Métodos Numéricos 1 (MN1)

Unidade 1: Teoria dos Erros
Parte 2: Aritmética de Ponto Flutuante

Joaquim Bento Cavalcante Neto

joaquimb@lia.ufc.br

Grupo de Computação Gráfica, Realidade Virtual e Animação (CRAb)
Departamento de Computação (DC)
Universidade Federal do Ceará (UFC)

CRAb

UFC

Representação em Ponto Flutuante

- Sistema utilizado por um computador para representar um número real **r**

$$r = m \times \beta^e$$

- Onde em um sistema (β, t, l, u), temos:
  - β ≥ 2 é a base em que a máquina opera
  - *e* é o expoente, número inteiro no intervalo [l,u]
  - $m = \pm 0.d_1 d_2 \ldots d_t$ é a mantissa
  - t é o número de dígitos na mantissa

$$0 \leq d_i \leq (\beta - 1); i = 0, \ldots, t, d_1 \neq 0$$

Representação em Ponto Flutuante: Exemplo 1 (Base decimal)

- Dado o sistema SPF (10, 3, -5, 5) temos:

$$\pm 0.d_1 d_2 d_3 \times 10^e, 0 \leq d_i \leq 9, d_1 \neq 0, e \in [-5, 5]$$

- Onde:
  - β = 10 é a base em que a máquina opera
  - *e* é o expoente, número inteiro no intervalo [-5,5]
  - $m = \pm 0.d_1 d_2 \ldots d_t$ é a mantissa
  - t = 3 é o número de dígitos na mantissa

Representação em Ponto Flutuante: Exemplo 1 (Base geral)

- Dado o sistema SPF (3, 2, -1, 2) temos:

$$\pm 0.d_1 d_2 \times 3^e, 0 \leq d_i \leq 2, d_1 \neq 0, e \in [-1, 2]$$

- Onde:
  - β = 3 é a base em que a máquina opera
  - *e* é o expoente, número inteiro no intervalo [-1,2]
  - $m = \pm 0.d_1 d_2 \ldots d_t$ é a mantissa
  - t = 2 é o número de dígitos na mantissa

Limites em Ponto Flutuante

- Dado um sistema (β, t, l, u), temos:
  - zero = $0.\underbrace{00...0}_{\textbf{t} \text{ vezes}} \times \beta^{l}$
  - menor número positivo, não nulo, exatamente representável: menor mantissa, menor expoente l
    - m = $0.\underbrace{100...0}_{\textbf{t}\text{-1 vezes}} \times \beta^{l}$
  - maior número positivo, não nulo, exatamente representável: maior mantissa, maior expoente u
    - M = $0.\underbrace{\delta\delta..\delta}_{\textbf{t} \text{ vezes}} \times \beta^{u}$, onde $\delta = \beta-1$

Limites em Ponto Flutuante: Exemplo 2 (Base decimal)

- Dado o sistema SPF (10, 3, -5, 5) temos:

$$\pm 0.d_1d_2d_3 \times 10^e, 0 \le d_i \le 9, d_1 \neq 0, e \in [-5, 5]$$

- zero: $0.\underbrace{000}_{\textbf{3}\text{ vezes}} \times 10^{-5}$

- menor número positivo: $m = 0.1\underbrace{00}_{\textbf{2}\text{ vezes}} \times 10^{-5} = 10^{-6}$

- maior número positivo: $M = 0.\underbrace{999}_{\textbf{3}\text{ vezes}} \times 10^{5} = 99900$

Limites em Ponto Flutuante: Exemplo 2 (Base geral)

- Dado o sistema SPF (3, 2, -1, 2) temos:

$$\pm 0.d_1 d_2 \times 3^e, 0 \leq d_i \leq 2, d_1 \neq 0, e \in [-1, 2]$$

- zero: $0.\underbrace{00}_{\textbf{2}\text{ vezes}} \times 3^{-1}$

- menor número positivo: $m = 0.1\underbrace{0}_{\textbf{1}\text{ vez}} \times 3^{-1}$

- maior número positivo: $M = 0.\underbrace{22}_{\textbf{2}\text{ vezes}} \times 3^{2}$

Extensão em Ponto Flutuante

- Dado um sistema (β, t, l, u), temos:
  - número máximo de mantissas positivas possíveis:
    - $mantissas_{+} = (\beta - 1) \times \beta^{t-1}$
  - número máximo de expoentes possíveis:
    - $exp_{possíveis} = u - l + 1$
  - número de elementos positivos representáveis:
    - $NR_{+} = mantissas_{+} \times exp_{possíveis}$
  - número total de elementos representáveis:
    - $NR_{t} = 2 \times NR_{+} + 1$

Extensão em Ponto Flutuante: Exemplo 3 (Base decimal)

- Dado o sistema SPF (10, 3, -5, 5) temos:
  - número máximo de mantissas positivas possíveis:
    - $mantissas_{+} = (\beta - 1) \times \beta^{t-1} = (10 - 1) \times 10^{3-1} = 9 \times 10^{2} = 900$
  - número máximo de expoentes possíveis:
    - $exp_{possíveis} = u - l + 1 = 5 - (-5) + 1 = 11$
  - número de elementos positivos representáveis:
    - $NR_{+} = mantissas_{+} \times exp_{possíveis} = 900 \times 11 = 9900$
  - número total de elementos representáveis:
    - $NR_{t} = 2 \times NR_{+} + 1 = 2 \times 9900 + 1 = 19801$

Extensão em Ponto Flutuante: Exemplo 3 (Base geral)

