CÁLCULO NUMÉRICO

"existem 10 tipos de pessoas no mundo..."

① Representação de números no computador

* bit : → 0 ou 1 (com ou sem corrente)
  → menor unidade de informação
  → números binários

* Representação dos números:

$$(1328)_{10} = 1 \times 10^3 + 3 \times 10^2 + 2 \times 10^1 + 8 \times 10^0$$

$$= 1000 + 300 + 20 + 8$$

⇒ em qualquer base:

$$(a_3 a_2 a_1 a_0)_\beta = (a_3\beta^3 + a_2\beta^2 + a_1\beta^1 + a_0\beta^0)_{10}$$

onde $0 \leq a_i \leq \beta - 1$

⇒ base 2: $(10010)_2 = (1 \times 2^4 + 0 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 0 \times 2^0)_{10}$

$$= (16 + 2)_{10} = (18)_{10}$$

- $(0)_2 = (0 \times 2^0)_{10} = (0)_{10}$
- $(1)_2 = (1 \times 2^0)_{10} = (1)_{10}$
- $(10)_2 = (1 \times 2^1 + 0 \times 2^0)_{10} = (2)_{10}$
- $(11)_2 = (1 \times 2^1 + 1 \times 2^0)_{10} = (3)_{10}$
- $(100)_2 = (1 \times 2^2 + 0 \times 2^1 + 0 \times 2^0)_{10} = (4)_{10}$

- $(28)_{10} = ?$

$$
\begin{array}{l}
28 \;\lfloor 2 \\
0 \quad 14 \;\lfloor 2 \\
\quad 0 \quad 7 \;\lfloor 2 \\
\qquad 1 \quad 3 \;\lfloor 2 \\
\qquad\quad 1 \quad 1 \;\lfloor 2 \\
\qquad\qquad 1 \quad 0
\end{array}
$$

$= (11100)_2$

- $x = 2 \rightarrow (10)_2 \rightarrow 2$ bits
- $x = 0{,}25$?
- $x = 1/3 = 0{,}333$?
- $x = \sqrt{2}$?
  - ↳ irracional, $\neq \frac{A}{B}$

} não tem solução exata no computador

números reais → $\mathbb{R}$ (infinitos)

inteiros $\mathbb{Z}$ (infinitos contáveis → $\mathbb{N}$)

$\mathbb{F}$ → ponto flutuante

⇒ $\mathbb{F}$ → ponto flutuante → número finito de casas decimais

↳ arredondamento

$$x = 1/3 = 0{,}3333$$

$$x = 2/3 = 0{,}6667$$

⇒ posição do número decimal não é fixa

$$x = (-1)^s (0.\underbrace{a_1 a_2 \ldots a_t}_{m}) \beta^e = (-1)^s m \beta^{e-t}$$

- β = base ≥ 2 (inteiro)
- m = mantissa (inteiro) $\quad 0 \leq a_i \leq \beta - 1$
- t = nº de dígitos da mantissa
- e = expoente (inteiro), $\quad$ • s = sinal $\begin{cases} 0 & \oplus \\ 1 & \ominus \end{cases}$

ex) • $250 = (-1)^0\, 0{,}25 \times 10^3 = (-1)^{0}\, 25 \times 10^{1}$

• $-0{,}33 = (-1)^1\, 0{,}33 \times 10^0 = (-1)^{1}\, 33 \times 10^{-2}$

• $0{,}001 = (-1)^0\, 0{,}1 \times 10^{-2} = (-1)^{0}\, 1 \times 10^{-3}$

obs: se $a_1 \neq 0$, representação é única.

* Padrão IEEE 754

$$(-1)^S\, 1.m \times 2^{e-b} \qquad E = e - b$$

viés (b)

↳ implícito (1)

• byte = 8 bits
• half = 16 bits (2 bytes)
• single = 32 bits (4 bytes)
• double = 64 bits (8 bytes)

half:

S	1	2	3	4	5	1	2	3	4	5	6	7	8	9	10
s	e					m									

Casos especiais:
$$\begin{cases} 0 \quad 00000 \quad 0000000000 \to (0)_{10} \\ 0 \quad 11111 \quad \text{"} \to +\infty \\ 1 \quad 11111 \quad \text{"} \to -\infty \end{cases}$$

