{2}.$$

Formula lui Newton ne dă relaţia:

$$x^{n+1} = x^n - \frac{5(x^n)^2 + (-1+5\sqrt{2})x^n - \sqrt{2}}{10x^n - 1 + 5\sqrt{2}} = \frac{5(x^n)^2 + \sqrt{2}}{10x^n - 1 + 5\sqrt{2}}$$

Pentru iteraţia iniţială x^0, putem folosi Teorema 5.3

$$1 + \max_{0\le k\le 1}\left|\frac{a_k}{a_n}\right| = 1 + \max\left\{\frac{\sqrt{2}}{5}, \sqrt{2} - \frac{1}{5}\right\} = \sqrt{2} + \frac{4}{5},$$

şi luăm $x^0 \ge \sqrt{2} + \frac{4}{5}$.

O altă variantă este să luăm chiar $x^0 \ge \sqrt{2} = r_3$. Vom considera acest ultim caz $x^0 = \sqrt{2}$. Obţimen că:

$$x^1 = \frac{10+\sqrt{2}}{15\sqrt{2}-1} = 0.56469097531924; \quad x^2 = 0.25675019718244; \quad x^3 = 0.20186407296630;$$

$$x^4 = 0.20000214764723; \quad x^5 = 0.20000000000286$$

S-ar părea că $r_2 = 0.2$(?).Vom încerca să vedem dacă este aşa printr-un calcul exact

$$p(0.2) = 0.04 - 0.04 - 2 + 2 = 0 \text{ (URAAA!)}$$

Am obţinut a doua cea mai mare rădăcină:

$$r_2 = 0.2.$$

Etapa 3: Din nou facem deflaţie. Folosind schema lui Horner împărţim pe $p_1(x)$ la $x - 0.2$.

5	$-1+5\sqrt{2}$	$-\sqrt{2}$
5	$-1+5\sqrt{2}+1$	0
b_1	b_0	$p(0.2)$

Tabelul 5.3: Schema lui Horner de determinare a restului şi a câtului împărţirii lui $p_1(x)$ la $x - 0.2$.

Obţinem (vezi tabelul 5.3): $p_1(x) = (x - 0.2)p_2(x)$ cu

$$p_2(x) = 5x + 5\sqrt{2}.$$

Acum obţinem direct cea de a treia rădăcină:

$$r_3 = -\sqrt{2}.$$

Cursul 6

Interpolare polinomială

Interpolarea şi extrapolarea sunt tehnici de obţinere a unor date noi pornind de la o mulţime cunoscută şi finită de date iniţiale. De obicei dorim să folosim interpolarea şi extrapolarea pentru predicție, în sensul că utilizăm date cunoscute pentru a obţine o estimare a datelor necunoscute. De exemplu, presupunem că avem următorul tabel care reprezintă temperatura aerului la patru ore diferite:

Ora	8	10	12	14
Temperatura	5	7	11	12

Am dori ca, folosind datele din tabel, să spunem cât este temperatura şi la alte ore, cum ar fi ora 11 sau ora 16, fără a face alte măsurători. Acest lucru se poate face folosind extrapolarea şi interpolarea.

Pentru a vedea care este diferența dintre cele două, trebuie să cunoaştem semnificaţia prefixelor "extra" și "inter" (de la originalele din latină). Prefixul "extra" înseamnă "afară" sau "pe lângă". Prefixul "inter" înseamnă "între" sau "printre". Prin urmare interpolarea prezice valoarea unor date noi care se află în mijlocul datelor iniţiale. În exemplul de mai sus, această valoare ar fi temperatura la ora 11. Din contră, extrapolarea prezice valoarea unor date noi din afara domeniului datelor iniţiale, cum ar fi temperatura la ora 16 în cazul nostru.

Există o mare varietate de metode de interpolare și extrapolare, bazate pe tendința generală care se observă în datele iniţiale. Pentru ambele metode, presupunem însă că, prin experimentare, eşantionare sau colectare de date, avem un număr finit, să zicem $n+1$, de date iniţiale sub forma unor perechi de numere:

$$(x_0, y_0), (x_1, y_1), ..., (x_n, y_n). \tag{6.1}$$

Primele componente, x_0, x_1,..., x_n, reprezintă o variabilă independentă iar componentele de pe poziţia a doua, y_0, y_1,..., y_n, reprezintă o variabilă dependentă. În exemplul anterior, valorile variabilei independente sunt 8, 10, 12 şi 14 şi reprezintă orele la care s-au măsurat temperaturile. Valorile variabilei dependente sunt 5, 7, 11 şi 12, indicând temperatura în funcţie de oră. Să mai menţionăm că valorile variabilei independente trebuie să fie numere reale diferite iar valorile variabilei dependente sunt numere reale oarecare.

În continuare, având datele (6.1), încercăm să formulăm un model pentru ele. Acesta ar putea fi reprezentat de o linie dreaptă care este cât mai aproape de date sau ar putea fi dat un alt tip de curbă care chiar treace prin acele date. În orice caz, vom avea o funcție care leagă variabila

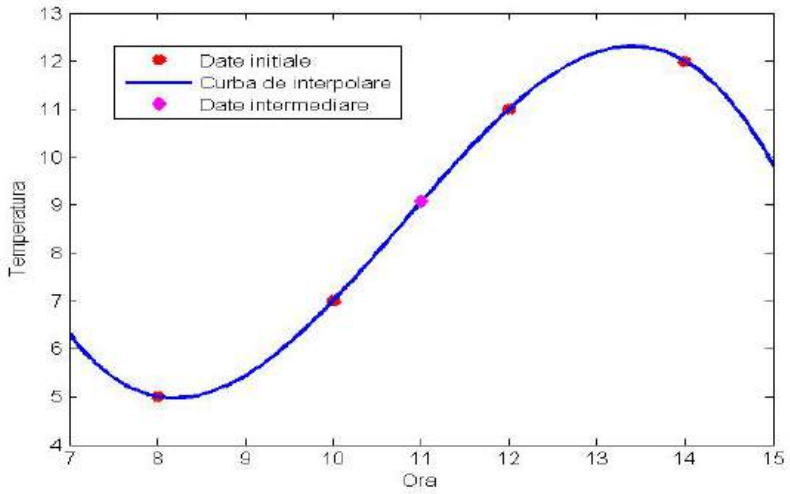


Figura 6.1: Curba de interpolare si evaluarea unei date intermediare

independentă cu variabila dependentă şi care ne dă valori adiţionale ale datelor, în afara celor iniţiale (vezi Figura 6.1).

Dintre cele două metode, interpolarea este preferată pentru că ea oferă o probabilitate mai mare de a obţine o estimare valabilă pentru noile date. Când folosim extrapolarea, presupunem că tendinţa observată va continuă pentru valori în afara intervalului pe care l-am folosit pentru a propune modelul nostru. Este posibil să nu fie cazul şi, prin urmare, trebuie să fim foarte atenţi atunci când folosim tehnici de extrapolare. În cele ce urmează vom aborda doar procedeul de interpolare. Mai mult, ne vom limita la o singură metodă de interpolare, cea mai simplă, numită *interpolare polinomială*.

În analiza numerică, interpolarea polinomială constă în următoarele: dat un set de puncte în plan de forma (6.1), vom căuta un polinom care trece prin toate aceste puncte şi care are cel mai mic grad posibil. Notăm cu $\mathbb{P}_n[x]$ mulţimea polinoamelor de grad cel mult n cu coeficienţi reali.

6.1 Polinomul de interpolare Lagrange

Fie $n+1$ numere reale diferite între ele x_0, x_1, ..., x_n şi $n+1$ numere reale arbitrare y_0, y_1, ..., y_n.

Definiţia 6.1. *Se numeşte polinom de interpolare pe nodurile* (x_0, y_0), (x_1, y_1), ..., (x_n, y_n) *un polinom* p_n *cu următoarele proprietăţi:*

1. $p_n(x_i) = y_i$, $0 \leq i \leq n$;
2. $\text{grad}\,(p_n) \leq n$.

Următoarea teoremă nu doar ne arată că există un unic polinom de interpolare, ci ne oferă şi o modalitate de construcţie a lui.

Teorema 6.1. *Există un unic polinom de interpolare* p_n *pe nodurile* (x_0, y_0), (x_1, y_1), ..., (x_n, y_n).

Proof. Existenţa: Definim mai întâi următoarele polinoame (numite polinoame elementare Lagrange):

$$l_i(x) = \prod_{\substack{k=0 \\ k\neq i}}^{n} \frac{x - x_k}{x_i - x_k} \qquad (0 \leq i \leq n).$$

Polinoamele elementare Lagrange au următoarele proprietaăţi:

1. Fiecare l_i este un polinom de grad n;
2. Pentru fiecare l_i avem următoarele relaţii

$$l_i(x_j) = \delta_{ij} = \begin{cases} 0 & \text{dacă } i \neq j \\ 1 & \text{dacă } i = j. \end{cases}$$

Definim acum polinomul p_n astfel

$$p_n(x) = \sum_{i=0}^{n} y_i l_i(x).$$

Ţinând cont de proprietăţile polinoamelor l_i se verifică uşor că p_n este un polinom de interpolare pe nodurile (x_0, y_0), (x_1, y_1), ..., (x_n, y_n).

Unicitatea: Fie p_n şi q_n două polinoame de interpolare pe nodurile (x_0, y_0), (x_1, y_1), ..., (x_n, y_n). Să notăm $r_n = p_n - q_n$. Polinomul r_n are următoarele proprietaăţi:

1. r_n este un polinom de grad cel mult n;
2. $r_n(x_i) = 0$, $0 \leq i \leq n$.

Din cele două proprietăţi de mai sus rezultă imediat că $r_n \equiv 0$ şi, prin urmare, există un unic polinom de interpolare. □

Observaţia 6.1. *Teorema precedentă ne arată o modalitate de a construi polinomul de interpolare p_n pe nodurile (x_0, y_0), (x_1, y_1), ..., (x_n, y_n):*

- *Se construiesc polinoamele elementare Lagrange:*

$$l_i(x) = \prod_{\substack{k=0 \\ k\neq i}}^{n} \frac{x - x_k}{x_i - x_k} \qquad (0 \leq i \leq n); \tag{6.2}$$

- *Se calculează suma care ne dă polinomul de interpolare:*

$$p_n(x) = \sum_{i=0}^{n} y_i \, l_i(x). \tag{6.3}$$

Aceasta este doar o posibilitate de a construi polinomul de interpolare. Scris în forma (6.3) *de mai sus, polinomul de interpolare va fi numit "polinomul de interpolare Lagrange pe nodurile (x_0, y_0), (x_1, y_1), ..., (x_n, y_n)".*

6.2 Polinomul de interpolare Newton

Dezavantajul formulei (6.3) este acela că, dacă mai adăugăm un nod, trebuie să luăm calculul de la început. Vom vedea o altă formă de scriere a polinomului de interpolare în care adăugarea de noi noduri se va face cu mai puţine calcule. Vom avea nevoie însă de o pregătire.

6.2.1 Formula Aitken-Neville

Începem cu următoarea formulă care ne permite construirea polinomului de interpolare recursiv, folosind două polinoame de interpolare pe mai puţine noduri.

Teorema 6.2 (Aitken-Neville)**.** *Dacă $i, j \in \{0, 1, ..., n\}$ sunt doi indici diferiţi, fie $q_{n-1,i}$ polinomul de interpolare pe nodurile (x_0, y_0), (x_1, y_1), ..., (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), ..., (x_n, y_n) şi fie $q_{n-1,j}$ polinomul de interpolare pe nodurile (x_0, y_0), (x_1, y_1), ..., (x_{j-1}, y_{j-1}), (x_{j+1}, y_{j+1}), ..., (x_n, y_n). Atunci, polinomul p_n de interpolare pe toate nodurile (x_0, y_0), (x_1, y_1), ..., (x_n, y_n) este dat de formula:*

$$p_n(x) = \frac{(x - x_i)q_{n-1,i}(x) - (x - x_j)q_{n-1,j}(x)}{x_j - x_i}. \tag{6.4}$$

Proof. Se arată uşor că p_n dat de (6.4) are următoarele proprietăţi:

- p_n este un polinom de grad cel mult n deoarece $q_{n-1,i}$ şi $q_{n-1,j}$, fiind polinoame de interpolare pe n noduri, sunt polinoame de grad cel mult $n - 1$.
- Pentru fiecare $k \in \{0, 1, ..., n\}$, $p_n(x_k) = y_k$ deoarece $q_{n-1,i}(x_k) = y_k$ dacă $k \neq i$ şi $q_{n-1,j}(x_k) = y_k$ dacă $k \neq j$.

Prin urmare, sunt îndeplinite cerinţele Definiţiei 6.1 şi p_n este polinomul de interpolare pe toate nodurile (x_0, y_0), (x_1, y_1), ..., (x_n, y_n). □

6.2.2 Diferenţe divizate

Ne plasăm în aceleşi ipoteze de mai înainte. Presupunem că dispunem de $n+1$ numere reale diferite între ele x_0, x_1, ..., x_n şi $n+1$ numere reale arbitrare y_0, y_1, ..., y_n. Vom defini un obiect matematic nou, numit diferenţe divizate.

Definiţia 6.2. *Se numesc diferenţe divizate următoarele numere reale definite prin recurenţă:*

$$y[x_i] := y_i,$$
(diferenţe divizate de ordin 0)

$$y[x_i, x_{i+1}, ..., x_{i+k}] := \frac{y[x_i, x_{i+1}, ..., x_{i+k-1}] - y[x_{i+1}, x_{i+2}, ..., x_{i+k}]}{x_i - x_{i+k}},$$
(diferenţe divizate de ordin $k \geq 1$).

Pentru calculul diferenţelor divizate trebuie să facem următorul tabel:

Observaţi că, pentru a avea diferenţele divizate de pe prima linie a tabelului 6.1, este nevoie să completăm întregul tabel.

Cele mai importante proprietăţi ale diferenţelor divizate sunt menţionate în următoarea propoziţie.

x_0	$y[x_0]=y_0$	$y[x_0,x_1]=\frac{y[x_0]-y[x_1]}{x_0-x_1}$	$y[x_0,x_1,x_2]=\frac{y[x_0,x_1]-y[x_1,x_2]}{x_0-x_2}$	...
x_1	$y[x_1]=y_1$	$y[x_1,x_2]=\frac{y[x_1]-y[x_2]}{x_1-x_2}$	$y[x_1,x_2,x_3]=\frac{y[x_1,x_2]-y[x_2,x_3]}{x_1-x_3}$	...
x_2	$y[x_2]=y_2$	$y[x_2,x_3]=\frac{y[x_2]-y[x_3]}{x_2-x_3}$	...	...
x_3	$y[x_3]=y_3$	...	...	...
...	...	...	...	...
x_{n-1}	$y[x_{n-1}]=y_{n-1}$	$y[x_{n-1},x_n]=\frac{y[x_{n-1}]-y[x_n]}{x_{n-1}-x_n}$	–	...
x_n	$y[x_n]=y_n$	–	–	...

