1- Noções básicas sobre erros e aritmética de ponto flutuante

Introdução

- O **Cálculo Numérico** corresponde a um conjunto de **ferramentas** ou **métodos** usados para se obter a solução de problemas matemáticos de **forma aproximada.**

- Se preocupa com a construção da solução e com a sua qualidade, isto é, o quão distante ela está da solução exata (mesmo que a solução exata seja desconhecida).

- Importante em diversas áreas de conhecimento:

  - Engenharias, Ciência da Computação, Economia, Medicina, Física, Química, Biologia, entre outras.

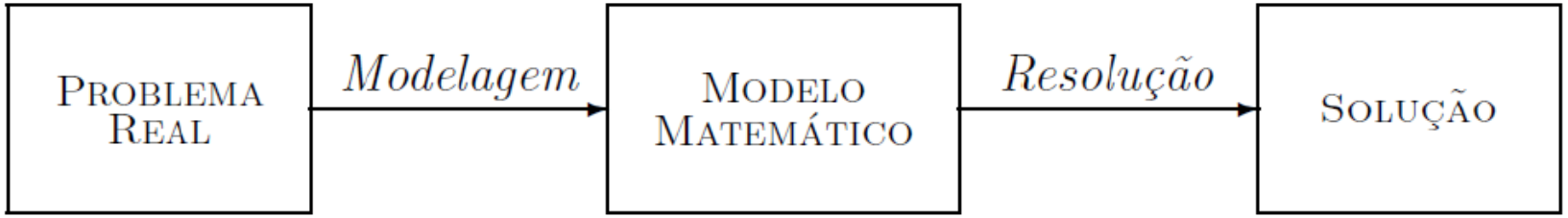
- **Objetivo geral:**

  - Prover fundamentação teórica para o desenvolvimento e implementação de métodos numéricos para a resolução de problemas.

Etapas na resolução de um problema

- Grande parte dos problemas matemáticos surge da necessidade de solucionar problemas da natureza, sendo que é possível descrever muitos fenômenos naturais por meio de **modelos matemáticos**.

- Etapas para solucionar um problema da natureza:



- **Modelagem do problema:** etapa inicial que consiste na representação do problema por um modelo matemático conveniente.

- **Resolução do modelo**: etapa em que se busca encontrar uma solução para o modelo matemático obtido na fase de modelagem.
  - Métodos analíticos (solução exata)
  - Métodos numéricos (solução aproximada)

Exemplos:

(1) Solução por método analítico

Um método analítico para determinar (quando existem) os zeros reais de uma função quadrática

$$f(x) = ax^2 + bx + c, \neq 0$$

é dado pela fórmula de Bhaskara, a saber:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Desse modo, os zeros reais de $f(x) = x^2 - 5x + 6$ são

$$x_1 = \frac{-(-5) - \sqrt{(-5)^2 - 4 \times 1 \times 6}}{2 \times 1} = 2 \text{ e } x_2 = \frac{-(-5) + \sqrt{(-5)^2 - 4 \times 1 \times 6}}{2 \times 1} = 3$$

(2) Solução por método numérico

Um método numérico para **determinar uma aproximação para a raiz quadrada de um número real** p, maior que 1, é o **algoritmo de Eudoxo**:

- Do fato que $p > 1$, temos que $1 < \sqrt{p} < p$.

- Escolhe-se, como uma primeira aproximação para $\sqrt{p}$, $x_0 = \frac{1+p}{2}$, ou seja, a média aritmética entre 1 e p. Pode-se mostrar que $\frac{p}{x_0} < \sqrt{p} < x_0$.

- Escolhe-se como uma nova aproximação $x_1 = \frac{\frac{p}{x_0}+x_0}{2}$, isto é, a média aritmética entre $\frac{p}{x_0}$ e x_0. Novamente, pode-se mostrar que $\frac{p}{x_1} < \sqrt{p} < x_1$.

- Continuando desse modo, pode-se construir uma sequência de aproximações dada por:

$$x_n = \begin{cases} (1+p)/2 & \text{se} \quad n = 0 \\ \left(\frac{p}{x_{n-1}} + x_{n-1}\right)/2 & \text{se} \quad n \geq 1 \end{cases}$$

Algoritmo de Eudoxo para $\sqrt{2}$		
n	x_n	x_n^2
0	1,50000000000000	2,25000000000000
1	1,41666666666667	2,00694444444444
2	1,41421568627451	2,00000600730488
3	1,41421356237469	2,00000000000451
4	1,41421356237310	2,00000000000000

Por que estudar Cálculo Numérico?

