Noções Básicas de Ponto Flutuante

Prof. Afonso Paiva

Departamento de Matemática Aplicada e Estatística
Instituto de Ciências Matemáticas e de Computação
USP – São Carlos

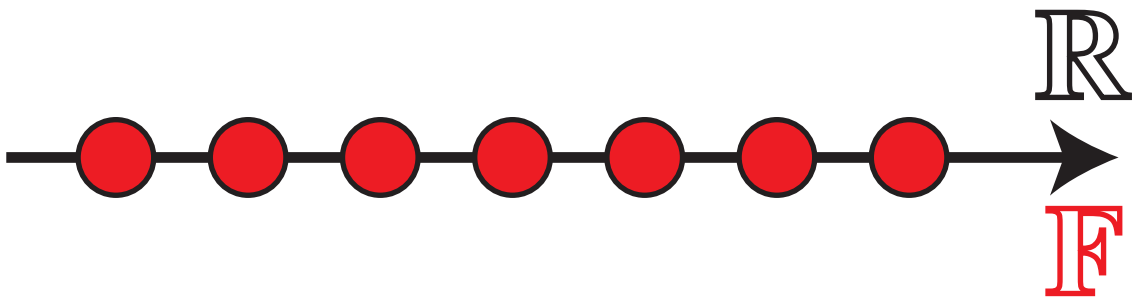
Métodos Numéricos e Computacionais I – SME0305

Motivação

Como representar os números reais no computador?

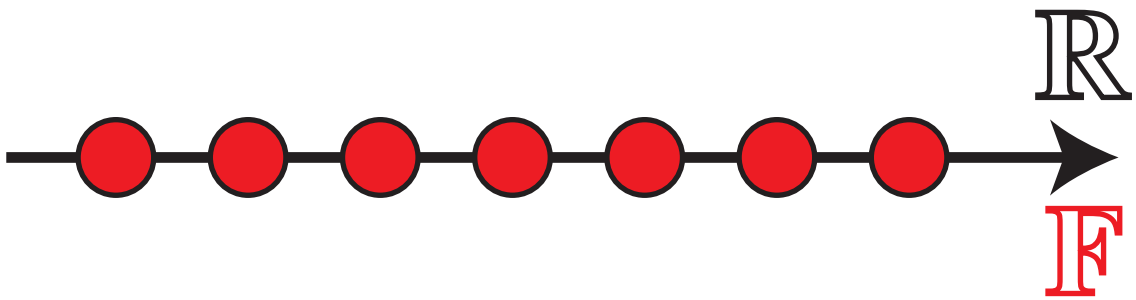
Motivação

Como representar os números reais no computador?



Motivação

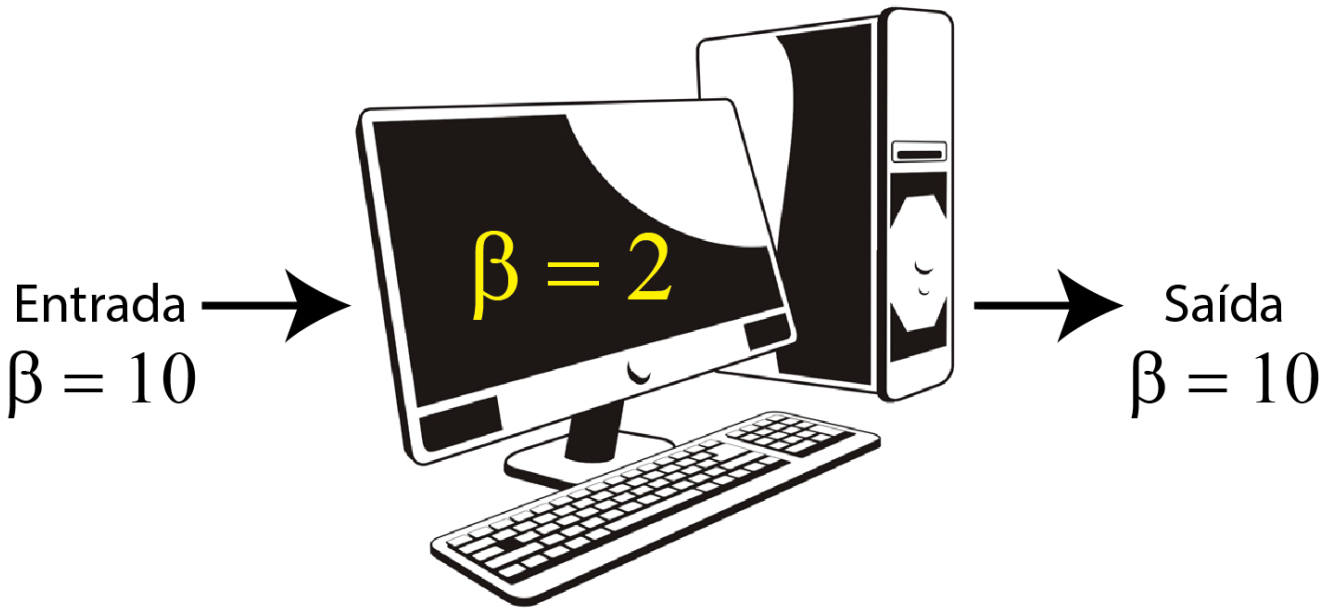
Como representar os números reais no computador?