...	$y[x_0,x_1,...,x_{n-1}]=\frac{y[x_0,x_1,...,x_{n-2}]-y[x_1,x_1,...,x_{n-1}]}{x_0-x_{n-1}}$	$y[x_0,x_1,...,x_n]=\frac{y[x_0,x_1,...,x_{n-1}]-y[x_1,x_1,...,x_n]}{x_0-x_n}$
...	$y[x_1,x_1,...,x_n]=\frac{y[x_1,x_2,...,x_{n-1}]-y[x_2,x_3,...,x_n]}{x_1-x_n}$	–
...	–	–
...	–	–
...	–	–
...	–	–
...	–	–

Tabelul 6.1: Tabelul diferenţelor divizate pe nodurile (x_0,y_0), (x_1,y_1), ..., (x_n,y_n).

Propoziţia 6.1. *Diferenţele divizate au următoarele proprietăţi:*

1. $y[x_0,\dots,x_n]=\sum_{j=0}^{n}\frac{y_j}{\prod_{0\le i\le n,\, i\neq j}(x_j-x_i)}$;
2. *Invarianţa la permutarea indicilor: pentru orice permutare $\sigma\in S_{\{0,\dots,k\}}$ avem că*
$$y[x_{\sigma(0)},\dots,x_{\sigma(k)}]=y[x_0,\dots,x_k];$$
3. *Linearitatea: Pentru orice două numere reale a şi b şi orice două sisteme de puncte (x_0,y_0), (x_1,y_1), ..., (x_n,y_n) şi (x_0,z_0), (x_1,z_1), ..., (x_n,z_n) avem că*
$$(ay+bz)[x_0,\dots,x_n]=a\,y[x_0,\dots,x_n]+b\,z[x_0,\dots,x_n];$$
4. *Regula lui Leibniz: Pentru orice două sisteme de puncte (x_0,y_0), (x_1,y_1), ..., (x_n,y_n) şi (x_0,z_0), (x_1,z_1), ..., (x_n,z_n) avem că*
$$(yz)[x_0,\dots,x_n]=\sum_{j=0}^{n}y[x_0,\dots,x_j]\,z[x_j,\dots,x_n];$$
5. *Proprietatea de medie: Dată o funcţie $f\in C^n[a,b]$, fie $n+1$ valori diferite din (a,b): $a<x_0<\dots<x_n<b$ şi nodurile $(x_0,f(x_0)),(x_1,f(x_1)),\dots,(x_n,f(x_n))$. Avem că există $c\in(x_0,x_n)$ astfel încât:*
$$f[x_0,\dots,x_n]=\frac{f^{(n)}(c)}{n!}.$$

Proof. Proprietăţile sunt elementare şi sunt lăsate ca exerciţii. □

6.2.3 Polinomul de interpolare Newton

Utilitatea diferenţelor divizate este pusă în evidenţă de următorul rezultat.

Teorema 6.3. *Polinomul p_n de interpolare pe nodurile (x_0, y_0), (x_1, y_1), ..., (x_n, y_n) este dat de următoarea formulă:*

$$\begin{aligned} p_n(x) = y[x_0] + y[x_0, x_1](x - x_0) + y[x_0, x_1, x_2](x - x_0)(x - x_1) \\ + \cdots + y[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1)\ldots(x - x_{n-1}). \end{aligned} \tag{6.5}$$

Proof. Vom demonstra formula (6.5) prin inducţie după n.

Pentru $n = 0$ avem un singur nod (x_0, y_0) iar polinomul de interpolare p_0 este constant y_0. Deci avem că

$$p_0(x) = y_0 = y[x_0],$$

şi formula (6.5) este adevărată.

Presupunem că formula (6.5) este adevărată pentru orice n noduri şi o demonstrăm pentru $n+1$ noduri (x_0, y_0), (x_1, y_1), ..., (x_n, y_n). Fie $q_{n-1,n}$ polinomul de interpolare pe nodurile (x_0, y_0),..., (x_{n-1}, y_{n-1}) şi $q_{n-1,n-1}$ polinomul de interpolare pe nodurile (x_0, y_0), ...,(x_{n-2}, y_{n-2}), (x_n, y_n). Formula Aitken-Neville (6.4) ne spune că

$$\begin{aligned} p_n(x) &= \frac{(x - x_n)q_{n-1,n}(x) + (x - x_{n-1})q_{n-1,n-1}(x)}{x_{n-1} - x_n} \\ &= q_{n-1,n}(x) + \frac{(x - x_{n-1})(q_{n-1,n}(x) - q_{n-1,n-1})}{x_{n-1} - x_n} \end{aligned}$$

Folosind ipoteza de inducţie pentru polinoamele $q_{n-1,n}$ şi $q_{n-1,n-1}$, avem că:

$$q_{n-1,n}(x) - q_{n-1,n-1} = (y[x_0, x_1, \ldots, x_{n-1}] - y[x_0, x_1, \ldots, x_{n-2}, x_n])(x - x_0)(x - x_1)\ldots(x - x_{n-2}),$$

de unde, ţinând cont de faptul că

$$\frac{y[x_0, x_1, \ldots, x_{n-1}] - y[x_0, x_1, \ldots, x_{n-2}, x_n]}{x_{n-1} - x_n} = y[x_0, x_1, \ldots, x_n],$$

obţinem că

$$p_n(x) = q_{n-1,n}(x) + y[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1)\ldots(x - x_{n-2})(x - x_{n-1}).$$

Tinând cont din nou de forma lui $q_{n-1,n}(x)$ dată de ipoteza de inducţie, obţinem din egalitatea de mai sus că p_n este dat de (6.5) şi demonstraţia se termină. □

Observaţia 6.2. *Vom numi polinomul de interpolare în forma* (6.5) *"polinomul de interpolare Newton". Vă atrag atenţia că* (6.3) *şi* (6.5) *sunt doar două forme diferite de a scrie acelaşi unic polinom de interpolare p_n. Deci, având aceleaşi noduri de interpolare, polinomul Lagrange este identic cu polinomul Newton, doar formele de scriere ale lor sunt diferite.*

Observaţia 6.3. *Avantajul polinomului de interpolare Newton este următorul: dacă avem deja polinomul de interpolare pe nodurile (x_0, y_0),..., (x_{n-1}, y_{n-1}) (notat $q_{n-1,n}$ în demonstraţia anterioară) atunci nu avem de calculat decât un termen nou, $y[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1)\ldots(x - x_{n-2})(x - x_{n-1})$, pentru a-l obţine pe p_n:*

$$p_n(x) = q_{n-1,n}(x) + y[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1)\ldots(x - x_{n-2})(x - x_{n-1}).$$

6.3 Exemple

- Determinaţi folosind metoda Lagrange polinomul de interpolare $p_3(x)$ pe nodurile $(0,4)$, $(1,3)$, $(4,-2)$ şi $(5,1)$.

  Avem că:
  $$p_3(x) = \sum_{i=0}^{3} y_i\, l_i(x).$$
  Cum
  $$l_0(x) = \frac{(x-1)(x-4)(x-5)}{(-1)(-4)(-5)} = -\frac{1}{20}(x^3 - 10x^2 + 29x - 20),$$
  $$l_1(x) = \frac{(x-0)(x-4)(x-5)}{(1)(-3)(-4)} = \frac{1}{12}(x^3 - 9x^2 + 20x),$$
  $$l_2(x) = \frac{(x-0)(x-1)(x-5)}{(4)(3)(-1)} = -\frac{1}{12}(x^3 - 6x^2 + 5x),$$
  $$l_3(x) = \frac{(x-0)(x-1)(x-4)}{(5)(4)(1)} = \frac{1}{20}(x^3 - 5x^2 + 4x).$$
  Se obţine că:
  $$p_3(x) = 4\left(-\frac{1}{20}\right)(x^3 - 10x^2 + 29x - 20) + 3\frac{1}{12}(x^3 - 9x^2 + 20x)$$
  $$+(-2)\left(-\frac{1}{12}\right)(x^3 - 6x^2 + 5x) + 1\frac{1}{20}(x^3 - 5x^2 + 4x)$$
  $$= \frac{4}{15}x^3 - \frac{3}{2}x^2 + \frac{7}{30}x + 4.$$

- Determinaţi folosind metoda Newton polinomul de interpolare $p_2(x)$ pe nodurile $(1,4)$, $(2,-1)$ şi $(5,-8)$.

  Avem că
  $$p_2(x) = y[x_0] + y[x_0,x_1](x-x_0) + y[x_0,x_1,x_2](x-x_0)(x-x_1) =$$
  $$y[x_0] + y[x_0,x_1](x-1) + y[x_0,x_1,x_1](x-1)(x-2).$$
  Cum,
  $$y[x_0] = 4,$$
  $$y[x_0,x_1] = \frac{4-(-1)}{1-2} = -5,$$
  $$y[x_1,x_2] = \frac{-1-(-8)}{2-5} = -\frac{7}{3},$$
  $$y[x_0,x_1,x_2] = \frac{y[x_0,x_1]-y[x_1,x_2]}{1-5} = \frac{(-5)-(-7/3)}{1-5} = \frac{2}{3}.$$
  Deci,
  $$p_2(x) = 4 - 5(x-1) + \frac{2}{3}(x-1)(x-2)$$
  $$= \frac{2}{3}x^2 - 7x + \frac{31}{3}.$$

Cursul 7

Interpolare polinomială (II)

În cursul precedent am văzut cum, date $n+1$ noduri (x_0, y_0), (x_1, y_1), ..., (x_n, y_n), se defineşte şi se construieşte polinomul de interpolare p_n asociat. O situaţie particulară, dar destul de des întâlnită, este aceea în care nodurile sunt puncte pe graficul unei funcţii. Avem următoarea definiţie.

Definiţia 7.1. *Date $n+1$ numere reale diferite între ele x_0, x_1, ..., x_n şi o funcţie f definită în aceste puncte, polinomul de interpolare p_n pe nodurile $(x_0, f(x_0))$, $(x_1, f(x_1))$,..., $(x_n, f(x_n))$ se va numi polinomul de interpolare a funcţiei f pe nodurile x_0, x_1, ..., x_n.*

Diferenţele divizate pe nodurile $(x_0, f(x_0))$, $(x_1, f(x_1))$,..., $(x_k, f(x_k))$ se vor nota $[f; x_0, \ldots, x_k]$.

Prin urmare, dată o funcţie f, polinomul de interpolare p_n din definiţia de mai sus coincide cu f în x_0, x_1, ..., x_n. Putem să ne imaginăm că funcţia "complicată" f este "înlocuită" de funcţia polinomială p_n mult mai simplă şi mai uşor de folosit. Vom mai introduce o definiţie.

Definiţia 7.2. *Interpolarea pe două noduri, x_0 şi x_1, se va numi interpolare liniară, interpolarea pe trei noduri, x_0, x_1 şi x_2, se va numi interpolare pătratică, iar interpolarea pe patru noduri, x_0, x_1, x_2 şi x_3, se va numi interpolare cubică.*

7.1 Estimarea erorii în interpolarea polinomială

Următoarea teoremă ne spune cum se poate estima diferenţa între o funcţie f şi polinomul său de interpolare p_n.

Teorema 7.1. *Fie $n+1$ puncte diferite x_0, x_1, ..., x_n din intervalul $[a, b]$, o funcţie $f : [a, b] \to \mathbb{R}$ şi polinomul de interpolare p_n a funcţiei f pe nodurile x_0, x_1, ..., x_n. Dacă $f \in C^{n+1}[a, b]$, atunci pentru fiecare $x \in [a, b]$ există $\xi \in (a, b)$ cu proprietatea că*

$$f(x) - p_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi) \prod_{i=0}^{n} (x - x_i). \tag{7.1}$$

Proof. Mai întâi, să observăm că egalitatea (7.1) se verifică pentru orice $x = x_i$, $0 \leq i \leq n$. Fie acum $x \in [a, b] \setminus \{x_0, x_1, ..., x_n\}$ fixat. Definim funcţia $g : [a, b] \to \mathbb{R}$,

$$g(t) = f(t) - p_n(t) - \prod_{i=0}^{n} \frac{(t - x_i)}{(x - x_i)} \left(f(x) - p_n(x) \right).$$

Funcţia g este de clasă $C^{n+1}[a,b]$ şi $g(x_i)=0$ pentru $0\le i\le n$. Cum, în plus, $g(x)=0$, rezultă că g se anulează în cel puţin $n+2$ puncte diferite din $[a,b]$. Folosind teorema lui Rolle, rezultă că g' se anulează în cel puţin $n+1$ puncte diferite din (a,b). Folosind din nou teorema lui Rolle, obţinem că g'' se anulează în cel puţin n puncte diferite din (a,b) ş.a.m.d.

Se obţine că $g^{(n+1)}$ se anulează cel puţin o dată într-un punct $\xi\in(a,b)$. Rezultă că

$$0=g^{(n+1)}(\xi)=f^{(n+1)}(\xi)-(n+1)!\prod_{i=0}^{n}\frac{1}{(x-x_i)}(f(x)-p_n(x)),$$

de unde se obţine imediat (7.1). □

Observaţia 7.1. *Fie $f\in C^\infty[a,b]$ o funcţie cu proprietatea că există $M>0$ astfel încât*

$$\sup_{x\in[a,b]}\left|f^{(k)}(x)\right|\le M^k\qquad(k\ge 0). \tag{7.2}$$

Pentru fiecare $n\ge 0$, considerăm $n+1$ puncte diferite $(x_i^n)_{0\le i\le n}\subset[a,b]$ şi fie p_n polinomul de interpolare a lui f pe nodurile $(x_i^n)_{0\le i\le n}$. Avem că

$$\lim_{n\to\infty}||f-p_n||_{C[a,b]}=0. \tag{7.3}$$

Într-adevăr, (7.3) *rezultă din* (7.1), (7.2) *şi din faptul că*

$$\lim_{n\to\infty}\frac{(M(b-a))^{n+1}}{(n+1)!}=0.$$

7.2 Interpolare Hermite

Fie $f:[a,b]\to\mathbb{R}$ o funcţie derivabilă în punctele diferite $x_0,x_1,...,x_n$ din (a,b). Problema interpolării Hermite constă în a determina un polinom H de grad minim care să coincidă cu f şi a cărui derivată să coincidă cu derivata lui f în nodurile $x_0,x_1,...,x_n$:

$$H(x_i)=f(x_i),\quad H'(x_i)=f'(x_i),\qquad 0\le i\le n. \tag{7.4}$$

Următoarea teoremă ne arată existenţa, unicitatea şi modalitatea de construcţie a polinomului H. De asemenea, ne oferă o estimare a erorii în interpolarea Hermite.

