

Engenharia de Controle e Automação



Cálculo Numérico Aula 3 - Noções Básicas sobre Erro



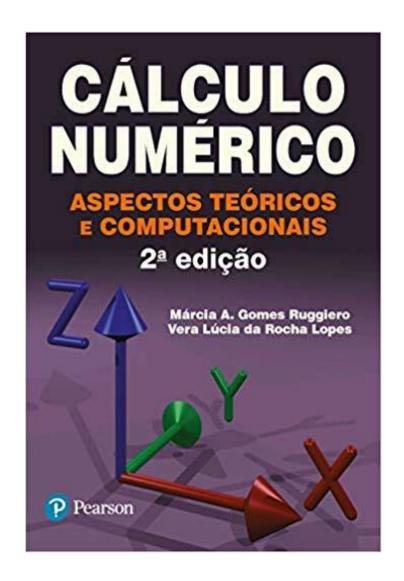
Prof. Hilário Tomaz Alves de Oliveira hilario.oliveira@ifes.edu.br

Agenda

- Representação de Números
 - Conversão de Números Nos Sistemas Decimal e Binário.
 - Aritmética de Ponto Flutuante.
- Erros
 - Erros Absolutos e Relativos.
 - Truncamento e arredondamento.

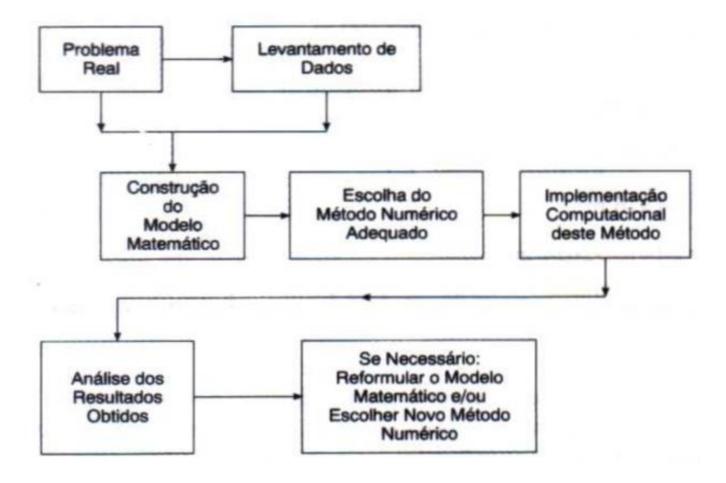
Leitura Recomendada

- Capítulo 1 Noções Básicas sobre Erros.
 - Seção 1.1;
 - Seção 1.2;
 - Seção 1.3.
 - Itens 1.3.2 e 1.3.3 (Fora).
- Lista de Exercícios 2 e Projeto de Implementação 1.
 - Disponíveis no Ambiente Virtual (Moodle).



Introdução

 A resolução de problemas envolvem várias fases que podem ser estruturadas da seguinte forma:



Introdução

- Muitas vezes os resultados finais estão distantes do que se espera obter.
 - Mesmo que todas as fases de resolução tenham sido realizadas corretamente.
- Os resultados obtidos dependem também:
 - Da precisão dos dados de entrada;
 - Da forma como estes dado são representados no computador;
 - Das operações numéricas efetuadas.

Representação dos Números

Representação dos Números – Exemplo 1

Calcular a área de uma circunferência de raio 100 metros.

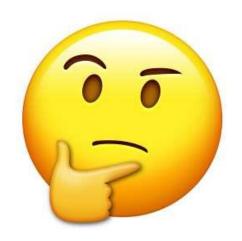
•
$$A = \pi * r^2$$

Resultados Obtidos:

a)
$$A = 31400 m^2$$

b)
$$A = 31416 m^2$$

c)
$$A = 31415,92654 m^2$$



Representação dos Números – Exemplo 2

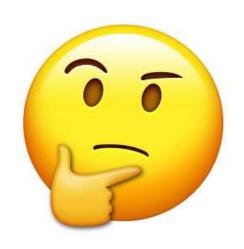
• Efetuar o seguinte somatório em uma calculadora e em um computador:

•
$$\sum_{i=1}^{30000} x_i$$

para
$$x_i = 0.5 \text{ e } x_i = 0.11$$

Resultados Obtidos:

- a) Para $x_i = 0.5$
 - Na calculadora \rightarrow S = 15000
 - No Computador \rightarrow S = 15000
- b) Para $x_i = 0.11$
 - Na calculadora \rightarrow S = 3300
 - No Computador \rightarrow S* = 3299,99691



Representação dos Números

No Exemplo 1:

- O número π não pode ser representado através de um número finito de dígitos decimais.
 - Π foi escrito como 3,14; 3,1416 e 3,141592654 respectivamente nos casos a), b) e c).
- No Exemplo 2:
 - A diferença dos resultados ocorre por conta da representação dos números na máquina utilizada.
 - $x_i = 0,11$ não possui representação finita na base 2.

- A base decimal é a que mais empregamos atualmente.
 - Antigamente, foram utilizadas outras bases, como a base 12, a base 60.
- Um computador opera no sistema binário (Base 2).
 - Entrada: Os dados estão em formato decimal;
 - Processamento:
 - A conversão para binário;
 - As operações são feitas na base 2;
 - Saída: Os dados são convertidos para decimal.

Considere os números:

•
$$(347)_{10} = 3 * 10^2 + 4 * 10^1 + 7 * 10^0$$

•
$$(10111)_2 = 1 * 2^4 + 0 * 2^3 + 1 * 2^2 + 1 * 2^1 + 1 * 2^0$$

- De modo geral, um número na base β .
 - $(a_j a_{j-1}, \dots, a_2, a_1, a_0)$, de forma que $0 \le a_k \le \beta 1$, $k = 1, 2, \dots, j$, pode ser escrito na forma polinomial:

•
$$a_i \beta^j + a_{i-1} \beta^{j-1} + \dots + a_2 \beta^2 + a_1 \beta^1 + a_0 \beta^0$$

Exemplo:

•
$$(10111)_2 = 1 * 2^4 + 0 * 2^3 + 1 * 2^2 + 1 * 2^1 + 1 * 2^0 \rightarrow (23)_{10}$$

Podemos colocar o número 2 em evidência:

•
$$(101111)_2 = 2 * (1 + 2 * (1 + 2 * (0 + 2 * 1))) + 1 = (23)_{10}$$

- Assim:
 - A representação de um número $(a_j a_{j-1}, \dots, a_2, a_1, a_0)$ na base 10, denotada por b_0 , é obtida através do seguinte processo:

- $b_j = a_j$
- $b_{j-1} = a_{j-1} + 2 * b_j$
- $b_{j-2} = a_{j-2} + 2 * b_{j-1}$

- $b_1 = a_1 + 2 * b_2$
- $b_0 = a_0 + 2 * b_1$

Para o número (10111)₂, a sequência obtida será:

•
$$b_4 = a_4 = 1$$

•
$$b_3 = a_3 + 2 * b_4 = 0 + 2 * 1 = 2$$

•
$$b_2 = a_2 + 2 * b_3 = 1 + 2 * 2 = 5$$

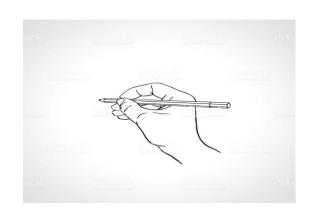
•
$$b_1 = a_1 + 2 * b_2 = 1 + 2 * 5 = 11$$

•
$$b_0 = a_0 + 2 * b_1 = 1 + 2 * 11 = 23$$

Exercício 1

1. Converta os seguintes números para a base decimal.

- a) $(110)_2$
- b) (1001)₂
- c) (1110011)₂



- Como fazer o processo inverso?
 - Converter um número inteiro representado na base 10 para o sistema binário?
- Considere o número (347)₁₀ e sua representação na base 2

$$(a_j a_{j-1} \dots a_1 a_0)_2$$

- $347 = 2 * (a_j * 2^{j-1} + a_{j-1} * 2^{j-2} + + a_2 * 2^1 + a_1) + a_0$
- 347 = 2 * 173 + 1 (resto da divisão entre 347 e 2)

• Repetimos este processo para o número $N_1 = 173$.