Definição (ponto flutuante)

O conjunto $\mathbb{F} = \{$ *representação de* $x \in \mathbb{R}$ *no computador* $\}$. *Cada número* $\overline{x} \in \mathbb{F}$ *é chamado de ponto flutuante.*

Conversão entre números binários e decimais



Conversão entre números binários e decimais

Definição (base numérica)

Dado $N \in \mathbb{Z}$, ele pode ser escrito em uma base β como

$$N = (d_k\, d_{k-1} \ldots d_1\, d_0)_\beta = d_k \times \beta^k + d_{k-1} \times \beta^{k-1} + \cdots + d_1 \times \beta^1 + d_0 \times \beta^0\,,$$

onde $0 \leq d_i \leq (\beta - 1)$, para $i = 0, \ldots, k$.

Conversão entre números binários e decimais

Definição (base numérica)

Dado $N \in \mathbb{Z}$, ele pode ser escrito em uma base β como

$$N = (d_k\, d_{k-1} \ldots d_1\, d_0)_\beta = d_k \times \beta^k + d_{k-1} \times \beta^{k-1} + \cdots + d_1 \times \beta^1 + d_0 \times \beta^0\,,$$

onde $0 \leq d_i \leq (\beta - 1)$, para $i = 0, \ldots, k$.

Exemplo (conversão binário $\Rightarrow$ decimal)

$$(10011)_2 =$$

Conversão entre números binários e decimais

Definição (base numérica)

Dado $N \in \mathbb{Z}$, ele pode ser escrito em uma base β como

$$N = (d_k\, d_{k-1} \ldots d_1\, d_0)_\beta = d_k \times \beta^k + d_{k-1} \times \beta^{k-1} + \cdots + d_1 \times \beta^1 + d_0 \times \beta^0\,,$$

onde $0 \leq d_i \leq (\beta - 1)$, para $i = 0, \ldots, k$.

Exemplo (conversão binário $\Rightarrow$ decimal)

$$(10011)_2 = 1 \times 2^4 + 1 \times 2^1 + 1 \times 2^0 = 19 = 1 \times 10^1 + 9 \times 10^0 = (19)_{10}$$

Conversão decimal ⇒ binário

Seja $N \in \mathbb{Z}$, escrito na base $\beta = 10$. Desejamos representar N como:

$$N = (d_k\, d_{k-1} \ldots d_1\, d_0)_2 = d_k \times 2^k + d_{k-1} \times 2^{k-1} + \cdots + d_1 \times 2^1 + d_0 \times 2^0$$

Conversão decimal ⇒ binário

Seja $N \in \mathbb{Z}$, escrito na base $\beta = 10$. Desejamos representar N como:

$$N = (d_k\, d_{k-1} \ldots d_1\, d_0)_2 = d_k \times 2^k + d_{k-1} \times 2^{k-1} + \cdots + d_1 \times 2^1 + d_0 \times 2^0$$

Logo

$$N = 2 \times \left(d_k \times 2^{k-1} + d_{k-1} \times 2^{k-2} + \cdots + d_2 \times 2^1 + d_1 \right) + d_0$$

Conversão decimal ⇒ binário

Seja $N \in \mathbb{Z}$, escrito na base $\beta = 10$. Desejamos representar N como:

$$N = (d_k\, d_{k-1} \ldots d_1\, d_0)_2 = d_k \times 2^k + d_{k-1} \times 2^{k-1} + \cdots + d_1 \times 2^1 + d_0 \times 2^0$$

Logo

$$N = 2 \times \left(d_k \times 2^{k-1} + d_{k-1} \times 2^{k-2} + \cdots + d_2 \times 2^1 + d_1 \right) + d_0$$

$$= 2 \times \left(2 \times \left(d_k \times 2^{k-2} + d_{k-1} \times 2^{k-3} + \cdots + d_2 \right) + d_1 \right) + d_0$$

Conversão decimal ⇒ binário

Seja $N \in \mathbb{Z}$, escrito na base $\beta = 10$. Desejamos representar N como:

$$N = (d_k\, d_{k-1} \ldots d_1\, d_0)_2 = d_k \times 2^k + d_{k-1} \times 2^{k-1} + \cdots + d_1 \times 2^1 + d_0 \times 2^0$$

Logo

$$N = 2 \times \left(d_k \times 2^{k-1} + d_{k-1} \times 2^{k-2} + \cdots + d_2 \times 2^1 + d_1 \right) + d_0$$

$$= 2 \times \left(2 \times \left(d_k \times 2^{k-2} + d_{k-1} \times 2^{k-3} + \cdots + d_2 \right) + d_1 \right) + d_0$$

Assim por diante até o quociente for igual a zero.

Conversão decimal ⇒ binário

Seja $N \in \mathbb{Z}$, escrito na base $\beta = 10$. Desejamos representar N como:

$$N = (d_k\, d_{k-1} \ldots d_1\, d_0)_2 = d_k \times 2^k + d_{k-1} \times 2^{k-1} + \cdots + d_1 \times 2^1 + d_0 \times 2^0$$

Logo

$$N = 2 \times \left(d_k \times 2^{k-1} + d_{k-1} \times 2^{k-2} + \cdots + d_2 \times 2^1 + d_1 \right) + d_0$$

$$= 2 \times \left(2 \times \left(d_k \times 2^{k-2} + d_{k-1} \times 2^{k-3} + \cdots + d_2 \right) + d_1 \right) + d_0$$

Assim por diante até o quociente for igual a zero.

Exemplo: Escreva 29 na base $\beta = 2$ (no quadro).

Conversão decimal ⇒ binário

Seja $N \in \mathbb{Z}$, escrito na base $\beta = 10$. Desejamos representar N como:

$$N = (d_k\, d_{k-1} \ldots d_1\, d_0)_2 = d_k \times 2^k + d_{k-1} \times 2^{k-1} + \cdots + d_1 \times 2^1 + d_0 \times 2^0$$

Logo

$$N = 2 \times \left(d_k \times 2^{k-1} + d_{k-1} \times 2^{k-2} + \cdots + d_2 \times 2^1 + d_1 \right) + d_0$$

$$= 2 \times \left(2 \times \left(d_k \times 2^{k-2} + d_{k-1} \times 2^{k-3} + \cdots + d_2 \right) + d_1 \right) + d_0$$

Assim por diante até o quociente for igual a zero.