Teorema 7.2. *Dacă $f:[a,b]\to\mathbb{R}$ este derivabilă în punctele diferite $x_0,x_1,...,x_n$ din (a,b), atunci există un unic polinom H_{2n+1} de grad cel mult $2n+1$ cu proprietăţile (7.4). Dacă, în plus, $f\in C^{2n+2}[a,b]$, pentru fiecare $x\in[a,b]$ există $\xi\in(a,b)$ cu proprietatea că*

$$f(x)-H_{2n+1}(x)=\frac{1}{(2n+2)!}f^{(2n+2)}(\xi)\prod_{i=0}^{n}(x-x_i)^2. \tag{7.5}$$

Proof. Să presupunem că dispunem, pentru fiecare $0\le j\le n$, de polinoame h_j şi $\widehat{h}_j$ de grad cel mult $2n+1$ cu următoarele proprietăţi

$$\begin{aligned}&h_j(x_i)=\delta_{ij},\quad h_j'(x_i)=0,\quad 0\le i\le n,\\ &\widehat{h}_j(x_i)=0,\quad \widehat{h}_j'(x_i)=\delta_{ij},\quad 0\le i\le n.\end{aligned} \tag{7.6}$$

Obţinem că

$$H_{2n+1}(x) = \sum_{j=0}^{n} f(x_j)h_j(x) + \sum_{j=0}^{n} f'(x_j)\widehat{h}_j(x) \tag{7.7}$$

este un polinom de grad cel mult $2n+1$ care verifică (7.4) şi demonstraţia existenţei se încheie.

Să construim acum polinoamle h_j şi $\widehat{h}_j$. Pentru aceasta folosim polinomul elementar Lagrange:

$$l_j(x) = \prod_{\substack{i=0 \\ i\neq j}} \frac{x - x_i}{x_j - x_i}.$$

Căutăm h_j şi $\widehat{h}_j$ de forma $h_j(x) = (ax+b)l_j^2(x)$ şi respectiv $\widehat{h}_j(x) = (\widehat{a}x + \widehat{b})l_j^2(x)$.

Evident, polinoamele de această formă au gradul cel mult $2n+1$ şi verifică

$$h_j(x_i) = h_j'(x_i) = \widehat{h}_j(x_i) = \widehat{h}_j'(x_i) = 0, \quad 0 \le i \le n, \quad i \neq j.$$

În primul rând avem că

$$0 = \widehat{h}_j(x_j) = (\widehat{a}x_j + \widehat{b})l_j^2(x_j), \quad 1 = \widehat{h}_j'(x_j) = \widehat{a}l_j^2(x_j) + 2(\widehat{a}x_j + \widehat{b})l_j(x_j)l_j'(x_j),$$

şi prin urmare

$$\widehat{h}_j(x) = (x - x_j)l_j^2(x). \tag{7.8}$$

În al doilea rând avem că

$$1 = h_j(x_j) = (ax_j + b)l_j^2(x_j), \quad 0 = h_j'(x_j) = a_j l_j^2(x_j) + 2(ax_j + b)l_j(x_j)l_j'(x_j),$$

de unde rezultă că

$$h_j(x) = (1 - 2(x - x_j)l_j'(x_j))l_j^2(x). \tag{7.9}$$

Astfel, am determinat h_j şi $\widehat{h}_j$ cu proprietăţile (7.6) şi existenţa polinomului hermite este demonstrată. Să mai observăm că

$$l_j'(x_j) = \sum_{\substack{i=0 \\ i\neq j}}^{n} \frac{1}{x_j - x_i},$$

şi prin urmare putem calcula $l_j'(x_j)$ fără a fi nevoie efectiv de derivarea lui l_j.

Pentru a demonstra unicitatea polinomului Hermite, fie H^1 şi H^2 două polinoame de grad cel mult $2n+1$ care verifică (7.4). Fie $H = H^1 - H^2$. H este un polinom de grad cel mult $2n+1$ care verifică

$$H(x_i) = H'(x_i) = 0, \quad 0 \le i \le n.$$

Din teorema lui Rolle obţinem că există y_i între x_i şi x_{i+1} cu proprietatea că $H'(y_i)$, pentru orice $0 \le i < n$. Prin urmare, H' se anulează în cel puţin $2n+1$ puncte diferite: $x_0, x_1, \ldots, x_n, y_0, \ldots, y_{n-1}$. Cum H' este polinom de grad cel mult $2n$, rezultă că $H' \equiv 0$. Rezultă că există $c \in \mathbb{R}$ astfel încât $H \equiv c$. Ţinând cont că H se anulează în cel puţin un punct (de exemplu x_0) rezultă că $H \equiv 0$.

Trecem acum la demonstrarea formulei de rest. Pentru aceasta, să observăm că egalitatea (7.5) se verifică pentru orice $x = x_i$, $0 \leq i \leq n$. Fie acum $x \in [a,b] \setminus \{x_0, x_1, ..., x_n\}$ fixat. Definim funcţia $g : [a,b] \to \mathbb{R}$,

$$g(t) = f(t) - H_{2n+1}(t) - \prod_{i=1}^{n} \frac{(t-x_i)^2}{(x-x_i)^2} \left(f(x) - H_{2n+1}(x)\right).$$

Funcţia g este de clasă $C^{2n+2}[a,b]$ şi $g(x_i) = g'(x_i) = 0$, $0 \leq i \leq n$. Cum, în plus, $g(x) = 0$, folosind teorema lui Rolle, rezultă că g' se anulează în cel puţin $2n+2$ puncte diferite din $[a,b]$.

Folosind din nou teorema lui Rolle, obţinem că g'' se anulează în cel puţin $2n+1$ puncte diferite din (a,b) ş.a.m.d.

În final se obţine că $g^{(2n+2)}$ se anulează cel puţin o dată într-un punct $\xi \in (a,b)$. Rezultă că

$$0 = g^{(2n+2)}(\xi) = f^{(2n+2)}(\xi) - (2n+2)! \prod_{i=0}^{n} \frac{1}{(x-x_i)^2} (f(x) - H_{2n+1}(x)),$$

de unde se obţine imediat (7.5). □

Vom folosi următoarea definiţie:

Definiţia 7.3. *Polinomul H_{2n+1} de grad cel mult $2n+1$ cu proprietăţile (7.4) ale cărui existenţă şi unicitate au fost demonstrate în Teorema 7.2 se numeşte polinomul de interpolare Hermite a funcţiei f pe nodurile x_0, x_1, ..., x_n.*

Observaţia 7.2. *Fie $f \in C[-1,1]$, $(x_i)_{0 \leq i \leq n}$ zerourile polinomului Cebîşev de grad $n+1$,*

$$x_i = \cos\left(\frac{(2i+1)\pi}{2(n+1)}\right),$$

iar H_{2n+1} polinomul de interpolare Hermite cu proprietatea că

$$H_{2n+1}(x_i) = f(x_i), \quad H'_{2n+1}(x_i) = 0, \qquad 0 \leq i \leq n.$$

H_{2n+1} se numeşte polinomul de interpolare Féjer-Hermite. Se poate demonstra că

$$\lim_{n \to \infty} ||f - H_{2n+1}||_{C[-1,1]} = 0. \tag{7.10}$$

Deci avem aceeaşi concluzie ca în Observaţia 7.1, dar fără a impune condiţii suplimentare asupra lui f.

7.3 Exemplu

Următorul extras este din tabelul corzilor aflat în cartea Almageste (din arabă, al-Magisti - Cea mai mare) sau Mathematike Syntaxis (din greacă, Tratat de matematică) scrisă în jurul anului 150 d.Hr. de către Claudiu Ptolemeu (cca. 87-165 d.Hr, Alexandria, Egipt) şi care mai mult de o mie de ani a constituit baza cunoştinţelor astronomice ale lumii islamice și europene:

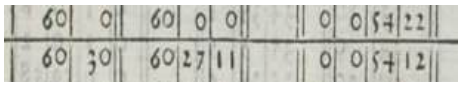


Figura 7.1: Extras din Almageste

Ne spune că (numerele sunt în baza 60, după tradiţia babiloniană):

$$\sin(30^\circ) \;=\; \sin\frac{\pi}{6} \;=\; \frac{1}{120}\left(60 + 0\frac{1}{60} + 0\frac{1}{3600}\right) \;=\; 0.5$$

60 0 din primele două coloane indică: coarda unghiului de 60° dintr-un cerc de rază 60 iar

60 0 0 din următoarele trei coloane este valoarea ei în baza 60

$$\text{“şaizecimi”} \;=\; \frac{1}{120}\left(0 + 0\frac{1}{60} + 54\frac{1}{3600} + 22\frac{1}{216000}\right) \;=\; 0.0001258;$$

0 0 54 23 din ultimele patru coloane este valoarea şaizecimilor în baza 60

$$\sin(30^\circ 30') \;=\; \sin\left(\frac{\pi}{6} + \frac{\pi}{320}\right) \;=\; \frac{1}{120}\left(60 + 27\frac{1}{60} + 11\frac{1}{3600}\right) \;=\; 0.5037755$$

$$\text{“şaizecimi”} \;=\; \frac{1}{120}\left(0 + 0\frac{1}{60} + 54\frac{1}{3600} + 12\frac{1}{216000}\right) \;=\; 0.0001255;$$

La ce servesc aceste cantităţi numite “şaizecimi” (în engleză “sixtieths”)?

Răspunsul este: pentru determinarea valorilor $\sin(30^\circ 1')$, $\sin(30^\circ 2')$, ..., $\sin(30^\circ 29')$ adică a valorilor sinusului pentru argumente situate la distanţă de $1'$ unul de altul, între 30° şi $30^\circ 30'$.

Cum se face acest lucru? Prin interpolare liniară, folosind cele două valori cunoscute din tabel $\sin(30^\circ)$ şi $\sin(30^\circ 30')$ precum şi a “şaizecimilor”. Mai precis:

$$\begin{aligned}
\sin(30^\circ j') = \sin\left(\frac{\pi}{6} + \frac{\pi j}{9600}\right) &\approx p_1\left(\frac{\pi}{6} + \frac{\pi j}{9600}\right);\\
p_1(x) = \sin\left(\frac{\pi}{6}\right)\frac{x - \frac{\pi}{6} - \frac{\pi}{320}}{-\frac{\pi}{320}} &+ \sin\left(\frac{\pi}{6} + \frac{\pi}{320}\right)\frac{x - \frac{\pi}{6}}{\frac{\pi}{320}};\\
p_1\left(\frac{\pi}{6} + \frac{\pi j}{9600}\right) = \sin\left(\frac{\pi}{6}\right)\frac{30 - j}{30} &+ \sin\left(\frac{\pi}{6} + \frac{\pi}{320}\right)\frac{j}{30}\\
= \sin\left(\frac{\pi}{6}\right) + j\,\underbrace{\frac{1}{30}\left[\sin\left(\frac{\pi}{6} + \frac{\pi}{320}\right) - \sin\left(\frac{\pi}{6}\right)\right]}_{\text{”şaizecimi”}}&.
\end{aligned}$$

Deci, folosind prima valoare ”şaizecimi” de mai sus, obţinem că:

$$\sin(30^\circ j') = \sin\left(\frac{\pi}{6} + \frac{\pi j}{9600}\right) \approx 0.5 + j * 0.0001258 \qquad (0 \le j \le 30). \tag{7.11}$$

În concluzie, ultimele patru coloane din tabelul corzilor cuprins în cartea Almageste de acum aproape 2000 de ani reprezintă nişte valori numite "şaizecimi" care ne permit, prin interpolare liniară, să aproximăm valori intermediare ale sinusului, de exemplu $\sin(30°1')$, $\sin(30°2')$, $\ldots \sin(30°29')$, care nu sunt date în tabel.

Aceasta este una dintre primele folosiri ale interpolării. Reprezintă, de asemenea, unul dintre exemplele de bază ale utilităţii interpolării: obţinerea de date noi, pe baza celor precizate în tabele.

Exerciţiul 7.3.1. *Estimaţi eroarea în procesul de interpolare descris mai sus.*

Proof. Vom folosi Teorema 7.1 cu $f(x) = \sin(x)$, $n = 1$, $x_0 = \frac{\pi}{6}$, $x_1 = \frac{\pi}{6} + \frac{\pi}{320}$. Avem că:

$$|\sin(x) - p_1(x)| = \left|\sin(\xi)\frac{1}{2!}\left(x - \frac{\pi}{6}\right)\left(x - \frac{\pi}{6} - \frac{\pi}{320}\right)\right|,$$

de unde, pentru $x = \frac{\pi}{6} + \frac{\pi j}{9600}$, obţinem că

$$\left|\sin\left(\frac{\pi}{6} + \frac{\pi j}{9600}\right) - p_1\left(\frac{\pi}{6} + \frac{\pi j}{9600}\right)\right| = \left|\sin(\xi)\frac{j(30-j)\pi^2}{2 \cdot 9600^2}\right|$$

$$\leq \left|\frac{j(30-j)\pi^2}{2 \cdot 9600^2}\right| \leq 2.7j(30-j)10^{-8} \leq 6.03 \times 10^{-6},$$

ultima valoare fiind obţinută luând $j = 15$, argument care ne dă valoarea maximă a lui $j(30-j)$ pe intervalul $[0, 30]$.

Prin urmare, eroarea în interpolare este de cel mult 6.03×10^{-6}. □

Cursul 8

Integrare numerică

Integrare numerică urmăreşte determinarea aproximativă, pe cale numerică, a valorii unei integrale definite. La analiză matematică aţi văzut că problema calculului unei integrale definite foloseşte noţiunea de primitivă: dacă pentru funcţia f definită în intervalul $[a, b]$ cunoaştem o primitivă a ei F, atunci integrala definită a acestei funcţii între limitele a şi b se calculează cu formula Newton-Leibniz:

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a).$$

În practică există însă numeroase situaţii în care primitiva F fie are o expresie complicată, fie nu poate fi determinată pe căi elementare. Pe de altă parte, este posibil ca funcţia f să fie dată sub forma unui tabel de valori, situaţie în care însăşi noţiunea de primitivă îşi pierde sensul. În asemenea cazuri, calculul integralei definite cu relaţia anterioară nu mai este posibil, fiind necesară aplicarea unor procedee de integrare numerică, numite şi metode de cuadratură.

În principiu toate metodele de cuadratură numerică calculează o valoare aproximativă a integralei definite pornind de la un număr finit de valori cunoscute ale funcţiei integrant f. Majoritatea metodelor de cuadratură aproximează funcţia integrant, pe întregul interval $[a, b]$ sau pe subdiviziuni ale acestuia, printr-un polinom şi determină valoarea integralei definite aplicând formula Newton-Leibnitz polinomului de aproximare.

Orice formulă de integrare (cuadratură) numerică va avea forma următoare

$$\int_a^b f(x)\,\mathrm{dx} \approx \sum_{i=0}^{n} f(c_i)b_i. \tag{8.1}$$

Numerele c_i se numesc noduri de integrare (şi sunt $n+1$ valori diferite între ele din $[a, b]$) iar b_i se numesc ponderi de integrare (fiind $n+1$ numere reale). Pentru a avea o formulă de integrare numerică precizată avem nevoie să cunoaştem numărul n, nodurile c_i şi ponderile b_i. Formula (8.1) ne dă o aproximare a integralei definite folosind doar valorile funcţiei integrant f în noduri de integrare c_i.

