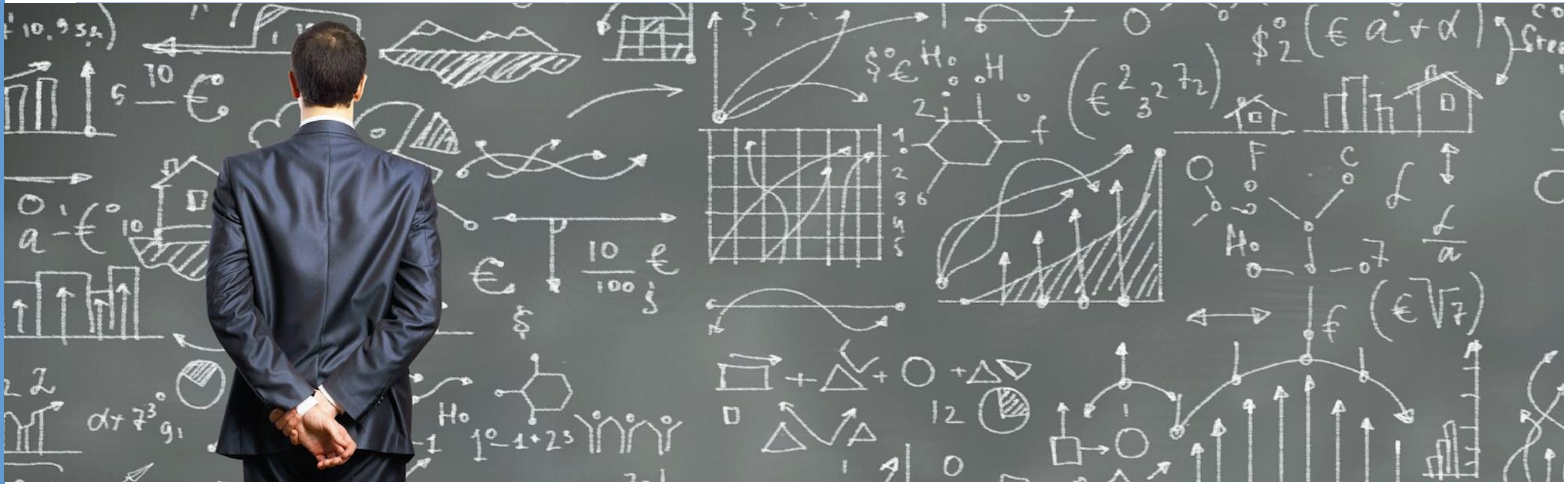
Métodos Numéricos

UNIVERSIDADE
FEDERAL DO CEARÁ



Unidade I: Ponto Flutuante e Erros

Representação dos Números

UNIVERSIDADE FEDERAL DO CEARÁ

- Os números empregados no calculo computacional podem ser de dois tipos: **números inteiros** e **números em “ponto flutuante”** (números reais da matemática, por exemplo 3.56 → 0.356×10^1).
- Os computadores atuais **representam os números internamente no formato binário**, como uma sequência de 0s e 1s.
- Apesar dessa representação ser **conveniente para as maquinas** é **antinatural para os seres humanos**, cujo sistema de numeração é o decimal.
  - Obs. No passado o nosso sistema de numeração já foi também na base 12 (ex. contar nas falanges dos dedos) na base 60 (ex. sistema horário).

Representação em Ponto Flutuante

UNIVERSIDADE FEDERAL DO CEARÁ

- 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:

$$m = \pm\, , d_1 d_2 d_3 \ldots d_t \times \beta^e$$

- Sendo:
  - $d_{i's}$: dígitos da parte fracionária, $d_1 \neq 0$, $0 \leq d_i \leq \beta\text{-}1$.
  - β: base (em geral 2, 10 ou 16).
  - t: número de dígitos na mantissa.
  - e: expoente inteiro.

Exemplos

- x=34,2 (decimal)
- β=10
- t=4

$$x = 0{,}3420 \times 10^2$$

- x=0,1 (decimal)
- β=2
- t=9

$$x = 0{,}110011001 \times 2^{-3}$$

Obs: $0{,}1_{10} = 0{,}0001100110011\ldots_2$

Representação em Ponto Flutuante

UNIVERSIDADE FEDERAL DO CEARÁ

- Nas máquinas digitais, **um digito binário é denominado BIT** (do inglês, binary digit).
- Um grupo de **oito bits** corresponde a **1 byte**.
- Dessa forma, percebemos que a representação dos números binários num computador é feita com um **número finito de bits**.
- A esse tamanho finito de bits é dado o nome **palavra de computador**.
- O tamanho da palavra do computador **depende de características internas** à arquitetura do mesmo.