•
$$173 = 2 * (a_j * 2^{j-1} + a_{j-1} * 2^{j-2} + \dots + a_2 * 2^1) + a_1$$

- Dessa forma obteremos o digito a₁, que será o resto da divisão entre
 N₁ e 2.
- Seguindo este raciocínio obteremos a sequência de números N_j e a_j .

- Exemplo: $(347)_{10}$
 - $N_0 = 347 \rightarrow 2 * 173 + 1 \rightarrow a_0 = 1$
 - $N_1 = 173 \rightarrow 2 * 86 + 1 \rightarrow a_1 = 1$
 - $N_2 = 86 \rightarrow 2 * 43 + 0 \rightarrow a_2 = 0$
 - $N_3 = 43 \rightarrow 2 * 21 + 1 \rightarrow a_3 = 1$
 - $N_4 = 21 \rightarrow 2 * 10 + 1 \rightarrow a_4 = 1$
 - $N_5 = 10 \rightarrow 2 * 5 + 0 \rightarrow a_5 = 0$

- $N_6 = 5 \rightarrow 2 * 2 + 1 \rightarrow a_6 = 1$
- $N_7 = 2 \rightarrow 2 * 1 + 0 \rightarrow a_7 = 0$
- $N_8 = 1 \rightarrow 2 * 0 + 1 \rightarrow a_8 = 1$

- Logo, $(347)_{10}$ na base 2 é $(a_8a_7a_6a_5a_4a_3a_2a_1a_0)_2$
 - $(a_8a_7a_6a_5a_4a_3a_2a_1a_0)_2 = 101011011$

 Podemos generalizar o processo de conversão de um número N seguindo o algoritmo ao lado.

Passo 0:
$$k = 0$$

 $N_k = N$

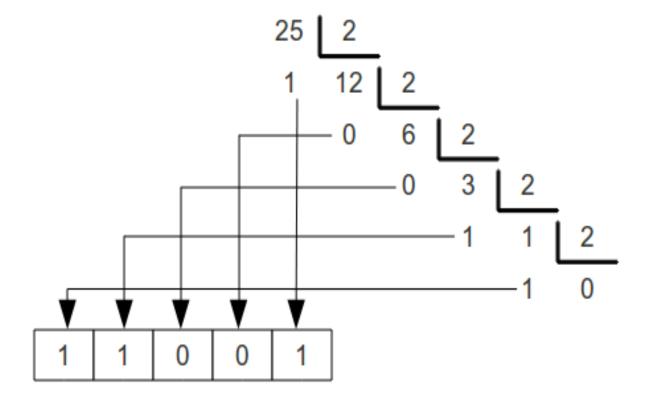
Passo 1: Obtenha
$$q_k e r_k$$
 tais que:

$$N_k = 2 \times q_k + r_k$$
Faça $a_k = r_k$

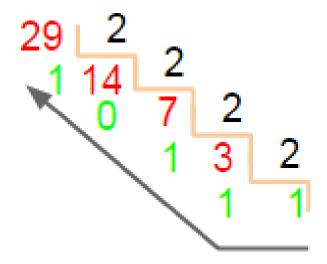
Passo 2: Se
$$q_k = 0$$
, pare.
Caso contrário, faça $N_{k+1} = q_k$.
Faça $k = k + 1$ e volte para o passo 1.

Uma outra forma de realizar a conversão de decimal para binário é realizar sucessivas divisões e armazenamento do resto das divisões.

• Exemplo: $(25)_{10} \rightarrow (X)_2$

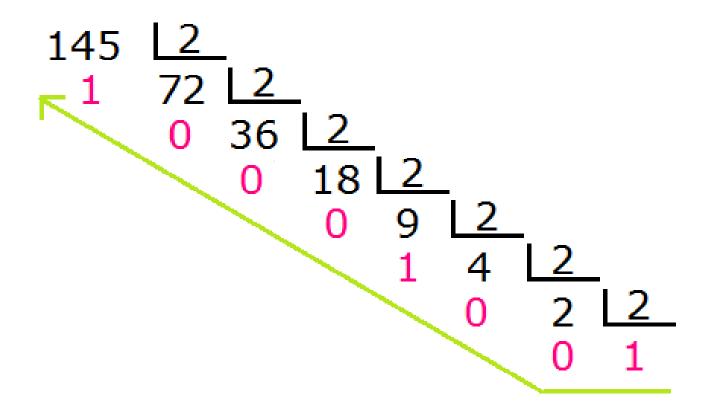


• Exemplo 2: $(29)_{10} \rightarrow (X)_2$



29 Decimal = 11101 Binário

• Exemplo 3: $(145)_{10} \rightarrow (X)_2$

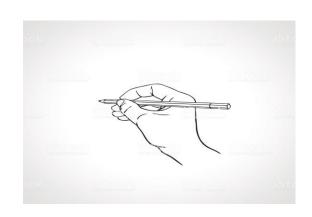


145 em decimal é 10010001 em binário.

Exercício 2

2. Converta os seguintes números para a base binária.

- a) $(35)_{10}$
- b) (100)₁₀
- c) $(357)_{10}$



• Um número real positivo x pode ser escrito como:

$$x = \underbrace{\sum_{i=0}^{n} a_i B^i}_{x_{\text{int}}} + \underbrace{\sum_{i=1}^{\infty} b_i B^{-i}}_{x_{\text{frac}}}$$

- onde a_i e b_i são, respectivamente, os coeficientes da parte inteira e fracionária do número x.
- Por exemplo:
 - $(132,56)_{10} = 1 \times 10^2 + 3 \times 10^1 + 2 \times 10^0 + 5 \times 10^{-1} + 6 \times 10^{-2}$

- Conversão da base 2 para decimal é similar ao caso inteiro.
- Conversão da base decimal para binária.
 - Converte-se a parte inteira.
 - Divisões sucessivas.
 - Converte-se a parte fracionária.
 - Multiplicações sucessivas.

- Consideremos agora a conversão de um número fracionário na base
 10 para a base 2.
- Sejam, por exemplo:
 - r = 0.125

→ representação Finita

• $s = 0.6666 \dots$

- → representação Infinita
- t = 0.414213562...
- → representação Infinita
- Dado um número entre 0 e 1 no sistema decimal, como obter sua representação binária?

- Considere o número r = 0,125, existem dígitos binários: d_1 , d_2 , ..., d_j , tais que $(0, d_1 d_2 d_j)_2$ será sua representação na base 2.
- Assim:
 - $(0.125)_{10} = d_1 * 2^{-1} + d_2 * 2^{-2} + ... + d_i * 2^{-j}$
 - Vamos multiplicar cada termo da expressão por 2:
 - $0.250 = d_1 + d_2 * 2^{-1} + ... + d_j * 2^{-j+1}$
 - d_1 será a parte inteira de 2 x 0,125 = 0,250 que é igual a zero
 - $d_1 = 0$.

• $d_2 * 2^{-1} + ... + d_j * 2^{-j+1}$

- → Parte fracionária
- Aplicando o mesmo procedimento para 0,250, temos:
 - $0.250 = d_2 * 2^{-1} + ... + d_j * 2^{-j+1}$ (* 2)
 - $0.5 = d_2 + d_3 * 2^{-1} + d_4 * 2^{-2} + ... + d_j * 2^{-j+2}$
 - $d_2 = 0$.
 - $0.5 = d_3 * 2^{-1} + d_4 * 2^{-2} + ... + d_i * 2^{-j+2}$ (* 2)

•
$$0.5 = d_3 * 2^{-1} + d_3 * 2^{-2} + ... + d_i * 2^{-j+2}$$
 (* 2)

- 1,0 = $d_3 + d_4 * 2^{-1} + ... + d_j * 2^{-j+3}$
 - $d_3 = 1$.
- Como a parte fracionária de 1,0 é igual a zero, o processo termina.
- Logo, obtivemos:
 - $d_1 = 0$ $d_2 = 0$ $d_3 = 1$
- Portanto:
 - $(0.125)_{10} = (0.001)_2$

- OBS: Um número real entre 0 e 1 pode ter representação finita no sistema decimal, mas representação infinita o sistema binário.
 - Exemplo: 0,1

- No caso geral, seja r um número entre 0 e 1 no sistema decimal e $(d_1d_2 \dots d_j)_2$ sua representação no sistema binário.
 - Os dígitos binários $d_1d_2 \dots d_j$ podem ser obtidos através do seguinte algoritmo:

Passo 0:
$$r_1 = r$$
; $k = 1$

Passo 1: Calcule
$$2r_k$$
.
Se $2r_k \ge 1$, faça: $d_k = 1$,
caso contrário, faça: $d_k = 0$

Passo 2: Faça
$$r_{k+1} = 2r_k - d_k$$
.
Se $r_{k+1} = 0$, pare.
Caso contrário:

Passo 3:
$$k = k + 1$$
.
Volte ao passo 1.

