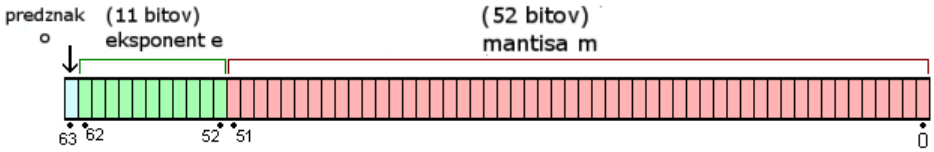
Poglavje 3

Plavajoča vejica



Slika 3.1: Plavajoča vejica

Zapis je oblike $(-1)^o(1.m)2^{e-1023}$, mantisa je v normalizirani obliki, eksponent je podan z zamikom. Več lahko najdete na tej strani.

Naloga 3.1 (Vaje) *Zapiši naslednja števila v dvojni natančnosti:*

(i). 2.71875,

(ii). -0.1,

(iii). 255.125.

Rešitev.

(i). Število najprej pretvorimo v dvojiški zapis. Lotimo se celega dela.

$$\begin{aligned} 2 &= 1*2+0 \\ 1 &= 0*2+1 \end{aligned}$$

$2 = 10_{(2)}$. Potem pa še decimalni del.

$$\begin{aligned} 0.71875*2 &= 1+0.4375 \\ 0.4375*2 &= 0+0.875 \\ 0.875*2 &= 1+0.75 \\ 0.75*2 &= 1+0.5 \\ 0.5*2 &= 1+0 \end{aligned}$$

Torej je $2.71875 = 10.10111_{(2)}$. Ker iščemo zapis normalizirane oblike $(-1)^o(1+m)2^{e-1023}$, moramo število deliti/množiti z dva dokler ne dobimo 1.m. Torej $10.10111_{(2)} = 1.01011_{(2)} * 2$ in $e - 1023 = 1$. Dobili smo $o = 0$, $m = 010111\underbrace{0\ldots0}_{46}$, $e = 10000000000_{(2)}$.

(ii). Število pretvorimo v dvojiški zapis.

$$\begin{aligned} 0.1 * 2 &= 0 + 0.2 \\ 0.2 * 2 &= 0 + 0.4 \\ 0.4 * 2 &= 0 + 0.8 \\ 0.8 * 2 &= 1 + 0.6 \\ 0.6 * 2 &= 1 + 0.2 \\ 0.2 * 2 &= 0 + 0.4 \\ 0.4 * 2 &= 0 + 0.8 \\ 0.8 * 2 &= 1 + 0.6... \end{aligned}$$

$0.1 = 0.000\overline{1100} = 1.100\overline{1100} * 2^{-4}$. Mantisa je dolžine 52, torej moramo na ustreznem mestu odrezati. Dobimo $o = 1$, $m = 100\underbrace{1100}_{11}11010$. Iz $e - 1023 = -4$, dobimo $e = 1019$, torej je $e = 01111111011$. Zadnje števke v mantisi so take zaradi zaokroževanja. Število, ki je najbližje $0. * 11001\underbrace{1}$, je $0. * 11010$. Tukaj $*$ označuje preostale decimalke.

(iii). Število pretvorimo v dvojiški zapis. Celi del:

$$\begin{aligned} 255 &= 2 * 127 + 1 \\ 127 &= 2 * 63 + 1 \\ &\vdots \\ 3 &= 2 * 1 + 1 \\ 1 &= 0 * 2 + 1 \end{aligned}$$

$255 = 11111111_{(2)}$. Še decimalni del:

$$\begin{aligned} 0.125 * 2 &= 0 + 0.25 \\ 0.25 * 2 &= 0 + 0.5 \\ 0.5 * 2 &= 1 + 0 \end{aligned}$$

$0.125 = 0.001_{(2)}$. Dobili smo $255.125 = 11111111.001_{(2)}$. Premaknemo decimalno piko, da dobimo normalizirano obliko. Dobimo $11111111.001_{(2)} = 1.1111111001 * 2^7$. Torej je $o = 0$, $m = 1111111001\underbrace{0\ldots0}_{42}$. Velja še $e - 1023 = 7$, torej je $e = 10000000110_{(2)}$.

■

Naj bo x število in $fl(x)$ najbližje predstavljivo število. Velja

$$fl(x) = x(1+\delta) \text{ in } |\delta| \leq u,$$

kjer je u osnovna zaokrožitvena napaka. Za predstavljivo število velja $fl(x) = x$.