Pentru a avea aproximări satisfăcătoare, nodurile şi ponderile de integrare numerică nu pot fi alese la întâmplare, ci trebuie să verifice o serie de criterii. În acest scop este folosită noţiunea de exactitate. Spunem că formula de integrare (8.1) este **exactă pentru o anumită funcţie** $\tilde{f}$ dacă

$$\int_a^b \tilde{f}(x)\,\mathrm{dx} = \sum_{i=0}^{n} \tilde{f}(c_i)b_i.$$

O formulă de integrare numerică poate fi exactă pentru anumite clase de funcţii, dar nu va fi niciodată exactă pentru orice funcţie. O noţiune des folosită este următoarea.

Definiţia 8.1. *O formulă de integrare numerică are* **ordin de precizie** k *dacă ea este exactă pentru orice polinom de grad mai mic sau egal cu* k *şi există un polinom de grad* $k+1$ *pentru care nu este exactă.*

Observaţia 8.1. *Folosing linearitatea integralei, putem arăta uşor că o formulă de integrare numerică este exactă pentru orice polinom de grad mai mic sau egal decât* r *dacă şi numai dacă este exactă pentru polinoamele elementare* $1, x, \ldots, x^r$. *Prin urmare, o formulă de integrare are ordin de precizie* k *dacă şi numai dacă este exactă pentru polinoamele elementare* $1, x, \ldots, x^k$, *dar nu mai este exactă pentru* x^{k+1}.

Următoarea teoremă ne arată că, dacă formula (8.1) are ordin de precizie suficient de mare, atunci ponderile sunt determinate în mod unic de nodurile de integrare.

Teorema 8.1. *Dacă formula de integrare numerică (8.1) are ordin de precizie* $k \geq n$ *atunci*

$$b_i = \int_a^b l_i(x)\,\mathrm{dx} \qquad (0 \leq i \leq n), \tag{8.2}$$

unde l_i *sunt polinoamele elementare Lagrange pe nodurile* $c_0, c_1, ..., c_n$.

Proof. Fie $(l_i)_{0\leq i\leq n}$ polinoamele elementare Lagrange pe nodurile $(c_i)_{0\leq i\leq n}$. Cum (8.1) este exactă pentru fiecare polinom l_i iar $l_i(c_j) = \delta_{ij}$, rezultă că

$$\int_a^b l_i(x)dx = \sum_{j=0}^{n} l_i(c_j)b_j = b_i \qquad (0 \leq i \leq n),$$

şi demonstraţia se încheie. □

Observaţia 8.2. *Teorema anterioară ne arată că, dacă formula (8.1) are ordin de precizie* $k \geq n$, *atunci*

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=0}^{n} f(c_i)b_i \;=\; \sum_{i=0}^{n} f(c_i)\int_a^b l_i(x)\,\mathrm{d}x \;=\; \int_a^b \left(\sum_{i=0}^{n} f(c_i)l_i(x)\right)\mathrm{d}x \;=\; \int_a^b p_n(x)\,\mathrm{d}x,$$

unde p_n *este polinoamul de interpolare a lui* f *pe nodurile* $c_0, c_1, ..., c_n$.

Să mai observăm că, reciproc, orice formulă de integrare pe $n+1$ *noduri de forma*

$$\int_a^b f(x)\,\mathrm{d}x \;\approx\; \int_a^b p_n(x)\,\mathrm{d}x, \tag{8.3}$$

are ordinul de precizie cel puţin n. *Într-adevăr, ştim că, dacă* f *este un polinom de grad mai mic sau egal ca* n, *atunci el coincide cu polinomul său de interpolare* p_n *şi deci formula* (8.3) *este exactă.*

În final, să notăm că prin (8.3) *înlocuim integrarea funcţiei* f *(care poate fi complicată) cu integrarea polinomului* p_n *(care este foarte uşor de făcut).*

Următoarea definiţie se impune ţinând cont de observaţia de mai sus.

Definiţia 8.2. *O formulă de integrare numerică este* **interpolatoare** *dacă ea este de forma* (8.3).

Prin urmare, o formulă de integrare numerică are ordin de precizie mai mare sau egal ca n dacă şi numai dacă este interpolatoare, iar în acest caz ponderile b_i din (8.1) sunt determinate în mod unic. De acum înainte vom lucra doar cu formule interpolatoare!

Întrebarea următoare este: cum trebuie alese nodurile c_i? S-au impus două strategii diferite de alegere a nodurilor şi, în conformitate, avem două clase importante de metode: Newton-Cotes sau Gauss.

8.1 Formule Newton-Cotes simple

Nodurile de integrare $(c_i)_{0\le i\le n}$ sunt puncte echidistante în intervalul $[a,b]$. Avem două posibilităţi, depinzând dacă sunt luate sau nu capetele intervalului $[a,b]$:

1. Formule închise:
$$h=\frac{b-a}{n},\quad c_i=a+ih\quad (0\le i\le n,\quad n\ge 1).$$
Observăm că $c_0=a$ şi $c_n=b$, deci capetele intervalului de integrare $[a,b]$ sunt noduri de integrare.

2. Formule deschise:
$$h=\frac{b-a}{n+2},\quad c_i=a+(i+1)h\quad (0\le i\le n,\quad n\ge 0).$$
În acest caz, $c_0=a+h$ şi $c_n=b-h$, deci capetele intervalului de integrare $[a,b]$ nu sunt printre nodurile de integrare.

Cu alegerea de mai sus a nodurilor, formulele interpolatoare rezultate se numesc Newton-Cotes.

Dat un număr n, regulile de mai sus ne dau, fiecare, câte o modalitate de alegere a nodurilor de integrare $(c_i)_{0\le i\le n}$. Reamintim că cerem ca formula de integrare numerică să aibă ordin de precizie cel puţin n (deci formula este interpolatoare), prin urmare ponderile $(b_i)_{0\le i\le n}$ se determină în mod unic. În acest caz, pentru fiecare $n\ge 1$, vom avea o metodă Newton-Cotes închisă şi una deschisă. Pentru $n=0$ avem doar o metodă deschisă. Vom descrie în continuare cele mai folosite metode Newton-Cotes.

8.1.1 Formula punctului mediu (formulă Newton-Cotes deschisă cu $n = 0$)

Este o formulă interpolatoare de tip Newton-Cotes deschisă cu $n = 0$. Prin urmare, singurul nod de integrare este $c_0 = \frac{a+b}{2}$ iar formula de aproximare se obţine astfel:

$$\int_a^b f(x)\,dx \approx \int_a^b p_0(x)\,\mathrm{d}x = \int_a^b f(c_0)l_0(x)\,\mathrm{d}x$$

$$= \int_a^b f\left(\frac{a+b}{2}\right)\,\mathrm{d}x = (b-a)f\left(\frac{a+b}{2}\right).$$

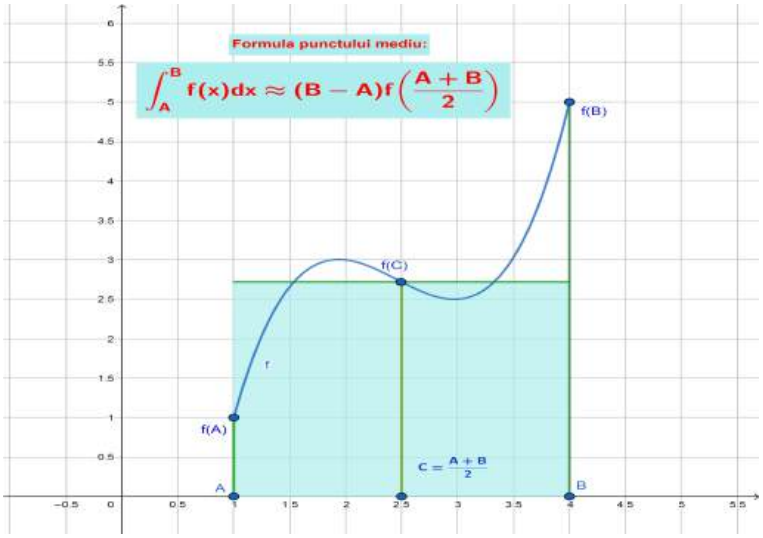


Figura 8.1: Formula punctului mediu pentru aproximarea integralei

Prin urmare, avem următoarea formulă de aproximare Newton-Cotes deschisă (formula punctului mediu, vezi Figura 8.1):

$$\int_a^b f(x)\,\mathrm{d}x \approx (b-a)f\left(\frac{a+b}{2}\right). \tag{8.4}$$

8.1.2 Formula trapezului (formulă Newton-Cotes închisă cu $n = 1$)

Este o formulă interpolatoare de tip Newton-Cotes închisă cu $n = 1$. Prin urmare nodurile de integrare sunt $c_0 = a$ şi $c_1 = b$ iar formula de aproximare se obţine astfel:

$$\int_a^b f(x)\,dx \approx \int_a^b p_1(x)\,\mathrm{d}x = \int_a^b \left(f(c_0)l_0(x) + f(c_1)l_1(x)\right)\,\mathrm{d}x$$

$$= \int_a^b \left(f(c_0)\frac{x-c_1}{c_0-c_1} + f(c_1)\frac{x-c_0}{c_1-c_0}\right)\,\mathrm{d}x = \int_a^b \left(f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a}\right)$$

$$= f(a)\frac{(x-b)^2}{2(a-b)}\Big|_a^b + f(b)\frac{(x-a)^2}{2(b-a)}\Big|_a^b = \frac{b-a}{2}\left(f(a)+f(b)\right).$$

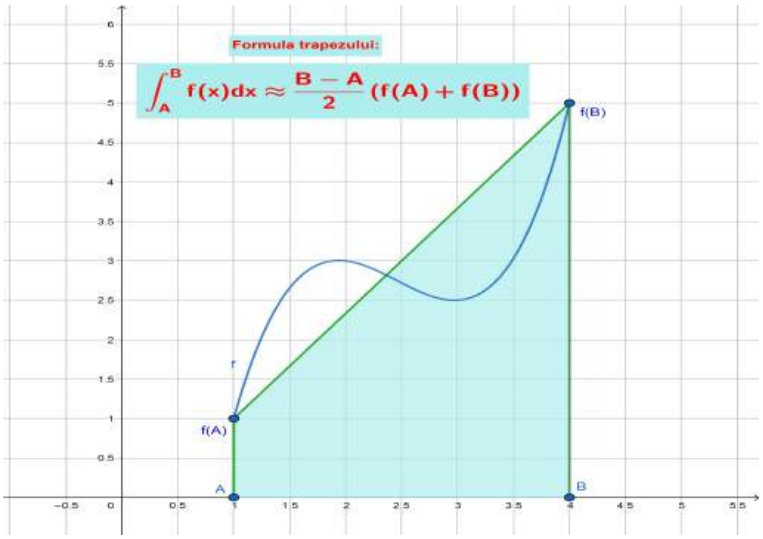


Figura 8.2: Formula trapezului pentru aproximarea integralei

Prin urmare, avem următoarea formulă de aproximare Newton-Cotes închisă (formula trapezului, vezi Figura 8.2):

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{2}\left(f(a)+f(b)\right). \tag{8.5}$$

8.1.3 Formula lui Simpson (formulă Newton-Cotes închisă cu $n = 2$)

Este o formulă interpolatoare de tip Newton-Cotes închisă cu $n = 2$. Prin urmare nodurile de integrare sunt $c_0 = a$, $c_1 = \frac{a+b}{2}$ şi $c_2 = b$ iar formula de aproximare se obţine astfel:

$$\int_a^b f(x)\,dx \approx \int_a^b p_2(x)\,\mathrm{d}x = \int_a^b \left(f(c_0)l_0(x) + f(c_1)l_1(x) + f(c_2)l_2(x)\right)\,\mathrm{d}x$$

$$= \int_a^b \left(f(c_0)\frac{(x-c_1)(x-c_2)}{(c_0-c_1)(c_0-c_2)} + f(c_1)\frac{(x-c_0)(x-c_2)}{(c_1-c_0)(c_1-c_2)} + +f(c_2)\frac{(x-c_0)(x-c_1)}{(c_2-c_0)(c_2-c_1)} \right)\,\mathrm{d}x$$

$$= \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right).$$

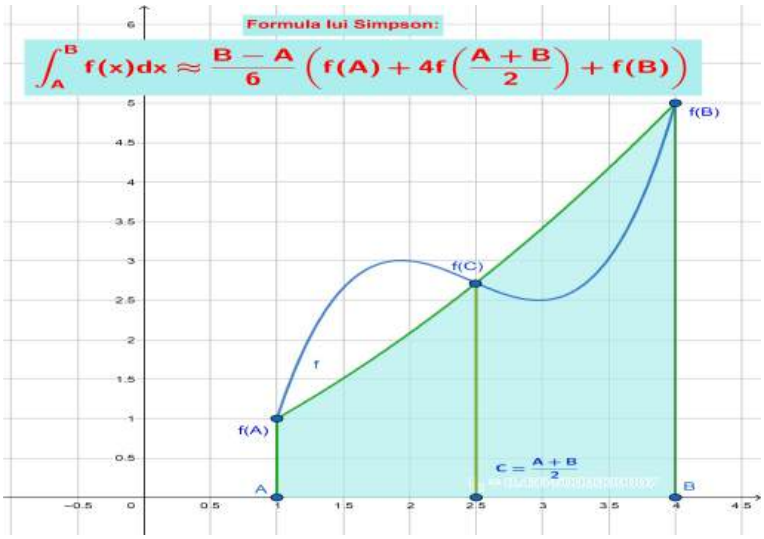


Figura 8.3: Formula lui Simpson pentru aproximarea integralei

Prin urmare, avem următoarea formulă de aproximare Newton-Cotes închisă (formula lui Simpson, vezi Figura 8.3):

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right). \tag{8.6}$$

Cursul 9

Integrare numerică (II)
Formule Newton-Cotes simple şi compuse cu rest

Reamintim că o formulă generală de integrare numerică este de tipul:

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=0}^{n} b_i\, f(c_i). \tag{9.1}$$

Dacă în (9.1) nodurile $(c_i)_{0\le i\le n}$ sunt alese echidistante (cu sau fără capetele a şi b) iar ponderile $(b_i)_{0\le i\le n}$ sunt date de integralele polinoamelor de interpolare elementare Lagrange (sau, echivalent, formula (9.1) este interpolatoare), atunci formula (9.1) este de tip Newton-Cotes. Am dedus în cursul precedent trei formule Newton-Cotes simple:

- $\displaystyle\int_a^b f(x)\,\mathrm{d}x \approx (b-a) f\left(\frac{a+b}{2}\right) := F_0(f)$
  Formula punctului mediu simplă - Formulă Newton-Cotes deschisă cu $n = 0$.

- $\displaystyle\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{2}\,(f(a)+f(b)) := F_1(f)$
  Formula trapezului simplă - Formulă Newton-Cotes închisă cu $n = 1$.

- $\displaystyle\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right) := F_2(f)$
  Formula lui Simpson simplă - Formulă Newton-Cotes închisă cu $n = 2$.