Exemplo: Escreva 29 na base $\beta = 2$ (no quadro).

Solução: $29 = (11101)_2$.

Algoritmo para conversão decimal para binária de números inteiros

Entrada: $N \in \mathbb{Z}$
Saída: $N = (d_k\, d_{k-1} \ldots d_1\, d_0)_2$

Algoritmo para conversão decimal para binária de números inteiros

Entrada: $N \in \mathbb{Z}$
Saída: $N = (d_k\, d_{k-1} \ldots d_1\, d_0)_2$

$k = 0$;
Calcule R e Q tal que $N = 2Q + R$;
$d_k = R$;

Algoritmo para conversão decimal para binária de números inteiros

Entrada: $N \in \mathbb{Z}$
Saída: $N = (d_k\, d_{k-1} \ldots d_1\, d_0)_2$

$k = 0$;
Calcule R e Q tal que $N = 2Q + R$;
$d_k = R$;

Enquanto $Q \neq 0$ faça
    $k = k + 1$;
    $N = Q$;
    Calcule R e Q tal que $N = 2Q + R$;
    $d_k = R$;
Fim do Enquanto

Conversão decimal para binária de números fracionários

Vamos supor que N é um número fracionário no intervalo $(0,1)$. Assim,

$$N = (0.d_1\, d_2 \ldots d_k)_2 = d_1 \times 2^{-1} + d_2 \times 2^{-2} + \cdots + d_k \times 2^{-k}$$

Conversão decimal para binária de números fracionários

Vamos supor que N é um número fracionário no intervalo $(0,1)$. Assim,

$$N = (0.d_1\, d_2 \ldots d_k)_2 = d_1 \times 2^{-1} + d_2 \times 2^{-2} + \cdots + d_k \times 2^{-k}$$

Multiplicando por 2 a equação acima, temos:

$$2 \times N = d_1 + \underbrace{d_2 \times 2^{-1} + \cdots + d_k \times 2^{-k+1}}_{N_1}$$

Conversão decimal para binária de números fracionários

Vamos supor que N é um número fracionário no intervalo $(0,1)$. Assim,

$$N = (0.d_1\, d_2 \ldots d_k)_2 = d_1 \times 2^{-1} + d_2 \times 2^{-2} + \cdots + d_k \times 2^{-k}$$

Multiplicando por 2 a equação acima, temos:

$$2 \times N = d_1 + \underbrace{d_2 \times 2^{-1} + \cdots + d_k \times 2^{-k+1}}_{N_1}$$

Portanto, d_1 é parte inteira de $(2 \times N)$ e a parte fracionária é N_1.

Conversão decimal para binária de números fracionários

Vamos supor que N é um número fracionário no intervalo $(0,1)$. Assim,

$$N = (0.d_1\, d_2 \ldots d_k)_2 = d_1 \times 2^{-1} + d_2 \times 2^{-2} + \cdots + d_k \times 2^{-k}$$

Multiplicando por 2 a equação acima, temos:

$$2 \times N = d_1 + \underbrace{d_2 \times 2^{-1} + \cdots + d_k \times 2^{-k+1}}_{N_1}$$

Portanto, d_1 é parte inteira de $(2 \times N)$ e a parte fracionária é N_1. Multiplicando N_1 por 2 em ambos os lados, temos:

$$2 \times N_1 = d_2 + d_3 \times 2^{-1} + \cdots + d_k \times 2^{-k+2}$$

Portanto, d_2 é parte inteira de $(2 \times N_1)$. Assim por diante até a parte fracionária for zero.

Conversão decimal para binária de números fracionários

Exemplo: Represente 0.125 na base $\beta = 2$ (no quadro).

Conversão decimal para binária de números fracionários

Exemplo: Represente 0.125 na base $\beta = 2$ (no quadro).

Solução: $0.125 = (0.001)_2$.

Conversão decimal para binária de números fracionários

Exemplo: Represente 0.125 na base $\beta = 2$ (no quadro).

Solução: $0.125 = (0.001)_2$.

Exemplo: Represente 3.8 na base $\beta = 2$ (no quadro).

Conversão decimal para binária de números fracionários

Exemplo: Represente 0.125 na base $\beta = 2$ (no quadro).

Solução: $0.125 = (0.001)_2$.

Exemplo: Represente 3.8 na base $\beta = 2$ (no quadro).

Solução: $3.8 = (11.110011001100\ldots)_2 = (11.\overline{1100})_2$.

Obs.: Note que a representação binária de um número decimal fracionário pode ser infinita.

Conversão decimal para binária de números fracionários

Exemplo: Represente 0.125 na base $\beta = 2$ (no quadro).

Solução: $0.125 = (0.001)_2$.

Exemplo: Represente 3.8 na base $\beta = 2$ (no quadro).

Solução: $3.8 = (11.110011001100\ldots)_2 = (11.\overline{1100})_2$.

Obs.: Note que a representação binária de um número decimal fracionário pode ser infinita.

Exercício

Faça um algoritmo para representar um número decimal fracionário em um número binário, e vice-versa.