- Os métodos numéricos desenvolvidos e estudados no cálculo numérico servem, em geral, para a aproximação da solução de problemas complexos que normalmente não são resolúveis por técnicas analíticas.

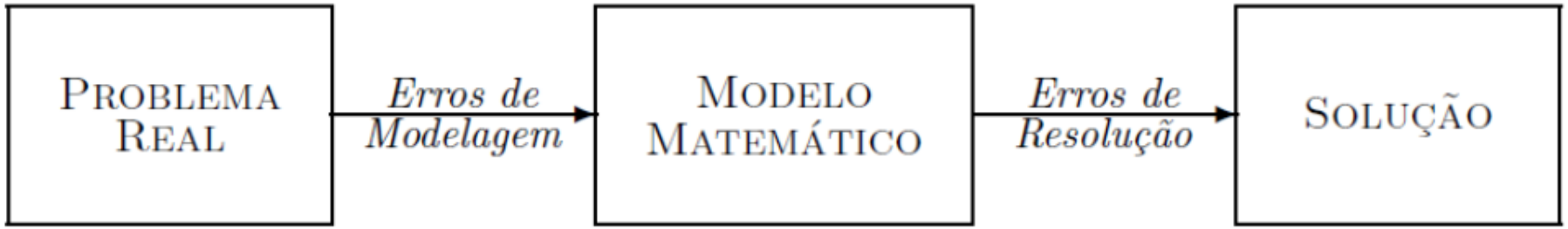
- **Objetivos específicos:**

  - Entender o que são *métodos numéricos* de aproximação, como e por que utilizá-los, e quando é esperado que eles funcionem.

  - Identificar problemas que requerem o uso de técnicas numéricas para a obtenção de sua solução.

  - Conhecer e aplicar os principais métodos numéricos para a solução problemas clássicos: obter zeros reais de funções reais, resolver sistemas de equações lineares, fazer interpolação polinomial, ajustar curvas e fazer integração numérica.

Erros nas etapas de resolução de um problema



- Não é raro acontecer que os resultados finais estejam distantes do que se esperaria obter, ainda que todas as fases de resolução tenham sido realizadas corretamente.

- **Erros de Modelagem:** Devido às simplificações no processo de modelagem matemática de um problema, que muitas vezes são necessárias, podem ocorrer erros na representação do fenômeno da natureza que estivermos analisando.

- **Erros de Resolução:** São erros devido ao fato dos equipamentos computacionais terem capacidade limitada para armazenar os dígitos significativos de valores numéricos, utilizados nas operações elementares de adição, multiplicação, subtração, divisão, etc.

Logo, é importante entender como os números são representados no computador e como as operações aritméticas são realizadas.

Representação de números

Representação em um sistema de bases

A representação de um número em uma base β é indicada pela soma dos dígitos dessa base por potências de β:

$$(N)_\beta = (d_n d_{n-1} d_{n-2} \ldots d_0, d_{-1} d_{-2} \ldots d_{-m})_\beta$$
$$= d_n\beta^n + d_{n-1}\beta^{n-1} + d_{n-2}\beta^{n-2} + \cdots + d_0\beta^0 + d_{-1}\beta^{-1} + d_{-2}\beta^{-2} + \cdots d_{-m}\beta^{-m}$$

onde os dígitos $d_j \in \{0, 1, 2, \ldots, \beta - 1\}$.

Sistema de numeração decimal ou base 10

Nesse caso todos os múltiplos e submúltiplos de um número são escritos com potencias de 10.

Exemplos:

$$1537 = (1537)_{10} = 7 \times 10^0 + 3 \times 10^1 + 5 \times 10^2 + 1 \times 10^3$$

$$36.189 = (36.189)_{10} = 9 \times 10^{-3} + 8 \times 10^{-2} + 1 \times 10^{-1} + 6 \times 10^0 + 3 \times 10^1$$

Sistema de numeração binário ou base 2

Nesse caso todos os múltiplos e submúltiplos de um número são escritos com potencias de 2.