Naloga 3.2 (Vaje) *Podan je tangens kota* $\tan(\alpha) = t$. *Poišči učinkovito metodo za izračun* $\cos(\alpha)$ *in* $\sin(\alpha)$ *brez uporabe* arctan . *Poizkusi se izogniti podkoračenju pri računanju v premični piki. Kje nam podkoračenje prinese največ težav? Kaj se zgodi pri kvadriranju* $t = 2^{-600}$ *v Matlabu?*

Rešitev. Spomnimo se trigonometričnih zvez $1 + \tan^2\alpha = \frac{1}{\cos^2\alpha}$ in $\sin^2\alpha + \cos^2\alpha = 1$. Definirajmo $r = \sqrt{1+t^2}$, $c = \frac{1}{r}$ in $s = \sqrt{1-c^2}$. Pri računanju $\sin\alpha$ bo lahko prišlo do podkoračenja pri majhnem t. Temu se izognemo na sledeči način

$$s = \sqrt{1-c^2} = \sqrt{1-\frac{1}{r^2}} = \sqrt{\frac{r^2-1}{r^2}} = \sqrt{\frac{t^2}{1+t^2}} = \frac{t}{r},$$

kjer se izognemo računanju t^2. ∎

Naloga 3.3 (Vaje) *Izračunati želimo vrednost funkcije* $f(x) = \sqrt{1+x}$ *pri* $x = \frac{1}{13}$, *pri čemer računamo v premični piki na pet decimalk natančno in z zaokrožanjem na najbližje predstavljivo število. Za računanje uporabimo prve tri člene razvoja* $f(x)$ *v taylorjevo vrsto. Oceni celotno napako in jo primerjaj z dejansko.*

Rešitev. Razvoj v taylorjevo vrsto je

$$\sqrt{1+x} = \sum_{i=0}^{\infty} \binom{\frac{1}{2}}{i} x^i = 1 + \binom{\frac{1}{2}}{1} x + \binom{\frac{1}{2}}{2} x^2 + \ldots = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{1}{6}f^3(\xi)x^3.$$

Napaka pri numeričnem računanju je sestavljena iz neodstranljive napake, napake metode in zaokrožitvene napake. Velja

$$D = D_n + D_m + D_z \text{ in seveda } |D| \le |D_n| + |D_m| + |D_z|.$$

- **Osnovna zaokrožitvena napaka**
  Velja $u = \frac{1}{2}10^{-5} = 5 \cdot 10^{-6}$ in $\overline{x} = fl(\frac{1}{13}) = 10^{-1} \cdot 0.76923$.

- **Neodstranljiva napaka**
  Namesto z $x = \frac{1}{13}$ računamo z $\overline{x} = 0.07692$, $D_n = f(x) - f(\overline{x})$. Velja

  $$|f(x) - f(\overline{x})| = |f'(\zeta)||x - \overline{x}| \text{ za } \zeta \text{ med } x \text{ in } \overline{x}.$$

  Ocenimo odvod

  $$f'(x) = \frac{1}{2\sqrt{1+x}} \le \frac{1}{2}.$$

  Tako dobimo oceno za

  $$|D_n| \le \frac{1}{2}|x - \overline{x}| \le \frac{1}{2}u = \frac{1}{2} \cdot 5 \cdot 10^{-6}.$$

  Točna vrednost

  $$D_n = \sqrt{1+\frac{1}{13}} - \sqrt{1+10^{-1}0.76923} \doteq 3.706 \cdot 10^{-8}.$$

- **Napaka metode**
  Namesto vrednosti $f(\overline{x})$ računamo $g(\overline{x})$.

  $$D_m = f(\overline{x}) - g(\overline{x}) = \sqrt{1+\overline{x}} - (1 + \frac{1}{2}\overline{x} - \frac{1}{8}\overline{x}^2) \le \frac{1}{6}f^{(3)}(\xi_{\overline{x}})\overline{x}^3 \le \frac{1}{16}\overline{x}^3 \le 2.85 \cdot 10^{-5}.$$

  Točna vrednost $D_m = 2.8444788 \cdot 10^{-6}$.

- **Zaokrožitvena napaka** nastane zaradi zaokroževanja med računanjem.
  Najprej moramo določiti vrstni red računanja, saj operacije niso več asociativne. Denimo, da najprej izračunamo produkte in nato še ostalo. Poizkusi oceniti napako. Oceni še celotno napako.

■