Numerične metode 1 - definicije, trditve in izreki

Oskar Vavtar

2020/21

Kazalo

1 NUMERIČNO RAČUNANJE

1.1 Uvod

Definicija 1.1 (Napaka)**.** Pri numeričnem računanju izračunamo numerični približek za točno rešitev. Razlika med približkom in točno vrednostjo je *napaka* približka. Ločimo *absolutno* in *relativno* napako.

- absolutna napaka = približek − točna vrednost
- relativna napaka = $\frac{\text{absolutna napaka}}{\text{točna vrednost}}$

Naj bo x točna vrednost, $\hat{x}$ pa približek za x.

- Če je $\hat{x} = x + d_a$, potem je $d_a = \hat{x} - x$ *absolutna napaka.*
- Če je $\hat{x} = x(1 + d_r)$, potem je $d_r = \frac{\hat{x}-x}{x}$ *relativna napaka.*

1.2 Premična pika

Definicija 1.2. Velja $\mathrm{fl}(x) = x(1 + \delta)$ za $|\delta| \leq u$, kjer je

$$u = \frac{1}{2}b^{1-t}$$

osnovna zaokrožitvena napaka:

- single: $u = 2^{-24} = 6 \times 10^{-8}$
- double: $u = 2^{-53} = 1 \times 10^{-16}$

Izrek 1. Če za število x velja, da $|x|$ leži na intervalu med najmanjšim in največjim *pozitivnim predstavljivim normaliziranim* številom, potem velja

$$\frac{|\mathrm{fl}(x) - x|}{|x|} \leq u.$$

1.3 Občutljivost problema

Definicija 1.3 (Občutljivost)**.** Če se rezultat pri majhni spremembi argumentov (*motnji* oz. *perturbaciji*) ne spremeni veliko, je problem *neobčutljiv*, sicer pa je *občutljiv.*

1.4 Vrste napak pri numeričnem računanju

Definicija 1.4. Pri numeričnem računanje se pojavijo 3 vrste napak:

1. Neodstranljive napake: Npr. ko podatek ni predstavljivo število. Namesto $y = f(x)$ lahko v najboljšem primeru izračunamo $\overline{y} = f(\overline{x})$, kjer je $\overline{x}$ najbližje predstavljivo število.

$$D_n = y - \overline{y} = f(x) - f(\overline{x})$$

2. Napaka metode: Npr. ko na voljo nimamo željene operacije. Namesto $f(\overline{x})$ potem izračunamo $\widetilde{y} = g(\overline{x})$, kjer je $g(x)$ približek za $f(x)$, kjer znamo vrednost g izračunati s končnim številom operacij.

$$D_m = \overline{y} - \widetilde{y} = f(\overline{x}) - g(\overline{x})$$

3. Zaokrožitvene napake: Pri izračunu $\widetilde{y} = g(\overline{x})$ lahko pri vsaki osnovni operaciji pride do zaokrožitvene napake, zato na koncu kot numeričen rezultat dobimo $\widehat{y}$.

$$D_z = \widetilde{y} - \widehat{y}$$

Skupna napaka:

$$D = D_n + D_m + D_z$$

V splošnem lahko ocenimo:

$$|D| \leq |D_n| + |D_m| + |D_z|$$

1.5 Stabilnost metode

Definicija 1.5. Če metoda za $\forall x$ vrne $\widehat{y}$, ki je *absolutno (relativno)* blizu točnemu y, je metoda *direktno stabilna.*

Če metoda za $\forall x$ vrne tak $\widehat{y}$, da $\exists \widehat{x}$ absolutno (relativno) blizu x, da je $\widehat{y} = f(\widehat{x})$ (točno), je metoda *obratno stabilna.*

V splošnem:

$$|\text{direktna napaka}| \leq |\text{občutljivost}| \cdot |\text{obratna napaka}|$$

1.6 Analiza zaokrožitvenih napak

1.6.1 Produkt $n+1$ predstavljivih števil

Algoritem. Dana so predstavljiva števila $x_0, x_1, \ldots, x_n$;
računamo $p = x_0 \cdot x_1 \cdot \ldots \cdot x_n$.