- O algoritmo anterior pode ou não parar após um número finito de passos.
 - Para $r = (0.125)_{10}$, teremos $r_4 = 0$ (critério de parada).
 - Já para $r = (0,1)_{10}$, teremos $r_1 = 0,1$.

• Para $r = (0,1)_{10}$, teremos:

$$k = 1$$
 $2r_1 = 0.2 \Rightarrow d_1 = 0$
 $r_2 = 0.2$

Passo 0:
$$r_1 = r$$
; $k = 1$

$$k = 2$$
 $2r_2 = 0.4 \Rightarrow d_2 = 0$
 $r_3 = 0.4$

Passo 1: Calcule
$$2r_k$$
.
Se $2r_k \ge 1$, fa

$$k = 3 2r_3 = 0.8 \Rightarrow d_3 = 0$$

Se
$$2r_k \ge 1$$
, faça: $d_k = 1$, caso contrário, faça: $d_k = 0$

$$k = 4$$
 $2r_4 = 1.6 \Rightarrow d_4 = 1$
 $r_5 = 0.6$

Passo 2: Faça
$$r_{k+1} = 2r_k - d_k$$
.
Se $r_{k+1} = 0$, pare.
Caso contrário:

$$k = 5 \quad 2r_5 = 1.2 \implies d_5 = 1$$

Passo 3:
$$k = k + 1$$
.
Volte ao passo 1.

$$r_6 = 0.2 = r_2$$

- Como $r_6 = r_2$, teremos que os resultados para k de 2 a 5 se repetirão e então: $r_{10} = r_6 = r_2 = 0.2$ e assim indefinidamente.
- Logo,
 - $(0,1)_{10} = (0,000110011\overline{0011}...)_2$
- E, portanto, o número $(0,1)_{10}$ não tem representação binária finita.
 - Isso pode ocasionar erros em cálculos efetuados em sistemas computacionais.

Caso Patriot Missile



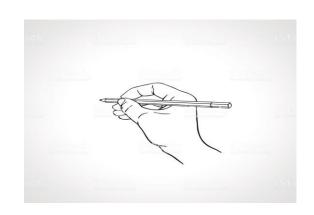
- Houve uma falha no radar do sistema
 Patriot e no software que o suportava.
 - Um erro de arredondamento no cálculo e na medição do tempo.

Patriot Missile

Exercício 3

3. Converta os seguintes números para a base 2.

- a) $(0,186)_{10}$
- *b)* (8,250)₁₀
- c) (213,11)₁₀



Conversão de Números nos Sistemas Decimal e Binário

• Considere agora um número entre 0 e 1 representado no sistema binário: $(r)_2 = (0, d_1 d_2 \dots d_i)_2$

- Como obter sua representação no sistema decimal?
- Para isso pode-se utilizar o seguinte algoritmo:
 - $(0,000111)_2 = 0 * 2^{-1} + 0 * 2^{-2} + 0 * 2^{-3} + 1 * 2^{-4} + 1 * 2^{-5} + 1 * 2^{-6}$

• = 1 *
$$\frac{1}{16}$$
 + 1 * $\frac{1}{32}$ + 1 * $\frac{1}{64}$ \rightarrow 0,10938

• $(0,000111)_2 = (0,10938)_{10}$

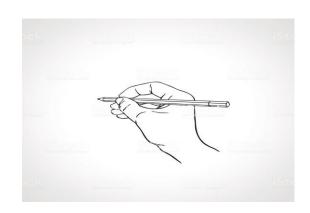
Conversão de Números nos Sistemas Decimal e Binário