Exemplos:

$$(10111)_2 = 1 \times 2^0 + 1 \times 2^1 + 1 \times 2^2 + 0 \times 2^3 + 1 \times 2^4$$

$$(10.1)_2 = 1 \times 2^{-1} + 0 \times 2^0 + 1 \times 2^1$$

Conversão de números: mudança de base

O exemplo a seguir ilustra como efetuar a conversão de números não negativos entre a base decimal e binária.

Exemplo: Mudar a representação dos números:

i) **1101** *da base 2, para a base 10,*

ii) **0.110** *da base 2, para a base 10,*

iii) **13** *da base 10, para a base 2,*

iv) **0.75** *da base 10, para a base 2,*

v) **3.8** *da base 10, para a base 2.*

Solução: Vejamos a seguir o procedimento a ser seguido para cada número.

i) **1101** que está na base 2, para a base 10.

Neste caso o procedimento é multiplicar cada algarismo do número na base 2 por potências crescente de 2, da direita para a esquerda e somar todas as parcelas. Assim:

$$1101 = 1 \times 2^0 + 0 \times 2^1 + 1 \times 2^2 + 1 \times 2^3 = 1 + 0 + 4 + 8 = 13 \ .$$

Logo, $(1101)_2 = (13)_{10}$.

ii) **0.110** que está na base 2, para a base 10.

Neste caso o procedimento é multiplicar cada algarismo do número na base 2, após o ponto, por potências decrescente de 2, da esquerda para a direita e somar todas as parcelas. Assim:

$$0.110 = 1 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} = \frac{1}{2} + \frac{1}{4} + 0 = 0.75 \ .$$

Logo, $(0.110)_2 = (0.75)_{10}$.

iii) **13** que está na base 10, para a base 2.

Neste caso o procedimento é dividir o número por 2. A seguir continuar dividindo o quociente por 2 até que o último quociente seja igual a 1. O número na base 2 será então obtido tomando-se o último quociente e todos os restos das divisões anteriores. Assim:

$$\begin{array}{r|l} 13 & 2 \\ \hline \mathbf{1} & 6 \quad \big| \quad 2 \\ \hline & \mathbf{0} \quad\quad 3 \quad \big| \quad 2 \\ \hline & \quad\quad\quad\; \mathbf{1} \quad\quad 1 \end{array}$$

Logo, $(13)_{10} = (1101)_2$.

iv) **0.75** que está na base 10, para a base 2.

Neste caso o procedimento é multiplicar a parte decimal por 2. A seguir continuar multiplicando a parte decimal do resultado obtido, por 2. O número na base 2 será então obtido tomando-se a parte inteira do resultado de cada multiplicação. Assim:

$$\begin{aligned} 0.75 \times 2 &= \mathbf{1}.50 \\ 0.50 \times 2 &= \mathbf{1}.00 \\ 0.00 \times 2 &= \mathbf{0}.00 \end{aligned}$$

Logo, $(0.75)_{10} = (0.110)_2$.

v) **3.8** que está na base 10, para a base 2.

O procedimento neste caso é transformar a parte inteira seguindo o item **iii)** o que nos fornece $(3)_{10} = (11)_2$ e a parte decimal seguindo o item **iv)**. Assim, obtemos:

$$\begin{aligned} 0.8 \times 2 &= \mathbf{1}.6 \\ 0.6 \times 2 &= \mathbf{1}.2 \\ 0.2 \times 2 &= \mathbf{0}.4 \\ 0.4 \times 2 &= \mathbf{0}.8 \\ 0.8 \times 2 &= \ldots \end{aligned}$$

Logo, $(3.8)_{10} = (11.11001100\ldots)_2$. Portanto o número $(3.8)_{10}$ não tem representação exata na base 2. Esse exemplo ilustra também o caso de erro de arredondamento nos dados.

Observação:

- No exemplo anterior, mudamos a representação de núumeros na base 10 para a base 2 e vice-versa. O mesmo procedimento pode ser utilizado para mudar da base 10 para uma outra base qualquer e vice-versa.

- A pergunta que surge naturalmente é: qual o procedimento para representar um número que está numa dada base β_1 em uma outra base β_2, onde β_1 e β_2 são diferentes da base 10? Nesse caso devemos seguir o seguinte procedimento: inicialmente representamos o número que está na base β_1 na base 10 e a seguir o número obtido na base 10, na base β_2.

Exemplo: Dado o número $(12.20)_4$ que está na base 4, representá-lo na base 3.

Solução: Assim, usando os procedimentos dados no exemplo 2.4, obtemos:

$$12 = 2 \times 4^0 + 1 \times 4^1 = 6.$$
$$0.20 = 2 \times 4^{-1} + 0 \times 4^{-2} = \frac{2}{4} = 0.5 \ .$$

Portanto: $(12.20)_4 = (6.5)_{10}$.

Agora:

$$\begin{array}{r|l} 6 & 3 \\ \hline 0 & 2 \end{array}$$

$$\begin{aligned} 0.5 \times 3 &= 1.5 \\ 0.5 \times 3 &= 1.5 \\ &\vdots \end{aligned}$$

Portanto: $(6.5)_{10} = (20.11\ldots)_3$. Logo $(12.20)_4 = (20.111\ldots)_3$. Observe que o número dado na base 4, tem representação exata na base 10, mas não na base 3.

Exercícios

(1) Converta os seguintes números binários para sua forma decimal:

(a) $(101101)_2$

(b) $(110101011)_2$

(c) $(0.1101)_2$

(d) $(111.01)_2$

(2) Converta os seguintes números decimais para sua forma binária:

(a) 37

(b) 64

(c) 3.25

(d) 0.3

Representação de números no computador:

Representação em Ponto Flutuante

- O computador, com sua memória finita, não é capaz de representar todos os números reais.
- Números como $\pi = 3.141592\ldots$ são aproximados.
- Nos computadores costuma-se usar uma representação denominada **representação em ponto flutuante.**
- Os parâmetros e propriedades que definem essa representação constituem um **sistema de ponto flutuante**.

Definição: A **representação em ponto flutuante de um número real x** é dada por

$$x = \pm d \times \beta^e,$$

onde d é a **mantissa**, β é a **base do sistema de numeração** e e é o **expoente**.

- A mantissa é um número na forma

$$(0.d_1 d_2 \ldots d_t)_\beta$$

onde t é o número de dígitos e $d_i \in \{0, 1, \ldots, (\beta - 1)\}$, $i = 1, \ldots, t$.

- O expoente e é definido no intervalo $[L, U]$.

Observações:

(**1**) Um número é dito **normalizado** quando $d_1 \neq 0$.

(**2**) O número **zero** pertence a qualquer sistema de ponto flutuante e é representado com mantissa igual a zero e expoente assumindo o menor valor possível.

(**3**) Trabalharemos apenas com números normalizados.

Um **sistema de ponto flutuante** será representado por

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

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.

Exemplos: (1) Considere o sistema de ponto flutuante dado por $F(10, 3, -5, 5)$.

(**i**) Nesse sistema o número 1.23 é representado por $1.23 = 0.123 \times 10^1$.

(**ii**) Nesse sistema o número 0.0418 é representado por $0.0418 = 0.418 \times 10^{-1}$.

(**iii**) O menor número, em valor absoluto, que esse sistema pode representar é

$$m = 0.100 \times 10^{-5} = \frac{1}{10} \times \frac{1}{10^5} = \frac{1}{10^6} = 10^{-6}.$$

(**iv**) O maior número, em valor absoluto, que esse sistema pode representar é

$$M = 0.999 \times 10^5 = 99900.$$

(2) Considere o sistema de ponto flutuante dado por $F(2, 2, -1, 2)$.

Os números que são representáveis nesse sistema são da forma:

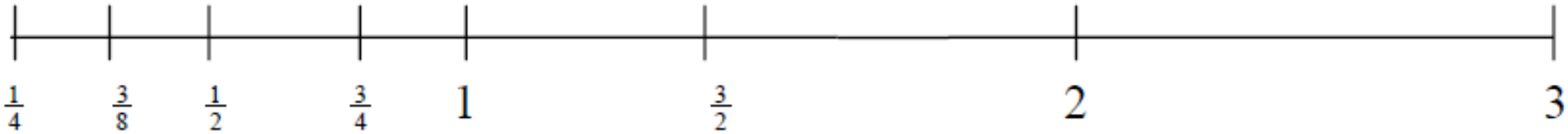
$$\pm 0.10 \times 2^e \text{ ou } \pm 0.11 \times 2^e, \text{ com } -1 \leq e \leq 2.$$

Em decimal, temos que $(0.10)_2 = \frac{1}{2}$ e $(0.11)_2 = \frac{3}{4}$.

Logo, os únicos números positivos que serão representados exatamente nesse sistema são:

$$\frac{1}{2} \times 2^e \text{ ou } \frac{3}{4} \times 2^e, \text{ com } -1 \leq e \leq 2,$$

ou seja, $\frac{1}{4}, \frac{1}{2}, 1, 2, \frac{3}{8}, \frac{3}{4}, \frac{3}{2}$ e 3.



Além desses números, os seus respectivos números negativos e o número zero também serão representados, totalizando 17 números.

Observações sobre a representação em Ponto Flutuante

Sejam m e M, respectivamente, o menor e o maior valor absoluto representável no sistema $F(\beta, t, L, U)$. Dado um número real x, temos que:

- Se $m \leq |x| \leq M$, então x pode ser representado nesse sistema e sua representação será feita através de um **arredondamento** ou **truncamento**;
  - Arredondamento: na base 10, remove-se os dígitos $d_{t+1}d_{t+2}$ ... e soma-se 1 ao dígito d_t, se $d_{t+1} \geq 5$
  - Truncamento: remove-se os dígitos $d_{t+1}d_{t+2}$ ...
- Se $|x| < m$, então x **não pode ser representado nesse sistema** e diz-se que ocorreu um *underflow;*
- Se $|x| > M$, então x **não pode ser representado nesse sistema** e diz-se que ocorreu um *overflow.*

Exemplo: Escreva, se possível, a representação dos números a seguir, no sistema de ponto flutuante $F(10, 3, -5, 5)$.

(**a**) 235.89 (**b**) -5.86231 (**c**) 0.003636

(**d**) 0.0000002 (**e**) 125340

Solução:

(**a**) Observe que $x = 235.89 = 0.23589 \times 10^3$. Como a mantissa possui 5 dígitos, então será necessário realizar um arredondamento ou truncamento. Temos que:

- Representação por Arredondamento: $x = 0.236 \times 10^3$
- Representação por Truncamento: $x = 0.235 \times 10^3$

(**b**) Observe que $x = -5.86231 = -0.586231 \times 10^1$. Como a mantissa possui 6 dígitos, então será necessário realizar um arredondamento ou truncamento. Temos que:

- Representação por Arredondamento: $x = -0.586 \times 10^1$
- Representação por Truncamento: $x = -0.586 \times 10^1$

(**c**) Observe que $x = 0.003636 = 0.3636 \times 10^{-2}$. Como a mantissa possui 4 dígitos, então será necessário realizar um arredondamento ou truncamento. Temos que:

- Representação por Arredondamento: $x = 0.364 \times 10^{-2}$

- Representação por Truncamento: $x = 0.363 \times 10^{-2}$

(**d**) Temos que $0.0000002 = 0.2 \times 10^{-6}$. Como o expoente e é menor que -5, então esse número não pode ser representado nesse sistema, ocorrendo um *underflow*.

(**e**) Temos que $125340 = 0.125340 \times 10^{6}$. Como o expoente e é maior que 5, então esse número não pode ser representado nesse sistema, ocorrendo assim um *overflow*.

Observações:

- Quanto maior o intervalo para o expoente e, maior será a faixa de números que um sistema de ponto flutuante pode representar;

- Quanto maior o número de algarismos para a mantissa, maior será a precisão da representação.

- O sistema de ponto flutuante é discreto, ao contrário do que ocorre com os números reais, que são contínuos.

Operações Aritméticas em Ponto Flutuante

Regras Operatórias:

- **Adição e Subtração:**
  A adição/subtração requer o alinhamento dos pontos decimais dos dois números. Deve-se ajustar o número de menor expoente para igualá-lo ao do outro número.

- **Multiplicação e Divisão:**
  Realiza-se operação nas mantissas e nos expoentes.

- Os valores devem ser representados no sistema utilizado.

- Os resultados devem ser arredondados ou truncados, de acordo com a definição do sistema.

Exemplos:

(1) Considere o sistema de ponto flutuante $F(10, 4, L, U)$ com arredondamento, onde os limitantes do expoente são ignorados. Sejam $x = 0.9370 \times 10^4$ e $y = 0.1272 \times 10^2$. Calcule, nesse sistema $x + y$ e xy.