obs: $e = \begin{cases} 00000 \\ 11111 \end{cases}$ são casos especiais

$(1)_{10} = (1)_2 \Rightarrow 0 \quad 01111 \quad 0000000000$

$= 1 \times 2^0 = (-1)^0\, 1{,}0 \times 2^0 \qquad S = 0,\ m = 0,\ E = 0$

$\Rightarrow$ half: 16 bits: 1(s), 5(e), 10(m) $\mathbb{F}(2, 11, -14, 15)$

- $e_{max} = (11110)_2 = (30)_{10}$
- $e_{min} = (00001)_2 = (1)_{10}$

números pequenos

- $e = [1, 30] \Rightarrow b = 15 \Rightarrow E = [-14, 15]$

números grandes

- $|x|_{max} = 0\ 11110\ 1111111111$

$$= (-1)^0 (1,1111111111)_2 \times 2^{(11110)_2 - (15)_{10}}$$

$$= [1 \times 2^0 + (1 - 2^{-10})] \times 2^{15}$$

$$= (2 - 2^{-10})\, 2^{15} = 65504 = 6{,}55 \times 10^4$$

- $|x|_{min} = 0\ 00001\ 0000000000$

$$= (-1)^0\ 1{,}0 \times 2^{(00001)_2 - (15)_{10}}$$

$$= 1 \times 2^{-14} = 1/16384 = 6{,}10 \times 10^{-5}$$

- precisão: 10 (mantissa) + 1 = 11 bits

$$\underbrace{(11111111111)_2}_{11\ \text{bits}} = \underbrace{(2047)_{10}}$$

a partir de 2048, precisa de 12 bits, logo, todos os números com 3 dígitos (até 999) são representados corretamente.

$$\boxed{p_{10} = 3}$$

$\Rightarrow$ single: 32 bits : 1(s), 8(e), $\underline{23}$(m) $\quad \mathbb{F}(2, 24, -126, 127)$

- $e_{max} = (11111110)_2 = (254)_{10}$
- $e_{min} = (00000001)_2 = (1)_{10}$
- $e = [1, 254] \Rightarrow b = 127 \Rightarrow E = [-\underline{126}, \underline{127}]$
- $|x|_{max} = (2 - 2^{-\underline{23}}) \times 2^{\underline{127}} = 3,4028235 \times 10^{38}$
- $|x|_{min} = 1 \times 2^{-\underline{126}} = 1,1754944 \times 10^{-38}$
- $p_{10} = 7$ (24 bits em $\beta = 10$)

$\Rightarrow$ double: 64 bits : 1(s), 11(e), 52(m) $\quad \mathbb{F}(2, 53, -1022, 1023)$

- $e_{max} = (11111111110)_2 = (2046)_{10}$
- $e_{min} = (00000000001)_2 = (1)_{10}$
- $e = [1, 2046] \rightarrow b = 1023 \rightarrow E = [-1022, 1023]$
- $|x|_{max} = (2 - 2^{-52})\, 2^{1023} = 1,797693 \cdots \times 10^{308}$
- $|x|_{min} = 1 \times 2^{-1022} = 2,2250 \cdots \times 10^{-308}$
- $p_{10} = 15$ (53 bits em $\beta = 10$)

$\mathbb{F}(\beta, t, E_{min}, E_{max}) \qquad E_{min} \leqslant E \leqslant E_{max}$

obs: Quarteroni : $\mathbb{F}(2, 53, \underline{-1021, 1024}) \Rightarrow (-1)^s\, \underline{0.1m}\, 2^{e-b}$

Aula 2

1 Representação de números no computador

$\Rightarrow$ ponto flutuante : $\mathbb{F}(\beta, t, \underbrace{E_{min}, E_{max}}_{\text{decimal}})$, $E_{min} \leq E \leq E_{max}$

* Padrão IEEE 754 : $x = (-1)^s \, \underbrace{1}_{\text{implícito}} , m \times 2^{e-b}$ ← viés

$$\begin{cases} \beta = \text{base} = 2 \\ t = \underset{\text{binária "real"}}{\text{precisão}} = m+1 \quad (m = \text{memória}) \quad t = \text{mantissa real} \\ E = \text{expoente} = \underbrace{e}_{\text{memória}} - b \end{cases}$$

• half : $\mathbb{F}(2, 11, -14, 15)$ $\begin{cases} 16 \text{ bits}: 1(s), 5(e), 10(m) \\ b = 15 \end{cases}$

• single : $\mathbb{F}(2, 24, -126, 127)$ $\begin{cases} 32 \text{ bits}: 1(s), 8(e), 23(m) \\ b = 127 \end{cases}$

• double : $\mathbb{F}(2, 53, -1022, 1023)$ $\begin{cases} 64 \text{ bits}: 1(s), 11(e), 52(m) \\ b = 1023 \end{cases}$

---

ex: $\begin{cases} x = (1)_{10} \\ x = (1)_2 \end{cases}$ $x = (-1)^0 \, 1,0 \times 2^0$ $s = 0, \; E = 0, \; m = 0$

$e = E + b = 0 + 15 = (15)_{10}$

$e = 1111$

$x_{half} = 0 \; 01111 \; 0000000000$

---

obs: Quarteroni: double: $\mathbb{F}(2, 53, \underline{-1021, 1024})$

pois $x = (-1)^s \, 0,\underline{1m} \, 2^{e-b}$.

