Cálculo Numérico

Aritmética de Ponto Flutuante e Noções de Erro

Representação em Ponto Flutuante

- Um sistema de ponto flutuante pode ser definido como

$$F(\beta, t, L, U)$$

onde

- β é a base do sistema
- t é o número de dígitos da mantissa
- L é o menor valor para o expoente
- U é o maior valor para o expoente

Representação em Ponto Flutuante

A representação de números reais mais utilizada em máquinas é a do ponto flutuante . Esse número tem três partes: o sinal, a parte fracionária (mantissa) e o expoente,

$$m = \pm\ ,d_1d_2d_3... d_t \times \beta^e$$

sendo

d_i's : dígitos da parte fracionária, $d_1 \neq 0$, $0 \leq d_i \leq \beta$-1
β: base (em geral 2, 10 ou 16),
t: no de dígitos na mantissa.
e: expoente inteiro.

Equivalente à: $(0{,}0001100110011001100110011.....)_2$

Ex.

x=34,2 (decimal); β=10; t=4	x=0,1 (decimal) ; β=2; t=9
$x=0{,}3420 \times 10^2$	$x=0{,}110011001 \times 2^{-3}$

Exemplo

- **Exemplo 7**
  Considerando o sistema $F(10, 3, -5, 5)$.
  Represente o número $1{,}23$ nesse sistema.

$1{,}23 = 0{,}123 \times 10^1$

Exemplo

- **Exemplo 8**
  Considerando o sistema $F(10, 3, -5, 5)$.
  Qual o menor número em valor absoluto que esse sistema pode representar?

$m = 0{,}100 \times 10^{-5}$

Exemplo

- **Exemplo 9**
  Considerando o sistema $F(10, 3, -5, 5)$.
  Qual o maior número que esse sistema pode representar?

$M = 0{,}999 \times 10^5$

Representação em Ponto Flutuante

- Sejam m e M, respectivamente, o menor e o maior valores absolutos representáveis no sistema $F(\beta, t, L, U)$
- Dado um número x, então
  - Se $m \leq |x| \leq M$, então o número pode ser representado no sistema
    - Os valores podem ser arredondados ou truncados
    - Truncamento: dígitos $d_{t+1}d_{t+2}\ldots$ são removidos
    - Arredondamento: na base 10, além de remover os dígitos $d_{t+1}d_{t+2}\ldots$, soma-se 1 ao dígito d_t se $d_{t+1} \geq 5$,
  - Se $|x| < m$, então o número não pode ser representado no sistema e diz-se que ocorre *underflow*
  - Se $|x| > M$, então o número não pode ser representado no sistema e diz-se que ocorre *overflow*

Exemplo

Exemplo 2

Considere o sistema $F(10, 3, [-2, 2])$. Represente nesse sistema, se possível, os números:

$$x_1 = 0.35, \quad x_2 = -5.17, \quad x_3 = 0.0123, \quad (1)$$
$$x_4 = 5390, \quad x_5 = 0.0003. \quad (2)$$

Resposta:

$x_1 = 0.350 \cdot 10^0$, $x_2 = -0.517 \cdot 10^1$, $x_3 = 0.123 \cdot 10^{-1}$.

O número $5390 = 0.539 \cdot 10^4$ não pode ser representado porque seu expoente é maior que 2. Tem-se *overflow*.
O número $0.0003 = 0.300 \cdot 10^{-3}$ não pode ser representado porque seu exponente é menor que -2. Tem-se um *underflow*.

Exemplo

Exemplo 3

Represente no sistema $F(10, 3, [-5, 5])$ os números

$$x_1 = 1234.56, \quad x_2 = -0.00054962, \quad x_3 = 0.9995,$$
$$x_4 = 123456.7, \quad x_5 = 0.0000001.$$

Resposta:

$\mathtt{fl}(x_1) = 0.123 \cdot 10^4$, $\mathtt{fl}(x_2) = -0.550 \cdot 10^{-3}$,
$\mathtt{fl}(x_3) = 0.100 \cdot 10^1$.

Para x_4 e x_5 tem-se *overflow* e *underflow*, respectivamente.

Para arredondar um número na base $\beta = 10$, devemos apenas observar o primeiro dígito a ser descartado. Se ele for menor que 5, deixamos os dígitos inalterados; Se ele é maior ou igual a 5, devemos somar 1 ao último dígito remanescente.