Solução:

(**i**) Alinhando os pontos decimais, obtemos:

$$x = 0.9370 \times 10^4 \text{ e } y = 0.001272 \times 10^4.$$

Logo,

$$x + y = 0.938272 \times 10^4$$

Como o sistema possui 4 dígitos na mantissa, aplicando o arredondamento, obtemos:

$$x + y = 0.9383 \times 10^4.$$

(**ii**) Temos que o resultado exato da operação é:

$$\begin{aligned} xy &= (0.9370 \times 10^4) \times (0.1272 \times 10^2) \\ &= (0.9370 \times 0.1272) \times 10^6 \\ &= 0.1191864 \times 10^6. \end{aligned}$$

Como o sistema possui 4 dígitos na mantissa, aplicando o arredondamento, obtemos:

$$xy = 0.1192 \times 10^6.$$

(2) Considere o sistema de ponto flutuante $F(10, 3, L, U)$ com arredondamento, onde os limitantes do expoente são ignorados.

(**a**) Some 4.32 e 0.064 nesse sistema.
(**b**) Multiplique 1235 por 0.016 nesse sistema.

Solução:

(**a**) No sistema dado, temos que

$$4.32 = 0.432 \times 10^1 \quad \text{e} \quad 0.064 = 0.640 \times 10^{-1}.$$

Logo,

$$\begin{aligned} 4.32 + 0.064 &= 0.432 \times 10^1 + 0.640 \times 10^{-1} \quad \text{(representação no sistema)} \\ &= 0.432 \times 10^1 + 0.0064 \times 10^1 \quad \text{(organização dos expoentes)} \\ &= (0.432 + 0.0064) \times 10^1 \\ &= 0.4384 \times 10^1 \end{aligned}$$

Como o sistema possui 3 dígitos, fazendo o arredondamento, obtemos $0.438 \times 10^1 = 4.38$.

(**b**) Temos que

- $1235 = 0.1235 \times 10^4 \Rightarrow$ No sistema dado: $1235 = 0.124 \times 10^4$
- $0.016 = 0.160 \times 10^{-1} \Rightarrow$ No sistema dado: $0.016 = 0.160 \times 10^{-1}$

Logo,

$$1235 \times 0.016 = (0.124 \times 10^4) \times (0.160 \times 10^{-1}) \quad \text{(representação no sistema)}$$

$$= (0.124 \times 0.160) \times (10^4 \times 10^{-1}) \quad \text{(operações)}$$

$$= 0.01984 \times 10^3$$

$$= 0.1984 \times 10^2$$

Como o sistema possui 3 dígitos, fazendo o arredondamento, obtemos $0.198 \times 10^2 = 19.8$.

(3) Considere o sistema de ponto flutuante com base $\beta = 10$ e 3 dígitos significativos, com arredondamento. Efetue as operações a seguir nesse sistema:

(a) $(11.4 + 3.18) + 5.05$ e $11.4 + (3.18 + 5.05)$

(b) $3.18 \times (5.05 + 11.4)$ e $3.18 \times 5.05 + 3.18 \times 11.4$

Solução: No sistema considerado, temos que:

- $11.4 = 0.114 \times 10^2$
- $3.18 = 0.318 \times 10^1$
- $5.05 = 0.505 \times 10^1$

(**a**)

- $(11.4 + 3.18) + 5.05$

Temos que: $11.4 + 3.18 = 0.114 \times 10^2 + 0.318 \times 10^1$

$$= 0.114 \times 10^2 + 0.0318 \times 10^2$$
$$= 0.1458 \times 10^2$$
$$= 0.146 \times 10^2$$

Daí,

$$(11.4 + 3.18) + 5.05 = 0.146 \times 10^2 + 0.505 \times 10^1$$
$$= 0.146 \times 10^2 + 0.0505 \times 10^2$$
$$= 0.1965 \times 10^2$$
$$= 0.197 \times 10^2$$

Desse modo, nesse sistema, $(11.4 + 3.18) + 5.05 = 19.7$.

(**a**)

- $11.4 + (3.18 + 5.05)$

Temos que: $3.18 + 5.05 = 0.318 \times 10^1 + 0.505 \times 10^1$

$$= 0.823 \times 10^1$$

Daí,

$$\begin{aligned} 11.4 + (3.18 + 5.05) &= 0.114 \times 10^2 + 0.823 \times 10^1 \\ &= 0.114 \times 10^2 + 0.0823 \times 10^2 \\ &= 0.1963 \times 10^2 \\ &= 0.196 \times 10^2 \end{aligned}$$

Desse modo, nesse sistema, $11.4 + (3.18 + 5.05) = 19.6$.

(b)

- $3.18 \times (5.05 + 11.4)$

Temos que: $5.05 + 11.4 = 0.505 \times 10^1 + 0.114 \times 10^2$

$$= 0.0505 \times 10^2 + 0.114 \times 10^2$$
$$= 0.1645 \times 10^2$$
$$= 0.165 \times 10^2$$

Daí,

$$3.18 \times (5.05 + 11.4) = 0.318 \times 10^1 \times 0.165 \times 10^2$$
$$= 0.05247 \times 10^3$$
$$= 0.5247 \times 10^2$$
$$= 0.525 \times 10^2$$

Desse modo, nesse sistema, $3.18 \times (5.05 + 11.4) = 52.5$.

(b)

- $3.18 \times 5.05 + 3.18 \times 11.4$

Temos que: $3.18 \times 5.05 = (0.318 \times 10^1) \times (0.505 \times 10^1)$

$$= 0.16059 \times 10^2$$

$$= 0.161 \times 10^2$$

E,

$$3.18 \times 11.4 = (0.318 \times 10^1) \times (0.114 \times 10^2)$$

$$= 0.036252 \times 10^3$$

$$= 0.36252 \times 10^2$$

$$= 0.363 \times 10^2$$

Logo,

$$3.18 \times 5.05 + 3.18 \times 11.4 = (0.161 \times 10^2) + (0.363 \times 10^2)$$

$$= 0.524 \times 10^2$$

Desse modo, nesse sistema, $3.18 \times 5.05 + 3.18 \times 11.4 = 52.4$.

Observações:

(i) Os exemplos anteriores mostram que ainda que as parcelas ou fatores de uma operação estejam representados exatamente no sistema de ponto flutuante, não se pode esperar que o resultado armazenado seja exato.

(ii) As operações aritméticas no sistema de ponto flutuante **não são nem associativas e nem distributivas.**

Tipos de Erros: Erro Absoluto e Erro Relativo

Definição: Seja $\bar{x}$ uma aproximação de x.

O **erro absoluto** é definido como

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

O **erro relativo** é definido como

$$\boxed{ER(\bar{x}) = \frac{|x - \bar{x}|}{|x|} = \frac{|EA(\bar{x})|}{|x|}.}$$

Observação: O Erro Relativo nos fornece mais informações sobre a qualidade do erro que estamos cometendo num determinado cálculo, uma vez que no Erro Absoluto não é levada em consideração a ordem de grandeza do valor calculado, enquanto que no Erro Relativo essa ordem é contemplada.

Exemplos: (1)

(a) Considere o valor exato $x = 2345.713$ e o valor aproximado $\bar{x} = 2345.000$.

O erro absoluto é

$$EA(\bar{x}) = |2345.713 - 2345.000| = 0.713.$$

O erro relativo é

$$ER(\bar{x}) = \frac{EA(\bar{x})}{|x|} = \frac{0.713}{2345.713} = 0.00030396\ (0.03\%)$$

(b) Considere o valor exato $x = 1.713$ e o valor aproximado $\bar{x} = 1.000$.

O erro absoluto é

$$EA(\bar{x}) = |1.713 - 1.000| = 0.713.$$

O erro relativo é

$$ER(\bar{x}) = \frac{EA(\bar{x})}{|x|} = \frac{0.713}{1.713} = 0.416229\ (41.6\%)$$

Observe que em a) e b) o erro absoluto é o mesmo, embora o erro cometido pela aproximação seja muito mais significativo no exemplo b).

(2) Considere o sistema de ponto flutuante $F(10, 4, L, U)$, com arredondamento. Seja $x = 1428.756$. Determine o erro absoluto e o erro relativo ao representar x nesse sistema.

