Premična pika

$$x = (-1)^S m \cdot b^e$$
$$m = c_0 + c_1 b^{-1} + c_2 b^{-2} + \cdots + c_t b^{-t}$$

S	...	predznak
b	...	baza, ponavadi 2
m	...	vrednost mantise
t	...	dolžina mantise
e	...	vrednost eksponenta $L \leq e \leq U$
c_i	...	števke v mejah $0 \leq c_i \leq b-1$

Sistem premične pike označimo z $P(b, t, L, U)$.

Standard IEEE

Eksponent je zapisan z odmikom:

$$E = e + \text{odmik}$$

Če je $E = 0$, uporabimo **denormiran zapis**:

$$x = (-1)^S (c_1 b^{-1} + c_2 b^{-2} + \cdots + c_t b^{-t}) \cdot b^{e+1}$$

Sicer pa **normiran zapis**:

$$x = (-1)^S (1 + c_1 b^{-1} + c_2 b^{-2} + \cdots + c_t b^{-t}) \cdot b^e$$

Če so vsi biti eksponenta 1 in vsi biti mantise 0, je $x = (-1)^S \infty$.

Če so vis biti eksponenta 1 in vsi biti mantise niso 0, je $x = \text{NaN}$.

- **Single precision** $b = 2$, $t = 23$, $L = -126$, $U = 127$, odmik: 127

predznak 1	*exponent* 8	*mantisa* 23

- **Double precision** $b = 2$, $t = 52$, $L = -1022$, $U = 1023$, odmik: 1023

predznak 1	*exponent* 11	*mantisa* 52

Zaokroževanje

Naj bo x pozitivno število z neskončnim zapisom

$$x = (c_1 b^{-1} + c_2 b^{-2} + \cdots + c_t b^{-t} + c_{t+1} b^{-t-1}) + \ldots b^e$$

Kandidata za približek $\text{fl}(x)$ sta:

$$x_- = (c_1 b^{-1} + c_2 b^{-2} + \cdots + c_t b^{-t}) b^e$$
$$x_+ = (c_1 b^{-1} + c_2 b^{-2} + \cdots + c_t b^{-t} + b^{-t}) b^e$$

Vzamemo tistega, ki je bljižje. Če sta enako blizu, izberemo tistega, ki ima zadnjo števko sodo.

Osnovna zaokrožitvena napaka

$$u = \frac{1}{2} b^{-t}$$
$$\text{fl}(x) = x(1 + \delta) \quad \text{za} \quad |\delta| \leq u$$
$$\frac{|\text{fl}(x) - x|}{|x|} \leq u$$

Napake pri numeričnem računanju

- **Neodstranljiva napaka** Namesto x imamo približek $\bar{x}$.

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

- **Napaka metode** Namesto funkcije f imamo približek g.

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

- **Zaokrožitvena napaka** Pri računanju $\tilde{y} = f(\bar{x})$ se pri vsaki operaciji se pojavi zaokrožitvena napaka. Namesto $\tilde{y}$ dobimo $\hat{y}$.

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

Celotna napaka je $D = |D_n| + |D_m| + |D_z|$.

Stopnja občutljivosti

Razmerje velikosti spremembe podatkov in spremembe rezultata.

Naj bo $f : \mathbb{R} \to \mathbb{R}$ zvezno odvedljiva funkcija in δx majhna motnja.

- **Absolutna občutljivost** f v točki x:

$$|f(x + \delta x) - f(x)| \approx |f'(x)| \cdot |\delta x|$$

- **Relativna občutljivost** f v točki x:

$$\frac{|f(x + \delta x) - f(x)|}{|f(x)|} \approx \frac{|f'(x)| \cdot |\delta x|}{|f(x)|}$$

Obratna in direktna stabilnost

- **Direktna stabilnost**: za vsak x direktna napaka ($|f(x) - f(x + \Delta x)|$) majhna (absolutno oz. relativno).
- **Obratna stabilnost**: za vsak x razlika Δx, ki bi nam dala pravi rezultat majhna.

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

Nelinearne enačbe

Iščemo ničle funkcije f.

- **Enostavne ničle**:

$$f(\alpha) = 0 \quad \text{in} \quad f'(\alpha) \neq 0$$

- **m-kratne ničle**:

$$f(\alpha) = f'(\alpha) = \cdots = f^{(m-1)}(\alpha) = 0$$

Občutljivost ničle

Naj bo α m-kratna ničla $\hat{\alpha}$ približek, da je $f(\hat{\alpha}) = \varepsilon$.