Sistema $\mathbb{F}(\beta, t, m, M)$

Seja $x \in \mathbb{R}$, podemos representar x em $\mathbb{F}$ numa base β da seguinte forma:

$$\overline{x} = \pm \underbrace{(0.d_1\, d_2 \ldots d_t)_\beta}_{\textbf{mantissa}} \times \beta^e \quad ,$$

com $1 \leq d_1 \leq \beta - 1$ e $0 \leq d_i \leq \beta - 1$, para $i = 2, \ldots, t$.

Sistema $\mathbb{F}(\beta, t, m, M)$

Seja $x \in \mathbb{R}$, podemos representar x em $\mathbb{F}$ numa base β da seguinte forma:

$$\overline{x} = \pm \underbrace{(0.d_1\, d_2 \ldots d_t)_\beta}_{\textbf{mantissa}} \times \beta^e,$$

com $1 \leq d_1 \leq \beta - 1$ e $0 \leq d_i \leq \beta - 1$, para $i = 2, \ldots, t$.

- O número inteiro t representa a quantidade de dígitos (significativos) na mantissa;

Sistema $\mathbb{F}(\beta, t, m, M)$

Seja $x \in \mathbb{R}$, podemos representar x em $\mathbb{F}$ numa base β da seguinte forma:

$$\overline{x} = \pm \underbrace{(0.d_1\, d_2 \ldots d_t)_\beta}_{\textbf{mantissa}} \times \beta^e,$$

com $1 \leq d_1 \leq \beta - 1$ e $0 \leq d_i \leq \beta - 1$, para $i = 2, \ldots, t$.

- O número inteiro t representa a quantidade de dígitos (significativos) na mantissa;
- O valor e é o **expoente**, número inteiro definido no intervalo $[m, M]$.

Sistema $\mathbb{F}(\beta, t, m, M)$

Exemplo: Qual é a quantidade de números não nulos que podemos representar no sistema $\mathbb{F}(2, 3, -2, 1)$? (no quadro)

Sistema $\mathbb{F}(\beta, t, m, M)$

Exemplo: Qual é a quantidade de números não nulos que podemos representar no sistema $\mathbb{F}(2, 3, -2, 1)$? (no quadro)

Solução: $2 \times 1 \times 2 \times 2 \times 4 = 32$ números.

Sistema $\mathbb{F}(\beta, t, m, M)$

Exemplo: Qual é a quantidade de números não nulos que podemos representar no sistema $\mathbb{F}(2, 3, -2, 1)$? (no quadro)

Solução: $2 \times 1 \times 2 \times 2 \times 4 = 32$ números.

Exemplo: Seja $\mathbb{F}(10, 3, -5, 5)$. Quais são o menor e maior número em valor absoluto que podemos representar nesse sistema? (no quadro).

Sistema $\mathbb{F}(\beta, t, m, M)$

Exemplo: Qual é a quantidade de números não nulos que podemos representar no sistema $\mathbb{F}(2, 3, -2, 1)$? (no quadro)

Solução: $2 \times 1 \times 2 \times 2 \times 4 = 32$ números.

Exemplo: Seja $\mathbb{F}(10, 3, -5, 5)$. Quais são o menor e maior número em valor absoluto que podemos representar nesse sistema? (no quadro).

Solução: O menor número $l = 10^{-6}$ e o maior $u = 99900$.

Sistema $\mathbb{F}(\beta, t, m, M)$

Exemplo: Qual é a quantidade de números não nulos que podemos representar no sistema $\mathbb{F}(2, 3, -2, 1)$? (no quadro)

Solução: $2 \times 1 \times 2 \times 2 \times 4 = 32$ números.

Exemplo: Seja $\mathbb{F}(10, 3, -5, 5)$. Quais são o menor e maior número em valor absoluto que podemos representar nesse sistema? (no quadro).

Solução: O menor número $l = 10^{-6}$ e o maior $u = 99900$.

Dessa forma os números $\overline{x} \in \mathbb{F}$ estaria limitados a:

$$l \leq |\overline{x}| \leq u$$

Sistema $\mathbb{F}(\beta, t, m, M)$

Exemplo: Qual é a quantidade de números não nulos que podemos representar no sistema $\mathbb{F}(2, 3, -2, 1)$? (no quadro)

Solução: $2 \times 1 \times 2 \times 2 \times 4 = 32$ números.

Exemplo: Seja $\mathbb{F}(10, 3, -5, 5)$. Quais são o menor e maior número em valor absoluto que podemos representar nesse sistema? (no quadro).

Solução: O menor número $l = 10^{-6}$ e o maior $u = 99900$.

Dessa forma os números $\overline{x} \in \mathbb{F}$ estaria limitados a:

$$l \leq |\overline{x}| \leq u$$

- se $|\overline{x}| < l$: erro de **underflow**

Sistema $\mathbb{F}(\beta, t, m, M)$

Exemplo: Qual é a quantidade de números não nulos que podemos representar no sistema $\mathbb{F}(2, 3, -2, 1)$? (no quadro)

Solução: $2 \times 1 \times 2 \times 2 \times 4 = 32$ números.

Exemplo: Seja $\mathbb{F}(10, 3, -5, 5)$. Quais são o menor e maior número em valor absoluto que podemos representar nesse sistema? (no quadro).

Solução: O menor número $l = 10^{-6}$ e o maior $u = 99900$.