TOČNO:

$\mathtt{p_0 = x_0}$
$\mathtt{i = 1,\ldots,n}$
$\quad \mathtt{p_i = p_{i-1} \cdot x_i}$
$\mathtt{p = p_n}$

NUMERIČNO:

$\widehat{\mathtt{p}}_0 = \mathtt{x_0}$
$\mathtt{i = 1,\ldots,n}$
$\quad \widehat{\mathtt{p}}_\mathtt{i} = \widehat{\mathtt{p}}_\mathtt{i-1} \cdot \mathtt{x_i} \cdot (1 + \delta_\mathtt{i}) \quad |\delta_\mathtt{i}| \leq \mathtt{u}$
$\widehat{\mathtt{p}} = \widehat{\mathtt{p}}_\mathtt{n}$

1.6.2 Skalarni produkt vektorjev dolžine n

Algoritem. Imamo vektorje *predstavljivih* števil $a = [a_1, \ldots, a_n]^T$, $b = [b_1, \ldots, b_n]^T$. Računamo $s = \langle b^T, a \rangle = \sum_{i=1}^{n} a_i b_i$.

TOČNO:

$$
\begin{aligned}
&\mathtt{s_0 = 0}\\
&\mathtt{i = 1,\ldots,n}\\
&\qquad \mathtt{p_i = a_i \cdot b_i}\\
&\qquad \mathtt{s_i = s_{i-1} + p_i}\\
&\mathtt{s = s_n}
\end{aligned}
$$

NUMERIČNO:

$$
\begin{aligned}
&\widehat{\mathtt{s}}_0 = 0\\
&\mathtt{i = 1,\ldots,n}\\
&\qquad \widehat{\mathtt{p}}_i = \mathtt{a_i \cdot b_i \cdot (1 + } \alpha_i\mathtt{)} && |\alpha_i| \leq u\\
&\qquad \widehat{\mathtt{s}}_i = (\widehat{\mathtt{s}}_{i-1} + \widehat{\mathtt{p}}_i) \cdot (1 + \beta_i) && |\beta_i| \leq u\\
&\widehat{\mathtt{s}} = \widehat{\mathtt{s}}_n
\end{aligned}
$$

Trditev 1. Računanje skalarnega produkta je *obratno* stabilno, ni pa *direktno* stabilno.

2 NELINEARNE ENAČBE

2.1 Uvod

Definicija 2.1. Naj bo α ničla funkcije f, ki je *zvezno odvedljiva* v okolici α:

- $f'(\alpha) \neq 0$: α je *enostavna* ničla
- $f'(\alpha) = 0$: α je *večkratna* ničla

  Če je f m-krat zvezno odveljiva in
  $$f'(\alpha) = f''(\alpha) = \ldots = f^{(m-1)}(\alpha) = 0, \quad f^{(m)} \neq 0,$$
  je α m-kratna ničla.

Trditev 2 (Občutljivost ničel)**.** Naj bo α enostavna ničla. Če v okolici $x = \alpha$ obstaja inverzna funkcija $\alpha = f^{-1}(0)$ v bistvu "računamo"vrednost inverzne funkcije. Občutljivost je enaka absolutni vrednosti odvoda inverzna funkcije:
$$|(f^{-1})'(0)| = \frac{1}{|f'(f^{-1}(0))|} = \frac{1}{|f'(\alpha)|}.$$

Večkratno ničlo lahko izračunamo le z natančnostjo $u^{\frac{1}{m}}$, kjer je m večkratnost ničle (za dvojno ničlo dobimo le polovico točnih decimalk, za trojno le tretjino...).

2.2 Bisekcija

Izrek 2. Če je f realna zvezna funkcija na $[a, b]$ in je $f(a) \cdot f(b) < 0$, potem $\exists \xi \in (a, b)$, da je $f(\xi) = 0$.