Operações Aritméticas em Ponto Flutuante

Operações Aritméticas em Ponto Flutuante

- Adição/Subtração
  - Deve-se ajustar o número de menor expoente para igualá-lo ao do outro número
- Multiplicação/Divisão
  - Realiza-se a operação nas mantissas e nos expoentes
- Os valores devem ser representados no sistema utilizado
- Os resultados devem ser truncados ou arredondados
  - Definição do sistema

Em um sistema de base 10 com $t = 4$, temos

$$\begin{aligned} 0{,}4370\times10^5 + 0{,}1565\times10^3 &= 0{,}4370\times10^5 + 0{,}0016\times10^5 \\ &= (0{,}4370 + 0{,}0016)\times10^5 \\ &= 0{,}4386\times10^5 \end{aligned}$$

Em um sistema de base 10 com $t = 4$, temos

$$\begin{aligned} 0{,}0000\times10^0 + 0{,}1428\times10^{-2} &= 0{,}0000\times10^0 + 0{,}0014\times10^0 \\ &= 0{,}0014\times10^0 \\ &= 0{,}1400\times10^{-2} \end{aligned}$$

Em um sistema de base 10 com $t = 4$, temos

$$\begin{aligned} 0,4370 \times 10^5 \times 0,1565 \times 10^3 &= (0,4370 \times 0,1565) \times 10^{5+3} \\ &= 0,6839 \times 10^{-1} \times 10^8 \\ &= 0,6839 \times 10^7 \end{aligned}$$

Exemplo

- **Exemplo 13**
  Seja o sistema $F(10, 2, L, U)$ com arredondamento; os limitantes do expoente são ignorados nesse exemplo. Some $4,32$ e $0,064$ nesse sistema.

- 0,43 x 10^1
- 0,64 x 10^{-1} = 0,0064 x 10^1

- (0,43 + 0,0064) x 10^1
- (0,4364) x 10^1
- (0,44) x 10^1

- 0,43 x 10^1
- 0,64 x 10^{-1} = 0,0064 x 10^1 = 0,01 x 10^1

- (0,43 + 0,01) x 10^1
- (0,44) x 10^1

Exemplo

- **Exemplo 14**
  Seja o sistema $F(10, 2, L, U)$ com arredondamento. Multiplique 1234 por $0,016$ nesse sistema.

- 1234 = 0,1234 x 10^4 = 0,12 x 10^4
- 0,016 = 0,16 x 10^{-1}

- (0,12 x 0,16) x 10^4 x 10^{-1}
- (0,0192) x 10^{4-1}
- (0,192) x 10^{-1} x 10^3
- (0,19) x 10^2

Noções Básicas Sobre Erros

Noções Básicas Sobre Erros

- Além disso, erros podem ser introduzidos ao representar números no computador
- Um número real x provavelmente será aproximado quando representado em ponto flutuante no computador
- É necessário definir medidas para calcular erros em aproximações
  - erro absoluto
  - erro relativo

Erro Absoluto

- Seja $\bar{x}$ uma aproximação de x, o erro absoluto é definido como

$$EA(\bar{x}) = |x - \bar{x}|$$

Exemplo

- **Exemplo 15**
  Seja o sistema $F(10, 4, L, U)$ com arredondamento. Qual o erro absoluto ao representar $x = 1428{,}756$ nesse sistema?
- Solução:
  $\bar{x} = 0{,}1429 \times 10^4 \Rightarrow EA(\bar{x}) = |1428{,}756 - 1429| = 0{,}244$

Exemplo

- **Exemplo 16**
  Seja o sistema $F(10, 4, L, U)$ com truncamento. Qual o erro absoluto ao representar $x = 1428{,}756$ nesse sistema?
- Solução:
  $\bar{x} = 0{,}1428 \times 10^4 \Rightarrow EA(\bar{x}) = |1428{,}756 - 1428| = 0{,}756$

Erro Relativo

- Seja $\bar{x}$ uma aproximação de x, o erro relativo é definido como
$$ER(\bar{x}) = \frac{|x - \bar{x}|}{|x|} = \frac{EA(\bar{x})}{|x|}$$
- dado $x \neq 0$.

Exemplo

- **Exemplo 17**
  Sejam $x_1 = 1000{,}5$, $\bar{x_1} = 1000{,}6$, $x_2 = 10{,}5$ e $\bar{x_2} = 10{,}6$. Nota-se que $EA(\bar{x_1}) = EA(\bar{x_2}) = 0{,}1$. Quais os erros relativos?
- Solução:
  $ER(\bar{x_1}) = \frac{0{,}1}{1000{,}6} \approx 0{,}9995 \times 10^{-4}$
  $ER(\bar{x_2}) = \frac{0{,}1}{10{,}6} \approx 0{,}9524 \times 10^{-2}$