- Exemplo: $(0,1100100)_2 = X_{10}$
 - $(0,1100100)_2 = 1 * 2^{-1} + 1 * 2^{-2} + 0 * 2^{-3} + 0 * 2^{-4} + 1 * 2^{-5} + 0 * 2^{-6} + 0 * 2^{-7}$
 - = 1 * $\frac{1}{2}$ + 1 * $\frac{1}{4}$ + 1 * $\frac{1}{32}$ \rightarrow 0,78125
 - $(0,1100100)_2 = (0,78125)_{10}$

Exercício 4

4. Converta os seguintes números para a base 10.

- *a*) (111,01)₂
- *b*) (101,10)₂



- Um computador ou calculadora representa um número real no sistema denominado Aritmética de Ponto Flutuante.
- Neste sistema, um número real r será representado na forma:
 - $\pm (.d_1d_2...d_t) \times \beta^e$
 - Onde:
 - β é a base em que a máquina opera;
 - t é o número de dígitos na mantissa; $0 \le d_j \le (\beta 1)$, $j = 1, ..., t, d_1 \ne 0$;
 - *e* é o expoente no intervalo [l,u].

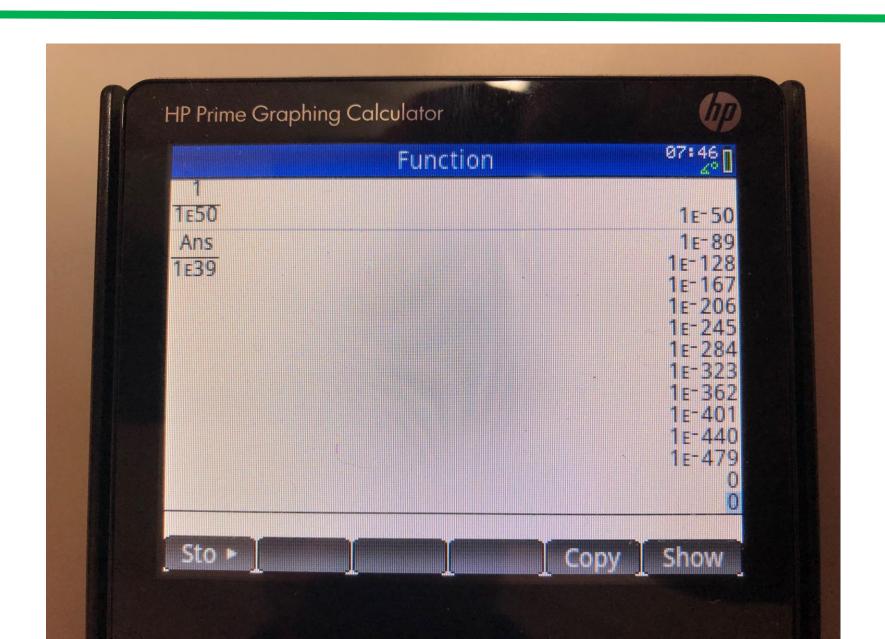
- Em qualquer máquina, apenas um subconjunto dos números reais é representado exatamente.
 - Portanto, a representação de um número real será realizada através de truncamento ou de arredondamento.
- Considere, por exemplo, uma máquina que opera no sistema:
 - $\beta = 10$; t = 3; $e \in [-5,5]$ \rightarrow F(10; 3; -5, 5)
 - Os números serão representados na seguinte forma nesse sistema:
 - $.d_1d_2d_3 \times 10^e, 0 \le d_j \le 9, e \in [-5,5]$

- O menor e o maior número, e valor absoluto, representado nesta máquina é:
 - $m = 0.100 \times 10^{-5} = 10^{-6}$
 - $M = 0.999 \times 10^5 = 99900$
- Considere o conjunto dos números reais \mathbb{R} e o seguinte conjunto:
 - $G = \{ x \in \mathbb{R} \mid m \le x \le M \}$
- Teremos 3 situação para analisar:

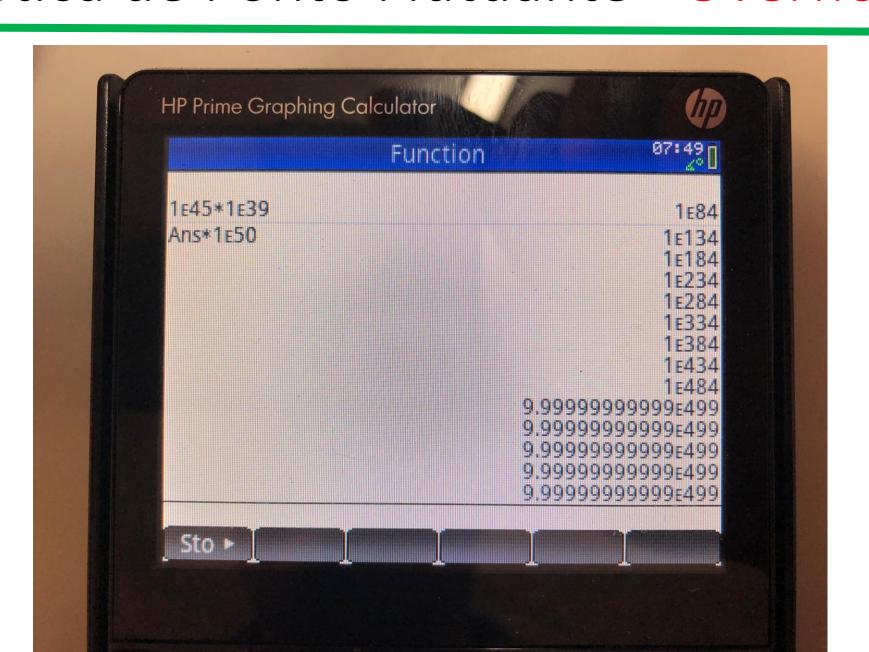
1. $x \in G$

- Exemplo: $x = 235.89 \rightarrow 0.23589 \times 10^3$ (OK)
- 2. |x| < m:
 - Exemplo: $x = 0.345 \times 10^{-7}$ (Underflow)
 - Expoente é menor do que -5.
- 3. |x| > M
 - Exemplo: $x = 0.875 \times 10^9$ (Overflow)
 - Expoente é maior do que 5.

Aritmética de Ponto Flutuante - Underflow



Aritmética de Ponto Flutuante - Overflow



Caso Ariane 5



Caso Ariane 5 — Erro



Ariane 501 Inquiry Board report page 4

The internal SRI software exception was caused during execution of a data conversion from 64-bit floating point to 16-bit signed integer value. The floating point number which was converted had a value greater than what could be represented by a 16-bit signed integer. This resulted in an Operand Error. The data conversion instructions (in Ada code) were not protected from causing an Operand Error, although other conversions of comparable variables in the same place in the code were protected.

Cálculo Numérico....

- Algumas Linguagens de Programação permitem que as variáveis sejam declaradas em precisão dupla.
 - Neste caso, esta variável será representada no sistema de aritmética de ponto flutuante da máquina, mas com aproximadamente o dobro de dígitos disponíveis na mantissa.
 - OBS: O tempo de execução e o consumo de memória aumentam significativamente.

Dar a representação dos números a seguir num sistema de aritmética de ponto flutuante de três dígitos para $\beta = 10$, m = -4 e M = 4.

x	Representação obtida por arredondamento	Representação obtida por truncamento		
1.25	0.125 × 10	0.125 × 10		
10.053	0.101×10^{2}	0.100×10^2		
-238.15	-0.238×10^3	-0.238×10^3		
2.71828	0.272×10	0.271×10		
0.000007	(expoente menor que -4)	=		
718235.82 (expoente maior que 4)		=		

Máquina e Aritmética	β	t	e_{min}	$e_{\rm max}$
Cray-1 Precisão Simples	2	48	-8192	8191
Cray-1 Precisão Dupla	2	96	-8192	8191
DEC VAX formato G	2	53	-1023	1023
Dupla				
DEC VAX formato D	2	56	-127	127
Dupla				
Calculadoras HP 28 e 48G	10	12	-499	499
IBM 3090 Precisão Simples	16	6	-64	63
IBM 3090 Precisão Dupla	16	14	-64	63
IBM 3090 Precisão	16	28	-64	63
Extendida				
IEEE Precisão Simples	2	24	-126	127
IEEE Precisão Dupla	2	53	-1022	1023
PDP 11	2	24	-128	127
Control Data 6600	2	48	-976	1070

Fonte: https://www1.univap.br/spilling/CN/CN_Capt1.pdf

Exercício 5

5. Considerando o sistema *F*(10; 3; -5; 5). Represente o número 1,23 nesse sistema. Qual o maior e menor número que pode ser representado nesse sistema?