Algoritem (Bisekcija)**.** Naj velja $f(a) \cdot f(b) < 0$ in $a < b$:

```
e = b - a
while e > ε
    e = e/2, c = a + e
    if sign(f(c)) = sign(f(a))
        a = c
    else
        b = c
```

2.3 Navadna iteracija

Algoritem (Navadna iteracija)**.**

```
izberi x_0
r = 0, 1, 2,...
    x_{r+1} = g(x_r)
```

Ustavitveni kriterij:

a) $r > r_{max}$ (prekoračeno število korakov)

b) $|r_{x+1} - r_x| < \varepsilon$

Izrek 3. Naj bo $\alpha = g(\alpha)$ in naj iteracijska funkcija g na intervalu $I = [\alpha - \delta, \alpha + \delta]$ za nek $\delta > 0$ zadošča Lipschitzovem pogoju

$$|g(x) - g(y)| \leq m|x - y| \quad \text{za} \quad x, y \in I, 0 \leq m < 1.$$

Potem za $\forall x_0 \in I$ zaporedje $x_{r+1} = g(x_r)$, $r = 0, 1, \ldots$ konvergira k α in velja

- $|x_r - \alpha| \ \leq \ m^r \cdot |x_0 - \alpha|$
- $|x_{r+1} - \alpha| \ \leq \ \dfrac{m}{1 - m} \cdot |x_{r+1} - x_r|$

Posledica. Naj bo $\alpha = g(\alpha)$, g *zvezno odvedljiva* in $|g'(\alpha)| < 1$. Potem $\exists \delta > 0$, da za $\forall x_0 \in [\alpha - \delta, \alpha + \delta]$ zaporedje $x_{r+1} = g(x_r)$ konvergira k α.

Definicija 2.2. Naj zaporedje $\{x_r\}$ konvergira proti α ($\lim_{r\to\infty} x_r = \alpha$). Pravimo, da zaporedje konvergira z redom konvergence p, če obstaja limita

$$\lim_{r\to\infty} \frac{|x_{r+1} - \alpha|}{|x_r - \alpha|^p} = C > 0.$$

Izrek 4. Naj bo iteracijska funkcija g p-krat *zvezno odvedljiva* v okolici negibne točke α. Če velja $g'(\alpha) = \ldots = g^{(p-1)}(\alpha) = 0$ in $g^{(p)}(\alpha) \neq 0$, potem zaporedje $x_{r+1} = g(x_r)$, $r = 0, 1, \ldots$, v okolici α konvergira z redom p. V primeru $p = 1$ mora za konvergenco veljati še $|g'(\alpha)| < 1$.

2.4 Tangentna metoda

Metoda.

$$x_{r+1} = x_r - \frac{f(x_r)}{f'(x_r)}$$

Konvergenca:

$$|e_{r+1}| \approx C \cdot |e_r|^2$$

Posledica. Če je f *dvakrat zvezno odvedljiva* v okolici ničle α, potem tangentna metoda za dovolj dober začetni približek x_0 vedno konvergira k α.

Izrek 5. Naj bo funkcija f na $I = [a, \infty)$ *dvakrat zvezno odvedljiva, naraščajoča in ima ničlo na* $\alpha \in I$. Potem je α edina ničla na I in za $\forall x_0 \in I$ tangentna metoda konvergira le k α.

2.5 Metode brez f'

Metoda (Sekantna metoda)**.**

$$x_{r+1} = x_r - \frac{f(x_r)(x_r - x_{r-1})}{f(x_r) - f(x_{r-1})}$$

Konvergenca:

$$|e_{r+1}| \approx C \cdot |e_r| \cdot |e_{r-1}|$$

Metoda (Mullerjeva metoda)**.** Skozi točke $(x_r, f(x_r))$, $(x_{r-1}, f(x_{r-1}))$, $(x_{r-2}, f(x_{r-2}))$ potegnemo kvadratni polinom $y = p(x)$ in za x_{r+1} vzamemo tisto ničlo polinoma p, ki je bližje x_r.