În cele ce urmează vom determina expresii ale diferenţelor între valorile exacte şi cele aproximative ale integralei. Acestea ne vor ajuta să estimăm eroarea în aproximare. Mai precis, vom determina, pentru fiecare $i \in \{0, 1, 2\}$ expresiile resturilor $R_i(f)$, astfel ca

$$\int_a^b f(x)\,\mathrm{d}x = F_i(f) + R_i(f). \tag{9.2}$$

9.1 Formula trapezului simplă cu rest

În formula trapezului simplă am aproximat integrala lui f prin integrala lui p_1, polinomul de interpolare a lui f pe nodurile a şi b:

$$\int_a^b f(x)\ \mathrm{dx} \approx \int_a^b p_1(x)\,\mathrm{d}x.$$

Pentru evaluarea restului considerăm $f \in C^2[a,b]$ şi ţinem cont de expresia erorii în interpolare:

$$f(x) - p_1(x) = \frac{1}{2!} f''(\xi(x))(x-a)(x-b),$$

unde $\xi(x)$ este un număr din (a,b) care depinde de x. Deci

$$R_1(f) := \int_a^b f(x)\ \mathrm{d}x - \int_a^b p_1(x)\ \mathrm{d}x = \int_a^b \frac{f^{(2)}(\xi(x))}{2!}(x-a)(x-b)\ \mathrm{d}x. \tag{9.3}$$

Cum $f \in C^2[a,b]$, există $m_2 = \inf_{x\in[a,b]} f^{(2)}(x)$, $M_2 = \sup_{x\in[a,b]} f^{(2)}(x)$ şi

$$\frac{M_2}{2}\int_a^b (x-a)(x-b)\ \mathrm{d}x \le \int_a^b \frac{f^{(2)}(\xi(x))}{2!}(x-a)(x-b)\ \mathrm{d}x \le \frac{m_2}{2}\int_a^b (x-a)(x-b)\ \mathrm{d}x.$$

Prin urmare

$$m_2 \le \frac{2}{\int_a^b (x-a)(x-b)} \int_a^b \frac{f^{(2)}(\xi(x))}{2!}(x-a)(x-b)\ \mathrm{d}x \le M_2.$$

Din continuitatea lui $f^{(2)}$ deducem existenţa unui număr $\xi \in [a,b]$ astfel ca

$$f^{(2)}(\xi) = \frac{2}{\int_a^b (x-a)(x-b)} \int_a^b \frac{f^{(2)}(\xi(x))}{2!}(x-a)(x-b)\ \mathrm{d}x. \tag{9.4}$$

Folosind (9.4) în (9.3), rezultă că

$$R_1(f) = \frac{f^{(2)}(\xi)}{2}\int_a^b (x-a)(x-b)\,\mathrm{d}x = -\frac{f^{(2)}(\xi)}{12}(b-a)^3. \tag{9.5}$$

Obţinem următoarea formulă a trapezului simplă cu rest:

$$\int_a^b f(x)\ \mathrm{d}x = \frac{b-a}{2}\left(f(a)+f(b)\right) - \frac{f^{(2)}(\xi)}{12}(b-a)^3. \tag{9.6}$$

9.2 Formula punctului mediu simplă cu rest

În formula punctului mediu simplă se aproximează integrala lui f cu integrala lui p_0, polinomul de interpolare a lui f pe nodul $c_0 = \frac{a+b}{2}$:

$$\int_a^b f(x)\ \mathrm{dx} \approx \int_a^b p_0(x)\,\mathrm{d}x = (b-a) f\left(c_0\right).$$

Pentru evaluarea restului, considerăm că $f \in C^2[a,b]$ şi folosim formula lui Taylor în c_0:

$$f(x) = f(c_0) + (x-c_0)f^{(1)}(c_0) + \frac{1}{2}f^{(2)}(\xi(x))(x-c_0)^2,$$

de unde, prin integrare şi observând că $\int_a^b (x-c_0)\,\mathrm{d}x = 0$, obţinem că

$$\int_a^b f(x)\,\mathrm{d}x = (b-a)f(c_0) + \frac{1}{2}\int_a^b f^{(2)}(\xi(x))(x-c_0)^2\,\mathrm{d}x = \frac{1}{2}f^{(2)}(\xi)\int_a^b (x-c_0)^2\,\mathrm{d}x = f^{(2)}(\xi)\frac{(b-a)^3}{24}.$$

Valoarea ξ de mai sus reprezintă un număr din $[a,b]$ rezultat din proprietatea de continuitate a lui $f^{(2)}$ ca la metoda trapezului.

Obţinem următoarea formulă a punctului mediu simplă cu rest:

$$\int_a^b f(x)\,\mathrm{d}x = (b-a)f\left(\frac{a+b}{2}\right) + \frac{f^{(2)}(\xi)}{24}(b-a)^3. \tag{9.7}$$

9.3 Formula lui Simpson

În formula lui Simpson simplă se aproximează integrala lui f cu integrala lui p_2, polinomul de interpolare a lui f pe nodurile $c_0 = a$, $c_1 = \frac{a+b}{2}$ şi $c_1 = b$:

$$\int_a^b f(x)\,\mathrm{dx} \approx \int_a^b p_2(x)\,\mathrm{d}x = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right).$$

Pentru evaluarea restului, considerăm că $f \in C^4[a,b]$ şi folosim formula lui Taylor în c_1:

$$f(x) = f(c_1) + (x-c_1)f'(c_1) + \frac{1}{2}f^{(2)}(c_1)(x-c_1)^2 + \frac{1}{6}f^{(3)}(c_1)(x-c_1)^3 + \frac{1}{24}f^{(4)}(\xi'(x))(x-c_1)^4.$$

Prin integrare, ţinând cont că $\int_a^b (x-c_1)\,\mathrm{d}x = \int_a^b (x-c_1)^3\,\mathrm{d}x = 0$, rezultă că

$$\int_a^b f(x)\,\mathrm{d}x = (b-a)f(c_1) + \frac{f^{(2)}(c_1)}{2}\int_a^b (x-c_1)^2\,\mathrm{d}x + \frac{f^{(4)}(\xi')}{24}\int_a^b (x-c_1)^4\,\mathrm{d}x =$$

$$= (b-a)f(c_1) + \frac{f^{(2)}(c_1)}{12}(b-a)^3 + \frac{f^{(4)}(\xi')}{1920}(b-a)^5,$$

unde ξ' este un număr din $[a,b]$.

Acum, folosind faptul că, pentru o funcţie $f \in C^4[a,b]$, avem că există $\xi'' \in [a,b]$ astfel încât

$$f^{(2)}(c_1) = \frac{4}{(b-a)^2}(f(a) - 2f(c_1) + f(b)) - \frac{(b-a)^2}{96}f^{(4)})(\xi'')$$

şi substituind în formula de mai sus avem că

$$\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{6}\left(f(a) + 4f(c_1) + f(b)\right) - \frac{f^{(4)})(\xi'')}{1152}(b-a)^5 + \frac{f^{(4)}(\xi')}{1920}(b-a)^5.$$

Cum $f \in C^4[a,b]$, rezultă că există $\xi \in [a,b]$ astfel încât

$$-\frac{1}{3}f^{(4)})(\xi'') + \frac{1}{5}f^{(4)}(\xi') = -\frac{2}{15}f^{(4)}(\xi).$$

Din ultima relaţie şi expresia integralei de mai sus, obţinem următoarea formulă a lui Simpson simplă cu rest:

$$\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right) - \frac{1}{2880}f^{(4)}(\xi)(b-a)^5. \tag{9.8}$$

Am demonstrat că restul în fiecare dintre cele trei formule de bază simple este dat de:

1. Formula punctului mediu:

$$R_0(f) = \frac{f^{(2)}(\xi)}{2!}\int_a^b x(x-\frac{a+b}{2})\,\mathrm{d}x = \frac{f^{(2)}(\xi)}{24}(b-a)^3.$$

2. Formula trapezului:

$$R_1(f) = \frac{f^{(2)}(\xi)}{2!}\int_a^b (x-a)(x-b)\,\mathrm{d}x = -\frac{f^{(2)}(\xi)}{12}(b-a)^3.$$

3. Formula lui Simpson:

$$R_2(f) = \frac{f^{(4)}(\xi)}{4!}\int_a^b x(x-a)(x-\frac{a+b}{2})(x-b)\,\mathrm{d}x = -\frac{f^{(4)}(\xi)}{2880}(b-a)^5.$$

9.4 Formulele Newton-Cotes compuse

Formulele Newton-Cotes prezentate mai sus nu permit obţinerea unor aproximări din ce în ce mai precise. Putem însă să realizăm acest lucru cu următoarea strategie:

- Împărţim intervalul de integrare $[a,b]$ în N subintervale egale $[x_i, x_{i+1}]$, $0 \le i \le N-1$, unde x_i sunt puncte echidistante în $[a,b]$: $x_i = a + i\dfrac{b-a}{N} \qquad (0 \le i \le N)$.



- Pe fiecare interval mic $[x_i, x_{i+1}]$ folosim una dintre formulele simple de mai sus şi însumăm.
- Obţinem o formulă de aproximare pe intervalul $[a,b]$ care va depinde de numărul de subintervale N. Vom vedea că această formulă oferă aproximari din ce în ce mai bune a când N este din ce în ce mai mare.

9.4.1 Formula punctului mediu compusă

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x \\ &= \sum_{i=0}^{N-1} \left[(x_{i+1}-x_i) f\left(\frac{x_i+x_{i+1}}{2}\right) + \frac{f^{(2)}(\xi_i)}{24}(x_{i+1}-x_i)^3 \right] \\ &= \frac{b-a}{N} \sum_{i=0}^{N-1} f\left(\frac{x_i+x_{i+1}}{2}\right) + \frac{(b-a)^3}{24N^3} \sum_{i=0}^{N-1} f^{(2)}(\xi_i) \\ &= \frac{b-a}{N} \sum_{i=0}^{N-1} f\left(\frac{x_i+x_{i+1}}{2}\right) + \frac{(b-a)^3}{24N^2} f^{(2)}(\xi).\end{aligned}$$

Prin urmare, avem următoarea formulă a punctului mediu compusă:

$$\int_a^b f(x)\ \mathrm{d}x = \underbrace{\frac{b-a}{N} \sum_{i=0}^{N-1} f\left(\frac{x_i+x_{i+1}}{2}\right)}_{Aproximare} + \underbrace{\frac{(b-a)^3}{24N^2} f^{(2)}(\xi)}_{Rest}. \tag{9.9}$$

Observăm că partea numită *Aproximare* se poate calcula iar partea numită *Rest* se majorează cu

$$\frac{(b-a)^3}{24N^2} \sup_{x\in[a,b]} \left| f^{(4)}(x) \right|$$

şi tinde la zero cu viteza cu care $\frac{1}{N^2}$ tinde la zero. În consecinţă avem o aproximare a integralei cu o eroare care tinde la zero când N tinde la infinit.

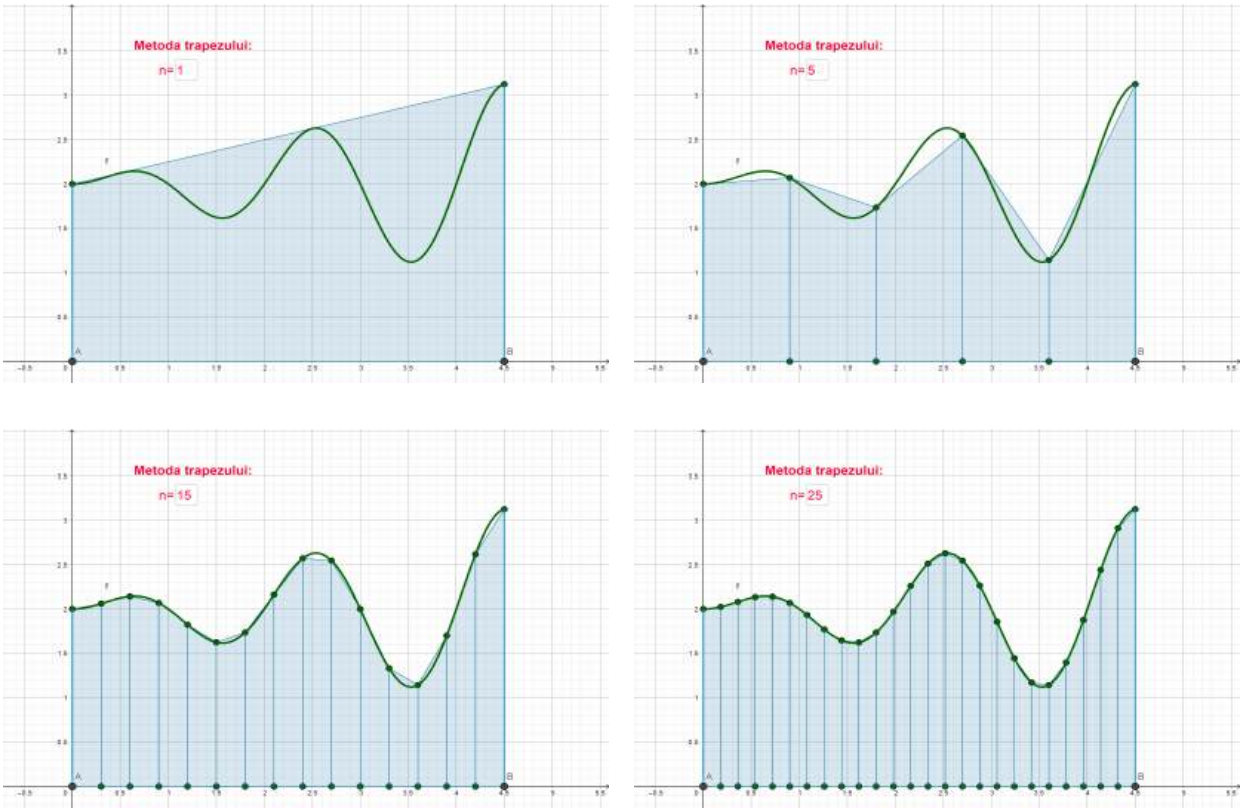


Figura 9.1: Formula trapezului compusă cu $N = 1, 5, 15, 25$

9.4.2 Formula trapezului compusă

$$\begin{aligned}
\int_a^b f(x)\,\mathrm{d}x &= \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x \\
&= \sum_{i=0}^{N-1} \left[\frac{x_{i+1}-x_i}{2} \left(f\left(x_i\right) + f\left(x_{i+1}\right) \right) - \frac{f^{(2)}(\xi_i)}{12} (x_{i+1}-x_i)^3 \right] \\
&= \frac{b-a}{2N} \sum_{i=0}^{N-1} \left(f\left(x_i\right) + f\left(x_{i+1}\right) \right) - \frac{(b-a)^3}{12N^3} \sum_{i=0}^{N-1} f^{(2)}(\xi_i) \\
&= \frac{b-a}{2N} \left(f(a) + f(b) + 2 \sum_{i=1}^{N-1} f\left(x_i\right) \right) - \frac{(b-a)^3}{12N^2} f^{(2)}(\xi).
\end{aligned}$$

Prin urmare, avem următoarea formulă a trepezului compusă:

$$\int_a^b f(x)\ \mathrm{d}x = \underbrace{\frac{b-a}{2N} \left(f(a) + f(b) + 2 \sum_{i=1}^{N-1} f\left(x_i\right) \right)}_{Aproximare} - \underbrace{\frac{(b-a)^3}{12N^2} f^{(2)}(\xi)}_{\text{Rest}}. \tag{9.10}$$

Observăm că partea numită *Aproximare* se poate calcula iar partea numită *Rest* se majorează cu

$$\frac{(b-a)^3}{12N^2} \sup_{x\in[a,b]} \left|f^{(2)}(x)\right|$$

şi tinde la zero cu viteza cu care $\frac{1}{N^2}$ tinde la zero. În consecinţă avem o aproximare a integralei cu o eroare care tinde la zero când N tinde la infinit. Viteza cu care eroarea tinde la zero în formula trapezului compusă este similară cu cea a formulei punctului mediu precedentă.

