CÁLCULO NUMÉRICO

Noções de aritmética de máquina

Gustavo Bono

Universidade Federal de Pernambuco
Centro Acadêmico do Agreste
Núcleo de Tecnologia
Caruaru - Brasil

2017.2

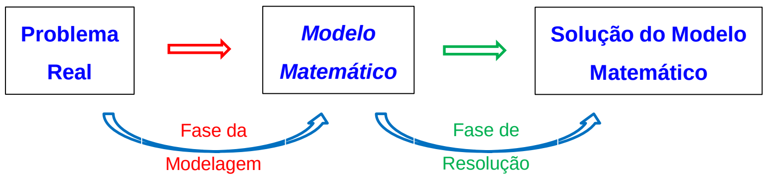
1) Noções de aritmética de máquina: *Erros em processos numéricos*

Problema Real ⟹ ***Modelo Matemático*** ⟹ **Solução do Modelo Matemático**

Fase da Modelagem — Fase de Resolução

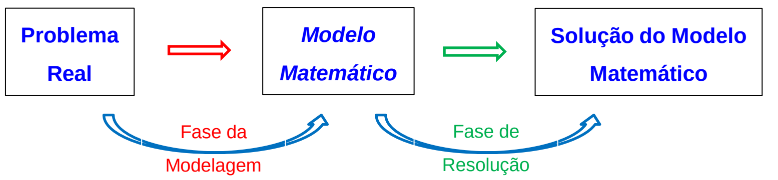
Erros na fase da modelagem:

São os erros decorrentes de simplificações, muitas vezes necessárias, para que o fenômeno da natureza (problemas com multifísica e multiescala no tempo e no espaço) que estivermos observando possa ser representado por um modelo matemático e que tenha condições de ser tratado com as ferramentas matemáticas disponíveis.



Erros na fase de resolução:

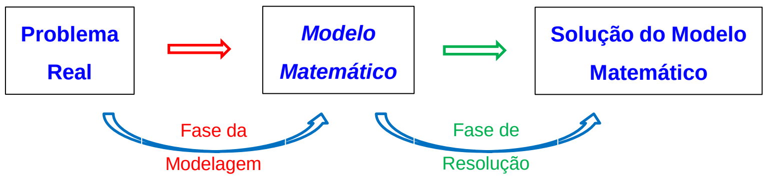
São erros provenientes da utilização de algum equipamento, como, por exemplo, um computador/calculadora, para processarmos os cálculos necessários à obtenção de uma solução para o modelo matemático. Tais erros ocorrem devido ao fato de os equipamentos terem capacidade limitada para armazenar os dígitos significativos de valores numéricos utilizados nas operações elementares (+ , - , x , /).



Erros na fase de resolução:

1) Erro de arredondamento: os números são representados em um computador através de um número finitos de *bits*. Consequentemente, números reais que têm uma mantissa mais longa do que o número de bits disponíveis para representá-los têm que ser encurtados.

1) Noções de aritmética de máquina: *Erros em processos numéricos*

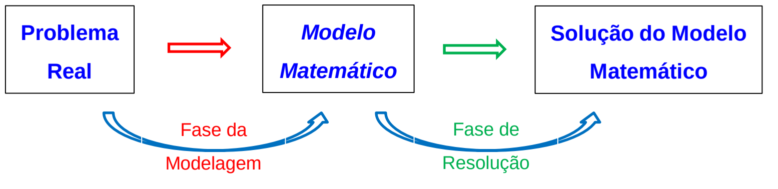


Erros na fase de resolução:

2) Erro de truncamento: quando representamos uma função através de uma série infinita e, por limitações do sistema de armazenamento de dados do equipamento, considerarmos apenas um número finito de termos, dizemos que estamos cometendo um erro de truncamento.

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

1) Noções de aritmética de máquina: *Erros em processos numéricos*



Erros na fase de resolução:

3) Erros na mudança da base: os equipamentos computacionais representam os valores numéricos no sistema binário. Assim, quando os dados numéricos presentes nos modelos matemáticos são lidos, estes são transformados em uma outra base de representação. Muitas vezes, esta transformação pode ser acometida de erros, em razão da limitação da representação do equipamento computacional que estamos utilizando para o processamento de dados numéricos.

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

1) Noções de aritmética de máquina: *Erros em processos numéricos*

Problema Real ⟹ ***Modelo Matemático*** ⟹ **Solução do Modelo Matemático**

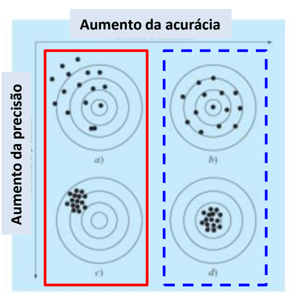
Fase da Modelagem — Fase de Resolução

Erros na fase de resolução:

4) Erros de representação: na construção de um equipamento computacional, uma questão importante a ser considerada em sua arquitetura é a forma que será adotada para representar os dados numéricos. Basicamente, na memória de um equipamento, cada número é armazenado em uma posição que consiste de um sinal que identifica se o número é (+) ou (-) e um número fixo e limitado de dígitos significativos.

1.1) Noções de aritmética de máquina: *Acurácia e Precisão*

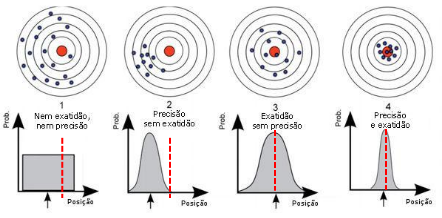
Os erros associados tanto aos cálculos quanto às medidas podem ser caracterizados com relação à sua **acurácia** e ***precisão***. A **acurácia** se refere a quão próximo o valor calculado ou medido está do valor verdadeiro. A ***precisão*** se refere a quão próximo os valores individuais calculados ou medidos estão uns dos outros.



Inacurácia é definida como um <u>desvio sistemático</u> da verdade.
Embora os pontos em (**C**) estejam agrupados mais juntos do que aqueles em (**A**), os dois são igualmente *inacurados* porque ambos estão centrados no quadrante superior esquerdo !!

