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:

$\mathtt{p}_0 = \mathtt{x}_0$
$\mathtt{i} = 1,\ldots,\mathtt{n}$
$\quad \mathtt{p}_\mathtt{i} = \mathtt{p}_{\mathtt{i-1}} \cdot \mathtt{x}_\mathtt{i}$
$\mathtt{p} = \mathtt{p}_\mathtt{n}$

NUMERIČNO:

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

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:

$$
\begin{aligned}
&\mathtt{s}_0 = 0\\
&\mathtt{i} = 1, \ldots, \mathtt{n}\\
&\quad \mathtt{p}_\mathtt{i} = \mathtt{a}_\mathtt{i} \cdot \mathtt{b}_\mathtt{i}\\
&\quad \mathtt{s}_\mathtt{i} = \mathtt{s}_{\mathtt{i-1}} + \mathtt{p}_\mathtt{i}\\
&\mathtt{s} = \mathtt{s}_\mathtt{n}
\end{aligned}
$$

NUMERIČNO:

$$
\begin{aligned}
&\widehat{\mathtt{s}}_0 = 0\\
&\mathtt{i} = 1, \ldots, \mathtt{n}\\
&\quad \widehat{\mathtt{p}}_\mathtt{i} = \mathtt{a}_\mathtt{i} \cdot \mathtt{b}_\mathtt{i} \cdot (1 + \alpha_i) && |\alpha_i| \leq u\\
&\quad \widehat{\mathtt{s}}_\mathtt{i} = (\widehat{\mathtt{s}}_{\mathtt{i-1}} + \widehat{\mathtt{p}}_\mathtt{i}) \cdot (1 + \beta_i) && |\beta_i| \leq u\\
&\widehat{\mathtt{s}} = \widehat{\mathtt{s}}_\mathtt{n}
\end{aligned}
$$

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...).