9.4.3 Formula lui Simpson compusă

$$\begin{aligned}
\int_a^b f(x)\,\mathrm{d}x &= \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x \\
&= \sum_{i=0}^{N-1} \left[\frac{x_{i+1}-x_i}{6} \left(f\left(x_i\right) + 4f\left(\frac{x_{i+1}+x_i}{2}\right) + f\left(x_{i+1}\right)\right) - \frac{f^{(4)}(\xi_i)}{2880}(x_{i+1}-x_i)^5 \right] \\
&= \frac{b-a}{6N} \sum_{i=0}^{N-1} \left(f\left(x_i\right) + 4f\left(\frac{x_{i+1}+x_i}{2}\right) + f\left(x_{i+1}\right)\right) - \frac{(b-a)^5}{2880N^5} \sum_{i=0}^{N-1} f^{(4)}(\xi_i) \\
&= \frac{b-a}{6N} \left(f(a) + f(b) + 2\sum_{i=1}^{N-1} f\left(x_i\right) + 4\sum_{i=0}^{N-1} f\left(\frac{x_i+x_{i+1}}{2}\right)\right) - \frac{(b-a)^5}{2880N^4} f^{(4)}(\xi).
\end{aligned}$$

Prin urmare, avem următoarea formulă a lui Simpson compusă:

(9.11)

$$\int_a^b f(x)\ \mathrm{d}x = \underbrace{\frac{b-a}{6N} \left(f(a) + f(b) + 2\sum_{i=1}^{N-1} f\left(x_i\right) + 4\sum_{i=0}^{N-1} f\left(\frac{x_i+x_{i+1}}{2}\right)\right)}_{Aproximare} \underbrace{- \frac{(b-a)^5}{2880N^4} f^{(4)}(\xi)}_{Rest}.$$

Observăm că partea numită *Aproximare* se poate calcula iar partea numită *Rest* se majorează cu

$$\frac{(b-a)^5}{2880N^4} \sup_{x\in[a,b]} \left|f^{(4)}(x)\right|$$

şi tinde la zero cu viteza cu care $\frac{1}{N^4}$ tinde la zero. În consecinţă avem o aproximare a integralei cu o eroare care tinde la zero când N tinde la infinit. Viteza cu care eroarea tinde la zero în formula lui Simpson compusă este mai mare decât în cele două formule precedente.

Cursul 10

Integrare numerică
Metoda de integrare Gauss

Ne amintim că o formulă de integrare numerică

$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=0}^{n} f(c_i)b_i, \tag{10.1}$$

este **exactă pentru o anumită funcţie** $\tilde{f}$ dacă, punând în loc de f pe $\tilde{f}$, în (10.1) avem egalitate. În plus, formula de integrare numerică are **ordin de precizie** k dacă ea este exactă pentru orice polinom de grad mai mic sau egal cu k şi există un polinom de grad $k+1$ pentru care nu este exactă. De asemenea, am văzut în cursurile precedente că formula (10.1) are ordin de precizie cel puţin n dacă şi numai dacă

$$b_i = \int_a^b l_i(x)\,\mathrm{d}x \qquad (0 \le i \le n), \tag{10.2}$$

unde l_i sunt polinoamele elementare Lagrange pe nodurile de integrare $c_0, c_1, ..., c_n$. Astfel, formula (10.1) este **interpolatoare**, adică este de forma

$$\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b p_n(x)\,\mathrm{d}x, \tag{10.3}$$

unde p_n este polinoamul de interpolare a lui f pe nodurile $c_0, c_1, ..., c_n$.

Prin urmare, ponderile de integrare $(b_i)_{0\le i\le n}$ în (10.1) se determină în mod unic pentru orice formulă cu ordin de precizie mai mare sau egal decât n (sau, echivalent, pentru orice formulă interpolatoare). În cele ce urmează dorim să răspundem la întrebarea: cum trebuie alese nodurile de integrare $(c_i)_{0\le i\le n}$ pentru ca formula (10.1) să aibă ordin de precizie maxim?

10.1 Polinoame ortogonale

În cele ce urmează vom nota cu $\mathbb{R}[X]$ mulţimea polinoamelor cu coeficienţi reali şi cu $\mathbb{P}_N[X]$ mulţimea polinoamelor de grad cel mult N cu coeficienţi reali.

Definiţia 10.1. *O pondere în intervalul (a,b) este o funcţie continuă $\omega : (a,b) \to \mathbb{R}$ cu proprietăţile:*

1. $\omega(x) > 0, \quad \forall x \in (a,b)$

2. $\int_a^b |x|^k \omega(x)\,\mathrm{d}x < \infty, \quad \forall k \geq 0.$

Exemple:

- $\omega(x) = 1$ în $(-1, 1)$
- $\omega(x) = \frac{1}{\sqrt{1-x^2}}$ în $(-1, 1)$
- $\omega(x) = \exp(-x)$ în $(0, \infty)$
- $\omega(x) = \exp(-x^2)$ în $(-\infty, \infty)$.

Definim spaţiul vectorial

$$\mathcal{S} = \left\{ f : (a,b) \to \mathbb{R} \,:\, f \text{ integrabilă şi } \int_a^b |f(x)|^2 \omega(x)\,\mathrm{d}x < \infty \right\}.$$

Pentru orice f şi g din $\mathcal{S}$ vom nota

$$\langle f, g \rangle = \int_a^b f(x) g(x) \omega(x)\,\mathrm{d}x, \tag{10.4}$$

şi-l vom numi produsul scalar al lui f cu g în $\mathcal{S}$.

Observaţia 10.1. *Datorită proprietăţilor unei ponderi, mulţimea $\mathbb{R}[X]$ este conţinută în $\mathcal{S}$.*

Observaţia 10.2. *Dacă $f \in \mathcal{S}$ este o funcţie continuă astfel încât $\langle f, f \rangle = 0$, atunci $f \equiv 0$.*

Următoarea noţiune va juca un rol important în cele ce urmează.

Definiţia 10.2. *Dată o pondere ω pe (a, b), un şir de polinoame $(P_n)_{n\geq 0} \subset \mathbb{R}[X]$ cu proprietăţile:*

1. *$grad\,(P_n) = n, \quad n \geq 0$*
2. *Coeficientul dominant al lui P_n este egal cu 1, $\quad n \geq 0$*
3. *$\langle P_n, P_m \rangle = 0, \quad n, m \geq 0,\ n \neq m$*

se numeşte **şir de polinoame ortogonale în raport cu ponderea ω.**

Observaţia 10.3. *Orice şir de polinoamele ortogonale $(P_n)_{n\geq 0}$ are următoarele proprietăţi elementare:*

1. *Pentru orice $n \geq 0$, $\langle P_n, P_n \rangle \neq 0$.*
2. *Pentru orice $N \geq 0$, mulţimea $(P_n)_{0\leq n\leq N}$ este liniar independentă;*
3. *Pentru orice $N \geq 0$, avem că*

$$\sum_{n=0}^{N} a_n P_n(x) = \sum_{n=0}^{N} b_n P_n(x) \Rightarrow a_n = b_n,\ 0 \leq n \leq N;$$

4. Pentru orice $N \geq 0$ există $(a_n^N)_{0\leq n\leq N}$ unici astfel ca

$$x^N = \sum_{n=0}^{N} a_n^N P_n(x).$$

(Inducţie după N)

5. Pentru orice $q \in \mathbb{P}_N[X]$ există unici $(a_n)_{0\leq n\leq N}$ astfel ca

$$q(x) = \sum_{n=0}^{N} a_n P_n(x).$$

În plus, $a_n = a_n(q) = \dfrac{\langle q, P_n\rangle}{\langle P_n, P_n\rangle}$;

6. Mulţimea $(P_n)_{0\leq n\leq N}$ este o bază ortogonală a spaţiului vectorial $\mathbb{P}_N[X]$;

7. Pentru orice $n > m$ avem că

$$\langle P_n, x^m\rangle = 0.$$

Mai mult, pentru orice polinom q de grad mai mic strict ca n avem că

$$\langle P_n, q\rangle = 0;$$

Următoarea teoremă ne arată existenţa şi unicitatea şirului de polinoame ortogonale. De asemenea, ne oferă o relaţie de recurenţă pe baza căreia se pot determina termenii şirului de polinoame ortogonale.

Teorema 10.1. *Dată o ponderе ω în (a, b) există un unic şir $(P_n)_{n\geq 0}$ de polinoame ortogonale în raport cu ponderea ω pe (a, b). În plus, polinoamele verifică relaţia de recurenţă*

$$P_{n+1}(x) = (x - a_{n+1})P_n(x) - b_{n+1}P_{n-1}(x), \quad n \geq 0 \tag{10.5}$$

unde, prin convenţie $P_{-1}(x) = 0$, prin definiţie $P_0(x) = 1$ şi

$$a_{n+1} = \frac{\langle xP_n(x), P_n(x)\rangle}{\langle P_n(x), P_n(x)\rangle} \quad (n \geq 0), \qquad b_{n+1} = \frac{\langle P_n(x), P_n(x)\rangle}{\langle P_{n-1}(x), P_{n-1}(x)\rangle} \quad (n \geq 1). \tag{10.6}$$

Proof. Vom determina termenii şirului $(P_n)_{n\geq 0}$ de polinoame ortogonale în raport cu ponderea ω pe (a, b) din aproape în aproape.

- Prin definiţie, $P_0 = 1$.
- Avem că $P_1(x) = x - a_1$ cu $a_1 \in \mathbb{R}$. Din relaţia $\langle P_0, P_1\rangle = 0$ obţinem $\langle a_1P_0, P_0\rangle = \langle x, P_0\rangle$, şi relaţiile de mai sus sunt verificare.
- Presupunem cunoscute $P_0, P_1, ..., P_n$ şi construim P_{n+1}. Polinomul $P_{n+1}(x) - xP_n(x)$ are grad mai mic sau egal decât n şi, conform Observaţiei 10.3, punctul *5.*, se poate scrie în mod unic sub forma

$$P_{n+1}(x) - xP_n(x) = \sum_{k=0}^{n} \alpha_k P_k(x).$$

Vom arăta că relaţiile de ortogonalitate 3. din Definiţia 10.2 a şirului de polinoame ortogonale determină în mod unic coeficienţii α_k. Într-adevăr,

$$0 = \langle P_{n+1}(x), P_n(x)\rangle = \langle xP_n(x) + \sum_{k=0}^{n} \alpha_k P_k(x), P_n(x)\rangle = \langle xP_n(x), P_n(x)\rangle + \langle \alpha_n P_n(x), P_n(x)\rangle$$

şi prin urmare

$$a_{n+1} = -\alpha_n = \frac{\langle xP_n(x), P_n(x)\rangle}{\langle P_n(x), P_n(x)\rangle}.$$

De asemenea, din condiţia

$$0 = \langle P_{n+1}(x), P_{n-1}(x)\rangle = \langle xP_n(x) + \sum_{k=0}^{n} \alpha_k P_k(x), P_n(x)\rangle =$$

$$= \langle xP_n(x), P_{n-1}(x)\rangle + \langle \alpha_{n-1} P_{n-1}(x), P_{n-1}(x)\rangle$$

rezultă că

$$-\alpha_{n-1} = \frac{\langle xP_n(x), P_{n-1}(x)\rangle}{\langle P_{n-1}(x), P_{n-1}(x)\rangle} = \frac{\langle P_n(x), xP_{n-1}(x)\rangle}{\langle P_{n-1}(x), P_{n-1}(x)\rangle} = \frac{\langle P_n(x), P_n(x) + q_{n-1}(x)\rangle}{\langle P_{n-1}(x), P_{n-1}(x)\rangle}.$$

q_{n-1} este un polinom de grad mai mic sau egal decât $n-1$ şi prin urmare $\langle q_{n-1}, P_n\rangle = 0$. Rezultă că

$$b_{n+1} = -\alpha_{n-1} = \frac{\langle P_n(x), P_n(x)\rangle}{\langle P_{n-1}(x), P_{n-1}(x)\rangle}.$$

În plus, pentru $j < n-1$,

$$0 = \langle P_{n+1}(x), P_j(x)\rangle = \langle xP_n(x) + \sum_{k=0}^{n} \alpha_k P_k(x), P_j(x)\rangle = \langle xP_n(x), P_j(x)\rangle + \langle \alpha_j P_j(x), P_j(x)\rangle =$$

$$= \langle P_n(x), xP_j(x)\rangle + \langle \alpha_j P_j(x), P_j(x)\rangle = \alpha_j \langle P_j(x), P_j(x)\rangle$$

şi prin urmare $\alpha_j = 0$.

Astfel, relaţiile (10.5)-(10.6) sunt verificate şi demonstraţia se încheie. □

Următoarea teoremă ne dă informaţii asupra rădăcinilor polinoamelor ortogonale.

Teorema 10.2. *Fiecare polinom P_n are toate rădăcinile reale şi simple în intervalul (a, b).*

Proof. Fie $x_1, ..., x_r$ rădăcinile reale din (a, b) ale lui P_n, unde P_n îşi schimbă semnul.

Dacă $r = n$ toate rădăcinile lui P_n sunt reale, simple şi aparţin lui (a, b).

Dacă $r < n$, definim $g(x) = (x - x_1)...(x - x_r)$ şi avem că

$$0 = \langle P_n(x), g(x)\rangle = \int_a^b P_n(x)g(x)\omega(x)\,\mathrm{d}x.$$

Aceasta este însă o contradicţie deoarece $P_n(x)g(x)$ este o funcţie continuă nenulă şi nu îşi schimbă semnul în (a, b). □

Observaţia 10.4. *Dacă $x_1 < x_2 < ... < x_{n-1}$ sunt rădăcinile lui P_{n-1} iar $y_1 < y_2 < ... < y_{n-1} < y_n$ sunt rădăcinile lui P_n atunci $-1 < y_1 < x_1 < y_2 < x_2 < ... < y_{n-1} < x_{n-1} < y_n < 1$.*

Am văzut că, pentru fiecare pondere ω în $[a, b]$, există un unic şir $(P_n)_{n\geq 0}$ de polinoame ortogonale în raport cu ponderea ω pe $[a, b]$. Cele mai întâlnite cazuri sunt următoarele:

- $(a, b) = (-1, 1)$ şi $\omega(x) = 1$. Polinoamele Legendre care verifică

$$P_0 = 1, \quad P_1(x) = x, \quad (n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x).$$

  Formula de recurenţă care dă coeficient dominat 1 este

$$P_{-1} = 0, \quad P_0 = 1, \quad P_1(x) = x, \quad P_n(x) = xP_{n-1}(x) - \frac{(n-1)^2}{(2n-3)(2n-1)}P_{n-2}(x).$$

- $(a, b) = (-1, 1)$ şi $\omega(x) = \frac{1}{\sqrt{1-x^2}}$. Polinoamele Cebîşev care verifică

$$T_0 = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

- $(a, b) = (-\infty, \infty)$ şi $\omega(x) = e^{-x^2}$. Polinoamele Hermite care verifică

$$H_0 = 1, \quad H_1(x) = 2x, \quad H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x).$$

10.2 Metoda de integrare Gauss

Avem acum toate ingredientele necesare pentru a răspunde la întrebarea de la început: cum trebuie alese nodurile de integrare $(c_i)_{0\leq i\leq n}$ pentru ca formula (10.1) să aibă ordin de precizie maxim?