Dessa forma os números $\overline{x} \in \mathbb{F}$ estaria limitados a:

$$l \leq |\overline{x}| \leq u$$

- se $|\overline{x}| < l$: erro de **underflow**
- se $|\overline{x}| > u$: erro de **overflow**

Padrão nos Computadores – IEEE 754-2008

	armazenamento				representação				
tipo	sinal	expoente	mantisa	bits	β	t	m	M	b
half	1	5	10	16	2	11	-14	15	15
single	1	8	23	32	2	24	-126	127	127
double	1	11	52	64	2	53	-1022	1023	1023

Padrão nos Computadores – IEEE 754-2008

	armazenamento				representação				
tipo	sinal	expoente	mantisa	bits	β	t	m	M	b
half	1	5	10	16	2	11	-14	15	15
single	1	8	23	32	2	24	-126	127	127
double	1	11	52	64	2	53	-1022	1023	1023

A representação **normalizada** de um número em ponto flutuante nesse padrão é feita da seguinte forma:

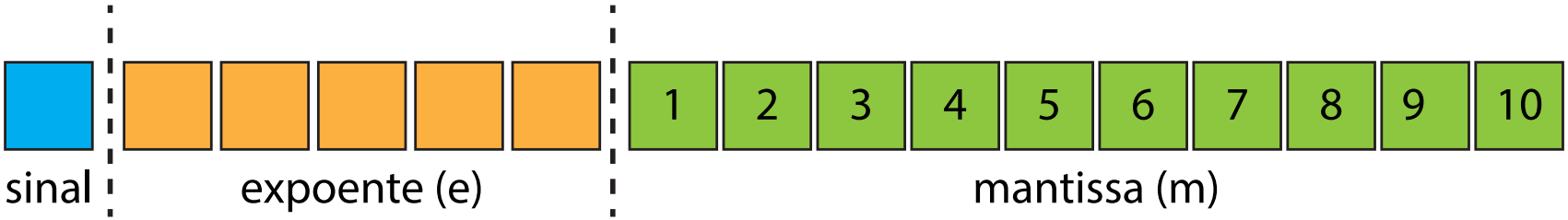
$$\overline{x} = (-1)^s \times (1.d_1 \ldots d_t)_2 \times 2^{e-b}$$

Padrão nos Computadores – IEEE 754-2008

Vamos analisar o padrão half (16 bits)

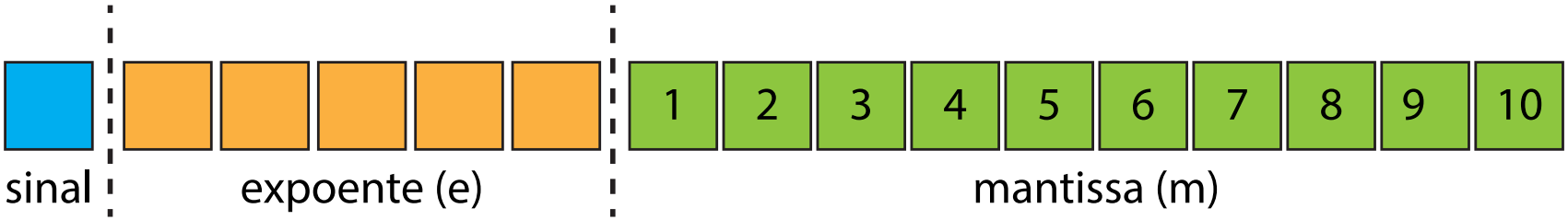
Padrão nos Computadores – IEEE 754-2008

Vamos analisar o padrão half (16 bits)



Padrão nos Computadores – IEEE 754-2008

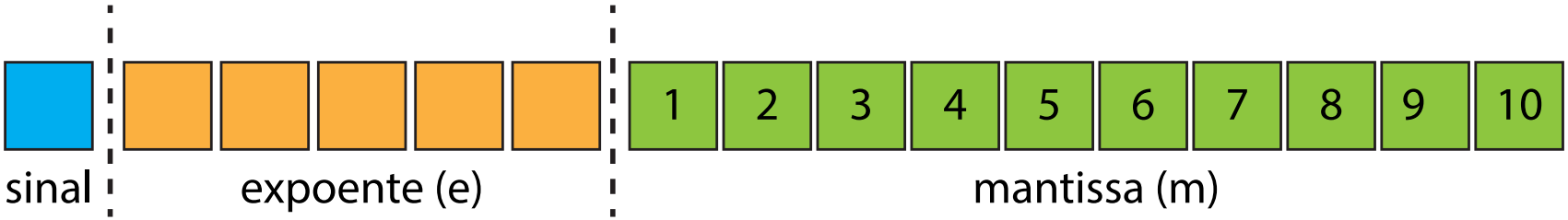
Vamos analisar o padrão half (16 bits)



Valores especiais:

Padrão nos Computadores – IEEE 754-2008

Vamos analisar o padrão half (16 bits)

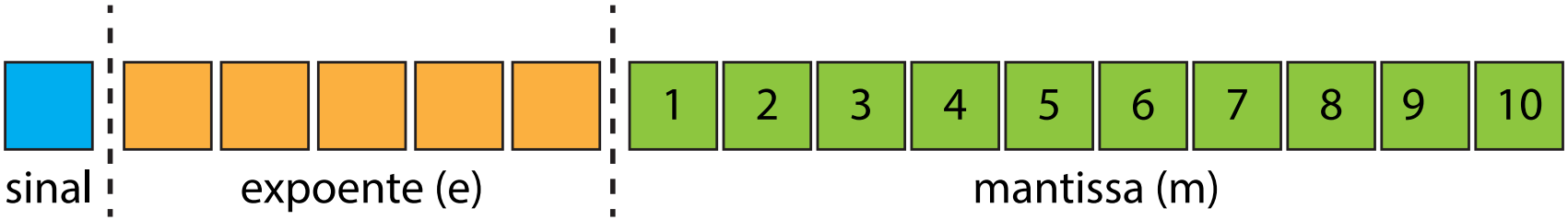


Valores especiais:

- $0 = 0\ 00000\ 0000000000$

Padrão nos Computadores – IEEE 754-2008

Vamos analisar o padrão half (16 bits)