As **imprecisões** (incertezas), por outro lado, referem-se à intensidade do espalhamento. Portanto, embora (B) e (D) sejam igualmente exatas (centradas no alvo), a última é mais precisa porque os pontos estão agrupados mais juntos.

1.1) Noções de aritmética de máquina: *Acurácia e Precisão*



A ***precisão*** é definida pelo *desvio padrão* de uma série de medidas de uma mesma amostra ou um mesmo ponto. *Quando maior o desvio padrão, menor é a precisão*. A precisão está relacionada com a incertezas a ... o c...

Exatidão (acurácia) é a aptidão de um instrumento p... or verdadeiro do mensurado. É a capacidade que o instru... er um resultado correto. Um equipamento exato é aque... s, nos fornece um valor médio que é próximo ao real ... ja elevado, ou seja, apresente baixa precisão.

A exatidão está relacionada às incertezas sistemáticas da medição. A exatidão pode ser avaliada através da calibração do instrumento.

1.2) Noções de aritmética de máquina: *Erros de arredondamento*

Algarismos Significativos

Os algarismos significativos são dígitos que iniciam com o dígito não nulo mais à esquerda e terminam com o dígito mais correto à direita.

$0,0018 \rightarrow 0,18\times10^{-2}$ ⟶ Tem 2 dígitos significativos

$0,001800 \rightarrow 0,1800\times10^{-2}$ ⟶ Tem 4 dígitos significativos

$0,03418 \rightarrow 0,3418\times10^{-1}$ ⟶ Tem 4 dígitos significativos

$0,45086 \rightarrow 0,45086\times10^{0}$ ⟶ Tem 5 dígitos significativos

$0,00035018 \rightarrow 3,5018\times10^{-4}$ ⟶ Tem 5 dígitos significativos

Mas que acontece com as medições ??

1.2) Noções de aritmética de máquina: *Erros de arredondamento*

Algarismos Significativos nas medições (incerteza nos dados de entrada/medidos)

$15{,}18 \pm 0{,}18$ mm

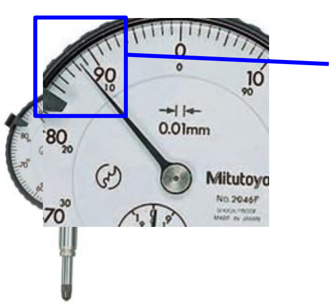
$5{,}39 \pm 0{,}33$ Kg

$12{,}15 \pm 0{,}05$ volts

1.2) Noções de aritmética de máquina: *Erros de arredondamento*

Algarismos Significativos nas medições

Os algarismos significativos de um número são aqueles que podem ser usados com confiança. Eles correspondem ao número de algarismos corretos mais um algarismo estimado.



No comparador observamos que a medida está entre **88** e **89 mm**.

Pode-se dizer com segurança que a medida é aproximadamente **88 mm**.

Entretanto, se fosse considerada até uma casa decimal, uma pessoa poderia dizer ***88,3 mm***, enquanto outra poderia dizer ***88,4 mm***.

Os **DOIS primeiros algarismos (88)** podem ser usados com confiança !!

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

1.2) Noções de aritmética de máquina: *Erros de arredondamento*

Algarismos Significativos nas medições

Os algarismos significativos de um número são aqueles que podem ser usados com confiança. Eles correspondem ao número de algarismos corretos mais um algarismo estimado.

Convenciona-se tomar o **algarismo estimado** como a metade da menor divisão de escala no aparelho de medida.

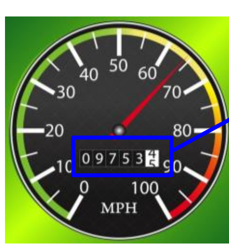
Portanto, a leitura consistiria de **TRÊS algarismos significativos**

88,5 mm

1.2) Noções de aritmética de máquina: *Erros de arredondamento*

Algarismos Significativos nas medições

Os algarismos significativos de um número são aqueles que podem ser usados com confiança. Eles correspondem ao número de algarismos corretos mais um algarismo estimado.



O hodômetro do carro forneceria quantos algarismos significativos ?

9753,45 km

SEIS algarismos significativos !!

EXEMPLO 1: erro de arredondamento

Vamos a resolver um sistema linear de 2 x 2 equações, de duas formas.

$$0,1036\,x + 0,2122\,y = 0,7381$$
$$0,2081\,x + 0,4247\,y = 0,9327$$

A) Vamos a considerar somente três dígitos significativos de precisão nos cálculos;

B) Vamos a considerar quatro dígitos significativos para o sistema original.

Arredondar um número x, por outro com um número menor de dígitos significativos, consiste em encontrar um número $\bar{x}$, pertencente ao sistema de numeração, tal que $|\bar{x} - x|$ seja o menor possível.

Norma ABNT NBR 5891 – Regras de arredondamento na numeração decimal

Arredondando a 2 algarismos decimais deveremos ter em atenção o terceiro e quarto decimal, conforme as regras apresentadas nos exemplos a seguir:

- O número 12,6**529** seria arredondado para 12,6**5** (uma vez que 29 é inferior a 50, então não se modifica);
- O número 12,8**6512** seria arredondado para 12,8**7** (uma vez que 512 é superior a 500, então incrementa-se uma unidade);
- O número 12,7**44623** seria arredondado para 12,7**4** (uma vez que 4623 é inferior a 5000, então não se modifica);
- O número 12,8**150** seria arredondado para 12,8**2** (uma vez que os algarismos seguintes são iguais a 50 e o anterior é ímpar, nesse caso 1, então incrementa-se uma unidade);
- O número 12,8**050** seria arredondado para 12,8**0** (uma vez que os algarismos seguintes são iguais a 50 e o anterior é par, nesse caso 0, então o anterior não se modifica).

1.2) Noções de aritmética de máquina: *Erros de arredondamento*

EXEMPLO 2: erro de arredondamento

Efetue as operações indicadas, fazendo o arredondamento após cada uma das operações:

$$(11,4+3,18)+5,06 \quad \text{e} \quad 11,4+(3,18+5,06)$$
$$\frac{3,18\times 11,4}{5,05} \quad \text{e} \quad \left(\frac{3,18}{5,05}\right)\times 11,4$$
$$3,18\times(5,06+11,4) \quad \text{e} \quad 3,18\times 5,06+3,18\times 11,4$$

EXEMPLO 3: erro de arredondamento

Somar $1/3$ dez vezes consecutivas, usando arredondamento.

1.2) Noções de aritmética de máquina: *Erros de arredondamento*

EXEMPLO 4: erro de arredondamento

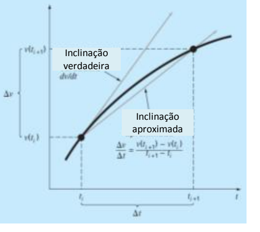
Avaliar o polinômio $P(x)=x^3-6x^2+4x-0,1$ no ponto 5,24 e compara com o resultado *exato*.

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

1.3) Noções de aritmética de máquina: *Erros de truncamento*

Surgem quando aproximamos um procedimento formado por uma sequencia infinita de passos através de um procedimento finito. Os erros de truncamento são aqueles que resultam do uso de uma aproximação no lugar de um procedimento matemática exato.

Por exemplo:



Aproximação da derivada da velocidade

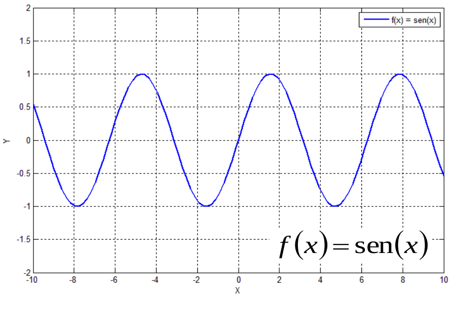
$$\frac{dv}{dt} = \lim_{\Delta t \to 0} \frac{\Delta v}{\Delta t} \qquad \frac{dv}{dt} \cong \frac{\Delta v}{\Delta t} = \frac{v(t_{i+1}) - v(t_i)}{t_{i+1} - t_i}$$

Expansão em séries de Taylor

$$\sin x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$$\cos x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$

1.3) Noções de aritmética de máquina: *Erros de truncamento*



Séries de Taylor

1.3) Noções de aritmética de máquina: *Erros de truncamento*

Teorema: Se f tiver uma representação (expansão) em série de potências em a, isto é, se:

$$f(x) = \sum_{n=0}^{\infty} c_n (x-a)^n \qquad |x-a| < R$$

então seus coeficientes são dados pela fórmula:

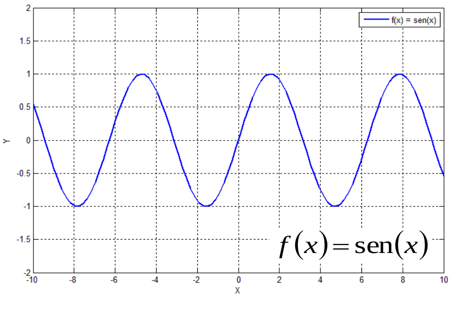
$$c_n = \frac{f^{(n)}(a)}{n!}$$

Série de Taylor da função *f* em *a* (em torno de a ou centrada em a):

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

$$= f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots$$

1.3) Noções de aritmética de máquina: *Erros de truncamento*

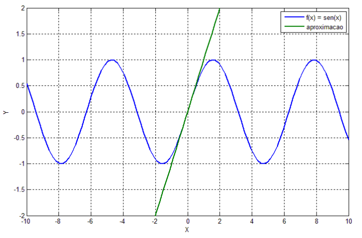


- para x = 0

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$$= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} \qquad \text{for all } x$$

1.3) Noções de aritmética de máquina: *Erros de truncamento*



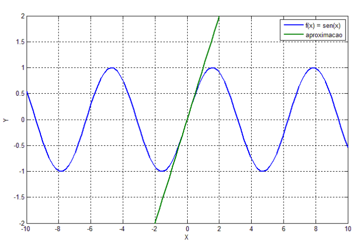
- para x = 0

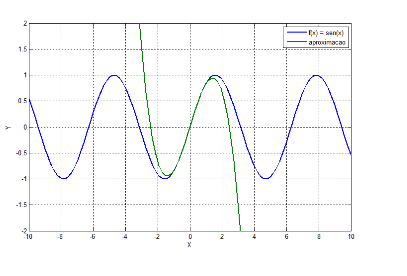
$$\boxed{\operatorname{sen}(x) = x} - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \frac{x^{13}}{13!} - \frac{x^{15}}{15!} + \cdots$$

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

Campus AGRESTE

1.3) Noções de aritmética de máquina: *Erros de truncamento*





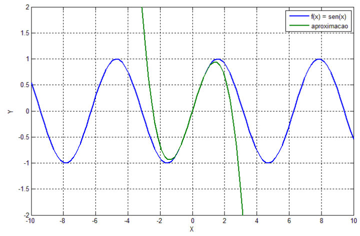
- para x = 0

$$\boxed{\operatorname{sen}(x) = x - \frac{x^3}{3!}} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \frac{x^{13}}{13!} - \frac{x^{15}}{15!} + \cdots$$

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

Campus AGRESTE

1.3) Noções de aritmética de máquina: *Erros de truncamento*



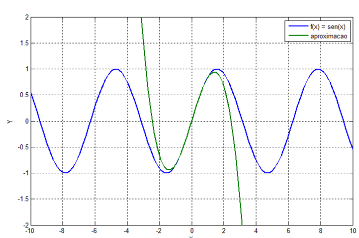
- para x = 0

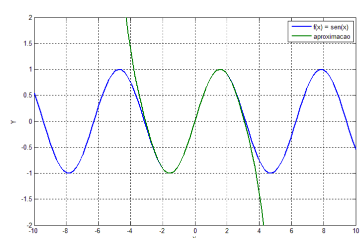
$$\boxed{\text{sen}(x)=x-\frac{x^3}{3!}}+\frac{x^5}{5!}-\frac{x^7}{7!}+\frac{x^9}{9!}-\frac{x^{11}}{11!}+\frac{x^{13}}{13!}-\frac{x^{15}}{15!}+\cdots$$

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

Campus AGRESTE

1.3) Noções de aritmética de máquina: *Erros de truncamento*





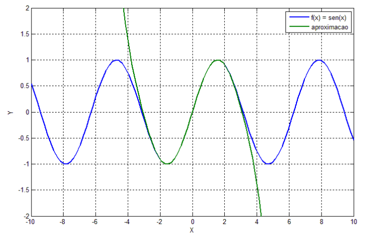
- para x = 0

$$\boxed{\text{sen}(x)=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}}+\frac{x^9}{9!}-\frac{x^{11}}{11!}+\frac{x^{13}}{13!}-\frac{x^{15}}{15!}+\cdots$$

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

Campus AGRESTE

1.3) Noções de aritmética de máquina: *Erros de truncamento*



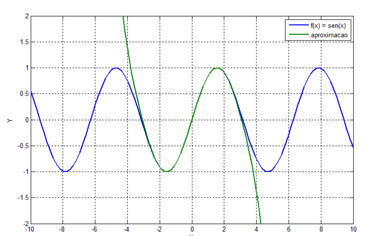
- para x = 0

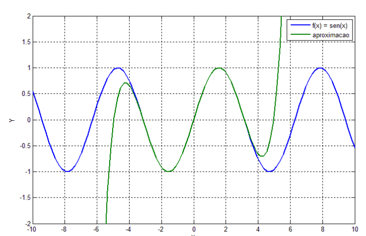
$$\boxed{\mathrm{sen}(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \frac{x^{13}}{13!} - \frac{x^{15}}{15!} + \cdots$$

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

Campus AGRESTE

1.3) Noções de aritmética de máquina: *Erros de truncamento*





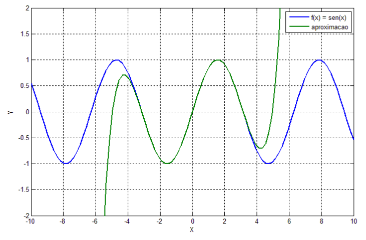
- para x = 0

$$\boxed{\mathrm{sen}(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!}} - \frac{x^{11}}{11!} + \frac{x^{13}}{13!} - \frac{x^{15}}{15!} + \cdots$$

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

Campus AGRESTE

1.3) Noções de aritmética de máquina: *Erros de truncamento*

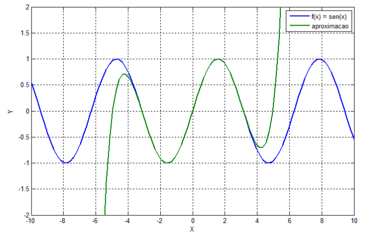


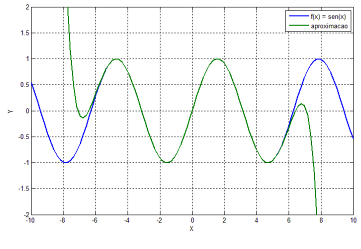
- para x = 0

$$\boxed{\text{sen}(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!}} - \frac{x^{11}}{11!} + \frac{x^{13}}{13!} - \frac{x^{15}}{15!} + \cdots$$

Campus AGRESTE

1.3) Noções de aritmética de máquina: *Erros de truncamento*





- para x = 0

$$\boxed{\text{sen}(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \frac{x^{13}}{13!} - \frac{x^{15}}{15!}} + \cdots$$

Campus AGRESTE

1.3) Noções de aritmética de máquina: *Erros de truncamento*

Em geral, é conveniente simplificar a série de Taylor definindo um tamanho do passo $h = x_{i+1} - x_i$ e expressando a equação como:

$$f(x_{i+1}) = f(x_i) + f'(x_i)(x_{i+1} - x_i) + \frac{f''(x_i)}{2!}(x_{i+1} - x_i)^2 + \frac{f^{(3)}(x_i)}{3!}(x_{i+1} - x_i)^3 + \cdots + \frac{f^{(n)}(x_i)}{n!}(x_{i+1} - x_i)^n + R_n$$

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \frac{f^{(3)}(x_i)}{3!}h^3 + \cdots + \frac{f^{(n)}(x_i)}{n!}h^n + R_n$$

onde o resto (R_n) agora é

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1}$$

Em geral, a expansão em série de Taylor de ordem *n* será exata para um polinômio de grau *n*. Para outras funções diferenciáveis (contínuas), como exponenciais e funções senoidais, um número finito de termos não resultará em uma estimativa exata !!! $\Rightarrow$ Ver exemplo do seno.

1.3) Noções de aritmética de máquina: *Erros de truncamento*

Para fins práticos, quantos termos serão suficientes na expansão em série de Taylor para ter uma boa aproximação ?

A avaliação de quantos termos são necessários para ficar suficientemente próximo é baseado no resto (R_n) da expansão:

$$R_n = \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1}$$

1) Não é conhecido exatamente → está entre x_i e x_{i+1}
2) É preciso determinar a $n+1$ ésima derivada de $f(x)$ $\Rightarrow$

Para fazer isso, é necessário conhecer *f(x)*

A equação do resto em geral é expressa como

$$\boxed{R_n = O(h^{n+1})}$$

1.3) **Noções de aritmética de máquina:** *Erros de truncamento*

$$\boxed{R_n = O(h^{n+1})}$$

Ou seja, o erro é proporcional ao tamanho do passo h elevado à $n+1$ ésima potência. Embora essa aproximação não tenha qualquer implicação no valor da derivada que multiplica h^{n+1}, ela é extremamente útil para julgar o erro comparativo de métodos numéricos baseados em expansões em ST.

Ordem do resto	Tamanho do passo	Erro
$O(h)$	h	$(erro)$
$O(h)$	$\frac{h}{2}$	$\left(\frac{erro}{2}\right)$
$O(h^2)$	$\frac{h}{2}$	$\left(\frac{erro}{4}\right)$

O erro de truncamento diminui com a adição de termos na ST. Em muitos casos, se h for suficientemente pequeno, o primeiro e alguns outros termos de baixa ordem serão necessários para uma estimativa adequada !!

1.3) **Noções de aritmética de máquina:** *Erros de truncamento*

EXEMPLO 5: erro de truncamento

Use expansões em séries de Taylor de ordem zero até ordem quatro para a aproximar a função

$$f(x) = -0{,}1x^4 - 0{,}15x^3 - 0{,}5x^2 - 0{,}25x + 1{,}2$$

A partir de $x_i = 0$ com $h = 1$. Isto é, faça uma previsão do valor da função em $x_{i+1} = 1$.

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{f''(x_i)}{2!}h^2 + \frac{f^{(3)}(x_i)}{3!}h^3 + \cdots + \frac{f^{(n)}(x_i)}{n!}h^n + R_n$$

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

Campus AGRESTE

1.4) Noções de aritmética de máquina: *Erros*

Suponha que se quer computar $f(x)$ para um dado x real. No cálculo prático x é aproximado por $\underline{x}$ pois o computador tem uma palavra finita. A diferença entre $(\underline{x} - x)$ é o ***erro inicial***, enquanto $\varepsilon_1 = f(\underline{x}) - f(x)$ é o ***erro propagado*** correspondente.

Normalmente, f é trocada por uma função mais simples f_1 (frequentemente uma expansão em ST). A diferença $\varepsilon_2 = f_1(\underline{x}) - f(\underline{x})$ é o ***erro de truncamento***. Os cálculos são feitos por um computador, portanto, não são operações exatas. Na realidade calculamos $f_2(\underline{x})$ no lugar de $f_1(\underline{x})$, o qual é um valor calculado errado de uma função errada com argumento errado. A diferença $\varepsilon_3 = f_2(\underline{x}) - f_1(\underline{x})$ é o ***erro de arredondamento propagado***.

O erro total é

$$\varepsilon = f_2(\underline{x}) - f(x)$$

EXEMPLO 6: erro de truncamento

Calcular $e^{1/3}$ fazendo todos os cálculos com 4 dígitos significativos.

1.4) Noções de aritmética de máquina: *Erros*

Um defeito que apresenta a definição do erro dado pela equação

$$Erro\_Real = \text{Solução Exata} - \text{Solução Numérica (aproximação)}$$

É que ela não leva em conta a ordem de grandeza do valor que está sendo examinado, por exemplo:

1) Erro de *1 cm* na medição de um parafuso
2) Erro de *1 cm* na medição de uma ponte

Uma questão fundamental é a quantificação dos erros que estamos sujeitos ao computar a solução de um dado problema. Para tanto, precisamos definir medidas de erros (ou de *exatidão*). As medidas de erro mais utilizadas são o **erro absoluto** e o **erro relativo**.

1.4) Noções de aritmética de máquina: *Erros*

Definição: Seja x um número real e $\bar{x}$ sua aproximação. O **erro absoluto** da aproximação $\bar{x}$ é definido como:

$$|x-\bar{x}|$$

O **erro relativo** da aproximação $\bar{x}$ é definida como:

$$\frac{|x-\bar{x}|}{|x|} \quad \text{com} \quad |x| \neq 0$$

Essa grandeza adimensional e independente de escalas indica quão grande é o erro em relação à solução exata !!

Observe que o **erro relativo** é adimensional e, muitas vezes, é dado em porcentagem. Mais precisamente, o erro relativo em porcentagem da aproximação $\bar{x}$ é dado por:

$$\frac{|x-\bar{x}|}{|x|}\times 100\%$$

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

Campus AGRESTE

1.4) Noções de aritmética de máquina: *Erros*

Definição: Seja x um número real e $\bar{x}$ sua aproximação. O **erro absoluto** da aproximação $\bar{x}$ é definido como:

$$|x-\bar{x}|$$

O **erro relativo** da aproximação $\bar{x}$ é definida como:

$$\frac{|x-\bar{x}|}{|x|} \quad \text{com} \quad |x| \neq 0$$

Os erros definidos acima não podem ser de fato determinados em problemas cuja solução requer o uso de métodos numéricos, já que a solução verdadeira não é conhecida. Contudo, essas grandezas podem ser úteis na verificação da precisão de diferentes métodos numéricos. Isso é feito com o emprego do método numérico na solução de problemas que têm solução analítica, avaliando-se com isso os erros reais.

VER arquivo COBEM 2007 !!

Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

Campus AGRESTE

1.4) **Noções de aritmética de máquina:** *Erros*

EXEMPLO 7: erro absoluto e relativo

Sejam o valor real x e sua aproximação $\overline{x}$. Determinar o erro absoluto e relativo para os seguintes casos:

caso	x	$\overline{x}$
(a)	$123456,789$	123000
(b)	$1,2345678$	$1,13$

1.5) **Noções de aritmética de máquina:** *Cancelamento catastrófico*

EXEMPLO 8: cancelamento catastrófico

Considere o problema de encontrar as raízes da equação de segundo grau:

$$x^2 + 300\,x - 0,014 = 0$$

Usando seis dígitos significativos.

1.5) Noções de aritmética de máquina: *Aritmética de ponto flutuante e sua representação*

As leis que regem as operações aritméticas, quando executadas em um computador, não obedecem à mesma estrutura dos números reais. Isso acontece porque uma máquina digital tem recursos finitos.

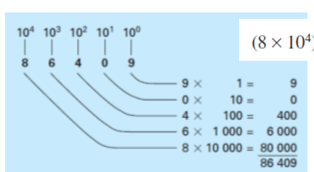
Os erros numéricos de arredondamento estão diretamente relacionados à maneira como os números são armazenados no computador. A unidade fundamental na qual a informação é representada é chamada de *palavra*, entidade, que consiste em uma sequencia de dígitos binários (*binary digits*), ou *bits*. Os números são tipicamente armazenados um uma ou mais *palavras*.

1.5) Noções de aritmética de máquina: *Sistemas numéricos*

SISTEMAS NUMÉRICOS

Um *sistema numérico* é meramente uma convenção para representar quantidades. Usualmente, utilizamos o ***sistema de numeração decimal*** ou o ***sistema numérico na base 10*** para representar números. Esse é um sistema de numeração posicional onde a posição do dígito indica a potência de 10 que o dígito está representando. A base é o número usado como referência para construir o sistema. O sistema na base 10 usa os 10 algarismos (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) para representar os números.

Para quantidades maiores, combinações desses algarismos básicos dever ser usadas, como a posição ou o valor do lugar especificando a ordem de grandeza. Por exemplo: no número **86409** há *oito grupos de 10000*, *seis grupos de 1000*, *quatro grupos de 100*, *zeros grupos de 10* e *nove unidades* a mais, ou seja:

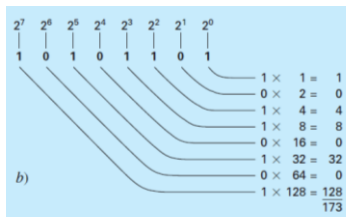


$$(8 \times 10^4) + (6 \times 10^3) + (4 \times 10^2) + (0 \times 10^1) + (9 \times 10^0) = 86\,409$$

Notação posicional !!

1.5) Noções de aritmética de máquina: *Sistemas numéricos*

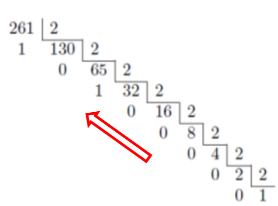
O ***sistema de numeração em base dois*** é chamado de ***binário*** e os algarismos binários são conhecidos como *bits*, do inglês *binary digits*. Um *bit* pode assumir dois valores distintos: **0** ou **1**. Da mesma forma como no sistema decimal, as quantidades podem ser representadas usando-se a notação posicional. Por exemplo, o número binário **10101101** é equivalente a **173** no *sistema decimal*.



1.5) Noções de aritmética de máquina: *Sistemas numéricos*

O ***sistema de numeração em base dois*** é chamado de ***binário*** e os algarismos binários são conhecidos como *bits*, do inglês *binary digits*. Um *bit* pode assumir dois valores distintos: **0** ou **1**. Da mesma forma como no sistema decimal, as quantidades podem ser representadas usando-se a notação posicional. Por exemplo, o número binário **10101101** é equivalente a **173** no *sistema decimal*.

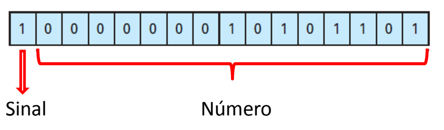
Para fazer as conversão do sistema decimal para a representação binária fazemos a divisão mostrada abaixo, considerando o número binário através dos restos das divisões escritos na ordem inversa da sua obtenção,. Por exemplo o número decimal **261** é equivalente a **100000101** no sistema binário.



Observe que com o método descrito não conseguimos descrever os números inteiros negativos !! A abordagem mais direta, chamada de *método dos valores com sinal*, utilizando o primeiro bit de uma palavra para indicar o sinal:

- **0** → número positivo (**+**)
- **1** → número negativo (**–**)

Os bits restantes são usados para armazenar o número. Por exemplo o número **– 173** em um computador de ***16 bits*** fica:



Nos computadores convencionais, emprega-se a *técnica do complemento de 2* para incorporar diretamente o sinal no valor absoluto do número em vez de fornecer um bit separado para representar mais ou menos.

EXEMPLO 9: intervalo de inteiros

Determinar o intervalo dos inteiros na base 10 que pode ser representado em um computador de 16 bits.

Binary (Base 2)	Decimal (Base 10)	Hexadecimal (Base 16)
0000	0	0
0001	1	1
0010	2	2
0011	3	3
0100	4	4
0101	5	5
0110	6	6
0111	7	7
1000	8	8
1001	9	9
1010	10	A
1011	11	B
1100	12	C
1101	13	D
1110	14	E
1111	15	F

EXEMPLO 10: sistemas de representação

Determinar no sistema decimal os números representados nos sistemas binário, quaternário e hexadecimal.

$$\begin{aligned}(1001{,}101)_2 &= (x_1)_{10}\\(301{,}2)_4 &= (x_2)_{10}\\(E2AC)_{16} &= (x_3)_{10}\end{aligned}$$

$$\begin{aligned}\text{sistema binário (base 2)} &= \{0,1\}\\\text{sistema quaternário (base 4)} &= \{0,1,2,3\}\\\text{sistema hexadecimal (base 16)} &= \{0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F\}\end{aligned}$$

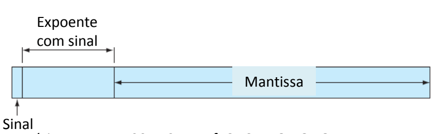
As quantidades fracionários são representadas tipicamente em computadores usando-se a forma de ***ponto flutuante***. Nessa abordagem, o número é expresso como uma parte fracionária, chamada de **mantissa** (*m*) ou significando, e uma parte inteira, chamada de **expoente** (*e*) ou **característica**, como em:

$$m \cdot b^e$$

onde ***m*** é a **mantissa**, ***b*** é a <u>*base do sistema numérico*</u> que está sendo usando e ***e*** é o **expoente**. Por exemplo:

$$156{,}78 \implies 0{,}15678 \times 10^3$$

A figura mostra uma forma na qual um número em ponto flutuante poderia ser armazenado em uma palavra. O primeiro bit fica reservado para o sinal, a próxima série de bits, para o **expoente** com sinal; e os últimos bits, para a **mantissa**.



A **mantissa** está usualmente *normalizada* se ela tiver o algarismo dominante nulo. Por exemplo:

$$\frac{1}{34} = 0,0294117647...$$

1) Sistema na base 10
2) Ponto Flutuante
3) Quatro casas decimais

$$0,0294 \times 10^{0}$$

$$0,0294 \times 10^{0}$$

Normalizamos para remover o zero dominante

$$0,2941 \times 10^{-1}$$

A consequência da normalização é que o valor absoluto de m é limitado. Isto é,

$$\frac{1}{b} \le m < 1$$

onde b é a base. Por exemplo, para um sistema na base 10, ***m*** iria variar entre 0,1 e 1, e para um sistema na base 2, entre 0,5 e 1.