Teorema 10.3. *Formula de integrare numerică (10.1) are ordin de precizie maxim $2n + 1$ care se obţine dacă şi numai dacă $(c_i)_{0\leq i\leq n}$ sunt zerourile polinomului P_{n+1} din şirul de polinoame ortogonale pe (a, b) în raport cu ponderea $\omega = 1$.*

Proof. Să remarcăm mai întâi că dacă f este un polinom de grad $r \geq n + 1$ şi p_n este polinomul său de interpolare pe nodurile $c_0, c_1, ..., c_n$, atunci

$$f(x) - p_n(x) = (x - c_0)...(x - c_n)q(x),$$

unde q este un polinom de grad $r - n - 1 \geq 0$.

Rezultă că formula (10.1) are ordin $r \geq n$ dacă şi numai dacă

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b p_n(x)\,\mathrm{d}x \Leftrightarrow \int_a^b (x - c_0)...(x - c_n)q(x)\,\mathrm{d}x = 0, \tag{10.7}$$

pentru orice polinom q de grad mai mic sau egal decât $r - n - 1$. Prin urmare polinomul $(x - c_0)...(x - c_n)$ este ortogonal oricărui polinom de grad mai mic sau egal decât $r - n - 1$.

De aici obţinem în primul rând că $r \leq 2n + 1$, altfel $(x - c_0)...(x - c_n)$ ar trebui să fie ortogonal sie însuşi.

Pe de altă parte, dacă $r = 2n+1$, relaţia de ortogonalitate (10.7) se verifică pentru orice polinom q de grad mai mic sau egal decât n dacă şi numai dacă $(x - c_0)...(x - c_n) = P_{n+1}(x)$, al $(n+1)-$lea polinom ortogonal în raport cu ponderea $\omega = 1$ pe (a, b). □

Următoarea teoremă ne oferă o expresie a erorii în metoda de integrare Gauss.

Teorema 10.4. *Dacă $f \in C^{2n+2}[a,b]$ iar p_n este polinomul de interpolare a lui f pe nodurile $a < c_0 < c_1 < ... < c_n < b$ care sunt şi zerourile polinomului P_{n+1} din şirul de polinoame ortogonale în raport cu ponderea $\omega = 1$ pe (a,b), atunci există $\xi \in (a,b)$ astfel încât*

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b p_n(x)\,\mathrm{d}x + \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\int_a^b \varphi_{n+1}^2(x)\,\mathrm{d}x. \tag{10.8}$$

Proof. Pornim de la formula de interpolare Hermite pe nodurile c_i

$$f(x) = H_{2n+1}(x) + \frac{f^{(2n+2)}(\xi(x))}{(2n+2)!}\varphi_{n+1}^2(x)$$

unde $H_{2n+1}(x)$ este polinomul de interpolare Hermite de grad $2n+1$. Obţinem că

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=0}^n f(c_i)\int_a^b h_i(x)\,\mathrm{d}x + \sum_{i=0}^n f'(c_i)\int_a^b \widehat{h}_i(x)\,\mathrm{d}x + \int_a^b \frac{f^{(2n+2)}(\xi(x))}{(2n+2)!}\varphi_{n+1}^2(x)\,\mathrm{d}x.$$

Folosind a doua formulă de medie, rezultă că există $\xi \in (a,b)$ astfel încât

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=0}^n f(c_i)\int_a^b h_i(x)\,\mathrm{d}x + \sum_{i=0}^n f'(c_i)\int_a^b \widehat{h}_i(x)\,\mathrm{d}x + \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\int_a^b \varphi_{n+1}^2(x)\,\mathrm{d}x.$$

Cum $\widehat{h}_i(c_j) = 0$, $0 \le j \le n$, avem că $\widehat{h}_i(x) = q_i(x)P_{n+1}(x)$ unde P_{n+1} este polinomul ortogonal de grad $n+1$ iar q_i polinom de grad n. Rezultă că

$$\int_a^b \widehat{h}_i(x)\,\mathrm{d}x = \int_a^b q_i(x)P_{n+1}(x)\,\mathrm{d}x = 0.$$

Pe de altă parte, $h_i(x) - l_i(x)$ se divide prin $P_{n+1}(x)$ şi prin urmare $h_i(x) - l_i(x) = q_i(x)P_{n+1}(x)$. Rezultă că

$$\int_a^b (h_i(x) - l_i(x))\,\mathrm{d}x = \int_a^b q_i(x)P_{n+1}(x)\,\mathrm{d}x = 0 \Leftrightarrow \int_a^b h_i(x)\,\mathrm{d}x = \int_a^b l_i(x)\,\mathrm{d}x.$$

Formula (10.8) este astfel demonstrată. □

Observaţia 10.5. *Din formula restului în (10.8) vedem că ordinul de precizie este $2n+1$.*

10.3 Determinarea formulei Gauss-Legendre de ordin 5

Dorim acum determinarea formulei de integrare pe trei noduri care să aibă ordinul de precizie maxim. Conform Teoremei 10.3, dacă $n = 2$, formula de integrare numerică (10.1) are ordin de precizie maxim 5 şi se obţine dacă $(c_i)_{0\le i\le n}$ sunt zerourile polinomului P_3 din şirul de polinoame ortogonale pe (a,b) în raport cu ponderea $\omega = 1$.

Formula de integrare se determină iniţial pe intervalul $(-1,1)$. Pentru aceasta, în primul rând, vom determina polinomul Legendre de grad 3 folosind formula de recurenţă (10.5)-(10.6).

- Polinomul Legrendre de grad zero este, prin definiţie:

$$P_0(x) = 1.$$

- Polinomul Legrendre de grad unu este, conform (10.5), $P_1(x) = (x - a_1)P_0(x) - b_1 P_{-1}(x)$, cu

$$a_1 = \frac{\langle xP_0(x), P_0(x)\rangle}{\langle P_0(x), P_0(x)\rangle} = \frac{\int_{-1}^{1} x\,\mathrm{d}x}{\int_{-1}^{1}\mathrm{d}x} = 0.$$

Prin urmare,

$$P_1(x) = x.$$

- Polinomul Legrendre de grad doi este, conform (10.5), $P_2(x) = (x - a_2)P_1(x) - b_2 P_0(x)$, iar

$$a_2 = \frac{\langle xP_1(x), P_1(x)\rangle}{\langle P_1(x), P_1(x)\rangle} = \frac{\int_{-1}^{1} x^3\,\mathrm{d}x}{\int_{-1}^{1} x^2\,\mathrm{d}x} = 0, \quad b_2 = \frac{\langle P_1(x), P_1(x)\rangle}{\langle P_0(x), P_0(x)\rangle} = \frac{\int_{-1}^{1} x^2\,\mathrm{d}x}{\int_{-1}^{1}\mathrm{d}x} = \frac{1}{3}.$$

Prin urmare,

$$P_2(x) = x^2 - \frac{1}{3}.$$

- Polinomul Legrendre de grad trei este, conform (10.5), $P_3(x) = (x - a_3)P_2(x) - b_3 P_1(x)$, iar

$$a_3 = \frac{\langle xP_2(x), P_2(x)\rangle}{\langle P_2(x), P_2(x)\rangle} = \frac{\int_{-1}^{1} x(x^2 - \frac{1}{3})^2\,\mathrm{d}x}{\int_{-1}^{1}(x^2 - \frac{1}{3})^2\,\mathrm{d}x} = 0, \quad b_3 = \frac{\langle P_2(x), P_2(x)\rangle}{\langle P_1(x), P_1(x)\rangle} = \frac{\int_{-1}^{1}(x^2 - \frac{1}{3})^2\,\mathrm{d}x}{\int_{-1}^{1} x^2\,\mathrm{d}x} = \frac{4}{15}.$$

Prin urmare,

$$P_3(x) = x^3 - \frac{3}{5}x.$$

Rădăcinile lui P_3 sunt nodurile din formula de integrare

$$c_0 = -\frac{3}{5}, \quad c_1 = 0, \quad c_2 = \frac{3}{5}. \tag{10.9}$$

Coeficienţii b_i se determină din faptul că formula (10.1) este exactă pentru 1, x şi x^2:

$$\sum_{i=0}^{2} b_i c_i^p = \frac{1}{p+1}\left(1 - (-1)^{p+1}\right), \quad 0 \le p \le 2.$$

Sistemul de determinare a coeficienţilor b_i este

$$\begin{cases} b_0 & + & b_2 & + & b_3 & = 2 \\ c_0 b_0 & + & c_1 b_1 & + & c_2 b_2 & = 0 \\ c_0^2 b_0 & + & c_1^2 b_1 & + & c_2^2 b_2 & = \frac{2}{3}. \end{cases} \tag{10.10}$$

şi prin urmare

$$b_0 = \frac{5}{9}, \quad b_1 = \frac{8}{9}, \quad b_2 = \frac{5}{9}. \tag{10.11}$$

Deducem următoarea formulă Gauss-Legendre de ordin 5 pe intervalul $(-1,1)$:

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx \frac{5}{9}f(-\sqrt{\frac{3}{5}}) + \frac{8}{9}f(0) + \frac{5}{9}f(\sqrt{\frac{3}{5}}). \tag{10.12}$$

Pentru un interval oarecare se aplică o schimbare de variabilă pentru a se obţine o integrală pe $(-1,1)$:

$$\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{a+b}{2} + \frac{b-a}{2}t\right)\,\mathrm{d}t.$$

Prin urmare formula Gauss-Legendre de ordin 5 este:

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{2}\left[\frac{5}{9}f\left(\frac{a+b}{2} - \frac{b-a}{2}\sqrt{\frac{3}{5}}\right) + \right. \\ \left. + \frac{8}{9}f\left(\frac{a+b}{2}\right) + \frac{5}{9}f\left(\frac{a+b}{2} + \frac{b-a}{2}\sqrt{\frac{3}{5}}\right)\right]. \tag{10.13}$$

Dăm în continuare şi formula Gauss-Legendre de ordin 7:

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{2}\Bigg[&\frac{1}{6}\left(-\sqrt{\frac{5}{6}}+3\right)f\left(\frac{a+b}{2} - \frac{b-a}{2}\sqrt{\frac{1}{7}\left(3+2\sqrt{\frac{6}{5}}\right)}\right) + \\ &+ \frac{1}{6}\left(\sqrt{\frac{5}{6}}+3\right)f\left(\frac{a+b}{2} - \frac{b-a}{2}\sqrt{\frac{1}{7}\left(3-2\sqrt{\frac{6}{5}}\right)}\right) + \\ &+ \frac{1}{6}\left(\sqrt{\frac{5}{6}}+3\right)f\left(\frac{a+b}{2} + \frac{b-a}{2}\sqrt{\frac{1}{7}\left(3-2\sqrt{\frac{6}{5}}\right)}\right) \\ &+ \frac{1}{6}\left(-\sqrt{\frac{5}{6}}+3\right)f\left(\frac{a+b}{2} + \frac{b-a}{2}\sqrt{\frac{1}{7}\left(3+2\sqrt{\frac{6}{5}}\right)}\right)\Bigg].\end{aligned} \tag{10.14}$$

Cursul 12

Metode numerice pentru ecuaţii diferenţiale ordinare

12.1 Teoria generală

Fie $t_0 \in \mathbb{R}$, $x^0 \in \mathbb{R}^d$, $a, b \in \mathbb{R}_+^*$ şi domeniul

$$D = D(t_0, x^0) = \left\{ (t, x) \in \mathbb{R} \times \mathbb{R}^d \;\middle|\; |t - t_0| \leq a, \quad ||x - x^0|| \leq b \right\}.$$

Pentru o funcţie $f : D \to \mathbb{R}^d$ considerăm ecuaţia diferenţială ordinară

$$\begin{cases} x'(t) = f(t, x(t)) \\ x(t_0) = x^0. \end{cases} \tag{12.1}$$

Definiţia 12.1. *Funcţia $x : [t_0 - \alpha, t_0 + \alpha] \to \mathbb{R}^d$ este soluţie a lui (12.1) pe $[t_0 - \alpha, t_0 + \alpha]$ dacă:*

1. *$|t - t_0| \leq a$ şi $||x(t) - x^0|| \leq b$ pentru orice $t \in [t_0 - \alpha, t_0 + \alpha]$*
2. *$x(t_0) = x^0$*
3. *x derivabilă pe $(t_0 - \alpha, t_0 + \alpha)$*
4. *$x'(t) = f(t, x(t))$ pentru orice $t \in (t_0 - \alpha, t_0 + \alpha)$.*

Teorema 12.1. Presupunem că sunt verificate următoarele condiţii

1. f este continuă pe D;
2. f este L-lipschitziană în a doua variabilă, adică

$$||f(t, x) - f(t, y)|| \leq L||x - y||, \quad \forall (t, x), (t, y) \in D;$$

Atunci ecuaţia diferencţială (12.1) are o soluţie unică $x(t)$ pe intervalul $[t_0 - \alpha, t_0 + \alpha]$, unde

$$\alpha = \inf \left\{ a, \frac{b}{M} \right\}, \quad M = \sup_{(t,x) \in D} ||f(t, x)||.$$

Observaţia 12.1. *Dacă f verifică doar condiţia 1), avem garantată existenţa dar nu şi unicitatea. De exemplu, dacă $f(t,x) = \sqrt[3]{x}$ avem două soluţii*

$$x(t) = 0$$

$$x(t) = \begin{cases} \frac{2}{3}t^{\frac{3}{2}} & t \geq 0 \\ 0 & t < 0. \end{cases}$$

12.2 Aproximarea soluţiilor ecuaţiilor diferenţiale ordinare

Fie $T > 0$ şi $x^0 \in \mathbb{R}^d$. Presupunem că (12.1) are o soluţie unică x pe intervalul $[t_0, t_0 + T]$.

Dat $n \in \mathbb{N}^*$, definim $h = T/n$ şi punctele echidistante $t_k = t_0 + kh$, $0 \leq k \leq n$.

