

# Numerične metode 1 - definicije, trditve in izreki

Oskar Vavtar

2020/21

## Kazalo

<b>1</b>	<b>NUMERIČNO RAČUNANJE</b>	<b>3</b>
1.1	Uvod . . . . .	3
1.2	Premična pika . . . . .	3
1.3	Občutljivost problema . . . . .	4
1.4	Vrste napak pri numeričnem računanju . . . . .	4
1.5	Stabilnost metode . . . . .	5
1.6	Analiza zaokrožitvenih napak . . . . .	5
1.6.1	Produkt $n + 1$ predstavljivih števil . . . . .	5
1.6.2	Skalarni produkt vektorjev dolžine $n$ . . . . .	6
<b>2</b>	<b>NELINEARNE ENAČBE</b>	<b>7</b>
2.1	Uvod . . . . .	7
2.2	Bisekcija . . . . .	7
2.3	Navadna iteracija . . . . .	8
2.4	Tangentna metoda . . . . .	9
2.5	Metode brez $f'$ . . . . .	10

# 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  $\text{fl}(x) = x(1 + \delta)$  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

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

### 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 računanju se pojavijo 3 vrste napak:

1. NEODSTRANLJIVE NAPAKE: Npr. ko podatek ni predstavljlivo število. Namesto  $y = f(x)$  lahko v najboljšem primeru izračunamo  $\bar{y} = f(\bar{x})$ , kjer je  $\bar{x}$  najbližje predstavljlivo število.

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

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

$$D_m = \bar{y} - \tilde{y} = f(\bar{x}) - g(\bar{x})$$

3. ZAOKROŽITVENE NAPAKE: Pri izračunu  $\tilde{y} = g(\bar{x})$  lahko pri vsaki osnovni operaciji pride do zaokrožitvene napake, zato na koncu kot numeričen rezultat dobimo  $\hat{y}$ .

$$D_z = \tilde{y} - \hat{y}$$

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  $\hat{y}$ , ki je *absolutno (relativno)* blizu točnemu  $y$ , je metoda *direktno stabilna*.

Če metoda za  $\forall x$  vrne tak  $\hat{y}$ , da  $\exists \hat{x}$  absolutno (relativno) blizu  $x$ , da je  $\hat{y} = f(\hat{x})$  (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  $x_0, x_1, \dots, x_n$ ; računamo  $p = x_0 \cdot x_1 \cdot \dots \cdot x_n$ .

TOČNO:

```
p0 = x0
i = 1, ..., n
    pi = pi-1 · xi
p = pn
```

NUMERIČNO:

```
 $\hat{p}_0 = x_0$ 
 $i = 1, \dots, n$ 
     $\hat{p}_i = \hat{p}_{i-1} \cdot x_i \cdot (1 + \delta_i) \quad |\delta_i| \leq u$ 
 $\hat{p} = \hat{p}_n$ 
```

### 1.6.2 Skalarni produkt vektorjev dolžine $n$

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

TOČNO:

```
s0 = 0
i = 1, ..., n
    pi = ai · bi
    si = si-1 + pi
s = sn
```

NUMERIČNO:

```
ŝ0 = 0
i = 1, ..., n
    p̂i = ai · bi · (1 + αi)    |αi| ≤ u
    ŝi = (ŝi-1 + p̂i) · (1 + βi)  |βi| ≤ u
ŝ = ŝn
```

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

## 2 NELINEARNE ENAČBE

### 2.1 Uvod

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

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

Če je  $f$   $m$ -krat zvezno odvedljiva in

$$f'(\alpha) = f''(\alpha) = \dots = f^{(m-1)}(\alpha) = 0, \quad f^{(m)} \neq 0,$$

je  $\alpha$   $m$ -kratna ničla.

**Trditev 2** (Občutljivost ničel). Naj bo  $\alpha$  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 inverzne 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 decimal, 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

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

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

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



**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  $\alpha$ .

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

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

**Izrek 4.** Naj bo iteracijska funkcija  $g$   $p$ -krat zvezno odvedljiva v okolici negibne točke  $\alpha$ . Če velja  $g'(\alpha) = \dots = g^{(p-1)}(\alpha) = 0$  in  $g^{(p)}(\alpha) \neq 0$ , potem zaporedje  $x_{r+1} = g(x_r)$ ,  $r = 0, 1, \dots$ , v okolici  $\alpha$  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  $\alpha$ , potem tangentna metoda za dovolj dober začetni približek  $x_0$  vedno konvergira k  $\alpha$ .

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

## 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*.