1.5) Noções de aritmética de máquina: *Sistemas numéricos – Representação em ponto flutuante*

A representação em ponto flutuante permite que tanto frações quanto números muito grandes sejam expressos em um computador.

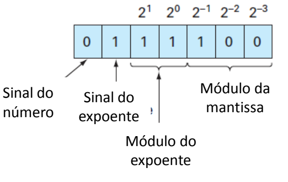
Números em ponto flutuante

1) Ocupam mais espaço e levam mais tempo para processar que os números inteiros,
2) Erro de arredondamento (mantissa com um número finito de algarismos significativos).

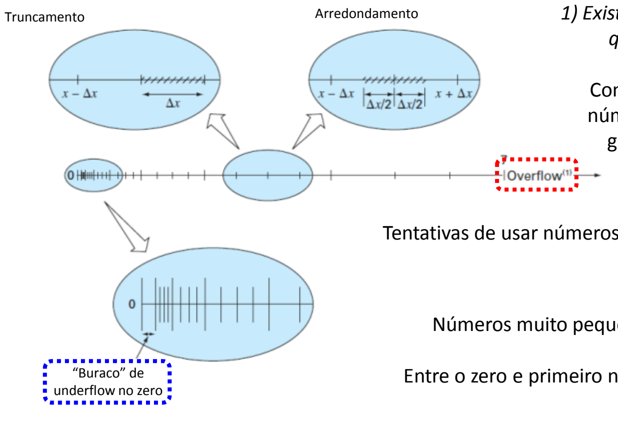
Cálculo Numérico – UFPE – 2017.2 – Prof. GUSTAVO BONO – UFPE

EXEMPLO 11: ponto flutuante

Crie um conjunto de números hipotéticos em ponto flutuante para uma máquina que armazena informação usando palavras de **7 bits**. Use o primeiro bit para o sinal do número, os próximos três para o sinal e o módulo do expoente, e os três últimos para o módulo da mantissa.



EXEMPLO 11: ponto flutuante

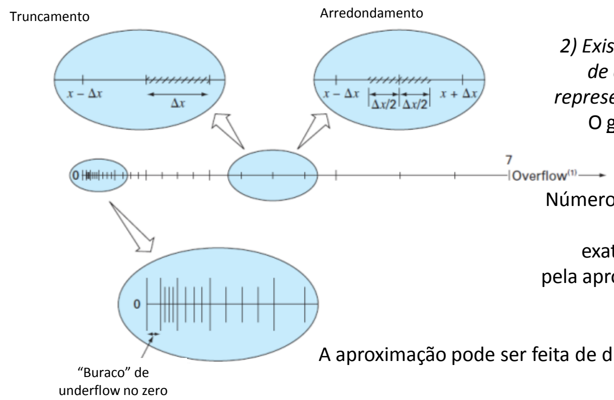


1) Existe um intervalo limitado de quantidades que podem ser representadas.
Como no caso dos inteiros, há números positivos e negativos grandes que não podem ser representados.

Tentativas de usar números fora do intervalor → ***erro de OVERFLOW***

Números muito pequenos tampouco podem ser representados.
Entre o zero e primeiro número positivo → ***erro de UNDERFLOW***

EXEMPLO 11: ponto flutuante



2) Existe apenas um número finito de quantidades que podem ser representadas dentro do intervalo. O grau de precisão (algarismos significativos) é limitado !!

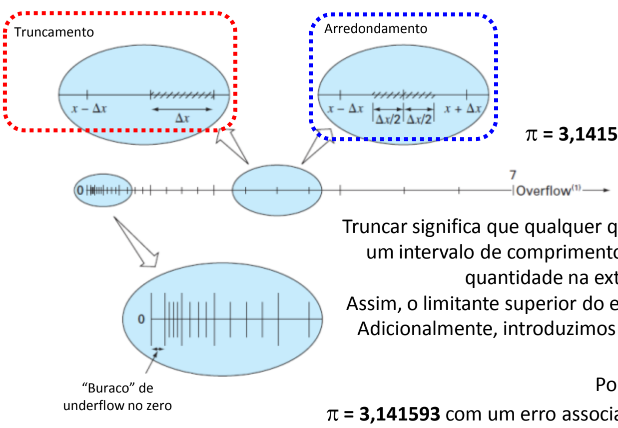
Números irracionais e racionais não podem ser representados exatamente. O erro introduzido pela aproximação denomina-se ***erro de quantização***.

A aproximação pode ser feita de duas maneiras: **truncamento** ou **arredondamento**.

Por exemplo, suponha que queremos armazenar o valor de π = 3,14159265358... em um sistema numérico na *base 10* com *sete algarismos significativos*.

1.5) Noções de aritmética de máquina: *Sistemas numéricos – Representação em ponto flutuante*

EXEMPLO 11: ponto flutuante



Por **truncamento** temos:

π = **3,141592** com um erro associado de **E_t = 0,00000065358....**

Truncar significa que qualquer quantidade que caia dentro de um intervalo de comprimento Δx será armazenada como a quantidade na extremidade inicial do intervalo. Assim, o limitante superior do erro para o truncamento é Δx. Adicionalmente, introduzimos um erro sistemático positivo.

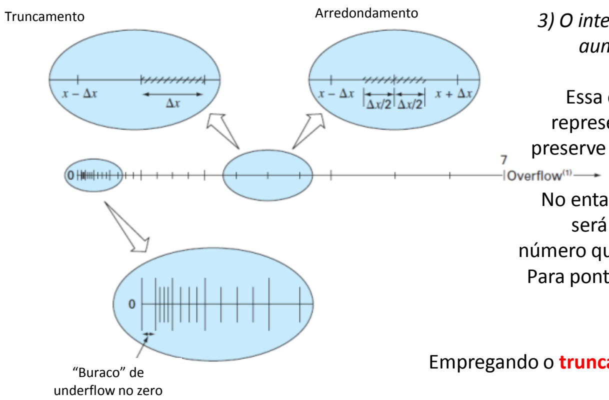
Por **arredondamento**, obtemos:

π = **3,141593** com um erro associado de **E_t = 0,00000035358....**

Arredondar significa que qualquer quantidade que caia dentro de um intervalo de comprimento Δx será representada pelo número permitido mais próximo.
Então, o limitante superior do erro para o arredondamento é $\Delta x / 2$.
Adicionalmente, nenhum erro sistemático é introduzido (erros + e -).

