Aritmética de Ponto Flutuante

Um número real $x \in \mathbb{R}$ pode ser escrito como

$$x = \pm \left(\frac{d_0}{\beta^0} + \frac{d_1}{\beta^1} + \cdots + \frac{d_t}{\beta^t} + \cdots \right) \times \beta^e$$

- β: Base
- $d_0, d_1, \ldots, d_t, \ldots$: dígitos dependem da base
- $e \in \mathbb{Z}$: expoente

Ex:

$$\frac{1}{3} = \left(\frac{3}{10^0} + \frac{3}{10^1} + \frac{3}{10^2} + \cdots + \frac{3}{10^t} + \cdots \right) \times 10^{-1}$$

$$= 3.3333\ldots \times 10^{-1} = 0.3333\ldots = 0.\overline{3}$$

- Sistema de Ponto Flutuante

$$\begin{cases} \beta \rightarrow \text{base} \\ t \rightarrow \text{precisão (nº de dígitos significativos} \\ l \rightarrow \text{Limitante inferior do expoente } e \\ u \rightarrow \text{Limitante superior do expoente } e. \end{cases}$$

$$fl(x) = \underbrace{sgn(x)}_{\text{sinal}} \times (0.\underbrace{\tilde{d}_0 \tilde{d}_1 \cdots \tilde{d}_{t-1}}_{\text{mantissa}}) \times \beta^{e}, \quad e \in [l, u]$$

Para base decimal ($\beta = 10$)

$$fl(x) = sgn(x) \cdot (0.d_0 d_1 \cdots d_{t-1}) \times 10^e$$

($d_0 \neq 0$)

($d_0 d_1 \cdots d_{t-1}$: mantissa finita)

em que t é a precisão, $0 \leq d_j \leq 9$, $d_j \in \mathbb{Z}$, $j = 0, \ldots, t-1$,

$d_0 \neq 0$, e $e \in [l, u] \cap \mathbb{Z}$

Lo Se $x \in \mathbb{R}$, temos infinitos dígitos de x em qualquer base β.

Como guardar em um computador com t dígitos de precisão o número ($\beta = 10$)

$$x = \pm(0.d_0 d_1 \cdots d_{t-1} d_t d_{t+1} \cdots) \times 10^e$$

não cabem todos os dígitos na máquina

Temos duas estratégias:
- Arredondamento
- Truncamento

(t dígitos)

- **Arredondamento**

$$fl_A(x) = \begin{cases} \pm(0.d_0d_1\cdots d_{t-1})\times 10^e, & \text{se } d_t < 5 \\ \pm(0.d_0d_1\cdots d_{t-1} + \frac{1}{10^{t-1}})\times 10^e, & \text{se } d_t \geq 5 \end{cases} \quad (p/2)$$

- **Truncamento**

$$fl_T(x) = \pm(0.d_0d_1\cdots d_{t-1})\times 10^e$$

Ex: $\beta = 10$, $t = 3$

x	trunc	Arred
5.672	5.67	5.67
-5.672	-5.67	-5.67
5.677	5.67	5.68
5.692	5.69	5.69
-5.695	5.69	-5.70

Ex. $x = \frac{8}{3} = 2.6666\ldots$ (t = 4)

$fl(x) = 2.667 \times 10^{0}$

$= 0.2667 \times 10^{1}$

$fl_4(x) = 2.666 \times 10^{0}$

$= 0.2666 \times 10^{1}$

→ normalizado

- 64 bits (double-dupla precisão)

  $\boxed{\beta = 2}$

  1 bit – sinal

  11 bits – expoente

  52 bits – mantissa (dígitos) ~ 16 dígitos decimais

$t \geq 16$ $\quad -1023 \quad 1024$

Ex: $\beta = 10$, $t = 3$ e $l = -5$, $u = 5$.

$d_0 \neq 0$ $\quad e \in [-5, 5]$

maior número representável (em módulo)

$$M = 0.999 \times 10^{5} = 99900$$

menor número representável (em módulo)

$$m = 0.100 \times 10^{-5} = 0.000001$$

Teorema: Seja $x \to fl(x) = g \times \beta^e$, com $x \neq 0$, g é a mantissa normalizada.

Então

$$E_A: \quad |x - fl(x)| \leq \begin{cases} \beta^{1-t} \times \beta^e, & \text{p/ truncamento} \\ \frac{1}{2}\beta^{1-t} \times \beta^e, & \text{p/ arredondamento} \end{cases}$$

$$E_R: \quad \frac{|x - fl(x)|}{|x|} \leq \begin{cases} \beta^{1-t}, & \text{p/ truncamento} \\ \frac{1}{2}\beta^{1-t}, & \text{p/ arredondamento} \end{cases}$$

Erro nas Operação de P. flutuante

Ex: $p=10$, $t=4$ e $[e,r)$ suficientes

$x = 0.937 \times 10^4$ e $y = 0.1272 \times 10^2$

$x+y \Rightarrow fl(fl(x) + fl(y))$

$$fe(x+y) = fe(0.937 \times 10^4 + 0.001272 \times 10^4)$$

$$= fl(0.9382\,72 \times 10^4)$$

$$= \begin{cases} 0.9382 \times 10^4 & (T) \\ 0.9383 \times 10^4 & (A) \end{cases}$$

$fp(x * y) = ?$

TAKEFA