

CI-202 - Métodos Numéricos

Aula 04 - Representação Ponto Flutuante

Professor Murilo V. G. da Silva

Departamento de Informática
Universidade Federal do Paraná

11/09/2024

Mantissa e expoente

- ▶ $0,125 = 0,001_2$
- ▶ $-5,875 = -101,111_2$
- ▶ $41 = 101001_2$

Vamos representar cada um destes números binários na seguinte forma:

- ▶ $x_b = m_b \cdot 2^e$
 - ▶ m_b é a mantissa em binário $0 \leq |m_b| < 1$
 - ▶ e é o expoente em decimal
- ▶ $0,001_2 = 0,1_2 \cdot 2^{-2}$
- ▶ $101,111_2 = 0,101111_2 \cdot 2^3$
- ▶ $101001_2 = 0,101001_2 \cdot 2^6$

Mantissa e expoente

- ▶ $-5,875 = 0,101111_2 \cdot 2^3$
- ▶ $0,125 = 0,1_2 \cdot 2^{-2}$
- ▶ $41 = 0,101001_2 \cdot 2^6$

Representando o número (mantissa e expoente) como uma única sequência de 16 bits:

- ▶ Um primeiro bit 0/1 correspondendo ao sinal da mantissa (+/-)
- ▶ Uma sequência de 10 bits que corresponde aos 10 primeiros bits da representação binária da parte fracionária da mantissa
- ▶ Um bit 0/1 correspondendo sinal do expoente (+/-)
- ▶ Uma sequência de 4 bits que corresponde ao expoente em binário

Mantissa e expoente

- ▶ $-5,875 = 0,101111_2 \cdot 2^3$
- ▶ $0,125 = 0,1_2 \cdot 2^{-2}$
- ▶ $41 = 0,101001_2 \cdot 2^6$

- ▶ $-5,875$
 - ▶ mantissa: 11011110000
 - ▶ expoente: 00011
- ▶ $0,125$
 - ▶ mantissa: 01000000000
 - ▶ expoente: 10010
- ▶ 41
 - ▶ mantissa: 01010010000
 - ▶ expoente: 00110

Observações:

- ▶ 10 bits da mantissa: pode ser necessário completar com zeros à direita
- ▶ 4 bits do expoente: pode ser necessário completar com zero à esquerda
- ▶ Números que precisam de mais de 10 bits para mantissa ou mais de 4 bits para o expoente não são representáveis
 - ▶ No caso da mantissa, pode-se usar truncamento para armazenar aproximação do número

Número ponto flutuante de 16 bits

$$(-1)^{s_1} \cdot 0, m_1 m_2 m_3 m_4 m_5 m_6 m_7 m_8 m_9 m_{10} \cdot 2^{(-1)^{s_2} \cdot e_1 e_2 e_3 e_4}$$

s_1	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	s_2	e_1	e_2	e_3	e_4
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	----------	-------	-------	-------	-------	-------

- ▶ s_1 : sinal do número da mantissa. 0 $\rightarrow$ positivo e 1 $\rightarrow$ negativo
- ▶ $m_1 \dots m_{10}$: mantissa com 10 bits significativos
- ▶ s_2 : sinal do expoente. 0 $\rightarrow$ positivo e 1 $\rightarrow$ negativo
- ▶ $e_1 \dots e_4$: expoente

Número ponto flutuante de 16 bits

- O maior valor possível para 16 bits

s_1	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	s_2	e_1	e_2	e_3	e_4
0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1

- $0,1111111111 \cdot 2^{1111} = 32.736$

Número ponto flutuante de 16 bits

- O menor valor possível para 16 bits

s_1	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	s_2	e_1	e_2	e_3	e_4
1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1

- $-0,1111111111 \cdot 2^{1111} = -32.736$

Número ponto flutuante de 16 bits

- Qual é a representação do valor zero?

s_1	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	s_2	e_1	e_2	e_3	e_4
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- Qual o menor valor possível **maior de que zero**?

s_1	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	s_2	e_1	e_2	e_3	e_4
0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1

- $0,1000000000 \cdot 2^{-1111} = 0,1 \cdot 2^{-15} = 0,000015259$

- Qual o **segundo menor** valor possível maior de que zero?

s_1	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	s_2	e_1	e_2	e_3	e_4
0	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1

- $0,1000000001 \cdot 2^{-15} = 0,000015289$

Exercício

- ▶ Dado um sistema de 8 bits em que:
 - ▶ mantissa tem 4 bits
 - ▶ menor expoente: $-7 = -111_2$
 - ▶ maior expoente: $+7 = +111_2$
- ▶ Teremos: $x = (0, d_1 d_2 d_3 d_4) \cdot 2^e$

s_1	d_1	d_2	d_3	d_4	s_2	e_1	e_2	e_3
-------	-------	-------	-------	-------	-------	-------	-------	-------

- ▶ Represente o valor 13.
- ▶ $13 = 1101_2 = 0,1101 \cdot 2^4 = 0,1101 \cdot 2^{100}$

s_1	d_1	d_2	d_3	d_4	s_2	e_1	e_2	e_3
0	1	1	0	1	0	1	0	0

Exercícios

1. Considere o sistema de 8 bits do exercício anterior
 - ▶ Represente o valor 0_{10} .
 - ▶ Represente o valor 1_{10} .
 - ▶ Represente o valor 15_{10} .
 - ▶ Represente o valor $0,1_{10}$.
2. Sejam m , e respectivamente a mantissa e o exponte obtidos da representação do número $0,1_{10}$ no sistema de 8 bits acima.
Converta esse número binário de volta para decimal.
3. Qual é o menor número representável nesta máquina?
4. Qual é o segundo e o terceiro menor número representável nesta máquina?
5. Qual é o maior e o segundo maior número representável nesta máquina?