Exercício Resolvido

- Representar os números abaixo em um sistema de aritmética de ponto flutuante F(10, 5, -5, 5). Utilize em cada caso a aproximação por arredondamento e por truncamento.
 - a) 1,38
 - b) 92,0637
 - c) -398,216
 - d) 3,141592
 - e) 0,000003
 - f) 21479312,95

Exercício Resolvido

a)
$$1,38 = 0,13800 \times 10^{1}$$
 (arredondamento=truncamento)
b) $92,0637 = 0,92064 \times 10^{2}$ (arredondamento)
 $0,92063 \times 10^{2}$ (truncamento)
c) $-398,216 = -0,39822 \times 10^{3}$ (arredondamento)
 $-0,39821 \times 10^{3}$ (truncamento)
c) $3,141592 = 0,31416 \times 10^{1}$ (arredondamento)
 $0,31415 \times 10^{1}$ (truncamento)

Exercício Resolvido

d) $0.0000003 = 0.30000 \times 10^{-6}$ (e < -5)

UNDERFLOW

e) $21479312,95 = 0,21479 \times 10^8 \text{ (e > 5)}$

OVERFLOW

Erros Absolutos e Relativos

Erro Absoluto

• Erro Absoluto é a diferença entre o valor exato de um número x e de seu valor aproximado \bar{x} .

•
$$EA_x = x - \bar{x}$$

- No geral, apenas o valor de \bar{x} é conhecido, o que impossibilita o valor exato do erro absoluto.
 - Define-se um limitante superior ou uma estimativa para o módulo do erro absoluto.

Erro Absoluto

- Por exemplo, sabendo-se que $\pi \in (3,13;3,15)$, tomaremos um valor dentro desse intervalo e teremos:
 - $|EA_x| = |\pi \bar{\pi}| < 0.01$
- Suponha a seguinte situação:
 - $\bar{x} = 2112,9 \text{ e } |EA_x| < 0,1$
 - $x \in (2112,8; 2113)$
 - $\bar{y} = 5.3 \text{ e } |EA_v| < 0.1$
 - $y \in (5,2;5,3)$

Erro Relativo

O Erro Relativo é definido como o erro absoluto dividido pelo valor real.

•
$$ER_{x} = \frac{EA_{x}}{\overline{x}} = \frac{x - \overline{x}}{\overline{x}}$$

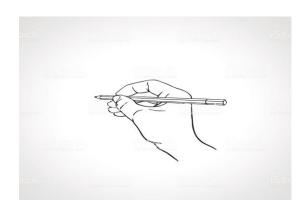
Usando o exemplo anterior, temos:

•
$$ER_x = \frac{EA_x}{\overline{x}} = \frac{0.1}{2112.9} \approx 4.7 \ x \ 10^{-5}$$

• ER_y =
$$\frac{EA_y}{\overline{y}} = \frac{0.1}{5.3} \approx 0.02$$

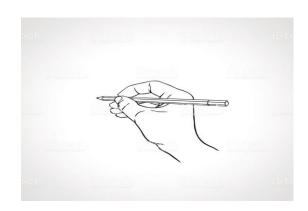
Exercício 6

6. Seja o sistema *F*(10; 4; *L*; *U*). Qual o erro absoluto ao representar *x* = 1428,756 nesse sistema usando truncamento e arredondamento?



Exercício 7

7. Sejam $x_1 = 1000,5$, $\overline{x_1} = 1000,6$, $x_2 = 10,5$ e $\overline{x_2} = 10,6$. Nota-se que $EA(\overline{x_1}) = EA(\overline{x_2}) = 0,1$. Quais os erros relativos?



Dúvidas



Próxima aula ...

Aula 4 – Zeros Reais de Funções Reais.