- Dado o sistema SPF (3, 2, -1, 2) temos:
  - número máximo de mantissas positivas possíveis:
    - $mantissas_{+} = (\beta - 1) \times \beta^{t-1} = (3-1) \times 3^{2-1} = (3-1) \times 3^{1} = 6$
  - número máximo de expoentes possíveis:
    - $exp_{possíveis} = u - l + 1 = 2 - (-1) + 1 = 4$
  - número de elementos positivos representáveis:
    - $NR_{+} = mantissas_{+} \times exp_{possíveis} = 6 \times 4 = 24$
  - número total de elementos representáveis:
    - $NR_{t} = 2 \times NR_{+} + 1 = 2 \times 24 + 1 = 25$

Aproximação em Ponto Flutuante

- Dado um sistema (β, t, l, u), seja o conjunto:

$$G = \{x \in \mathbb{R} \,|\, m \leq |x| \leq M\}$$

- Dado um número real x, três casos ocorrem:
  - x ∈ G e x não é representável:
    - **truncamento** ou **arredondamento**
  - |x| < m: underflow
  - |x| > M: overflow

Aproximação em Ponto Flutuante: Exemplo 4 (Base decimal)

- Dado o sistema SPF (10, 3, -5, 5) temos:

$$\pm 0.d_1 d_2 d_3 \times 10^e, 0 \leq d_i \leq 9, d_1 \neq 0, e \in [-5, 5]$$

- zero: 0.000×10^{-5}
- menor número positivo: $m = 0.100 \times 10^{-5} = 10^{-6}$
- maior número positivo: $M = 0.999 \times 10^5 = 99900$
- representando $x = 235.89$:
    - $x = 0.23589 \times 10^3$, $5 > t = 3$; $0.235 \times 10^3 < x < 0.236 \times 10^3$
    - truncamento: $x = 0.235 \times 10^3$, arredondamento: $x = 0.236 \times 10^3$

Aproximação em Ponto Flutuante: Exemplo 4 (Base geral)

- Dado o sistema SPF (3, 2, -1, 2) temos:

$$\pm 0.d_1 d_2 \times 3^e, 0 \le d_i \le 2, d_1 \neq 0, e \in [-1, 2]$$

- zero: 0.00×3^{-1}
- menor número positivo: $0.10 \times 3^{-1} = (1 \times 3^{-1} + 0 \times 3^{-2}) \times 3^{-1} = \frac{1}{9}$
- maior número positivo: $0.22 \times 3^2 = (2 \times 3^{-1} + 2 \times 3^{-2}) \times 3^2 = 8$
- representando x = 0.708:
  - $x = 0.708 \times 3^0$, $2 > t = 0$; $0.70 \times 3^0 < x < 0.71 \times 3^0$
  - truncamento: $x = 0.70 \times 3^0$, arredondamento: $x = 0.71 \times 3^0$

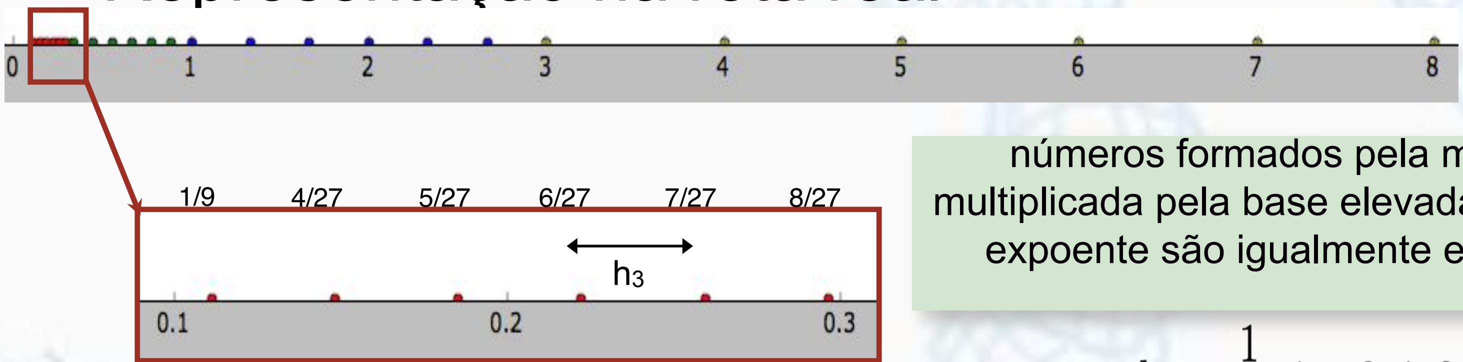
Aproximação em Ponto Flutuante
Exemplo 4 (Base geral)

- Dado o sistema SPF (3, 2, -1, 2) temos:
  - Os elementos representáveis pertencem ao conjunto

$$R = \left\{ x; x \in \left[\frac{1}{9}, 8\right] \cup \left[-8, -\frac{1}{9}\right] \cup \{0\} \right\}$$

- m x 3^{-1}, h_3 = 1/27
- m x 3^{0}, h_2 = 1/9
- m x 3^{1}, h_1 = 1/3
- m x 3^{2}, h_0 = 1

  - Representação na reta real



números formados pela mantissa multiplicada pela base elevada ao mesmo expoente são igualmente espaçados

$$h_i = \frac{1}{3^i}; i = 0, 1, 2, 3$$

Precisão simples e dupla

- simples: 32 bits
  - 1 bit sinal
  - 8 bits expoente
  - 23 bits mantissa
- dupla: 64 bits
  - 1 bit sinal
  - 11 bits expoente
  - 52 bits mantissa