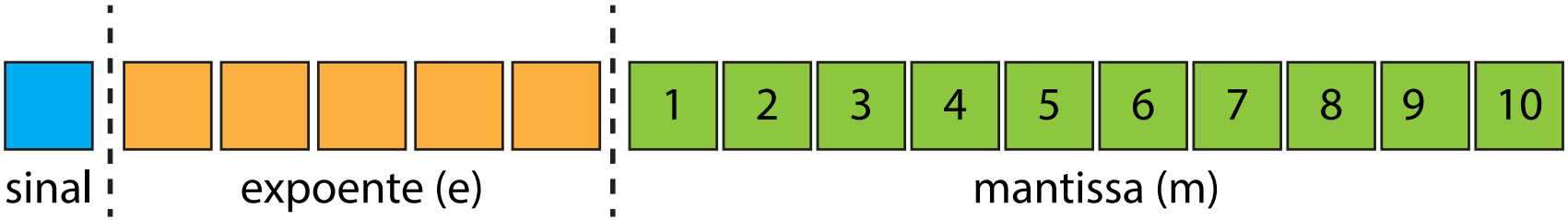


Valores especiais:

- $0 = 0\ 00000\ 0000000000$
- $\text{inf} = 0\ 11111\ 0000000000$

Padrão nos Computadores – IEEE 754-2008

Vamos analisar o padrão half (16 bits)

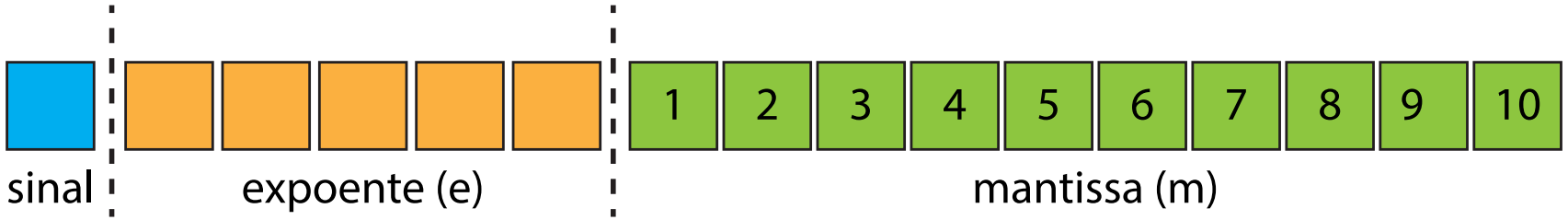


Valores especiais:

- $0 = 0\ 00000\ 0000000000$
- $\text{inf} = 0\ 11111\ 0000000000$
- $\text{-inf} = 1\ 11111\ 0000000000$

Padrão nos Computadores – IEEE 754-2008

Vamos analisar o padrão half (16 bits)



Valores especiais:

- $0 = 0\ 00000\ 0000000000$
- inf $= 0\ 11111\ 0000000000$
- -inf $= 1\ 11111\ 0000000000$

Dessa forma, os valores disponíveis para o expoente e variam de $1 = (00001)_2$ a $30 = (11110)_2$.

Padrão nos Computadores – IEEE 754-2008

Exemplo: Qual número é $\bar{x}_1 = 0\ 10100\ 1010010000$? (no quadro)

Padrão nos Computadores – IEEE 754-2008

Exemplo: Qual número é $\overline{x}_1 = 0\ 10100\ 1010010000$? (no quadro)

Solução: $\overline{x}_1 = 52.5$.

Padrão nos Computadores – IEEE 754-2008

Exemplo: Qual número é $\overline{x}_1 = 0\ 10100\ 1010010000$? (no quadro)

Solução: $\overline{x}_1 = 52.5$.

Qual seria o próximo número após $\overline{x}_1$ a ser representado nesse formato?

Padrão nos Computadores – IEEE 754-2008

Exemplo: Qual número é $\bar{x}_1 = 0\ 10100\ 1010010000$? (no quadro)

Solução: $\bar{x}_1 = 52.5$.

Qual seria o próximo número após $\bar{x}_1$ a ser representado nesse formato?

$$\bar{x}_2 = 0\ 10100\ 1010010001 = 52.53125$$

Padrão nos Computadores – IEEE 754-2008

Exemplo: Qual número é $\bar{x}_1 = 0\ 10100\ 1010010000$? (no quadro)

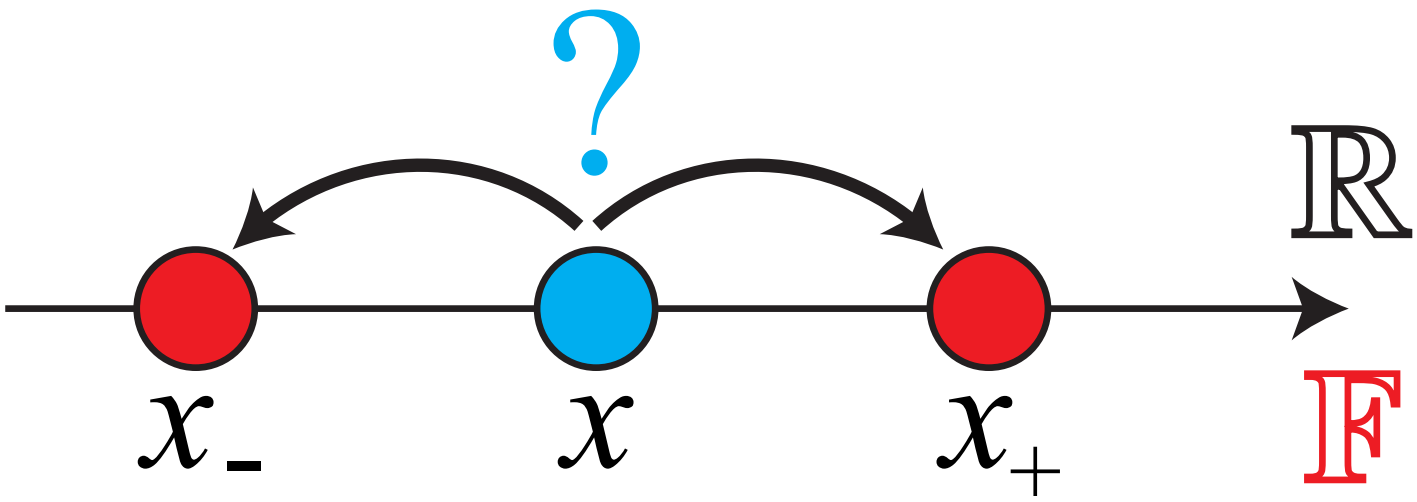
Solução: $\bar{x}_1 = 52.5$.