Teoria dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

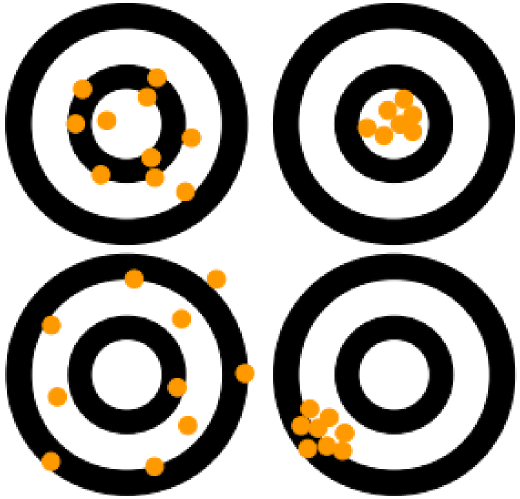
- Na busca de uma solução do modelo matemático por meio do calculo numérico **temos o surgimento de erros através de diversas fontes**.
- Além disso, **toda medida experimental apresenta uma incerteza** e desta forma a solução do problema pode ser influenciada pela mesma.
- Como consequência, **métodos numéricos podem chegar a resultados distantes do que se esperaria** ou mesmo fornecer **respostas sem nenhuma relação com a solução** do problema original.

Teoria dos Erros

- Precisão x Exatidão
  - Conceitos erroneamente tratados como sinônimos no cotidiano.
    - **Exatidão:** Grau de concordância entre o resultado de uma medição e um valor verdadeiro mensurado.
    - **Precisão:** Grau de concordância entre resultados de medição obtidos sob as mesmas condições.

- preciso
+ exato

+ preciso
+ exato



- preciso
- exato

+ preciso
- exato

Teoria dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- Agora analisamos uma aplicação prática desses conceitos em números.
  - Exatidão
    - É governada pelos **erros no método numérico empregado**. Assim, se os números π_1 =3,1416304958 e π_2 =3,1415809485 almejam ambos a representar o número π=3,141592654 . . . , o número π_2 possui maior exatidão que π_1 , embora ambos possuam a mesma precisão.
  - Precisão
    - A precisão de um número é governada pelo **número de dígitos empregados** na representação e na álgebra. Assim, a constante π será representada com maior precisão utilizando 8 bytes do que utilizando 4 bytes para armazenar o número.

Teoria dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- Para medir a acurácia de um número podemos utilizar como ferramentas de cálculo o conceito de erros absoluto e relativo.
  - **Erro Absoluto:** Diferença entre o valor exato de um número e o seu valor aproximado.

$$EA_x = \Delta x = \left|x_{exato} - x_{aprox}\right| = |x - \bar{x}|$$

  - **Erro Relativo:** Trata-se do erro absoluto dividido pelo valor verdadeiro.

$$ER_x = \frac{EA_x}{x} = \frac{\Delta x}{x} = \frac{|x - \bar{x}|}{x}$$

Teoria dos Erros

- Dado um número x já na forma normalizada que não possua representação exata no sistema $F[b, n, e_{min}, e_{max}]$. Pode-se escrever x como:

$$x = (0, d_1 d_2 \ldots d_n) \times b^e + gx \times b^{e-n}, com\ 0 \leq g_x < 1$$

  - Onde g_x é a parcela que não pode ser incluída em sua representação.

- Dessa forma, existem duas formas de realizarmos a aproximação.
  - Truncamento
  - Arredondamento

Teoria dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- Truncamento
  - Consistem em simplesmente ignorar o gx. Assim,

$$\bar{x} = (\mathbf{0}, \boldsymbol{d_1 d_2} \dots \boldsymbol{d_n}) \times \boldsymbol{b^e}$$

- Arredondamento
  - No arredondamento, executa-se a seguinte operação,

$$\bar{x} = \begin{cases} (\mathbf{0}, \boldsymbol{d_1 d_2} \dots \boldsymbol{d_n}) \times \boldsymbol{b^e}, se\ |g_x| < 1/2 \\ (\mathbf{0}, \boldsymbol{d_1 d_2} \dots (\boldsymbol{d_n} + \mathbf{1})) \times \boldsymbol{b^e}, se\ |g_x| \geq 1/2 \end{cases}$$

Teoria dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- Dado um sistema em ponto flutuante fictício com 4 algarismos na mantissa e base 10.
- Ex. $x = 0{,}037 \times 10^4$, $y = 0{,}1272 \times 10^2$, calcule x+y.

No procedimento da adição em ponto flutuante devemos alinhar as casas decimais de ambos os números igualando os expoentes ao maior expoente presente na soma.

Teoria dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- Dado uma sequência de operações algébricas é importante observar **como o erro se propaga** ao logo destas operações consecutivas.

- Em um sistema de representação em ponto flutuante qualquer, **a soma de dois números exatos fornecerá um resultado exato**?