Konvergenca:

$$|e_{r+1}| \approx C \cdot |e_r| \cdot |e_{r-1}| \cdot |e_{r-2}|$$

Metoda (Inverzna interpolacija)**.** Zamenjamo vlogi x in y in vzamemo kvadratni polinom $x = \mathcal{L}(y)$, ki gre skozi točke $(x_r, f(x_r))$, $(x_{r-1}, f(x_{r-1}))$, $(x_{r-2}, f(x_{r-2}))$. Za x_{r+1} vzamemo

$$x_{r+1} = \mathcal{L}(0).$$

Red konvergence je enak kot pri *Mullerjevi metodi.*

3 SISTEMI LINEARNIH ENAČB

3.1 Oznake in definicije

Definicija 3.1. Sistem n linearnih enačb z n neznankami pišemo v obliki

$$Ax = b, \quad A \in \mathbb{R}^{n\times n}\ (\mathbb{C}^{n\times n}),\ x, b \in \mathbb{R}^n\ (\mathbb{C}^n).$$

Definicija 3.2. Skalarni produkt vektorjev x in y je enaka

a) $x, y \in \mathbb{R}^n$:

$$y^T x \; = \; \sum_{i=1}^{n} x_i y_i \; = \; \langle x, y\rangle \; = \; \langle y, x\rangle$$

b) $x, y \in \mathbb{C}^n$:

$$y^H x \; = \; \sum_{i=1}^{n} x_i \overline{y_i} \; = \; \langle x, y, \rangle \; = \; \overline{\langle y, x\rangle}$$

Definicija 3.3. Množenje vektorja x z matriko A:

a)

$$y_i \; = \; \sum_{k=1}^{n} a_{ik} x_k \; = \; \alpha_i^T x$$

b)

$$y \; = \; \sum_{i=1}^{n} x_i a_i$$

Definicija 3.4. $A \in \mathbb{R}^{n\times n}$ je *nesingularna*, če (ekvivalentno):

a) $\det(A) \neq 0$

b) obstaja inverz A^{-1}, da je $A^{-1}A = AA^{-1} = I$

c) $rang(A) = n$

d) za $\forall x \neq 0$ je $Ax \neq 0$

e) $\ker(A) = \{x \mid Ax = 0\} = \{0\}$

Definicija 3.5. Matrika je *simetrično pozitivno definitna*, če $A = A^T$ in $x^T Ax > 0$ za $x \neq 0$.

Definicija 3.6. Če za $x \neq 0$ velja $Ax = \lambda x$, je λ lastna vrednost in x lastni vektor. Vsaka matrika ima n lastnih vrednosti, ki so ničle karakterističnega polinoma

$$p(\lambda) := \det(A - \lambda I).$$

3.2 Vektorske in matrične norme

Definicija 3.7. Vektorska norma je preslikava $\|.\| : \mathbb{C}^n \to \mathbb{R}$, da velja:

1) Nenegativnost:
$$\|x\| \geq 0, \quad \|x\| = 0 \Leftrightarrow x = 0$$

2) Homogenost:
$$\|\alpha x\| = |\alpha| \cdot \|x\|$$

3) Trikotniška neenakost:
$$\|x + y\| \leq \|x\| + \|y\|$$

Definicija 3.8. Matrična norma je preslikava $\|.\| : \mathbb{C}^{n\times n} \to \mathbb{R}$, da velja:

1)
$$\|A\| \geq 0, \quad \|A\| = 0 \Leftrightarrow A = 0$$

2)
$$\|\alpha A\| = |\alpha| \cdot \|A\|$$

3)
$$\|A + B\| \leq \|A\| + \|B\|$$

4) Submultiplikativnost:

$$\|A \cdot B\| \leq \|A\| \cdot \|B\|$$

za $\forall A, B \in \mathbb{C}^{n \times n}$ in $\forall \alpha \in \mathbb{C}$.

Lema 1. Če je za poljubno vektorsko normo $\|.\|_v$ definiramo

$$\|A\| := \max_{x \neq 0} \frac{\|Ax\|_v}{\|x\|_v},$$

je to matrična norma.

Lema 2.

$$\|A\|_1 \;=\; \max_{j=1,\ldots,n} \|a_j\|_1 \;=\; \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|$$

Lema 3.

$$\|A\|_2 \;=\; \sigma_1(A) = \max_{i=1,\ldots,n} \sqrt{\lambda_i(A^H A)}$$