Solução: Temos que

$$x = 1428.756 = 0.1428756 \times 10^4 \Rightarrow \text{No sistema dado: } \bar{x} = 0.1429 \times 10^4 = 1429.$$

- Erro absoluto: $EA(\bar{x}) = |x - \bar{x}| = |1428.756 - 1429| = 0.244$
- Erro relativo: $ER(\bar{x}) = \frac{|x-\bar{x}|}{|x|} = \frac{EA(\bar{x})}{|x|} = \frac{0.244}{1428.756} \approx 0.00017\ (0.017\ \%)$

Observação: Na prática, o valor de x geralmente não é conhecido. Assim, utiliza-se uma medida de erro entre aproximações:

$$|x_{k+1} - x_k| \text{ e } \frac{|x_{k+1}-x_k|}{|x_{k+1}|}.$$

Efeitos Numéricos

Além dos erros causados pela representação no computador, existem certos efeitos numéricos que contribuem para aumentar os erros:

(**i**) Somar (ou subtrair) números com ordens de grandeza muito diferentes;

(**ii**) Cancelamento;

(**iii**) Propagação de erro e Instabilidade Numérica.

(i) Somar (ou subtrair) números com ordens de grandeza muito diferentes

As operações de soma e subtração podem não ter o efeito desejado.

Por exemplo, ao somar 0.1 e 5000 no sistema $F(10, 4, L, U)$, obtém-se:

$$\begin{aligned} 0.1 + 5000 &= 0.1000 \times 10^0 + 0.5000 \times 10^4 \\ &= 0.00001 \times 10^4 + 0.5000 \times 10^4 \\ &= 0.50001 \times 10^4 \\ &= 0.5000 \times 10^4 \quad \text{(arredondando ou truncando)} \end{aligned}$$

(ii) Cancelamento

- Ocorre na subtração de dois números muito parecidos;
- Na aritmética de ponto flutuante, para calcular $x - y$, deve-se igualar os expoentes dos números;
- Quando x e y são próximos, vários zeros aparecem no final da mantissa do resultado normalizado;
- Ocorre assim, uma perda de dígitos significativos.

Por exemplo, ao realizar a subtração $\sqrt{9876} - \sqrt{9875}$ no sistema $F(10, 10, L, U)$, temos que:

- $\sqrt{9876} = 0.9937806599 \times 10^2$ e $\sqrt{9875} = 0.9937303457 \times 10^2$
- $\sqrt{9876} - \sqrt{9875} = 0.0000503142 \times 10^2$

Fazendo a normalização, obtemos o resultado $0.5031420000 \times 10^{-2}$.
Observe que os quatro zeros no final da mantissa não tem significado e assim, perdemos 4 casas decimais.

(iii) Propagação de erro e Instabilidade numérica

- Se um resultado intermediário de um cálculo é contaminado por um erro de arredondamento, este erro pode influenciar todos os resultados subsequentes que dependem desse resultado intermediário;

- Cada novo resultado intermediário introduz um novo erro de arredondamento.

- É de se esperar portanto, que todos esses erros influenciem o resultado final.

- Instabilidade Numérica ocorre se os erros intermediários tem uma influencia muito grande no resultado final.

- É necessário analisar como esses erros se propagam:
  - **erro ilimitado:** se acumulam a uma taxa crescente e a sequência de operações é considerada **instável**;
  - **erro limitado:** se acumulam a uma taxa decrescente e a sequência de operações é considerada **estável**.

- Em outros casos (como em processos iterativos) os erros intermediários podem ter um efeito desprezível no resultado final (**algoritmos estáveis**).

Exercícios

(1) Escreva, se possível, a representação por arredondamento e por truncamento dos números a seguir, no sistema de ponto flutuante $F(10, 3, -4, 4)$.

(a) $x_1 = 1.25$ **(b)** $x_2 = 10.053$ **(c)** $x_3 = -238.15$

(d) $x_4 = 2.71828\ldots$ **(e)** $x_5 = 0.000007$ **(f)** $x_6 = 718235.82$

(2) Seja um sistema de aritmética de ponto flutuante de quatro dígitos e base decimal. Dados os números: $x = 0{,}7237 \times 10^4$, $y = 0.2145 \times 10^{-3}$ e $z = 0.2585 \times 10^{-1}$ efetue as operações:

(a) $x + y + z$

(b) $\frac{xy}{z}$.