

Sistema de Ponto Flutuante

Prof. Americo Cunha

Universidade do Estado do Rio de Janeiro – UERJ

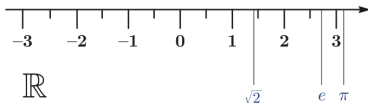
americo.cunha@uerj.br

www.americocunha.org



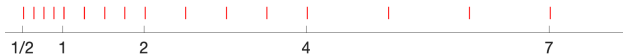
Números no Papel

Uma quantidade infinita de números ao longo de uma reta contínua



Números no Computador

Uma quantidade finita de números ao longo de uma reta discreta



A noção de sistema de ponto flutuante

O *sistema de ponto flutuante* $F \subset \mathbb{R}$ é subconjunto dos reais cujos elementos tem o seguinte formato:

$$\underbrace{\text{fl}(x)}_{\text{float de } x} = \pm \underbrace{(0, d_1 d_2 \cdots, d_t)_\beta}_{\text{mantissa}} \times \beta^e,$$

onde os dígitos $\{d_i\}_{i=1}^t$ são inteiros tais que $0 \leq d_i \leq \beta - 1$ e $d_1 \neq 0$.

O sistema é caracterizado por quatro números inteiros:

- a *base* $\beta > 1$ (também chamada de *radix*),
- a *precisão* $t \geq 1$ (quantidade de dígitos significativos), e
- o *intervalo do expoente* $m \leq e \leq M$.

Notação:

$$F(\beta, t, m, M)$$

ou

$$(\beta, t, m, M)$$



Um sistema de ponto flutuante simplista

Considere o sistema de ponto flutuante

$$F(\beta, t, m, M) = F(2, 3, -1, 3)$$

- Quais números são representáveis nesse sistema?
- Qual o menor número representável (em módulo e não nulo)?
- Qual o maior número representável (em módulo)?



Um sistema de ponto flutuante simplista

$$f1(x) = \pm (0, d_1 d_2 d_3)_2 \times 2^e, \quad d_1 = 1, \quad d_2, d_3 \in \{0, 1\}, \quad -1 \leq e \leq 3$$



Um sistema de ponto flutuante simplista

$$fl(x) = \pm (0, d_1 d_2 d_3)_2 \times 2^e, \quad d_1 = 1, \quad d_2, d_3 \in \{0, 1\}, \quad -1 \leq e \leq 3$$

$$\pm (0, 100)_2 \times 2^{-1} =$$

$$\pm (0, 100)_2 \times 2^{+0} =$$

$$\pm (0, 100)_2 \times 2^{+1} =$$

$$\pm (0, 100)_2 \times 2^{+2} =$$

$$\pm (0, 100)_2 \times 2^{+3} =$$

$$\pm (0, 101)_2 \times 2^{-1} =$$

$$\pm (0, 101)_2 \times 2^{+0} =$$

$$\pm (0, 101)_2 \times 2^{+1} =$$

$$\pm (0, 101)_2 \times 2^{+2} =$$

$$\pm (0, 101)_2 \times 2^{+3} =$$

$$\pm (0, 110)_2 \times 2^{-1} =$$

$$\pm (0, 110)_2 \times 2^{+0} =$$

$$\pm (0, 110)_2 \times 2^{+1} =$$

$$\pm (0, 110)_2 \times 2^{+2} =$$

$$\pm (0, 110)_2 \times 2^{+3} =$$

$$\pm (0, 111)_2 \times 2^{-1} =$$

$$\pm (0, 111)_2 \times 2^{+0} =$$

$$\pm (0, 111)_2 \times 2^{+1} =$$

$$\pm (0, 111)_2 \times 2^{+2} =$$

$$\pm (0, 111)_2 \times 2^{+3} =$$



Um sistema de ponto flutuante simplista

$$fl(x) = \pm (0, d_1 d_2 d_3)_2 \times 2^e, \quad d_1 = 1, \quad d_2, d_3 \in \{0, 1\}, \quad -1 \leq e \leq 3$$

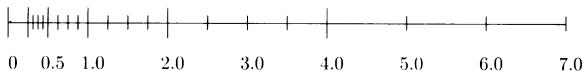
$\pm (0, 100)_2 \times 2^{-1}$	$=$	$\pm 0,2500$	$\pm (0, 110)_2 \times 2^{-1}$	$=$	$\pm 0,3750$
$\pm (0, 100)_2 \times 2^{+0}$	$=$	$\pm 0,5000$	$\pm (0, 110)_2 \times 2^{+0}$	$=$	$\pm 0,7500$
$\pm (0, 100)_2 \times 2^{+1}$	$=$	$\pm 1,0000$	$\pm (0, 110)_2 \times 2^{+1}$	$=$	$\pm 1,5000$
$\pm (0, 100)_2 \times 2^{+2}$	$=$	$\pm 2,0000$	$\pm (0, 110)_2 \times 2^{+2}$	$=$	$\pm 3,0000$
$\pm (0, 100)_2 \times 2^{+3}$	$=$	$\pm 4,0000$	$\pm (0, 110)_2 \times 2^{+3}$	$=$	$\pm 6,0000$
$\pm (0, 101)_2 \times 2^{-1}$	$=$	$\pm 0,3125$	$\pm (0, 111)_2 \times 2^{-1}$	$=$	$\pm 0,4375$
$\pm (0, 101)_2 \times 2^{+0}$	$=$	$\pm 0,6250$	$\pm (0, 111)_2 \times 2^{+0}$	$=$	$\pm 0,8750$
$\pm (0, 101)_2 \times 2^{+1}$	$=$	$\pm 1,2500$	$\pm (0, 111)_2 \times 2^{+1}$	$=$	$\pm 1,7500$
$\pm (0, 101)_2 \times 2^{+2}$	$=$	$\pm 2,5000$	$\pm (0, 111)_2 \times 2^{+2}$	$=$	$\pm 3,5000$
$\pm (0, 101)_2 \times 2^{+3}$	$=$	$\pm 5,0000$	$\pm (0, 111)_2 \times 2^{+3}$	$=$	$\pm 7,0000$



Um sistema de ponto flutuante simplista

$$fl(x) = \pm (0, d_1 d_2 d_3)_2 \times 2^e, \quad d_1 = 1, \quad d_2, d_3 \in \{0, 1\}, \quad -1 \leq e \leq 3$$

$\pm (0, 100)_2 \times 2^{-1}$	$=$	$\pm 0,2500$	$\pm (0, 110)_2 \times 2^{-1}$	$=$	$\pm 0,3750$
$\pm (0, 100)_2 \times 2^0$	$=$	$\pm 0,5000$	$\pm (0, 110)_2 \times 2^0$	$=$	$\pm 0,7500$
$\pm (0, 100)_2 \times 2^1$	$=$	$\pm 1,0000$	$\pm (0, 110)_2 \times 2^1$	$=$	$\pm 1,5000$
$\pm (0, 100)_2 \times 2^2$	$=$	$\pm 2,0000$	$\pm (0, 110)_2 \times 2^2$	$=$	$\pm 3,0000$
$\pm (0, 100)_2 \times 2^3$	$=$	$\pm 4,0000$	$\pm (0, 110)_2 \times 2^3$	$=$	$\pm 6,0000$
$\pm (0, 101)_2 \times 2^{-1}$	$=$	$\pm 0,3125$	$\pm (0, 111)_2 \times 2^{-1}$	$=$	$\pm 0,4375$
$\pm (0, 101)_2 \times 2^0$	$=$	$\pm 0,6250$	$\pm (0, 111)_2 \times 2^0$	$=$	$\pm 0,8750$
$\pm (0, 101)_2 \times 2^1$	$=$	$\pm 1,2500$	$\pm (0, 111)_2 \times 2^1$	$=$	$\pm 1,7500$
$\pm (0, 101)_2 \times 2^2$	$=$	$\pm 2,5000$	$\pm (0, 111)_2 \times 2^2$	$=$	$\pm 3,5000$
$\pm (0, 101)_2 \times 2^3$	$=$	$\pm 5,0000$	$\pm (0, 111)_2 \times 2^3$	$=$	$\pm 7,0000$



Um sistema de ponto flutuante simplista

- 41 números são representáveis:
40 números tabela anterior, mais o zero!
- O menor número representável (em módulo e não nulo) é
$$L = (0, 100)_2 \times 2^{-1} = 0,25.$$
- O maior número representável (em módulo) é
$$U = (0, 111)_2 \times 2^3 = 7,0.$$



Alguns fatos sobre sistemas de ponto flutuante

No sistema de ponto flutuante $F(\beta, t, m, M)$:

- O menor número representável (em módulo e não nulo) é

$$L = (0, 1 \underbrace{00 \cdots 0}_{t-1 \text{ vezes}})_\beta \times \beta^m$$

- O maior número representável (em módulo) é

$$U = (0, \underbrace{(\beta - 1)(\beta - 1) \cdots (\beta - 1)}_{t \text{ vezes}})_\beta \times \beta^M$$

- O número zero admite diversas representações:

$$\text{fl}(0) = (0, \underbrace{00 \cdots 0}_t)_\beta \times \beta^e, \quad m \leq e \leq M$$



A geometria de um sistema de ponto flutuante

- A região de **underflow** é definida por

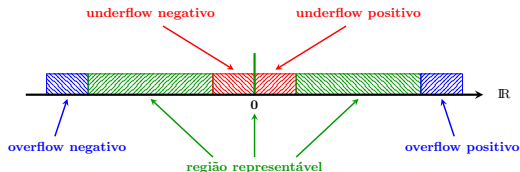
$$\mathcal{U} = \{f1(x) \in F(\beta, t, m, M) \mid |f1(x)| < L \text{ e } f1(x) \neq 0\}$$

- A região de **overflow** é definida por

$$\mathcal{O} = \{f1(x) \in F(\beta, t, m, M) \mid |f1(x)| > U\}$$

- A **região representável** é definida por

$$\mathcal{R} = \{f1(x) \in F(\beta, t, m, M) \mid L \leq |f1(x)| \leq U\} \cup \{0\}$$



* Figura elaborada por Marcus Vinicius Issa.

Representação exata ou aproximada

O real a seguir *não admite representação exata* em $F(\beta, t, m, M)$:

$$x = (0, d_1 d_2 \cdots d_{t-1} d_t d_{t+1} d_{t+2} \cdots)_\beta \times \beta^e$$

Isso ocorre porque *x tem mais de t dígitos!*

Nesses casos uma *representação não exata (aproximação)* do real x em $F(\beta, t, m, M)$ se faz necessária.

Existem duas estratégias para construir tal aproximação:

- *Truncamento*
- *Arredondamento*



Truncamento ou arredondamento?

- Truncamento

$$\text{fl}(x) = \pm (0, d_1 d_2 \cdots, d_t)_\beta \times \beta^e$$

- Arredondamento

$$\text{fl}(x) = \begin{cases} \pm (0, d_1 d_2 \cdots d_t)_\beta \times \beta^e & \text{se } d_{t+1} < \beta/2 \\ \pm (0, d_1 d_2 \cdots d_t + \beta^{-t})_\beta \times \beta^e & \text{se } d_{t+1} > \beta/2 \end{cases}$$

Se $d_{t+1} = \beta/2$, arredonda-se para o número par mais próximo



Truncamento ou arredondamento?

- Truncamento (+ rápido)

$$\text{fl}(x) = \pm (0, d_1 d_2 \cdots, d_t)_\beta \times \beta^e$$

- Arredondamento

$$\text{fl}(x) = \begin{cases} \pm (0, d_1 d_2 \cdots d_t)_\beta \times \beta^e & \text{se } d_{t+1} < \beta/2 \\ \pm (0, d_1 d_2 \cdots d_t + \beta^{-t})_\beta \times \beta^e & \text{se } d_{t+1} > \beta/2 \end{cases}$$

Se $d_{t+1} = \beta/2$, arredonda-se para o número par mais próximo



Truncamento ou arredondamento?

- Truncamento (+ rápido)

$$\text{fl}(x) = \pm (0, d_1 d_2 \cdots, d_t)_\beta \times \beta^e$$

- Arredondamento (+ preciso)

$$\text{fl}(x) = \begin{cases} \pm (0, d_1 d_2 \cdots d_t)_\beta \times \beta^e & \text{se } d_{t+1} < \beta/2 \\ \pm (0, d_1 d_2 \cdots d_t + \beta^{-t})_\beta \times \beta^e & \text{se } d_{t+1} > \beta/2 \end{cases}$$

Se $d_{t+1} = \beta/2$, arredonda-se para o número par mais próximo



Erros numa representação em ponto flutuante

- Erro absoluto

$$|x - fl(x)| \leq \begin{cases} 2\epsilon_M \beta^e, & \text{truncamento} \\ \epsilon_M \beta^e, & \text{arredondamento} \end{cases}$$

- Erro relativo (para $x \neq 0$)

$$\frac{|x - fl(x)|}{|x|} \leq \begin{cases} 2\epsilon_M, & \text{truncamento} \\ \epsilon_M, & \text{arredondamento} \end{cases}$$

- $\epsilon_M = \frac{1}{2} \beta^{-t}$ é denominado *precisão da máquina*
(*menor número representável tal que $1 + \epsilon_M \neq 1$*)



IEEE 754 - 2008 (o padrão dos computadores modernos)

Padrão técnico para cálculo de ponto flutuante, estabelecido em 1985 pelo Instituto de Engenheiros Elétricos e Eletrônicos (IEEE)

$$fl(x) = \underbrace{(-1)^s}_{\text{sinal}} \underbrace{(1, b_1 b_2 \cdots, b_{t-1})_2}_{\text{mantissa}} \times 2^{e+M}$$

onde $s = 0$ se $x \geq 0$, $s = 1$ se $x < 0$ e $b_j \in \{0, 1\}$

tipo	sinal	expoente	mantissa	bits	β	t	m	M	ϵ_M
half	1	5	10	16	2	11	-14	15	$\approx 10^{-03}$
single	1	8	23	32	2	24	-126	127	$\approx 10^{-07}$
double	1	11	52	64	2	53	-1022	1023	$\approx 10^{-16}$



IEEE Standard for Floating-Point Arithmetic, in IEEE Std 754-2008, Aug. 29 2008.

<https://doi.org/10.1109/IEEESTD.2008.4610935>

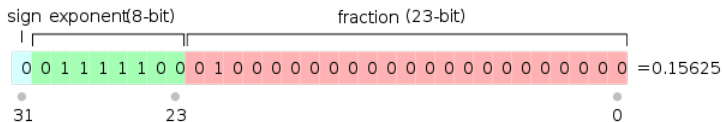


IEEE 754 - 2008 (o padrão dos computadores modernos)

Considere o número $x = (0,15625)_{10} = (1,01)_2 \times 2^{-3}$.

- **precisão simples (single)**

$$fl(x) = (-1)^0 (1,0100000000000000000000000000)_2 \times 2^{-3+127}$$

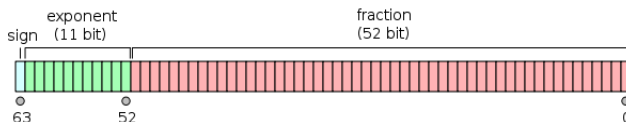


* Figura obtida em https://en.wikipedia.org/wiki/Floating-point_arithmetic



IEEE 754 - 2008 (o padrão dos computadores modernos)

- precisão dupla (double)



- valores especiais

- Zero — e = todos 0, mantissa = todos 0, s arbitrário
- $+\infty$ — e = todos 1, mantissa = todos 0, s = 0
- $-\infty$ — e = todos 1, mantissa = todos 0, s = 1
- NaN — e = todos 1, mantissa $\neq 0$ (not a number)

* Figura obtida em https://en.wikipedia.org/wiki/Double-precision_floating-point_format



Para pensar em casa ...

Exercício teórico:

Considere o sistema de ponto flutuante $(\beta, t, m, M) = (10, 3, -3, 4)$, onde a parcela que não pode ser incorporada à mantissa é truncada.

- Qual a região de *overflow* desse sistema?
- Qual a região de *underflow* desse sistema?
- Qual a representação de $\alpha = 10^{-5}$ nesse sistema?

Exercício computacional:

Implemente no GNU Octave um algoritmo para calcular a *precisão da máquina* ϵ_M .



Como citar esse material?

A. Cunha, *Sistema de Ponto Flutuante*, Universidade do Estado do Rio de Janeiro – UERJ, 2020.

Essas notas de aula podem ser compartilhadas nos termos da licença Creative Commons BY-NC-ND 3.0, com propósitos exclusivamente educacionais.