Qual seria o próximo número após $\bar{x}_1$ a ser representado nesse formato?

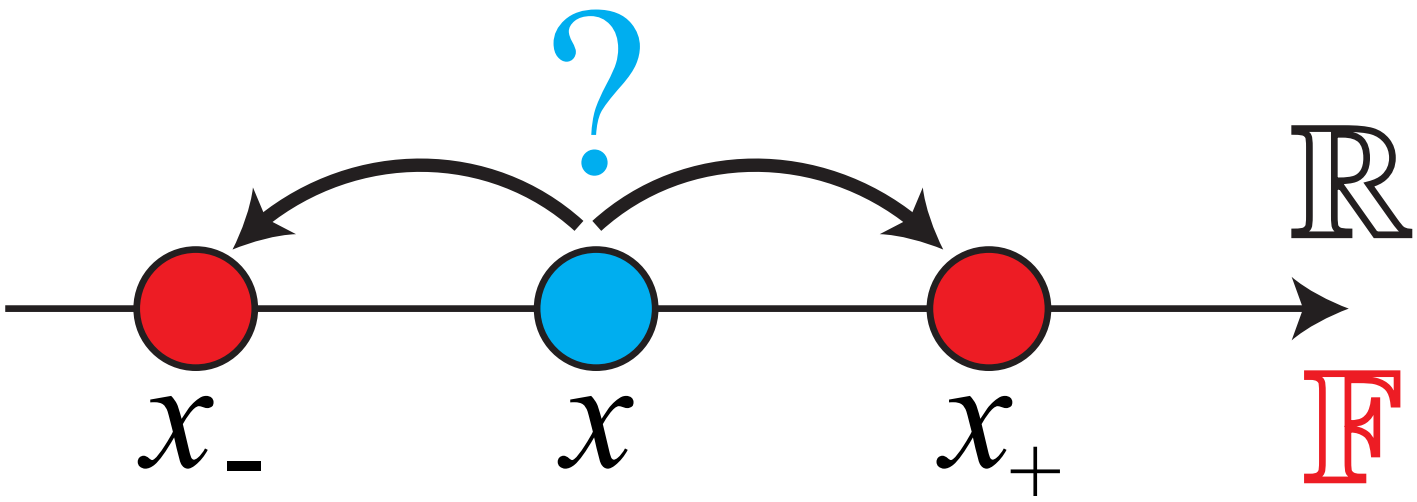
$$\bar{x}_2 = 0\ 10100\ 1010010001 = 52.53125$$

Como representar $x = 52.51$ nesse sistema?

Arredondamento de Ponto Flutuante



Arredondamento de Ponto Flutuante



Solução: Basta escolher $\overline{x} \in \mathbb{F}$ mais perto de x, isto é, $dist(x, \overline{x}) = |x - \overline{x}|$ seja mínima.

Arredondamento de Ponto Flutuante

Sejam $\beta = 10$ e $t = 3$. Dado $x = 37.29$. Logo,

Arredondamento de Ponto Flutuante

Sejam $\beta = 10$ e $t = 3$. Dado $x = 37.29$. Logo,

$$x = \underbrace{0.372}_{f_x} \times 10^2 + \underbrace{0.9}_{g_x} \times 10^{-1}$$

Arredondamento de Ponto Flutuante

Sejam $\beta = 10$ e $t = 3$. Dado $x = 37.29$. Logo,

$$x = \underbrace{0.372}_{f_x} \times 10^2 + \underbrace{0.9}_{g_x} \times 10^{-1}$$

Critério de Arredondamento: ponto flutuante mais próximo

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

Arredondamento de Ponto Flutuante

Sejam $\beta = 10$ e $t = 3$. Dado $x = 37.29$. Logo,

$$x = \underbrace{0.372}_{f_x} \times 10^2 + \underbrace{0.9}_{g_x} \times 10^{-1}$$

Critério de Arredondamento: ponto flutuante mais próximo

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

Usando o critério acima, onde $e = 2$, temos:

$$x = 0.372 \times 10^2 + 10^{2-3} = 0.372 \times 10^2 + 0.001 \times 10^2 = 0.373 \times 10^2$$

Arredondamento de Ponto Flutuante

Sejam $\beta = 10$ e $t = 3$. Dado $x = 37.29$. Logo,

$$x = \underbrace{0.372}_{f_x} \times 10^2 + \underbrace{0.9}_{g_x} \times 10^{-1}$$

Critério de Arredondamento: ponto flutuante mais próximo

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

Usando o critério acima, onde $e = 2$, temos:

$$x = 0.372 \times 10^2 + 10^{2-3} = 0.372 \times 10^2 + 0.001 \times 10^2 = 0.373 \times 10^2$$

Observação: **truncamento** de x é quando sempre fazemos $g_x = 0$, isto é, $\overline{x} = f_x \times 10^e$. Nesse caso o arredondamento é feito na direção de 0.