$|x|_{min} = 2^{E_{min}}$, $|x|_{max} = (2 - 2^{-\overset{t-1}{m}}) \, 2^{E_{max}}$

* precisão em decimal:

- half: 11 bits em $\beta = 2 \Rightarrow (11111111111)_2 = (2047)_{10}$

  $p_{10} = 3$ (até 999).

- single: 24 bits em $\beta = 2 \Rightarrow (111\ldots11)_2 = 16.777.215$

  $p_{10} = 7$ (até 9.999.999)

- double: 53 bits em $\beta = 2 \Rightarrow (111\ldots11)_2 = ?$

  $p_{10} = 15$

$$\begin{cases} * \ |x| > |x|_{max} \Rightarrow \text{overflow } (\pm\infty) \\ * \ |x| < |x|_{min} \Rightarrow \text{underflow (zero)} \end{cases}$$

---

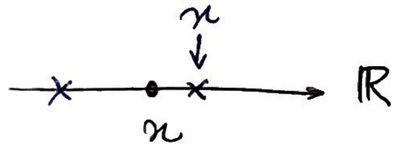
* exemplo memória: 1 milhão de dados

- half: $10^6 \times 2$ bytes ≈ 2 Mb
- single: $10^6 \times 4$ bytes ≈ 4 Mb
- double: $10^6 \times 8$ bytes ≈ 8 Mb

  ↳ usado pela RAM ou armazenamento binário

  ↳ ASCii → caracteres que representam números ⇒ ~ 3x arquivo binário.

(1.1) Erro em ponto flutuante:



- $EA_x = |x - \hat{x}|$ → erro absoluto, minimizado na escolha de $\hat{x}$ (arredondamento)

- $ER_x = \dfrac{|x - \hat{x}|}{|x|}$ → erro relativo, ~ constante em ponto flutuante ex) $\beta = 10,\ p_{10} = 4$

$\Rightarrow$ ponto fixo:

a) $x = 3507,6$; $\hat{x} = 3508$; $EA_x = 0,4$; $ER_x = \dfrac{0,4}{3507,6} \simeq 1,1 \times 10^{-4}$

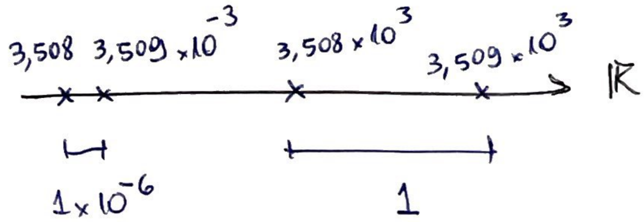
b) $x = 0,0035076$; $\underline{\hat{x} = 0,004}$; $EA_x = 0,0004924$; $ER_x \approx 0,14$

$\Rightarrow$ ponto flutuante:

a) $x = 3507,6 = 3,5076 \times 10^{3}$; $\hat{x} = 3,508 \times 10^{3}$; $EA_x = 0,4$; $ER_x \approx \underline{1,1 \times 10^{-4}}$

b) $x = 0,0035076 = 3,5076 \times 10^{-3}$; $\underline{\hat{x} = 3,508 \times 10^{-3}}$; $EA_x = 0,0004 \times 10^{-3}$; $ER_x = \dfrac{0,0004 \times 10^{-3}}{3,5076 \times 10^{-3}} \simeq \underline{1,1 \times 10^{-4}}$

obs: quanto maior o valor absoluto, maior o erro absoluto e o espaçamento na reta $\mathbb{F}$



---

* operações aritméticas com ponto flutuante

a) associativa: $(2+3)+4 = 2+(3+4)$; $(2 \times 3) \times 4 = 2 \times (3 \times 4)$ $\quad \times\ \mathbb{F}$

b) comutativa: $x+y = y+x$ ✓ F

$xy = yx$

c) distributiva: $2(1+3) = 2 \cdot 1 + 2 \times 3$ X F

erros de arredondamento a cada operação.

ex) $\begin{cases} (23,4 + 5,18) + 3,05 = 31,7 \\ 23,4 + (5,18 + 3,05) = 31,6 \end{cases}$

ex) $\begin{cases} 3,18\,(5,05 + 11,4) = 52,5 \\ 3,18 \cdot 5,05 + 3,18 \times 11,4 = 52,4 \end{cases}$

ex) $$\frac{(1+x)-1}{x} = 1$$

se $x = 1 \times 10^{-15}$ (próximo da precisão de máquina)

$$\frac{(1+x)-1}{x} = 1,1102\ldots$$ erro de 11% !

obs: - trabalhe com números na ordem de 1 !

- Cap 1 Quarteroni: