Parte 1

CÁLCULO NUMÉRICO

Prof. Maria Clara Schuwartz Ferreira
mclarascferreira@gmail.com

INSTITUTO FEDERAL
Espírito Santo

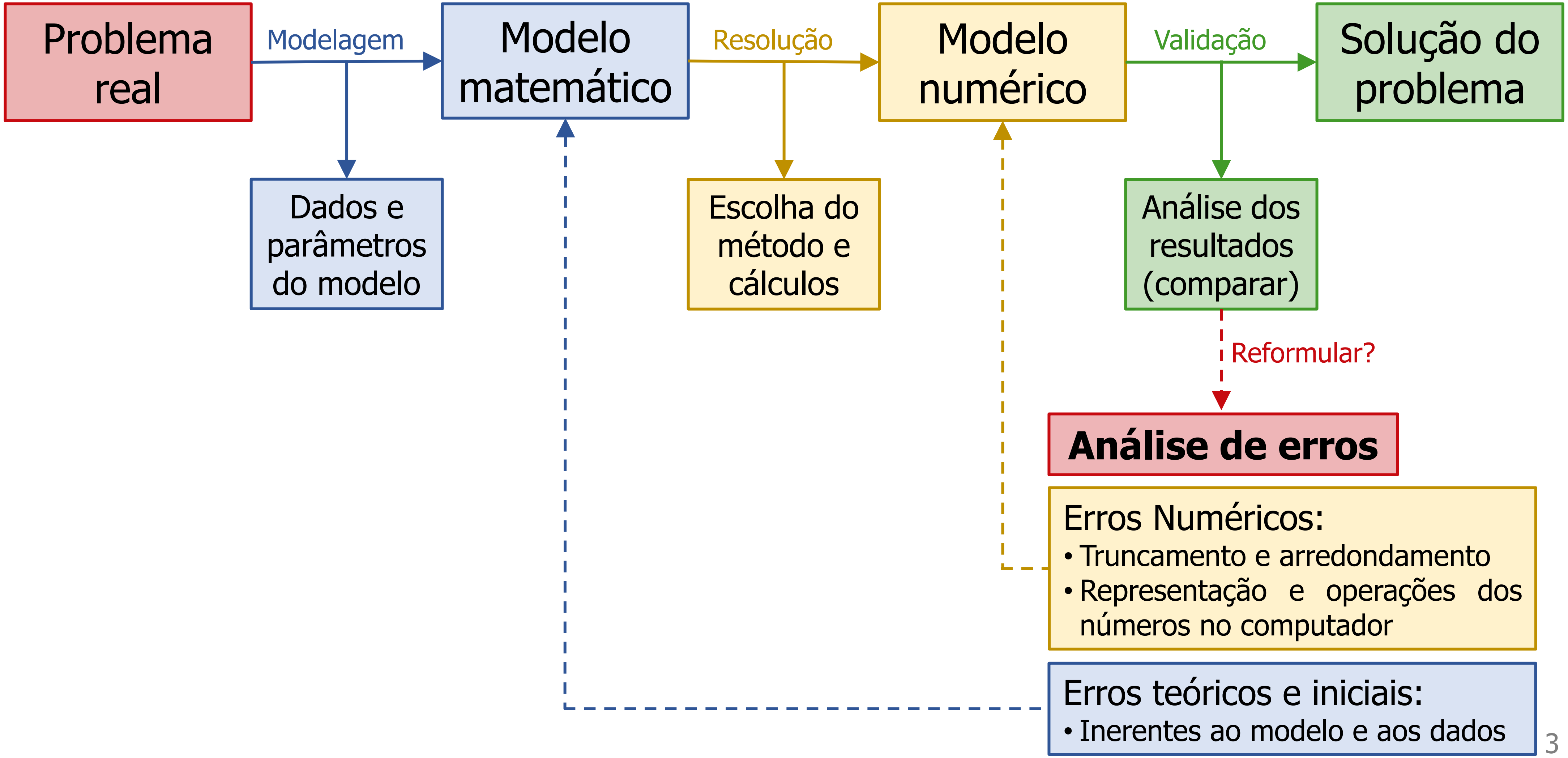
INSTITUTO FEDERAL
Espírito Santo

Conteúdo da Semana 1

1. Plano de Ensino
2. Introdução
3. Representação numérica
4. Conversão entre bases
5. Aritmética de ponto flutuante
6. Análise de erros

Introdução

- Etapas na solução de um problema



Conteúdo da Semana 1

1. Plano de Ensino
2. Introdução
3. Representação numérica
4. Conversão entre bases
5. Aritmética de ponto flutuante
6. Análise de erros

- Motivação
- Decomposição numérica
- Alguns sistemas numéricos

Representação numérica

Representação numérica

- Sistema numérico indo-arábico: representação posicional é importante!

(Ouça o Podcast "A História dos Algarismos (SciCast #306)")

- Decomposição do número
  - Base 10

$$(2173)_{10} = 2.10^3 + 1.10^2 + 7.10^1 + 3.10^0$$

$$(2{,}173)_{10} = 2.10^0 + 1.10^{-1} + 7.10^{-2} + 3.10^{-3}$$

- Em outras bases funciona de maneira similar:
  - Base 2

$$(10111)_2 = 1.2^4 + 0.2^3 + 1.2^2 + 1.2^1 + 1.2^0 = 23$$

  - Base 4

$$(21{,}3)_4 = 2.4^1 + 1.4^0 + 3.4^{-1} = 9{,}75$$

Representação numérica

Decimal	Binário	Octal	Hexadecimal
0	0	0	0
1	1	1	1
2	10	2	2
3	11	3	3
4	100	4	4
5	101	5	5
6	110	6	6
7	111	7	7
8	1000	10	8
9	1001	11	9
10	1010	12	A
11	1011	13	B
12	1100	14	C
13	1101	15	D
14	1110	16	E
15	1111	17	F
16	10000	20	10
⋮	⋮	⋮	⋮

Conteúdo da Semana 1

1. Plano de Ensino
2. Introdução
3. Representação numérica
4. Conversão entre bases
5. Aritmética de ponto flutuante
6. Análise de erros

- Conversão para a base decimal
- Decimal para binária
- Decimal para outras bases
- Bases que são potências entre si
- Resumo

Conversão entre bases

- Em uma base β qualquer:

$$\left(a_j a_{j-1} \cdots a_1 a_0\right)_\beta = a_j \beta^j + a_{j-1}\beta^{j-1} + \cdots + a_1\beta^1 + a_0\beta^0$$

com $0 \leq a_k \leq \beta - 1$.

Exemplo: Uma caixa alienígena com o número 25 gravado na tampa foi entregue a um grupo de cientistas. Ao abrirem a caixa, encontraram 17 objetos. Considerando que o alienígena tem um formato humanóide, quantos dedos ele tem nas duas mãos?

Solução :

$(25)_b \quad = (17)_{10}$

$2.b^1 + 5.b^0 = 17$

$b = (17 - 5)/2 = 6$

Conversão entre bases

- Exemplo: Um sistema ternário tem "trits", cada "trit" assumindo o valor 0, 1 ou 2. Quantos "trits" são necessários para representar um número de seis bits?

Solução :

Maior número que pode ser representado com 6 bits = 63

$$2.\left(3^{n} + 3^{n-1} + \cdots + 3^{0}\right) \;\geq\; 63$$

$$2.\left(\frac{3^{n+1}-1}{3-1}\right) \geq 63$$

$3^{n+1} \geq 64$ (64 está entre 3^3 e 3^4)

$n = 3$

Conversão entre bases

Exemplo: Vimos que $(10111)_2 = 1.2^4 + 0.2^3 + 1.2^2 + 1.2^1 + 1.2^0 = 23$

Conversão da base binária para base decimal ✓

Como converter da base decimal para a base binária?

$$
\begin{aligned}
23 &= 1 + 2(1.2^3 + 0.2^2 + 1.2^1 + 1) \\
&= 1 + 2\left(1 + 2(1.2^2 + 0.2^1 + 1)\right) \\
&= 1 + 2\left(1 + 2\left(1 + 2(1.2 + 0)\right)\right) \\
&= 1 + 2\left(1 + 2\left(1 + 2(0 + 2(1))\right)\right) \\
&= (10111)_2
\end{aligned}
$$

Exemplo: $(40)_{10} = (?)_2$

$$= (101000)_2$$

Conversão entre bases

- Número fracionário:
  - Para converter um número com parte inteira e parte fracionária, fazer a conversão de cada parte, separadamente.

$$\underbrace{a_n b^n + a_{n-1} b^{n-1} + \ldots + a_0 b^0}_{\textbf{Parte } \textbf{\textit{Inteira}}} + \underbrace{a_{-1} b^{-1} + a_{-2} b^{-2} + \ldots + a_{-m} b^{-m}}_{\textbf{Parte } \textbf{\textit{Fracionária}}}$$

- Exemplo: Converter 0,625 decimal para binário

$$0{,}625 = b_1 . 2^{-1} + b_2 . 2^{-2} + b_3 . 2^{-3} + \cdots \xrightarrow{\times 2} 1{,}25 = b_1 + b_2 . 2^{-1} + b_3 . 2^{-2} + \cdots \quad \boxed{\rightarrow \boldsymbol{b_1 = 1}}$$

$$0{,}25 = b_2 . 2^{-1} + b_3 . 2^{-2} + \cdots \xrightarrow{\times 2} 0{,}5 = b_2 + b_3 . 2^{-1} + \cdots \quad \boxed{\rightarrow \boldsymbol{b_2 = 0}} \quad \boxed{\rightarrow \boldsymbol{b_3 = 1}}$$

Resultado: $(0{,}625)_{10} = (0{,}\mathbf{101})_2$

Conversão entre bases

- Exemplo: $(8{,}375)_{10} = (\ ?\ )_2$

0,375	0,750	0,500	0,000
x 2	x 2	x 2	
0,750	1,500	1,000	

$$\boxed{8{,}375_{10} = 1000{,}011_2}$$

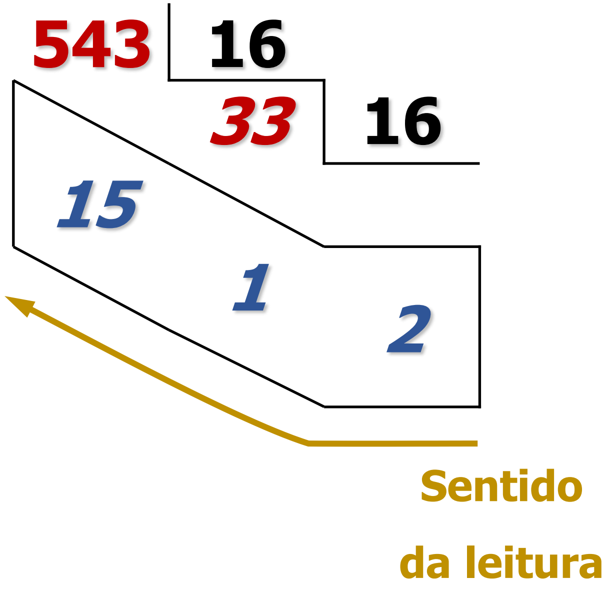
- Exemplo: $(0{,}11)_{10} = (?)_2$

$0{,}22 \rightarrow 0{,}44 \rightarrow 0{,}88 \rightarrow 1{,}76 \rightarrow 1{,}52 \rightarrow 1{,}04 \rightarrow 0{,}08 \rightarrow 0{,}16 \rightarrow 0{,}32 \rightarrow 0{,}64 \rightarrow 1{,}28 \rightarrow 0{,}56 \rightarrow$
$1{,}12 \rightarrow 0{,}24 \rightarrow 0{,}48 \rightarrow 0{,}96 \rightarrow 1{,}92 \rightarrow 1{,}84 \rightarrow 1{,}68 \rightarrow 1{,}36 \rightarrow 0{,}72 \rightarrow 1{,}44 \rightarrow 0{,}88 \rightarrow$ repetir

$$(0{,}11)_{10} = (0{,}00011100001010001111010\overline{11100001010001111101})_2$$

Conversão entre bases

- Exemplo: Decimal para hexadecimal



O resto *15* é representado pela letra *F*

$(543)_{10} = (21F)_{16}$

Conversão entre bases

- Binário para octal

Agrupa-se o número binário de 3 em 3 dígitos. $(2^3 = 8)$

- Octal para binário

Utiliza-se um procedimento inverso.

Analogamente...

- Binário para hexadecimal

Agrupa-se o número binário de 4 em 4 dígitos. $(2^4 = 16)$

Hexadecimal para binário

Utiliza-se um procedimento inverso.

Conversão entre bases

- Exemplo: Binário e hexadecimal

$0110110101001010001100010101 0011_2$

$(6\ D\ 4\ A\ 3\ 1\ 5\ 3)_{16}$

Binário	Hexadecimal
0	0
1	1
10	2
11	3
100	4
101	5
110	6
111	7
1000	8
1001	9
1010	A
1011	B
1100	C
1101	D
1110	E
1111	F
10000	10
⋮	⋮

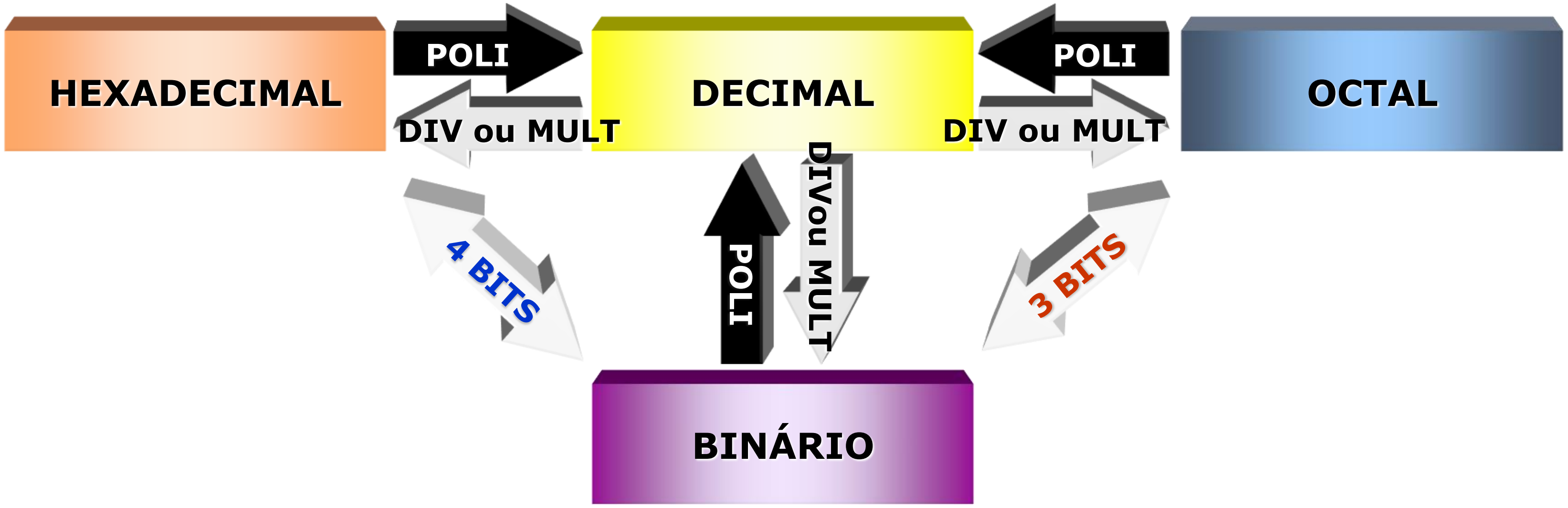
- Exemplo: $(1001001000,1011011)_2 = (?)_{16}$

$(001001001000,10110110)_2 = (?)_{16}$ $\boxed{= (248,B6)_{16}}$

10 = 2, 0100 = 4, 1000 = 8, 1011 = B, 0110 = 6

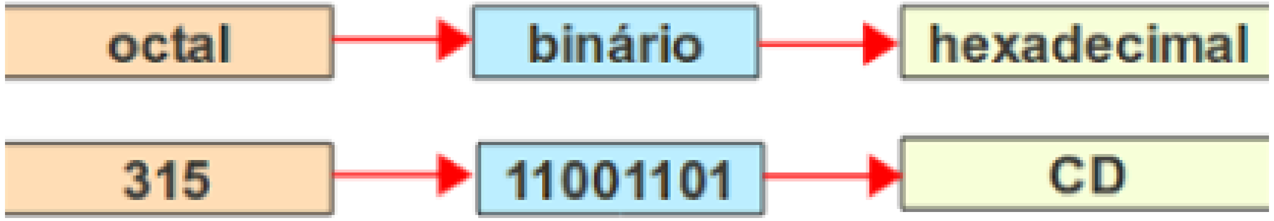
Conversão entre bases

- Divisões sucessivas (parte inteira) e multiplicações sucessivas (parte decimal) pela base do sistema para o qual se deseja converter o número
- Decomposição polinomial do número a ser convertido
- Agrupamento de bits



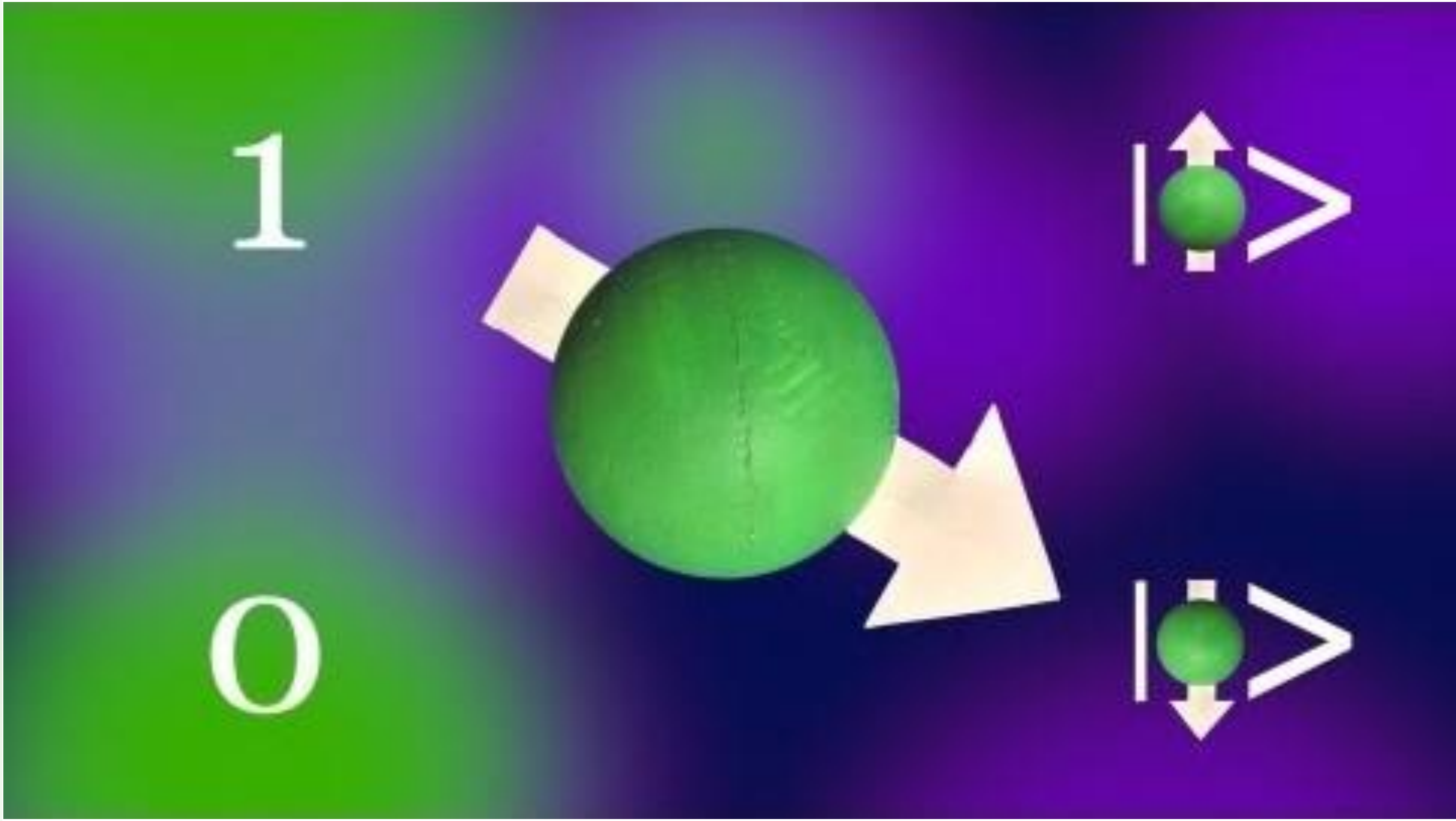
Conversão entre bases

- Exemplo: Octal e Hexadecimal
  - Não é direta, pois não há relação de potências entre as bases 8 e 16.



Existem 10 tipos de pessoas no mundo: as que entendem binário e as que não entendem.

Computadores QUÂNTICOS



Parte 1 – Fim

CÁLCULO NUMÉRICO

Prof. Maria Clara Schuwartz Ferreira
mclarascferreira@gmail.com

INSTITUTO FEDERAL
Espírito Santo

Parte 2

CÁLCULO NUMÉRICO

Prof. Maria Clara Schuwartz Ferreira
mclarascferreira@gmail.com

INSTITUTO FEDERAL
Espírito Santo

INSTITUTO FEDERAL
Espírito Santo

Conteúdo da Semana 1

1. Plano de Ensino
2. Introdução
3. Representação numérica
4. Conversão entre bases
5. Aritmética de ponto flutuante
6. Análise de erros

- Representação
- Gama representável
- Erros na representação
- Amplitude e precisão
- Normalização
- Operações aritméticas

Aritmética de ponto flutuante

- A área de memória reservada para armazenar um número em um computador (chamada palavra) é geralmente divida em 3:

Sinal	**Expoente**	**Mantissa**
+ ou −	nº inteiro	parte fracionária

- Um número x em um sistema de Ponto Flutuante $FP(\beta, n, e_1, e_2)$ é:

$$x = \pm(0, d_1 d_2 \cdots d_n) \, . \, \beta^e$$

β é a base;

d_i com $0 \leq d_i \leq \beta - 1$ são os dígitos da parte fracionária (mantissa), $d_1 \neq 0$;

n é o número de dígitos na mantissa;

e é um expoente inteiro.

e_1 e e_2 são o menor e o maior expoente possível, respectivamente, para o sistema.

INSTITUTO FEDERAL
Espírito Santo

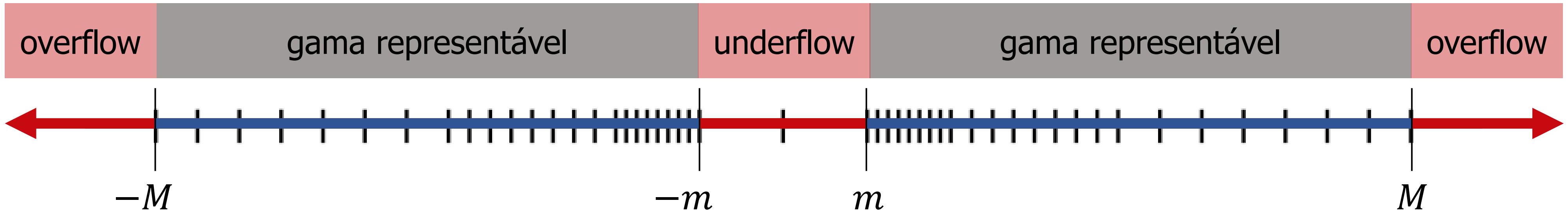
Aritmética de ponto flutuante

- Sistema de pontos flutuantes $FP(\beta, n, e_1, e_2)$ é:

$$x = \pm(0, d_1 d_2 \cdots d_n)\ .\ \beta^e$$

Menor número em valor absoluto m: $\boxed{m = 0{,}10 \cdots 0\ .\ \beta^{e_1}}$

Maior número em valor absoluto M: $\boxed{M = 0, aa \cdots a\ .\ \beta^{e_2}}$ onde $a = \beta - 1$



Aritmética de ponto flutuante

- $FP(\beta, n, e_1, e_2)$ é:

$$x = \pm(0, d_1 d_2 \cdots d_n) \,.\, \beta^e$$

E se a mantissa for $d_1 d_2 \cdots d_n d_{n+1} \cdots$?

Truncamento

Ignoram-se algarismos a partir do índice $n + 1$

Arredondamento

Se $0, d_{n+1} \cdots < 0{,}5 \Rightarrow$ mantém-se o algarismo da posição n (arred. para baixo)

Se $0, d_{n+1} \cdots > 0{,}5 \Rightarrow$ soma-se uma unidade ao algarismo da posição n (arred. para cima)

Se $0, d_{n+1} \cdots = 0{,}5 \Rightarrow$ arred. para cima ou para baixo ficando o algarismo da posição n par.

INSTITUTO FEDERAL
Espírito Santo

Aritmética de ponto flutuante

- Exemplo: Considere o sistema $FP(10, 3, -2,2)$. Represente neste sistema:

a) $0,35 \ = 0,350.10^{0}$

b) $0,0123 = 0,123.10^{-1}$

c) $5,175 = 0,518.10^{1}$

d) $5391,3$ overflow!

e) $0,0003$ underflow!

Aritmética de ponto flutuante

Exemplo: Quantos números podem ser representados de forma exata no sistema $FP(2,3,-1,2)$?

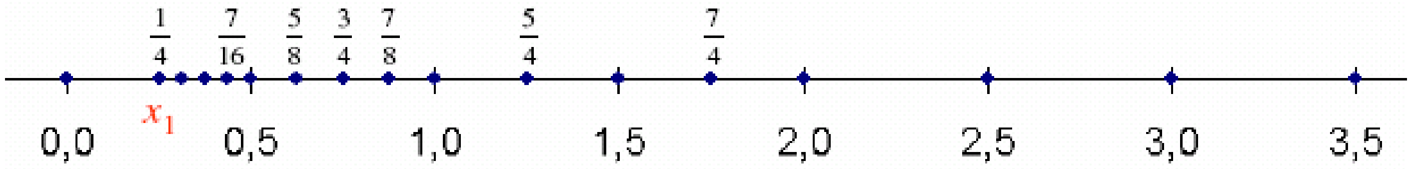
16 números positivos

16 números negativos

Número 0

Total: 33 números!

		Mantissas			
e	b^e	0.100	0.101	0.110	0.111
-1	½	1/4	5/16	3/8	7/16
0	1	1/2	5/8	3/4	7/8
1	2	1	5/4	3/2	7/4
2	4	2	5/2	3	7/2



Em um sistema qualquer $FP(\beta, n, e_1, e_2)$:

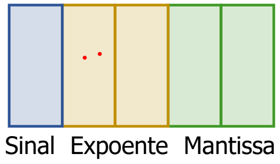
$$\boxed{N = 2.(\beta - 1).\beta^{n-1}.(e_2 - e_1 + 1) + 1}$$

INSTITUTO FEDERAL
Espírito Santo

Aritmética de ponto flutuante

- Exemplo: Computador hipotético de base 10 com tamanho de palavra de 5 dígitos com 1 dígito para o sinal, 2 para o expoente (um deles utilizado para seu sinal) e 2 para a mantissa.

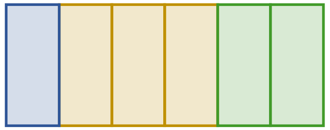
$\Rightarrow FP(10, 2, -9{,}9)$



$$\boxed{m = 0{,}10\,.\,10^{-9}}$$

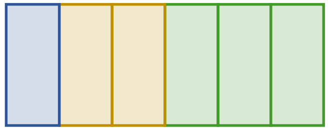
$$\boxed{M = 0{,}99\,.\,10^{9}}$$

E se o tamanho da palavra fosse 6?



EXPOENTE $\boxed{m = 0{,}10.\,10^{-99} \text{ e } M = 0{,}99.\,10^{99}}$

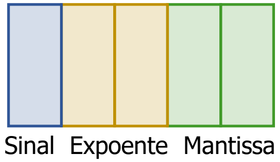
$\Rightarrow$ Maior ganho na amplitude!
(muita diferença em m e M)



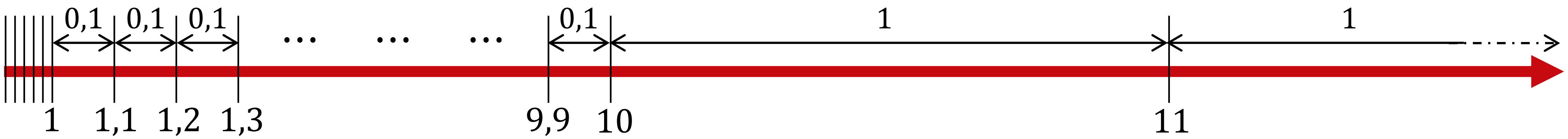
MANTISSA $\boxed{m = 0{,}100.\,10^{-9} \text{ e } M = 0{,}999.\,10^{9}}$

$\Rightarrow$ Ganho de precisão!
(menos arredondamento e truncamento)

Aritmética de ponto flutuante



- A distribuição dos números na reta real não é uniforme



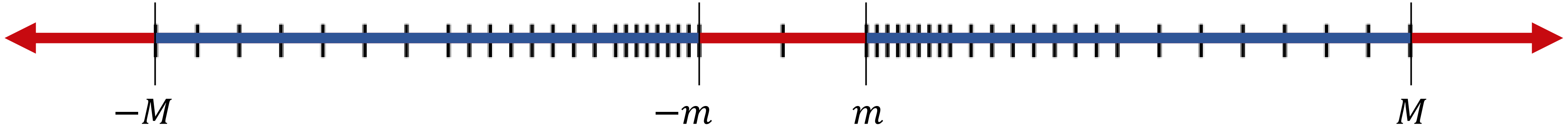
- No exemplo anterior:

Números entre 0,1 e 1: espaçamento de 0,01

Números entre 1 e 10: espaçamento de 0,1

Números entre 10 e 100: espaçamento de 1

O erro por arredondamento é proporcional ao módulo do número

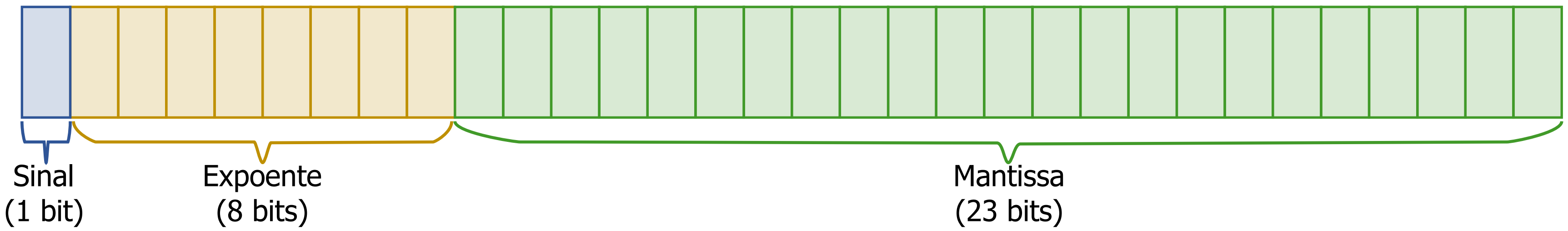


Aritmética de ponto flutuante

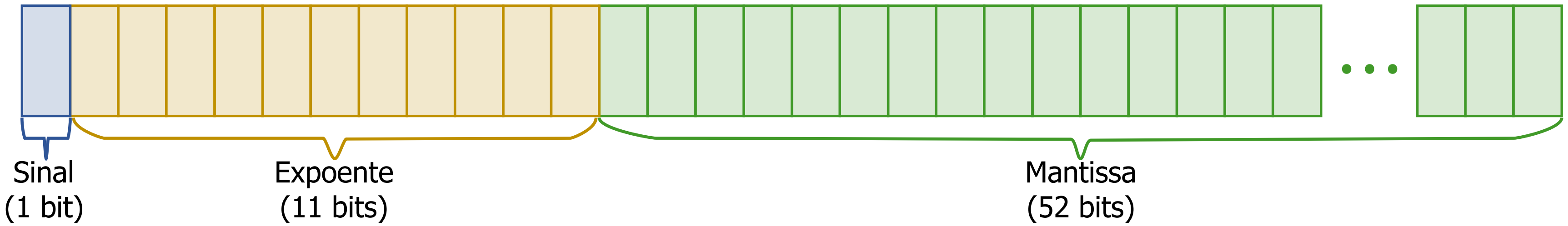
- Padrão IEEE 754

- Os computadores atuais utilizam base binária com tamanho de palavra:

- de 32 bits (precisão simples) ou



- de 64 bits (precisão dupla):



Aritmética de ponto flutuante

- Soma e subtração:
  - Alinhar na maior base entre os números (os dígitos do número de menor expoente devem ser deslocados).
- Exemplo:

$$4{,}32 + 0{,}064 = .43 \times 10^{1} + .64 \times 10^{-1} = \begin{array}{lll} & .43 & \times 10^{1} \\ + & .0064 & \times 10^{1} \\ = & .4364 & \times 10^{1} \\ \rightarrow & .44 & \times 10^{1}. \end{array}$$

  - Resultado foi 4,4 em vez de 4,383!
- Após a operação, normalizar e arredondar (se necessário) o número.

Aritmética de ponto flutuante

- As operações não são associativas, nem comutativas e nem distributivas
- Exemplo: $x_1=0{,}3491\times10^4$, $x_2= 0{,}2345\times10^0$
- $(x_2 + x_1) - x_1 \stackrel{?}{=} x_2 + (x_1 - x_1)$

Máquina com mantissa de tamanho 7:

- $(x_2+x_1) - x_1 =(0{,}0000234\times10^4+0{,}3491\times10^4) -0{,}3491\times10^4 =$ $(0{,}3491234\times10^4) - 0{,}3491\times10^4 = (0{,}0000234\times10^4) = 0{,}234\times10^0$
- $x_2+(x_1 - x_1)=0{,}2345\times10^0 +(0{,}3491\times10^4 -0{,}3491\times10^4) = 0{,}2345\times10^0 + 0 =$ $0{,}2345\times10^0$

$\Rightarrow$ Cuidado com o cancelamento aditivo! ($a + b$ com $a \ll b$ ou $b \ll a$)

Aritmética de ponto flutuante

- Cuidado com o cancelamento subtrativo! ($a - b$ com a $\approx b$)
- Exemplo: $372 - 371 = .37 \times 10^3 - .37 \times 10^3 =$

$$\begin{array}{rl} & .37 \times 10^3 \\ - & .37 \times 10^3 \\ = & .00 \times 10^3 \\ \rightarrow & .00 \times 10^0. \end{array}$$

- Podem-se perder algarismos significativos
- Pode conduzir a erros elevados
- Possível minorar rearranjando cálculos

Aritmética de ponto flutuante

- Multiplicação e divisão:
  - Multiplicar (ou dividir) os significandos (a máquina utiliza $2n$ dígitos) e adicionar (ou subtrair) os expoentes para em seguida o resultado ser normalizado e arredondado, se necessário.
- Exemplo: $1234 \times 0{,}016 = .12.10^4 \times .16.10^{-1} =$

$$\begin{array}{lll} & .12 & .10^4 \\ \times & .16 & .10^{-1} \\ = & .0192 & .10^3 \\ \rightarrow & .19 & .10^2. \end{array}$$

  - Resultado foi 19 em vez de 19,744!

INSTITUTO FEDERAL
Espírito Santo

Conteúdo da Semana 1

1. Plano de Ensino
2. Introdução
3. Representação numérica
4. Conversão entre bases
5. Aritmética de ponto flutuante
6. Análise de erros

- Tipos de erros
- Exatidão e precisão
- Erros absolutos e relativos
- Propagação de erros
- Sistemas mal condicionados

Análise de erros

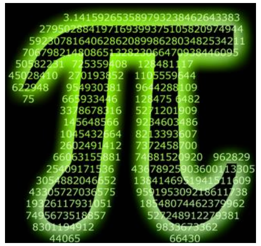
- Inerentes

  Erros na modelagem ou dados de entrada

- Erro de truncamento

$$e^x \cong \sum_{k=0}^{n-1} \frac{x^k}{k!}$$

- Erro de arredondamento



Análise de erros

- A precisão está associado ao número de dígitos com os quais uma grandeza é representada em uma máquina.

- A exatidão está associado a quantificar a aproximação do valor computado com o valor real (medida de perfeição).