Erro de Arredondamento

- **Erro absoluto:** $EA_x = |x - \overline{x}|$
- **Erro relativo:** $ER_x = |x - \overline{x}| / |x|$

Erro de Arredondamento

- **Erro absoluto:** $EA_x = |x - \bar{x}|$
- **Erro relativo:** $ER_x = |x - \bar{x}| / |x|$

Exemplo: Calcule o erro absoluto e relativo quando aproximamos $x = 37.29$ em $\bar{x} = 37.3$

Erro de Arredondamento

- **Erro absoluto:** $EA_x = |x - \bar{x}|$
- **Erro relativo:** $ER_x = |x - \bar{x}| / |x|$

Exemplo: Calcule o erro absoluto e relativo quando aproximamos $x = 37.29$ em $\bar{x} = 37.3$

Solução: $EA_x = |x - \bar{x}| = |37.29 - 37.3| = 0.01$ e
$ER_x = EA_x / |x| = 0.01 / |37.29| \approx 0.27 \times 10{-}3$

Erro de Arredondamento

- **Erro absoluto:** $EA_x = |x - \bar{x}|$
- **Erro relativo:** $ER_x = |x - \bar{x}| / |x|$

Exemplo: Calcule o erro absoluto e relativo quando aproximamos $x = 37.29$ em $\bar{x} = 37.3$

Solução: $EA_x = |x - \bar{x}| = |37.29 - 37.3| = 0.01$ e
$ER_x = EA_x / |x| = 0.01 / |37.29| \approx 0.27 \times 10{-}3$

Definição (precisão de máquina)

A precisão de máquina é a distância entre 1 e o próximo ponto flutuante maior do que 1 e é dada por $\epsilon_M = \beta^{1-t}$. O tal `eps` *do MATLAB.*

Operações Aritméticas com Ponto Flutuante

Sejam $\overline{x}, \overline{y} \in \mathbb{F}$. As operações com ponto flutuante são definidas da seguinte maneira:

1. $\overline{x} \oplus \overline{y} = \overline{\overline{x} + \overline{y}}$
2. $\overline{x} \ominus \overline{y} = \overline{\overline{x} - \overline{y}}$
3. $\overline{x} \otimes \overline{y} = \overline{\overline{x} \times \overline{y}}$
4. $\overline{x} \underline{\div} \overline{y} = \overline{\overline{x} \div \overline{y}}$

No final de cada operação faz **arredondamento**.

Operações Aritméticas com Ponto Flutuante

Sejam $\overline{x}, \overline{y} \in \mathbb{F}$. As operações com ponto flutuante são definidas da seguinte maneira:

1. $\overline{x} \oplus \overline{y} = \overline{\overline{x} + \overline{y}}$
2. $\overline{x} \ominus \overline{y} = \overline{\overline{x} - \overline{y}}$
3. $\overline{x} \otimes \overline{y} = \overline{\overline{x} \times \overline{y}}$
4. $\overline{x} \underline{\div} \overline{y} = \overline{\overline{x} \div \overline{y}}$

No final de cada operação faz **arredondamento**.

Exemplo: Seja $\beta = 10$ e $t = 3$. Verifique se $(23.4 \oplus 5.18) \oplus 3.05 = 23.4 \oplus (5.18 \oplus 3.05)$ e $3.18 \otimes (5.05 \oplus 11.4) = 3.18 \otimes 5.05 \oplus 3.18 \otimes 11.4$

Operações Aritméticas com Ponto Flutuante

Sejam $\overline{x}, \overline{y} \in \mathbb{F}$. As operações com ponto flutuante são definidas da seguinte maneira:

1. $\overline{x} \oplus \overline{y} = \overline{\overline{x} + \overline{y}}$
2. $\overline{x} \ominus \overline{y} = \overline{\overline{x} - \overline{y}}$
3. $\overline{x} \otimes \overline{y} = \overline{\overline{x} \times \overline{y}}$
4. $\overline{x} \underline{\div} \overline{y} = \overline{\overline{x} \div \overline{y}}$

No final de cada operação faz **arredondamento**.

Exemplo: Seja $\beta = 10$ e $t = 3$. Verifique se
$(23.4 \oplus 5.18) \oplus 3.05 = 23.4 \oplus (5.18 \oplus 3.05)$ e
$3.18 \otimes (5.05 \oplus 11.4) = 3.18 \otimes 5.05 \oplus 3.18 \otimes 11.4$

Solução: operações em $\mathbb{F}$ não são associativa e nem distributivas, pois:

$$(23.4 \oplus 5.18) \oplus 3.05 = 31.7 \neq 31.6 = 23.4 \oplus (5.18 \oplus 3.05)$$

$$3.18 \otimes (5.05 \oplus 11.4) = 52.5 \neq 52.4 = 3.18 \otimes 5.05 \oplus 3.18 \otimes 11.4$$

Operações Aritméticas com Ponto Flutuante

Exercício:

Considere a função abaixo:

$$f(x) = (x-1)^7 = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1$$

Plote usando o MATLAB e compare os gráficos de $f_1(x) = (x-1)^7$ e $f_2(x) = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1$ usando 100 pontos igualmente espaçados no intervalo $[1 - 2 \times 10^{-8}, 1 + 2 \times 10^{-8}]$. Explique o que acontece com os dois gráficos.