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 $m \leq |x| \leq M$.

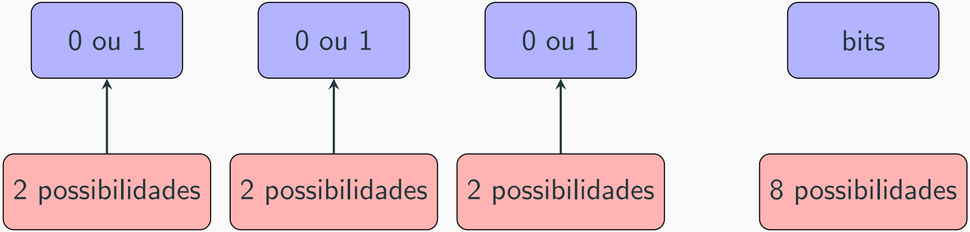
Observações:

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*.

Aritmética de Ponto Flutuante

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:

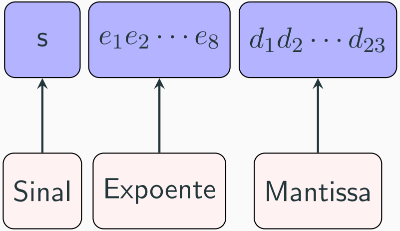
Binário	000	001	010	011	100	101	110	111
Decimal	0	1	2	3	-4	-3	-2	-1

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

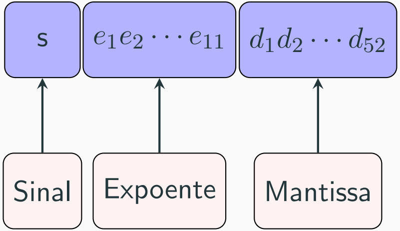
Overflow

Aritmética de Ponto Flutuante

Em um computador de 32 bits



Em um computador de 64 bits



Análise de Erros

Definição
Se $x \in \mathcal{F}(\beta, t, m, M))$ então ele pode ser representado como

$$x = f_x \times 10^e + g_x \times 10^{e-t},$$

onde $0.1 \leq f_x < 1$ e $0 \leq g_x < 1$.

Exemplo:

Seja $x = 234.57$ e $t = 4$. logo

$$\begin{aligned} x &= 234.57 \\ &= 0.23457 \times 10^3 \\ &= (0.2345 + 0.00007) \times 10^3 \\ &= 0.2345 \times 10^3 + 0.00007 \times 10^3 \\ &= 0.2345 \times 10^3 + 0.7 \times 10^{-1} \end{aligned}$$

Definição de Truncamento
Seja $\mathcal{F}(10, t, m, M)$ uma máquina e x um número que em geral não pode ser representado em $\mathcal{F}$ de forma exata. Quando isso ocorre, devemos utilizar uma aproximação $\overline{x}$ para x. Assim, se x é tal que

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \text{ para } m \leq e \leq M,$$

onde $0.1 \leq f_x < 1$ e $0 \leq g_x < 1$ então a operação de truncamento gera uma aproximação $\overline{x}$, de x, da forma

$$\overline{x} = f_x \times 10^e.$$

Definição de Arredondamento
Seja $\mathcal{F}(10,t,m,M)$. No caso de obtermos uma aproximação, $\overline{x}$, de um número $x = f_x \times 10^e + g_x \times 10^{e-t}$, usando arredondamento, teremos que analisar g_x de forma que

$$\overline{x} = \begin{cases} f_x \times 10^e & \text{se } g_x < 1/2 \\ f_x \times 10^e + 10^{e-t} & \text{se } g_x \geq 1/2 \end{cases}$$

Exemplo:

Considere uma máquina $\mathcal{F}(10,3,-5,5)$. Vamos representar $x = 45.8787$ em $\mathcal{F}$. De fato,

$$x = 45.8787 = 0.458 \times 10^2 + 0.787 \times 10^{-1}$$

Fazendo o arredondamento

$$\overline{x} = 0.458 \times 10^2 + 10^{-1} = 0.459 \times 10^2 = 0.459$$

Erro Absoluto

Definição

Seja $x \in \mathbb{R}$ e $\overline{x}$ sua aproximação. O erro absoluto, cometido na representação de x por $\overline{x}$ é definido por

$$EA_x = x - \overline{x}.$$

Exemplo

O erro absoluto cometido na aproximação de π por $\overline{\pi} = 3.14$ é

$$|EA_\pi| = |\pi - \overline{\pi}| = |\pi - 3.14| \leq 0.01.$$

Erro Relativo

Considere dois números $x = 1991.67$ e $y = 3.67$. Se aproximarmos x e y por $\overline{x} = 1991.7$ e $\overline{y} = 3.7$ teremos

$$|EA_x| = |EA_y| = 0.03.$$

No entanto, os dois números estão aproximados da “mesma forma” ?

Erro Relativo

Considere dois números $x = 1991.67$ e $y = 3.67$. Se aproximarmos x e y por $\overline{x} = 1991.7$ e $\overline{y} = 3.7$ teremos

$$|EA_x| = |EA_y| = 0.03.$$

No entanto, os dois números estão aproximados da "mesma forma" ? Qual aproximação está mais precisa ?