1.5) **Noções de aritmética de máquina:** *Sistemas numéricos – Representação em ponto flutuante*

EXEMPLO 11: ponto flutuante



3) O intervalo entre os números, Δx, aumenta quando o módulo dos números cresce.
Essa característica permite que a representação em ponto flutuante preserve os algarismos significativos.

No entanto, o erro de quantificação será proporcional ao módulo do número que está sendo representado. Para pontos flutuantes normalizados, temos:

Empregando o **truncamento**, $\frac{|\Delta x|}{|x|} \leq \mathscr{E}$

e para o **arredondamento**, $\frac{|\Delta x|}{|x|} \leq \frac{\mathscr{E}}{2}$

Onde o ***épsilon da máquina*** pode ser calculado como:

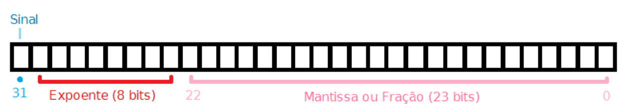
$$\mathscr{E} = b^{1-t}$$

1.5) **Noções de aritmética de máquina:** *Ponto flutuante – norma IEEE 754*

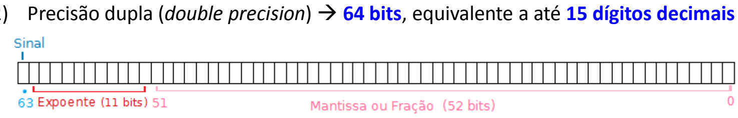
O padrão IEEE 754 (definido pelo *Institute of Electrical and Electronics Engineers*) foi adotado em 1985 e desde então passou por algumas modificações, e define algumas regras de normalização a serem seguidas nas operações e representações de números binários com **ponto flutuante**. Antes disso, cada fabricante de computadores e outros dispositivos, possuía um formato de representação diferente.

O padrão IEEE define a precisão da representação numérica como:

1) Precisão simple (*single precision*) → **32 bits**, equivalente a até **7 dígitos decimais** !!

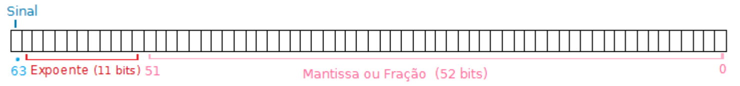


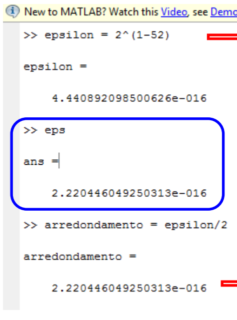
2) Precisão dupla (*double precision*) → **64 bits**, equivalente a até **15 dígitos decimais** !!



1.5) Noções de aritmética de máquina: *Ponto flutuante – norma IEEE 754*

Considerando precisão dupla o *épsilon da máquina* vale:



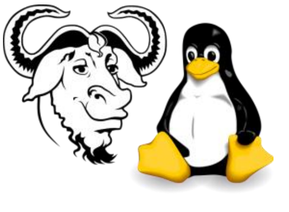


$$\mathscr{E} = b^{1-t}$$

$$\frac{|\Delta x|}{|x|} \leq \frac{\mathscr{E}}{2}$$

Campus AGRESTE

1.6) Erros com implementações / algoritmos



Campus AGRESTE

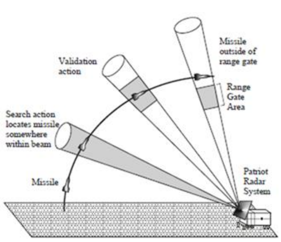
1.6) Erros com implementações / algoritmos

a) Falha na plataforma offshore Sleipnir (23/08/1991)

- A plataforma afunda devido a uma falha em uma parede, resultando em uma rachadura e vazamento.
- Motivo: combinação de erros no programa de análise pelo MEF (NASTRAN), que subestimou (47 %) a tensão na parede.

1.6) Erros com implementações / algoritmos

b) Falha na bateria de mísseis Patriot (25/01/1991)



- Na Guera do Golfo, um míssil *Scud* (M=5, 1700m/s) acertou o acampamento americano, matando 28 pessoas e ferindo centenas.
- O tempo era medido em décimos de segundo (1/10).

$$\frac{1}{10} = \frac{1}{2^4} + \frac{1}{2^5} + \frac{1}{2^8} + \frac{1}{2^9} + \frac{1}{2^{12}} + \frac{1}{2^{13}} + \ldots$$

$$\frac{1}{10} \to 0.0001100110011001100110011001100\ldots$$

$$\frac{1}{10} \to 0.00011001100110011001100$$

$$\text{error} \quad 0.000000000000000000000000011001100\ldots$$

- Motivo: O valor (1/10) ao ser representado em binário não termina. Acúmulo do erro no software após longo tempo do sistema rodando levou à falha.

1.6) Erros com implementações / algoritmos

b) Falha na Ariane 5 (04/06/1996)

- O foguete ARIANE 5 da ESA explode a uma altitude de aproximadamente 3700 m (37s) após o lançamento na Guiana Francesa.
- Em torno de U$ 7 bilhões foram investidos no desenvolvimento. O foguete e a carga foram avaliados em U$ 500 milhões.
- O programa do sistema de referencia inercial tinha um erro.
- Motivo: um número de 64 bits de ponto flutuante era convertido em um inteiro de 16 bits com sinal. Falha na conversão para números maiores que 32767 (maior inteiro representado com 16 bits).

Dúvidas ??

Obrigado

Críticas e sugestões serão bem-vindas, pois assim poderão ser melhoradas as aulas/slides.
bonogustavo@gmail.com