Aritmética de Ponto Flutuante

Márcio Antônio de Andrade Bortoloti
mbortoloti@uesb.edu.br
https://mbortoloti.github.io

Cálculo Numérico

Departamento de Ciências Exatas e Tecnológicas - DCET
Universidade Estadual do Sudoeste da Bahia

Sumário

Aritmética de Ponto Flutuante

Definição
Um sistema de representação numérica em uma máquina, $\mathcal{F}(\beta, t, l, u)$ será chamado de *Aritmética de Ponto Flutuante*. Nesse sistema, um número r será representado da forma

$$r = \pm(\cdot d_1 d_2 \cdots d_t) \times \beta^e,$$

onde

- β é a base;
- t é o número de dígitos na mantissa;
- $0 \leq d_j \leq (\beta - 1)$, $j = 1, \cdots, t$ e $d_1 \neq 0$;
- e é o expoente no intervalo $[l, u]$.

Exemplo:

Considere uma máquina que opera no sistema $\mathcal{F}(10, 3, -5, 5)$. Os números serão representados da segiunte forma, neste sistema,

$$0.d_1d_2d_3 \times 10^e, \quad e \in [-5, 5], \quad 0 \leq d_j \leq 9 \quad \text{e } d_1 \neq 0.$$

- Qual o menor número, em valor absoluto (diferente de zero), que pode ser representado nessa máquina? $m = 0.100 \times 10^{-5} = 10^{-6}$.
- E o maior ? $M = 0.999 \times 10^5 = 99900$

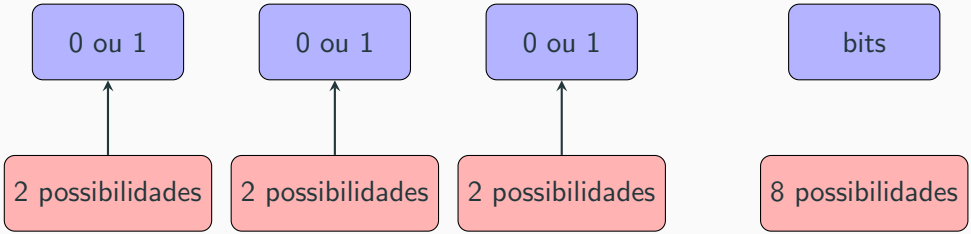
Assim, se $x \in \mathcal{F}(10, 3, -5, 5)$ então $m \leq |x| \leq M$.

Observações:

1. Se $x = 123.456 = 0.123456 \times 10^3$ então x não pode ser representado de forma exata em $\mathcal{F}(10, 3, -5, 5)$.
   Neste caso é necessário aplicar um processo de truncamento ou arredondamento (veremos isso logo mais!).
2. Note que não existe nenhum número entre 0.123×10^2 e 0.124×10^2 que pertença a $\mathcal{F}(10, 3, -5, 5)$.
3. Se $|x| < m$ então x não poderá ser representado em $\mathcal{F}(10, 3, -5, 5)$. Neste caso dizemos que ocorre *underflow*.
4. Se $|x| > M$ então x não poderá ser representado em $\mathcal{F}(10, 3, -5, 5)$. Neste caso dizemos que ocorre *overflow*.

Observações:

- Em um computador padrão considera-se $\beta = 2$. Isso implica que $d_i = 0$ ou $d_i = 1$.
- Em um computador padrão de 3 bits tem-se



Aritmética de Ponto Flutuante

Em um computador de 3 bits pode ser definido:

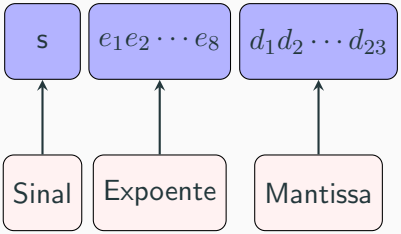
Binário	000	001	010	011	100	101	110	111
Decimal	0	1	2	3	-4	-3	-2	-1

$$\begin{array}{r} 001 \\ +\ 010 \\ \hline 011 \end{array} \qquad \begin{array}{r} 1 \\ +\ 2 \\ \hline 3 \end{array} \qquad\qquad \begin{array}{r} 001 \\ +\ 011 \\ \hline 100 \end{array} \qquad \begin{array}{r} 1 \\ +\ 3 \\ \hline -4 \end{array}$$

Overflow

Aritmética de Ponto Flutuante

Em um computador de 32 bits



Em um computador de 64 bits