Če f razvijemo v Taylorjevo vrsto okoli α in vzamemo prvih $m + 1$ členov dobimo:

$$\varepsilon \doteq \frac{f^{(m)}(\alpha)}{m!} (\hat{\alpha} - \alpha)^m \qquad |\hat{\alpha} - \alpha| \doteq \sqrt[m]{\frac{\varepsilon \cdot m!}{|f^{(m)}(\alpha)|}}$$

Bisekcija

vhod: funkcija $f : [a, b] \to \mathbb{R}$, $f(a)f(b) < 0$, natancnost ε
izhod: nicla funkcije f

dokler $|b - a| > \varepsilon$:
  $c \leftarrow \frac{a+b}{2}$
  ce $\text{sign}(f(c)) = \text{sign}(f(a))$:
    $a \leftarrow c$
  sicer :
    $b \leftarrow c$
vrni c

c je približek ničle α. Velja

$$|\alpha - c| \leq \frac{b - a}{2^m} \leq \varepsilon$$

Za natančnost ε potrebujemo $\log_2 \left(\frac{b-a}{\varepsilon} \right)$ korakov.

Navadna iteracija

Rešujemo $f(x) = 0$. Enačbo pretvorimo v $x = g(x)$. Načinov je veliko:

- $g(x) = f(x) + x$
- $g(x) = cf(x) + x$
- $g(x) = h(x)f(x) + x$ kjer je $h(x)$ funkcija, ki nima ničle v α.

Izrek o konvergenci navadne iteracije

Naj bo α negibna točka za g in naj g na intervalu $[\alpha - d, \alpha + d]$ $(d > 0)$ zadošča Lipschitzovemu pogoju:

$$\exists m \in [0, 1) \ \forall x, y \in I \ : |g(x) - g(y)| \leq m|x - y|$$

tedaj je g *skrčitev* na I.

Potem za vsak $x_0 \in I$ zaporedje $x_{r+1} = g(x_r)$ konvergira k α in velja:

$$|x_r - \alpha| \leq \frac{m}{1 - m} |x_r - x_{r-1}|$$

Posledica: Če je g zvezno odvedljiva v α in velja $g'(\alpha) < 1$, obstaja interval I, ki vsebuje α, da za vsak $x_0 \in I$ zaporedje konvergira k α.

- Če je $|g'(\alpha)| < 1$, je α *privlačna negibna točka*
- Če je $|g'(\alpha)| > 1$, je α *odbojna negibna točka*

Če je α odbojna za g, je privlačna za g^{-1}:
$g(x) = x \implies x = g^{-1}(x)$

Hitrost konvergence

$p > 0$ je red konvergence, če $\exists C_1, C_2 > 0$, da za vse dovolj pozne člene zaporedja $x_{r+1} = g(x_r)$ velja:

$$C_1 |x_r - \alpha|^p \leq |x_{r+1} - \alpha| \leq C_2 |x_r - \alpha|^p$$

Vsak korak se št. decimalk pomnoži s p.

Naj bo g v okolici α p-krat zvezno odvedljiva in naj velja $g(\alpha) = \alpha$, $g'(\alpha) = \cdots = g^{(p-1)}(\alpha) = 0$ in $g^{(p)}(\alpha) \neq 0$. Tedaj je red konvergence enak p.

Standardni redi konvergence

$p = 1$	...	linearna konvergenca
$p = 2$	...	kvadratična konvergenca

Tangentna (Newtonova) metoda

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

Red konvergence:

- 2, če je α enkratna ničla $f'(\alpha) \neq 0$
- 3, če je $f'(\alpha) \neq 0$ in $f''(\alpha) = 0$
- 1, če je α večkratna ničla $f'(\alpha) = 0$

Sekantna metoda

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

Red konvergence: $\frac{\sqrt{5}+1}{2} \approx 1.62$

Metoda (f, f', f'')

$$x_{r+1} = x_r - \frac{f(x_r)}{f'(x_r)} - \frac{f''(x_r)f^2(x_r)}{2f'^3(x_r)}$$

Red konvergence: 3 *(pri predpostavkah)*

Müllerjeva metoda

Na x_r, x_{r-1}, x_{r-2} napnemo parabolo, ničla parabole je naslednji približek.

$$p(x) = a(x - x_r)^2 + b(x - x_r) + c$$

$$x_{r+1} = x_r - \frac{2c}{b + \text{sign}(b)\sqrt{b^2 - 4ac}}$$

Hallejeva metoda

$$x_{r+1} = x_r - \frac{2f(x_r)f'(x_r)}{2f'(x_r)^2 - f(x)f''(x)}$$