De asemenea, notăm $x^k = x(t_k)$ şi ne propunem să aproximăm aceste valori

$$x^k \approx u^k, \quad 0 \leq k \leq n. \tag{12.2}$$

Definiţia 12.2. *Vom numi* **metodă de aproximare a soluţiei lui (12.1) pe intervalul** $[t_0, t_0 + T]$ *o aplicaţie $\mathcal{A}$ care, pentru fiecare $n \in \mathbb{N}^*$, ne dă $\mathcal{A}(n, x^0) = (u^0, u^1, ..., u^n) \in \mathcal{M}_{d\times(n+1)}(\mathbb{R})$.*

Dacă notăm eroarea (globală)

$$e^j = x^j - u^j, \quad 0 \leq j \leq n \tag{12.3}$$

avem că:

- Metoda $\mathcal{A}$ este **convergentă** dacă

$$\lim_{n\to\infty} \sup_{0\leq j\leq n} ||e^j|| = 0.$$

- Metoda $\mathcal{A}$ este **convergentă de ordin s** dacă

$$\sup_{0\leq j\leq n} ||e^j|| \leq Ch^s, \quad \forall n \geq 1.$$

12.3 Metoda Euler

Deducere:

Din $x'(t) = f(t, x(t))$ prin integrare între t_k şi t_{k+1} deducem că

$$x^{k+1} - x^k = \int_{t_k}^{t_{k+1}} f(t, x(t))dt \approx (t_{k+1} - t_k)f(t_k, x(t_k)) = hf(t_k, x^k).$$

Acest lucru ne conduce la a considera metoda de aproximare a lui Euler:

$$\begin{cases} u^0 = x^0 \\ u^{k+1} = u^k + hf(t_k, u^k), \quad k \geq 0. \end{cases} \tag{12.4}$$

Interpretare geometrică

u^1 se obţine intersectând tangenta la graficul soluţiei în $u^0 = x^0$ cu dreapta $t = t_1$,

$$y - x^0 = x'(t_0)(t - t_0) \cap t = t_1 \Leftrightarrow y = x^0 + f(t_0, x^0)(t - t_0) \cap t = t_1 \Leftrightarrow u^1 = u^0 + hf(t_0, u^0).$$

Pentru paşii următori, se foloseşte dreapta care trece prin (t_k, u^k) şi are panta $f(t_k, u^k)$, ca aproximare a tangentei la graficul soluţiei $x(t)$ în (t_k, x^k) care ar avea panta $x'(t_k) = f(t_k, x^k)$. Această dreapă se intersectează cu $t = t_{k+1}$:

$$y - u^k = f(t_k, x^k)(t - t_0) \cap t = t_{k+1} \Leftrightarrow u^{k+1} = u^k + hf(t_k, u^k).$$

Eroare locală

Pentru fiecare $j \geq 0$ definim eroarea locală a metodei Euler prin

$$\tau^{j+1} = x^{j+1} - \left(x^j + hf(t_j, x^j)\right), \quad 0 \leq j \leq n-1. \tag{12.5}$$

Să notăm că, dacă la pasul j am avea $u^j = x^j$ (acest lucru este întotdeauna adevărat pentru $j = 0$, de exemplu), atunci

$$\tau^{j+1} = x^{j+1} - u^{j+1}$$

adică τ^{j+1} este eroarea comisă la pasul $j + 1$ dacă la pasul anterior avem valoarea exactă. Este o eroare de un singur pas. De aici şi numele de eroare locală.

Observaţia 12.2. *Să observăm că*

$$\tau^{j+1} = (x^{j+1} - x^j) - hf(t_j, x^j) = \int_{t_j}^{t_{j+1}} f(t, x(t))dt - hf(t_j, x^j)$$

şi prin urmare τ^{j+1} este eroarea la aproximarea integralei la acest pas.

Propoziţia 12.3.1. *Avem că eroarea locală este de ordin doi*

$$\sup_{0 \leq j \leq n-1} ||\tau^{j+1}|| \leq \frac{M_2}{2} h^2 := Mh^2, \tag{12.6}$$

unde $M_2 = \sup_{(t,x) \in D} |(f_t + f_x f)(t, x)|$.

Demonstraţie. Folosind dezvoltarea Taylor, obţinem că

$$\tau^{j+1} = x^{j+1} - \left(x^j + hf(t_j, x^j)\right) = x(t_{j+1}) - x(t_j) - hx'(t_j) = \frac{h^2}{2} x''(\theta_j)$$

unde $t_j \leq \theta_j \leq t_{j+1}$. □

Teorema 12.2. Avem că

$$\sup_{0 \leq j \leq n} ||x^j - u^j|| \leq \frac{M_2}{2L} e^{LT} h. \tag{12.7}$$

(eroarea globală este de ordin unu).

Demonstraţie. Folosind (12.5), obţinem că

$$x^{j+1}-u^{j+1}=\tau^{j+1}+x^j+hf(t_j,x^j)-u^j-hf(t_j,u^j)=\tau^{j+1}+\left(x^j-u^j\right)+h\left(f(t_j,x^j)-f(t_j,u^j)\right).$$

Prin urmare

$$\|e^{j+1}\|\leq Mh^2+(1+hL)\|e^j\|\leq ...\leq Mh^2\left(1+(1+hL)+...+(1+hL)^j\right)+(1+hL)^{j+1}\|e^0\|=$$
$$=Mh^2\frac{(1+hL)^{j+1}-1}{hL}<\frac{M}{L}he^{(j+1)\ln(1+hL)}\leq\frac{M}{L}e^{LT}h.$$

□

12.4 Metoda Euler modificată

$$u^{k+1}=u^k+hf\left(t_k+\frac{h}{2},u^k+\frac{h}{2}f(t_k,u^k)\right) \tag{12.8}$$

Definim eroarea locală prin

$$\tau^{k+1}=x^{k+1}-\left[x^k+hf\left(t_k+\frac{h}{2},x^k+\frac{h}{2}f(t_k,x^k)\right)\right] \tag{12.9}$$

Propoziţia 12.4.1. *Eroarea locală în metoda Euler modificată este de ordin trei:*

$$\sup_{0\leq k\leq n-1}\|\tau^{k+1}\|=\mathcal{O}(h^3), \tag{12.10}$$

adică există o contantă $M>0$ astfel încât

$$\sup_{0\leq k\leq n-1}||\tau^{k+1}||\leq Mh^3. \tag{12.11}$$

Demonstraţie. Observăm că

$$\begin{aligned}\tau^{k+1}&=x^{k+1}-\left(x^k+hf\left(t_k+\frac{h}{2},x^k+\frac{h}{2}f(t_k,x^k)\right)\right)\\&=x(t_k+h)-x(t_k)-hf(t_k,x(t_k))-\frac{h^2}{2}\left(f_t(t_k,x(t_k))+(f_xf)(t_k,x(t_k))\right)\\&+h\left[f(t_k,x(t_k))+\frac{h}{2}\left(f_t(t_k,x(t_k))+(f_xf)(t_k,x(t_k))\right)-f\left(t_k+\frac{h}{2},x^k+\frac{h}{2}f(t_k,x^k)\right)\right]\end{aligned}$$

Cum, potrivit Teoremei lui Taylor, avem că

$$f\left(t_k+\frac{h}{2},x^k+\frac{h}{2}f(t_k,x^k)\right)=f(t_k,x^k)+\frac{h}{2}f_t(t_k,x^k)+\frac{h}{2}(f_xf)(t_k,x^k)+\mathcal{O}(h^2),$$

şi

$$\begin{aligned}x(t_k+h)&=x(t_k)+hx'(t_k)+\frac{h^2}{2}x''(t_k)+\mathcal{O}(h^3)\\&=x(t_k)+hf(t_k,x(t_k))+\frac{h^2}{2}\left(f_t(t_k,x(t_k))+(f_xf)(t_k,x(t_k))\right)+\mathcal{O}(h^3),\end{aligned}$$

obţinem că

$$\tau^{k+1}=x(t_k+h)-x(t_k)-hx'(t_k)-\frac{h^2}{2}x''(t_k)+\mathcal{O}(h^3)=\mathcal{O}(h^3).$$

□

12.5 Metoda Euler-Cauchy

$$u^{k+1} = u^k + \frac{h}{2}\left[f(t_k, u^k) + f\left(t_k + h, u^k + hf(t_k, u^k)\right)\right] \tag{12.12}$$

Propoziţia 12.5.1. *Eroarea locală în metoda Euler-Cauchy este de ordin trei:*

$$\sup_{0\le k\le n-1} \|\tau^{k+1}\| = \mathcal{O}(h^3), \tag{12.13}$$

adică există o contantă $M > 0$ astfel încât

$$\sup_{0\le k\le n-1} ||\tau^{k+1}|| \le Mh^3. \tag{12.14}$$

Observaţia 12.3. *Să observăm că în metoda Euler modificată*

$$\begin{aligned}\tau^{k+1} &= x^{k+1} - x^k - hf\left(t_k + \frac{h}{2}, x^k + \frac{h}{2}f(t_k, x^k)\right)\\ &= \int_{t_k}^{t_{k+1}} f(t, x(t))dt - hf\left(t_k + \frac{h}{2}, x^k + \frac{h}{2}f(t_k, x^k)\right),\end{aligned}$$

ceea ce se poate interpreta ca fiind eroarea la aproximarea integralei cu o formulă de tip punct mediu.

Analog, în metoda Euler-Cauchy am avea că

$$\begin{aligned}\tau^{k+1} &= x^{k+1} - x^k - \frac{h}{2}\left(f(t_k, x^k) + f\left(t_k + h, x^k + hf(t_k, x^k)\right)\right)\\ &= \int_{t_k}^{t_{k+1}} f(t, x(t))dt - \frac{h}{2}\left(f(t_k, x^k) + f\left(t_k + h, x^k + hf(t_k, x^k)\right)\right),\end{aligned}$$

ceea ce se poate interpreta ca fiind eroarea la aproximarea integralei cu o formulă de tip trapez.

12.6 Metode Runge-Kutta

Deducere

Din $x'(t) = f(t, x(t))$ prin integrare între t_k şi t_{k+1} deducem că:

$$x^{k+1} - x^k = \int_{t_k}^{t_{k+1}} f(t, x(t))\,\mathrm{dt} = \int_{t_k}^{t_k+h} f(t, x(t))\,\mathrm{dt} = h\int_0^1 f(t_k + hs, x(t_k + hs))\,\mathrm{ds}$$

$$\approx h\sum_{j=1}^{\nu} b_j f(t_k + hc_j, x(t_k + hc_j))$$

unde $c_j \in [0,1]$ şi $b_j \in \mathbb{R}$ cu $1 \le j \le \nu$ (integrala se aproximează printr-o combinayie de valori ale lui f). Cum valorile $f\left(t_k + hc_j, x(t_k + hc_j)\right)$ nu sunt disponibile, le vom aproxima

$$x(t_k + hc_j) \approx \xi^j, \quad f(t_k + hc_j, x(t_k + hc_j) \approx f\left(t_k + hc_j, \xi^j\right) := K^j,$$

astfel:

$$\text{(12.15)} \qquad \begin{cases} c_1 = 0, \quad \xi^1 = u^k, \quad K^1 = f(t_k, u^k); \\ \\ c_j \text{ dat }, \quad \xi^j = u^k + h\sum_{l=1}^{j-1} a_{jl}K^l, \quad K^j = f\left(t_k + hc_j, \xi^j\right). \end{cases}$$

Acest lucru ne conduce la a considera metoda de aproximare Runge-Kutta:

$$\text{(12.16)} \qquad \begin{cases} u^0 = x^0 \\ u^{k+1} = u^k + h\sum_{j=1}^{\nu} b_j K^j \quad k \ge 0. \end{cases}$$

Vom numi:

$$\begin{cases} \nu \text{ ordin Runge-Kutta} \\ \\ (c_j)_{1\le j\le \nu} \text{ noduri Runge-Kutta} \\ \\ (b_j)_{1\le j\le \nu} \text{ ponderi Runge-Kutta} \\ \\ (a_{jl})_{1\le j\le \nu,\, 1\le l\le j-1} \text{ matrice Runge-Kutta.} \end{cases}$$

Există mai multe moduri de a alege aceşti parametri. Vom proceda astfel:

- Se alege ν. Cu cat ν este mai mare cu atat ordinul de convergenţă este mai mare (de fapt ν va fi ordinul de convergenţă a metodei).
- Se dezvoltă $x(t_k + h)$ în serie Taylor până la ordinul ν:

$$x(t_k + h) = x(t_k) + hx'(t_k) + \frac{h^2}{2}x''(t_k) + \mathcal{O}(h^3)$$

$$\approx u^k + hf(t_k, u^k) + \frac{h^2}{2}\left(f_t(t_k, u^k) + f_x(t_k, u^k)f(t_k, u^k)\right) + \mathcal{O}(h^3).$$

- Se dezvoltă $h\sum_{j=1}^{2} b_j f\left(t_k + hc_j, x(t_k + hc_j)\right)$ în serie Taylor până la ordinul ν:

$$h\sum_{j=1}^{\nu} b_j f\left(t_k + hc_j, x(t_k + hc_j)\right) = h\left[b_1 f(t_k + hc_1, x(t_k + hc_1)) + b_2 f\left(t_k + hc_2, x(t_k + hc_2)\right)\right]$$

$$\approx hb_1K^1 + hb_2K^2 = hb_1 f(t_k, u^k) + hb_2 f(t_k + hc_2, u^k + ha_{21} f(t_k, u^k))$$
$$= hb_1 f(t_k, u^k) + hb_2\Big(f(t_k, u^k) + h(c_2 f_t(t_k, u^k) + a_{21} f_x(t_k, u^k) f(t_k, u^k))\Big) + \mathcal{O}(h^3).$$

- Identificând coeficienţii se obţin valorile parametrilor:

$$\begin{cases} b_1 + b_2 = 1 \\ b_2 c_2 = \frac{1}{2} \\ b_2 a_{21} = \frac{1}{2}. \end{cases}$$

  Posibile alegeri:

  – $b_1 = 0$, $b_2 = 1$, $c_2 = \frac{1}{2}$, $a_{21} = \frac{1}{2}$ (metoda Euler modificată)

$$u^{k+1} = u^k + \frac{h}{2} f\left(t_k + \frac{h}{2}, u^k + \frac{h}{2} f(t_k, u^k)\right). \tag{12.17}$$

  – $b_0 = b_1 = \frac{1}{2}$, $c_2 = 1$, $a_{21} = 1$ (metoda Euler-Cauchy)

$$u^{k+1} = u^k + \frac{h}{2}\left(f(t_k, u^k) + f\left(t_k + h, u^k + hf(t_k, u^k)\right)\right). \tag{12.18}$$

Cele două scheme de mai sus sunt de ordin 2. Cea mai folosită este însă schema de ordin 4:

$$\begin{cases} u^{k+1} = u^k + \frac{h}{6}(K^1 + 2K^2 + 2K^3 + K^4) \\ K^1 = f\left(t_k, u^k\right) \\ K^2 = f\left(t_k + \frac{h}{2}, u^k + \frac{h}{2}K^1\right) \\ K^3 = f\left(t_k + \frac{h}{2}, u^k + \frac{h}{2}K^2\right) \\ K^4 = f\left(t_k + h, u^k + hK^3\right). \end{cases} \tag{12.19}$$