Para responder a pergunta vamos usar a seguinte definição:

Erro Relativo

Definição
O Erro Relativo, ER_x, cometido na aproximação de x por $\overline{x}$ é definido como

$$ER_x = \frac{EA_x}{\overline{x}} = \frac{x - \overline{x}}{\overline{x}}$$

Voltando ao exemplo ...

Se $x = 1991.67$ e $y = 3.67$ as aproximações $\overline{x} = 1991.7$ e $\overline{y} = 3.7$ cometem erros relativos da ordem de

$$|ER_x| = \frac{|EA_x|}{|\overline{x}|} = \frac{0.03}{1991.7} = 1.506250941 \times 10^{-5}.$$

$$|ER_y| = \frac{|EA_y|}{|\overline{y}|} = \frac{0.03}{3.7} = 0.810810810 \times 10^{-2}.$$

Teorema

Sejam $x \in \mathbb{R}$ e $\mathcal{F}(10, t, m, M)$ uma máquina. Os erros absoluto e relativo cometidos na aproximação de x por $\overline{x}$, utilizando truncamento, são da ordem de

$$|EA_x| = |x - \overline{x}| < 10^{e-t} \quad \text{e} \quad |ER_x| = \frac{|EA_x|}{|\overline{x}|} < 10^{-t+1}.$$

Prova:

Note que

$$x = f_x \times 10^e + g_x \times 10^{e-t},$$

onde $0.1 \leq f_x < 1$ e $0 \leq g_x < 1$.

Usando o truncamento, tem-se

$$\overline{x} = f_x \times 10^e.$$

Logo

$$
\begin{aligned}
|EA_x| &= |x - \overline{x}| \\
&= |f_x \times 10^e + g_x \times 10^{e-t} - f_x \times 10^e| \\
&= |g_x| \times 10^{e-t} \\
&< 10^{e-t} \quad (|g_x| < 1)
\end{aligned}
$$

Agora, o erro relativo ...

$$
\begin{aligned}
|ER_x| &= \frac{|EA_x|}{|\overline{x}|} \\
&= \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e} \\
&< \frac{10^{e-t}}{0.1 \times 10^e} \\
&< 10^{-t+1}
\end{aligned}
$$

Teorema
Sejam $x \in \mathbb{R}$ e $\mathcal{F}(10, t, m, M)$ uma máquina. Os erros absoluto e relativo cometidos na aproximação de x por $\overline{x}$, utilizando arredondamento, são da ordem de

$$|EA_x| =\leq 0.5 \times 10^{e-t} \quad \text{e} \quad ER_x = 0.5 \times 10^{-t+1}.$$

Prova:

Note que

$$x = f_x \times 10^e + g_x \times 10^{e-t},$$

onde $0.1 \leq f_x < 1$ e $0 \leq g_x < 1$ e

$$\overline{x} = \begin{cases} f_x \times 10^e & \text{se } |g_x| < 1/2 \\ f_x \times 10^e + 10^{e-t} & \text{se } |g_x| \geq 1/2 \end{cases}$$

Se $g_x < 1/2$ então

$$
\begin{aligned}
|EA_x| &= |x - \overline{x}| = |g_x| \times 10^{e-t} \\
&< \frac{1}{2} \times 10^{e-t}
\end{aligned}
$$

E também

$$
\begin{aligned}
|ER_x| &= \frac{|EA_x|}{|\overline{x}|} \\
&= \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^{e}} \\
&< \frac{0.5 \times 10^{e-t}}{0.1 \times 10^{e}} \\
&< \frac{1}{2} \times 10^{-t+1}
\end{aligned}
$$

Análise de Erros

Se $g_x \geq 1/2$ então

$$
\begin{aligned}
|EA_x| &= |x - \overline{x}| \\
&= |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})| \\
&= |g_x \times 10^{e-t} - 10^{e-t}| \\
&= |g_x - 1| \times 10^{e-t} \\
&\leq \frac{1}{2} \times 10^{e-t}
\end{aligned}
$$

E também

$$
\begin{aligned}
|ER_x| &= \frac{|EA_x|}{|\overline{x}|} \leq \frac{1/2 \times 10^{e-t}}{|f_x \times 10^e + 10^{e-t}|} \\
&< \frac{1/2 \times 10^{e-t}}{|f_x| \times 10^e} < \frac{1/2 \times 10^{e-t}}{0.1 \times 10^e} < \frac{1}{2} \times 10^{-t+1}
\end{aligned}
$$

Operações em Aritmética de Ponto Flutuante

- O arredondamento não é muito utilizado, pois mesmo acarretando erros menores, ele aumenta o tempo de execução de um programa.
- Mesmo que x e y estejam representados de forma exata, a soma $x + y$, por exemplo, também gera erros numéricos.
- **Exemplo:** Sejam $x = 0.234 \times 10^5$ e $y = 0.567 \times 10^2$ em uma máquina $\mathcal{F}(10, 3, -5, 5)$. Então

$$
\begin{aligned}
x + y &= 0.234 \times 10^5 + 0.567 \times 10^2 \\
&= 0.234 \times 10^5 + 0.000567 \times 10^5 \\
&= (0.234 + 0.000567) \times 10^5 \\
&= 0.234567 \times 10^5 \\
&= 0.234 \times 10^5 \quad \text{(se truncarmos)} \\
&= 0.235 \times 10^5 \quad \text{(se arredondarmos)}
\end{aligned}
$$