Análise de erros

- Erro absoluto (EA) – Medir quão próximo uma quantidade $\bar{x}$ está do valor exato x, ou seja, medir a exatidão de uma quantidade.

$$\boxed{|EA| = |x - \bar{x}|}$$

- Erro relativo (ER) – É o quociente entre o erro absoluto é o valor real $(x \neq 0)$

$$\boxed{|ER| = \frac{|x - \bar{x}|}{x}}$$

- Limitante ε – estimativa para o módulo do erro absoluto $\boxed{|EA| \leq \varepsilon}$

Notação: $x = \bar{x} \pm \varepsilon \;\Rightarrow\; x \in (\bar{x} - \varepsilon, \bar{x} + \varepsilon)$

$$\pi = 3.14 \pm 0.002 \quad \Leftrightarrow \quad \pi \in [3.138, 3.142]$$

Análise de erros

- Propagação do erro na adição e subtração

Sejam ε_x e ε_y erros absolutos máximo de x e y, respectivamente. Então:

$$\begin{matrix} x = \bar{x} \pm \varepsilon_x \\ y = \bar{y} \pm \varepsilon_y \end{matrix} \Rightarrow \quad x + y = \bar{x} + \bar{y} \pm (\varepsilon_x + \varepsilon_y) \Rightarrow \boxed{\varepsilon_{x+y} = \varepsilon_x + \varepsilon_y}$$

$$\boxed{|EA_{x+y}| \leq \varepsilon_x + \varepsilon_y}$$

$$\boxed{|ER_{x+y}| \leq \frac{\varepsilon_x + \varepsilon_y}{|x + y|}}$$

- Analogamente para a subtração: $\boxed{\varepsilon_{x-y} = \varepsilon_x + \varepsilon_y}$

$$\boxed{|EA_{x-y}| \leq \varepsilon_x + \varepsilon_y}$$

$$\boxed{|ER_{x-y}| \leq \frac{\varepsilon_x + \varepsilon_y}{|x - y|}}$$

Análise de erros

- Propagação do erro na multiplicação

Sejam ε_x e ε_y erros absolutos máximo de x e y, respectivamente. Então:

$$\begin{aligned} x &= \bar{x} \pm \varepsilon_x \\ y &= \bar{y} \pm \varepsilon_y \end{aligned} \quad \Rightarrow \quad x.y = (\bar{x} \pm \varepsilon_x)(\bar{y} \pm \varepsilon_y)$$

$$\Rightarrow \quad x.y = \bar{x}\bar{y} \pm (\bar{x}\varepsilon_y + \bar{y}\varepsilon_x + \varepsilon_x\varepsilon_y)$$

Considerando $\varepsilon_x\varepsilon_y$ desprezível, temos:

$$\varepsilon_{xy} = \bar{x}\varepsilon_y + \bar{y}\varepsilon_x$$

$$|EA_{xy}| \leq \bar{x}\varepsilon_y + \bar{y}\varepsilon_x$$

$$\boxed{|ER_{xy}| \leq \frac{\varepsilon_x}{|\bar{x}|} + \frac{\varepsilon_y}{|\bar{y}|}}$$

Análise de erros

- Propagação do erro na divisão:

$$x = \bar{x} \pm \varepsilon_x \qquad\qquad y = \bar{y} \pm \varepsilon_y \qquad \Rightarrow \qquad \frac{x}{y} = \frac{\bar{x} \pm \varepsilon_x}{\bar{y} \pm \varepsilon_y} = \frac{\bar{x} \pm \varepsilon_x}{\bar{y}} \frac{1}{1 \pm \frac{\varepsilon_y}{\bar{y}}}$$

Mas $\frac{1}{1 \pm \frac{\varepsilon_y}{\bar{y}}} = 1 \pm \frac{\varepsilon_y}{\bar{y}} \pm \left(\frac{\varepsilon_y}{\bar{y}}\right)^2 \pm \left(\frac{\varepsilon_y}{\bar{y}}\right)^3 \pm \cdots$ (Soma dos termos de uma PG)

Desprezando os termos da série com expoente maior que 1, temos:

$$\varepsilon_{\frac{x}{y}} = \frac{\bar{x}\varepsilon_y + \bar{y}\varepsilon_x}{\bar{y}^2}$$

$$\left|EA_{\frac{x}{y}}\right| \leq \frac{\bar{x}\varepsilon_y + \bar{y}\varepsilon_x}{\bar{y}^2}$$

$$\boxed{\left|ER_{\frac{x}{y}}\right| \leq \frac{\varepsilon_x}{|\bar{x}|} + \frac{\varepsilon_y}{|\bar{y}|}}$$

Análise de erros

- Sistemas mal condicionados: pequenas perturbações nos dados originais podem induzir alterações expressivas no resultado.
- Sistemas bem condicionados: pequenas perturbações nos dados originais provocam alterações pouco significativas no resultado.

- Características de um algoritmo numérico de boa qualidade:
  - Inexistência de erro lógico;
  - Inexistência de erro operacional (erros de ”underflow” ou ”overflow”);
  - O algoritmo deve ter uma quantidade finita de cálculos (iterações);
  - Existência de um critério de exatidão (análise de erros);
  - Independência de máquina;
  - Precisão infinita: os limites do erro devem convergir a zero;
  - Eficiência: obter respostas corretas do problema no menor custo possível.

Parte 2 – Fim

CÁLCULO NUMÉRICO

Prof. Maria Clara Schuwartz Ferreira
mclarascferreira@gmail.com

INSTITUTO FEDERAL
Espírito Santo