Teoria dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- $x_1 = 0{,}3491 \times 10^4$
- $x_2 = 0{,}2345 \times 10^0$

- $(x_2 + x_1) - x_1$

$$= (0{,}2345 \times 10^0 + 0{,}3491 \times 10^4) - 0{,}3491 \times 10^4$$

$$= 0{,}3491 \times 10^4 - 0{,}3491 \times 10^4$$

$$= 0{,}0000$$

- $x_1 = 0{,}3491 \times 10^4$
- $x_2 = 0{,}2345 \times 10^0$

- $x_2 + (x_1 - x_1)$

$$= 0{,}2345 \times 10^0 + (0{,}3491 \times 10^4 - 0{,}3491 \times 10^4)$$

$$= 0{,}2345 \times 10^0 - 0{,}0000$$

$$= 0{,}2345$$

Análise dos Erros

- Adição
  - Erro Absoluto

$$x + y = (\bar{x} + EA_x) + \left(\bar{y} + EA_y\right) = (\bar{x} + \bar{y}) + \left(EA_x + EA_y\right)$$

$$EA_{x+y} = EA_x + EA_y$$

Análise dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- Adição
  - Erro Relativo

$$ER_{x+y} = \frac{EA_{x+y}}{x+y} = \frac{EA_x}{x} \cdot \frac{x}{x+y} + \frac{EA_y}{y} \cdot \frac{y}{x+y}$$

$$ER_{x+y} = ER_x \cdot \frac{x}{x+y} + ER_y \cdot \frac{y}{x+y}$$

$$ER_{x+y} = \frac{ER_x}{1+\frac{y}{x}} + \frac{ER_y}{1+\frac{x}{y}}$$

Análise dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- Adição
  - Erro Relativo
    - x >> y

$$ER_{x+y} = \frac{ER_x}{1 + \frac{y}{x}} + \frac{ER_y}{1 + \frac{x}{y}}$$

$$ER_{x+y} \approx ER_x$$

Análise dos Erros

- Adição
  - Erro Relativo
    - x << y

$$ER_{x+y} = \frac{ER_x}{1+\frac{y}{x}} + \frac{ER_y}{1+\frac{x}{y}}$$

$$ER_{x+y} \approx ER_y$$

Análise dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- Adição
  - Erro Relativo
    - x ≈ y

$$ER_{x+y} = \frac{ER_x}{1 + \frac{y}{x}} + \frac{ER_y}{1 + \frac{x}{y}}$$

$$ER_{x+y} \approx \frac{1}{2}(ER_x + ER_y)$$

Análise dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- Adição
  - Erro Relativo

$$ER_{x+y} \approx Max\ (ER_x, ER_y)$$

Análise dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- Subtração
  - Erro Absoluto

$$x - y = (\bar{x} + EA_x) - (\bar{y} + EA_y) = (\bar{x} + \bar{y}) - (EA_x + EA_y)$$

$$EA_{x-y} = |EA_x - EA_y|$$

Análise dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- Subtração
  - Erro Relativo

$$ER_{x-y} = \left|\frac{EA_{x-y}}{x-y}\right| = \left|\frac{EA_x}{x} \cdot \frac{x}{x-y} + \frac{EA_y}{y} \cdot \frac{y}{x-y}\right|$$

$$ER_{x-y} = \left|ER_x \cdot \frac{x}{x-y} + ER_y \cdot \frac{y}{x-y}\right|$$

$$ER_{x-y} = \left|\frac{ER_x}{1-\frac{y}{x}} + \frac{ER_y}{1-\frac{x}{y}}\right|$$

Análise dos Erros

- Adição
  - Erro Relativo
    - x >> y

$$ER_{x-y} = \left| \frac{ER_x}{1 - \frac{y}{x}} + \frac{ER_y}{1 - \frac{x}{y}} \right|$$

$$ER_{x-y} \approx ER_x$$

Análise dos Erros

UNIVERSIDADE FEDERAL DO CEARÁ

- Adição
  - Erro Relativo
    - x << y

$$ER_{x-y} = \left| \frac{ER_x}{1 - \frac{y}{x}} + \frac{ER_y}{1 - \frac{x}{y}} \right|$$

$$ER_{x-y} \approx ER_y$$

Análise dos Erros

- Adição
  - Erro Relativo
    - x ≈ y

$$ER_{x-y} = \left| \frac{ER_x}{1 - \frac{y}{x}} + \frac{ER_y}{1 - \frac{x}{y}} \right|$$

$$Se\ \mathbf{1} - \frac{\boldsymbol{y}}{\boldsymbol{x}} \ll \mathbf{1}\ e\ \mathbf{1} - \frac{\boldsymbol{x}}{\boldsymbol{y}} \ll \mathbf{1}$$

$$ER_{x+y} \gg \ (ER_x + ER_y)$$

Este resultado mostra claramente como o erro relativo pode se tornar muito grande quando **X ≈ Y** . Isto ocorre porque a subtração de dois números muito próximos entre si resulta em um numero cuja representação ocorre nos últimos dígitos da mantissa, resultando em um grande erro de arredondamento.

Análise dos Erros

- Desafio
  - Encontre as expressões de erros para a multiplicação e para a divisão.

Exercícios

UNIVERSIDADE FEDERAL DO CEARÁ

- Seja um sistema que opera em aritmética de ponto flutuante de t = 4, na base 10, calcule os erros absolutos e relativos por truncamento e arredondamento:
  a) 123,456
  b) 374,3 + 3,345
  c) 124,34 + 0,1234
  d) 22,12 x 0,123
  e) 0,3212 x 12,32